Université de Lille

Thèse

pour obtenir le grade de :

Docteur

En

Spécialité: Traitement du Signal et des Images

Par

Martin Rouault

Titre de la thèse

Méthodes de Monte Carlo avec distributions de Gibbs

Monte Carlo methods with Gibbs distributions

Thèse soutenue le 19 mars 2026 devant le jury d'examen :

Présidente	Alice Guionnet	DR CNRS, UMPA, ENS Lyon
Rapporteur	Djalil Chafaï	Professeur des universités, ENS Paris
Rapporteur	Nicolas Chopin	Professeur, ENSAE, IP Paris
Examinateur	Carlos Beltrán	Full Professor, Universidad de Cantabria
Examinatrice	Anna Korba	Professeure assistante, ENSAE, IP Paris
Directeur de thèse	Rémi Bardenet	DR CNRS, CRIStAL, Université de Lille
Co-directrice de thèse	Mylène Maïda	Professeure des universités, Université de Lille

Thèse préparée dans le laboratoire CRIStAL

Ecole Doctorale MADIS-631

Université de Lille — CRIStAL Centre de Recherche en Informatique, Signal et Automatique de Lille — cnrs — centralelille — Laboratoire Paul Painlevé

"Il est quelquefois sans inconvénient de remettre à plus tard son travail. Mais, s'il s'agit des baobabs, c'est toujours une catastrophe."

Le Petit Prince, Antoine de Saint–Exupéry

Remerciements

Cette thèse n'est rien sans toutes les personnes qui ont illuminé mon quotidien, qui m'ont tant donné et inspiré durant ces trois années. Au moment d'écrire ces remerciements, je suis partagé entre la joie du travail accompli, et la tristesse de laisser derrière moi ces personnes que j'aime. Puissent ces quelques lignes, les plus importantes de ce manuscrit, leur rendre justice.

Tout d'abord, je souhaite remercier Nicolas et Djalil d'avoir accepté de rapporter cette thèse, tâche souvent ingrate. Nicolas, merci d'avoir guidé mes premiers pas dans le monde des méthodes de Monte Carlo, d'abord sur les bancs de l'ENSAE, puis en stage. Merci de m'avoir transmis dès le début une vision enthousiaste et humaine de la recherche. J'en profite également pour remercier Sébastien Da Veiga qui a piqué ma curiosité en prononçant pour la première fois le doux acronyme MCMC à mes oreilles lors d'une discussion en stage de première année. Djalil, merci pour ton suivi et tes conseils depuis le début de cette thèse. Cette dernière n'aurait pas vu le jour sans ta production scientifique antérieure.

Merci également à Alice et Anna de m'avoir fait l'honneur d'intégrer ce jury de thèse, vos travaux respectifs sont source d'inspiration. Gracias Carlos por ser parte de mi tribunal, es un honor poder exponer mi tesis delante de ti.

Rémi et Mylène, merci pour votre encadrement chaleureux, pour la liberté et la précieuse confiance que vous m'avez accordées. Merci d'avoir été des pilliers infaillibles tout au long de ce numéro d'équilibriste, d'avoir cru en moi sans relâche lorsque je ramais et d'avoir su me remotiver lorsque le doute s'immisçait. Merci pour votre patience face à mon absurde et ma ponctualité hasardeuse ! Rémi, tu m'as dit au début de la rédaction qu'une bonne thèse est une thèse derrière laquelle on peut prendre l'aspiration : j'espère avoir su rendre hommage au peloton ! Merci pour ta positivité, ton humour et ton enthousiasme contagieux en toute circonstance, scientifique ou autre. Mylène, désolé de t'avoir assomée d'acronymes obscurs et de commandes Git ! Merci pour ta bienveillance et ta protection constante, pour tes nombreuses relectures sur les parties les plus techniques, et de m'avoir entraîné vers ce crossover insoupçonné à partir de mes cours de matrices aléatoires à Orsay ! Merci à tous les deux de m'avoir tant appris et inspiré humainement et scientifiquement; merci d'avoir rendu cette aventure amusante.

Je mesure aussi la chance qui est la mienne d'avoir pu grandir au chaud dans un environnement de travail si stimulant et agréable. Merci aux membres de l'équipe SIgMA et autres Cristaux pour votre énergie au quotidien, des bêtises toujours pertinentes de pause café aux journées d'équipe (19 colonels tout de même !), sans oublier bien sûr les séances olympiques de bad et de course à pied. Merci en particulier à Barbara, ma grande soeur de thèse Diala, Michaël, Juan, Alex, Thibaut, Pierre P., Stéphane, Jenny experte malgré toi de Notilus, Jan, Yigit, Jérémie, Vincent, Christelle, Thong, Aurélien, Charles, Oriane, Nicolas T., Hugo G., Jean, Maxence, Pierre, Patrick (à quand le championnat du monde de rameur indoor ?), les frères Popov Louis et Pierre-Antoine, Gaël, Eve, Ludovic pour les conseils running et l'ED, et à Sonia pour son sang froid

face à mes réservations de dernière minute sur Notilus malgré mon golden badge ! Désolé d'avoir malmené le cluster Zeus, c'était pour la science ! Merci aux aficionados de processus ponctuels et thèmes connexes du GT-PP et de l'équipe proba-stat, en particulier David C., Clément, Serge, Guillaume, Adèle, Tom, Benoit, Nicolas C., Adam, Stephan pour les cours d'info Q, Laurence, Alexandre, Chi, David D. (mention spéciale pour les cours de raquette) : les DPPs ont un avenir radieux ! Merci à Raphaël (maitre flan et actionnaire principal de la Pazienza) pour toutes les journées au MEGA, les légendaires TPs yaourts et ton implication auprès des "jeunes".

Merci à la communauté MEGA pour le vent de fraîcheur des journées à l'IHP ou à Toulouse, en particulier Charlie DG, Hugo, Thomas, Younes, Ella, Sandrine et Michel. Merci à toutes les belles rencontres en conférence, trop nombreuses pour être listées, qui ont parsemées cette thèse : des GRETSI aux GeoSto en passant par Saint-Flour (en particulier Fu-Hsuan !) et le CIRM des Bayésiens (sauf le scorpion de la salle de bains, je te tiens à l'œil...). Merci à Barbara (encore !!), Slim, Manon Michel et Pierre Jacob pour leurs précieux conseils en conférence.

Je suis également tellement redevable envers les amitiés si naturelles qui ont germé pendant cette thèse, et qui font mon bonheur ! Merci à Nico (dans tous les coups fourrés même sous botte de sept lieues !), Vlad (co-explorateur des restaus Lillois, respect pour m'avoir supporté tout ce temps en co-bureau), Thomas le goat même en version aigri, coach et mécano Etienne, Rony (s/o pour le template !), François 1er king of pedantix, Mike, Maxime, Juliette, Markus, Rafaël (on y croit pour le permis !), Chris, Keenan, Emma (Verstappen du pilotage de thèse !), Antoine, Cyriaque et Quentin Nadal. Merci de m'avoir apprivoisé et tant appris, merci pour les rires, la tendresse, le soutien, et toutes les bêtises si importantes. Merci pour le soleil que vous avez apporté dans ma vie au cours de ces trois années, et pour les aventures en tous genres passées ou à venir ! Comme dirait maître Gims : "je garde espoir de te revoir, dans un proche avenir" !

Merci aux indéboulonnables Amiénois pour votre fidélité : Victor, Fabien, Théo, Quentin, Pierrick, Amélie, Damien, Jeanne, Lucie, Tobias, et aux Toulousaings de longue date Gauthier, Erwan et Dylan; la vieillesse ennemie et la distance n'auront pas su ternir la joie des moments passés ensembles (poète...). Merci aussi à la team StatML Hugo, Farid, Julien, Jacob, Paul, à Maël, Léo, Robin, Solène et tous les autres pour les bouts de chemin accomplis ensembles entre Tours et Saclay ! Merci à Lucas pour les sorties vélos (quand les freins fonctionnent !), et aussi à Matthieu, Adri et la tribu IMT.

Enfin, merci à ma famille : mon père, ma mère, mes grands-parents, Inés, Jules premier du nom, Inés et Valentín. Merci pour votre soutien inconditionnel et pour votre amour malgré mes failles et mes contradictions, merci d'être lumière quand toutes les autres s'éteignent.

Finalmente, gracias a mi familia : mi padre, mi madre, mis abuelos, Inés, Jules primero, Inés y Valentín. Gracias por vuestro apoyo y amor sin condición, gracias por ser luz cuando las demás se apagan.

Table of contents

Résumé court en français 10

Abstract 11

1 Introduction 13

1.1 Motivations 14

1.2 Overview of the contributions 16

1.2.1 Monte Carlo with Gibbs measures 17

1.2.2 Approximation of the kernel embedding 20

1.3 Manuscript organization 22

2 A tour of numerical integration methods 23

2.1 Introduction 24

2.1.1 Quantifying integration errors 24

2.1.2 Lower bounds 30

2.2 Two crude quadratures 33

2.2.1 Gaussian quadrature 33

2.2.2 Independent sampling 34

2.3 Markov chain Monte Carlo 37

2.3.1 Metropolis–Hastings algorithms 38

2.3.2 Bias of MCMC algorithms 40

2.3.3 Asymptotic variance and integration guarantees 41

2.4 Quasi Monte Carlo 43

2.4.1 Low-discrepancy sequences 43

2.4.2 Randomized QMC 45

2.5 Variational Bayes 48

2.5.1 The variational approximation 48

2.5.2 Stein Variational Gradient Descent 50

2.6 Kernel methods 52
2.6.1 Energy minimization and kernel herding . . . 53
2.6.2 Energy minimization with Stein discrepancies . . . 54
2.6.3 Kernel quadrature and volume sampling . . . 56
2.6.4 Thinning algorithms . . . 57
2.7 Conclusion 62
3 Repulsive point processes for quadrature . . . 64
3.1 Generalities on point processes 65
3.1.1 Definitions . . . 65
3.1.2 Negative dependence in point processes . . . 67
3.2 Some examples 71
3.2.1 Binomial and Poisson point processes . . . 71
3.2.2 (Perturbed) Lattices . . . 72
3.2.3 Eigenvalues of random matrices . . . 73
3.2.4 Determinantal point processes . . . 75
3.3 Gibbs measures 77
3.3.1 Motivations and definitions . . . 77
3.3.2 Large deviations and concentration . . . 79
3.3.3 Gaussian fluctuations in low dimensions . . . 82
3.4 Conclusion 83
4 Monte Carlo with Gibbs measures . . . 84
4.1 Introduction 85
4.2 Related work 85
4.2.1 Concentration for independent sampling and MCMC . . . 85
4.2.2 Concentration for kernel methods and Coulomb gases . . . 86
4.3 A concentration inequality for Gibbs measures 87
4.3.1 Definitions . . . 87
4.3.2 Main results . . . 89
4.3.3 Application to Monte Carlo integration . . . 90

4.4 Numerical experiments 95
4.4.1 Sampling from the Gibbs measure . . . 95
4.4.2 Coverage of Monte Carlo integration . . . 102
4.4.3 Worst-case integration errors . . . 104
4.4.4 Variance of linear statistics . . . 106
4.5 Proofs 107
4.5.1 Proof of Proposition 4.3.5 . . . 107
4.5.2 Proof of Proposition 4.3.10 . . . 108
4.5.3 Proof of Theorem 4.3.7 . . . 109
4.5.4 Proof of Proposition 4.3.13 . . . 112
4.6 Conclusion 112
5 Approximation of the kernel embedding . . . 114
5.1 Introduction 115
5.2 Related work 117
5.2.1 Monte Carlo integration and KSDs . . . 117
5.2.2 Large deviations . . . 117
5.2.3 Extrema of the potential of a Coulomb gas . . . 118
5.3 Definitions and main results 119
5.4 Experiments 127
5.4.1 Approximately sampling from the Gibbs measure . . . 128
5.4.2 Bayesian classification . . . 130
5.4.3 Worst-case integration error and variance . . . 132
5.5 Proofs: quenched large deviations 135
5.5.1 Reminder on LDPs for Gibbs measures, after (García-Zelada, 2019) 136
5.5.2 Proof of Theorem 5.3.7 (non-singular kernels) . . . 139
5.5.3 Proof of Theorem 5.3.13 (Coulomb kernel) . . . 140
5.6 Proofs: uniform convergence for the Coulomb potential field 141
5.6.1 Proof of Proposition 5.3.14 . . . 142
5.6.2 Proof of Proposition 5.3.15 . . . 144

5.7 Proof of auxiliary lemmas 148

5.7.1 Preliminaries 148
5.7.2 Auxiliary lemmas for Proposition 5.3.14 151
5.7.3 Auxiliary proofs for Proposition 5.3.15 155

5.8 Conclusion 157

6 Conclusion and further work 159

6.1 Further work 160

6.1.1 Investigating background quality 160
6.1.2 A fast CLT for bounded kernels 163
6.1.3 Sampling from Gibbs distributions 167

6.2 Openings 169

6.2.1 Smale's 7th problem 169
6.2.2 Integration in RKHSs 170
6.2.3 Gibbs measures and parallel MCMC 170
6.2.4 Mean-field theory for neural networks 172

Résumé court en français

Méthodes de Monte Carlo avec distributions de Gibbs

Les algorithmes Monte Carlo par chaînes de Markov (MCMC) sont des algorithmes d'intégration numérique en grande dimension utiles pour l'inférence bayésienne. Le principle est d'échantillonner une chaine de Markov bien choisie dans le domaine d'intégration, et on calcule la moyenne de l'intégrande sur les points de la chaine. Cependant, les méthodes MCMC sont lentes : elles produisent des estimateurs dont l'erreur quadratique moyenne décroit au mieux en $1/n$, où n est le nombre d'étapes dans l'échantillon de la chaine de Markov. Lorsque la fonction à intégrer est couteuse à évaluer, il est alors pertinent de dépenser davantage de budget computationnel dans la construction d'un estimateur à l'erreur quadratique plus petite pour le même nombre n d'évaluations de l'intégrande. C'est l'objectif que l'on s'est fixé dans cette thèse.

Suivant l'intuition que, dans un ensemble de points aléatoires, la répulsion amène de la réduction de variance, nous proposons dans cette thèse de choisir les nœuds d'intégration selon une distribution de Gibbs, inspirée de la physique statistique. Dans un premier temps, nous étudions les propriétés de convergence de l'estimateur associé pour l'intégration numérique. En particulier, nous montrons une inégalité de concentration pour de telles distributions de Gibbs avec une interaction par paires générale mais bornée. À niveau de précision fixé sur l'erreur d'intégration, cette dernière implique qu'un plus petit nombre de points n est nécessaire pour obtenir les mêmes régions de confiance que pour l'estimateur MCMC. Nous illustrons numériquement ces résultats en échantillonnant approximativement ces mesures de Gibbs dans des exemples-jouets.

Simuler des points de façon approchée selon cette distribution de Gibbs requiert néanmoins d'être en capacité de calculer le potentiel d'interaction de la mesure cible à intégrer. Dans un second temps, nous construisons une procédure algorithmique en deux étapes incluant une approximation empirique de ce potentiel. Si l'approximation est naturelle, l'analyse mathématique de cette nouvelle procédure est délicate. Nous parvenons à montrer que les garanties obtenues précédemment sont conservées, du moins asymptotiquement, en établissant un principe conditionnel de grandes déviations incluant cette approximation.

Mots clés - Monte Carlo, Intégration numérique, Distributions de Gibbs.

Abstract

Monte Carlo methods with Gibbs distributions

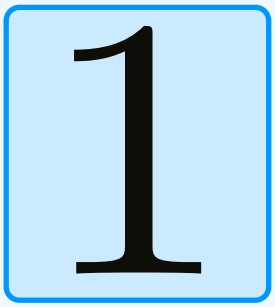
Markov chain Monte Carlo algorithms (MCMC) are numerical integration algorithms that are ubiquitous in high-dimensional statistical inference and Bayesian machine learning. The crux is to sample a carefully-chosen Markov chain in the domain of integration, and average the evaluations of the integrand along that chain. However, MCMC methods are slow: the resulting estimators have a mean squared error that decreases as $1/n$, where n is the number of time steps in the Markov chain. When the integrand function is expensive to evaluate, we prefer to spend more computational budget on designing an estimator with smaller quadratic error for the same number n of integrand evaluations. This is the goal of this thesis.

Following the intuition that repulsiveness within a set of integration nodes brings variance reduction, we propose in this thesis to choose the n nodes according to a Gibbs distribution coming from statistical physics. In a first part, we study the convergence properties of a natural associated estimator for numerical integration. To do so, we prove a fast concentration inequality for such Gibbs distributions with general but bounded pairwise interaction. Given a fixed precision level on the integration error, this means that a smaller number of points n is required to obtain the same confidence regions as MCMC. We empirically illustrate these results by approximately sampling from such Gibbs measures on toy examples.

Approximately sampling points from this Gibbs distribution however requires computing the so-called interaction potential of the target distribution. In a second part, we thus consider a two-step algorithmic procedure relying on an approximation of this potential. While this approximation is natural, the mathematical analysis of this new procedure is delicate. We are able to show that the guarantees previously obtained still hold, at least asymptotically, by establishing a large deviation principle that includes this approximation.

Keywords - Monte Carlo, Numerical integration, Gibbs distributions.

Introduction



"Va Numérobis, construis ce palais ! Tu as trois mois."

Cléopâtre

1.1 Motivations 14

1.2 Overview of the contributions 16

1.2.1 Monte Carlo with Gibbs measures 17

1.2.2 Approximation of the kernel embedding 20

1.3 Manuscript organization 22

1.1 Motivations

In the context of this thesis, we investigate numerical integration methods to compute quantities of the form

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}\pi(x), \tag{1.1}$$

where π is a probability measure on $\mathbb{R}^d$, $d \geq 1$, called the *target*, and the function to integrate f is called the *integrand*. Quantities like (1.1) are usually *intractable*, meaning that we do not know how to arrive at a satisfying closed form expression with pen and paper. The crux of numerical integration methods, also called *quadratures*, is thus to design *nodes* $x_1, \ldots, x_n \in \mathbb{R}^d$ and *weights* $w_1, \ldots, w_n \in \mathbb{R}$ such that

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}\pi(x) \approx \sum_{i=1}^{n} w_i f(x_i), \tag{1.2}$$

while being cheap in terms of computational time and memory. Relatedly, in the *sampling* problem, we are interested in producing a discrete measure of the form

$$\mu_n = \sum_{i=1}^{n} w_i \delta_{x_i}, \tag{1.3}$$

such that μ_n is close to the target distribution π in the sense that (1.2) holds uniformly over a large set of test functions f.

Numerical integration is now routine in computational statistics (Robert, 2007) and probabilistic machine learning (Murphy, 2022). We quickly highlight a few key applications that will motivate us throughout the thesis.

Posterior sampling in Bayesian inference

We assume that we observe data $y_1, \ldots, y_N \in \mathcal{Y}$. In parametric Bayesian inference (Robert, 2007), we consider a marginal distribution $p(\theta)\,\mathrm{d}\theta$ for a set of parameters $\theta \in \Theta = \mathbb{R}^d$, called the *prior* distribution. Then, we model the joint distribution of the observations $y_1, \ldots, y_N$ and θ by

$$p(y_1, \ldots, y_N, \theta) = L(y_1, \ldots, y_N; \theta)\, p(\theta), \tag{1.4}$$

where $L(y_1, \ldots, y_N; \theta)$ is the probability density function (pdf) of the observations $y_1, \ldots, y_N$ conditionally on the parameters θ, with respect to a reference measure on $\mathcal{Y}$. We will often abusively refer to $L(y_1, \ldots, y_N; \theta)$ as the *likelihood*.

Under the model (1.4), the conditional pdf of θ given the observations $y_1, \ldots, y_N$ writes

$$p(\theta | y_1, \ldots, y_N) = \frac{L(y_1, \ldots, y_N; \theta)\, p(\theta)}{\int_{\mathbb{R}^d} L(y_1, \ldots, y_N; \theta) p(\theta)\,\mathrm{d}\theta}. \tag{1.5}$$

The probability measure (1.5) is called the *posterior* distribution. It can be used to infer the parameters θ of the model given observations $y_1, \dots, y_N$, for instance by computing the mean of the posterior distribution given by

$$\mathbb{E}\left[\theta | y_1, \dots, y_N\right] = \int \theta \, \mathrm{d}p(\theta | y_1, \dots, y_N). \tag{1.6}$$

More generally, a wealth of situations require computing integrals of the form

$$\mathbb{E}\left[f(\theta) | y_1, \dots, y_N\right] = \int f(\theta) \, \mathrm{d}p(\theta | y_1, \dots, y_N), \tag{1.7}$$

for some function $f : \mathbb{R}^d \to \mathbb{R}$. For instance, for $\alpha \in (0, 1)$, Bayesian scientists are usually interested in designing a *credibility region* $C_\alpha \subset \mathbb{R}^d$, such that

$$\int \mathbb{1}_{C_\alpha}(\theta) \, \mathrm{d}p\left(\theta | y_1, \dots, y_N\right) \geq 1 - \alpha. \tag{1.8}$$

Designing such a C_α requires computing integrals of the form (1.7) for many indicator functions $f = \mathbb{1}_C$ of increasing regions C centered around the posterior mean (1.6), until we reach coverage $1 - \alpha$ in (1.8). Since quantities like (1.7) are usually intractable, quadrature rules like (1.2) come into play.

Applications of Bayesian inference are numerous, for instance in inverse problems where we wish to recover a signal from noisy observation data (Kaipio and Somersalo, 2005). These tasks are now ubiquitous in modern science, from astrophysics to imaging sciences and computational biology. Yet, these settings are challenging: the posterior (1.5) is often high-dimensional, while the normalization constant

$$\int_{\mathbb{R}^d} L(y_1, \dots, y_N; \theta) p(\theta) \, \mathrm{d}\theta \tag{1.9}$$

is unknown in most cases.

In addition, practitioners usually want to be able to quantify their uncertainty about the approximation (1.2), see for instance (Briol et al., 2019). For instance, probabilistic error controls can be given for *Monte Carlo* methods (Robert and Casella, 2004), where the points $x_1, \dots, x_n$ in (1.2) are suitably chosen at random in $\mathbb{R}^d$. In particular, Markov chain Monte Carlo (MCMC) produces points $x_1, \dots, x_n$ by iterating a well-chosen Markov kernel, and comes with *confidence intervals* of magnitude $n^{-1/2}$ in (1.2), with few restrictions on the integrand f. This rate has the impressive feature of being independent of the dimension d, but is often deemed slow in practice. We will see in Chapter 2 that this can be improved, upon restricting to smaller classes of integrands in (1.2), e.g. smooth functions in Sobolev spaces or functions belonging to a reproducing kernel Hilbert space (RKHS) of given kernel K.

Distribution compression

In distribution compression, or *thinning* (Riabiz et al., 2022; Chopin and Ducrocq, 2021), we assume that we already have access to samples $x_1, \dots, x_n$, for instance approximately drawn from a target π using MCMC. Then, the goal is to consider only a small subset of size $m < n$ of $\{x_1, \dots, x_n\}$ while preserving the same statistical guarantees as for the n initial points when it comes to approximating π.

There can be various motivations for this. For instance, when n is large, subsampling a small number m of nodes represents a save in storage cost. Second, MCMC algorithms are biased and may require many iterations t_0, called *burn-in*, to reach equilibrium. Removing the iterations of the chain before step t_0 can then help provide more accurate estimates in (1.2).

Finally, and most importantly in this thesis, the integrands f in (1.2) can be costly to evaluate, and the number of evaluations n becomes the computational bottleneck. For instance in Bayesian inference, f can contain a loss function and parts of the likelihood. Each evaluation of the likelihood may then take a few hours for population models in ecology (Purves et al., 2013). In cardiac cell biology, each evaluation of the likelihood triggers a whole-heart simulation whose cost can range from a few minutes to thousands of CPU hours (Johnstone et al., 2016; Niederer et al., 2011; Augustin et al., 2016; Strocchi et al., 2020).

In these cases, we would thus really like to keep n small. In particular, evaluating the integrand f in (1.2) only on m points rather than on the n original nodes represents a drastic save in time, when m is small compared to n. Doing so while preserving an integration error of similar magnitude in the inference pipeline is the grail of the thinning literature. This most likely requires a careful choice of the m selected nodes, and the time savings justifies spending more computational budget to pick those nodes. We will discuss examples of such choices in Chapter 2. Alternately, this also directly motivates drawing points $x_1, \dots, x_n$ with errors of smaller magnitude than MCMC with the same fixed number of points n, even if designing these new nodes requires more computational budget.

1.2 Overview of the contributions

For the ease of reading, we provide here a quick overview of the main contributions of the thesis.

We have produced three papers, one journal paper currently under review for the *SIAM journal on Mathematics of Data science*, one conference paper published in the proceedings of the *30th GRETSI Colloquium*, and another paper submitted to a journal in probability and statistics. Finally, we have a fourth paper in preparation, that will also be submitted to a journal in probability and statistics.

In this thesis, we consider π to be a compactly supported probability measure on $\mathbb{R}^d$.

We propose to draw the nodes $x_1, \dots, x_n$ in (1.2) from the *Gibbs distribution*

$$\mathbb{P}^{V^\pi}_{n,\beta_n}(x_1, \dots, x_n) = \frac{1}{Z^{V^\pi}_{n,\beta_n}} \exp\left(-\beta_n H_n(x_1, \dots, x_n)\right) \mathrm{d}x_1 \dots \mathrm{d}x_n, \tag{1.10}$$

with

$$H_n(x_1, \dots, x_n) = \frac{1}{2n^2} \sum_{i \neq j} K(x_i, x_j) + \frac{1}{n} \sum_{i=1}^{n} V^\pi(x_i), \tag{1.11}$$

where $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a function called *interaction* kernel, $V^\pi : \mathbb{R}^d \to \mathbb{R}$ is a function called *confinement potential* chosen to ensure convergence in (1.2) for integrands in the RKHS $\mathcal{H}_K$ of kernel K, namely:

$$V^\pi(z) = -\int K(z, x) \, \mathrm{d}\pi(x) + \Phi(z), \tag{1.12}$$

with $\Phi(z)$ a quadratic function outside the support of π for instance.

In (1.10), $Z^{V^\pi}_{n,\beta_n}$ is a normalization constant that makes $\mathbb{P}^{V^\pi}_{n,\beta_n}$ be a probability measure, while β_n is an *inverse temperature* parameter, a typical choice will be $\beta_n = n^2$.

1.2.1 Monte Carlo with Gibbs measures

In this section, we give an overview of the main contributions of the following papers, which will be described at length in Chapter 4 and Section 5.4.2.

- Rouault, Martin, Bardenet, Rémi, and Maïda, Mylène (Feb. 18, 2024). *Monte Carlo with kernel-based Gibbs measures: Guarantees for probabilistic herding*. arXiv: 2402.11736.
- Rouault, Martin, Bardenet, Rémi, and Maïda, Mylène (Aug. 25, 2025). "Intégrer plusieurs fonctions simultanément avec des mesures de Gibbs". In: *30° Colloque sur le traitement du signal et des images*. Issue: 2025-1751. Strasbourg: GRETSI - Groupe de Recherche en Traitement du Signal et des Images, p. 1157–1160.

The first one is currently under revision for the *SIAM journal on Mathematics of Data Science*. The second one is a short communication that was presented at the *30th GRETSI Colloquium, Strasbourg*. *GRETSI* is the French national association on signal and image processing, with its biennial meeting being the main French-speaking event in signal processing (400+ participants).

In (Rouault et al., 2024), we first introduce Monte Carlo integration based on nodes $x_1, \dots, x_n$ drawn from the joint distribution (1.10). We will see in Chapter 3 that our choice (1.10) is naturally motivated by the statistical physics literature, and can be seen as a probabilistic relaxation of kernel-based optimization methods trying to achieve minimal integration errors. The Gibbs measure (1.10) favors configurations of

nodes $x_1, \dots, x_n$ with low energy (1.11), with a strength given by β_n. The *energy* (1.11) balances two effects: a repulsive pairwise interaction that makes the nodes spread in the integration domain, and a confinement term that attracts the nodes towards the regions of large probability under the target π.

Our main result is a fast concentration inequality at speed β_n for the worst-case integration error

$$\mathrm{MMD}_K\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \pi\right) := \sup_{\|f\|_{\mathcal{H}_K}\leq 1} \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f \,\mathrm{d}\pi\right|, \tag{1.13}$$

under the assumption that K is **bounded**, where MMD stands for *maximum mean discrepancy*. In particular, Corollary 4.3.11 in this manuscript reads

$$\mathbb{P}_{n,\beta_n}^{V^\pi}\left(\mathrm{MMD}_K\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \pi\right) > r\right) \leq \exp\left(-\frac{1}{8}\beta_n r^2\right), \tag{1.14}$$

for all $n \geq 2$, as soon as

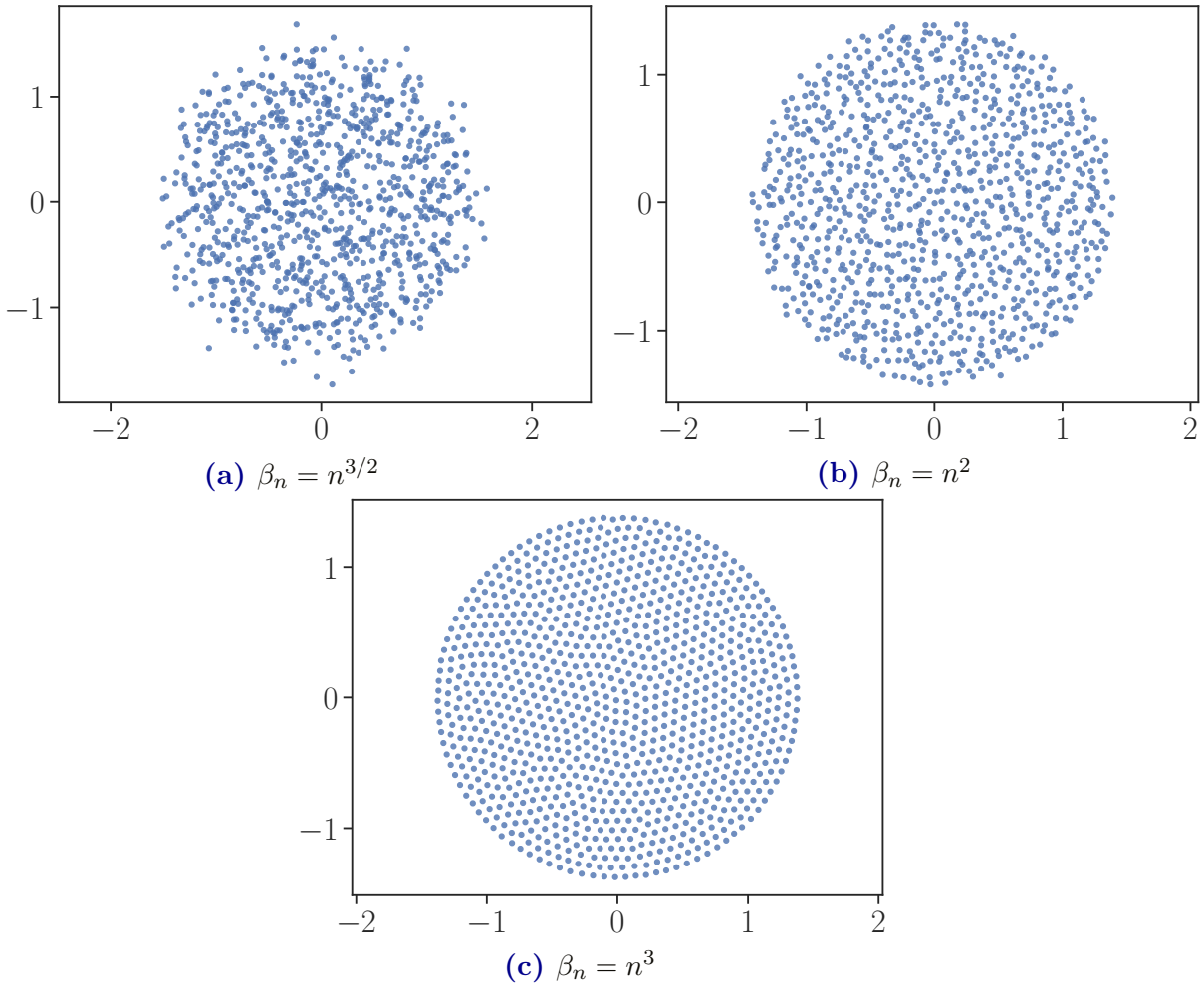
$$r \geq 4\max\left(\sqrt{c_2}n^{-1/2}, \sqrt{c_1}\,(\beta_n/n)^{-1/2}\right), \tag{1.15}$$

with c_1 and c_2 scaling at most like $\mathcal{O}(d\log d)$ in the dimension. Replacing the MMD by e.g. the bounded Lipschitz distance, similar results had already been obtained by (Chafaï et al., 2018) when K is the singular Coulomb interaction.

While not improving on the known $n^{-1/2}$ rate for the MMD obtained when using MCMC to sample from π directly, the speed β_n in the right-hand side of (1.14) is very satisfying. In fact, this is much faster than the corresponding speed n for crude MCMC, when β_n grows faster than n. This fast concentration inequality can be interpreted as follows: at fixed precision level for the worst-case integration error (1.13), our Gibbsian quadrature provides *tighter* confidence regions that crude MCMC targeting π with the same number of points n, see Corollary 4.3.12 in this manuscript. Moreover, *tighter* is quantified by how fast β_n grows compared to n.

The main computational bottlenecks and numerical experiments for our quadrature (1.10) will be presented in detail in Section 4.4.1, and we will provide a simple MCMC algorithm to approximately sample from (1.10) that enables us to validate our results on toy examples. Recall that using MCMC to sample from (1.10) is fair if our goal is to reduce the number n of evaluations of the integrand f. In Figure 1.1, we show $n = 1000$ points approximately drawn from our Gibbs distribution on a two-dimensional example, for different inverse temperature schedules β_n.

In Figure 1.2a we illustrate our Corollary 4.3.12. We consider $n = 100$ points from our Gibbs measure (1.10) for $\beta_n = n^2$, described by the blue curve in Figure 1.2a. We show that our Gibbs measure indeed provides tighter coverage than MCMC with $n = 100$



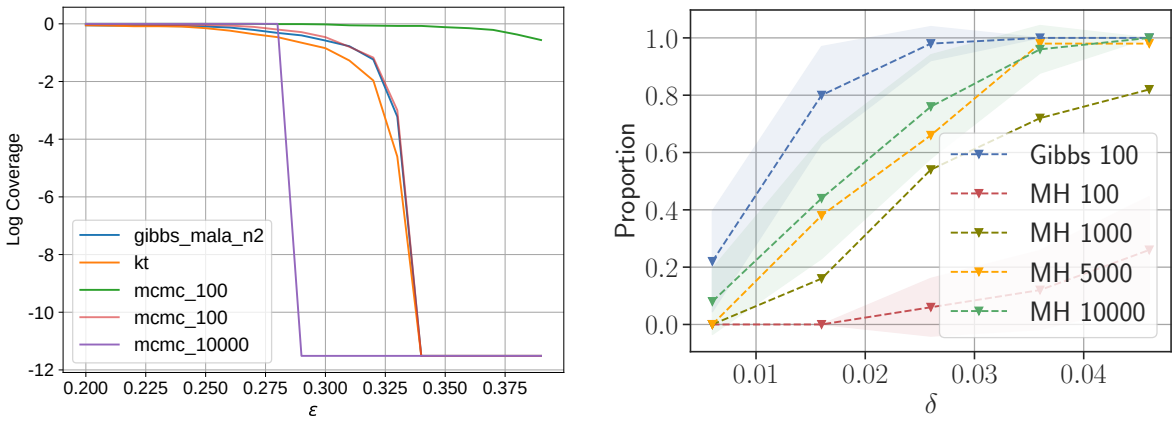
(a) $\beta_n = n^{3/2}$ **(b)** $\beta_n = n^2$

(c) $\beta_n = n^3$

Fig. 1.1: Three independent approximate samples of $\mathbb{P}^V_{n,\beta_n}$, with different temperature schedules.

points (green curve), and stands alongside state-of-the-art thinning methods such as kernel thinning (Dwivedi and Mackey, 2024) in terms of coverage (orange curve).

In (Rouault et al., 2025), we illustrated our improvements in coverage on a toy example coming from Bayesian classification which we describe in details in Section 5.4.2. In the latter, we need to approximate integrals of the form (1.7) simultaneously for M integrands f. In Figure 1.2b, we show that our estimator with $n = 100$ points performs as well as a crude MCMC estimator with $n = 10\,000$ nodes. Again, this is a toy example, but this supports the claim that quadrature with Gibbs distributions like (1.10) can help reduce the number of evaluations of the integrand f while preserving statistical properties for inference.



(a) **(b)**

Fig. 1.2: (1.2a) Coverage comparisons between $\mathbb{P}^{V^\pi}_{n,\beta_n}$, MCMC and kernel thinning, and (1.2b) Proportion of runs where the M estimators fall simultaneously in the confidence regions of size δ.

1.2.2 Approximation of the kernel embedding

In this section, we give an overview of the contributions of the following paper, which will be described in the manuscript in Chapter 5.

- Bardenet, Rémi, Maïda, Mylène, and Rouault, Martin (Aug. 2, 2025). *Quenched large deviations for Monte Carlo integration with Coulomb gases*. arXiv: 2508.01392[cs].

The latter has been submitted to a journal in probability and statistics, and the different author order compared to the papers of the previous section reflects the different publication habits of the probability and machine learning communities.

We can see from (1.12) that computing the density of (1.10) requires prior knowledge of an integral with respect to π. This seems to defeat our initial purpose, and we circumvent this issue in our numerical experiments by approximating this quantity using MCMC again. In our paper (Bardenet et al., 2025), our goal was to theoretically justify this approximation.

To do so, we introduce the approximation of (1.12) given by

$$V_n(z) = -\int K(z,x)\,\mathrm{d}\nu_n(x) + \Phi(z), \tag{1.16}$$

where ν_n is a discrete measure approximating π, in the sense that

$$\nu_n \to \pi, \tag{1.17}$$

almost surely. Conditionally on ν_n, that we call *background*, we then define the so-called

quenched Gibbs distribution by

$$\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}(x_1,\dots,x_n,\nu_n(\omega)) = \frac{1}{Z^{V_n}_{n,\beta_n}} \exp\left(-\beta_n H^{\mathrm{Q}}_n(x_1,\dots,x_n)\right) \mathrm{d}x_1 \dots \mathrm{d}x_n, \tag{1.18}$$

with

$$H^{\mathrm{Q}}_n(x_1,\dots,x_n) = \frac{1}{2n^2}\sum_{i\neq j} K(x_i,x_j) + \frac{1}{n}\sum_{i=1}^n V_n(x_i). \tag{1.19}$$

The "Q" in H^{Q}_n and $\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}$ stands for *quenched*, and (1.18) corresponds to the Gibbs measure (1.10) where we replaced π by the background ν_n in the right-hand side of (1.12).

In (Bardenet et al., 2025), we show that for almost all ω, for any $r > 0$, there exists $n_0(r)$ such that for all $n \geq n_0(r)$,

$$\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}\left(\mathrm{MMD}_K\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \pi\right) > r; \nu_n(\omega)\right) \leq \exp\left(-\frac{1}{8}\beta_n r^2\right), \tag{1.20}$$

as long as K is bounded, see Theorem 5.3.7 in this manuscript. We will prove it by relying on large deviation machinery for Gibbs measures developed in (Chafaï et al., 2014; García-Zelada, 2019).

This bound supports our approximation in the experiments of Figure 1.2. We could be positively surprised by the fact that such a bound still holds with fast speed β_n regardless of the quality of the background ν_n. Note however that (1.20) is only asymptotic as opposed to (1.14).

Actually, a crude bound (Corollary 6.1.1 in this manuscript) will suggest that when e.g. $\beta_n \geq cn^2$ and the number M_n of atoms in ν_n grows at least like n^2, we still have the fast non-asymptotic bound

$$\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}\left(\mathrm{MMD}_K\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \pi\right) > r; \nu_n(\omega)\right) \leq \exp\left(-c\beta_n r^2\right), \tag{1.21}$$

with high probability, and for any $r > cn^{-1/2}$. Yet, choosing $M_n = n^2$ makes the algorithmic procedure expensive, and we were able to obtain strong coverage with much smaller values of M_n in practice, e.g. in Figure 1.2b.

In the case of the singular Coulomb interaction

$$g(x,y) = \frac{1}{|x-y|^{d-2}}, \tag{1.22}$$

we prove an analogous result to (1.20) in Theorem 5.3.13 of this manuscript when $d \geq 3$ and with the MMD again replaced by the bounded Lipschitz distance in (1.20). For

technical reasons, we this time need ν_n to be produced by reweighting the realization of a first Gibbs measure (1.10) with interaction g and tractable confinement V, e.g. quadratic on the support of π. The proof is however much more involved in this case and the definition of (1.16) differs a bit, see Chapter 5.

Finally, we currently have another paper in preparation where our goal is to provide fast central limit theorems for points drawn from the Gibbs measure (1.10) with bounded interaction K. I believe this would be an important milestone to support the use of Gibbs measures for uncertainty quantification. We will discuss it in more detail in Chapter 6, along with remaining limitations of our approach, e.g. sampling efficiently from our Gibbs distributions.

1.3 Manuscript organization

In Chapter 2, we review the statistical guarantees of numerical integration methods existing in the literature. We focus on three main questions. First, are these guarantees good, or even optimal, and in what sense? Second, what is the computational complexity of the algorithmic procedure used to build the quadrature nodes? Third, how do these algorithms scale with the dimension d, either in the approximation error or in the computational complexity?

In Chapter 3, we give an introduction to the repulsive point processes appearing in the probability literature, especially in statistical physics and random matrix theory. Those will motivate us to define our Gibbs distributions at the end of the chapter. In particular, we emphasize that in many examples, the strong guarantees for numerical integration with those point processes are a trace of the repulsion between the points of the system.

Chapter 4 is the beating heart of this thesis. We introduce our numerical quadrature based on Gibbs measures, and state our concentration theorem. We then discuss the practical aspects of this quadrature and illustrate our results on toy examples.

In Chapter 5, our goal is to include the approximation of the kernel embedding into our statistical guarantees. To do so, we prove that the Gibbs measure based on this approximation satisfies a large deviation principle with the same speed as the original Gibbs measure, thus justifying our approximation. We discuss extensions and illustrate again our results on toy examples, and on a simple example coming from Bayesian classification.

Finally, we conclude this thesis in Chapter 6. We give a few words about ongoing work, and about the limitations that remain to be tackled. We also address the use of Gibbs distributions in some different frameworks of computational statistics and Bayesian machine learning.

A tour of numerical integration methods

"Le malheur fait bien sa propagande : indépendance, indépendance."

Clair de femme, Romain Gary

2

2.1 Introduction 24

2.1.1 Quantifying integration errors 24

2.1.2 Lower bounds 30

2.2 Two crude quadratures 33

2.2.1 Gaussian quadrature 33

2.2.2 Independent sampling 34

2.3 Markov chain Monte Carlo 37

2.3.1 Metropolis–Hastings algorithms 38

2.3.2 Bias of MCMC algorithms 40

2.3.3 Asymptotic variance and integration guarantees 41

2.4 Quasi Monte Carlo 43

2.4.1 Low-discrepancy sequences 43

2.4.2 Randomized QMC 45

2.5 Variational Bayes 48

2.5.1 The variational approximation 48

2.5.2 Stein Variational Gradient Descent 50

2.6 Kernel methods 52

2.6.1 Energy minimization and kernel herding 53

2.6.2 Energy minimization with Stein discrepancies 54

2.6.3 Kernel quadrature and volume sampling 56

2.6.4 Thinning algorithms 57

2.7 Conclusion 62

2.1 Introduction

2.1.1 Quantifying integration errors

Throughout this chapter, π is a probability distribution on $\mathbb{R}^d$ with a density with respect to the Lebesgue measure that we will also denote by π. Numerical integration methods aim at giving guarantees on the worst-case integration error

$$\mathrm{E}_{\mathcal{F}}(\nu) := \sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}(\nu - \pi) \right|, \tag{2.1}$$

obtained by replacing the target distribution π by a *tractable* distribution ν when integrating over a set of test functions $\mathcal{F}$. By *tractable*, we mean that we know how to compute the integrals of interest with respect to ν. This usually calls for approximating distributions ν with finite support and of the form $\nu_n = \sum_{i=1}^n w_i \delta_{x_i}$, with well-chosen *quadrature nodes* $x_1, \dots, x_n$ and *weights* $w_1, \dots, w_n$, in which case

$$\mathrm{E}_{\mathcal{F}}(\nu_n) = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n w_i f(x_i) - \int f \, \mathrm{d}\pi \right|. \tag{2.2}$$

We will see instances of such methods in Sections 2.2, 2.3, 2.4 and 2.6. Alternately, it is also possible to rely on approximate distributions ν with infinite support, see Section 2.5, in which case one more generally seeks to minimize a dissimilarity between the probability measures.

The choice of a set of integrands $\mathcal{F}$ has many consequences. First, the wider the class $\mathcal{F}$, the harder it is to provide a small error (2.1), since we ask for a uniform approximation over *all* the integrands in $\mathcal{F}$. We will indeed see in Section 2.1.2 that the best we can get can significantly downgrade when $\mathcal{F}$ grows. This is unfortunate, because a Bayesian scientist wishing for guarantees on approximate credibility intervals will need $\mathcal{F}$ to contain all the indicator functions over the support of π, see Chapter 1. Second, choosing $\mathcal{F}$ means choosing the functions we allow ourselves to consider in the numerical integration task, which often means how much structure or regularity we are ready to assume in the application we target. We will see in Section 2.1.2 that increasing the regularity or the structure of the functions in $\mathcal{F}$ will upgrade the guarantees we can get on the error (2.1).

As a baseline, we give a naive way to perform quadrature for a target π supported in the hypercube $[0,1]^d$. Consider the $n = m^d$ deterministic points on the grid $\left(\frac{1}{m}\mathbb{Z} \cap [0,1)\right)^d$, and consider weights $w_i = \frac{1}{n}\pi(x_i)$ in (2.2). Then, setting $g(x) = f(x)\pi(x)$, we can

decompose

$$\begin{aligned}\sum_{i=1}^n w_i f(x_i) - \int f \,\mathrm{d}\pi &= \frac{1}{n}\sum_{i=1}^n g(x_i) - \int_{[0,1]^d} g(x)\,\mathrm{d}x \\ &= \frac{1}{n}\sum_{0\le k_1,\dots,k_d\le m-1} g\left(\frac{k_1}{m},\dots,\frac{k_d}{m}\right) - \sum_{0\le k_1,\dots,k_d\le m-1}\int_{[\frac{k_1}{m},\frac{k_1+1}{m}]\times\cdots\times[\frac{k_d}{m},\frac{k_d+1}{m}]} g(x)\,\mathrm{d}x \\ &= \sum_{0\le k_1,\dots,k_d\le m-1}\int_{[\frac{k_1}{m},\frac{k_1+1}{m}]\times\cdots\times[\frac{k_d}{m},\frac{k_d+1}{m}]} \left\{g\left(\frac{k_1}{m},\dots,\frac{k_d}{m}\right) - g(x)\right\}\mathrm{d}x. \end{aligned} \tag{2.3}$$

Taylor expanding g at order 1 in each integral, we get that (2.3) is $\mathcal{O}(m^{-1}) = \mathcal{O}(n^{-1/d})$ for all $\mathcal{C}^1$ test functions, assuming that π is $\mathcal{C}^1$ as well. Actually, we will see in Section 2.1.2 that this is the best one can get uniformly for $\mathcal{C}^1$ functions, for any choice of nodes $x_1,\dots,x_n$ and weights $w_1,\dots w_n$. This exponential dependence in the dimension is called the *curse of dimensionality*: it implies that the number of nodes n must grow exponentially with the dimension d to reach the same worst-case-error as in dimension 1. This is thus prohibitive for moderate to high dimensions d appearing in modern applications, see Chapter 1.

There are two main ways to weaken the objective in order to escape this curse of dimensionality. First, one can assume more smoothness on the integrands, or replace this smoothness assumption by a structural assumption on the class $\mathcal{F}$. Second, one can restrict to classes $\mathcal{F}$ that are singletons, meaning that we only seek to give integration guarantees for a single test function f with some given regularity.

Some usual worst-case integration errors

We first give some usual choices for classes of integrands $\mathcal{F}$ apart from singletons. The first set of examples are inducing standard distances on the space of probability measures, where we assume only low regularity on the integrands f on general domains.

■ **Example 2.1.1** Let

$$\mathcal{F} = \{\|f\|_\infty \le 1\} \tag{2.4}$$

be the unit ball of the space of bounded functions. Then, the error (2.1) becomes

$$\mathrm{E}_{\mathcal{F}}(\nu) = \mathrm{TV}(\nu,\pi), \tag{2.5}$$

the total variation distance between probability measures. We will also sometimes use the notation $\|\nu-\pi\|_{\mathrm{TV}}$ for (2.5). The total variation enables to integrate indicator functions, but this is a very strong metric and thus often hard to control. ♦

■ **Example 2.1.2** Let

$$\mathcal{F} = \{\|f\|_{\mathrm{Lip}} \le 1\} \tag{2.6}$$

be the unit ball of the space of Lipschitz functions. Then

$$\mathrm{E}_{\mathcal{F}}(\nu) = W_1(\nu, \pi) = \inf_{\gamma \in \Pi(\nu,\pi)} \iint |x - y| \, \mathrm{d}\gamma(x, y), \tag{2.7}$$

where $\Pi(\nu, \pi)$ is the set of couplings between ν and π (Villani, 2009). This is the 1-Wassertein distance between probability measures, and its convergence is equivalent to the weak convergence plus the convergence of the first moment. More generally, we can as well look at the unit ball of the space of α-Hölder functions given by

$$\mathcal{F} = \{\|f\|_{C^{0,\alpha}} \leq 1\}, \tag{2.8}$$

where

$$\|f\|_{C^{0,\alpha}} = \sup_{x \neq y} \frac{|f(y) - f(x)|}{|y - x|^\alpha}. \tag{2.9}$$

In that case, $\mathrm{E}_{\mathcal{F}}(\nu)$ is also known as the Zolotarev distance between ν and π (Rachev, 1991), and we recover the 1-Wassertein distance for $\alpha = 1$. ♦

■ **Example 2.1.3** Let

$$\mathcal{F} = \{\|f\|_{\mathrm{Lip}} \leq 1, \|f\|_\infty \leq 1\} \tag{2.10}$$

be the space of functions bounded by 1 both in supremum norm and in Lipschitz norm. Then

$$\mathrm{E}_{\mathcal{F}}(\nu) = d_{\mathrm{BL}}(\nu, \pi) \tag{2.11}$$

is the Bounded–Lipschitz distance between probability measures, which metrizes the weak convergence of probability measures (Dudley, 2002). Again, we can also look at the more general class given by

$$\mathcal{F} = \{\|f\|_{C^{0,\alpha}} \leq 1, \|f\|_\infty \leq 1\}. \tag{2.12}$$

♦

We turn to the general example of Sobolev spaces, where we can tune the smoothness of the function f, usually on the hypercube.

■ **Example 2.1.4** Let

$$\mathcal{F} = \{f \in L^p : \sum_{|\alpha| \leq k} \|\partial^\alpha f\|_{L^p} \leq 1\} \tag{2.13}$$

be the unit ball of the Sobolev space of parameters $1 \leq p < \infty$ and $k \in \mathbb{N}^*$, where ∂^α denotes the mixed partial derivative of order given by the multi index α. For $\alpha = (\alpha_1, \ldots, \alpha_d)$, $|\alpha|$ is defined by

$$|\alpha| = \sum_{j=1}^{d} \alpha_j. \tag{2.14}$$

We will go back in Section 2.1.2 to the case $p = +\infty$, in which case

$$\mathcal{F} = \{f \, : \, \|\partial^\alpha f\|_{L^\infty} \leq 1 \, \forall |\alpha| \leq k\}. \tag{2.15}$$

In the case $p = 2$ and for any k, one can alternately define this space as

$$\mathcal{F} = \left\{ f \in L^2 \, : \, \int \left(1 + |\omega|^2\right)^k |\hat{f}(\omega)|^2 \, \mathrm{d}\omega \leq 1 \right\}, \tag{2.16}$$

where $\hat{f}$ is the Fourier transform of f (Brezis, 2011), meaning that the smoothness assumption on f translates as a decay condition on its Fourier transform. ♦

It might happen than we are more comfortable making a structural assumption on the class $\mathcal{F}$ rather than a smoothness assumption. This happens for instance when the integrands of interest are expensive-to-evaluate functions as presented in Chapter 1: in that case building a smooth approximation of f is too expensive.

Instead, if we are given a set of integrands of interest $\Phi_1, \ldots, \Phi_M$, we can already be happy to integrate functions f that lie in the span of $\Phi_1, \ldots, \Phi_M$, meaning that $f(x) = \sum_{i=1}^M a_i \Phi_i(x)$ for some $a_1, \ldots, a_M \in \mathbb{R}$ and $x \in \mathbb{R}^d$. If $g(x) = \sum_{i=1}^M b_i \Phi_i(x)$, we can define the dot product

$$\langle f, g \rangle = \sum_{i=1}^{M} \sum_{j=1}^{M} a_i b_j \Phi_i(x) \Phi_j(x), \tag{2.17}$$

and define $\mathcal{F}$ as the set of functions f such that $\langle f, f \rangle \leq 1$. Reproducing kernel Hilbert spaces (RKHS, Berlinet and Thomas-Agnan, 2004) are functional spaces that implement this idea for a possibly infinite number of features M.

Definition 2.1.5 — Reproducing kernel Hilbert space. Consider a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that is symmetric, non-negative and positive semidefinite (s.d.p), meaning that

$$\sum_{i=1}^{p} \sum_{j=1}^{p} a_i a_j K(x_i, x_j) \geq 0$$

for any $p \in \mathbb{N}_*$ and any $a_1, \ldots, a_p \in \mathbb{R}$ and $x_1, \ldots, x_p \in \mathbb{R}^d$. We consider the class $\mathcal{H}_0$ of functions f that are interpolations of features based on K, i.e. $f(x) = \sum_{i=1}^p a_j K(x, x_j)$. We endow $\mathcal{H}_0$ with the scalar product induced by the reproducing

property $\langle K(x,.), f\rangle = f(x)$, meaning that

$$\langle f, g\rangle_{\mathcal{H}_0} := \sum_{i=1}^{p}\sum_{j=1}^{q} a_i b_j K(x_i, x_j),$$

if $f = \sum_{i=1}^{p} a_i K(., x_i)$ and $g = \sum_{j=1}^{q} b_j K(., x_j)$. Then, the RKHS of kernel K is the Hilbert space $(\mathcal{H}_K, \langle ., .\rangle_{\mathcal{H}_K})$ obtained by closure of $\mathcal{H}_0$ with respect to this scalar product. It satisfies the reproducing property, meaning that $\langle K(x,.),\, f\rangle_{\mathcal{H}_K} = f(x)$ for all $f \in \mathcal{H}_K$ and $x \in \mathbb{R}^d$ (Berlinet et al., 2004).

Worst-case integration errors based on RKHSs will be key in this thesis.

■ **Example 2.1.6** Let K be as in Definition 2.1.5, and let

$$\mathcal{F} = \{\|f\|_{\mathcal{H}_K} \leq 1\} \tag{2.18}$$

be the unit ball of the RKHS induced by K. The worst-case integration error associated to $\mathcal{F}$ induces a distance between probability measures, called the maximum mean discrepancy (MMD)

$$\mathrm{MMD}_K(\nu, \pi) := \mathrm{E}_{\mathcal{F}}(\nu). \tag{2.19}$$

See for instance (Sriperumbudur et al., 2009) for detailed properties of this distance depending on the properties of kernel K. ♦

The smoothness of the functions in $\mathcal{H}_K$ is induced by the properties of K. In fact, depending on the properties of K, various expressions are available for the norm $\|f\|_{\mathcal{H}_K}$, we give them in the following.

Proposition 2.1.7 — Steinwart and Scovel (2012), Theorem 3.1. Let K be as in Definition (2.1.5) and measurable. If $\mathcal{H}_K$ can be compactly and injectively embedded into $L^2(\mu)$ for some measure μ, then K admits a Mercer decomposition, meaning that for all $x, y \in \mathbb{R}^d$,

$$K(x,y) = \sum_{m=1}^{+\infty} \sigma_m \Phi_m(x)\Phi_m(y), \tag{2.20}$$

where $(\sigma_m)_{m\geq 1}$ are positive real number, the $(\Phi_m)_{m\geq 1}$ form an orthonormal basis of $L^2(\mu)$ and the convergence in the sum is pointwise. Moreover,

$$\|f\|^2_{\mathcal{H}_K} = \sum_{m\geq 1} \sigma_m^2 \left|\langle f\,,\, \Phi_m\rangle_{L^2(\mu)}\right|^2. \tag{2.21}$$

While this decomposition can be explicit for standard kernels defined only on a compact subset of $\mathbb{R}^d$, the assumptions of Proposition 2.1.7 can be restrictive for kernels defined on

$\mathbb{R}^d$ and one needs to fall back on weaker versions of the Mercer decomposition (Steinwart et al., 2012). For translation invariant kernels, we can rephrase again the norm in Fourier space.

Proposition 2.1.8 — Wendland (2004), Theorem 10.12. Let K be as in Definition (2.1.5) and measurable. Assume further that K is *integrally strictly positive definite*, meaning that

$$\iint K(x,y)\,\mathrm{d}(\mu-\nu) > 0, \tag{2.22}$$

for any finite and non-negative Borel measures μ and ν. Assume furthermore that K is translation-invariant i.e. that for all $x, y \in \mathbb{R}^d$, $K(x,y) = \Psi(x-y)$ with $\Psi \in L^1(\mathbb{R}^d)$. Then

$$\|f\|^2_{\mathcal{H}_K} = \int \frac{|\hat{f}(\omega)|^2}{\hat{\Psi}(\omega)}\,\mathrm{d}\omega. \tag{2.23}$$

We see in (2.21) and (2.23) the regularity of the functions in $\mathcal{H}_K$ is determined by the spectral properties of the kernel K. Usual choices of K in probabilistic machine learning (Rasmussen and Williams, 2005) include for instance Gaussian kernels

$$K_\sigma(x,y) = \exp\left(-\frac{1}{2\sigma^2}|x-y|^2\right), \tag{2.24}$$

for $\sigma > 0$, inverse multiquadratic (IMQ) kernels

$$K_{a,s}(x,y) = \frac{1}{(a^2+|x-y|^2)^{s/2}}, \tag{2.25}$$

for $a > 0$ and $s > 0$, or the general family of Matérn kernels

$$K_{\nu,l}(x,y) = \left(\frac{\sqrt{2\nu}|x-y|}{l}\right)^\nu K_\nu\left(\sqrt{2\nu}\frac{|x-y|}{\ell}\right), \tag{2.26}$$

where $\nu > 0$, $\ell > 0$ and K_ν is the modified Bessel function of the second kind.

In particular, for $p = 2$ and $k > d/2$, we can show that the Sobolev space (2.16) is a RKHS with Matérn kernel with parameters $\ell = 1$ and $\nu = k - d/2$ (Novak et al., 2018).

Probabilistic integration errors

A large class of modern quadrature methods rely on points $x_1, \ldots, x_n$ chosen at random to build the approximating measure $\nu_n = \sum_{i=1}^n w_i \delta_{x_i}$ in (2.1), meaning that $x_1, \ldots, x_n$ are random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then $\mathrm{E}_{\mathcal{F}}(\nu_n)$ in (2.2) is random, and we can leverage probabilistic tools to quantify the error.

Measuring integration error is often done in one of the following ways. First, one can stick to the worst-case integration error over a large class of functions $\mathcal{F}$, and try to bound $\mathrm{E}_{\mathcal{F}}(\nu_n)$ with high probability under the sampling of $x_1, \dots, x_n$. Second, we can abandon pointwise bounds on $\mathrm{E}_{\mathcal{F}}(\nu_n)$ and rather try to bound the averaged worst-case integration error given by $\left(\mathbb{E}\left[\mathrm{E}_{\mathcal{F}}(\nu_n)^p\right]\right)^{1/p}$, often with $p = 1$ or 2. Finally, we can remove the supremum from inside the average, consider the mean squared error (or quadratic risk) for a single integrand

$$\sqrt{\mathbb{E}\left[\left|\sum_{i=1}^{n} w_i f(x_i) - \int f \mathrm{d}\pi\right|^2\right]}, \tag{2.27}$$

and try to show that this is small for all f belonging to a class of functions $\mathcal{F}$. This amounts to consider the singleton $\{f\}$ as the function class of the original worst-case integration error (2.2), which reduces a lot the objective. Yet, one then has access to the bias-variance decomposition given by

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{i=1}^{n} w_i f(x_i) - \int f \mathrm{d}\pi\right|^2\right] = \\ \left(\mathbb{E}\left[\sum_{i=1}^{n} w_i f(x_i)\right] - \int f \,\mathrm{d}\pi\right)^2 + \mathrm{Var}\left[\sum_{i=1}^{n} w_i f(x_i)\right].\end{aligned} \tag{2.28}$$

and probabilistic tools such as central limit theorems.

In the three cases, we should keep in mind that we reduced the objective compared to the minimization of the worst-case error for every realization of $x_1, \dots, x_n$: we either ask for a result with high probability, or ask for a small expectation of the worst-case error, or even ask for a small expectation of the integration error for a single integrand to get finer results. When comparing rates of convergence in the next sections, we should then keep in mind that we do not always talk about the same way of measuring the integration error. We now turn to some lower bounds on the integration error to have in mind what can be achieved at best.

2.1.2 Lower bounds

While there is to our knowledge no explicit lower bound for the integration error in a general setting, some results are available for standard classes of integrands on $[0,1]^d$, see (Novak et al., 2018) for a survey of existing results.

We first consider deterministic sequences of points $x_1, \dots, x_n$. We define the optimal worst-case error by

$$e\left(n, \mathcal{F}\right) = \inf_{\substack{x_1,\dots,x_n \in \mathbb{R}^d \\ w_1,\dots,w_n \in \mathbb{R}}} \mathrm{E}_{\mathcal{F}}\left(\sum_{i=1}^{n} w_i \delta_{x_i}\right). \tag{2.29}$$

First, $e(n, \mathcal{F})$ can be computed exactly when $\mathcal{F}$ is the unit ball of the set of Lipschitz functions on the hypercube.

Theorem 2.1.9 — Novak (2016), Theorem 2. Consider π to be the uniform measure on $[0,1]^d$ and

$$F_d = \{f : [0,1]^d \to \mathbb{R}, \, |f(x) - f(y)| \leq \|x - y\|_\infty\}, \tag{2.30}$$

the space of functions on the hypercube with oscillation bounded by 1. Then

$$e(n, F_d) = \frac{d}{2+d} n^{-1/d} \tag{2.31}$$

for $n = m^d$ with $m \in \mathbb{N}^*$.

This means that we cannot escape the curse of dimensionality on such a large set of integrands, and the naive grid mentioned in Section 2.1.1 is already optimal up to a constant. We thus consider classes of functions with more regularity such as the Sobolev classes

$$W_{p,k} = \{f : [0,1]^d \to \mathbb{R} : \sum_{|\alpha| \leq k} \|\partial^\alpha f\|_{L^p} \leq 1\}, \tag{2.32}$$

for $p \geq 1$ and

$$W_{\infty,k} = \{f : [0,1]^d \to \mathbb{R} : \, \|\partial^\alpha f\|_\infty \leq 1 \, \forall |\alpha| \leq k\}. \tag{2.33}$$

Recall that for a multi-index $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$, we denote

$$|\alpha| = \sum_{j=1}^d \alpha_j, \tag{2.34}$$

$$\|\alpha\|_\infty = \max_{j=1,\dots,d} \alpha_j. \tag{2.35}$$

For the Sobolev space (2.32), it is known (Novak, 2016; Novak et al., 2018) that

$$e(n, W_{p,k}) = \mathcal{O}\left(n^{-k/d}\right), \tag{2.36}$$

as soon as $p \geq 2$ and $k > d/2$, and that the order is the same for $W_{\infty,k}$. This means that we can hope to escape the curse of dimensionality by assuming enough smoothness on the integrands. Relatedly, we will also look at optimal results for the unit ball of the space of functions with L^p mixed derivatives of order k

$$W_{p,k}^{mix} = \{f : [0,1]^d \to \mathbb{R} : \|\partial^\alpha f\|_{L^p} \leq 1 \, \forall \|\alpha\|_\infty \leq k\}, \tag{2.37}$$

for $k \in \mathbb{N}$ and $1 \leq p \leq +\infty$. In that case, we know (Novak, 2016, Theorem 6) that the optimal lower bound is

$$e\left(n, W_{p,k}^{mix}\right) = \mathcal{O}\left(\frac{\log^{\frac{d-1}{2}}(n)}{n^k}\right) \tag{2.38}$$

for e.g. $1 < p < +\infty$ and $k \in \mathbb{N}^*$. The cases $p = 1$ and $p = +\infty$ are still open problems. The gain in (2.38) compared to (2.36) should not be surprising since we encapsulate functions with smoothness up to order kd. However, it comes at the price of a power of logarithmic term that can be quite an issue, we will discuss this in Section 2.4.

On the other hand, consider a kernel K that admits a Mercer decomposition $K(x, y) = \sum_{m=1}^{+\infty} \sigma_m \Phi_m(x)\Phi_m(y)$ as in Proposition 2.1.7, and consider the worst-case integration error (2.19) over $\mathcal{H}_K$. Then it is known (Pinkus, 1985) that the supremum of this worst-case integration error over all L^2 target distributions π is lower bounded by $\mathcal{O}(\sqrt{\sigma_{n+1}})$, i.e.

$$\sup_{\|\pi\|_{L^2} \leq 1} e\left(n, \mathcal{H}_K\right)^2 \geq \sigma_{n+1}. \tag{2.39}$$

In particular, we can hope to rely on the kernel structure to achieve small $\mathcal{O}(\sqrt{\sigma_{n+1}})$ integration errors, see Section 2.6. Note that this lower bound deals with quadrature rules that are good uniformly over all L^2 target distributions π, meaning that this is just an upper bound for the worst-case integration error for a single fixed target π. As an example, for the Sobolev RKHS class defined in (2.16) with smoothness $k > d/2$, one can show (Widom, 1963) that $\sigma_{n+1} = \mathcal{O}\left(n^{-2s/d}\right)$ whereas the decay can be exponential for the RKHS induced by a Gaussian kernel K, see for instance (Chatalic et al., 2025) for more references.

Turning to randomized sequences of points $x_1, \ldots, x_n$, we consider the minimal mean squared error (2.27) over points $x_1, \ldots, x_n$ and weights $w_1, \ldots, w_n$. Then, we usually define the optimal averaged error $e^{ran}(n, \mathcal{F})$ by the supremum of this minimal mean squared error over all integrands $f \in \mathcal{F}$, see (Novak, 2016) for a rigorous definition. When $\mathcal{F}$ is the unit ball of $L^p([0,1]^d)$ with $p \geq 2$, it is known (Novak, 2016, Page 6) that $e^{ran}(n, \mathcal{F})$ scales like $\mathcal{O}(n^{-1/2})$. This means that considering randomized sequences of points helps bypass the curse of dimensionality without having to assume much smoothness from the integrands. Yet, the vanilla $n^{-1/2}$ rate is often deemed to be slow. In the case of Sobolev spaces,

$$e^{ran}(n, W_{\infty,k}) = \mathcal{O}(n^{-1/2-k/d}), \tag{2.40}$$

(Novak, 2016, Theorem 3). Similarly, we have the following optimal bound for mixed Sobolev spaces (Ullrich, 2017)

$$e^{ran}\left(n, W_{p,k}^{mix}\right) = \mathcal{O}(n^{-1/2-k}). \tag{2.41}$$

Note that there is no more logarithmic penalty in the rate, as opposed to the deterministic setting (2.38). To my knowledge, there is however no general result for the optimal averaged error when $\mathcal{F}$ is a RKHS induced by a kernel K.

When comparing to $e(n, \mathcal{F})$ in (2.29), we should keep in mind that we reduced the objective from controlling the worst-case integration error to the averaged error for a single integrand. In particular, we always have

$$e^{ran}(n, \mathcal{F}) \leq e(n, \mathcal{F}). \tag{2.42}$$

In the next section, we give two crude examples of quadrature rules, one deterministic and the other based on independent random points, that illustrate common challenges in matching the lower bounds.

2.2 Two crude quadratures

We focus here on Gaussian quadrature and independent random sampling, that respectively achieve the lower bounds of Section 2.1.2 and are thus standard baselines for other quadratures.

2.2.1 Gaussian quadrature

An early quadrature method is Gaussian quadrature on the unit interval $[-1, 1]$ (Gauss, 1815). In that case, for a given family of orthonormal polynomials $(\phi_k)_{k\in\mathbb{N}}$ with respect to π and with respective degrees k, $x_1, \ldots, x_n$ are set to be the zeroes of ϕ_n. Then we set the weights $w_i = \frac{1}{K_n(x_i,x_i)}$ where $K_n(x, y) = \sum_{k=0}^{n-1} \phi_k(x)\phi_k(y)$ is the so-called Christoffel–Darboux kernel, and consider the approximating measure $\nu_n = \sum_{i=1}^{n} w_i \delta_{x_i}$. See for instance (Gautschi, 2006) for a detailed presentation. Computing the quadrature nodes and weights can be done in general in $\mathcal{O}(n^2)$ steps (Gautschi, 2006, Chapter 2).

An important consequence of the choice of such quadrature nodes and weights is that for any polynomial function f of degree less than $2n - 1$, we have exactly

$$\sum_{i=1}^{n} w_i f(x_i) = \int f \, \mathrm{d}\pi, \tag{2.43}$$

meaning that the integration error is zero for such polynomials. Still, this is a very limited class of functions. While we know from Jackson's approximation theorem that the worst-case error $\mathrm{E}_{\mathcal{C}^1}(\nu_n)$ is the optimal $\mathcal{O}(1/n)$ (Theorem 2.1.9), optimal guarantees for Gauss quadrature for other classes of integrands are hard to prove in general. Another drawback is that Gauss quadrature is prohibitive in higher dimensions d which are often the one of interest in modern applications. In theory, Tchakaloff's theorem and its generalizations (Putinar, 1997) state that for any positive Borel measure on $\mathbb{R}^d$ with compact support and for a fixed degree k, there always exists an integer n_k and a

quadrature rule $(x_i, w_i)_{1\leq i\leq n}$ with $n \leq n_k \in \mathbb{N}^*$ such that $\mathrm{E}_{\mathcal{P}_k}(\nu_n) = 0$, where $\mathcal{P}_k$ is the class of polynomials $p(y_1, \ldots, y_d)$ with degree smaller than k. This result is however not practical at all. In practice, to use Gaussian quadrature, one typically restricts to product measures $\pi = \mu_1 \otimes \cdots \otimes \mu_d$ of measures on the unit interval $[-1, 1]$ and take the Cartesian product of univariate quadratures for each μ_i. Such a construction then yields an error $\mathrm{E}_{\mathcal{C}^1}(\nu_n) = \mathcal{O}\left(n^{-1/d}\right)$ that is optimal in the sense of Theorem 2.1.9, but prohibitive in practice when d is large.

2.2.2 Independent sampling

Monte Carlo methods are numerical integration methods that rely on points $x_1, \ldots, x_n$ chosen at random. The most natural way to approximate $\int f \,\mathrm{d}\pi$ using points chosen at random is to set uniform weights $w_i = 1/n$ and to draw $x_1, \ldots, x_n$ independently and identically distributed (i.i.d) according to π. Then we have access to a wealth of probabilistic results to quantify the integration error.

Worst-case integration error

First, Sanov's theorem gives us asymptotic control on the worst-case integration error for large classes of integrands $\mathcal{F}$.

Theorem 2.2.1 — Sanov's theorem, (Sanov, 1957; Csiszár, 2006). Assume that $x_1, \ldots, x_n$ are drawn i.i.d from π, and consider the empirical measure $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Then,

$$-\inf_{\mu\in\mathring{A}} H(\mu||\pi) \leq \varliminf_{n\to+\infty} \frac{1}{n}\log \mathbb{P}\left(\mu_n \in A\right) \leq -\inf_{\mu\in\overline{A}} H(\mu||\pi). \tag{2.44}$$

for any Borel set A in the set of probability measures $\mathcal{P}(\mathbb{R}^d)$ equipped with the topology of weak convergence. $H(\mu||\pi)$ denotes the relative entropy (or Kullback-Leibler divergence)

$$H(\mu||\pi) = \int \frac{\mu}{\pi}\log\left(\frac{\mu}{\pi}\right)\,\mathrm{d}\pi, \tag{2.45}$$

when μ has density μ/π, and $H(\mu||\pi) = +\infty$ otherwise.

Setting $A = B_{\mathbf{d}}(\pi, r)^c$, for any distance $\mathbf{d}$ on probability measures, Equation (2.44) yields

$$\varliminf_{n\to+\infty} \frac{1}{n}\log \mathbb{P}\left(\mathbf{d}(\mu_n, \pi) > r\right) \leq \alpha_r, \tag{2.46}$$

with $\alpha_r = \inf\{H(\mu||\pi) \,:\, \mathbf{d}(\mu, \pi) \geq r\}$.

In order to give an explicit bound in (2.46), we need to compare the distance $\mathbf{d}$ and the relative entropy H. For instance, assuming that π satisfies the entroy-distance inequality

$$\mathbf{d}(\mu, \pi)^2 \leq CH(\mu, \pi) \tag{2.47}$$

for all $\mu \in \mathcal{P}(\mathbb{R}^d)$, we have

$$\alpha_r > \frac{1}{C} r^2. \tag{2.48}$$

From Equation (2.46), we get that for any $\epsilon > 0$, there exists n_0 depending on ϵ such that for any $n \geq n_0$

$$\mathbb{P}\left(\mathbf{d}(\mu_n, \pi) > \epsilon\right) \leq \exp\left(-\frac{1}{C} n r^2\right). \tag{2.49}$$

The right-hand side is the tail decay of a Gaussian distribution with variance $\sigma^2 = C/n$. Moreover, the lower bound in Equation (2.44) tells us that the rate in n is sharp. In other words, for any $n \geq n_0(\epsilon)$,

$$\mathbf{d}(\mu_n, \pi) \leq \epsilon \tag{2.50}$$

with probability larger than $1 - \exp\left(-\frac{1}{C} n \epsilon^2\right)$. The entropy transport inequality Equation (2.47) is satisfied for instance for any $\pi \in \mathcal{P}(\mathbb{R}^d)$ if $\mathbf{d}$ is the total variation distance

$$\mathbf{d}(\mu, \pi) = \mathrm{TV}(\mu, \pi) := \sup_{\|f\|_\infty \leq 1} \left| \int f \, \mathrm{d}(\nu - \pi) \right|, \tag{2.51}$$

as in Example 2.1.1 and $C = 1/2$. In fact, (2.47) is exactly Pinsker's inequality in that case.

Non-asymptotic counterparts of inequalities like (2.49) can be obtained through concentration inequalities. For instance, for the Wassertein distance of Example 2.1.2 and under exponential moments assumptions for π[1], Fournier et al. (2015) show that for all $d' > d$, there exists n_0 and a constant $K > 0$ such that for all $n \geq n_0$,

$$\mathbb{P}(W_1(\mu_n, \pi) > \epsilon) \leq \exp\left(-Kn \min(\epsilon, \epsilon^2)\right), \tag{2.52}$$

for any $\epsilon > \left(\frac{n_0}{n}\right)^{1/(d'+2)}$. Equivalently, for a fixed $\delta > 0$ and $n \geq n_0$,

$$W_1(\mu_n, \pi) \leq \frac{\log\left(\frac{1}{\delta}\right)}{Kn} \vee \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{Kn}}, \tag{2.53}$$

with probability larger than $1 - \delta$ and $\delta > \exp(-K n_0^{1/(d'+2)} n^{1-2/(d'+2)})$. The dependence of K and n_0 is however not explicit in the dimension d.

Results have been obtained as well for the MMD of Example 2.1.6. Consider a kernel K as in Definition 2.1.5 such that $|K(x, y)| \leq C$ for all $x, y \in \mathbb{R}^d$. It is then known (Gretton et al., 2006, Theorem 4) that the MMD (2.19) for i.i.d points is $\mathcal{O}(n^{-1/2})$ in probability. Furthermore, this rate is known to be optimal for the MMD with translation invariant kernel K and i.i.d points $x_1, \ldots, x_n$ (Tolstikhin et al., 2016, Theorem 6). This dimension independent rate is a feature coming from the kernel structure we assumed on the integrand space.

[1]Those can be weakened into moment assumptions or replaced by assuming T_1 inequalities (Fournier and Guillin, 2015).

Averaged worst-case error

Regarding the average worst-case error of an i.i.d quadrature, one can show (Chewi et al., 2024, Proposition 2.1) that for $d \geq 3$ and any π supported on $[0,1]^d$,

$$\mathbb{E}\left[W_1(\mu_n, \pi)\right] \lesssim \sqrt{d} n^{-1/d}, \tag{2.54}$$

where $\lesssim$ means that the inequality is true up a multiplicative constant. Moreover, the rate is $n^{-1/2}$ up to logarithmic terms in dimensions one and two. From Theorem 2.1.9, we know that the rate is n in (2.54) is optimal for $d \geq 3$.

On the other hand, for the MMD we have again the dimension independent rate

$$\mathbb{E}\left[\mathrm{MMD}_K(\mu_n, \pi)\right] \leq n^{-1/2}, \tag{2.55}$$

assuming that $|K(x,x)| \leq 1$ for all $x \in \mathbb{R}^d$ (Chewi et al., 2024, Theorem 2.22). Actually, we can even get $\mathbb{E}\left[\mathrm{MMD}_K(\mu_n, \pi)\right] = o\left(n^{-1/2}\right)$ by choosing optimal weights $w_1, \dots, w_n$, under moment assumptions and for a specific choice of kernel built from K and π (Li et al., 2024). The rate is however not explicit. From Section 2.1.2, we know this time that the rate in (2.55) is suboptimal in general and there is thus room for improvement.

Mean squared error for single integrands

Looking rather at single integrands, we can directly express the bias variance decomposition (2.28) for i.i.d points $x_1, \dots, x_n$ as

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f \mathrm{d}\pi\right|^2\right] = \frac{1}{n}\mathrm{Var}\left[f(x_1)\right] \tag{2.56}$$

for any $f \in L^2(\pi)$, since $\mathbb{E}[f(x_i)] = \int f \,\mathrm{d}\pi$ for all $1 \leq i \leq n$.

This means that the mean squared error behaves exactly like $n^{-1/2}$ at fixed n, which is optimal on L^2 functions from Section 2.1.2. An impressive feature is that this does not downgrade with the dimension d, while encapsulating a very large class of functions with no smoothness or structural assumption. This is possible only because we asked for much less in terms of guarantees.

Similarly, we also have access to the standard central limit theorem for i.i.d variables which is key for uncertainty quantification.

Theorem 2.2.2 — CLT for i.i.d variables. For any $f \in L^2(\pi)$,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f \,\mathrm{d}\pi\right) \to \mathcal{N}\left(0, \int f^2 \,\mathrm{d}\pi\right). \tag{2.57}$$

This means that for any f that is only in $\mathrm{L}^2(\pi)$ the integration error $\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f \,\mathrm{d}\pi$ asymptotically behaves as a centered Gaussian with variance scaling as n^{-1}.

Still, in many practical situations of Bayesian inference, we are usually unable to generate samples $x_1, \dots, x_n$ that are distributed i.i.d according to π. One way to circumvent this issue is to assume that we are able to draw from a distribution q whose support contains the one of π. Then, considering the weights

$$w_i = \frac{W(x_i)}{\sum_{i=1}^n W(x_i)} \tag{2.58}$$

with $W(x) = \pi(x)/q(x)$, Slutsky's lemma along with the CLT (2.57) applied to the function $W.f$ yields

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n w_i f(x_i) - \int f \,\mathrm{d}\pi\right) \to \mathcal{N}\left(0, \int f^2 \frac{\pi}{q}\,\mathrm{d}\pi\right). \tag{2.59}$$

This is called (self-normalized) importance sampling, and the self-normalization is here to avoid computing the normalization constant of π when computing the weights. Some issues remain with this trick: we still need to be able to sample i.i.d from q, and importance sampling is also known to suffer from a curse of dimensionality (Chopin and Papaspiliopoulos, 2020, Chapter 8.7). Yet, some modern Monte Carlo methods such as Sequential Monte Carlo methods (SMC, Chopin et al., 2020) build upon importance sampling to gain in scalability in high dimensions or for complex targets.

We devote the next section to Markov chain Monte Carlo algorithms, that aim to mimic the properties of i.i.d sampling without knowing how to draw points from π or any suitable proposal q that keeps the mean square error small.

2.3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC, Robert et al., 2004) is a celebrated family of algorithms to draw points $x_1, \dots, x_n$ at random to perform numerical integration with respect to π without relying on i.i.d sampling. The idea is to replace the unfeasible i.i.d draw of $x_1, \dots, x_n$ by iterating n steps of a Markov kernel P that leaves π invariant, starting from an arbitrary point x_0 drawn from an initial distribution ξ. Under a classical assumption known as Harris ergodicity, the Markov kernel then satisfies

$$\mathrm{TV}\left(\xi P^n, \pi\right) \to 0, \tag{2.60}$$

$\mathbb{P}_\xi$-almost surely for any initial distribution ξ (Douc et al., 2014), and where $\mathbb{P}_\xi$ denotes the law of the Markov chain starting from $X_0 \sim \xi$. This means in particular that

$$\frac{1}{n}\sum_{i=1}^n f(x_i) \longrightarrow \int f \,\mathrm{d}\pi \tag{2.61}$$

almost surely, when $x_1, \dots, x_n$ are a realization of the n steps of the Markov chain with Markov kernel P, for any starting distribution. This suggests that such $x_1, \dots, x_n$ are

a reasonable choice of nodes for numerical integration. MCMC has been successfully applied in many of contexts — historically for the simulation of hard-sphere models (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953) — and designing efficient MCMC algorithms is still an active field of research.

The main features of MCMC methods is that they are able to reproduce the guarantees of i.i.d sampling, and that they do not require to know the normalization constant of the density of the target π, which is key for instance in Bayesian statistics, see Chapter 1. We first introduce basic MCMC algorithms that we will refer to in this thesis, i.e. we give choices of P such that π is invariant and P is Harris ergodic.

2.3.1 Metropolis–Hastings algorithms

Random walk Metropolis–Hastings

We describe in the following the famous random walk Metropolis–Hastings (MH) algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953; Hastings, 1970).

Definition 2.3.1 Given a target π and a proposal distribution $q(\cdot|x)$, we consider the Markov kernel

$$P(x, A) = \int_A \alpha(x, y)\, \mathrm{d}q(y|x) + \left(1 - \int_{\mathbb{R}^d} \alpha(x, y)\, \mathrm{d}q(y|x)\right) \delta_x(A), \tag{2.62}$$

for all Borel set $A \subset \mathbb{R}^d$ and where

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}. \tag{2.63}$$

Under mild assumptions on the target π and on the proposal q, this Markov kernel is known to be Harris ergodic with unique invariant distribution given by π (Robert et al., 2004, Chapter 7.3.2). Actually, π is even reversible for this Markov kernel. Furthermore, the acceptance ratio $\alpha(x, y)$ does not depend on the normalization constant in π, which is key e.g. in Bayesian inference or for the simulation of physical systems, see Chapter 1.

The simplest choice for the proposal q is to consider a Gaussian proposal

$$x_{k+1}|x \sim \mathcal{N}(0, \alpha^2 I_d) \tag{2.64}$$

where α is a step-size to be tuned. In other words, at each step, the proposed point is given by $x_{k+1} = x_k + \alpha\zeta$ where ζ is an isotropic Gaussian random variable. The sampling cost is only $\mathcal{O}(n)$ evaluations of π in time complexity, where n is the number of Metropolis–Hastings steps.

On the one hand, a small step-size will result in small jumps: the Markov kernel will often accept the new point x_{k+1}, but the chain will move very slowly and many iterations

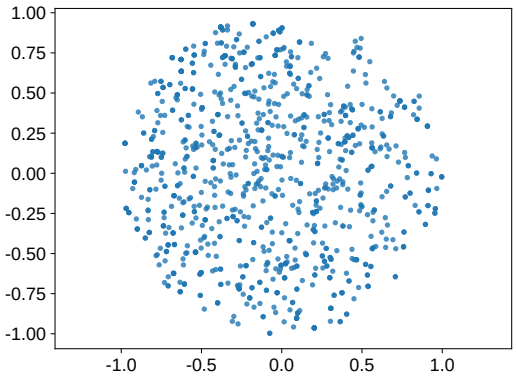


Fig. 2.1: $n = 1000$ **points from the history of a Metropolis–Hastings chain.**

will be needed to converge to regions of high probability under π, especially if π is a complicated target, e.g. multimodal or very spiked around some local maxima where the Markov chain can be trapped.

On the other hand, a large step-size will results in big jumps: the Markov kernel will reject the updates and the Markov chain won't move. Tuning the step size α is thus key, and it has been recommended (Gelman et al., 1996; Breyer and Roberts, 2000) that targeting an acceptance rate around 0.234 should be optimal in terms of limiting variance, as is scaling α^2 as $1/d$.

In practice, for a complicated target such as a very spiked π, one relies on an optimization algorithm beforehand to identify all the modes or local maxima and initialize the Markov chain with a warm start in those modes. It is also often a good idea to concatenate various Markov kernels from (2.64) with different choices of α to combine exploration within a mode and jumps to go from one mode to another.

In Figure 2.1, we show the history of a Metropolis–Hastings chain with proposal (2.64), for a two-dimensional uniform target π, and $n = 1\,000$ iterations. We see that the points tend to cluster and leave holes in the integration domain: trying to produce points that spread better in the integration domain will be the starting point of most methods we will consider in the rest of this chapter.

We now introduce gradient-based proposals that we will as well use in the rest of the thesis.

Metropolis-Adjusted Langevin

Consider the overdamped Langevin diffusion

$$\mathrm{d}X_t = \nabla \log \pi \mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

where B_t is a standard Brownian motion. It is known that π is the stationary measure for such a diffusion, see for instance (Roberts and Tweedie, 1996b, Theorem 2.1). If

one has access to the gradient of $\log \pi$ (which does not depend on the normalization constant of the density π), one can approximately sample from π by discretizing the Langevin diffusion above. The Euler scheme discretization gives the following Unadjusted Langevin algorithm (ULA):

$$x_{k+1} = x_k + \alpha \nabla \log \pi(x_k) + \sqrt{2\alpha}\zeta, \tag{2.65}$$

where α is a discretization step-size that needs to be tuned and ζ is a standard Gaussian. Due to the discretization error however, this sampling scheme is biased in the sense that the stationary distribution of ULA is not π anymore. It is moreover very sensitive to the choice of α (Roberts et al., 1996b), so one usually prefers to add a Metropolis rejection sampling step, which gives the Metropolis Adjusted Langevin Algorithm (MALA, Roberts et al., 1996b): this is the Metropolis–Hastings algorithm of Definition 2.3.1, with proposal $q(\cdot|x)$ given by

$$x_{k+1}|x_k \sim \mathcal{N}(x_k + \alpha \nabla \log \pi(x_k), 2\alpha I_d). \tag{2.66}$$

The idea is that this proposal is informed by the gradient of the log target, so that the updates will already be done in a direction that tends to increase the probability under π.

2.3.2 Bias of MCMC algorithms

MCMC algorithms only produce points that are approximately drawn according to π, they thus introduce a bias that one needs to control, i.e. the bias term

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} f(x_k)\right] - \int f \,\mathrm{d}\pi = \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}_\xi\left[P^k f\right] - \int f \,\mathrm{d}\pi\} \tag{2.67}$$

is non-zero, where ξ is the law of the initialization x_0. In order to get rates of convergence for this quantity, we are interested in the distance of the law of the $k-th$ iterate ξP^k to the target π. We usually choose this distance to be the total variation distance to get strong convergence guarantees and provide credibility regions, see Section 2.1.1. First, if π is invariant, then one is able to show that the total variation distance from ξP^k to π decreases with k for any ξ (Robert et al., 2004, Proposition 6.52). Assume now that the Markov kernel is geometrically ergodic, i.e. that for all $x \in \mathbb{R}^d$ there exists a constant $0 < \kappa_x < 1$ such that

$$\|P^k(x, .) - \pi\|_{\mathrm{TV}} \leq C\kappa_x^k. \tag{2.68}$$

Then, as soon as f is bounded, for any starting point x,

$$\mathbb{E}_x\left[\frac{1}{n}\sum_{k=1}^{n} f(x_k)\right] - \int f \,\mathrm{d}\pi \leq \frac{C}{n(1-\kappa_x)}, \tag{2.69}$$

where $\mathbb{E}_x$ denotes expectation with respect to $\mathbb{P}_x = \mathbb{P}_{\delta_x}$. Proving geometric ergodicity as in (2.68) usually needs to be tailored to each pair of target π and Markov kernel. Geometric ergodicity for MH with Gaussian proposal or MALA have been studied in (Roberts and Tweedie, 1996a; Roberts et al., 1996b), and sufficient conditions are usually strong log-concavity and mild smoothness for π. If we are in a geometrically ergodic case, a key quantity to assess the convergence of MCMC algorithms is the idealized stopping criterion given by the mixing time

$$\mathcal{T}_{\mathrm{TV}}(\xi, \delta) = \inf\{k \geq 1 : \|\xi P^k - \pi\|_{\mathrm{TV}} < \delta\}, \tag{2.70}$$

for a fixed precision $\delta > 0$ and initial distribution ξ. The mixing time gives the number of steps that are required in the MCMC algorithm to reach error δ. Obtaining bounds that encapsulate the effect of the dimension d has been the topic of much effort in the MCMC community. For instance, Dwivedi et al. (2019) show that when the target π is strongly log-concave, the mixing time for e.g. MALA scales polynomially in the dimension, meaning that we only need a polynomial number of MALA iterations to ensure that the last iterate is close to the target measure in total variation. We will discuss such results in more detail for Metropolis–Hastings algorithms in Chapter 4 and Chapter 6.

2.3.3 Asymptotic variance and integration guarantees

We now turn to the study of numerical integration errors with points produced by a Markov chain. Dembo and Zeitouni (2010, Theorem 6.3.8) provide large deviations for the empirical measure of Markov chains at the same speed n than for i.i.d points (Theorem 2.2.1), with explicit expressions for the rate function, under some involved hypothesis. This suggests that we should expect the same behavior in terms of n as the one we had for i.i.d points $x_1, \ldots, x_n$ in Section 2.2.2.

Bounds in expectation for the worst-case integration error of Markov chains have been obtained e.g. in (Bolley et al., 2007, Theorem 15) under moment conditions.

Theorem 2.3.2 — Bolley et al. (2007). Let $r > 2$, assume that π has a finite q-th moment for some $q > r/(r-1)$. Denote by ξ the initial distribution of the Markov chain, and by $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ the empirical measure of the chain. Under fast mixing conditions on the Markov chain, there exists a constant C such that for all $n \geq 1$,

$$\mathbb{E}_\xi\left[W_1(\mu_n, \pi)\right] \leq C\left(n^{-1/d} + n^{-(q_r-1)/q_r}\right), \tag{2.71}$$

where $q_r = q(r-1)/r$.

The first term in Equation (2.71) is the one we expect when comparing with the i.i.d result (2.54), while the second is an additional error term that is small when q is large.

Concentration inequalities have also been obtained in the MMD_K distance (2.19) for fast mixing Markov chain.

Theorem 2.3.3 — Dwivedi and Mackey (2021). Assume that K is bounded, measurable, and defines a RKHS as in Definition 2.1.5. Under fast mixing conditions on the Markov chain, mainly geometric ergodicity with stationary distribution π, there exists a $\mathbb{P}$-almost everywhere finite function c such that for any n and $\delta \in (0,1)$,

$$\mathrm{MMD}_K\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, \pi\right) \leq \sqrt{\frac{c(x_0)\|K\|_\infty \log(e/\delta)}{n}}, \tag{2.72}$$

with probability larger than $1-\delta$, where x_0 is the initial state of the Markov chain.

Li et al. (2024, Theorem 1) have further proved that the MMD is still $\mathcal{O}(n^{-1/2})$ in probability even for optimally chosen weights $w_1, \ldots, w_n$, under additional moment assumptions and for a specific form of kernel built from K and π. As in the i.i.d case in Section 2.2.2, the rate for the MMD in (2.72) is dimension independent as opposed to (2.71). Again, this is not surprising since we assumed much more structure in the set of integrands.

Finally, the CLT holds for Markov chains with the same rate as i.i.d points under mixing conditions. We give a rough statement here, some mild assumptions are being omitted for the ease of presentation.

Theorem 2.3.4 — Robert et al. (2004), Theorem 6.67. Assume that the chain has invariant distribution π and has geometric ergodicity. Assume that $f \in L^{2+\epsilon}(\pi)$ for some $\epsilon > 0$. Then

$$\sqrt{n}\left(\frac{1}{n}\sum f(x_i) - \int f \,\mathrm{d}\pi\right) \to \mathcal{N}(0, \sigma_f^2) \tag{2.73}$$

with $\sigma_f^2 = \mathrm{Var}_\pi[f(x_0)] + 2\sum_{k=1}^{+\infty} \mathrm{Cov}\left(f(x_0), f(x_k)\right)$.

As in the i.i.d setting (Theorem 2.2.2), the impressive features are the absence of the dimension in the rate in n, and the low regularity we assume on the integrand f. Moreover, from a practical point of view, the CLT allows to empirically quantify the error of the estimator $\frac{1}{n}\sum f(x_i)$ by replacing the limiting variance by estimates based on the same sample $x_1, \ldots, x_n$, see for instance (Flegal and Jones, 2010). Finally, together with (2.69), (2.73) means that the mean squared error (2.27) scales asymptotically as $\mathcal{O}(n^{-1/2})$.

Yet, the price to pay for these weak assumptions is a rate $n^{-1/2}$ that is deemed to be slow when the space of integrands exhibits some regularity or structure, see Section 2.1.2. In the sequel, QMC methods (Section 2.4) will try to give better integration

guarantees in these spaces. On the other hand, Variational Bayes methods (Section 2.5) will let go of integration guarantees and focus on providing fast scalable algorithms to produce nodes $x_1, \dots, x_n$ that approximate π more loosely.

2.4 Quasi Monte Carlo

Quasi Monte Carlo algorithms (QMC, Dick et al., 2013) aim at providing low integration errors by assuming more regularity or structure in the integrands we consider. More precisely, the objective is to try to match the lower bounds presented in Section 2.1.2 by building integration nodes $x_1, \dots, x_n$ that spread well in the integration domain. We first begin by low-discrepancy sequences that are deterministic sequences of points. Then, we look at Randomized Quasi Monte Carlo (RQMC) methods that choose the points $x_1, \dots, x_n$ at random, either by adding random perturbations to low-discrepancy sequences or by jointly drawing them from a distribution that spreads the points with strong negative dependence in the integration domain.

2.4.1 Low-discrepancy sequences

The justification for low-discrepancy sequences is the so-called Koksma–Hlawka inequality, see for instance (Dick et al., 2013).

Theorem 2.4.1 — Koksma-Hlawka inequality. Let $f \in \mathrm{L}_2([0,1]^d)$. Then for any $x = (x_1, \dots, x_n) \in [0,1]^{nd}$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{[0,1]^d} f(y)\, \mathrm{d}y \right| \leq \Delta_*(x) \|f\|_{\mathrm{HK}}, \tag{2.74}$$

where $\Delta_*(x)$ is the star-discrepancy of the point set x, defined below, and $\|f\|_{\mathrm{HK}}$ is the Hardy-Krause variation of the integrand f.

Definition 2.4.2 Let $x = (x_1, \dots, x_n)$. We define the local discrepancy at x by

$$\mathcal{D}_x(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(x_i) - \lambda(A)$$

where λ is the d-dimensional Lebesgue measure, and $A \subset [0,1]^d$. $\mathcal{D}_x(A)$ counts the excess in the number of points of x falling in A compared to the volume of the set A. Then the star-discrepancy is defined as

$$\Delta_*(x) = \sup_{A = \prod_{j=1}^{d} [0, u_j] \subset [0,1]^d} |\mathcal{D}_x(A)|. \tag{2.75}$$

On the one hand, the decoupling in Equation (2.74) means that the same sequence $x_1, \dots, x_n$ will give a small integration error for all integrands f, that are smooth enough. The Hardy–Krause variation is quite hard to express in general, but we can say the following: if f has continuous mixed first partial derivatives,

$$\|f\|_{\mathrm{HK}} = \sum_{\alpha:\|\alpha\|_\infty \leq 1} \|\partial^\alpha f\|_{L^1}, \tag{2.76}$$

(Dick et al., 2013), meaning that the Hardy-Krause variation typically involves derivatives of order d of f. In that case, we see from (2.74) that the star-discrepancy bounds the worst-case error in $W_{1,1}^{mix}$ defined in (2.37), and we are thus interested in building $x = (x_1, \dots, x_n)$ such that (2.75) is small, hence the *low-discrepancy* terminology.

On the other hand, the quantities in the right-hand side of (2.74) are quite hard to manipulate: computing the star discrepancy, for instance, is an NP-hard problem (Gnewuch et al., 2009). Hence, it is hard to get empirical assessments for low integration errors. Moreover, if we want to integrate with respect to a non-uniform target density, we can consider the test function $g(x) = f(x)\pi(x)$, but this assumes in turn regularity on the density of π, and that we are able to compute the normalization constant of π in the estimator $\frac{1}{n}\sum_{i=1}^n f(x_i)\pi(x_i)$. This can be unwelcome, see Chapter 1.

Recall that optimal lower bounds for the worst-case error in $W_{1,1}^{mix}$ are still an open problem. Yet, it is known (Dick and Pillichshammer, 2010, Theorem 3.20) that

$$\Delta_*(x_1, \dots, x_n) \geq c_d \frac{\log^{\frac{d-1}{2}}(n)}{n} \tag{2.77}$$

for $d \geq 2$, and $\Delta_*(x_1, \dots, x_n) \geq \frac{1}{2n}$ for $d = 1$, and we can thus expect similar bounds for the worst-case error in $W_{1,1}^{mix}$. Actually, other versions of the Koksma–Hlawka inequality (2.74) have been developed in the L^p setting for $1 < p < \infty$, replacing the supremum in the star discrepancy (2.75) by an L^q norm with $1/p + 1/q = 1$ and the Hardy-Krause variation (2.76) by

$$\|f\|_{p,1}^{mix} = \sum_{\alpha:\|\alpha\|_\infty \leq 1} \|\partial^\alpha f\|_{L^p}, \tag{2.78}$$

see (Dick et al., 2013, Theorem 3.9). In that case the L^p discrepancy exactly bounds the worst-case error in $W_{p,1}^{mix}$, and recall that we know (2.38) that the optimal lower bound is

$$e\left(n, W_{p,1}^{mix}\right) = \mathcal{O}\left(\frac{\log^{\frac{d-1}{2}}(n)}{n}\right). \tag{2.79}$$

Achieving the rate n^{-1} would be very satisfactory since it is much faster than the $n^{-1/2}$ error rate of MCMC; this is possible because we assumed some kind of smoothness at order d for the integrands.

Yet, the term $\log^{\frac{d-1}{2}}(n)$ in the numerator means that the number of points n needs to scale exponentially in the dimension d to get meaningful integration results with rate n^{-1} from (2.79). In fact, the quantity in (2.79) can otherwise be bigger than the MCMC scaling $n^{-1/2}$, and a simple computation shows that the function $\Phi_s(n) = \log^s(n)/n$ is actually even increasing for $n < \exp(s)$.

Apart from the fact that evaluating the integrand an exponential number of times can be prohibitive in complex Bayesian models (Chapter 1), we could argue that if we are ready to draw a number of points n that is exponential in the dimension, we could as well be already satisfied with the guarantees of MCMC. Since those are lower bounds, this is a fundamental limitation for all QMC methods based on minimizing L^q discrepancies for $q > 1$. We will see in Section 2.6 another version of Koksma–Hlawka inequality (2.74), this time taking advantage of a kernel structure.

We briefly describe a simple example of a low-discrepancy sequence, called the Halton sequence. For any integers $b > 2$ and n, define $v_{b,n} = \sum n_i b^{-i-1}$ where $n = \sum n_i b^i$ is the decomposition in base b, then the Halton sequence is given by the concatenation $x = (v_{b_1,n}, \dots, v_{b_d,n})$, where $b_1, \dots, b_d$ are the d first prime numbers. Then, it is known (Dick et al., 2013) that

$$\Delta_*(x_1, \dots, x_n) \leq \mathcal{O}\left(\frac{\log^d(n)}{n}\right),$$

which is close to optimal in regard of (2.77).

Building well-spread deterministic set of points is still the topic of ongoing research. For instance, modern QMC constructions such as digital nets (Dick et al., 2010, Chapters 4 and 5) can achieve $\mathcal{O}(\frac{(\log n)^{kd}}{n^k})$ integration error upon assuming that $f \in W_{2,k}^{mix}$ with $k \in \mathbb{N}^*$. This is again close to optimal in view of (2.38). We now turn to randomized set of points $(x_1, \dots, x_n)$ to improve the integration guarantees.

2.4.2 Randomized QMC

Randomized Quasi Monte Carlo methods are QMC methods that build up on well-spread *random* nodes $x_1, \dots, x_n$. We present two ways of doing so. The first one builds up on low-discrepancy sequences via random perturbations, while the second one seeks to sample from a joint distribution on $x_1, \dots, x_n$ that favors diversity between the points. We end with an alternative way to reduce integration errors for smooth functions, relying on control variates.

Randomization of low-discrepancy sequences

Consider a deterministic set of points $a_1, \dots, a_n \in \mathbb{R}^d$ which we decompose on each coordinate in base b i.e. $a_i^j = \sum_{k \geq 1} a_{i,j,k} b^{-k}$ for $1 \leq i \leq n$ and $1 \leq j \leq d$. Randomizations of the points $a_1, \dots, a_n$ in base b have been investigated in various ways. For instance, one can simply add i.i.d uniform perturbations to the coefficients $a_{i,j,k}$ modulo b. If f has finite Hardy–Krause variation and $a_1, \dots, a_n$ are chosen to be

instances of digital nets, this leads to a variance of order $\mathcal{O}(n^{-2}\log^{2d} n)$ (L'Ecuyer, 2016). Again, this improves on the n^{-1} scaling of the MCMC variance, when n grows exponentially in d. This rate can be even further improve for instance with scrambled nets, where the coefficients $a_{i,j,k}$ are randomized through uniform random permutations, which leads to a variance of order $\mathcal{O}(n^{-3}\log^{2d} n)$ (Owen, 1997), or even $\mathcal{O}(n^{-2k-1}\log^{(k+1)d} n)$ with higher order scrambled nets (Dick, 2011). We know from (2.41) that this is close to optimal, but we can this time hope to get rid of the logarithmic term, which is another big argument in favor of RQMC against vanilla QMC.

Determinantal point processes

Determinantal point processes (DPPs, Macchi, 1972; Soshnikov, 2000), and in particular projection DPPs of rank n, give joint distributions on points $x_1, \dots, x_n$ that incorporate negative correlations given by the determinant of an $n \times n$ kernel matrix $(K_n(x_i, x_j))_{1\le i,j\le n}$. We focus here on a specific case, the so-called Orthogonal Polynomial Ensembles (OPEs), and we will give a more general definition of DPPs in Chapter 3. For a given reference measure μ on $[-1,1]^d$ that is a product measure

$$\mu = \mu_1 \otimes \cdots \otimes \mu_d, \tag{2.80}$$

we consider the Christoffel–Darboux kernel

$$K_n(x,y) = \sum_{k=0}^{n-1} \Psi_k(x)\Psi_k(y) \tag{2.81}$$

where (Ψ_k) are the multivariate orthogonal polynomials associated to the measure μ. The DPP of kernel K_n then gives the joint distribution

$$x_1, \dots, x_n \sim \det\left(K_n(x_i,x_j)_{1\le i,j\le n}\right) \mathrm{d}\mu(x_1) \dots \mathrm{d}\mu(x_n). \tag{2.82}$$

and satisfies

$$\mathbb{E}\left[\sum_{i=1}^{n} f(x_i)\right] = \int f(x) K_n(x,x)\,\mathrm{d}\mu(x), \tag{2.83}$$

see Chapter 3.

For instance, in Figure 2.2, we use the DPPy package (Gautier et al., 2019b) to show a realization of a two-dimensional *multivariate Jacobi ensemble*, an example of (2.82), with $n = 1\,000$ points. We see that the points are well-spread inside the square, while sticking to the marginal distributions given by the orange curves on top of the axes.

Monte Carlo methods based on this choice of nodes have been investigated in (Bardenet and Hardy, 2020). The idea is that configurations like (2.82) bring variance reduction, and can thus be used for fast Monte Carlo integration after reweighting. In particular, there is a CLT with fast rate, which we give a rough statement of.

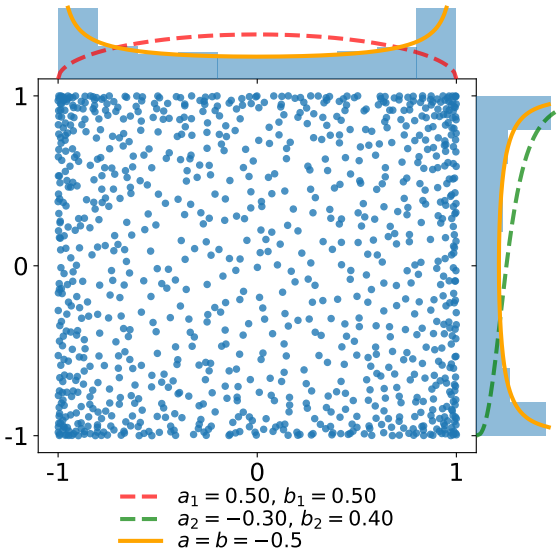


Fig. 2.2: $n = 1000$ **points sampled from a multivariate Jacobi ensemble.**

Theorem 2.4.3 — (Bardenet et al., 2020). Consider the target distribution π. Assume that $x_1, \dots, x_n$ are drawn from the distribution (2.82), with $\mu_1 = \cdots = \mu_d = q(x)\,\mathrm{d}x$. Then for any $f \in \mathcal{C}^1\left([-1,1]^d\right)$ with support strictly included in $[-1,1]^d$,

$$\sqrt{n^{1+1/d}}\left(\sum_{i=1}^{n} \frac{f(x_i)}{K_n(x_i,x_i)}\frac{\pi(x_i)}{q(x_i)} - \int f\,\mathrm{d}\pi\right) \to \mathcal{N}(0,\sigma^2_{f,\pi}), \tag{2.84}$$

where $\sigma_{f,\pi}$ can be expressed in terms of the Fourier coefficients of $f\pi$.

Thanks to (2.83), we know that the estimator in (2.84) is unbiased, and thus that the mean squared error (2.27) behaves asymptotically as $\mathcal{O}(n^{-1/2-1/2d})$.

From (2.40), we know that the rate in (2.84) is close to optimal, but there is still a gap compared to the optimal $n^{-1/2-1/d}$ rate. This gap has been recently closed when $[-1,1]^d$ is replaced by a compact and complex manifold (Lemoine and Bardenet, 2024), yet at the moment this is only proved for n roughly of the form l^d for some integer l. While this rate does improve on the $n^{-1/2}$ rate of MCMC, it does downgrade with the dimension, but this is again the best we can hope for $\mathcal{C}^1$ integrands. Sampling from joint distributions like (2.82) can be done in cubic time using the so-called HKPV algorithm (Hough et al., 2009). Still, this algorithm involves rejection sampling steps, it has been assessed empirically that the number of rejections must grow as $2^d n \log n$ if we want the rejection rate to be dimension-independent (Gautier et al., 2019a).

Variance reduction with control variates

Control variates (Glynn and Szechtman, 2002) are a popular way to reduce the variance of Monte Carlo estimators. The idea is to modify the estimator $n^{-1}\sum_{i=1}^n f(x_i)$ into

$$\frac{1}{n}\sum_{i=1}^{n}\{f(x_i)-\Phi_n(x_i)\}, \tag{2.85}$$

to approximate $\int f\,\mathrm{d}\pi$. A good choice of control variate Φ_n satisfies the following. First, we usually ask for Φ_n to have mean zero under π, so that the bias term in (2.27) remains untouched. Second, Φ_n must be tractable, and it is important in the context of expensive-to-evaluate integrands f that computing Φ_n pointwise does not require many evaluations of f. Finally, Φ_n is to be tailored to yield small variance, and thus small mean squared error for the estimator (2.85).

Recently, various control variates methods have achieved optimal or close to optimal rates for smooth functions on the hypercube $[0,1]^d$. For instance, assuming that we know how do draw $x_1,\dots,x_n$ i.i.d from π, Leluc et al. (2024) considered $\Phi_n(x)$ as the sum of f over nodes $x_1,\dots,x_n$ whose Voronoï cells contains x, and obtained optimal $n^{-1/2-\alpha/d}$ mean squared error on α-Hölder functions with $0<\alpha<1$. In another line of work, Oates et al. (2017b,a) have obtained near-optimal $n^{-1/2-k/d+\epsilon}$ mean squared errors on C^k functions and for arbitrary small $\epsilon>0$. This relies again on $\mathcal{O}(n)$ points that are either i.i.d from π or drawn from a Markov chain targeting π with strong assumptions (uniform ergodicity), and the computation of (2.85) is cubic. Finally, combining control variates with local Taylor expansions of f, Chopin and Gerber (2024) have obtained optimal worst-case errors and mean squared errors on C^k functions. For now, this optimal rate is only proved for $n=l^d$ where l is an integer, so that n must still scale exponentially in the dimension d at the moment. The price to pay for is however that (2.85) relies on a quantity which computation is exponential in d, yet independent of f.

In a nutshell, RQMC algorithms usually improve on the slow $n^{-1/2}$ rate of MCMC, usually at the price of a heavier algorithmic complexity or smoothness assumptions. On the other hand, we quickly describe in the next section Variational Bayes algorithms, which focus on building fast scalable algorithms to approximately sample from the target distribution π, often at the cost of losing integration guarantees.

2.5 Variational Bayes

2.5.1 The variational approximation

Variational inference (Blei, Kucukelbir, and McAuliffe, 2017) has been trending in the framework of Bayesian inference. While we saw that MCMC provides a very general and simple method to sample from unnormalized probability distributions along with quadrature guarantees, it tends to be slow either in a theoretical way because of the

$n^{-1/2}$ rate, or in practice because of long "burn-in" phases to reach equilibrium especially for complex targets.

For these reasons, practitioners exchanged asymptotic correctness (i.e. exact convergence to the target distribution) for increasing sampling speed in experiments, this is called approximate MCMC. The idea is that the reduction in Monte Carlo variance from faster sampling can outweigh the bias introduced. Variational inference follows this philosophy. It focuses on the sampling task, and trades worst-case integration guarantees for the minimization of a dissimilarity measure $\mathcal{D}(q||\pi)$ over a set of tractable probability distributions $q \in \mathcal{Q}$.

For instance, most variational inference methods rely on the choice

$$\mathcal{D}(q||\pi) = \mathrm{H}(q||\pi) := \mathbb{E}_q \left[\log \frac{q(x)}{\pi(x)} \right], \tag{2.86}$$

assuming that q and π have densities: this is the Kullback–Leibler divergence between probability measures. First, a simple computation shows that the minimization of (2.86) does not depend on the normalization constant of π. This is key for Bayesian inference. Second, a gradient can be defined for the Kullback-Leibler divergence on the space of probability measures endowed with the Wassertein-2 metric (Chewi et al., 2024), so that this minimization problem is usually amenable to fast scalable optimization algorithms such as stochastic gradient techniques (Hoffman et al., 2013) for standard choices of $\mathcal{Q}$.

Such standard choices for the set $\mathcal{Q}$ of distributions involve Gaussian distributions with varying mean and variance, or product measures $q(x) = \prod_{j=1}^{d} q_j(x^j)$, depending on the application of interest. In those cases, π might not belong to the class $\mathcal{Q}$ so that there is indeed a bias introduced, but this can be sufficient for the purpose of inference especially if the convergence to the minimizer is fast.

Still, there are few theoretical guarantees for the resulting sampling algorithm in the general case. In particular, obtaining convergence results is quite involved but has been the topic of recent works (Alquier et al., 2016, 2020; M. Lambert et al., 2023). For instance, M. Lambert et al. (2023) show that for a Gaussian family $\mathcal{Q}$ and a strongly log-concave target π, a continuous gradient algorithm in the Wassertein-2 metric converges at exponential speed. This is turn leads to a $t^{-1/2}$ speed for the convergence of the t-th iterate of the discretized gradient algorithm to the minimizer of Equation (2.86), yet again the latter can be quite different from the target π. Moreover, the number of iterations to reach a given neighborhood of the optimum scales like $\mathcal{O}(d)$ only, which supports the claim that VB algorithms scale to high dimensions.

From the point of view of numerical integration, we would further wish $\mathcal{Q}$ to encapsulate empirical distributions of the form $\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$. We would then like to show that the solution to Equation (2.86) converges to π as n grows to infinity with good rates in terms of n, and that it is still the case for the t-th iterate of a fast optimization algorithm. A first issue in this approach is that the quantity in (2.86) is ill-defined for such q since

there is no density. A way to tackle this problem is to rely on the Stein Variational Gradient Descent algorithm which can be seen as a refinement of kernel smoothing of $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ (Chewi et al., 2024).

2.5.2 Stein Variational Gradient Descent

Stein Variational Gradient Descent (SVGD, Liu and Wang, 2019) is a Variational inference algorithm leveraging the Stein method and Stein operators, which we quickly introduce (see Anastasiou et al., 2022 for a survey). We focus on the Langevin–Stein operator defined as follows for a function $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ and a distribution π:

$$\mathcal{A}_\pi \Phi(x) = \nabla_x \log \pi(x)\Phi(x) + \nabla_x \Phi(x). \tag{2.87}$$

Consider a tractable distribution q, meaning that we know to integrate with respect to q, or at least to sample from q. We consider the minimization of (2.86) with the set $\mathcal{Q}$ given by the laws of random variables of type $Z = T(X)$ with $X \sim q$ and $T(x) = x + \epsilon\Phi(x)$. Denoting by q_T the density of Z, Liu et al. (2019) show that

$$\partial_\epsilon H(q_T||\pi)|_{\epsilon=0} = -\mathbb{E}_{x\sim q}\left[Tr\left(\mathcal{A}_\pi \Phi(x)\right)\right]. \tag{2.88}$$

This means that for this choice of $\mathcal{Q}$, we can minimize (2.86) by initializing the algorithm with the tractable distribution q, and choosing at each step the new distribution Z induced by the function Φ that gives a minimal (2.88) over a function class $\mathcal{F}$. In fact, the right-hand side of (2.88) can also be used as a natural dissimilarity measure between probability distributions q and π:

$$\mathrm{S}(q,\pi) = \max_{\Phi\in\mathcal{F}} \left|\mathbb{E}_{x\sim q}\left[Tr\left(\mathcal{A}_\pi \Phi(x)\right)\right]\right|, \tag{2.89}$$

called the *Stein discrepancy*. The definition of (2.89) is motivated by the following: under mild boundary conditions on Φ and π, it is known (Anastasiou et al., 2022) that the following integration by parts identity is satisfied

$$\mathbb{E}_{x\sim\pi}\left[\mathcal{A}_\pi \Phi(x)\right] = 0. \tag{2.90}$$

We then often say that Φ is in the Stein class of π, and (2.90) is called a Stein identity.

Considering a positive semidefinite kernel K with associated RKHS $\mathcal{H}_K$ (Definition 2.1.5), a standard example is to take $\mathcal{F}$ to be the space of functions

$$\mathcal{F}_K = \{\Phi = (f_1, \ldots, f_d) \in \mathcal{H}_K^d : \sum_{i=1}^{d} \|f_i\|_{\mathcal{H}_K} \leq 1\}. \tag{2.91}$$

In that case, it is known (Anastasiou et al., 2022) that (2.89) is maximized at

$$\Phi^* = \mathbb{E}_{x\sim q}\left[\mathcal{A}_\pi K(x,\cdot)\right]. \tag{2.92}$$

The Stein discrepancy is then called "kernel Stein discrepancy" (KSD) and denoted by $\mathrm{KSD}(q, \pi)$.

Following this observation, at each step t, SVGD produces n points $(x_i^t)_{1\leq i\leq n}$ based on the updates $x_i^{t+1} = x_i^t + \epsilon_t \hat{\Phi}^*(x_i^t)$, with (x_i^0) sampled from q, ϵ_l is a step-size decreasing to 0 sufficiently fast, and

$$\hat{\Phi}^*(x) = \frac{1}{n}\sum_{j=1}^{n}\left[K(x_j^t, x)\nabla_{x_j^t}\log\pi(x_j^t) + \nabla_{x_j^t}K(x_j^t, x)\right] \tag{2.93}$$

is the discretization of Φ^* obtained by replacing the expectation by the average over the sample at step t. Again, this does not require knowing the normalization constant of π, and we can see that each step requires $\mathcal{O}(n^2)$ operations.

Yet, SVGD still lacks strong theoretical support. Guarantees are often quantified in terms of the kernel Stein discrepancy. From (2.92), we have the following alternative expression for the KSD.

Definition 2.5.1 Let K define a RKHS as in Definition 2.1.5. The Kernel Stein Discrepancy is the dissimilarity measure given by

$$\mathrm{KSD}(q, \pi) = \sqrt{\mathbb{E}_{x\sim q, y\sim q}\left[K_\pi(x, y)\right]}, \tag{2.94}$$

where K_π is the so-called Stein kernel

$$K_\pi(x, y) = Tr\left(\mathcal{A}_{\pi,x}\mathcal{A}_{\pi,y}K(x, y)\right) \tag{2.95}$$

in which we denoted by $\mathcal{A}_{\pi,x}$ the Langevin–Stein operator (2.87) differentiating on the variable x.

In particular, the Stein kernel K_π has a closed form, depending only on $\nabla\log\pi$ and K (Anastasiou et al., 2022), and the Stein identity (2.90) implies that no averages with respect to π are needed to compute the quantity (2.94). It can be shown (Gorham and Mackey, 2020) that the KSD is a relevant metric on probability measures, since its convergence implies the weak convergence of probability measures and the convergence in usual metrics such as the Wassertein (2.1.2) and bounded Lipschitz distances (2.1.3). We will discuss in more details whether good rates on the KSD are meaningful for numerical integration in Section 2.6.

For SVGD, Korba et al. (2020) have shown a decay of the Kullback-Leibler divergence between the t-th iterate μ_t of an idealized continuous version of the algorithm (i.e. with $n \to +\infty$) and the target π, along with $\mathcal{O}(t^{-1/2})$ rates for the KSD between μ_t and π. Moreover, Shi and Mackey (2023) have shown what is to our knowledge the first rate of convergence for the actual updates of SVGD described above, see (Balasubramanian et al., 2024) for an overview of recent results. Namely, considering $\mu_n^t = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^t}$,

they show that for each t,

$$\min_{1\le r\le t} \mathrm{KSD}(\mu_n^r, \pi) = \mathcal{O}((\log\log n\delta)^{-1}), \tag{2.96}$$

with probability larger than $1 - c\delta$, yet this slow rate is not expected to be sharp at all. We should also mention that it has been shown in (Ba et al., 2021) that SVGD suffers from a curse of dimensionality for the estimation of the variance of the target π.

Actually, we can wonder whether we can perform variational inference directly for dissimilarity measures of the kind of the KSD instead of the Kullback-Leibler divergence, and this is the objective of the next section.

2.6 Kernel methods

Consider the class of energies or functionals

$$I_K(\nu - \pi) = \iint K(x,y)\,\mathrm{d}(\nu-\pi)^{\otimes 2}(x,y), \tag{2.97}$$

where K is a kernel with associated Reproducing Kernel Hilbert Space (RKHS) as in Definition 2.1.5. In particular, the KSD (2.94) is the square root of this (2.97) with kernel K_π, up to an additive constant depending only on K and π.

The energy (2.97) balances two effects: I_K is small when the point distribution ν is close to π but exhibits negative dependence encoded by the interaction K. Minimization algorithms for such functionals arose at the interface of QMC and VB. The goal is to combine small integration errors with fast minimization algorithms of a dissimilarity measure between an empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and a target distribution π. In fact, the following duality formula shows that minimizing I_K amounts to minimize the worst-case integration error in the RKHS $\mathcal{H}_K$.

Theorem 2.6.1 — Sriperumbudur et al. (2009). Let K be a measurable kernel defining a RKHS $\mathcal{H}_K$. Let $\mathcal{P}_K$ be the set of probability distributions ν such that

$$\int \sqrt{K(x,x)}\,\mathrm{d}\nu(x) < \infty. \tag{2.98}$$

Then for all $\mu, \nu \in \mathcal{P}_K$,

$$I_K(\mu-\nu) = \left(\sup_{\|f\|_{\mathcal{H}_K}\le 1} \left|\int f\,\mathrm{d}(\mu-\nu)\right|\right)^2 := \mathrm{MMD}_K^2(\mu,\nu). \tag{2.99}$$

Actually, showing that the worst-case integration in the RKHS is bounded by the square root of the energy I_K is nothing more than an instance of the Cauchy–Schwarz inequality in the RKHS $\mathcal{H}_K$.

In particular, (2.99) means that the worst-case integration error (2.19) in $\{\|f\|_{\mathcal{H}_K} \leq 1\}$ is given by $\sqrt{I_K(\nu - \pi)}$. Using the RKHS structure, we can thus hope to take the best of both worlds between leveraging low energy sequences that yield small worst-case integration errors and designing fast scalable algorithms for the minimization of I_K thanks to its simple form. Finally, denoting by

$$U_K^\mu(z) = \int K(x, z) \, \mathrm{d}\mu(z) \tag{2.100}$$

the *kernel embedding* or *potential* of $\mu \in \mathcal{P}(\mathbb{R}^d)$, the reproducing property in the RKHS $\mathcal{H}_K$ implies that

$$I_K(\mu - \nu) = \|U_K^\mu - U_K^\nu\|_{\mathcal{H}_K}^2. \tag{2.101}$$

Thus, energy minimization of $\mu \to I_K(\mu - \pi)$ also amounts to approximate U_K^π in the functional space $\mathcal{H}_K$.

We first present deterministic optimization algorithms trying to minimize the energy functional I_K, either by gradient descent schemes or greedy sequential minimization. We then discuss kernel quadrature where one tries to optimize the weights and picks the nodes $x_1, \ldots, x_n$ jointly according to a well-chosen distribution to ensure small integration errors. We end the section with a word on thinning algorithms that try to take advantage of the kernel structure to reduce the number of evaluations of the integrand f while preserving small integration errors.

2.6.1 Energy minimization and kernel herding

A natural motivation for energy minimization algorithms (apart from approximatively sampling from π) are herding algorithms that we first discuss. Herding algorithms historically arose as deterministic optimization algorithms to learn Markov random fields (MRFs) (Welling, 2009; Y. Chen and Welling, 2010). Kernel herding produces samples following the updates

$$x_{t+1} \in \arg\max_x \langle w_t \,,\, K(x, \cdot) \rangle_{\mathcal{H}_K} \tag{2.102}$$

$$w_{t+1} = w_t + \mathbb{E}_{x \sim \pi} [K(x, \cdot)] - K(x_{t+1}, \cdot), \tag{2.103}$$

where $\mathcal{H}_K$ denotes the RKHS induced by the kernel K (Definition 2.1.5). In particular, we can see from (2.101) that the updates (2.102) perform greedy minimization of $I_K\left(\frac{1}{t}\sum_{i=1}^t w_i \delta_{x_i} - \pi\right)$ (Y. Chen et al., 2012).

Kernel herding has been further proved to be a special case of the more general Frank–Wolfe conditional gradient algorithm (Bach et al., 2012), for which the following guarantees have been obtained. Denote by $\mathcal{M} \subset \mathcal{H}_K$ the convex hull of elements $K(x, \cdot)$ for all $x \in \mathbb{R}^d$, and denote by R its diameter and by $\mu_t = \frac{1}{t}\sum_{i=1}^t w_i \delta_{x_i}$ the empirical

measure at the t-th step of (2.102). Without further assumptions, it is known (Bach et al., 2012) that

$$\frac{1}{2} I_K(\mu_t - \pi) \leq 4R^2/t, \tag{2.104}$$

which induces a slow MCMC-like $t^{-1/2}$ decay for the MMD between μ_t and π.

From the point of view of QMC methods and the lower bound (2.39), we would want to achieve a faster rate. Actually, if the RKHS $\mathcal{H}_K$ is finite dimensional, it can be proved that $I_K(\mu_t - \pi)$ decays at least in $\mathcal{O}(t^{-2})$ or even exponentially fast for a generalization of kernel herding (Y. Chen et al., 2012; Bach et al., 2012). Still, the assumptions made to get these rates are in fact never true in the infinite dimensional setting (Bach et al., 2012), although it has been observed in experiments (Y. Chen et al., 2012) that a faster rate indeed holds.

Moreover, the kernel herding algorithm comes with practical issues. First, the update (2.102) requires to compute the kernel embedding U_K^π at each step to update the weights. This is usually approximated via MCMC or by averaging over previous sample points, but such an approximation is not included in the rate guarantees for the algorithm. Second, Bach et al. (2012) provide an example where the maximization step in (2.102) is NP-hard. In practice, Y. Chen et al. (2012) and Bach et al. (2012) rely on zero-th order optimization methods and thus usually scale poorly with the dimension.

Other gradient descent or greedy minimization algorithms have been studied for I_K in the literature. For instance, Arbel et al. (2019) perform standard and noisy gradient descent on (2.97). Mak and Joseph (2018) build a sequence of points such that

$$\mathrm{MMD}_K(\mu_n, \pi) = \mathcal{O}\left(n^{-1/2} \log^{-(\alpha-1)/2}\right), \tag{2.105}$$

under some assumption involving $\alpha > 1$, and for a specific choice of interaction K. In the latter, the computation of U_K^π is bypassed by assuming that we have access to i.i.d draws of π. A particular focus has been put on the case where K is replaced by the Stein kernel K_π (2.95). We describe a few key approaches in the following.

2.6.2 Energy minimization with Stein discrepancies

Consider again the energy I_K in (2.97). Expanding the integrals, we easily obtain

$$I_K(\nu - \pi) = \iint K(x,y)\, \mathrm{d}\nu^{\otimes 2}(x,y) - 2\int U_K^\pi(z)\, \mathrm{d}\nu(z) + \iint K(x,y)\, \mathrm{d}\pi^{\otimes 2}(x,y), \tag{2.106}$$

and we thus see that minimizing I_K will require to compute the potential (2.100) of the target distribution π. The kernel Stein discrepancy (2.94, Anastasiou et al., 2022) provides a way to bypass this computation. Indeed, relying on the Stein identity (2.90)

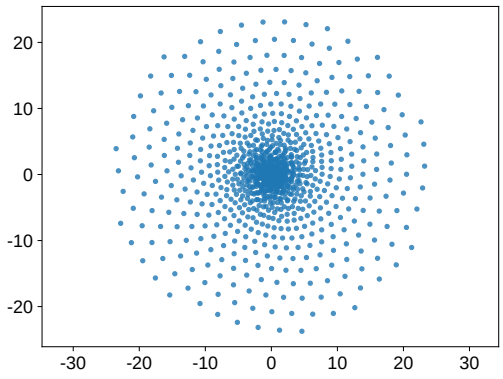


Fig. 2.3: $n = 1000$ **points produced through the KSD descent algorithm with** $T = 50\,000$ **iterations.**

and on integration by parts, it can be shown (Anastasiou et al., 2022) that $U^{\pi}_{K_\pi} = 0$ under mild conditions over K and π. In that case, minimizing I_{K_π} can be done without evaluating kernel embeddings with respect to π.

Optimization algorithms for the KSD have been investigated for instance in (Korba et al., 2021) with gradient descent schemes, or in (W. Y. Chen et al., 2018, 2020) via greedy optimization algorithms. For instance, we show in Figure 2.3 $n = 1000$ particles produced via the KSD descent algorithm[2] of (Korba et al., 2021), for a Gaussian target measure. We see that the points are very well-ordered in the space, with more density in the regions of large mass under the target. In particular, (W. Y. Chen et al., 2018, 2020) show a $n^{-1/2}$ MCMC-like decay for the KSD, which gives worst-case integration guarantees for functions in the RKHS of kernel K_π due to (2.99).

Yet, using the KSD instead of the full MMD (2.99) comes with some issues from the point of view of numerical integration. In fact, while the kernel K_π has explicit form, it is unclear what functions belong to the RKHS $\mathcal{H}_{K_\pi}$ for given K and π, or what the link with the original RKHS $\mathcal{H}_K$ is. Bénard et al. (2023) further present some pathological consequences. For numerical integration, not being able to prescribe the functions we allow ourselves to integrate is a bit disappointing.

We could however argue that the KSD or its variants are known (Gorham et al., 2020; Kanagawa et al., 2025) to control classical convergences on probability measures, such as convergence in bounded Lipschitz distance or Wasserstein distance which are worst-case integration errors on broad function classes, see Section 2.1.1. Explicit bounds have for instance been recently investigated in (Kanagawa et al., 2025). Yet, simple algorithms such as MCMC are already close to optimal for the Wassertein distance (2.3), and there are to our knowledge to explicit bound that enables to transfer rates on the KSD to

[2] https://github.com/pierreablin/ksddescent

rates on the integration error in the original $\mathcal{H}_K$ at the moment. On a side note, recent upper bounds on the eigenvalues σ_{n+1} from (2.20) have been obtained for the Stein kernel K_π, for large classes of the original σ_{n+1} for K, see (Li et al., 2024). In view of (2.39), we can thus also have a rough idea of optimal integration errors in the RKHS $\mathcal{H}_{K_\pi}$.

Gradient-based algorithms for the minimization of the MMD or the KSD are still an active field of research, and a lot of effort has been put up to design generalizations of KSDs with better properties, see (Kanagawa et al., 2025) for instance. We now turn to two alternative approaches to provide faster convergence rates for the MMD.

2.6.3 Kernel quadrature and volume sampling

Conditionally on the points $x_1, \dots, x_n$, optimal kernel quadrature algorithms try to minimize the worst-case integration in the RKHS by jointly optimizing non-uniform weights $w_1, \dots, w_n$. This differs for instance from standard kernel herding (2.102) where the weights are chosen sequentially via gradient updates. Given $x = (x_1, \dots, x_n) \in \mathbb{R}^{dn}$, the energy

$$I_K\left(\sum_{i=1}^n w_i \delta_{x_i} - \pi\right) = \tag{2.107}$$

$$\iint K(x,y)\,\mathrm{d}\pi - 2\sum_{i=1}^n w_i U_K^\pi(x_i) + \sum_{i,j} w_i w_j K(x_i, x_j) \tag{2.108}$$

is minimized for

$$w_n^* = (K(x_i,x_j))_{i,j}^{-1} \left(U_K^\pi(x_1), \dots, U_K^\pi(x_n)\right)^\top. \tag{2.109}$$

One natural approach is thus to select $x_1, \dots, x_n$ in such a way that we are able to prove small integration errors by choosing the approximation $\nu_n = \sum_{i=1}^n w_{n,i}^* \delta_{x_i}$. Assume that K admits the Mercer decomposition

$$K(x,y) = \sum \sigma_m \Phi_m(x) \Phi_m(y), \tag{2.110}$$

as in Proposition 2.1.7. Inspired by the OPEs (2.82), Belhadji et al. (2019) show that if K_n is the truncation of K onto the first n eigenvectors in (2.110), then drawing $x_1, \dots, x_n$ from the determinantal point process (2.82) with kernel K_n and any reference probability μ yields

$$\mathbb{E}\,\mathrm{MMD}_K^2\left(\sum (w_n^*)_i \delta_{x_i}, \pi\right) \leqslant 2\sigma_{n+1} + 2n^2 r_n, \tag{2.111}$$

where $r_n = \sum_{m \geq n+1} \sigma_m$. As soon as σ_m decays only polynomially in m, e.g. for Sobolev spaces, the second term of (2.111) dominates. In that case, the bound is then suboptimal in view of the lower bound (2.39).

Yet, optimal bounds can be obtained for a slightly different choice of nodes $x_1, \dots, x_n$, as stated in the following.

Theorem 2.6.2 — Belhadji et al. (2020), Theorem 4. Assume that the kernel K satisfies a Mercer decomposition as in Proposition 2.1.7. Let μ be a fixed probability measure, and consider the joint distribution

$$x_1, \dots, x_n \sim \frac{1}{Z} \det(K(x_i, x_j)) \, \mathrm{d}\mu(x_1) \, \dots \, \mathrm{d}\mu(x_n) \tag{2.112}$$

called continuous volume sampling. Then, for any $\pi \in L^2(\mu)$,

$$\mathbb{E} \, \mathrm{MMD}_K^2 \left(\sum (w_n^*)_i \delta_{x_i}, \pi \right) \leq \sigma_n (1 + \beta_n), \tag{2.113}$$

where β_n is this time bounded as soon as σ_m decays at least polynomially in m.

This time, the rate matches the optimal one from (2.39). While we are quite happy with this fast rate, two comments are again needed from the numerical integration perspective.

First, computing the weights (2.109) requires again to compute the kernel embedding U_K^π. Moreover, one can show (Belhadji et al., 2020) that the joint distribution (2.112) is nothing more than a mixture of determinantal distributions of the form (2.82). Without further argument, sampling from this mixture is cubic in n and may require $2^d n \log n$ rejection steps, see the discussion in Section 2.4.2.

Finally, a drawback of this quadrature with optimal weights is that there is little hope for a CLT as in (2.84), since each $(w_n^*)_i$ is a function of the whole set of points $x_1, \dots, x_n$.

2.6.4 Thinning algorithms

Thinning algorithms aim at reducing the number of evaluations of the integrand f while keeping small integration errors. Once we produce a sample $x_1, \dots, x_n$ using e.g. an MCMC algorithm, we can be interested in subselecting a small number of points $x_{s_1}, \dots, x_{s_m}$ that summarize well the initial sample. In particular, we are usually interested in the regime $m \ll n$ where the number of evaluation of the integrand f in $\frac{1}{n} \sum_{i=1}^n f(x_i)$ is drastically reduced from n to m, see Chapter 1 for motivations.

An easy way to do so is to choose the indices $s_1, \dots, s_m$ i.i.d uniformly among $\{1, \dots, n\}$. Assuming for instance that

$$\mathrm{MMD}_K \left(\frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \pi \right) \in \mathcal{O}(n^{-1/2}), \tag{2.114}$$

with high probability, and choosing $m = \sqrt{n}$, one gets

$$\mathrm{MMD}_K \left(\frac{1}{m} \sum_{j=1}^m \delta_{x_{s_j}}, \pi \right) \in \mathcal{O}(n^{-1/4}), \tag{2.115}$$

with high probability, (Dwivedi et al., 2021). Designing subsamples with $m \ll n$ while preserving the rate in (2.114) is a challenging task, see for instance (Riabiz et al., 2022; Chopin et al., 2021) for practical algorithms.

Actually, thinning can as well serve for sampling or numerical integration: if we are ready to spend some computational budget to draw n points using e.g. MCMC, and if we are able to ensure that the integration error over m subsampled points does not downgrade to the $m^{-1/2}$ rate, then this yields a numerical integration algorithm with better-than-i.i.d integration error over m points. This is for instance reasonable when the integrand f is expensive to evaluate as explained in Chapter 1: in that case we are happy to spend more computational budget to draw the m points rather than evaluating the integrand over $n \gg m$ nodes.

We quickly describe two classes of such algorithms that provably obtain small integration errors for the MMD in terms of the subsampled points. The first one iterates a halving procedure to split the dataset $x_1, \dots, x_n$ into smaller data sets while ensuring that the energy of the point configuration is close to optimal in terms of the MMD at each step. The second one builds on the optimal quadrature weights from Section 2.6.3 and draws nodes independently using a kernel-related notion of importance.

Kernel thinning

Let $x_1, \dots, x_n$ be drawn from a geometrically ergodic Markov chain targeting π, and consider uniform weights $w_i = 1/n$. Recall from (2.72) that we then have the following decay for the MMD:

$$\mathrm{MMD}_K\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, \pi\right) = \mathcal{O}(n^{-1/2}), \tag{2.116}$$

with high probability.

In kernel thinning (Dwivedi et al., 2021, 2024), the objective is to split $\{x_1, \dots, x_n\}$ into two equal-size sets $\mathcal{S}_1$ and $\mathcal{S}_2$ while preserving small integration error in each of the subsets. To do so, one considers for each $i = 1, \dots, n/2$ the couple (x_{2i}, x_{2i+1}) and assigns x_{2i} to $\mathcal{S}_1$ and x_{2i+1} to $\mathcal{S}_2$ with a probability p_i that accounts for the integration error obtained when choosing this assignment rather than the other way around.

Repeating this halving operation m times with m of the order of $\log_2(n)$, yields $\sqrt{n}$ subsets of $\{x_1, \dots, x_n\}$ each of size $\sqrt{n}$. Finally, we iteratively refine the result by choosing the final subset via a greedy minimization of

$$\mathrm{MMD}_K\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{\sqrt{n}}\delta_{x_{s(j)}}, \frac{1}{\sqrt{n}}\sum_{j=1}^{\sqrt{n}}\delta_{x_{s^*(j)}}\right), \tag{2.117}$$

where $x_{s^*(1)}, \dots, x_{s^*(\sqrt{n})}$ is the best of the $\sqrt{n}$ subsets obtained after the m halving rounds, in terms of MMD with respect to $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. We plot the resulting point set of

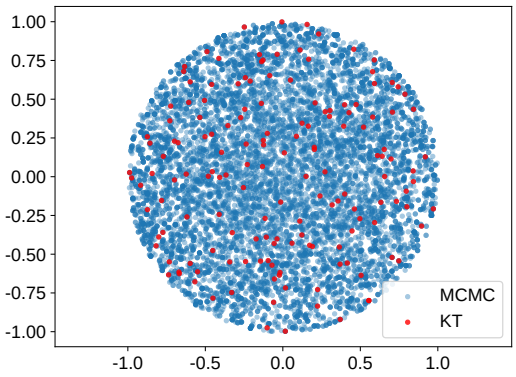


Fig. 2.4: $\sqrt{n} = 100$ **points produced through the kernel thinning (KT) algorithm, with** $n = 10\,000$ **MCMC input points.**

the kernel thinning algorithm[3] in Figure 2.4, with $n = 10\,000$ MCMC inputs points and thus $\sqrt{n} = 100$ thinned points, on a simple example where π is the uniform measure on the two-dimensional unit ball.

One can then show fast rates on the resulting MMD, depending on the properties of K and π. We subselect a few specific cases given in (Dwivedi et al., 2024, Table 3) in which the guarantees indeed improve on the i.i.d rate.

Theorem 2.6.3 — Dwivedi et al. (2024). Let π have subgaussian tails, and let $K(x, y) = \exp\left(-\frac{1}{2\sigma^2}|x - y|^2\right)$ be a Gaussian kernel. Then

$$\mathrm{MMD}_K\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{\sqrt{n}}\delta_{x_{s(j)}}, \frac{1}{\sqrt{n}}\sum_{j=1}^{\sqrt{n}}\delta_{x_{s^*(j)}}\right) = \mathcal{O}\left(\frac{\log^{3d/4+1} n}{\sqrt{nc_n^d}}\right), \tag{2.118}$$

with high probability, and where $c_n = \log\log n$.

Let π have sub-exponential tails and let $K(x, y) = \frac{1}{(1+|x-y|^2/\gamma^2)^\nu}$ be an inverse multiquadratic kernel with $\nu > 0$. Then

$$\mathrm{MMD}_K\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{\sqrt{n}}\delta_{x_{s(j)}}, \frac{1}{\sqrt{n}}\sum_{j=1}^{\sqrt{n}}\delta_{x_{s^*(j)}}\right) = \mathcal{O}\left(\frac{\log^{d+1} n}{\sqrt{n}}\right), \tag{2.119}$$

with high probability.

As for QMC-like guarantees in Section 2.4, the rate $n^{-1/2}$ is very good for a quadrature

[3] https://github.com/microsoft/goodpoints

with $\sqrt{n}$ points, but the bounds (2.118) and (2.119) are only interesting when the number of input points n is at least exponential in the dimension d. In practice, there is nonetheless empirical evidence (Dwivedi et al., 2021, 2024) that kernel thinning scales to moderate dimensions.

From (2.39), we should keep in mind that there is still room for improvement on the rates e.g. for the Gaussian kernel. Yet, it has been shown in (Phillips and Tai, 2020) that when compressing n input points to $\sqrt{n}$ final nodes, we must suffer a MMD error that is at least of the order of $n^{-1/2}$. In that sense, with this number of input points, the rates of Theorem 2.6.3 cannot be improved apart from the logarithmic term.

Finally, kernel thinning also comes with fast integration guarantees for single integrands in a given RKHS.

Theorem 2.6.4 — Dwivedi et al. (2024), Theorem 1. Assume for instance that K is bounded and defines a RKHS $\mathcal{H}_K$ (Definition 2.1.5). Then, for any $f \in \mathcal{H}_K$,

$$\left| \frac{1}{\sqrt{n}} \sum_{j=1}^{\sqrt{n}} f(x_s(j)) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right| = \mathcal{O}\left(\sqrt{\frac{\log n}{n}} \right), \tag{2.120}$$

with high probability.

Combining (2.120) with (2.116) yields a fast $n^{-1/2}\sqrt{\log n}$ rate for numerical integration with $\sqrt{n}$ points: this is a quadratic improvement over the MCMC CLT (2.73) for functions in $\mathcal{H}_K$. To my knowledge, this is the only existing result to obtain fast rates for the integration guarantees of single integrands in the RKHS $\mathcal{H}_K$, and I think this is a very strong achievement. Note in particular that the rate does not depend on the dimension d this time. We will discuss this result in more detail in Section 6.1.2.

The cost of the whole procedure is dominated by $\mathcal{O}(n^2) = \mathcal{O}(m^4)$ in time for the kernel evaluations and $\mathcal{O}(n \min(d, n)) = \mathcal{O}(m^2 \min(d, m^2))$ in memory due to the need to store either the input data $x_1, \ldots, x_n$ or the kernel matrix $K(x_i, x_j)$. Refinements of this procedure (Shetty et al., 2022; Li et al., 2024) have led to computational costs that are only $\mathcal{O}(n(\log n)^3)$ in time and $\mathcal{O}(\sqrt{n}(\log n)^2)$ in memory.

In the next section, we consider another thinning algorithm, that achieves optimal rates for the MMD at the price of a larger number of input points.

Thinning via leverage score sampling

Assume that we are given $x_1, \ldots, x_n$ drawn i.i.d from π. We consider the optimal quadrature weights given by

$$\hat{w}_n^* = (K(x_i, x_j))_{i,j}^{-1} \left(U_K^{\mu_n}(x_1), \ldots, U_K^{\mu_n}(x_n) \right)^\top, \tag{2.121}$$

where $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. This is similar to (2.109) but with μ_n replacing π. In particular, note that this time no kernel embedding evaluation with respect to π is required.

Define $\mathbf{K}_n = (K(x_i, x_j))_{1\leq i,j\leq n}$ and consider the ridge leverage score defined as

$$\ell_\lambda(i) = (\mathbf{K}_n(\mathbf{K}_n + \lambda n I)^{-1})_{ii}, \tag{2.122}$$

for $1 \leq i \leq n$. Chatalic et al. (2025) proposed to draw for each $j = 1, \dots, m$, $s(j)$ i.i.d from $\{1, \dots, n\}$ with

$$\mathbb{P}(s(j) = i) = \widehat{\ell_\lambda}(i), \tag{2.123}$$

where $\widehat{\ell_\lambda}(i)$ is an approximation of $\ell_\lambda(i)$ that reduces computational costs. Building the subset $(x_{s(j)})_{1\leq j\leq m}$ can then be done in cubic time, and we have the following result.

Theorem 2.6.5 — Chatalic et al. (2025). Let K admit a Mercer decomposition as in Proposition 2.1.7.

If the spectrum of K decays polynomially $\sigma_i \leq a_\gamma i^{-1/\gamma}$, and $m = \mathcal{O}(n^\gamma \log^{1-\gamma} n)$,

$$\mathrm{MMD}_K\left(\sum_{j=1}^m (\hat{w}_n^*)_{s(j)} x_{s(j)}, \pi\right) = \mathcal{O}\left(\frac{\log^{1/(2\gamma)} m}{m^{1/(2\gamma)}}\right), \tag{2.124}$$

with high probability.

If the spectrum of K decays exponentially $\sigma_i \leq a_\beta e^{-\beta i}$, choosing $m = \Omega(\log^2 n)$,

$$\mathrm{MMD}_K\left(\sum_{j=1}^m (\hat{w}_n^*)_{s(j)} x_{s(j)}, \pi\right) = \mathcal{O}\left(\frac{m^{1/4}}{\exp\left(\sqrt{m}/c\right)}\right), \tag{2.125}$$

with high probability, and where $m = \Omega(\log^2 n)$ means that m must be exactly of this order.

In the Gaussian case for instance, the exponential decay in (2.125) is much better than the corresponding m^{-1} rate obtained in (2.118). This is not that surprising, since we mimicked the optimal choice of weights (2.109) in (2.121) whereas we stuck to uniform weights in kernel thinning. In fact, we already saw in Section 2.6.3 that optimizing the weights is one of the ingredients to achieve fast rates and even lower bounds for the MMD, whereas the guarantees we get on the MMD for minimization methods relying on uniform weights are usually less kernel-dependent, see Section 2.6.1, Section 2.6.2 and (Dwivedi et al., 2024, Table 3).

For Sobolev RKHSs, we saw in Section 2.1.2 that we have polynomial decay with $\gamma = d/(2s)$. This means that (2.124) achieves the optimal $\mathcal{O}(m^{-s/d})$ rate for m integration nodes (2.36), up to a logarithmic term $\log^{s/d} m$. Again, this logarithmic term means that the bound (2.124) is only interesting when m is exponential in s/d, this seems less of a worry than assuming that m needs to scale exponentially in d.

Yet, if we want to use this result for numerical integration with m nodes, we see from the choice of m in (2.125) that the number of i.i.d input points n must scale exponentially in the number of integration nodes m for the Gaussian kernel, whereas we only needed m^2 input points from a geometrically ergodic Markov chain for kernel thinning with any kernel (Theorem 2.6.3).

2.7 Conclusion

In a nutshell, classical integration rules such as Gaussian quadrature (Section 2.2.1) or independent sampling (Section 2.2.2) provide optimal quadrature bounds under low assumptions on the integrands: they achieve the optimal $n^{-1/d}$ worst-case error e.g. on $\mathcal{C}^1([0,1]^d)$ or Lipschitz functions for the worst-case error.

On the other hand, independent sampling achieves the dimension independent and optimal $n^{-1/2}$ rate for the mean squared error on L^2 functions, and makes uncertainty quantification possible through probabilistic results such as CLTs.

MCMC (Section 2.3) mimics the guarantees of independent sampling without assuming that we are able to draw points from π, but rather relying on $\mathcal{O}(n)$ evaluations of the unnormalized density π. It is thus often a very strong candidate algorithm to perform numerical integration in moderate to high dimensional Bayesian settings. Yet, those rates suffer either from a curse of dimensionality for the worst-case error, or are deemed to be too slow in practice when it is not computationally feasible to evaluate the integrand on a large number of nodes n, see Chapter 1 for examples of complex Bayesian models.

Variational Bayes (Section 2.5) proposes a promising alternative to MCMC for sampling algorithms that scale to high dimensions, yet there is still a large theoretical gap to be filled. In particular, we only have access to weak guarantees on the approximation to the target.

One way to circumvent these issues is to assume that the integrands are smoother. Assuming that the integrands belong to a Sobolev space $W_{p,k}^{mix}$ with dominated mixed smoothness of order k, QMC methods (Section 2.4) build point configurations that are close to the optimal $\mathcal{O}(n^{-k}\log^{(d+1)/2} n)$ rate, which is seemingly fast. The logarithmic power in the optimal rate means however that the number of nodes n must be exponential in d to get meaningful integration guarantees, and vanilla QMC methods thus poorly scale even to moderate dimensions.

Randomized QMC methods (Section 2.4.2) rather try to achieve optimal guarantees for the mean squared error. Again, a common assumption is that the integrand belongs to $C^k([0,1]^d)$ or to $W_{p,k}^{mix}$, in which case the optimal rates are respectively given by $\mathcal{O}(n^{-1/2-k/d})$ and $\mathcal{O}(n^{-1/2-k})$, with no logarithmic penalty.

Randomization of low discrepancy sequences achieve close to optimal

$\mathcal{O}(n^{-k-1/2}\log^{(k/2+1/2)d} n)$ rates on $W_{p,k}^{mix}$, which still requires n to be exponential in d. On the other hand, optimal constructions with control variates or determinantal point processes requires $\mathcal{O}(n^3)$ operations but still an exponential number of operations in the dimension d. Randomized QMC are thus quite tailored for low dimensional settings: the gain in the optimal rates become marginal for moderate and high dimensions, and practical constructions still scale exponentially with the dimension at the moment.

Assuming that the set of integrands rather exhibit a RKHS structure, kernel methods (Section 2.6) aim to balance scalable algorithms and small integration errors in the RKHS, i.e. small MMD. In that case, the worst-case integration error in the RKHS coincides with an energy functional on probability distributions. Moreover, we should keep in mind that the RKHS structure assumed for the set of integrands allows to get rid of the dimension in the rates, for instance MCMC achieves $\mathcal{O}(n^{-1/2})$ worst-case integration error in the RKHS. Minimizing this energy with zero-th or first order optimization methods has led to efficient algorithms in low and moderate dimensions, yet with no theoretical support on any improvement over MCMC guarantees.

Relying on n^2 input points from an ergodic Markov chain, kernel thinning has led to faster than i.i.d guarantees on the MMD, yet only when n is again exponential in the dimension in most cases. Still, there is empirical evidence that kernel thinning scales to moderate dimensions.

Finally, optimizing the quadrature weights, volume sampling and thinning with leverage scores have achieved optimal or close to optimal worst-case errors in RKHSs. The current price to pay is either an exponential number of rejection steps in the dimension or a very large number of i.i.d input points.

To my knowledge, strong guarantees for the mean squared error of single integrands in a given RKHS have only been obtained by kernel thinning (Theorem 2.6.4) and have otherwise remained unexplored in the literature. In particular, we currently have no knowledge of what a general optimal bound would be for this quantity.

The starting-block of the work conducted in this thesis is the following. Inspired by quadrature with determinantal point processes and the literature in statistical physics that we will present in the next chapter, we will draw the nodes $x_1, \dots, x_n$ from a joint distribution that favors point configurations with low MMD.

Our general goal will be to provide guarantees for the worst-case integration error or the mean squared error in RKHSs with such random nodes $x_1, \dots, x_n$.

Leveraging probabilistic tools such as large deviation principles and concentration inequalities, we hope to improve on MCMC guarantees while staying computationally tractable in low and moderate dimensions. Choosing uniform quadrature weights $w_1, \dots, w_n$, our long term goal would be to obtain fast CLTs for the integration error of functions belonging to a given RKHS, thus giving strong mean squared error guarantees for integrands in the RKHS along with uncertainty quantification.

Repulsive point processes for quadrature

"Devant lui, à cinq pieds du sol une forme cauchemardesque se dresse menaçante et toute vibrante d'une terrifiante énergie intérieure."

Blake et Mortimer : Les Trois Formules du Professeur Satō, Edgar P. Jacobs

3

3.1 Generalities on point processes 65

3.1.1 Definitions 65

3.1.2 Negative dependence in point processes 67

3.2 Some examples 71

3.2.1 Binomial and Poisson point processes 71

3.2.2 (Perturbed) Lattices 72

3.2.3 Eigenvalues of random matrices 73

3.2.4 Determinantal point processes 75

3.3 Gibbs measures 77

3.3.1 Motivations and definitions 77

3.3.2 Large deviations and concentration 79

3.3.3 Gaussian fluctuations in low dimensions 82

3.4 Conclusion 83

3.1 Generalities on point processes

3.1.1 Definitions

We first give some general notions on point processes, which can be informally thought of as a random point clouds. Such processes have been widely used for modelling in spatial statistics (Stoyan et al., 1995; Moller and Waagepetersen, 2003).

Definition 3.1.1 We say that $\gamma \subset \mathbb{R}^d$ is a locally finite configuration of points in $\mathbb{R}^d$ if, for any compact set $K \subset \mathbb{R}^d$, $|\gamma \cap K| < +\infty$.

We denote by $\mathrm{Conf}\left(\mathbb{R}^d\right)$ the space of locally finite configurations, and we endow it with the σ-algebra generated by the counting maps

$$N_K : \gamma \in \mathrm{Conf}\left(\mathbb{R}^d\right) \mapsto |\gamma \cap K|$$

for any compact K.

A point process X on $\mathbb{R}^d$ is a random variable taking values in $\mathrm{Conf}\left(\mathbb{R}^d\right)$, meaning that for each ω, $X(\omega)$ is a locally finite configuration on $\mathbb{R}^d$.

Alternately, we can as well work with the following equivalent definition in terms of random measures (Last and Penrose, 2018).

Definition 3.1.2 Consider the space $\mathcal{U}_{\mathrm{atom}}\left(\mathbb{R}^d\right)$ of atomic measures of the form $\sum_{i\in I} \delta_{x_i}$ where $(x_i)_{i\in I} \in \left(\mathbb{R}^d\right)^I$ with I countable. We endow it with the σ-algebra generated by the counting maps $\mu \in \mathcal{U}_{\mathrm{atom}}\left(\mathbb{R}^d\right) \mapsto \mu(K)$ for any compact $K \subset \mathbb{R}^d$.

A point process X on $\mathbb{R}^d$ is a random variable taking values in $\mathcal{U}_{\mathrm{atom}}\left(\mathbb{R}^d\right)$, and we say that X is simple when $X = \sum_{i\in I} \delta_{x_i}$ has no multiple points almost-surely.

For a point configuration $\gamma = \{x_1, \ldots, x_n\}$ with $x_1, \ldots, x_n \in \mathbb{R}^d$, we will extensively use the following notation.

Definition 3.1.3 — Empirical measure. For $x_1, \ldots, x_n \in \mathbb{R}^d$, we define the empirical measure by

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}. \tag{3.1}$$

A point process X can be characterized by the following factorial moment measures. We begin with the first-order case.

Definition 3.1.4 The first factorial moment measure — or intensity measure — ρ_1 is the measure that satisfies

$$\mathbb{E}\left[\sum_{x\in X} f(x)\right] = \int f(x)\,\mathrm{d}\rho_1(x), \tag{3.2}$$

for any non-negative bounded measurable test function f.

When ρ_1 has a density with respect to a reference measure μ, we often abuse notations by denoting the density — or intensity function — by ρ_1 as well, and we have

$$\mathbb{E}\left[\sum_{x\in X} f(x)\right] = \int f(x)\rho_1(x)\,\mathrm{d}\mu(x). \tag{3.3}$$

Knowledge of the intensity ρ_1 of a point process is already enough to implement a numerical integration method. Assume that we are interested in computing an approximation to $\int f(x)\,\mathrm{d}x$ for f supported on a compact K, and consider a point process X the intensity measure of which has a density ρ_1 with respect to the reference measure μ given by the Lebesgue measure. We then sample the realization of a point process X in a bounded window W — say the smallest ball containing K — and compute the estimator

$$\sum_{x\in X\cap W} \frac{f(x)}{\rho_1(x)}. \tag{3.4}$$

In view of (3.3), this is an unbiased estimator of $\int f\,\mathrm{d}\pi$. By definition of a point process, the sum in (3.4) is finite, yet with a possibly random number of points given by $|X\cap W|$. The expected number of points is given by

$$\rho_1(W) = \int_W \rho_1(x)\,\mathrm{d}x. \tag{3.5}$$

When X has n points $\{x_1,\dots,x_n\}$ in W almost surely, (3.4) takes the usual Monte Carlo form

$$\sum_{i=1}^{n} \frac{f(x_i)}{\rho_1(x_i)}. \tag{3.6}$$

We now define higher-order factorial moment measures.

Definition 3.1.5 For $k\geq 2$, the k-th factorial moment measure of X is defined as the

measure M_k satisfying

$$\mathbb{E}\left[\sum_{x_1,\dots,x_k\in X}^{\neq} f(x_1,\dots,x_k)\right] = \int f(x_1,\dots,x_k)\,\mathrm{d}M_k(x_1,\dots,x_k), \tag{3.7}$$

for any non-negative and bounded measurable test function f, and where the superscript in the sum means that we only consider k-uples with distinct points $x_1,\dots,x_k \in X$. If M_k has a density ρ_k with respect to $\mu^{\otimes k}$, this density ρ_k is called the k-th correlation function of the point process X.

The case $k = 2$ is usually particularly interesting: in that case ρ_2 is called the pair correlation function.

We saw in Sections 2.4.2 and 2.6 that a solution to get optimal quadratures is to consider negatively dependent points $x_1,\dots,x_n$, i.e. points repelling each other. Intuitively, when the integrand f is smooth, we don't gain information by evaluating it on points that are close to each other. We would thus rather design point $x_1,\dots,x_n$ that spread well in the integration domain in order to maximize the information we gain through n evaluations of f. In the following, we give various encodings of negative dependence for point processes and then turn to some standard examples.

3.1.2 Negative dependence in point processes

Negative dependence can be informally defined by the inequality

$$\mathbb{P}\left(x \in A\cap X \text{ and } y \in A\cap X\right) < \mathbb{P}(x \in A\cap X)\mathbb{P}(y \in A\cap X), \tag{3.8}$$

for any bounded region $A \subset \mathbb{R}^d$ and for any $x \neq y \in \mathbb{R}^d$. This means that the probability of jointly observing two distinct points in a given window A is always smaller than the marginal probabilities of observing each point in A independently. Equivalently,

$$\mathbb{P}\left(x \in A\cap X \,|\, y \in A\cap X\right) < \mathbb{P}(x \in A\cap X) \tag{3.9}$$

meaning that there is repulsion between the points of X: the observation of y in A reduces the probability to observe another point in that region. Formally, we can take the following inequality as a definition by summing over all pairs of points in X:

$$\mathbb{E}\left[\sum_{x_1\neq x_2\in X} \mathbb{1}_{x_1\in A, x_2\in B}\right] \leq \mathbb{E}\left[\sum_{x\in X}\mathbb{1}_{x\in A}\right]\mathbb{E}\left[\sum_{x\in X}\mathbb{1}_{x\in B}\right], \tag{3.10}$$

for all bounded $A, B \subset \mathbb{R}^d$, i.e.

$$M_2(A\times B) < \rho_1(A)\rho_2(B). \tag{3.11}$$

Given $x, y \in \mathbb{R}^d$, using (3.11) for all $\epsilon > 0$ and $A_\epsilon = B(x,\epsilon)$ and $B_\epsilon = B(y,\epsilon)$ gives

$$\rho_2(x,y) < \rho_1(x)\rho_1(y). \tag{3.12}$$

Again, this can be taken as a definition for negative correlation, and the left-hand side of (3.12) corresponds to the case where the points of X are independent. Correlations functions thus quantify the repulsion between the points of X locally. Still, they are generally not explicit and providing good estimates is often a hard task. For $X = \{x_1, \dots, x_n\}$, other quantities of interest are for instance the minimal separation or particle gap

$$r(X) = \min\{|x_i - x_j| \, : \, i \neq j\}, \tag{3.13}$$

or the local discrepancy

$$\mathrm{D}(\Omega, \mu) = |X \cap \Omega| - n\mu(\Omega), \tag{3.14}$$

that measures the mismatch in the number of particles between X and a reference distribution μ locally. Intuitively, $r(X)$ is large when there is repulsion between the points, while $\mathrm{D}(\Omega, \mu)$ is small for all bounded Ω when the points are "well-spread". Those are however complicated functions in $x_1, \dots, x_n$ and are usually hard to control. We usually rather turn to simpler functions given by averages of *test functions* f over a realization of X

$$\langle f, \mu_n \rangle := \frac{1}{n} \sum_{i=1}^{n} f(x_i) = \int f \, \mathrm{d}\mu_n, \tag{3.15}$$

where μ_n denotes the empirical measure (3.1). Quantities like (3.15) are called *linear statistics* in the literature of point processes and statistical physics; we will focus on these quantities in the manuscript for purpose of numerical integration. In the following, we quickly survey probabilistic machinery where a trace of negative dependence can be found.

Deviations and worst-case errors

First, we are usually interested in quantifying the deviations of the linear statistics $\langle f, \mu_n \rangle$ from its equilibrium value $\langle f, \mu \rangle = \int f \, \mathrm{d}\mu$. This will be our main object of study in this manuscript, with $\mu = \pi$ given by the target distribution of numerical integration.

We can rely on the theory of large deviations (LDPs, Dembo et al., 2010). Roughly, we say that the family of measures $(\mu_n)_{n \in \mathbb{N}}$ satisfies a large deviation principle with rate function I and speed a_n if, for any μ,

$$\mathbb{P}\left(\mu_n \approx \mu\right) \sim \exp\left(-a_n I(\mu)\right). \tag{3.16}$$

(3.16) means that μ_n gets exponentially close to distributions μ with low *energy* $I(\mu)$. Again, we will be interested in the case where μ_n is the empirical measure (3.1) of a point process and $\mu = \pi$ is the target distribution.

Definition 3.1.6 Let $I : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R} \cup \{\pm\infty\}$ be a functional over probability measures. We endow $\mathcal{P}(\mathbb{R}^d)$ with the topology of weak convergence. We say that I is a good rate function when it has compact sublevel sets.

We say that the family of measures $(\mu_n)_n$ on $(\Omega, \mathcal{A}, \mathbb{P})$ satisfies a large deviation principle at speed a_n and with good rate function I if for any Borel set $A \subset \mathcal{P}(\mathbb{R}^d)$,

$$-\inf_{\mu \in \mathring{A}} I(\mu) \leq \underline{\lim} \frac{1}{a_n} \log \mathbb{P}(\mu_n \in A) \leq \overline{\lim} \frac{1}{a_n} \log \mathbb{P}(\mu_n \in A) \leq \inf_{\mu \in \overline{A}} I(\mu). \tag{3.17}$$

We already presented an instance of large deviations principle in Section 2.2.2 with Sanov's theorem when $x_1, \ldots, x_n$ are drawn i.i.d from π (2.44). To wit again, Sanov's theorem states that in this case, μ_n satisfies a large deviation principles at speed n and with good rate function given by the relative entropy between μ and π.

We further showed in Section 2.2.2 that Sanov's theorem can be used to give bounds on some standard worst-case integration errors, e.g. the total variation distance between μ_n and π. More generally, assume that π is a minimizer of the rate function I, and that we are able to lower bound I

$$I(\mu) > C \mathrm{d}(\mu, \pi)^2 \tag{3.18}$$

where d is a worst-case integration error

$$\mathrm{d}(\mu, \pi) = \mathrm{E}_{\mathcal{F}}(\mu) = \sup_{f \in \mathcal{F}} |\langle f, \mu \rangle - \langle f, \pi \rangle| , \tag{3.19}$$

for some functional space $\mathcal{F}$, see Section 2.1.1. Then, for any $r > 0$, there exists $n_0(r) \in \mathbb{N}^*$ such that for any $n \geq n_0(r)$,

$$\mathbb{P}\left(\mathrm{d}(\mu_n, \pi) > r\right) \leq \exp\left(-C a_n r^2\right). \tag{3.20}$$

If a_n grows faster than n, this means that, asymptotically, the tail decay of $\mathrm{d}(\mu_n, \pi)$ is faster than the one we get with i.i.d points (2.49). If we are interested in simultaneously approximating integrals $\int f \, \mathrm{d}\pi$ with $f \in \mathcal{F}$, at fixed precision level $r > 0$, this means that fewer points are required to achieve the same guarantee as with points $x_1, \ldots, x_n$ drawn i.i.d from π. We will say that there is tighter coverage at fixed precision level $r > 0$.

From a practical perspective, we usually prefer non-asymptotic results, and we would thus like n_0 to be independent of r. This calls for concentration inequalities of the type

$$\mathbb{P}\left(\mathrm{d}(\mu_n, \pi) > r\right) \leq \exp\left(-c\, a_n r^2\right), \tag{3.21}$$

for some constant $c > 0$ and for any $n \geq n_0$ with n_0 fixed. Such inequalities usually come with restrictions on the precision level r due to the existence of lower bounds on

$\mathrm{d}(\mu_n, \pi)$, see Section 2.1.2 and (2.52). Again, an inequality like (3.21) with a_n growing faster than n means tighter coverage for numerical integration at fixed precision level r than with points drawn independently, which is again a trace of negative dependence in the point process.

We are also often interested in studying directly the fluctuations of the system, meaning the variance of linear statistics (3.15)

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right]. \tag{3.22}$$

This is especially the case if the want to study the mean squared error (2.27) of numerical integration.

Variance of linear statistics

For smooth test functions, the control of linear statistics is often the first step towards central limit theorems of the form

$$\sqrt{a_n}\left(\langle f, \mu_n\rangle - \int f \,\mathrm{d}\pi\right) \to \mathcal{N}\left(0, \sigma_f^2\right), \tag{3.23}$$

with $\sigma_f^2 > 0$, for f belonging to a certain class of functions $\mathcal{F}$. In the i.i.d case, recall again that the standard CLT (Theorem 2.2.2) yields (3.23) with $a_n = n$ and $\mathcal{F} = L^2(\pi)$. Gaussian fluctuations like (3.23) with a_n growing faster than n implies smaller asymptotic variance than in the i.i.d case. Intuitively, this happens again when the points $x_1, \ldots, x_n$ are well-spread in the integration domain: in that case the points are closer to be deterministic than in the i.i.d case and there is thus less variance in the asymptotically unbiased estimate (3.15).

For a general point process X, another way to quantify fluctuations is to study the variance of the number of points falling in a ball of radius R

$$\mathrm{Var}\left[N_{B(0,R)}(X)\right], \tag{3.24}$$

as R grows to infinity. This is the key study object for hyperuniform systems. Hyperuniformity (Torquato, 2018) has emerged as a trendy property in a lot of fields, from complex material science to statistical physics.

Definition 3.1.7 We say that X is a hyperuniform point process whenever

$$\mathrm{Var}\left[\mathrm{N}_{B(0,R)}(X)\right] = o\left(|B(0,R)|\right), \tag{3.25}$$

as $R \to +\infty$, meaning that the variance of the number of points falling in the ball of radius R increases slower than the volume of the ball as the radius goes to infinity.

Again,(3.25) is to be seen as a comparison to the independent case. If all the points X were independent of one another, one can show that $\mathrm{Var}\left[\mathrm{N}_{B(0,R)}(Y)\right]$ scales as $|B(0,R)|$.

Thus, hyperuniformity accounts for cancellations in large scale fluctuations compared to the independent case, i.e the full disorder case. One can further show (Ghosh and Lebowitz, 2016; Beck, 1987) that the minimal possible growth is given by

$$\mathrm{Var}\left[\mathrm{N}_{B(0,R)}(Y)\right] \sim |\partial B(0,R)|. \tag{3.26}$$

Hyperuniformity has recently been investigated for the purpose of Monte Carlo integration (Hawat, 2023). In fact, assume for instance that

$$\mathrm{Var}\left[\mathrm{N}_{B(0,R)}(Y)\right] = \mathcal{O}(R^{d-1}), \tag{3.27}$$

and that $\rho_1 = 1$ for ease of presentation. Consider a compact set $A \subset W$ for a finite observation window W. Moreover, consider the dezoomed point process $X = n^{-1/d}Y$ which has intensity $n/|W|$, meaning that X has $n/|W|$ points in average. Then

$$\mathrm{Var}\left[\frac{|W|}{n}\sum_{x\in X\cap W}\mathbb{1}_A(x)\right] = \mathcal{O}(n^{-1-1/d}), \tag{3.28}$$

see (Hawat, 2023) for instance. Hence, the Monte Carlo estimator based on the points of X has a smaller variance than the crude Monte Carlo estimator based on i.i.d points with variance n^{-1} (2.56). Such results are however only available for indicator functions, but this is already motivating if the goal is to build credibility intervals for Bayesian settings (Chapter 1). In the next section, we present a few standard point processes, with an emphasis on negatively dependent ones.

3.2 Some examples

3.2.1 Binomial and Poisson point processes

Consider $n \in \mathbb{N}^*$ and μ a measure on $\mathbb{R}^d$. One of the most simple examples of points processes is the Binomial point process given by $X = \sum_{i=1}^n \delta_{x_i}$ for $x_1, \dots, x_n$ drawn i.i.d from μ. Numerical integration with the Binomial point process just amounts to i.i.d sampling. The correlation functions are constant and trivial in that case, with reference measure μ: $\rho_1 = n$ and

$$\rho_k = \binom{n}{k} \tag{3.29}$$

for $k \leq n$, meaning that there is no dependence between the points. For i.i.d points, we already saw Sanov's theorem (2.44) and the standard CLT (2.57).

Informally, Poisson point processes are analogous to Binomial point processes with a random number of points.

Definition 3.2.1 — Poisson point process. A Poisson point process with reference measure μ is a point process X satisfying the following properties.

- For each disjoint bounded regions $A, B \subset \mathbb{R}^d$, the counting random variables $N_A(X)$ and $N_B(X)$ are independent,
- For each bounded region A, $N_A(X)$ follows a Poisson distribution with parameter $\mu(A)$.

Moreover, if W is finite window i.e. $\mu(W) < +\infty$, we can give the following construction. Consider $(x_i)_{i\geq 1}$ i.i.d from μ and let κ be distributed according to a Poisson distribution of parameter $\mu(W)$. Then $X = \sum_{i=1}^{\kappa} \delta_{x_i}$ is a Poisson point process on W with reference measure μ. We can thus see the Binomial point process as a Poisson point process conditioned on having n points.

Again, the correlation functions for the Poisson point process are constant and there is no dependence between the points: we say that the Poisson point process corresponds to the *maximal disorder* case. Moreover, we know (Ghosh et al., 2016) that the Poisson point process is not hyperuniform, in fact it is the baseline of non hyperuniform point process:

$$\mathrm{Var}\left[N_X(B(0,R))\right] = |B(0,R)|, \tag{3.30}$$

This means that a hyperuniform point process is a system of points for which the variance of the number of points in a ball increases slower than for the Poisson point process (i.e. the independent case).

In Figure 3.1, we show a realization of a Poisson point process in a bounded window W, with μ given by the Lebesgue measure. We see that the Poisson point process leaves holes in the observation window, as a consequence of the independence between the points.

3.2.2 (Perturbed) Lattices

Intuitively, the Poisson point process corresponds to the full disorder case: there is no correlation between the points. On the other extreme of the spectrum, lattice configurations such as $X = \{z + U : z \in \mathbb{Z}^d\}$ with U uniform correspond to maximal order for a random system. In fact, such lattices configurations are known (Ghosh et al., 2016) to achieve the lower bound (3.26) for hyperuniformity, meaning that

$$\mathrm{Var}\left[N_X(B(0,R))\right] = |\partial B(0,R)|.$$

A more general class of interesting random lattice-like configurations are the so-called perturbed lattice that write $X = \{z + U_z : z \in \mathbb{Z}^d\}$, where $(U_z)_{z\in\mathbb{Z}^d}$ is a random field of identically distributed but possibly dependent random variables. It has been shown (Dereudre et al., 2025) that this is still a hyperuniform point process in dimensions one

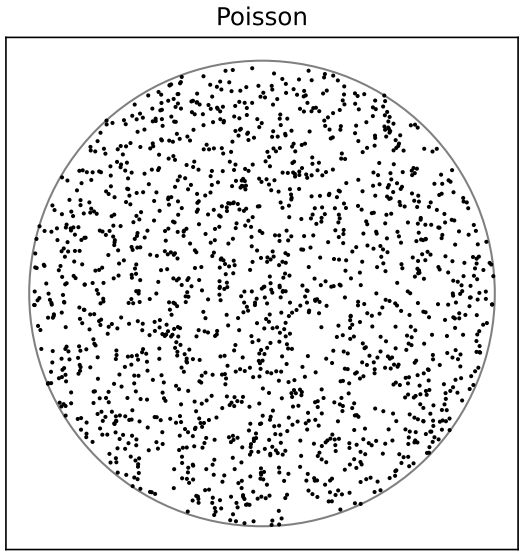


Fig. 3.1: Realization of a Poisson point process in a bounded window.

and two if the d-th moment of the law of the perturbation is finite, but that this can fail to be hyperuniform in dimensions larger than three. Actually, in dimension 2, it has been shown (Lachièze-Rey and Yogeshwaran, 2025) that, under mild integrability conditions, hyperuniformity implies being a perturbed lattice. Considering random perturbations of well-ordered point patterns should make us think of RQMC methods in numerical integration (Section 2.4.2).

We now turn to two examples of point processes that are intermediate between the full order and the full disorder case.

3.2.3 Eigenvalues of random matrices

A broad class of interesting point process — usually in dimensions one and two — is given by point patterns obtained as random zeros. It is useful to have in mind that the original Gaussian quadrature was designed using the zeros of an orthogonal basis of functions in L^2 (Section 2.2.1), so that numerical quadrature based on random zeros can be though of as a randomized Gaussian quadrature. We give some examples coming from random matrix theory.

Consider the $n \times n$ matrix $G = (G_{i,j})$ with random entries, and the point process $\sum_{\lambda \in Sp(G)} \delta_\lambda$ where $Sp(G)$ denotes the set of eigenvalues of G, i.e. the zeros of the characteristic polynomial of G. For instance, let G be symmetric or Hermitian i.e. $G_{i,j} = \overline{G_{j,i}}$ and

$$\begin{cases} G_{i,j} = \frac{Z_{i,j}}{\sqrt{n}} \text{ if } i < j, \\ G_{i,i} = \frac{Y_i}{\sqrt{n}}, \end{cases} \tag{3.31}$$

where $(Z_{i,j})_{i<j}$ and $(Y_i)_i$ are two independent families of i.i.d random variables. Such random matrices are called Wigner matrices (Anderson et al., 2010). A particularly interesting case is when the variables $Z_{i,j}$ and Y_i are Gaussian. In particular, when $G_{i,j} \sim \mathcal{N}(0,1)$ for all $i < j$ and $Y_i \sim \mathcal{N}(0,2)$, (3.31) is called the Gaussian Orthogonal Ensemble (GOE). Similarly, when the univariate Gaussian distribution are replaced by complex Gaussian distributions, (3.31) is called the Gaussian Unitary Ensemble (GUE). In these cases, we can show (Anderson et al., 2010) that the n eigenvalues of G are almost-surely distinct real numbers $\lambda_1, \dots, \lambda_n$. Moreover, the law of $\lambda_1, \dots, \lambda_n$ is given by the probability distribution on $\mathbb{R}^n$ defined as

$$\mathbb{P}_{n,\beta}(\lambda_1, \dots, \lambda_n) \propto \prod_{i \neq j} |\lambda_i - \lambda_j|^\beta \exp\left(-\beta n \sum_{i=1}^n \frac{|\lambda_i|^2}{4}\right) \mathrm{d}\lambda_1 \dots \mathrm{d}\lambda_n, \tag{3.32}$$

with $\beta = 1$ for the GOE and $\beta = 2$ for the GUE (Anderson et al., 2010, Theorem 2.5.2), and where $\propto$ means that (3.32) is defined up to a normalization constant that makes $\mathbb{P}_{n,\beta}$ a probability measure. Actually, one can study this probability distribution for any $\beta > 0$, and this is called a Gaussian β-ensemble or one dimensional log-gas. This joint distribution balances two effects: a repulsion between the eigenvalues, and a Gaussian confinement term preventing the eigenvalues to go far from a bounded region of the space.

Similarly, when we drop the symmetry constraint on the random matrix, a classical example is the so-called Ginibre ensemble where $G_{i,j}$ is a complex Gaussian distribution with variance $\frac{1}{\sqrt{n}}$ for all $1 \leq i, j \leq n$. Then, the n eigenvalues of G are almost-surely distinct complex numbers, and their joint law is the probability distribution on $\mathbb{R}^{2n}$ defined as

$$\mathbb{P}_{n,\beta}(\lambda_1, \dots, \lambda_n) \propto \prod_{i \neq j} |\lambda_i - \lambda_j|^\beta \exp\left(-\beta n \sum_{i=1}^n \frac{|\lambda_i|^2}{2}\right) \mathrm{d}\lambda_1 \dots \mathrm{d}\lambda_n, \tag{3.33}$$

for $\beta = 2$, see (Bordenave and Chafaï, 2012). Again, this is also well-defined for general $\beta > 0$, and this is also called the two-dimensional log gas, or two-dimensional one component plasma.

The limiting distribution and fluctuations for the empirical measure (3.1) of such point processes are now well understood. In particular, the fast rates of convergence that we present in the following can be seen as a trace of the repulsion between the eigenvalues of the system. Wigner's theorem (Wigner, 1958; Anderson et al., 2010) states that in the case of Gaussian β-ensembles, μ_n converges in law to the semicircle distribution $\mathrm{d}\mu_{sc}(x) = 1_{|x|<2} \frac{1}{2\pi} \sqrt{4 - x^2}\, \mathrm{d}x$ almost surely, while in the case of the Ginibre ensemble the circular law states that the empirical measure converges in law to the uniform distribution on the unit disk almost surely, see (Bordenave et al., 2012) for a long history of this result. One can further show that the empirical measure $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i}$ satisfies a large deviation principle with speed $n^2 \gg n$ that accounts for the repulsion between

the points: this was proved for one-dimensional log gases (3.32) in (Ben Arous and Guionnet, 1997) and for two-dimensional log gases for $\beta = 2$ in (Ben Arous and Zeitouni, 1998) and general $\beta > 0$ for instance in (Chafaï et al., 2014).

Furthermore, central limit theorems at the fast speed n of the form

$$n\left(\langle f\,,\mu_n\rangle - \int f\,\mathrm{d}\mu_{eq}\right) \to \mathcal{N}(m_{f,\beta}, \sigma^2_{f,\beta}), \tag{3.34}$$

have been obtained for β-ensembles in (Johansson, 1998; Borot and Guionnet, 2013; Bekerman et al., 2018), in which case μ_{eq} is the semicircle law. Similar results have been obtained in the two-dimensional case for the Ginibre ensemble (Rider and Virág, 2007) with $\beta = 2$, and for general $\beta > 0$ in (Leblé and Serfaty, 2018). In those cases, μ_{eq} is the uniform distribution on the unit disk. Results like (3.34) often require smoothness assumptions on f, for instance C^1 with growth conditions in (Rider et al., 2007). See Section 3.3.3 for a more detailed statement in the case of the two-dimensional log gas.

Eigenvalues of random matrices have been the topic of a lot of studies, and one can for instance study similar point processes obtained with more general entry laws or other matrix models (e.g. covariance matrices, different symmetries) (Anderson et al., 2010; Tao, 2012). Actually, a lot of effort has been put to investigate the fact that many of those models exhibit similar global or local fluctuations regardless of the entry law, this is known as *universality*.

Still, these models are either one or two-dimensional systems of points whereas we want to target higher dimensions for the purpose of numerical integration. Determinantal point processes give a larger class of general point processes encompassing repulsion between the points.

3.2.4 Determinantal point processes

Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ be a kernel function. The determinantal point process (DPP, Macchi, 1975; Soshnikov, 2000) of kernel K with reference measure μ is the point process whose correlation functions are given by

$$\rho_k(x_1, \dots, x_k) = \det\left((K(x_i, x_j))_{1\leq i,j\leq k}\right) \tag{3.35}$$

for any $k \geq 1$. In particular, we know that this exists when K defines an adjoint and trace class operator $\mathcal{K}$ whose eigenvalues belong to $[0, 1]$, see (Soshnikov, 2000). Assuming that K is Hermitian, meaning that $K(x, y) = \overline{K(y, x)}$, we then see that

$$\rho_2(x, y) = K(x, x)K(y, y) - |K(x, y)|^2 < K(x, x)K(y, y) = \rho_1(x)\rho_1(y), \tag{3.36}$$

meaning that determinantal point processes are repulsive point processes by nature, and the repulsion is encoded in the choice of the kernel K.

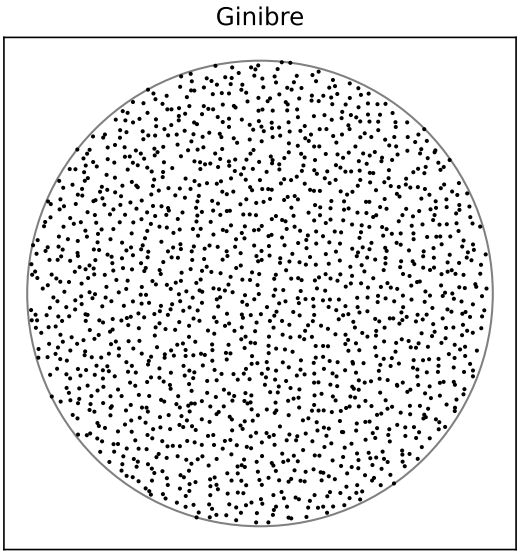


Fig. 3.2: A realization of the Ginibre point process in a bounded window.

Determinantal point process appear in a variety of situations in probability and mathematical physics, see for instance (Hough et al., 2006) for a survey. For instance, the Ginibre ensemble (3.33) is a DPP with kernel

$$K_{\mathrm{G}_n}(x, y) = \sum_{k=0}^{n} \frac{(\overline{x}y)^k}{k!} \tag{3.37}$$

and Gaussian reference measure μ (Ghosh et al., 2016). By letting $n \to +\infty$, we get the infinite Ginibre ensemble defined as the DPP with Gaussian reference measure and kernel

$$K_{\mathrm{G}_\infty}(x, y) = e^{\overline{x}y}. \tag{3.38}$$

The infinite Ginibre ensemble is known to be hyperuniform (Ghosh et al., 2016). We show a realization of this point process in Figure 3.2 in a bounded window. Compared to the Poisson point process of Figure 3.1, we see that the points are better spread in the window, as a consequence of the repulsion in the point process.

In Section 2.4.2, we defined OPEs which are determinantal point process with reference measure μ and kernel

$$K_n(x, y) = \sum_{k=0}^{n-1} \Psi_k(x)\Psi_k(y), \tag{3.39}$$

where (Ψ_k) are the multivariate orthogonal polynomials associated to the reference measure μ. We saw in Section 2.4.2 that we can get a fast CLT for linear statistics

based on OPEs. Moreover, kernel quadrature based on DPPs achieve optimal worst-case integration errors in RKHSs (Section 2.6.3), yet with few hopes for a CLT. Sampling from those point processes can usually be done in cubic time (Hough et al., 2006), yet with an exponential number of rejection steps in terms of the dimension d at the moment (Gautier et al., 2019a). This motivates the investigation of other repulsive point processes suited for numerical integration. In this thesis, we study Gibbs measures that generalize (3.33) to any dimension and with tunable interaction kernel K.

3.3 Gibbs measures

3.3.1 Motivations and definitions

Consider n points $x_1, \dots x_n \in \mathbb{R}^d$. We will sometimes use vocabulary inspired from statistical physics, for instance $x_1, \dots, x_n$ will be called *particles*. Assuming that the particles interact pairwise through an interaction kernel K, and are confined in a domain through a confinement potential V, the *energy* of the point configuration $x_1, \dots, x_n$ reads

$$H_n(x_1, \dots, x_n) = \frac{1}{2n^2} \sum_{i \neq j} K(x_i, x_j) + \frac{1}{n} \sum_{i=1}^{n} V(x_i). \tag{3.40}$$

The normalizations are chosen so that the two terms are of the same order, meaning that there is balance between interaction and confinement. Some usual choices of K such as the Gaussian kernel, inverse multiquadratic kernels and Matérn kernels have been presented in Section 2.1.1. Note in particular that all those choices give bounded interaction kernels meaning that $K(x, x) < \infty$ for all $x \in \mathbb{R}^d$.

On the other hand, physicists rather study models with interaction kernels K that involve singularities, such as the logarithmic interaction

$$K_{\log}(x, y) = -\log|x - y|, \tag{3.41}$$

or the Riesz interaction

$$K_s(x, y) = \frac{1}{|x - y|^s}, \tag{3.42}$$

for $s > 0$. In particular, the Coulomb interaction defined as

$$g(x, y) = \begin{cases} -\log|x - y| \text{ if } d = 1, 2 \\ \frac{1}{|x-y|^{d-2}} \text{ if } d \geq 3 \end{cases} \tag{3.43}$$

is the electrostatic interaction appearing in classical models for electromagnetism or condensed matter physics (Serfaty, 2018).

In all these examples, since K is non-negative, the interaction part in the energy (3.40) is smaller when the points $x_1, \dots, x_n$ don't get close to one another, with a strength

given by K. Much effort, either algorithmic or theoretical, has been put up in the statistical physics and approximation theory communities to try to find deterministic point configurations X that are optimal in the sense of achieving low interaction energy $\sum_{x,y\in X} K(x,y)$ for a given kernel K. See for instance (Serfaty, 2024) for a survey, we will go back to those questions in Chapter 6.

Moreover, for a given target distribution π, we will see that the potential V can be tuned so that (3.40) is exactly the MMD (2.97) between the empirical measure (3.1) associated to $x_1, \dots, x_n$ and π, which we want to minimize in numerical integration.

To find deterministic points $x_1, \dots, x_n$ achieving the minimum, we already presented in Section 2.6 gradient descent updates of the form

$$X_n^{(t+1)} = X_n^{(t)} - \gamma \nabla H_n(X_n^t), \tag{3.44}$$

with $X_n^{(t)} = (x_1^{(t)}, \dots, x_n^{(t)})$. Yet, there is little theoretical support for faster than i.i.d rates of convergence, see Section 2.6. In statistical physics, a canonical way to draw points with low energy (3.40) is to consider the following Boltzmann–Gibbs distribution.

Definition 3.3.1 Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be an interaction kernel and $V : \mathbb{R}^d \to \mathbb{R}$ a confinement potential. Let $\beta_n \geq cn$. The Gibbs measure is defined as the joint probability distribution $\mathbb{P}^V_{n,\beta_n}$ on $\mathbb{R}^{dn}$ with density

$$\mathrm{d}\mathbb{P}^V_{n,\beta_n}(x_1, \dots, x_n) = \frac{1}{Z^V_{n,\beta_n}} \exp\left(-\beta_n H_n(x_1, \dots, x_n)\right) \mathrm{d}x_1 \dots \mathrm{d}x_n, \tag{3.45}$$

where Z^V_{n,β_n} is a normalization constant — also called partition function in the statistical physics literature — making $\mathbb{P}^V_{n,\beta_n}$ a probability distribution.

$\mathbb{P}^V_{n,\beta_n}$ favors point configurations $(x_1, \dots, x_n)$ with low energy $H_n(x_1, \dots, x_n)$, with a strength induced by the inverse temperature β_n. At high temperature, meaning low β_n, H_n has a small impact on the probability $\mathbb{P}^V_{n,\beta_n}$ and there should not be strong dependence between the points $x_1, \dots, x_n$. At low temperature, meaning large β_n, $\mathbb{P}^V_{n,\beta_n}$ is very spiked around the minimizers of the energy H_n hence configurations with large probability are configurations with very low energy which implies strong dependence between the points.

However, as opposed to the case of determinantal point processes (Section 3.2.4), correlations functions ρ_k are not explicit for Gibbs measures in general, we refer the reader to (Lewin, 2022) for a survey on estimates for such quantities. Yet, this joint distribution is natural for multiples reasons.

First, consider a noisy version of the updates (3.44)

$$X_n^{(t+1)} = X_n^{(t)} - \gamma \nabla \beta_n H_n(X_n(t)) + \sqrt{2\gamma}\zeta, \tag{3.46}$$

with $\zeta \sim \mathcal{N}(0, I_d)$ and where $\beta_n > 0$ is the inverse temperature. Combining the proposal (3.46) with a Metropolis acceptance step to ensure good convergence of the algorithm yields the MALA updates (2.66) which converge to $\mathbb{P}^V_{n,\beta_n}$ as $t \to +\infty$. Relatedly, from the statistical physics perspective, $\mathbb{P}^V_{n,\beta_n}$ is the unique minimizer of "Energy($\mathbb{P}$) − Temperature × Entropy($\mathbb{P}$)" which writes

$$\int H_n(x_1, \dots, x_n) \, \mathrm{d}\mathbb{P}(x_1, \dots, x_n) \, \mathrm{d}x_1 \dots \mathrm{d}x_n \tag{3.47}$$

$$+ \frac{1}{\beta_n} \int \mathbb{P}(x_1, \dots, x_n) \log \mathbb{P}(x_1, \dots, x_n) \, \mathrm{d}x_1 \dots \mathrm{d}x_n, \tag{3.48}$$

over probability distributions $\mathbb{P}$ on $\left(\mathbb{R}^d\right)^n$ (Lewin, 2022).

Finally, we have actually already encountered examples of such Gibbs measures in the context of eigenvalues of random matrices (Section 3.2.3). Indeed, consider the one and two-dimensional cases, and the logarithmic interaction (3.41) along with $\beta_n = \beta n^2$ for $\beta > 0$. Then $\mathbb{P}^V_{n,\beta_n}$ coincides with the laws (3.32) and (3.33) of eigenvalues of β-ensembles and Ginibre ensemble (for $\beta = 2$) with quadratic choices of the potential V. This means that we can hope to obtain similar fast LDPs and central limit theorems as the ones presented in Section 3.2.3, and use them for numerical integration.

In the next two sections, we quickly give the main known results of interest in terms of large deviations, concentration inequalities and central limit theorems.

3.3.2 Large deviations and concentration

In this section, we consider a general dimension $d \geq 1$. We first introduce a continuous version of the energy (3.40) that will appear in the upcoming results. Moreover, we show that for a good choice of V, it can be connected to the MMD (2.97). We then state known large deviation and concentration results for points drawn from the Gibbs measure (3.45).

Definition 3.3.2 — Energies. Consider a general interaction kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and a general confinement potential $V : \mathbb{R}^d \to \mathbb{R}$. Whenever they are well-defined, we introduce the following quantities, for signed Borel measures μ, ν on $\mathbb{R}^d$. The *interaction potential*, or *kernel embedding*, of μ, is defined as

$$U^\mu_K(z) = \int K(z, y) \, \mathrm{d}\mu(y), \tag{3.49}$$

where $z \in \mathbb{R}^d$. The *interaction energy* between μ and ν is defined as

$$I_K(\mu, \nu) = \iint K(x, y) \, \mathrm{d}\mu(x) \, \mathrm{d}\nu(y). \tag{3.50}$$

When $\mu = \nu$, we simply write $I_K(\mu) = I_K(\mu, \mu)$. Finally, we let

$$I_K^V(\mu) = \frac{1}{2} \iint \{K(x,y) + V(x) + V(y)\} \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y). \tag{3.51}$$

Up to diagonal terms, $H_n(x_1, \dots, x_n)$ would just be the energy I_K^V of the empirical measure $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Moreover, for a target distribution π, setting $V(x) = -U_K^\pi(z)$ would yield

$$I_K^V(\mu) = \frac{1}{2} \left(\iint K(x,y) \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y) - 2 \int U_K^\pi(x) \, \mathrm{d}\mu(x) \right) \tag{3.52}$$

$$= \frac{1}{2} \left(\mathrm{MMD}_K(\mu, \pi) - \iint K(x,y) \, \mathrm{d}\pi(x) \, \mathrm{d}\pi(y) \right). \tag{3.53}$$

Thus, with this choice of V, the Gibbs measure (3.45) favors configurations $(x_1, \dots, x_n)$ that yield low MMD with respect to the target distribution π. We will however need to add a second term in V to keep the points confined in a compact domain, so that the energy I_K^V admits π as a unique minimizer and the Gibbs measure (3.45) is well-defined.

For the ease of presentation, we now focus in the rest of the section on the case of singular Riesz interactions (3.42). We will go back to the case of general but bounded interactions in Chapter 4.

Theorem 3.3.3 — Chafaï et al. (2014). Let $d \geq 1$, and consider the Riesz interaction K_s given by (3.42) for $0 < s < d$. Assume that V is continuous and that $V(x) \to +\infty$ when $|x| \to \infty$. Then the energy $I_{K_s}^V$ is a good rate function (Definition 3.1.6).

Assume furthermore that there exists $c > 0$ such that

$$\int \exp\left(-cV(x)\right) \, \mathrm{d}x < +\infty, \tag{3.54}$$

and consider the inverse temperature $\beta_n \geq 2cn$. Then the Gibbs measure (3.45) is well-defined, meaning that $0 < Z_{n,\beta_n}^V < +\infty$.

Finally, assuming that $\beta_n / n \to +\infty$, the sequence of empirical measures $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ with $x_1, \dots, x_n \sim \mathbb{P}_{n,\beta_n}^V$ satisfies a LDP (Definition 3.1.6) with speed β_n and rate function

$$I(\mu) = I_{K_s}^V(\mu) - \inf_{\nu \in \mathcal{P}(\mathbb{R}^d)} I_{K_s}^V(\nu).$$

In the rest of the manuscript, $\beta_n / n \to +\infty$ will often be denoted by $\beta_n \gg n$. Similar treatment for logarithmic interactions in dimensions 1 and 2 can be found in (Ben Arous et al., 1997; Ben Arous et al., 1998). Moreover, the result of (Chafaï et al., 2014) actually also applies to general but bounded interactions. A consequence of Theorem 3.3.3 is that, at fixed V, μ_n converges to μ_V almost-surely. The following corollary then motivates

using Gibbs measures for numerical integration.

Corollary 3.3.4 — Chafaï et al. (2014). Let $d \geq 1$, consider the Riesz interaction K_s given by (3.42) for $0 < s < d$ and $\beta_n \gg n$. Let π be any target distribution with compact support S_π and L^p density for some $p > d/(d-s)$. Let $R > 0$ be such that $S_\pi \subset B(0, R)$ and define the confinement potential

$$V^\pi(x) = -U^\pi_{K_s}(x) + \Phi(x), \tag{3.55}$$

where

$$\Phi(x) = \begin{cases} 0 \text{ if } |x| < R \\ |x|^2 - R^2 \text{ otherwise} \end{cases} \tag{3.56}$$

Then V^π satisfies the assumptions of Theorem 3.3.3 and $I^{V^\pi}_{K_s}$ has a unique minimizer on probability distributions given by π.

In the Coulomb case, such a choice of V has a long history in the statistical physics literature when π is the uniform distribution. In that case, the Gibbs measure (3.45) is often called *jellium*, see for instance (Chafaï et al., 2020; Butez et al., 2025, 2022). Again, we will present a similar statement for general but bounded interactions in Chapter 4. The following concentration inequalities provide non-asymptotic versions of this LDP, along with a connection to standard worst-case integration errors or distance between probability measures, depending on the choice of s.

Theorem 3.3.5 — Chafaï et al. (2018). Assume that $d \geq 3$, $\beta_n = \beta n^2$ with $\beta > 0$ and consider the Riesz kernel K_s with $s = d - 2$, i.e. the Coulomb interaction (3.43). We work under the same assumptions on V as in Theorem 3.3.3, and we further assume that it has uniformly bounded second order derivatives.

Then, there exists positive constants c_0 and c_1 such that for any $n \geq 2$ and any $r > c_0 n^{-1/d}$,

$$\mathbb{P}^V_{n,\beta_n}\left(\mathrm{d}_{\mathrm{BL}}(\mu_n, \mu_V) > r\right) \leq \exp\left(-c_1 \beta n^2 r^2\right), \tag{3.57}$$

where μ_V is the unique minimizer of $I^V_{K_s}$ and d_{BL} is the bounded-Lipschitz distance (2.11).

In dimension 2, Chafaï et al. (2018) give a similar result but for the logarithmic interaction (3.41), which is the Coulomb interaction in dimension 2. Moreover, (3.57) can be generalized to more general temperature $\beta_n \gg n$ instead of only $\beta_n = \beta n^2$, see Chapter 5. The case of general $0 < s < d$ is given by the following result.

Theorem 3.3.6 — García-Zelada and Padilla-Garza (2023). Let $d \geq 1$, $\beta_n \gg n$ and consider the Riesz kernel K_s with $0 < s < d$. We work under similar assumptions that in Theorem 3.3.5, see García-Zelada et al. (2023) for details. Then, for any $\alpha > (d-s)/2$, there exists positive constants c_0 and c_1 such that for any $n \geq 2$ and any $r > c_0 n^{-\alpha/d}$,

$$\mathbb{P}^V_{n,\beta_n}\left(\mathrm{d}_\alpha(\mu, \mu_V)\right) \leq \exp\left(-c_1 \beta_n r^2\right), \tag{3.58}$$

where μ_V is again the unique minimizer of $I^V_{K_s}$ and

$$\mathrm{d}_\alpha(\mu, \nu) = \sup_{f \in \mathcal{F}_\alpha} \left| \int f \, \mathrm{d}(\mu - \nu) \right|$$

with

$$\mathcal{F}_\alpha = \{\|f\|_{\mathcal{C}^{0,\alpha}} \leq 1, \, \|f\|_\infty \leq 1\},$$

and $\|f\|_{\mathcal{C}^{0,\alpha}}$ denotes the α-Hölder norm (2.9).

In Chapter 4, we will obtain similar concentration inequalities for general but bounded interactions K, this time for the MMD with kernel K between μ_n and μ_V. We will discuss the results for numerical integration. In the following section, we quickly describe what is known in terms of CLTs for Gibbs measures of the form (3.45).

3.3.3 Gaussian fluctuations in low dimensions

Central limit theorems have been obtained mainly in dimensions one (Bekerman et al., 2018) and two (Leblé et al., 2018), for the logarithmic interaction (3.41), under smoothness assumptions on V and on the test function f. The statements and proof techniques being quite involved, we will be less general, and we will just give an example to illustrate what can be hoped. To wit, Leblé et al. (2018) give the following result in the two-dimensional case.

Theorem 3.3.7 — Leblé et al. (2018). Let $d = 2$, and $\beta_n = \beta n^2$ with $\beta > 0$. Consider the logarithmic interaction $K_{\log}$ (3.41). Let Σ denote the support of the unique minimizer μ_V of $I^V_{K_{\log}}$. Assume that $f \in \mathcal{C}^{2,1}(\Sigma)$. If $V \in \mathcal{C}^{3,1}(\Sigma)$, then there exists m_f and $v_f > 0$ such that

$$\sqrt{\beta} n \left(\frac{1}{n} \sum_{i=1}^n f(x_i) - \int f \, \mathrm{d}\mu_V \right) \underset{n \longrightarrow +\infty}{\to} \mathcal{N}(m_f, v_f), \tag{3.59}$$

in distribution, $\mathbb{P}$-a.s.

In (Leblé et al., 2018), m_f and v_f can be described in terms of Σ and f.

Again, this faster rate of convergence is a trace of the repulsion between the points. CLTs for Coulomb or Riesz interactions in higher dimensions are still an open question. Conditional results have been obtained, upon proving beforehand a fine asymptotic expansion for the partition function Z^V_{n,β_n} in terms of n (Serfaty, 2023). There are to our knowledge yet no results that give CLTs for general interactions such as bounded kernels used in machine learning.

3.4 Conclusion

Repulsive point processes give a natural way to draw points at random and leverage probabilistic tools to get faster-than-i.i.d convergence rates for numerical integration as a consequence of the repulsion. In particular, Gibbs measures provide a probabilistic counterpart to kernel method aiming at minimizing the MMD with respect to the target distribution (Section 2.6).

In this thesis, our goal is to leverage Gibbs measure to design numerical integration methods with improved guarantees compared to the i.i.d case. In Chapter 4, we provide a fast concentration result for Gibbs measure general bounded interaction kernels K and discuss the consequences for numerical integration. Yet, this relies on a choice of the confinement V similar to (3.55) which involves an integral with respect to our target distribution π that needs to be approximated as well.

In Chapter 5, we thus deal with the issue of the approximation of the potential U^π_K, which is present as well in all MMD minimization algorithms (Section 2.6). In Chapter 6, we will discuss promising approaches to obtain fast central limit theorems for linear statistics of Gibbs measures with bounded interaction kernels K in any dimension.

Monte Carlo with Gibbs measures

"I guess I'm learning,
I must be warmer now."

The show must go on, Queen

4

4.1 Introduction 85

4.2 Related work 85

4.2.1 Concentration for independent sampling and MCMC . . . 85

4.2.2 Concentration for kernel methods and Coulomb gases . . . 86

4.3 A concentration inequality for Gibbs measures 87

4.3.1 Definitions . . . 87

4.3.2 Main results . . . 89

4.3.3 Application to Monte Carlo integration . . . 90

4.4 Numerical experiments 95

4.4.1 Sampling from the Gibbs measure . . . 95

4.4.2 Coverage of Monte Carlo integration . . . 102

4.4.3 Worst-case integration errors . . . 104

4.4.4 Variance of linear statistics . . . 106

4.5 Proofs 107

4.5.1 Proof of Proposition 4.3.5 . . . 107

4.5.2 Proof of Proposition 4.3.10 . . . 108

4.5.3 Proof of Theorem 4.3.7 . . . 109

4.5.4 Proof of Proposition 4.3.13 . . . 112

4.6 Conclusion 112

4.1 Introduction

The content of this chapter is based on the following paper, which is under revision for the *SIAM journal on Mathematics of Data Science*.

- Rouault, Martin, Bardenet, Rémi, and Maïda, Mylène (Feb. 18, 2024). *Monte Carlo with kernel-based Gibbs measures: Guarantees for probabilistic herding*. arXiv: 2402.11736.

We investigate the statistical properties of points drawn from the Gibbs measure defined in (4.1) by

$$\mathbb{P}^V_{n,\beta_n}(x_1,\dots,x_n) = \frac{1}{Z^V_{n,\beta_n}} \exp\left(-\beta_n H_n(x_1,\dots,x_n)\right) \mathrm{d}x_1 \dots \mathrm{d}x_n, \tag{4.1}$$

with

$$H_n(x_1,\dots,x_n) = \frac{1}{2n^2}\sum_{i\neq j} K(x_i,x_j) + \frac{1}{n}\sum_{i=1}^n V(x_i). \tag{4.2}$$

In particular, we show in Theorem 4.3.7 that for general but bounded choices of K, the empirical measure of points jointly drawn from (4.1) satisfies a fast concentration inequality at speed β_n. Moreover, we show in Proposition 4.3.10 that the external confinement V can be tuned so that this concentration inequality gives confidence regions around target integrals that are tighter than for crude MCMC, see Corollary 4.3.11.

In Section 4.2 we quickly recall existing results that we discussed in Chapters 2 and 3. In Section 4.3, we rigorously define our objects of study and state our main concentration theorem for general Gibbs measures along with consequences for Monte Carlo integration. In Section 4.4, we describe a practical algorithm for numerical integration and discuss the main computational bottlenecks. We then give numerical experiments on toy examples in moderate dimensions to illustrate our result.

4.2 Related work

For the reader to have a few elements of comparison in mind, we quickly pinpoint related work from Chapter 2 and Chapter 3.

4.2.1 Concentration for independent sampling and MCMC

We first recall the results discussed in Section 2.2.2 for i.i.d points and Section 2.3 for MCMC.

For the Wasserstein distance, concentration inequalities with exponential decay at speed n have been obtained for i.i.d points (2.52). Moreover, bounds in expectation of optimal order $n^{-1/d}$ are available for i.i.d points (2.54). A similar bound with an additional error term is also known for ergodic Markov chains (2.71).

We will rather seek bounds on the MMD (2.19). Assume that K defines a RKHS (Definition 2.1.5) and is bounded, e.g. $|K(x,x)| \leq C$ for some $C > 0$ and all $x \in \mathbb{R}^d$. Then, we know that when the nodes $x_1, \dots, x_n$ are i.i.d, the MMD with kernel K is exactly of order $n^{-1/2}$ in probability for translation invariant kernels (Tolstikhin et al., 2016, Theorem 6), and that it is bounded by $n^{-1/2}$ in expectation (2.55). For ergodic Markov chains, we further have an explicit concentration inequality with speed n (2.72), which leads as well to an $n^{-1/2}$ rate for the MMD.

Overall, the reader can keep in mind that the smoother test functions used in the MMD (2.19) allows to get rid of the dimension in the rate.

4.2.2 Concentration for kernel methods and Coulomb gases

We now turn to faster concentration inequalities or expectation bounds for the MMD presented in Section 2.6.

Assume that K admits a Mercer decomposition as in (2.20). In Theorem 2.6.2, we saw that nodes $x_1, \dots, x_n$ drawn with continuous volume sampling (2.112) achieve the $\sqrt{\sigma_n}$ rate for the MMD, where σ_n is defined in (2.20). This rate is optimal, in the sense that $\sigma_{n+1} \sim \sigma_n$ for most kernels K used in machine learning, and $\sqrt{\sigma_{n+1}}$ is the best MMD rate that can be achieved uniformly over L^2 targets π using n points (2.39). For now, continuous volume sampling however comes with a practical issue, namely an exponential dependence in the dimension d in the sampling procedure, see Section 2.6.3. Thinning with leverage score sampling also achieves close-to-optimal MMD in probability for Gaussian and Sobolev kernels (Theorem 2.6.5), yet assuming that we have access to a very large number of i.i.d input points.

Assuming that we have access to only n^2 input points from an ergodic Markov chain, kernel thinning provides fast, yet suboptimal, rates for the MMD in probability (Theorem 2.6.3). Yet, due to logarithmic penalties, the bounds of Theorem 2.6.3 are only interesting when the number of input points n^2 is at least exponential with the dimension d. There is however empirical evidence that kernel thinning scales to moderate dimensions d (Dwivedi et al., 2021).

In this chapter, our goal is to provide a concentration inequality for the MMD with general bounded kernels K for points $x_1, \dots, x_n$ drawn from a Gibbs measure (4.1), with fast and dimension-independent concentration speed. Large deviations at speed β_n have been obtained for general bounded kernels in (Chafaï et al., 2014), and we provide here a non-asymptotic counterpart.

We emphasize that we do *not* claim or hope for faster rates than the standard $n^{-1/2}$ one for the MMD. We rather obtain tighter coverage than with i.i.d or MCMC points, as a consequence of the faster concentration speed.

We are inspired by Theorem 3.3.5 which gives a fast concentration inequality in the Coulomb case. In this case, due to the singularity, the MMD is replaced by worst-case

errors on bounded Lipschitz functions, and our goal is to provide a similar inequality for the MMD with general but bounded kernels K. Maybe surprisingly, while our proof follows the lines of (Chafaï et al., 2018), we were able to considerably simplify the more technical arguments.

4.3 A concentration inequality for Gibbs measures

We first introduce our assumptions and state Theorem 3.3.3 for general interaction kernels following (Chafaï et al., 2014).

4.3.1 Definitions

We will extensively work with the energies introduced in Definition 3.3.2. Furthermore, we will work under the following assumptions on K and V.

The first one restricts our class of interaction kernels, insisting that points should repel, but that the interaction cannot be singular.

Assumption 4.3.1 K is symmetric, non-negative, continuous, and bounded on the diagonal: there exists some constant $C \geq 0$ such that $K(x,x) \leq C < \infty$ for all $x \in \mathbb{R}^d$.

Assumption 4.3.1 in particular ensures that $I_K(\mu)$ is well-defined for any probability measure, with possibly infinite value. Our next assumption excludes pathological cases where I_K does not induce a distance on probability distributions.

Assumption 4.3.2 K is integrally strictly positive definite (ISPD), i.e. $I_K(\mu) > 0$ for any non-zero finite signed measure Borel μ.

First, Assumptions 4.3.1 and 4.3.2 allow most kernels used in machine learning (Rasmussen et al., 2005), like the Gaussian (2.24) or isotropic Matérn kernels (2.26), as well as truncated singular kernels like the multiquadratic kernel (2.25). Second, under Assumptions 4.3.1 and 4.3.2, K is finite on the diagonal and ISPD, so that it is in particular positive definite. We can then consider the RKHS $\mathcal{H}_K$ induced by the kernel K as in Definition 2.1.5. An easy consequence of the Cauchy-Schwarz inequality in $\mathcal{H}_K$ is that for any $x, y \in \mathbb{R}^d$, $0 \leq K(x,y) \leq C$. In particular, $I_K(\mu,\nu)$ and $U_K^\mu(z)$ from Definition 3.3.2 are well-defined and finite for all finite signed Borel measures; see (Pronzato and Zhigljavsky, 2020) for more details. We henceforth denote by $\mathcal{E}_K$ (respectively $\mathcal{E}_K^V$) the set of finite signed Borel measures with finite interaction energy $I_K(\mu)$ (respectively, with finite energy $I_K^V(\mu)$). We recall the following duality formula, which we already stated in Proposition 4.3.3 and links energy minimization with quadrature guarantees for integrands in the unit ball of $\mathcal{H}_K$.

Proposition 4.3.3 — Sriperumbudur et al., 2009. Under Assumptions 4.3.1 and 4.3.2,

for probabilities μ, ν in $\mathcal{E}_K$, let

$$\mathrm{MMD}_K(\mu, \nu) = \sup_{\|f\|_{\mathcal{H}_K} \leq 1} \left| \int f \, \mathrm{d}(\mu - \nu) \right|. \tag{4.3}$$

Then

$$\mathrm{MMD}_K(\mu, \nu) = \left(I_K(\mu - \nu)\right)^{1/2}. \tag{4.4}$$

We add an assumption to make sure that V is strong enough a confining term. Together with Assumption 4.3.1, this makes I_K^V well-defined for any probability measure, with possibly infinite value. Note that, unlike I_K, we shall only evaluate I_K^V on probability measures.

Assumption 4.3.4 V is lower semi-continuous, finite everywhere and $V(x) \to +\infty$ when $|x| \to +\infty$. Moreover, there exists a constant $c > 0$ such that

$$\int \exp\left(-cV(x)\right) \, \mathrm{d}x < \infty. \tag{4.5}$$

We are now ready to consider the minimizers of I_K^V. Recall that our goal will be to tune V so that I_K^V has a unique minimizer given by the target distribution π.

Proposition 4.3.5 Let K satisfy Assumptions 4.3.1 and 4.3.2, and V satisfy Assumption 4.3.4. Then I_K^V is lower semi-continuous, has compact level sets and

$$I_K^V(\mu) > -\infty \tag{4.6}$$

for any probability distribution μ on $\mathbb{R}^d$. Moreover, if $\mu \in \mathcal{E}_K^V$, then $I_K(\mu)$ and $\int |V| \, \mathrm{d}\mu$ are finite, and

$$I_K^V(\mu) = \frac{1}{2} I_K(\mu) + \int V \, \mathrm{d}\mu. \tag{4.7}$$

I_K^V is strictly convex on the convex non-empty set $\mathcal{E}_K^V$, I_K^V has a unique minimizer μ_V over the set of probability measures on $\mathbb{R}^d$, called the *equilibrium measure*, and the support of μ_V is compact.

The proof borrows from (Chafaï et al., 2014) and (Pronzato et al., 2020); see Section 4.5.

Without further ado, we can state our concentration results for our Gibbs measure in the next section.

4.3.2 Main results

We saw in Section 2.6 that herding-like algorithms rely on finding a configuration of points $\{x_1, \dots, x_n\}$ that minimizes the interaction energy $I_K(\frac{1}{n}\sum_{i=1}^n \delta_{x_i} - \pi)$. In this chapter, we rather consider points drawn from the Gibbs measure (4.1) that favors small values for an empirical proxy of this interaction energy.

Definition 4.3.6 — Gibbs measure. Let K satisfy Assumptions 4.3.1 and 4.3.2, and V satisfy 4.3.4. Let $\beta_n \geq 2cn$, where c is the constant of Assumption 4.3.4. The measure $\mathbb{P}^V_{n,\beta_n}$ defined on $\mathbb{R}^{dn}$ by (4.1) is a *bona fide* probability measure. In particular, $Z^V_{n,\beta_n} \in (0, +\infty)$ (Chafaï et al., 2014, 2018).

Recall that we denote by μ_n the empirical measure (3.1) of the points $x_1, \dots, x_n$. When $x_1, \dots, x_n \sim \mathbb{P}^V_{n,\beta_n}$, we know that μ_n converges to the equilibrium measure μ_V, with a large deviation speed given by β_n (Chafaï et al., 2014). We are able to give non-asymptotic guarantees on this convergence via a concentration inequality, which is the main result of the chapter.

Theorem 4.3.7 — Concentration inequality. Let K satisfy Assumptions 4.3.1 and 4.3.2, and V satisfy Assumption 4.3.4. Further assume that the associated equilibrium measure μ_V has finite entropy

$$S(\mu_V) = -\int \log \mu_V \, \mathrm{d}\mu_V. \tag{4.8}$$

Let $\beta_n \geq 2cn$, where c is the constant of Assumption 4.3.4. Then, for any $n \geq 2$ and for any $r > 0$,

$$\mathbb{P}^V_{n,\beta_n}\left(I_K\left(\mu_n - \mu_V\right) > r^2\right) \leq \exp\left(-\frac{1}{4}\beta_n r^2 + n(c_1 + \frac{\beta_n}{n^2}c_2)\right), \tag{4.9}$$

with c_1, c_2 given by

$$c_1 = cI^V_K(\mu_V) + \log \int_{\mathbb{R}^d} \exp\left(-cV(x)\right) \mathrm{d}x - S(\mu_V), \tag{4.10}$$

$$c_2 = \frac{1}{2}C - \frac{1}{2}I_K(\mu_V), \tag{4.11}$$

and C is the constant of Assumption 4.3.1.

We emphasize again that by Proposition 4.3.3, (4.9) provides a non-asymptotic confidence interval for the worst-case quadrature error in the unit ball of the RKHS $\mathcal{H}_K$. Note that the bound is only useful in the regime $\beta_n/n \longrightarrow +\infty$, where the temperature $1/\beta_n$ goes down quickly enough. A classical choice of temperature scale is $\beta_n = \beta n^2$ where $\beta > 0$, this is the standard regime of random matrix theory (Section 3.2.3) and the one studied

in (Chafaï et al., 2018). We can rephrase a bit to get a more explicit sub-Gaussian decay in the bound.

Corollary 4.3.8 Under the assumptions of Theorem 4.3.7, let further $\beta_n \gg n$. Then, for any $n \geq 2$ and for any

$$r \geq 4 \max\left(\sqrt{c_2} n^{-1/2}, \sqrt{c_1}\left(\beta_n / n\right)^{-1/2}\right), \tag{4.12}$$

$$\mathbb{P}^V_{n,\beta_n}\left(\mathrm{MMD}_K\left(\mu_n, \mu_V\right) > r\right) \leq \exp\left(-\frac{1}{8}\beta_n r^2\right). \tag{4.13}$$

The proof of Corollary 4.3.8 is straightforward from Theorem 4.3.7, itself proved in Section 4.5.

A few remarks are in order. First, when $\beta_n \geq v n^2$ for some constant $v > 0$, the condition (4.12) on r becomes $r > u_0 n^{-1/2}$ for some constant $u_0 > 0$.

Second, integrating (4.13) over all $r > 0$ yields

$$\mathbb{E}\left[\mathrm{MMD}_K\left(\mu_n, \mu_V\right)\right] \leq \frac{u_1}{\sqrt{n}} + \frac{u_2}{\sqrt{\beta_n}} \tag{4.14}$$

for some constants $u_1, u_2 > 0$, and we thus recover the usual "slow" $n^{-1/2}$ rate for the MMD obtained for independent sampling, MCMC, or deterministic herding and its variants (Section 2.6). The improvement from our result compared to MCMC or i.i.d points rather comes at the level of coverage. In fact, under i.i.d or MCMC sampling, results like (2.52) or (2.72) feature n instead of β_n in the right-hand side of (4.13), and $\beta_n / n \to +\infty$. This means that for a fixed precision level $r > 0$ for the MMD, fewer points will be needed with a Gibbs measure to achieve the same coverage as i.i.d or MCMC points, as soon as the condition (4.12) on r still holds. This fast-increasing coverage probability of our confidence interval is a trace of the repulsion in the Gibbs measure.

I now explain in concrete terms how Corollary 4.3.8 implies that Monte Carlo integration with our Gibbs measure and with respect to an arbitrary target distribution π outperforms crude Monte Carlo.

4.3.3 Application to Monte Carlo integration

Let $d \geq 1$ and π be a probability measure on $\mathbb{R}^d$, which we assume to be our target. If we want to use Corollary (4.3.8) for Monte Carlo integration, we need to make sure that $\mu_V = \pi$. We will work under the following assumption.

Assumption 4.3.9 The support $S_\pi \subset \mathbb{R}^d$ of π is compact, and π has finite entropy, in the sense that π has a density π' w.r.t. Lebesgue, and that

$$S(\pi) := -\int \log \pi'(x) \mathrm{d}\pi(x) < \infty. \tag{4.15}$$

The following proposition shows that, for a given kernel K, we can choose V so that $\mu_V = \pi$, assuming prior computation of the kernel embedding $U_K^\pi(z)$ for all $z \in \mathbb{R}^d$.

Proposition 4.3.10 Let K satisfy Assumptions 4.3.1 and 4.3.2, and π satisfy Assumption 4.3.9. In particular, there exists $R > 0$ such that $S_\pi \subset B(0, R)$, where $B(0, R)$ is the closed Euclidean ball.

Let $\Phi : \mathbb{R}^d \to \mathbb{R}_+$ be any continuous function such that $\Phi = 0$ on $\partial B(0, R)$, $\Phi(z) \to +\infty$ as $|z| \to +\infty$ and

$$\int_{\{|x|>R\}} e^{-\Phi(x)} \, \mathrm{d}x < \infty. \tag{4.16}$$

Moreover, set

$$V^\pi(z) = \begin{cases} -U_K^\pi(z) \text{ when } z \in B(0, R), \\ -U_K^\pi(z) + \Phi(z) \text{ otherwise.} \end{cases} \tag{4.17}$$

Then, V^π satisfies Assumption 4.3.4 with $c = 1$ and $\mu_{V^\pi} = \pi$.

This is a standard result, which relies on the so-called Euler-Lagrange characterization of the equilibrium measure. We give a proof in Section 4.5, which is inspired by Corollary 1.4 of (Chafaï et al., 2014), who treat the more difficult case of singular interactions.

A classical choice of Φ is $\Phi(z) = (|z|^2 - R^2)_+$. In machine learning terms, Proposition 4.3.10 says that a suitably penalized kernel embedding is a good choice of confining potential. This should be no surprise, since we already discussed in Section 3.3.2 that with this choice of V, the energy (4.2) is a discrete version of the MMD (2.97).

Of course, this choice of V requires the ability to evaluate the kernel embedding U_K^π, and we fall back here onto a standard limitation in the herding literature, see Section 2.6. While closed form expression for U_K^π are available (Briol et al., 2019, Table 1) for simple pairs of kernel and target (K, π), there is no general expression.

In the literature, two main approaches are used to bypass this computation. First, one can rely on i.i.d or MCMC sampling to approximate U_K^π. For instance, kernel thinning (Section 2.6.4) relies on n^2 MCMC input points for this computation. Another approach is to build a new kernel K_π from K and π such that $U_{K_\pi}^\pi$ is zero: this is the Stein kernel approach presented in Section 2.6.2. We will use MCMC to approximate U_K^π in our experiments (Section 4.4). Moreover, we will show in Chapter 5 that this does not downgrade our guarantees asymptotically, nor non-asymptotically when the number of MCMC points in the approximation scales like n^2, see Section 6.1.1.

Following Proposition 4.3.10, the following corollary to Theorem 4.3.7 gives a

concentration inequality with speed β_n, this time for the MMD between the empirical measure of points drawn from our Gibbs measure and the target π.

Corollary 4.3.11 Let $\beta_n \gg n$, $d \geq 1$ and π satisfy Assumption 4.3.9. Let V^π be the confinement in (4.17), and let K satisfy Assumptions 4.3.1 and 4.3.2.

Then, for any $n \geq 2$ and any

$$r \geq 4 \max\left(\sqrt{c_2} n^{-1/2}, \sqrt{c_1}\left(\beta_n / n\right)^{-1/2}\right), \tag{4.18}$$

$$\mathbb{P}^V_{n,\beta_n}\left(\mathrm{MMD}_K\left(\mu_n, \pi\right) > r\right) \leq \exp\left(-\frac{1}{8}\beta_n r^2\right), \tag{4.19}$$

where

$$c_1 = -\frac{1}{2} I_K(\pi) - S(\pi) + \log \int_{\mathbb{R}^d} \exp\left(-V^\pi(x)\right) \mathrm{d}x, \tag{4.20}$$

$$c_2 = \frac{1}{2} C - \frac{1}{2} I_K(\pi). \tag{4.21}$$

The last expressions for c_1 comes from the fact that $c = 1$ in Proposition 4.3.10 and that

$$\begin{aligned} I_K^{V^\pi}(\pi) &= \frac{1}{2} \iint K(x,y) \,\mathrm{d}\pi(x) \,\mathrm{d}\pi(y) - \iint K(x,y) \,\mathrm{d}\pi(x) \,\mathrm{d}\pi(y) + \int_{|x|>R} \Phi(x) \,\mathrm{d}\pi(x) \\ &= -\frac{1}{2} I_K(\pi). \end{aligned} \tag{4.22}$$

We finally rephrase Corollary 4.3.11 in terms of coverage for numerical integration.

Corollary 4.3.12 We work under the assumptions of Corollary 4.3.11. Let $x_1, \ldots, x_n \sim \mathbb{P}^{V^\pi}_{n,\beta_n}$ and let further $\epsilon > 0$ and $\delta \in (0,1)$. For any n such that

$$\epsilon \geq 4 \max\left(\sqrt{c_2} n^{-1/2}, \sqrt{c_1}\left(\beta_n / n\right)^{-1/2}\right) \tag{4.23}$$

and $\exp\left(-\beta_n \epsilon^2 / 8\right) \leq \delta$, with probability larger than $1 - \delta$,

$$\sup_{\|f\|_{\mathcal{H}_K} \leq 1} \left| \int f \,\mathrm{d}(\mu_n - \pi) \right| \leq \epsilon. \tag{4.24}$$

Compared to (2.52) and (2.72), for a fixed worst-case integration error ϵ, fewer points are required under $\mathbb{P}^{V^\pi}_{n,\beta_n}$ than under i.i.d or MCMC samples from π to reach confidence level δ. In particular, β_n replaces n in the constraint that links δ and ϵ, while the constraint on ϵ stays the same as soon as β_n is bigger than a constant times n^2. I think this is quite satisfying, if the goal is to provide tight confidence regions using a small number of points for target integrals.

A few comments are in order. First our assumptions on the RKHS are generic, and one can see that the concentration result (4.19) does not depend on the kernel K or the target π except for the constants c_1 and c_2. This is to be expected from the statistical physics literature, since Gibbs measures typically exhibit *universality*: the fluctuations or local properties of the point process are expected to be independent of its global parameters. While this universality seems appealing at first sight, we will see in Section 6.1.2 that this can also be a curse if the goal is to get optimal results.

Second, a strength of our result is that the constants and speeds are quite explicit. We could be worried that the constants c_1 and c_2 hide exponential dependencies in the dimension: I will show you in the following proposition that this is not the case.

Proposition 4.3.13 We work under the assumptions of Corollary 4.3.11. Recall that C is a positive constant such that $0 \leq K(x,x) \leq C$ for all $x \in \mathbb{R}^d$, and that R is a positive constant such that the support of π is included in $B(0,R)$. We further choose $\Phi(x) = \|x\|^2 - R^2$ in Proposition 4.3.10. Then,

$$c_1 \leq 2C - S(\pi) + d\left(\frac{\sqrt{2}+1}{2}\log(\pi) + \log R\right) + R^2 - \log\Gamma\left(d/2+1\right), \tag{4.25}$$

$$c_2 \leq \frac{1}{2}C, \tag{4.26}$$

where Γ is Euler's Gamma function, meaning that

$$\log\Gamma(d/2+1) = \mathcal{O}\left(d\log d\right) \tag{4.27}$$

as $d \to +\infty$ using Stirling's formula. Moreover,

$$-S(\pi) := \int \log\pi'(x)\,\mathrm{d}\pi(x) \leq \|\log\pi'\|_\infty, \tag{4.28}$$

as soon as the density π' of π is continuous on its compact support S_π.

The proof is given in Section 4.5. We thus see from Proposition 4.3.13 that there is no intrinsic bad dependence of the constants in the dimension d. Of course, one could choose a kernel K which is close to be singular such as a regularization of (3.43), so that the constant C blows up exponentially with d, but this issue would transfer to all kernel-based methods for such a K. The same goes with choosing π such that R scales very badly with the dimension d. Moreover, the bounds in Proposition 4.3.13 are far from sharp, and we completely discarded the negative energy terms in (4.20).

Another high-level remark is that the concentration bounds in Corollary 4.3.12 and Corollary 4.3.11 suggest that considering β_n as large as possible would provide the best guarantee in terms of coverage. This is to be tempered with the fact that sampling

points from the Gibbs measure (4.1) is intuitively harder when β_n is very large.

The elephant in the room is that there is no general algorithm to draw points $x_1, \dots, x_n$ from $\mathbb{P}^{V^\pi}_{n,\beta_n}$, and we rely on MCMC to do so in this thesis, see Section 4.4. When β_n is very large, $\mathbb{P}^{V^\pi}_{n,\beta_n}$ becomes very spiked around the minimizers of the energy (4.2), which is typically a bad situation for MCMC algorithms, meaning that many iterations might be needed to get a good approximated sample. Quantifying how this approximated sampling interplays with the statistical guarantees of Corollary 4.3.11 is thus crucial.

The following corollary gives a general, yet still unpractical, result for the statistical guarantees obtained when approximately sampling from $\mathbb{P}^{V^\pi}_{n,\beta_n}$.

Corollary 4.3.14 We work under the assumptions of Corollary 4.3.11. Consider a Markov kernel P_n on $\mathbb{R}^{dn}$ and an initial distribution ξ_n, and define the mixing time

$$\mathcal{T}_{\mathrm{TV},n,\beta_n}(\xi_n, \delta) = \inf\{T \geq 1 \,:\, \|\xi_n P_n^T - \mathbb{P}^{V^\pi}_{n,\beta_n}\|_{\mathrm{TV}} < \delta\}. \tag{4.29}$$

We denote by $x_1^{(T)}, \dots, x_n^{(T)}$ the state of the Markov chain with kernel P_n started at ξ_n after T iterations, and by $\mu_n^{(T)}$ the associated empirical measure. Let $\epsilon > 0$ and $\delta \in (0,1)$. Let n be such that

$$\epsilon \geq 4 \max\left(\sqrt{c_2} n^{-1/2}, \sqrt{c_1}\,(\beta_n/n)^{-1/2}\right) \tag{4.30}$$

and

$$2\exp\left(-\beta_n \epsilon^2/8\right) \leq \delta. \tag{4.31}$$

Further assume that

$$T \geq \mathcal{T}_{\mathrm{TV},n,\beta_n}(\xi_n, \delta/2). \tag{4.32}$$

Then, with probability larger than $1-\delta$,

$$\sup_{\|f\|_{\mathcal{H}_K} \leq 1} \left| \int f \,\mathrm{d}\left(\mu_n^{(T)} - \pi\right) \right| \leq \epsilon. \tag{4.33}$$

The proof just comes from the inequality

$$\mathbb{P}\left(\mathrm{MMD}_K\left(\mu_n^{(T)}, \pi\right) > \epsilon\right) \leq \mathbb{P}^{V^\pi}_{n,\beta_n}\left(\mathrm{MMD}_K\left(\mu_n, \pi\right) > \epsilon\right) + \|P_n^T \xi - \mathbb{P}^{V^\pi}_{n,\beta_n}\|_{\mathrm{TV}}, \tag{4.34}$$

by definition of the total variation distance (2.51), and by applying Corollary 4.3.12 with $\delta/2$.

Corollary 4.3.14 gives tighter coverage than with n i.i.d points, this time for the true updates used by the algorithm (Section 4.4).

Yet, to my knowledge, there are no usable bounds for $\mathcal{T}_{\mathrm{TV},n,\beta_n}(\xi_n,\delta)$ in the MCMC literature. In my opinion, this is the main limitation of our approach. Nonetheless, I believe that we could be able to understand mixing times for specific Markov kernels tailored to our Gibbs measure, we will go back to this in Chapter 6.

For now, we will use standard Markov kernels P_n in Section 4.4 along with some heuristics to try to understand the dependence in β_n. We now turn to a practical description of the sampling algorithm, discuss the main bottlenecks, and illustrate our results on toy examples.

4.4 Numerical experiments

4.4.1 Sampling from the Gibbs measure

Markov kernels

As we just discussed, there is no known algorithm to sample exactly from (4.1) for generic K and V, so we resort to MCMC with target $\mathbb{P}^{V^\pi}_{n,\beta_n}$. In the case of singular interactions, Hamiltonian Monte Carlo algorithms have been investigated in (Chafaï and Ferré, 2019) to sample from the Gibbs measure (4.1). In this thesis, we will simply rely on MALA (2.66) with target $\mathbb{P}^{V^\pi}_{n,\beta_n}$. This is Metropolis–Hastings with proposal

$$y|y_t \sim \mathcal{N}\left(y_t - \alpha\beta_n \nabla H_n(y_t), 2\alpha I_{dn}\right), \tag{4.35}$$

where $y \in (\mathbb{R}^d)^n$, and α is a user-tuned step size parameter. To approximately sample from $\mathbb{P}^{V^\pi}_{n,\beta_n}$, we run T MALA iterations and keep *only* the $T-$th iteration as our approximate sample $(x_1^{(T)},\dots,x_n^{(T)}) \in \mathbb{R}^{dn}$. The target integral is then approximated by averaging over the n points $x_1^{(T)},\dots,x_n^{(T)}$ as in (4.33). An approximate sample from our Gibbs measure thus costs $\mathcal{O}(Tn^2)$, since each evaluation of the density (4.1) requires computing pairwise interactions at each MALA iteration. This is the standard cost for kernel methods such as kernel herding or gradient descents on the energy (4.2), see Section 2.6.

On the other hand, our baseline comparison will be vanilla MCMC, targeting π as in Section 2.3. The latter runs a Markov chain on $\mathbb{R}^d$ with target π and n iterations, and averages over the chain to approximate the target integral, thus using n quadrature nodes in $\mathbb{R}^d$ as well. Of course, our MALA chain (4.35) runs in much higher dimension $(d \times n)$, but we emphasize that we are not comparing the convergence of this MALA chain to $\mathbb{P}^{V^\pi}_{n,\beta_n}$ versus that of vanilla MCMC to its target π. Rather, we compare how our estimator based on a single sample from $\mathbb{P}^{V^\pi}_{n,\beta_n}$ (obtained after $T \gg n$ MALA iterations) compares to the estimator averaging over the n vanilla MCMC iterations.

Like any other kernel method trying to minimize (2.97), our procedure clearly requires more CPU time, yet what we are comparing is the performance of both rules as a function of n. This is fair in the setting of expensive-to-evaluate integrands f, where the computational bottleneck is the number n of times that we have to evaluate f

and one can usually afford a larger CPU time to pick the nodes, see Chapter 1. The important point is that the output n nodes result in a confidence interval for integral approximation that is tighter than with n "classical" MCMC iterations, which is what our Corollary 4.3.14 claims.

We now discuss how we choose the step-size α and give heuristics on mixing times for $\mathbb{P}^{V^\pi}_{n,\beta_n}$.

Scaling the step-size

We choose the MALA step size as $\alpha = \alpha_0 \beta_n^{-1}$, and α_0 is manually tuned at the beginning of each run so that acceptance reaches 50%. Having α decrease at least as β_n^{-1} intuitively avoids the distance between two consecutive MALA states to grow with n; see the MALA proposal (4.35). This choice is motivated by the fact that we do not wish for the number of iterations T in Corollary 4.3.14 to scale with β_n to reach a good approximation of $\mathbb{P}^{V}_{n,\beta_n}$ and thus π.

In fact, consider a target measure $\mu = e^{-h(x)}\,\mathrm{d}x$ on $\mathbb{R}^{d'}$ which is L-smooth and m-strongly log-concave in the sense that h is $\mathcal{C}^2$ and

$$m I_{d'} \preccurlyeq \nabla^2 h \preccurlyeq L I_{d'} \tag{4.36}$$

with $L > m > 0$. Recall that for any initial distribution ξ and any $\delta > 0$, we denote by

$$\mathcal{T}_{\mathrm{TV}}(\xi, \delta) = \inf\{T \geq 1 : \|\xi P^T - \pi\|_{\mathrm{TV}} < \delta\} \tag{4.37}$$

the mixing time of the Markov chain with Markov kernel P, in total variation. In particular, we write $\mathcal{T}^{\mathrm{MALA}}_{\mathrm{TV}}(\xi, \delta)$ the mixing time for MALA.

Assuming that the initial distribution ξ of the Markov chain is already close to μ — this is called a *warm start* — Dwivedi et al. (2019, Theorem 1) show that

$$\mathcal{T}^{\mathrm{MALA}}_{\mathrm{TV}}(\xi, \delta) = \mathcal{O}\left(\max\{d'\kappa, \sqrt{d'}\kappa^{3/2}\} \log\left(\frac{1}{\delta}\right)\right), \tag{4.38}$$

where $\kappa = L/m$, as soon as

$$\alpha = \min\left\{\frac{\sqrt{m}}{c(\delta, d')L\sqrt{d'L}}, \frac{1}{Ld'}\right\}, \tag{4.39}$$

where $c(\delta, d')$ is a function of δ and d'. Similar statements are available for random-walk Metropolis–Hastings described in (2.64), see (Dwivedi et al., 2019, Theorem 2). The polynomial dependence in the dimension d' support the claim that MCMC algorithms are able scale to moderate to high dimensions with a good initialization and in the convex setting.

In our case, the target is $\mathbb{P}^{V^\pi}_{n,\beta_n}$. We thus have $d' = dn$ and m and L would depend linearly on β_n. If the energy H_n in (4.2) was to satisfy (4.36) with m_n and L_n, then so would

the log-density of $\mathbb{P}^{V^\pi}_{n,\beta_n}$ with $\tilde{m}_n = \beta_n m_n$ and $\tilde{L}_n = \beta_n L_n$. Thus, $\kappa = \tilde{L}_n/\tilde{m}_n = L_n/m_n$ would be independent of the choice of β_n and so would the mixing time (4.38).

From (4.39), this requires to consider a step-size

$$\alpha = \beta_n^{-1} \min\left\{ \frac{\sqrt{m_n}}{c(\delta, dn) L_n \sqrt{dn L_n}}, \frac{1}{n L_n d} \right\}, \tag{4.40}$$

thus depending on β_n like β_n^{-1}.

This is only heuristic for numerous reasons. First, $\mathbb{P}^{V^\pi}_{n,\beta_n}$ is not at all log-concave in general, let alone strongly log-concave. Even if it was only weakly log-concave in the sense that (4.36) is satisfied with $m = 0$, which is still an unreasonable assumption for $\mathbb{P}^{V^\pi}_{n,\beta_n}$, then Dwivedi et al. (2019) show that the mixing time is $\mathcal{O}(d'^2 L^{3/2}/\delta^{3/2})$. Thus, there would this time be a dependence on β_n coming from $L = \tilde{L}_n = \beta_n L_n$. In Chapter 6, we will discuss recent works trying to depart from the log-concave setting, yet the mentioned results are still far from ready-to-use in our case.

More importantly, with a small step-size α or large β_n, many MALA iterations might be necessary to reach a warm start, i.e a region of large mass under $\mathbb{P}^{V^\pi}_{n,\beta_n}$ (or any other target measure). In practice, what we could recommend is to first run a fast-scaling deterministic method such as SVGD (Section 2.5.2) or KSD descent (Section 2.6.2), which are in a sense the warmest starts available in the literature, and use their result to initialize our algorithm, in order to benefit from our concentration guarantees. In the end, these mixing properties are more a feature from MALA than from our Gibbs measure, and there likely are dedicated sampling algorithms that will outperform MALA at approximating $\mathbb{P}^{V}_{n,\beta_n}$, see Chapter 6.

Nonetheless, there is thus a trade-off between large β_n (i.e. concentrated $\mathbb{P}^{V^\pi}_{n,\beta_n}$ and small confidence regions) and sampling cost. We advocate the use of $\beta_n \propto n^2$, which we could satisfyingly implement and which already provably outperforms vanilla MCMC in our theoretical result and in the toy experiments of this chapter. In practice, we have been able to experiment up to $\beta_n \propto n^5$ but not higher, as numerical errors then pop up.

Approximation of the kernel embedding

Evaluating the kernel embedding is key to design the Gibbs measure $\mathbb{P}^{V^\pi}_{n,\beta_n}$ that concentrates around π. Since $U^\pi_K(z)$ does not usually have a closed form, a standard way (Pronzato et al., 2020) to approximate it is to run a long Metropolis–Hastings chain of length M targeting π and average $K(z, .)$ along the chain. Recall that this is fair for expensive-to-evaluate target integrands f, since no additional evaluation of f is involved. From the computational point of view, since any evaluation of U^π_K needs to be replaced by a sum of size M, the computational complexity of each iteration of MALA becomes $\mathcal{O}(n^2 + nM)$.

If one insists on avoiding having to approximate U^π_K, the Stein kernel K_π associated to K and π has emerged as a tool to bypass this approximation when π is either fully

supported on $\mathbb{R}^d$, or when π has compact support up to a smooth modification of the kernel K, see Section 2.6.2. In fact, we have that

$$U^{\pi}_{K_\pi}(z) = 0, \tag{4.41}$$

for $\pi-$almost all z by design. Nonetheless, as discussed in Section 2.6.2, changing K for the induced Stein kernel changes the RKHS of integrands in a nontrivial way, which is not desirable for numerical integration.

We thus stick in this chapter to the standard Monte Carlo approximation for U^{π}_{K}. Our experiments on toy models empirically validate this approximation, since we recover known statistical properties of the target Gibbs measure; see Figure 4.7 with a reasonably small number of Monte Carlo iterations M.

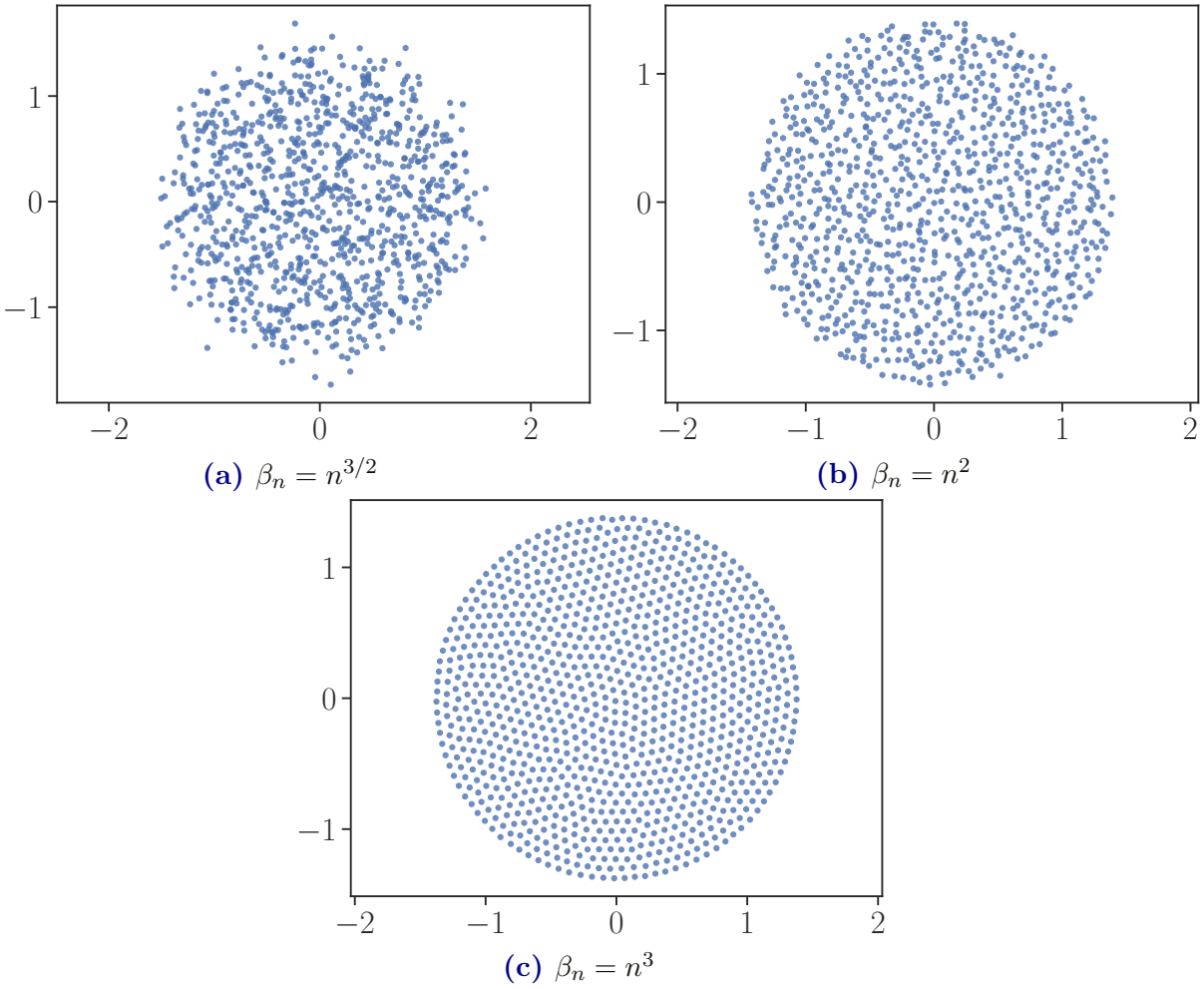
In Chapter 5, we will indeed show that this approximation does not downgrade the guarantees for our Gibbs measure, at least asymptotically. In the same vein as kernel thinning (Section 2.6.4), we will also show in Chapter 6 that, with a very crude bound, $M = n^2$ input points are enough to preserve the same non-asymptotic guarantees as in Corollary 4.3.14. With this scaling, the complexity of MALA becomes however $\mathcal{O}(n^3)$ per iteration, so we will stick in the experiments to a smaller number of input MCMC points M.

Toy examples

We show in Figure 4.1 how points sampled from $\mathbb{P}^{V}_{n,\beta_n}$ look like in dimension $d = 2$, for the quadratic potential $V : x \mapsto |x|^2/2$, and for different scalings of the inverse temperature β_n. We consider the truncated logarithmic kernel $K(x, y) = -\log\left(|x - y|^2 + \epsilon^2\right)$, where ϵ is a truncation parameter that we set to $\epsilon = 10^{-2}$, and set the number of particles n to 1000. For $\epsilon = 0$, the equilibrium measure is known to be uniform on the unit disk, and we expect it to be close to uniform in the truncated case as well. In each panel of Figure 4.1, we show the state of the MALA chain after $T = 5000$ iterations, using Python and Jax (Bradbury et al., 2018).

We observe that the three empirical measures indeed approximate the uniform distribution on the disk, with more regular spacings as the inverse temperature grows. Figure 4.1b already shows a more regular arrangement of the points than under i.i.d. draws from the uniform distribution, while the lattice-like structure of Figure 4.1c is a manifestation of what physicists call *crystallization* (Serfaty, 2018): the Gibbs measure is concentrated around minimizers of the energy.

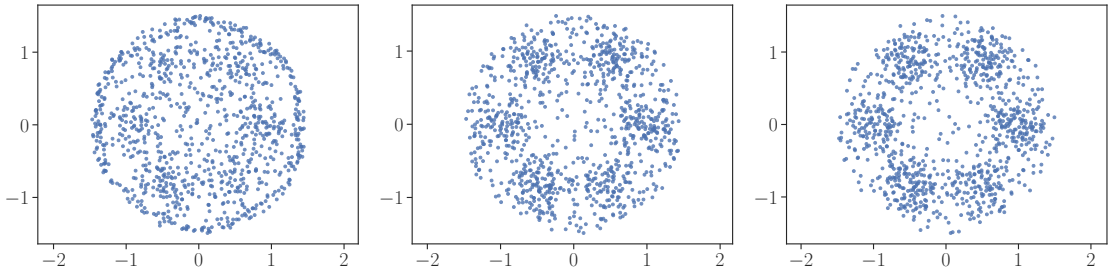
Like most MCMC samplers, the performance of MALA is influenced by the absence of a warm start and the multimodality of the target. We run an experiment targeting a mixture of 2D Gaussians, and illustrate how classical MCMC techniques like annealing tackle these issues. More precisely, we set π to be a balanced mixture of six two-dimensional truncated Gaussian distributions with radius 0.5, variance 0.1 and whose centers are evenly spaced on the unit circle. We stick to the truncated logarithmic interaction kernel as in the experiment of Figure 4.1, with $n = 1\,000$ and $\beta_n = n^2$,



(a) $\beta_n = n^{3/2}$ **(b)** $\beta_n = n^2$

(c) $\beta_n = n^3$

Fig. 4.1: Three independent approximate samples of $\mathbb{P}^V_{n,\beta_n}$, with different temperature schedules.



(a) $T = 5,000$ **(b)** $T = 10,000$ **(c)** $T = 15,000$

Fig. 4.2: Cold Gaussian start

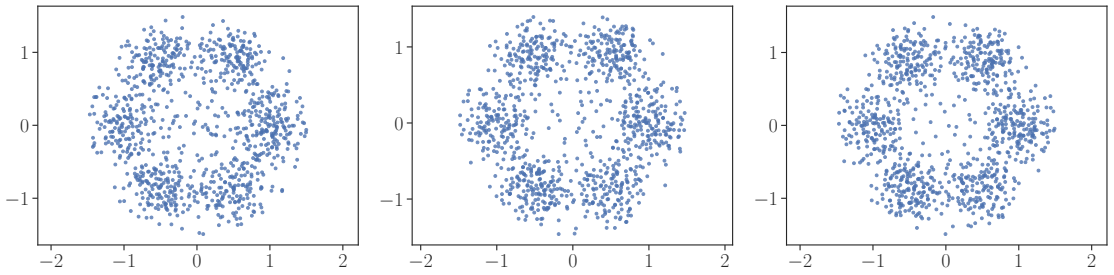
while the kernel embedding is approximated with $M = 10\,000$ iterations of MCMC as explained earlier.

Such multimodal target distributions are simple challenging toy examples in which one might suffer from a cold start and slow mixing. Nonetheless, if we take n large enough and an initial distribution with a large support, we can expect a few of the n points to be close to each mode of π, and the MALA chain to keep an appropriate proportion in each mode of π. However, as for any optimization algorithm targeting an objective like a worst-case integration error, we could suffer from a bad initialization with no particles in an isolated mode of π, and many MALA iterations might be needed to reach a good approximation of $\mathbb{P}^V_{n,\beta_n}$ and thus π.

In Figure 4.2, we plot the state of the MALA chain at three different iteration numbers, for a cold Gaussian start with mean 0 and variance 1. We then consider in Figure 4.3 a warm start obtained through $50,000$ iterations of KSD descent presented in Section 2.6.2, with Gaussian kernel and whose target is a mixture of Gaussian distributions centered in each of the modes of π with variance 0.1.

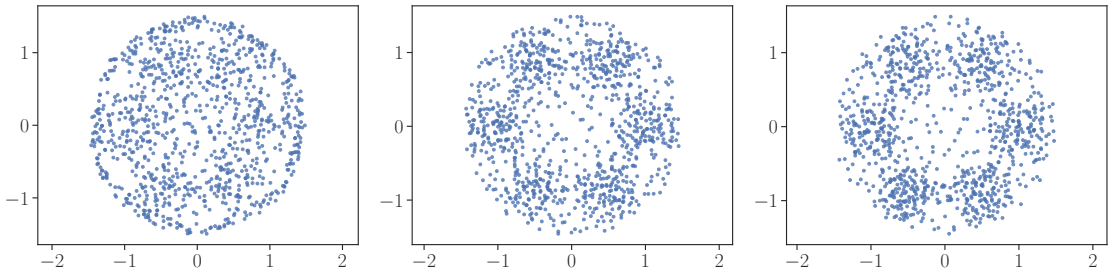
Another classical complementary solution for multimodal targets is to rely on annealing. To wit, we introduce a sequence of tempered targets $\pi_k(x) \propto e^{-t_k \log \pi(x)}$, for a sequence $(t_k)_{1 \leq k \leq l}$ going from 0.1 to 1. We successively use the result of the sampling algorithm on π_k as the initialization for the sampling algorithm on π_{k+1}, up to $\pi_l = \pi$. We choose here a naive annealing scheme $l = 10$ and $t_k = k/10$. In Figure 4.5, we then run our MALA updates *after* $50,000$ iterations of KSD descent for which annealing on π is used. In Figure 4.4, we run the MALA updates starting from a cold Gaussian start, but progressively annealing π in the updates. In detail, for each k, we approximate the kernel embedding of π_k using MCMC and then run MALA (for $T/10$ iterations, to keep our global number of MALA iterations constant) targeting the associated Gibbs measure as usual, and use the last iteration to initialize the same MALA updates for π_{k+1}, up to $\pi_l = \pi$ where the target measure of MALA becomes $\mathbb{P}^{V^\pi}_{n,\beta_n}$.

It appears clear, as expected, that the warm start helps in a sense that fewer iterations are needed to reach a well spread distribution in each of the modes, even if a cold start



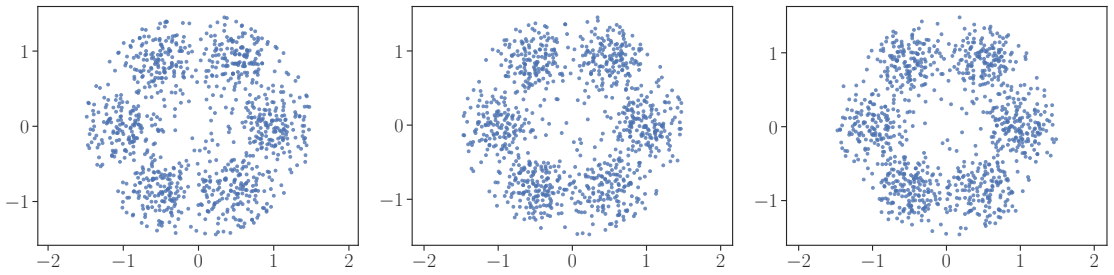
(a) $T = 5,000$ **(b)** $T = 10,000$ **(c)** $T = 15,000$

Fig. 4.3: **Warm KSD descent start**



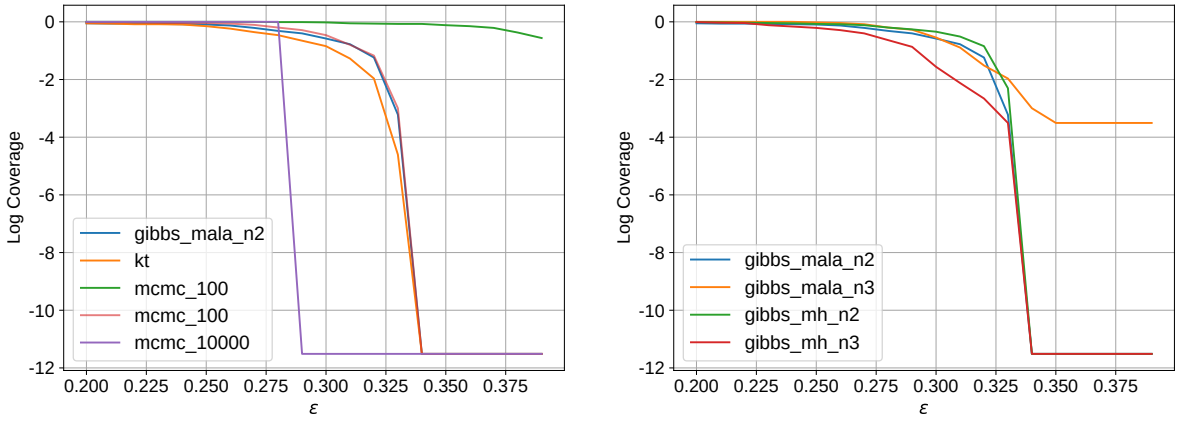
(a) $T = 5,000$ **(b)** $T = 10,000$ **(c)** $T = 15,000$

Fig. 4.4: **Cold Gaussian start (with annealing on π)**



(a) $T = 5,000$ **(b)** $T = 10,000$ **(c)** $T = 15,000$

Fig. 4.5: **Warm KSD descent start (with annealing on π)**



(a) $\mathbb{P}^{V^\pi}_{n,\beta_n}$ **versus MCMC**

(b) MH and MALA kernels for sampling $\mathbb{P}^{V^\pi}_{n,\beta_n}$

Fig. 4.6: Coverage comparisons between $\mathbb{P}^{V^\pi}_{n,\beta_n}$**, MCMC and kernel thinning (4.6a) and for different sampling schemes for** $\mathbb{P}^{V^\pi}_{n,\beta_n}$ **(4.6b).**

from a Gaussian distribution remains quite reasonable in that case in terms of number of MALA iterations. In this toy model, the effect of annealing on π does not seem tremendous.

A general remark is that we do not expect our method to perform well in cases where fast algorithms as SVGD (Section 2.5.2) or KSD descent (Section 2.6.2) fail. Another high level comment is that, in challenging applications of Monte Carlo integration to the physical sciences, we rarely trust an MCMC chain to find all the modes of a target. We usually repeatedly run a local optimization algorithm beforehand, with many random initial states, and cluster its output to find the modes of the target. Then only can we validate that our MCMC chain has visited all modes a sufficient number of times. Similarly, we only recommend using our algorithm after having heuristically identified all modes of π.

In the next sections, we focus on illustrating the guarantees obtained in Section 4.3 on toy examples.

4.4.2 Coverage of Monte Carlo integration

We first illustrate our main result from Corollary 4.3.14 stating that at a fixed precision level $\epsilon > 0$, Gibbs measures provide tighter confidence regions than vanilla Metropolis–Hastings targeting π with the same number of points.

We consider π to be a Gaussian distribution in dimension $d = 10$, with variance $\sigma^2 = 1$ and truncated at 0 outside $B(0, 2\sigma)$. We further consider a Gaussian kernel K (2.24) with $\sigma = 1$. We want to empirically evaluate the coverage obtained by each method. To do so, for each integration rule, we obtain 100 independent samples $(x_1, \dots, x_n)$ and

compute the proportion of samples such that

$$\mathrm{MMD}_{\mathrm{K}}(\mu_n, \pi) > \epsilon, \tag{4.42}$$

for each value of $\epsilon > 0$, where μ_n is the empirical measure (3.1). The MMD is approximated by replacing the integrals with respect to π in (2.106) by averages over MCMC chains targeting π with $T = 10^6$ iterations and independent of all the rest. This gives an estimate of the coverage

$$\mathbb{P}\left(\mathrm{MMD}_{\mathrm{K}}(\mu_n, \pi) > \epsilon\right) \tag{4.43}$$

for each quadrature rule, and we compare those estimates for different values of ϵ.

We compare the following methods, all initialized with the same distribution and run independently of one another. In Figure 4.6a, we consider the history of an MH chain of length n targeting π, with an isotropic Gaussian proposal tuned to reach 50% acceptance, for $n \in \{100, 1000, 10000\}$. We then consider $n = 100$ points $x_1, \ldots, x_n$ obtained via the kernel thinning algorithm (Section 2.6.4) with n^2 input points drawn again from an MH chain targeting π. Finally, we consider approximate samples from $\mathbb{P}_{n,\beta_n}^{V^\pi}$ after $T = 10\,000$ MALA updates described in Section 4.4.1 with $n = 100$ and $\beta_n = n^2$, with $M = 1\,000$ points drawn from an MH chain targeting π to approximately compute the kernel embedding U_K^π.

We see that $n = 100$ points approximately drawn from our Gibbs measure with temperature $\beta_n = n^2$ compare to $n = 100$ points obtained from kernel thinning in terms of coverage, and that both largely outperform crude MCMC with the same number of points. Yet, we would expect from Section 2.6.4 and Corollary 4.3.14 that both methods obtain similar coverages than vanilla MCMC with $n = 10\,000$ iterations. We see in Figure 4.6a that this is not the case, as MCMC with $n = 10\,000$ overperforms, while both Gibbs measure and kernel thinning with $n = 100$ points achieve guarantees comparable to MCMC with $n = 1\,000$ points.

The reasons can be many. First, one can expect Metropolis–Hastings to mix fast for such a simple choice of target π, and we will see in Section 5.4 a toy example in Bayesian classification where our Gibbs measure with n points indeed performs similarly to MCMC with n^2 points. Second, since we work at fixed n and dimension $d = 10$, the constants in (4.33) and (2.118) might play an important role in the observed results. Finally, for the Gibbs measure, we only used a small number $T = 10\,000$ of MALA iterations, and a small number $M = 1\,000$ of MCMC input points.

We particularly highlight the fact that we are able to match the results of kernel thinning in terms of coverage, for $\beta_n = n^2$ and with 10 times fewer input points, since we only chose $M = 1\,000$. These results have been obtained using CPUs with 2 GB of RAM. Each run of kernel thinning with $n = 100$ points took 15 seconds in average, while the MALA iterations to approximately sample from our Gibbs measure took around 65 seconds per run on average. Moreover, computationally cheaper, yet less readable,

versions of kernel thinning have been recently designed (see Section 2.6.4). We recall again that the main bottleneck is the number n of evaluations of the integrand f, so that runtimes are secondary.

In Figure 4.6b, we compare again the coverages obtained for $n = 100$ points approximately drawn from the Gibbs measure $\mathbb{P}^{V^\pi}_{n,\beta_n}$, using either MALA (4.35) or the simple random walk Metropolis–Hastings proposal (2.64) with target $\mathbb{P}^{V^\pi}_{n,\beta_n}$. We do so for the two temperature schedules $\beta_n = n^2$ and $\beta_n = n^3$, each time approximating the potential U^π_K with $M = 1\,000$ MCMC points.

At temperature $\beta_n = n^2$, we see that MALA performs slightly better than random walk MH, which is to be expected since the proposal (4.35) is informed by the log density of $\mathbb{P}^{V^\pi}_{n,\beta_n}$. We see that the coverage obtained in Figure 4.6b slightly improves when we increase $\beta_n = n^2$ to $\beta_n = n^3$ for points approximately drawn using random-walk MH. This is what we expect after a sufficient number of iterations according to Corollary 4.3.14. This is however not the case when using MALA, which means that more iterations are indeed needed in that case when going from $\beta_n = n^2$ to $\beta_n = n^3$ to produce good approximate samples from $\mathbb{P}^{V^\pi}_{n,\beta_n}$. Thus, while a nice behavior can be expected for the mixing time (4.29) in terms of β_n when using random-walk MH, this might not be the case with MALA. We will discuss ways to improve MALA in Chapter 6, along with other choices of Markov kernels that might help the sampling.

Note also that each run to approximately draw from our Gibbs measure with random-walk Metropolis–Hastings took around 15 seconds on average, which is much cheaper than with MALA and comparable to the vanilla version of kernel thinning that we used. This is due to the fact that the proposal (4.35) requires additional computation of the gradient of H_n at each step.

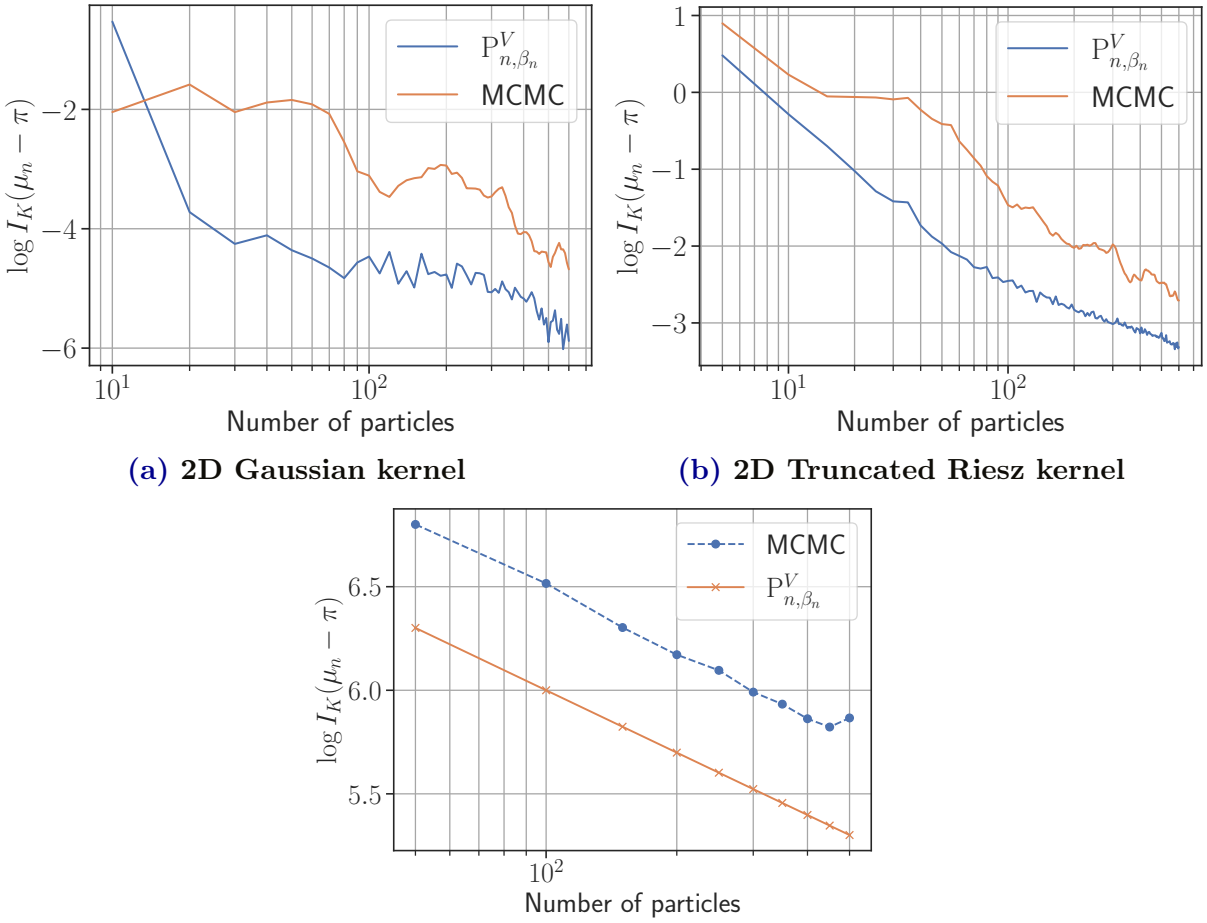
4.4.3 Worst-case integration errors

In this experiment, we take for π the uniform measure on the unit ball of $\mathbb{R}^d$, with $d = 3$. We compare, for various values n of the number of quadrature nodes, the worst-case integration error I_K of (i) the empirical measure μ_n^{MCMC} of an MH chain of length n targeting π, with an isotropic Gaussian proposal with variance $0.05 I_d$, and of (ii) the empirical measure μ_n of an approximate sample of $\mathbb{P}^{V^\pi}_{n,\beta_n}$ with $\beta_n = n^2$. For the latter, we run MALA for $T = 5000$ iterations. We consider two interaction kernels, the Gaussian kernel and the truncated Riesz kernel,

$$K_1(x,y) = \exp\left(-|x-y|^2/2\right), \text{ and } K_2(x,y) = \left(|x-y|^2 + 0.1^2\right)^{-(d-2)/2}. \tag{4.44}$$

For $K \in \{K_1, K_2\}$, U^π_K is again replaced by

$$M^{-1} \sum_{i=1}^{M} K(\cdot, z_i), \tag{4.45}$$



(a) 2D Gaussian kernel

(b) 2D Truncated Riesz kernel

(c) 10D Truncated Riesz kernel

Fig. 4.7: Energy comparisons between $\mathbb{P}^{V\pi}_{n,\beta_n}$ and MCMC samples.

where $z_1, \dots, z_M$ are a sample from an MH chain of length $M = 1\,000$ targeting π, independent from any other sample.

Figure 4.7a and 4.7b show the results for K_1 and K_2, respectively. For each value of n, we plot an independent approximation of $I_K(\mu - \pi)$ for $\mu \in \{\mu_n^{\mathrm{MCMC}}, \mu_n\}$ the empirical measure of either the baseline MH chain or an approximate draw of our Gibbs measure. Again, the energy $I_K(\mu_n - \pi)$ is approximated with large MCMC samples to estimate the integrals with respect to π. These experiments where obtained using GPU running time on Google Colab, with 12 GB of RAM.

We observe that $I_K(\cdot - \pi)$ decays at the same rate under the approximation of $\mathbb{P}^V_{n,\beta_n}$ as for MCMC samples. This was expected, since our concentration bound (4.3.8) only recovers the n^{-1} rate for the energy, the improvement rather being on confidence regions at fixed worst-case integration precision level. We see nonetheless that the approximated energy (and hence, the worst-case integration error) is consistently smaller by about

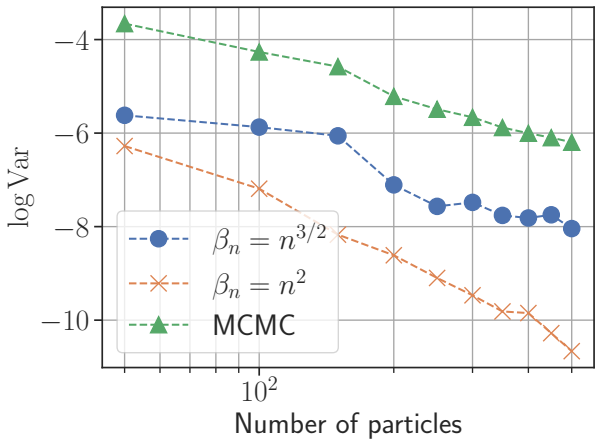


Fig. 4.8: Variance comparison between $\mathbb{P}^{V^\pi}_{n,\beta_n}$ and MCMC samples.

a factor 3 under $\mathbb{P}^V_{n,\beta_n}$, which is also expected since $\mathbb{P}^V_{n,\beta_n}$ favors small values of I_K by definition.

In Figure 4.7c, we consider the same setup with truncated Riesz kernel, but we set π to be a truncated Gaussian distribution on the unit ball in dimension $d = 10$ and keep $\beta_n = n^2$. We observe the same results as in Figure 4.7b: the energy decay rate is the same as MCMC, but our approach improves on the MMD by about a factor 3, even in moderate dimension ($d = 10$) and with a relatively low number of MALA iterations ($T = 5\,000$).

4.4.4 Variance of linear statistics

We know from classical CLT arguments that when $x_1, \dots, x_n$ are drawn from an MCMC chain targeting π, the variance of $n^{-1}\sum_{i=1}^n f(x_i)$ scales like n^{-1}, under appropriate assumptions on f (2.73). While this is not at all implied by our concentration bound (Corollary 4.3.11), analogies with the statistical physics literature (see Section 3.3.3) make us expect a CLT to hold for our Gibbs measure, at rate ${\beta_n}^{-1/2}$, at least for some temperature schedules (β_n) and smooth enough integrands. Such a result would imply asymptotic confidence intervals for single integrands of width decreasing like $\beta_n^{-1/2}$, a faster decay than standard Monte Carlo.

To assess whether this expectation is reasonable, we consider again the setting where π is uniform on the unit ball in $d = 3$, the kernel is the truncated Riesz kernel K_2, and the integrand is $f : x \longrightarrow K(x, 0)$, which naturally belongs to $\mathcal{H}_K$. For each value of n in Figure 4.8, we run 100 independent MH chains of length n targeting π, and plot the empirical variance of the 100 ergodic averages. Similarly, we run 100 independent MALA chains targeting $\mathbb{P}^{V^\pi}_{n,\beta_n}$, for $T = 5000$ iterations each, and plot the empirical variance of our estimator across these 100 approximate samples. Again, $U^\pi_{K_2}$ is approximated through long MH chains. These experiments were obtained using GPU running time on Google Colab, with 12 GB of RAM.

We observe in Figure 4.8 that the variance is noticeably smaller under the Gibbs measure, for both temperature schedules. Moreover, the rate of decay appears faster, at least in the "usual" temperature schedule $\beta_n = n^2$. It is hard to be more quantitative, as the fact we use MALA, and with a fixed number of iterations across all values of n, may impact the convergence we see here. Still, the experiment supports our belief that a fast CLT holds. We recall that obtaining such a fast CLT e.g. for integrands in the RKHS $\mathcal{H}_K$ is our long term goal, and we will go back to this in Section 6.1.2.

4.5 Proofs

4.5.1 Proof of Proposition 4.3.5

We first recall the result.

Proposition 4.5.1 Let K satisfy Assumptions 4.3.1 and 4.3.2, and V satisfy Assumption 4.3.4. Then

1. I_K^V is lower semi-continuous, has compact level sets and
$$I_K^V(\mu) > -\infty \tag{4.46}$$
for any probability distribution μ on $\mathbb{R}^d$.
2. If $\mu \in \mathcal{E}_K^V$, then $I_K(\mu)$ and $\int |V| \, \mathrm{d}\mu$ are finite, and
$$I_K^V(\mu) = \frac{1}{2} I_K(\mu) + \int V \, \mathrm{d}\mu. \tag{4.47}$$
3. I_K^V is strictly convex on the convex non-empty set $\mathcal{E}_K^V$.
4. I_K^V has a unique minimizer μ_V over the set of probability measures on $\mathbb{R}^d$, called the *equilibrium measure*, and the support of μ_V is compact.

The proof is a simple application of known results.

1. This is a consequence of the first point of Theorem 1.1 in (Chafaï et al., 2014);
2. This is given by the second point of Lemma 2.2 in (Chafaï et al., 2014);
3. This is a consequence of the ISPD assumption on the kernel, using Lemma 3.1 in (Pronzato et al., 2020). To see that $\mathcal{E}_K^V$ is non-empty, we can simply consider Dirac measures δ_x, since K is bounded and V is finite everywhere.
4. This is a consequence of points 1 and 3 using the arguments of Section 4.1 of (Chafaï et al., 2014).

4.5.2 Proof of Proposition 4.3.10

The proof of Proposition 4.3.10 relies on the so-called Euler-Lagrange equations, which we recall here.

Lemma 4.5.2 Let K be a kernel satisfying Assumptions 4.3.1 and 4.3.2, and V be an external potential satisfying Assumption 4.3.4. Set

$$C_V = I_K(\mu_V) + \int V \, \mathrm{d}\mu_V. \tag{4.48}$$

Then the equilibrium measure μ_V has compact support, and a probability measure ν satisfies $\nu = \mu_V$ if and only if ν has compact support and there exists a constant C such that

(i) $U_K^\nu(z) + V(z) \geq C$ for all $z \in \mathbb{R}^d$;

(ii) $U_K^\nu(z) + V(z) \leq C$ for all $z \in supp(\nu)$.

In that case, $C = C_V$.

Proof. Let us first check that μ_V satisfies this characterization. We already know from Proposition 4.3.5 that μ_V exists, is unique, and has compact support. We can use the same procedure as (Chafaï et al., 2014, Theorem 1.2, proof of item 5) : considering the directional derivative of I_K^V and using the fact that μ_V is the minimizer, their equation (4.5) yields that, for any probability distribution $\nu \in \mathcal{E}_K^V$,

$$\int (V + U_K^{\mu_V} - C_V) \, \mathrm{d}\nu \geq 0. \tag{4.49}$$

Since Dirac measures δ_z have finite interaction energy $I_K^V(\delta_z)$, we get Point (i) by taking $\nu = \delta_z$. The second point is obtained exactly as in the second part of the proof of item 5 of Theorem 1.2 of (Chafaï et al., 2014). Finally, the converse implication can be similarly obtained along the lines of the proof of item 6 of Theorem 1.2 in (Chafaï et al., 2014), using the strict convexity of the energy functional in Proposition 4.3.5. ■

We are now ready to prove Proposition 4.3.10, by checking that the Euler-Lagrange equations are satisfied for π and V^π, and that V^π satisfies the assumptions of Proposition 4.3.5.

First note that since the kernel K is nonnegative and bounded on the diagonal by assumption, Cauchy-Schwarz in the RKHS $\mathcal{H}_K$ implies that $K(x, y) \leq C$ for all $x, y \in \mathbb{R}^d$. As a consequence, since K is further assumed to be continuous, Lebesgue's dominated convergence theorem yields that $z \longmapsto U_K^\pi(z)$ is finite everywhere and continuous. In particular, V^π is continuous.

Moreover, the bound on K induces $0 \leq U_K^\pi(z) \leq C$ for any $z \in \mathbb{R}^d$, so that $V^\pi(z) \to +\infty$ when $|z| \longrightarrow +\infty$. Finally, the integrability assumption of Assumption 4.3.4 is satisfied

with $c = 1$ by our assumption on Φ and because U_K^π is bounded by C. Hence, V^π satisfies Assumption 4.3.4 and the equilibrium measure μ_{V^π} is well-defined. We conclude upon noting that, since $\Phi \geq 0$, Proposition 4.5.2 yields that $\mu_{V^\pi} = \pi$.

4.5.3 Proof of Theorem 4.3.7

The proof will follow the one of (Chafaï et al., 2018), with some notable simplifications. We first compute a lower bound on the partition function Z^V_{n,β_n}, which generalizes the one of (Chafaï et al., 2018) to bounded kernels.

Proposition 4.5.3 Let K be a kernel satisfying Assumptions 4.3.1 and 4.3.2, and V be an external potential satisfying Assumption 4.3.4. Assume that the associated equilibrium measure μ_V has finite entropy, i.e.

$$S(\mu_V) = -\int \log \mu_V' \mathrm{d}\mu_V < \infty, \tag{4.50}$$

where μ_V' is the density of μ_V w.r.t. the Lebesgue measure. Then for $n \geq 2$, we have

$$Z^V_{n,\beta_n} \geq \exp\left\{-\beta_n I_K^V(\mu_V) + n\left(\frac{\beta_n}{2n^2} I_K(\mu_V) + S(\mu_V)\right)\right\}. \tag{4.51}$$

Proof. The idea is to rephrase a bit the discrete energy H_n and use Jensen's inequality.

We let $X_n = (x_1, \ldots, x_n)$ for brevity, and we start by writing

$$n^2 H_n(X_n) = \frac{1}{2}\sum_{i\neq j}\{K(x_i, x_j) + V(x_i) + V(x_j)\} + \sum_{i=1}^n V(x_i). \tag{4.52}$$

Then

$$\begin{aligned}
\log Z^V_{n,\beta_n} &= \log \int_{\left(\mathbb{R}^d\right)^n} \exp\left(-\beta_n H_n(X_n)\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n \\
&\geq \log \int_{E_V^n} \exp\left(-\frac{\beta_n}{2n^2}\sum_{i\neq j}\{K(x_i, x_j) + V(x_i) + V(x_j)\}\right) \\
&\times \exp\left(-\sum_{i=1}^n \left(\frac{\beta_n}{n^2} V(x_i) + \log \mu_V'(x_i)\right)\right) \mathrm{d}\mu_V(x_1) \ldots \mathrm{d}\mu_V(x_n),
\end{aligned} \tag{4.53}$$

where

$$E_V^n = \{(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n : \prod_{i=1}^n \mu_V'(x_i) > 0\}. \tag{4.54}$$

Using Jensen's inequality, we get

$$\begin{aligned}\log Z^V_{n,\beta_n} &\geq -\frac{\beta_n}{n^2}\sum_{i\neq j}\frac{1}{2}\int_{E_V^n}\left(K(x_i,x_j)+V(x_i)+V(x_j)\right)\mathrm{d}\mu_V(x_1)\ldots\mathrm{d}\mu_V(x_n)\\ &\quad-\sum_{i=1}^{n}\int_{E_V^n}\left(\frac{\beta_n}{n^2}V(x_i)+\log\mu_V'(x_i)\right)\mathrm{d}\mu_V(x_1)\ldots\mathrm{d}\mu_V(x_n)\\ &=-\frac{\beta_n}{n^2}n(n-1)I_K^V(\mu_V)-\frac{\beta_n}{n}\int V\mathrm{d}\mu_V+nS(\mu_V).\end{aligned}\tag{4.55}$$

Using the definition

$$I_K^V(\mu_V)=\frac{1}{2}I_K(\mu_V)+\int V\mathrm{d}\mu_V,\tag{4.56}$$

we get the result. ■

To bound $I_K(\mu_n-\mu_V)$, we shall further use the following lemma, which is again inspired by (Chafaï et al., 2018).

Lemma 4.5.4 Let K and V be a kernel and an external potential satisfying Assumptions 4.3.1 to 4.3.4. Let μ be any probability measure of finite energy, $I_K^V(\mu)<+\infty$. Then

$$I_K(\mu-\mu_V)\leq 2\left(I_K^V(\mu)-I_K^V(\mu_V)\right).\tag{4.57}$$

Proof. We write

$$I_K(\mu-\mu_V)=\iint K(x,y)\,\mathrm{d}(\mu-\mu_V)^{\otimes 2}=I_K(\mu)-2I_K(\mu,\mu_V)+I_K(\mu_V).\tag{4.58}$$

With the notation of Proposition 4.5.2, we know that $U_K^{\mu_V}(z)+V(z)=C_V$ for all z in the support of μ_V, and that $U_K^{\mu_V}(z)+V(z)\geq C_V$ in general. Using Fubini, we thus see that

$$\begin{aligned}I_K(\mu,\mu_V)+\int V\,\mathrm{d}\mu&=\int\left\{U_K^{\mu_V}(z)+V(z)\right\}\mathrm{d}\mu(z)\\ &\geq C_V=I_K(\mu_V)+\int V\,\mathrm{d}\mu_V.\end{aligned}\tag{4.59}$$

Plugging this into (4.58) yields the result. ■

We are now ready to prove Theorem 4.3.7. Recall that we work under the assumption $\beta_n\geq 2cn$ where c is the constant of Assumption 4.3.4. Consider a Borel set $A\subset\left(\mathbb{R}^d\right)^n$. For brevity, recall that for $(x_k)_{k\leq n}\in A$, we write $\mu_n=\frac{1}{n}\sum_{i=1}^n\delta_{x_i}$ and $X_n=(x_1,\ldots,x_n)$.

Let $\eta > 0$ and $X_n \in A$. The key idea is to split

$$H_n(X_n) = (1-\eta)H_n(X_n) + \eta H_n(X_n). \tag{4.60}$$

The first part will be compared with the energy $I_K^V(\mu_n)$, while the second part will be kept to ensure integrability. Remembering that

$$n^2 H_n(X_n) = n^2 I_K^V(\mu_n) - \frac{1}{2}\sum_{i=1}^n K(x_i, x_i), \tag{4.61}$$

we write, by definition,

$$\begin{aligned}
\mathbb{P}^V_{n,\beta_n}(A) &= \frac{1}{Z^V_{n,\beta_n}} \int_A \exp\left(-\beta_n H_n(X_n)\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n \\
&= \frac{1}{Z^V_{n,\beta_n}} \exp\left(-\beta_n(1-\eta) I_K^V(\mu_V)\right) \int_A \exp\left(-\beta_n(1-\eta)\left(I_K^V(\mu_n) - I_K^V(\mu_V)\right)\right) \\
&\quad \times \exp\left(\frac{\beta_n}{2n^2}(1-\eta)\sum_{i=1}^n K(x_i,x_i) - \beta_n \eta H_n(X_n)\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n.
\end{aligned} \tag{4.62}$$

Using Proposition 4.5.3, we continue

$$\begin{aligned}
\mathbb{P}^V_{n,\beta_n}(A) &\le \exp\left(-n\left(S(\mu_V) + \frac{\beta_n}{2n^2} I_K(\mu_V)\right) + \beta_n \eta I_K^V(\mu_V)\right) \\
&\quad \times \exp\left(-\beta_n(1-\eta)\inf_A\left(I_K^V(\mu_n) - I_K^V(\mu_V)\right)\right) \\
&\quad \times \int_{\mathbb{R}^d} \exp\left(\frac{\beta_n}{2n^2}(1-\eta)\sum_{i=1}^n K(x_i,x_i) - \beta_n\eta H_n(X_n)\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n.
\end{aligned} \tag{4.63}$$

As noted in the proof of Proposition 4.5.2, $0 \le K \le C$, so that the last integral in (4.63) is easily bounded,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \exp\left(\frac{\beta_n}{2n^2}(1-\eta)\sum_{i=1}^n K(x_i,x_i) - \beta_n\eta H_n(X_n)\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n \\
&\quad \le \exp\left(\frac{\beta_n}{2n}(1-\eta)C\right)\left(\int_{\mathbb{R}^d} \exp\left(-\frac{\beta_n}{n}\eta V(x)\right)\mathrm{d}x\right)^n.
\end{aligned} \tag{4.64}$$

Now we choose a particular value for η, namely $\eta = cn/\beta_n$ where c is the constant of Assumption 4.3.4. Further let

$$C_2 = \log \int_{\mathbb{R}^d} \exp\left(-cV(x)\right) \mathrm{d}x, \tag{4.65}$$

which is finite by assumption. Setting $c_1 = cI_K^V(\mu_V) + C_2 - S(\mu_V)$ and $c_2 = \frac{1}{2}C - \frac{1}{2}I_K(\mu_V)$, (4.63) yields

$$\mathbb{P}^V_{n,\beta_n}(A) \le \exp\left(-(\beta_n - cn)\inf_A\left(I_K^V(\mu_n) - I_K^V(\mu_V)\right) + nc_1 + \frac{\beta_n}{n}c_2\right). \tag{4.66}$$

Note that c_1 and c_2 are indeed finite by definition of μ_V, since $\mathcal{E}_K^V$ is non-empty. The comparison inequality of Lemma 4.5.4 applied to (4.66), with

$$A = \{I_K(\mu_n - \mu_V) > r^2; \mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}\}, \tag{4.67}$$

yields Theorem 4.3.7, upon noting that $\beta_n - cn \geq \beta_n/2$ by assumption.

4.5.4 Proof of Proposition 4.3.13

Proof. We use the fact that $0 \leq K(x,y) \leq C$ for all $x, y \in \mathbb{R}^d$ to get $I_K(\pi) \geq 0$ and $c_2 \leq C/2$. Moreover, using the definition of V^π in Proposition 4.3.10, we have the bound

$$\int_{\mathbb{R}^d} \exp\left(-V^\pi(z)\right) \mathrm{d}x \leq \int_{|x|<R} \exp(C)\mathrm{d}x + \int_{|x|>R} \exp(C)\exp\left(-|x|^2 + R^2\right) \mathrm{d}x, \tag{4.68}$$

since $U_K^\pi(x) \leq C$ for all $x \in \mathbb{R}^d$. For the first term, we use the fact that the volume of $B(0,R)$ is given by

$$|B(0,R)| = \frac{\pi^{d/2}}{\Gamma\left(d/2+1\right)} R^d. \tag{4.69}$$

For the second term, we bound the tail of the Gaussian distribution

$$\int_{|x|>R} \exp\left(-|x|^2\right) \mathrm{d}x \leq \int \exp\left(-|x|^2\right) \mathrm{d}x \leq \left(\sqrt{2}\pi\right)^{d/2}. \tag{4.70}$$

This yields

$$c_1 \leq -S(\pi) + 2C + d\left(\frac{(1+\sqrt{2})\log\pi}{2} + \log R\right) - \log\Gamma\left(d/2+1\right) + R^2. \tag{4.71}$$

■

4.6 Conclusion

In this chapter, we provided a fast concentration inequality for Gibbs measures with general but bounded interaction kernel K, with a fairly easy proof, at least compared to the classical results in statistical physics that inspired us. Moreover, this Gibbs measure can be tuned to favor point configurations with low MMD, i.e. worst-case integration error in the RKHS of kernel K. Drawing points from this joint distribution thus allows for tighter confidence regions at fixed precision level $\epsilon > 0$ than crude Monte Carlo, while preserving the same $n^{-1/2}$ rate for the MMD. We illustrated this improvement on toy examples in moderate dimension, and compared with state-of-the-art methods such as kernel thinning.

A strong additional argument in favor of a Gibbs measure would be a CLT with a fast rate. We showed experimental evidence that supports this expectation, and will come back to this in Chapter 6.

Two limitations of our approach that deserve further inquiry are the impact of using an approximation of the kernel embedding U_K^π and an MCMC sampler, here MALA. Understanding how the accuracy of the MALA approximation relates to n and β_n would help find the right trade-off between statistical accuracy and computational cost. Moreover, a careful experimental comparison at fixed budget with e.g. kernel thinning and volume sampling (Section 2.6.3) would be interesting. Finally, integrating a tractable approximation of U_K^π without loss on the guarantees would be an important improvement, and this is the topic of the next chapter.

Approximation of the kernel embedding



“ — Ouf !... Le danger est passé !... Je respire !... Je puis vous avouer maintenant que j’ai eu bien peur...

— Nous aurions été vraiment réduits en miettes ?...

— Si ce n’est que ça !... Mais c’est beaucoup plus grave !... Oui, je puis vous le dire à présent : si mes prévisions ne s’étaient pas réalisées, j’aurais dû recommencer tous mes calculs !...”

Les aventures de Tintin: On a marché sur la Lune, Hergé

5.1 Introduction 115

5.2 Related work 117

5.2.1 Monte Carlo integration and KSDs 117

5.2.2 Large deviations 117

5.2.3 Extrema of the potential of a Coulomb gas 118

5.3 Definitions and main results 119

5.4 Experiments 127

5.4.1 Approximately sampling from the Gibbs measure 128

5.4.2 Bayesian classification 130

5.4.3 Worst-case integration error and variance . . 132

5.5 Proofs: quenched large deviations 135

5.5.1 Reminder on LDPs for Gibbs measures, after (García-Zelada, 2019) 136

5.5.2 Proof of Theorem 5.3.7 (non-singular kernels) 139

5.5.3 Proof of Theorem 5.3.13 (Coulomb kernel) 140

5.6 Proofs: uniform convergence for the Coulomb potential field 141

5.6.1 Proof of Proposition 5.3.14 142

5.6.2 Proof of Proposition 5.3.15 144

5.7 Proof of auxiliary lemmas 148

5.7.1 Preliminaries 148

5.7.2 Auxiliary lemmas for Proposition 5.3.14 . . . 151

5.7.3 Auxiliary proofs for Proposition 5.3.15 155

5.8 Conclusion 157

5.1 Introduction

The content of this chapter is based on the following papers

- Bardenet, Rémi, Maïda, Mylène, and Rouault, Martin (Aug. 2, 2025). *Quenched large deviations for Monte Carlo integration with Coulomb gases*. arXiv: 2508.01392[cs].
- Rouault, Martin, Bardenet, Rémi, and Maïda, Mylène (Aug. 25, 2025). "Intégrer plusieurs fonctions simultanément avec des mesures de Gibbs". In: *30° Colloque sur le traitement du signal et des images*. Issue: 2025-1751. Strasbourg: GRETSI - Groupe de Recherche en Traitement du Signal et des Images, p. 1157–1160.

We recall that we are interested in the statistical properties of the Gibbs measure defined by

$$\mathbb{P}^V_{n,\beta_n}(y_1,\dots,y_n) = \frac{1}{Z^V_{n,\beta_n}} \exp\left(-\beta_n H_n(y_1,\dots,y_n)\right) \mathrm{d}y_1 \dots \mathrm{d}y_n, \tag{5.1}$$

with

$$H_n(y_1,\dots,y_n) = \frac{1}{2n^2}\sum_{i\neq j} K(y_i,y_j) + \frac{1}{n}\sum_{i=1}^n V(y_i), \tag{5.2}$$

and

$$V = V^\pi(z) = \begin{cases} -U^\pi_K(z) \text{ when } z \in B(0,R) \\ -U^\pi_K(z) + \Phi(z) \text{ otherwise} \end{cases}, \tag{5.3}$$

where

$$U^\pi_K(z) = \int K(z,y)\,\mathrm{d}\pi(y). \tag{5.4}$$

Computing the potential (5.3) requires evaluating U^π_K in (5.4), which requires integrating w.r.t. π, seemingly defeating our initial purpose of numerical integration. In this chapter, we address this issue, completely removing the need to evaluate U^π_K while preserving the better-than-Monte-Carlo integration guarantees of Gibbs measures.

Formally, let

$$\nu_n = \sum_{i=1}^{M_n} w_i \delta_{x_i} \tag{5.5}$$

be a random M_n-atom measure on $\mathbb{R}^d$, called the *background*. Conditionally on ν_n, let $y_1,\dots,y_n$ be drawn from the Gibbs measure (5.1) with confining potential

$$V^{\nu_n}(z) = -U^{\nu_n}_K(z) + \Phi(z) = -\sum_{i=1}^{M_n} w_i K(z,x_i) + \Phi(z), \tag{5.6}$$

that is (5.3) where we replace the target measure π by the atomic measure ν_n. Following the vocabulary of random environment models in probability, we call the conditional distribution $\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}$ of $y_1, \dots, y_n$ given ν_n a *quenched* Gibbs measure. In this chapter, we show a large deviation principles at speed β_n for this quenched Gibbs measure.

Seeing Gibbs measures as an avatar of MMD minimization, we believe that, outside thinning algorithms, ours is the first work that formally includes the approximation of the kernel embedding U^π_K in the analysis of its quadrature, for general kernel K. We hope that this can inspire the analysis of non-Gibbs-based MMD minimization algorithms.

Our first result deals with a bounded interaction kernel K in the Gibbs measure. We show in Theorem 5.3.7 that, on any event such that ν_n converges to π, the distribution $\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}$ of the quadrature $y_1, \dots, y_n$ almost surely satisfies a LDP at speed β_n (Definition 3.1.6), just like the Gibbs measure with intractable potential (5.3). Consequently, we avoid the evaluation of U^π_K – an issue of most MMD minimization methods– at no statistical cost on the LDP. Note that it is computationally easy to build a ν_n that converges almost surely to π using standard Monte Carlo techniques, such as self-normalized importance sampling.

Our second result covers the more difficult case of the (singular) Coulomb kernel $K = g$ in (3.43). This time, we show in Theorem 5.3.13 that the quenched Gibbs measure almost surely satisfies a LDP at speed β_n if the atoms of ν_n are themselves n points drawn from a Gibbs measure with Coulomb interaction, under weak requirements on the equilibrium measure, and with importance weights compensating for the mismatch between that equilibrium measure and the target π. Here again, we obtain the same statistical guarantee on the worst-case integration error without ever needing to evaluate the intractable potential U^π_K.

We provide *en route* a control on the uniform convergence of the approximation of the potential. This control is key to deal with the singularity of the Coulomb interaction. For a well-chosen regularization K_ζ of the Coulomb kernel, we show in Proposition 5.3.14 and Proposition 5.3.15 that

$$\sup_{z \in \mathbb{R}^d} \left| U^{\nu_n}_{K_\zeta}(z) - U^\pi_g(z) \right| \to 0, \tag{5.7}$$

almost surely as $n \to +\infty$. Note in particular that the regularization is indeed required to keep this supremum finite. We believe this intermediate result is of independent interest since the extrema of the potential field generated by a Coulomb gas have recently drawn attention, with exact scalings obtained at first order in dimension 2 (G. Lambert et al., 2024; Peilen, 2024), more on this below.

In Section 5.2, we quickly recall existing results. Then, we introduce our main results in Section 5.3 and turn to numerical experiments in Section 5.4. The last sections of the chapter are devoted to proofs.

5.2 Related work

5.2.1 Monte Carlo integration and KSDs

Computing $U_K^\pi(z)$ has been up to our knowledge one of main limitations of MMD minimization algorithms (Section 2.6). One standard way to bypass this computation is to rely on a large MCMC chain targeting π (Pronzato et al., 2020), which is fair if the computational bottleneck is the number n of evaluations of the test function f. This can be the case for instance in computational biology when each evaluation of the likelihood of the model can range from a few minutes in cardiac electrophysiology models (Johnstone et al., 2016) to a few hours for population models in ecology (Purves et al., 2013), see again Chapter 1. Yet, this usually comes at the price of losing theoretical integration guarantees for Monte Carlo integration with respect to π, expect in the case of thinning algorithms. For instance, in kernel thinning (Section 2.6.4), the fast MMD bounds in Theorem 2.6.3 for n integration nodes are obtained upon assuming that we have access to n^2 samples from an ergodic Markov chain targeting π

Kernel Stein Discrepancies (KSDs, Anastasiou et al., 2022) have been introduced as a way to bypass this computation, as presented in Section 2.6.2. Indeed, given a kernel K and a target measure π, one can define a new kernel K_π with closed form such that $U_{K_\pi}^\pi(z) = 0$ for all z under a mild regularity assumption on K when π vanishes. The latter can always be ensured even when π does not vanish up to multiplying the kernel K by a vanishing weight function at the boundary of the integration domain (Oates et al., 2017a). The kernel Stein discrepancy is then defined as the square root of the MMD in (2.97) with kernel K_π. Yet, replacing K by K_π replaces the set of test functions $\mathcal{H}_K$ by $\mathcal{H}_{K_\pi}$, and there is little knowledge about the kind of functions contained in $\mathcal{H}_{K_\pi}$. There is evidence that the quadrature is influenced in that case by pathological test functions that oscillate wildly in regions where π puts little mass (Bénard et al., 2023). Still, much effort has been put up to analyze the properties of the KSD (Gorham et al., 2020; Kanagawa et al., 2025), and it can for instance be shown that, under some assumptions on K and π, $\mathcal{H}_{K_\pi}$ is still large enough to imply convergence in Wassertein or bounded Lipschitz distance. Since our goal is to give integration guarantees in the original $\mathcal{H}_K$ (see Section 2.7), we stick in this thesis to approximating U_K^π.

5.2.2 Large deviations

Large deviations for Gibbs measures with pairwise interactions as in (5.1) have been thoroughly studied by Chafaï et al. (2014) and García-Zelada (2019).

In particular, combining Proposition 4.3.10 with (Chafaï et al., 2014, Theorem 1.1) establishes that for a very general class of bounded interaction kernels K, for any $r > 0$, there exists $n_0(r)$ such that for any $n \geq n_0(r)$, the worst-case integration error on the unit ball of the RKHS $\mathcal{H}_K$ is small with overwhelming probability, i.e.

$$\mathbb{P}_{n,\beta_n}^{V^\pi}\left(\mathrm{MMD}_K\left(\frac{1}{n}\sum_{i=1}^n \delta_{y_i}, \pi\right) > r\right) \leq \exp\left(-\frac{1}{2}\beta_n r^2\right). \tag{5.8}$$

Similarly, taking for K the (unbounded) Coulomb interaction (3.43), combining Theorem 3.3.3 and Corollary 3.3.4 with (Chafaï et al., 2018, Theorem 1.2) yields

$$\mathbb{P}_{n,\beta_n}^{V^\pi}\left(\mathrm{d_{BL}}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{y_i},\pi\right)>r\right)\leq\exp\left(-c\beta_n r^2\right), \tag{5.9}$$

for some constant $c>0$, where $\mathrm{d_{BL}}$ is the bounded Lipschitz distance (2.11). The results (5.8) and (5.9) are to compare to classical results for i.i.d. quadrature nodes, such as Sanov's theorem (Theorem 2.2.1), where β_n in the exponential is replaced by n. This fast convergence when β_n grows faster than n motivated our study of numerical quadratures based on the Gibbs measure $\mathbb{P}_{n,\beta_n}^{V^\pi}$ in Chapter 4.

In our work, we need to further include a random environment, i.e. an empirical approximation to the potential. Actually, García-Zelada (2019, Theorem 4.1 and Theorem 4.2) gives a list of sufficient conditions to derive large deviations principles in varying environments, with applications to Gibbs measures. Yet, the latter applications deal with Gibbs measures for which we cannot pre-specify the equilibrium measure, an important requirement for Monte Carlo integration where the equilibrium measure should be (at least an approximation of) the target integration measure π. Our proofs thus consist in checking the sufficient conditions of (García-Zelada, 2019) for the potential (5.6). Checking these conditions in our setting is however no sinecure, especially in the case of the Coulomb interaction kernel.

5.2.3 Extrema of the potential of a Coulomb gas

In dimension 1 and 2, the two-dimensional Coulomb interaction kernel

$$g(x,y)=-\log|x-y|, \tag{5.10}$$

along with $V(x)=|x|^2/2$ and $\beta_n=2n^2$, yields a Gibbs measure that also describes the distribution of the eigenvalues of random matrices, respectively the Gaussian unitary and Ginibre ensembles, see Section 3.2.3. In this case, the interaction potential $\sum_{i=1}^n g(z,x_i)$ generated at point z by points $x_1,\dots,x_n$ drawn from (5.1) coincides with the logarithm of the modulus of the characteristic polynomial of the random matrix, evaluated in z. The behavior of the maximum of this quantity has been a topic of interest, following the famous conjectures of Fyodorov et al. (2012) in the one-dimensional case. In dimension 2, for general V and $\beta_n=\beta n^2$, it has been for instance shown by G. Lambert (2020), G. Lambert et al. (2024), and Peilen (2024) that

$$\sup_{z\in D(x,r)}\frac{n}{\log n}\left\{\frac{1}{n}\sum_{i=1}^{n}g(z,x_i)-\int g(z,x)\,\mathrm{d}\mu_V(x)\right\}\underset{n\to+\infty}{\to}\frac{1}{\sqrt{\beta}} \tag{5.11}$$

in probability, where μ_V is the unique minimizer of I_K^V in Definition 3.3.2 and $D(x,r)$ is any disk strictly contained in the support of μ_V. Obtaining a similar behavior in dimension $d\geq 3$ is still an open question up to our knowledge, but we show as a

byproduct of our proof in Proposition 5.3.14 that for general V and β_n growing faster than n,

$$\sup_{z\in\mathbb{R}^d}\left|\frac{1}{n}\sum_{i=1}^{n}K_\zeta(z,x_i)-\int g(z,x)\,\mathrm{d}\mu_V(x)\right|\underset{n\to+\infty}{\to}0, \tag{5.12}$$

$\mathbb{P}$-almost surely, where K_ζ is a regularized version of the Coulomb interaction g (3.43). Obtaining the exact scaling for the leading order as in (5.11) requires fine estimates on the fluctuations of the system at the second order, a hot topic in the field for $d \geq 3$ (Serfaty, 2023).

5.3 Definitions and main results

We fix once and for all a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that we will use throughout the chapter. We will again extensively use the energies defined in Definition 3.3.2. We will work under the following assumptions on π and K, which as similar to the ones used in Chapter 4.

Let $d \geq 1$ and π be a probability measure on $\mathbb{R}^d$, which we call again the *target* distribution.

Assumption 5.3.1 The support S_π of π is compact. In particular, there exists $R > 0$ such that $S_\pi \subset B(0, R)$. Furthermore, we assume that π has a continuous density π' on the interior of S_π and that it is bounded on S_π.

Throughout the chapter, some particular kernels will play an important role.

Definition 5.3.2 For $x, y \in \mathbb{R}^d$, when $d \geq 3$, the Coulomb interaction kernel (3.43) is given by

$$g(x,y)=\frac{1}{|x-y|^{d-2}}. \tag{5.13}$$

Assumption 5.3.3 — Non-singular kernel. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a non-negative continuous, symmetric interaction kernel, such that $C := \sup_x K(x,x) < +\infty$. Furthermore, K is integrally strictly positive definite; i.e. $I_K(\mu-\nu) > 0$ for any finite signed Borel measures μ and ν.

For a kernel K satisfying Assumption 5.3.3, the interaction energy I_K between two distributions coincides with the squared worst-case integration error for test functions in the reproducing kernel Hilbert space $\mathcal{H}_K$ of kernel K, see Proposition 4.3.3. This is for instance true for Gaussian (2.24) or Matérn (2.26) kernels, or for the regularized Riesz

kernels

$$K_{s,\epsilon}(x,y) = \frac{1}{(|x-y|^2+\epsilon^2)^{s/2}}. \tag{5.14}$$

In particular, for $n \geq 1$ and $\zeta > 0$, we will also use the following non-singular regularization of the Coulomb interaction g in (5.13),

$$K_\zeta(x,y) = \frac{1}{(|x-y|^2+n^{-2\zeta})^{\frac{d-2}{2}}}. \tag{5.15}$$

When dealing with the Coulomb interaction (5.13), we will also need to consider Gibbs measures (3.45) with V differing from (5.6) to design our two-step approximation. We recall that, as in Chapter 4, we then work under the following assumption that makes the Gibbs measure (3.45) well-defined.

Assumption 5.3.4 Let $V : \mathbb{R}^d \mapsto \mathbb{R}$ be a continuous function such that $V(x) \to +\infty$ as $|x| \to +\infty$. Without loss of generality, we will always assume that $V \geq 0$. Furthermore, we assume that there exists $c' > 0$ such that

$$\int \exp\left(-c'V(x)\right) \mathrm{d}x < +\infty. \tag{5.16}$$

Under Assumption 5.3.4, we know from (Chafaï et al., 2014) that the energy I_K^V has a unique minimizer over the set of probability distributions, called the equilibrium measure μ_V. Furthermore, μ_V then has finite energy and compact support. Outside the intractable case given by V^π in (5.3), the expression of μ_V is not explicit in general. Yet, when $K = g$, a few choices of potential V do lead to explicit expressions. For instance, for $V(x) = \frac{d-2}{2R^d}|x|^2$, μ_V is the uniform measure on $B(0,R)$, see (Chafaï et al., 2014, Corollary 1.3). This is the main reason why the Coulomb kernel g will play an important role in the upcoming results, along with the fact that the Coulomb interaction enables to bound integration errors for functions with low regularity, see Theorem 3.3.5. Yet, dealing with g involves some additional theoretical difficulties.

We now define our new Gibbs measure based on the approximated confinement V_n (5.6). We call the Gibbs measure with potential V_n the *quenched* Gibbs measure, as it is defined conditionlly on ν_n.

Definition 5.3.5 — Quenched Gibbs measure. Under Assumptions 5.3.1 and 5.3.4, further assume that $\beta_n \gg n$, and let $n \geq 2$. Let ν_n be any random finitely supported measure on $\mathbb{R}^d$. We define a function $V_n : \mathbb{R}^d \to \mathbb{R}$ as follows. When K satisfies Assumption 5.3.3, we let

$$V_n(z) := \begin{cases} -U_K^{\nu_n}(z) \text{ for } z \in B(0,R) \\ -U_K^{\nu_n}(z) + \Phi(z) \text{ otherwise} \end{cases}, \tag{5.17}$$

where $\Phi(z) = [|z|^2 - R^2]_+$. Similarly, when $d \geq 3$ and $K = g$ in (5.13), let $\zeta > 0$ and

$$V_n(z) := \begin{cases} -U^{\nu_n}_{K_\zeta}(z) \text{ for } z \in B(0,R) \\ -U^{\nu_n}_{K_\zeta}(z) + \Phi(z) \text{ otherwise} \end{cases}, \tag{5.18}$$

where K_ζ is the regularized interaction kernel defined in (5.15). We are now ready to define the quenched Gibbs measure as the Gibbs measure with random confinement V_n, i.e.

$$\mathrm{d}\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}(y_1,\dots,y_n;\nu_n) = \mathrm{d}\mathbb{P}^{V_n}_{n,\beta_n}(y_1,\dots,y_n). \tag{5.19}$$

We denote by H^{Q}_n the associated discrete energy i.e.

$$H^{\mathrm{Q}}_n(y_1,\dots,y_n) = \frac{1}{2n^2}\sum_{i\neq j} K(y_i,y_j) + \frac{1}{n}\sum_{i=1}^n V_n(y_i), \tag{5.20}$$

and by $Z^{V_n}_{n,\beta_n}$ the associated partition function. If $y_1,\dots,y_n \sim \mathbb{P}^{\mathrm{Q}}_{n,\beta_n}$, we denote by $\mu^{\mathrm{Q}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ the quenched empirical measure.

Proposition 5.3.6 Under the assumptions of Definition 5.3.5, the quenched Gibbs measure $\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}(\,\cdot\,;\nu_n(\omega))$ is well-defined, for $\mathbb{P}$-almost all ω.

Proof. It is enough to show that V_n satisfies Assumption 5.3.4, $\mathbb{P}$-almost surely, since this implies that $Z^{V_n}_{n,\beta_n}$ is positive and finite $\mathbb{P}$-almost surely; see (Chafaï et al., 2014). In both cases of Equation (5.17) and Equation (5.18), V_n is clearly continuous and satisfies $V_n(z) \to +\infty$ as $|z| \to +\infty$, since K is either bounded or decays to 0. Moreover, as the kernel embedding term $U^{\nu_n}_K$ is continuous and bounded in each case, we clearly have

$$\int \exp\left(-c' V_n(z)\right) \mathrm{d}z \leq C \int \exp\left(-c'\Phi(z)\right) \mathrm{d}z < +\infty \tag{5.21}$$

for all $c' > 0$, so that Assumption 5.3.4 is satisfied. ■

The atoms $x_1,\dots,x_n$ of ν_n in Definition 5.3.5 act as a random environment (or *background*) that approximates π. Then, the new points $y_1,\dots,y_n \sim \mathbb{P}^{\mathrm{Q}}_{n,\beta_n}$ repel each other, while the random environment acts as a confinement, attracting the nodes $y_1,\dots y_n$ towards regions of high probability under the target distribution π. Crucially, all of this happens without ever needing to evaluate the intractable potential U^π_K in (5.3). Our goal is now to show that for $\mathbb{P}$-*almost-all background realization*, the quenched empirical measure still converges to the target distribution π at the same speed as (5.9) and (5.8). Without further assumption, we can already state our main result for bounded kernels.

Theorem 5.3.7 Assume that π satisfies Assumption 5.3.1, and assume that $\beta_n \gg n$. Let K be a non-singular kernel satisfying Assumption 5.3.3. Let $(\nu_n)_n$ be any family of discrete probability measures on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\nu_n \to \pi$, $\mathbb{P}$-almost surely.

Then for $\mathbb{P}$-almost all ω, the quenched empirical measure $\mu_n^{\mathrm{Q}} = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, where $y_1, \dots, y_n \sim \mathbb{P}^{\mathrm{Q}}_{n,\beta_n}(.\,; \nu_n(\omega))$, satisfies a large deviation principle (Definition 3.1.6) as $n \to +\infty$, at speed β_n and with good rate function

$$\mu \mapsto I_K^{V^\pi}(\mu) - I_K^{V^\pi}(\pi), \tag{5.22}$$

whose minimizer is π. More precisely, for $\mathbb{P}$-almost all ω, for any Borel set $A \subset \mathcal{P}(\mathbb{R}^d)$,

$$\begin{aligned} -\inf_{\mu \in \mathring{A}} \{I_K^{V^\pi}(\mu) - I_K^{V^\pi}(\pi)\} &\leq \varliminf_{n\to+\infty} \frac{1}{\beta_n} \log \mathbb{P}^{\mathrm{Q}}_{n,\beta_n}\left(\mu_n^{\mathrm{Q}} \in A; \nu_n(\omega)\right) \\ &\leq \varlimsup_{n\to+\infty} \frac{1}{\beta_n} \log \mathbb{P}^{\mathrm{Q}}_{n,\beta_n}\left(\mu_n^{\mathrm{Q}} \in A; \nu_n(\omega)\right) \\ &\leq -\inf_{\mu \in \overline{A}} \{I_K^{V^\pi}(\mu) - I_K^{V^\pi}(\pi)\}. \end{aligned} \tag{5.23}$$

As a consequence, for $\mathbb{P}$-almost all ω,

1. μ_n^{Q} converges in law to π,
2. For any $r > 0$ and n large enough,

$$\begin{aligned} \mathbb{P}^{\mathrm{Q}}_{n,\beta_n}\left(\mathrm{MMD}_K\left(\frac{1}{n}\sum_{i=1}^n \delta_{y_i}, \pi\right) > r; \nu_n(\omega)\right) &= \mathbb{P}^{\mathrm{Q}}_{n,\beta_n}\left(\sqrt{I_K(\mu_n^{\mathrm{Q}} - \pi)} > r; \nu_n(\omega)\right) \\ &\leq \exp(-\frac{1}{2}\beta_n r^2), \end{aligned} \tag{5.24}$$

where I_K and $\mathsf{E}_{\mathcal{H}_K}$ are respectively defined in Definition (3.3.2) and (2.19).

The proof of the LDP is in Section 5.5. The first consequence is immediate, and the second one follows from Proposition 4.3.3, along with the fact that

$$MMD_K\left(\mu_n^{\mathrm{Q}}, \pi\right)^2 = I_K(\mu_n^{\mathrm{Q}} - \pi) \leq 2\left(I_K^{V^\pi}(\mu_n^{\mathrm{Q}}) - I_K^{V^\pi}(\pi)\right); \tag{5.25}$$

see Lemma 4.5.4 and Proposition 4.3.3. Taking a step back, Theorem 5.3.7 yields tighter simultaneous confidence intervals for Monte Carlo integration of functions in the unit ball of $\mathcal{H}_K$ than with points drawn i.i.d from π or using MCMC, yet only asymptotically. We stress the fact that we put no condition on the background ν_n, other than the almost sure convergence to π. Our LDP is thus valid even in the case where ν_n is the empirical measure of the history of a suitable MCMC chain up to time n. This justifies a common practical heuristic of replacing the intractable potential by a Monte Carlo estimate obtained using a long MCMC run, as we did in Section 4.4.

As discussed in Chapter 4, from Theorem 5.3.7, one might consider taking an inverse temperature β_n as large as possible in order to maximize the gain in the tail bound. Yet, one still has to produce samples from the quenched Gibbs measure with inverse temperature β_n in order to provide Monte Carlo estimators. Heuristically, this becomes more and more difficult as β_n becomes large as particles drawn from the Gibbs measure then try to get close to the deterministic minimizers of the energy H_n^{Q} in Definition 5.3.5. We thus expect a tradeoff between sampling accuracy or complexity and tail bound guarantees governed by β_n, which we observed on toy examples in Section 4.4.

Motivated by simultaneous confidence intervals for the wider class of bounded Lipschitz functions, we now turn to our second main result, covering the case when $K = g$ is the Coulomb kernel and $d \geq 3$. In that case, to give guarantees as in Theorem 5.3.7, we will need to assume that the background ν_n is itself a realization of a Gibbs measure (3.45) with Coulomb interaction g and known equilibrium measure μ_V. We shall thus need another set of assumptions, always in conjunction with Assumption 5.3.1 that guarantees that π is compactly supported.

Assumption 5.3.8 μ_V has a continuous and lower bounded density $\mu_V' \geq c > 0$ on the support S_π of π.

Assumption 5.3.9 V is $\mathcal{C}^2$ with uniformly bounded second-order derivatives and μ_V has finite classical entropy

$$S(\mu_V) = -\int \log \mu_V \, \mathrm{d}\mu_V. \tag{5.26}$$

Assumption 5.3.10 The density π' of π vanishes on the boundary ∂S_π.

When V is the quadratic potential $V(z) = \frac{d-2}{2R^d}|z|^2$, for instance, Assumptions 5.3.8 and 5.3.9 are satisfied as soon as $S_\pi \subset B(0,R)$ since μ_V is the uniform measure on $B(0,R)$ (Chafaï et al., 2014, Corollary 1.3). We shall also need weights in our Monte Carlo estimator of the potential in the case $K = g$.

Definition 5.3.11 Under Assumptions 5.3.1 and 5.3.8, we define

$$W(x) := \begin{cases} \frac{\pi'(x)}{\mu_V'(x)} \text{ if } x \text{ is in the interior of } S_\pi \\ 0 \text{ otherwise} \end{cases}. \tag{5.27}$$

For $x_1, \ldots, x_n \in \mathbb{R}^d$, we set, for $1 \leq i \leq n$,

$$w(x_i) := \begin{cases} \frac{W(x_i)}{\sum_{i=1}^n W(x_i)} \text{ if there exists } j \text{ such that } \pi'(x_j) > 0 \\ 0 \text{ otherwise.} \end{cases}. \tag{5.28}$$

We have the following convergence for the weighted empirical measures.

Proposition 5.3.12 Under Assumptions 5.3.1, 5.3.8 and 5.3.10, consider any $x_1, \dots, x_n \in \mathbb{R}^d$ such that $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i} \to \mu_V$. Then, for n large enough, there exists j s.t. $\pi'(x_j) > 0$, and

$$\nu_n := \sum_{i=1}^n w(x_i)\delta_{x_i} \tag{5.29}$$

is a well-defined probability measure for n large enough and $\nu_n \to \pi$ as $n \to \infty$.

Proof. W is clearly continuous and compactly supported under those assumptions. The empirical distribution $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ converges in law to μ_V, so that

$$\frac{1}{n}\sum_{i=1}^n W(x_i) \underset{n\to+\infty}{\to} \int W(x)\,\mathrm{d}\mu_V(x) = 1. \tag{5.30}$$

Hence, for n large enough, if $x_1, \dots, x_n \sim \mathbb{P}^V_{n,\beta_n}$, there exists i s.t $W(x_i) > 0$ (or equivalently $\pi'(x_i) > 0$) and ν_n is well-defined.

For any bounded and continuous test function f, we have

$$\int f\,\mathrm{d}\nu_n = \frac{\frac{1}{n}\sum_{i=1}^n f(x_i)W(x_i)}{\frac{1}{n}\sum_{i=1}^n W(x_i)}. \tag{5.31}$$

Since fW is bounded and continuous, Slutsky's lemma then implies that ν_n converges in law to π. ■

Note that since computing w is invariant to a multiplication of π' by a constant, ν_n does not depend on the normalization constant of the target π, and so neither does any linear statistic w.r.t ν_n. In the following theorem, we chose $x_1, \dots, x_n$ as a realization of a Coulomb gas with external potential V, and we show that for this choice of ν_n in (5.29), we have the desired LDP for the quenched Gibbs measure.

Theorem 5.3.13 Let $d \geq 3$ and g be the Coulomb kernel (5.13). Under Assumptions 5.3.1, 5.3.4, 5.3.8, 5.3.9 and 5.3.10, consider $\beta_n \geq un^{1+\delta}$ for $\delta > 0$ and $u > 0$ (in particular $\beta_n \gg n$). For any $n \geq 2$, let $x_1, \dots, x_n \sim \mathbb{P}^V_{n,\beta_n}$ and let ν_n be as in (5.29).

Consider the kernel K_ζ in (5.15) with

$$0 < \zeta < \frac{\delta \wedge \frac{2}{d}}{2d}. \tag{5.32}$$

Moreover, consider the quenched empirical measure $\mu_n^{\mathrm{Q}} = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, where $y_1, \dots, y_n \sim \mathrm{d}\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}(.\,; \nu_n(\omega))$ in Definition 5.3.5 with kernel K_ζ.

Then for $\mathbb{P}$-almost all ω, μ_n^{Q} satisfies a large deviation principle (Definition 3.1.6) as $n \to +\infty$, at speed β_n and with good rate function

$$\mu \mapsto I_g^{V^\pi}(\mu) - I_g^{V^\pi}(\pi), \tag{5.33}$$

whose minimizer is π. More precisely, for $\mathbb{P}$-almost all ω,

$$\begin{aligned} -\inf_{\mu \in \mathring{A}} \{I_g^{V^\pi}(\mu) - I_g^{V^\pi}(\pi)\} &\leq \varlimsup_{n \to +\infty} \frac{1}{\beta_n} \log \mathbb{P}_{n,\beta_n}^{\mathrm{Q}} \left(\mu_n^{\mathrm{Q}} \in A; \nu_n(\omega)\right) \\ &\leq -\inf_{\mu \in \overline{A}} \{I_g^{V^\pi}(\mu) - I_g^{V^\pi}(\pi)\}. \end{aligned} \tag{5.34}$$

As a consequence, for $\mathbb{P}$-almost all ω:

1. μ_n^{Q} converges in law to π,
2. There exists $c > 0$ such that for any $r > 0$ and n large enough,

$$\mathbb{P}_{n,\beta_n}^{\mathrm{Q}} \left(\mathrm{d_{BL}} \left(\frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}, \pi \right) > r; \, \nu_n(\omega) \right) \leq \exp(-c\beta_n r^2). \tag{5.35}$$

The proof is again in Section 5.5. The first consequence is immediate, and the second one follows from the Coulomb transport inequality (Chafaï et al., 2018, Theorem 1.2). Again, Theorem 5.3.13 gives tighter asymptotic simultaneous confidence intervals for Monte Carlo integration of bounded Lipschitz functions than with points drawn i.i.d from π or using MCMC. In that case however, we need the first approximation ν_n to π to be built as well on a Coulomb gas, i.e. to exhibit fast converging properties, this is due to the fact that the quality of ν_n interplays with the regularization that we need to make to deal with the singularity of g. Moreover, the fact that the regularized Coulomb interaction (5.15) satisfies Assumption 5.3.3 will play an important role in the proof. In particular, in dimensions 1 and 2, the natural counterpart to (5.15) is given by

$$K'_\zeta(x,y) = -\frac{1}{2} \log \left(|x-y|^2 + n^{-2\zeta} \right), \tag{5.36}$$

which fails to satisfy Assumption 5.3.3 since $I_K(\mu) > 0$ only for finite signed Borel measures with total mass 0 (Pronzato et al., 2020). Still, it would be interesting to try to circumvent this issue in the proofs.

We will use a general result on LDPs for Gibbs measures from (García-Zelada, 2019) to get Theorem 5.3.7 and Theorem 5.3.13. The case of non-singular interactions is relatively easy to handle, whereas more control is needed in the Coulomb case because of the regularization. In particular, the key control will be given by Proposition 5.3.15. In the latter, we prove uniform convergence of the regularized potential generated by

the Coulomb gas $x_1, \dots, x_n$. We first begin with an unweighted version of the desired result, which we will use in the proof of the weighted version and which we believe to be interesting per se as discussed in Section 5.2.

Proposition 5.3.14 Let $d \geq 3$ and g be the Coulomb kernel. Under Assumptions 5.3.4 and 5.3.9, let $\beta_n \geq u n^{1+\delta}$ for some $\delta > 0$ and $u > 0$. Let $x_1, \dots, x_n \sim \mathbb{P}^V_{n,\beta_n}$ with $K = g$ and

$$0 < \zeta < \frac{\delta \wedge \frac{2}{d+2}}{d-2}, \tag{5.37}$$

and let μ_n denote the associated empirical measure.

Then $\mathbb{P}$-almost surely,

$$A_n := \sup_{z \in \mathbb{R}^d} \left| U^{\mu_n}_{K_\zeta}(z) - U^{\mu_V}_g(z) \right| \underset{n \to +\infty}{\to} 0. \tag{5.38}$$

We give a sketch of proof, to be fleshed out in Section 5.6.1. We first note that K_ζ satisfies Assumption 5.3.3, so that $K_\zeta(z, .)$ is in the RKHS $\mathcal{H}_{K_\zeta}$ for all z and has norm $n^{\zeta(d-2)/2}$ in $\mathcal{H}_{K_\zeta}$. In particular, the duality formula from Proposition 4.3.3 yields

$$\sup_{z \in \mathbb{R}^d} \left| U^{\mu_n}_{K_\zeta}(z) - U^{\mu_V}_{K_\zeta}(z) \right| \leq \sqrt{n^{\zeta(d-2)} I_{K_\zeta}(\mu_n - \mu_V)}. \tag{5.39}$$

Then, one can show (Lemma 5.6.3) that I_{K_ζ} is always smaller than I_g. Since the points $x_1, \dots, x_n$ are generated from a Coulomb gas with interaction g, one should thus be able to use concentration properties of Gibbs measures (Lemma 5.6.4) to show that $n^{\zeta(d-2)} I_g$ goes to zero. However, due to the singularity of g, $I_g(\mu_n - \mu_V)$ is infinite. We bypass this issue by convolving each Dirac δ_{x_i} in μ_n with a uniform distribution on a small ball of radius $n^{-\epsilon}$ to make sense of the energy I_g. Denoting by $\mu_n^{(\epsilon)}$ this convolved empirical measure, we thus show that

$$n^{\zeta(d-2)} I_{H_\zeta}(\mu_n^{(\epsilon)} - \mu_V) \leq n^{\zeta(d-2)} I_g(\mu_n^{(\epsilon)} - \mu_V) \to 0 \tag{5.40}$$

for a suitable choice of ϵ and ζ. It then remains to show that the error terms vanish when one replaces $U^{\mu_n}_{K_\zeta} - U^{\mu_V}_g$ by $U^{\mu_n^{(\epsilon)}}_{K_\zeta} - U^{\mu_V}_{K_\zeta}$, which we do in Lemma 5.6.1 and Lemma 5.6.2.

We now state the same result for the weighted measures, the proof of which relies on Proposition 5.3.14 and will be used in the proof of Theorem 5.3.13.

Proposition 5.3.15 Let $d \geq 3$ and g be the Coulomb kernel (5.13). Under Assumptions 5.3.1, 5.3.4, 5.3.8, 5.3.9 and 5.3.10, let $\beta_n \geq u n^{1+\delta}$ for some $\delta > 0$ and

$u > 0$. Let $x_1, \dots, x_n \sim \mathbb{P}^V_{n,\beta_n}$ with $K = g$ and

$$0 < \zeta < \frac{\delta \wedge \frac{2}{d}}{2d}, \tag{5.41}$$

and let $\nu_n = \sum_{i=1}^n w(x_i)\delta_{x_i}$ be the weighted empirical measure defined in (5.29).

Then $\mathbb{P}$-almost surely,

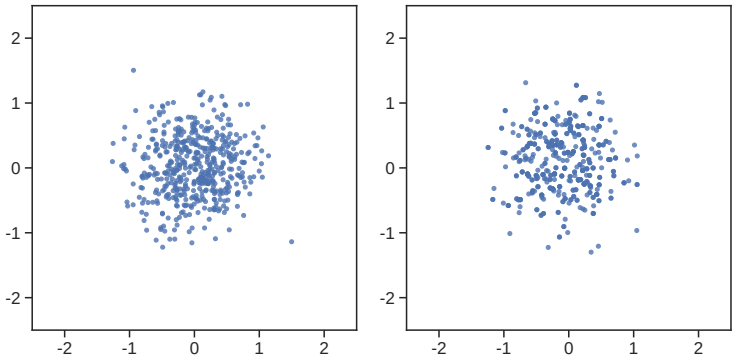
$$B_n := \sup_{z \in \mathbb{R}^d} \left| U^{\nu_n}_{K_\zeta}(z) - U^{\pi}_{g}(z) \right| \underset{n \to +\infty}{\to} 0. \tag{5.42}$$

The constraint on ζ is upper bounded by $\mathcal{O}(1/d^2)$ both in Proposition 5.3.14 and Proposition 5.3.15, but it is not expected to be optimal at all.

We give again a sketch of proof, with details to be found in Section 5.6.2. The beginning of the proof is in the same vein as for Proposition 5.3.14: we show using Lemma 5.6.2 and Lemma 5.6.5 that it is enough (i.e. the error terms vanish) to study $U^{\nu_n^{(\epsilon)}}_{K_\zeta} - U^{\pi}_{K_\zeta}$ where $\nu_n^{(\epsilon)}$ is the weighted empirical measure with each Dirac convolved by a uniform distribution on $B(0, n^{-\epsilon})$. For that purpose, we need instead of $\nu_n^{(\epsilon)}$ to consider $\nu_n^{\epsilon,\eta}$ in which the weights are a smoothed version of W obtained by convolving with a Gaussian mollifier. Then, we wish to decouple the weights from the unweighted measure $\mu_n^{(\epsilon)}$ and μ_V to apply once again the concentration properties for the Gibbs measure from which the points are generated. To do so, a key idea is to use Lemma 5.7.1 that rephrases the energy $I_K(\mu - \nu)$ in Fourier space as an L^2 norm between the characteristic functions of μ and ν with respect to a spectral measure induced by K. For $K = K_\zeta$ and $K = g$, explicit expressions and useful estimates for this spectral measure are given in Lemma 5.7.2. In particular, using Lemmas 5.7.1 and 5.7.2 and Proposition 5.3.14, we show in Lemma 5.6.6 that we can split this L^2 norm in two terms: one vanishing term and the other one controlled by quantities of the form $n^{2\zeta d} I_g(\mu_n^{(\epsilon)} - \mu_V)$ which converges to 0 again for suitable choices of ϵ and ζ.

5.4 Experiments

First, we emphasize that since we relied on MCMC in Chapter 4 to approximate the potential U^{π}_K, the experiments presented therein (Section 4.4) already illustrate our results for non-singular interactions (Theorem 5.3.7). In particular, we highlighted in Section 4.4.2 that Monte Carlo integration with the quenched Gibbs measure (5.3.5) indeed provides tighter coverage than crude MCMC, at fixed precision on the integration error. We quickly recall how we sample from Gibbs measures such as (5.3.5). We then illustrate our main results on a toy example for Bayesian classification, and give additional experiments to discuss the difficulty of sampling from the quenched Gibbs measure with MALA.



(a) Gibbs, $\beta_n = n^2$ **(b) MCMC**

Fig. 5.1: First two coordinates of an approximate sample from (5.1a) the quenched Gibbs measure with $\beta_n = n^2$ and 3D truncated Gaussian target π, (5.1b) MCMC with 3D truncated Gaussian target π.

5.4.1 Approximately sampling from the Gibbs measure

In this section, we first recall our crude method to sample approximately from the quenched Gibbs measure.

We use MALA in the same way as in Section 4.4.1 to sample approximately from the quenched Gibbs measure (5.3.5). Recall that each of the T MALA iterations thus costs $\mathcal{O}(n^2 + nM_n)$ for the quenched Gibbs measure of Definition 5.3.5, where M_n is the number of atoms of ν_n.

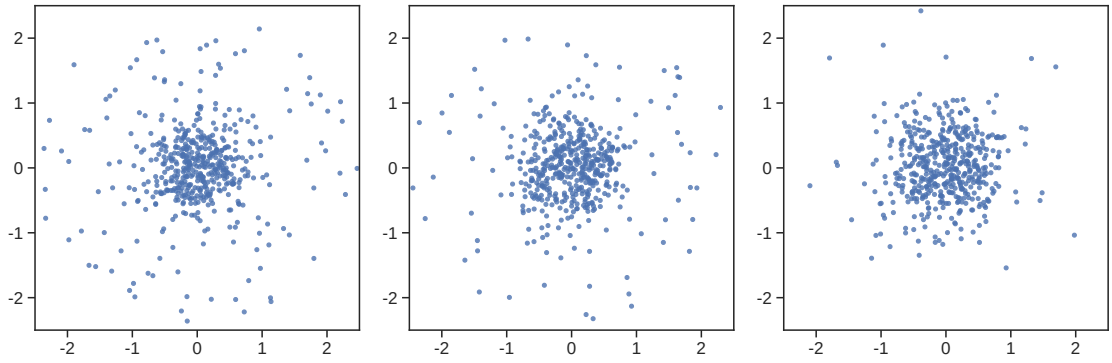
We consider again a toy example where $d = 3$ and $n = 500$. The target distribution π is a centered Gaussian distribution with covariance $\sigma^2 I_d$, truncated at $|x| > 5\sigma$, and with $\sigma = 0.5$. We consider the quenched Gibbs measure in Definition 5.3.5 with $\beta_n = n^2$, and interaction kernel the regularized Riesz kernel $K_{s,\epsilon}$ in (5.14),

$$K_{s,\epsilon}(x, y) = \frac{1}{(|x - y|^2 + \epsilon^2)^{s/2}}, \tag{5.43}$$

with $s = d - 2$ and $\epsilon = 0.1$. As for the background, we set $\nu_n = \frac{1}{M_n} \sum_{i=1}^{M_n} \delta_{x_i}$ where $x_1, \dots, x_{M_n}$ is a realization of the history of a Metropolis–Hastings chain targeting π, with $M_n = 1\,000$.

In Figure 5.1a, we show the first two spatial coordinates of the 50 000-th iterate of a MALA chain approximating the Gibbs measure. For comparison, in Figure 5.1b, we show the first two spatial coordinates of an approximate sample $y_1, \dots, y_n$ from π, obtained as the history $y_1, \dots, y_n$ of a Metropolis–Hastings chain targeting π with Gaussian proposal $N(0, \alpha I_d)$, where α is tuned to reach 50% acceptance, and we removed 5 000 burn-in iterations.

We see in Figure 5.1 that both methods stay confined in the regions of large probability under the target π, as expected. Moreover, the approximated Gibbs samples in



(a) $T = 50\,000$ **(b)** $T = 100\,000$ **(c)** $T = 200\,000$

Fig. 5.2: First two coordinates of an approximate sample from the quenched Gibbs measure, with $\beta_n = n^3$ and increasing number T of MALA iterations.

Figure 5.1a are more diverse than a Metropolis-Hastings sample, let alone because about half the proposed points in the latter are rejected, explaining the many superimposed disks in Figure 5.1b. On top of these rejections causing autocorrelation in the MCMC chain, we expect the interaction term in the Gibbs measure (5.3.5) to favor diverse configurations. However, this promotion of diversity is balanced by 1) the use of an MCMC chain as a confining background, 2) the near-singular interaction kernel we chose which makes sampling hard, 3) the constraint that a significant proportion of the points should remain in the bulk of the Gaussian target, as well as 4) the fact that we project a 3-dimensional sample onto a plane. This explains why repulsiveness is less visually striking in Figure 5.1a than in Figure 4.1b, where we considered an approximate Gibbs sample when the target π is the 2D uniform measure on $B(0,1)$, with interaction kernel $-\frac{1}{2}\log(|x-y|^2)$, exact potential U^π and $T = 5\,000$ MALA iterations.

In Figure 5.2, we plot the first two coordinates in the same setup as Figure 5.1a, with the same realization of the background, but this time with an inverse temperature of $\beta_n = n^3$. In Figure 5.2a and Figure 5.2b, we see that points spread a lot in the domain of the target, even in regions of low probability under π. In light of Figure 5.2c, where outer regions of low probability under π now contain a matching fraction of the points, one can understand the overspread in Figures 5.2a and 5.2b as a consequence of MALA needing more iterations to yield an accurate approximation to its Gibbs target. This supports the intuitive claim that, outside warm-start cases in log-concave setting, higher β_n comes with a worse mixing time for MALA: the lower the temperature, the more difficult it is to approximately sample a Gibbs measure, see the discussion in Section 4.4.1. There indeed seems to be a tradeoff in choosing a large β_n for better statistical guarantees, and keeping β_n low to preserve tractable approximate sampling procedures. Recall from Section 4.4 that we recommend $\beta_n = n^2$ in practice, which already yields improved statistical guarantees over independent and MCMC quadratures, and has remained computationally tractable in our experiments. We now turn to a toy

example coming from Bayesian classification.

5.4.2 Bayesian classification

In this section, we give a practical statistical illustration that a quenched Gibbs quadrature provides better coverage of simultaneous confidence intervals for a given number of integration nodes n, compared to classical MCMC.

First, we define the target integration task starting from the foundational expected utility principle; this is classical in Bayesian statistics, we include it for completeness. Consider a training data set $(z_j, t_j)_{j=1}^N$ where $z_j \in \mathcal{Z} = \mathbb{R}^d$ encodes the features of the j-th individual and $t_j \in \mathcal{T} = \{1, \dots, C\}$ is the associated class label. An *action* is the choice of a classifier $g : \mathcal{Z}^N \times \mathcal{T}^N \times \mathcal{Z} \to \mathcal{T}$. For a new point (z, t) with unknown class t, if one considers a joint probability measure p on $z_1, t_1, \dots, z_N, t_N, z, t$ and the so-called $0-1$ utility function, the expected utility of the action g is given by

$$-\int 1_{t \neq g(z_{1:N}, t_{1:N}, z)} \mathrm{d}p(z_{1:N}, t_{1:N}, z, t), \tag{5.44}$$

where $z_{1:N}$ is short for $(z_1, \dots, z_N)$. The Bayes action or classifier is the function $g^\star$ which maximizes the expected utility (5.44). It is easy to see that it is given by

$$g^\star : z_{1:N}, t_{1:N}, z \mapsto \arg\max_{k=1}^{C} p(t = k | z_{1:N}, t_{1:N}, z). \tag{5.45}$$

In other words, the Bayes classifier predicts the class k with the highest conditional probability under p. Making the classical assumption that p factorizes as

$$p(z_{1:N}, t_{1:N}, z, t) = \int p(y) \cdot \prod_{j=1}^{N} p(t_j | z_j, y) p(z_j) \cdot p(t | z, y) p(z) \mathrm{d}y, \tag{5.46}$$

where we introduced a parameter $y \in \mathcal{Y}$ and a reference measure $\mathrm{d}y$, the conditional in (5.45) writes

$$p(t = k | z_{1:N}, t_{1:N}, z) = \int p(t = k | z, y) \pi(y) \mathrm{d}y, \tag{5.47}$$

where

$$\pi(y) \propto \prod_{j=1}^{N} p(t_j | z_j, y) p(y)$$

is the density of the *posterior* probability measure on the parameter y. Using a Monte Carlo method to replace π by an approximation $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$, we approximate the Bayes action by

$$g_n^{\mathrm{MC}} : z_{1:N}, t_{1:N}, z \mapsto \arg\max_{k=1}^{C} \int p(t = k | z, y) \mathrm{d}\mu_n(y). \tag{5.48}$$

One way to guarantee that the expected utility (5.44) of g_n^{MC} is close to the one of $g^\star$ is to ensure that the C integrals in the argmax of (5.48) simultaneously approximate the C integrals in the argmax of (5.45). Furthermore, if the dataset on which we want to perform classification contains M test points instead of only one and we want to ensure that we approximate all predictions of the Bayes classifier using a single Monte Carlo run $y_1, \dots, y_n$, then there are $M \times C$ integrals to simultaneously approximate. In that case, Theorem 5.3.7 and Theorem 5.3.13 show that the asymptotic confidence regions obtained with a single run $y_{1:n}^{\mathrm{Q}}$ from the quenched Gibbs measure are tighter than the ones obtained for a single run $y_{1:n}^{\mathrm{MCMC}}$ from MCMC, depending on which smoothness one wants to assume on $p(t = k|z, y)$. We empirically illustrate in the following that it is also the case at finite n and for a MCMC-approximated sample of the quenched Gibbs measure, using MALA as in Section 4.4.1.

To specify our setting, we consider binary logistic regression, where $C = 2$, $\mathcal{Z} = \mathbb{R}^2$, $\mathcal{T} = \{0, 1\}$, $d = 3$, and we assume p to factorize as in (5.46) with

$$p(t = 1|z, y) = 1 - p(t = 0|z, y) \propto \exp\left(y^\top [z, 1]\right), \tag{5.49}$$

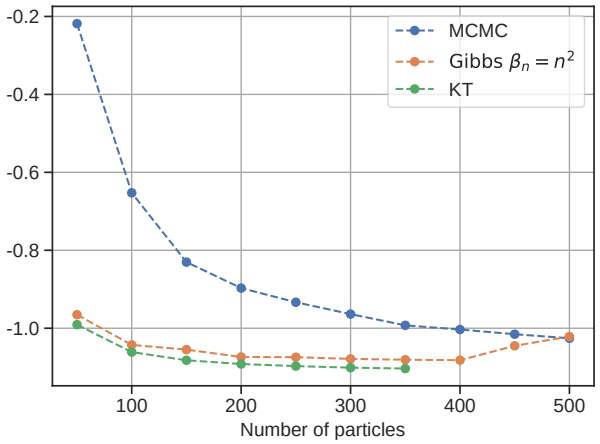
where $[z, 1]$ stands for x concatenated with 1 to include an intercept. To keep the posterior π compactly supported, we consider the prior $p(y)$ to be a three-dimensional Gaussian distribution with covariance $\sigma^2 I_3$ truncated at $|x| > 5\sigma$, with $\sigma = 0.5$ as before. We mimic a toy dataset from the documentation of the blackjax[1] library: the features $z_{1:N}$ are drawn from a balanced mixture of two isotropic Gaussian distributions, and the class $t_{1:N}$ corresponds to the component membership. We consider $N = 50$ individuals for the training set, and $M = 10$ test individuals to classify. Computing the Bayes classifier for these 10 individuals then requires computing $M(C - 1) = 10$ integrals with respect to π. In order to be able to compare with a reference value for the Bayes classifier, we first approximate the conditional $p(t = 1|z_{1:N}, t_{1:N}, z)$ in (5.47), for each z in the test set, using a long Metropolis–Hastings chain targeting π, with 10^5 iterations. Then, we compare the results to the Monte Carlo approximation

$$\hat{p}_n(t = 1|z_{1:N}, t_{1:N}, z) = \frac{1}{n} \sum_{i=1}^{n} p(t = 1|z, y_i), \tag{5.50}$$

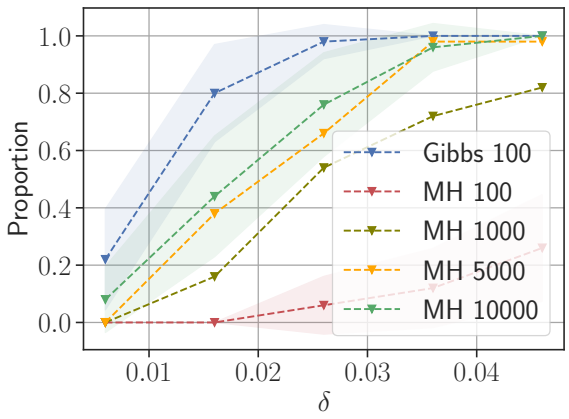
when $y_1, \dots, y_n$ is i) the history of a Metropolis–Hastings chain targeting π after 500 burn-in iterations, ii) an approximate sample from the quenched Gibbs measure, after $T = 10\,000$ MALA iterations, with regularized Riesz kernel with $s = d - 2$ and $\epsilon = 0.1$, at inverse temperature $\beta_n = n^2$, and with background ν_n of size $M_n = 1000$ given by an i.i.d subsample of the history of a Metropolis–Hastings chain. We build 50 independent samples for each method, and compute for each value of δ the proportions $\mathfrak{p}_{\delta,\mathrm{MCMC}}$ and $\mathfrak{p}_{\delta,\mathrm{Gibbs}}$ of realizations of samples for which the M estimators (5.50) simultaneously fall in the M confidence regions

$$[\hat{p}_n(t = 1|z_{1:N}, t_{1:N}, z) - \delta\,;\, \hat{p}_n(t = 1|z_{1:N}, t_{1:N}, z) + \delta]$$

[1] https://blackjax-devs.github.io/sampling-book/models/logistic_regression.html



(a) Decay of the energies.



(b) Proportions for Bayesian classification.

Fig. 5.3: (5.3a) Quantile of order 90% of $I_K(\mu_n - \pi) - I_K(\pi)$, in log scale, for points obtained via MCMC, the quenched Gibbs measure, and kernel thinning. (5.3b) Proportion of runs where the M estimators fall simultaneously in the confidence regions of size δ.

corresponding to the M values z in the test set. We plot $\mathfrak{p}_{\delta,\mathrm{MCMC}}$ and $\mathfrak{p}_{\delta,\mathrm{Gibbs}}$ for different values of δ in Figure 5.3b when $n = 100$ for ii) and for different values of n for i). We use Gaussian confidence intervals at confidence level 95% with Bonferroni correction for those 5 values of δ.

As expected, the Gibbs measure yields tighter simultaneous confidence intervals. In particular, 100 particles drawn from the quenched Gibbs measure exhibit a coverage quality that is at least the same the history of a Metropolis–Hastings chain of size 10 000.

5.4.3 Worst-case integration error and variance

In this section, we give additional experiments similar to the ones of Section 4.4 on the decay of the worst-case integration error and the variance of linear statistics, highlighting limitations of MALA to sample from Gibbs distributions such as (5.3.5).

We consider the same setting as in Section 5.4.1 where π is a three-dimensional Gaussian distribution $\mathcal{N}(0, \sigma^2 I_3)$, truncated at $|x| > 5\sigma$ with $\sigma = 0.5$. We estimate the worst-case integration error of various quadratures $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ as n grows.

We first evaluate $\mathrm{E}_{\mathcal{H}_K}(\mu_n) = \sqrt{I_K(\mu_n - \pi)}$ when $K = K_{s,\epsilon}$ again with $s = d - 2$ and $\epsilon = 0.1$. In particular, we compute

$$I_K(\mu_n - \pi) - I_K(\pi) = I_K(\mu_n) - 2I_K(\mu_n, \pi),$$

using a long MH chain targeting π to approximate the integral with respect to π, with size $M = 90,000$ and $10,000$ burn-in iterations. We show in Figure 5.3a how $\mathrm{E}_{\mathcal{H}_K}(\mu_n)$

decays when n grows for i) μ_n^{MCMC}, where the (y_i) is the history of an MH chain targeting π after 5000 burn in iterations, ii) μ_n^{Q}, where the (y_i) are an approximate sample from the quenched Gibbs measure with $T = 10\,000$ iterations, temperature $\beta_n = n^2$, kernel $K = K_{s,\epsilon}$ and $\nu_n = \frac{1}{M_n}\sum_{i=1}^{M_n} \delta_{x_i}$ where $M_n = 1,000$ and the x_i are the history of an MH chain targeting π again, iii) μ_n^{KT} where the (y_i) are a subset of size n of an n^2 MCMC chain $x_1, \dots x_{n^2}$ targeting π. In the latter method, the subset if obtained via the kernel thinning algorithm of Section 2.6.4, who proved that the $(y_i)_{i=1,\dots,n}$ preserve the same sample quality as the original $(x_i)_{i=1,\dots,n^2}$ with high probability. We compute the 90%-quantile of the worst-case error, over 100 independent realizations of the quadrature for each method and each cardinality n, with fixed background for ii) and iii).

As expected, we see that both kernel thinning and our method outperform MCMC in terms of worst-case integration error, at least up to $n = 350$ particles. For kernel thinning, we used the vanilla version presented in (Dwivedi et al., 2021) which requires storing all the n^4 pairwise interactions for the underlying MCMC background of size n^2, which is why we could not run the experiment for more than $n = 350$ due to storage issues. Note however that recent refinements of this algorithm might help circumvent this issue (Shetty et al., 2022).

Beyond $n = 350$, we observe that the energy increases for the quenched Gibbs measure, becoming similar to the MCMC error, which we attribute again to the fact that $T = 10\,000$ MALA iterations are not enough to approximate the Gibbs measure for that large a number of particles.

In the following, we provide further experiments that investigate the impact of the background, as well as a conjectured faster decaying variance for linear statistics.

In Figure 5.4, we run the same experiment as in Figure 5.3a, where we vary the size M_n of the MCMC background (x_i) to see whether there is an impact on the worst-case integration error. We also add to the comparison a background ν_n obtained via reweighting the realization of a Coulomb gas $x_1, \dots, x_n$ as in Equation (5.29), with quadratic potential V so that μ_V is the uniform measure on $B(0, 5\sigma)$. We use $T = 10\,000$ iterations of MALA to approximately sample that Coulomb background.

When the background is drawn using MCMC, we see a slight improvement between a background of size n and a background of (larger) fixed size 1000 or 5000. When the background is drawn from a first well-spread Coulomb gas with reweighting as in (2.58), we observe a small loss in terms of energy, especially for small values of n. We conjecture that this is due to the quality of the estimation when using weights to approximate the potential with a small number of points n.

Then, we look at the effect of increasing the temperature schedule β_n in Figure 5.5a for MCMC background with size $M_n = 1000$, and in Figure 5.5b for a reweighted Coulomb background $x_1, \dots, x_n$ with uniform equilibrium measure. Again, we clearly see a downgrade in terms of energy for lower temperature schedules which is again due to the fact that $T = 10\,000$ are not enough to sample from the Gibbs measure in those

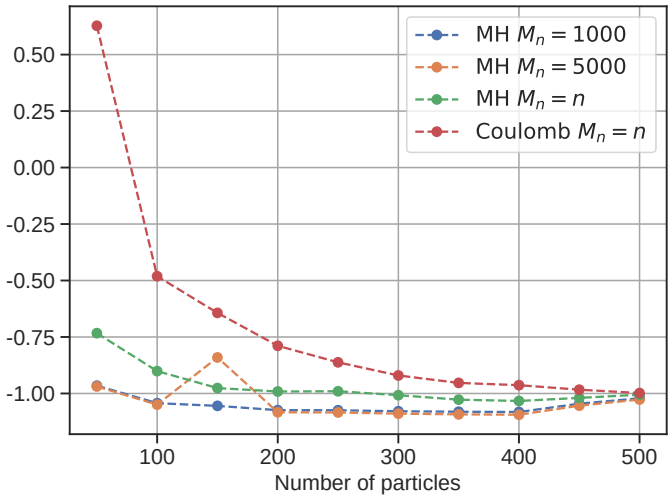
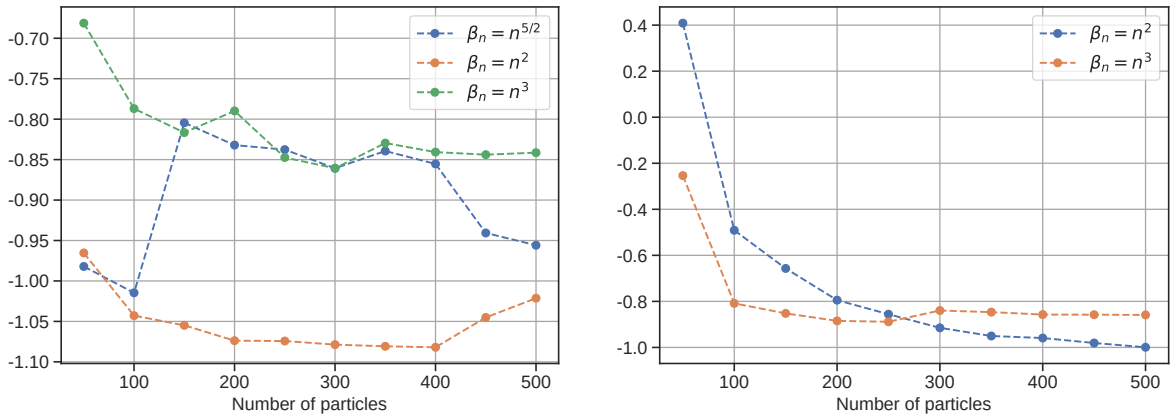


Fig. 5.4: Quantile of order 90% of $I_K(\mu_n - \pi) - I_K(\pi)$ for different background of the quenched Gibbs measure.

regimes as we saw in Figure 5.2. Running the same experiments with $T = 200\,000$ MALA iterations for instance is very costly, and one could argue that it is not reasonable for a three-dimensional truncated Gaussian target.



(a) MCMC background **(b) Reweighted Coulomb background**

Fig. 5.5: Quantile of order 90% of $I_K(\mu_n - \pi) - I_K(\pi)$ for different temperature schedules for the quenched Gibbs measure.

We now look at the variance of linear statistics with respect to μ_n for the three different methods (i-iii) used for Figure 5.3a. We conjecture that Gibbs measures can significantly improve on MCMC in that criterion, as illustrated as well in Section 4.4, although we defer the corresponding theoretical investigation to further work. We consider a simple test function in $\mathcal{H}_{K_{s,\epsilon}}$ given by $f(x) = \sum_{i=1}^{10} a_i K_{s,\epsilon}(x, z_i)$ where the (z_i) are chosen

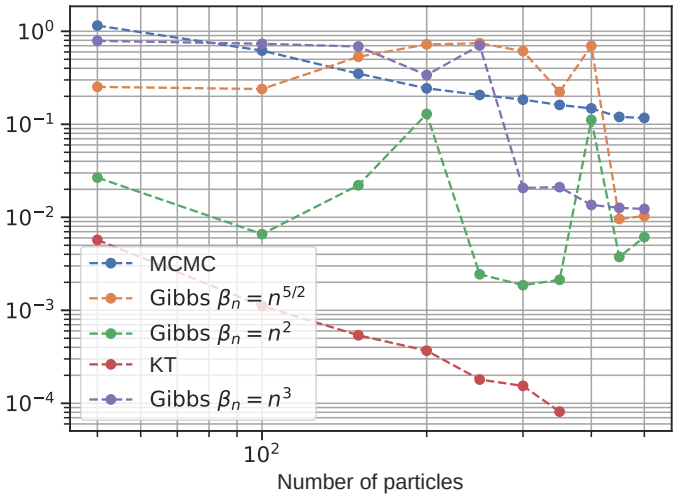


Fig. 5.6: Empirical variance of $\frac{1}{n}\sum_{i=1}^n f(x_i)$ in log scale for points obtained via MCMC, the quenched Gibbs measure with different temperature schedules, and kernel thinning.

uniformly at random on $[-1,1]^d$ and the a_i are chosen uniformly at random on $[-1,1]$. We compute in Figure 5.6 the empirical estimate of the variance $\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n f(y_i)\right]$ over 100 independent realizations of $y_1,\dots,y_n$ for each method and each value of n, with MCMC background of size $M_n = 1\,000$ and $T = 10\,000$ MALA iterations for the quenched Gibbs measure.

Kernel thinning clearly exhibits a faster rate of convergence than MCMC, as it is expected from the fast rate of Theorem 2.6.4 for integrands in the RKHS. On the other hand, the variance of the Gibbs estimator can –when tuned well– fall between that of MCMC and kernel thinning, yet it is hard to conclude anything for the rates corresponding to Gibbs measures. This is again related to the difficulty of tuning the MALA sampler as the Gibbs target changes.

5.5 Proofs: quenched large deviations

García-Zelada (2019) established a LDP for Gibbs measures in a quite general framework. In Section 5.5.1, we recall his results in the case when $\frac{\beta_n}{n}$ converges to infinity (that we denote by $\beta_n \gg n$) and detail in particular the conditions that we need to check to apply the results in our context. In Section 5.5.2, we check these conditions in the case of non-singular kernels and conclude the proof of Theorem 5.3.7. To further prove Theorem 5.3.13, where the kernel is the Coulomb kernel g, we will need Proposition 5.3.15, whose proof is the main technical input of the present paper and is postponed to Section 5.6.1.

5.5.1 Reminder on LDPs for Gibbs measures, after (García-Zelada, 2019)

Equip the set $\mathcal{P}(\mathbb{R}^d)$ of probability measures on $\mathbb{R}^d$ with the topology of weak convergence. The following theorem is a general large deviations result for Gibbs measures, due to (García-Zelada, 2019).

Theorem 5.5.1 — García-Zelada (2019), Corollary 1.3. For a reference probability measure $\nu \in \mathcal{P}\left(\mathbb{R}^d\right)$, we consider the unnormalized Gibbs measure

$$\mathrm{d}\gamma_n(y_1, \ldots, y_n) = \exp^{-\beta_n W_n(y_1,\ldots,y_n)} \, \mathrm{d}\nu^{\otimes n}(y_1, \ldots, y_n),$$

with $W_n : \left(\mathbb{R}^d\right)^n \to (-\infty, +\infty]$ symmetric and measurable. We extend W_n to all probability measures by defining

$$\tilde{W}_n(\mu) = \begin{cases} W_n(y_1, \ldots, y_n) \text{ if } \mu \text{ is atomic with } \mu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i} \\ +\infty \text{ otherwise} \end{cases} . \tag{5.51}$$

Consider the following conditions.

1. *(Stable sequence)* There exists a constant $C \in \mathbb{R}$ such that for all n, $W_n \geq C$ uniformly over $\left(\mathbb{R}^d\right)^n$.
2. *(Confining sequence)* For any increasing sequence of natural numbers (n_j) and any sequence of probability measures (μ_j) on $\mathbb{R}^d$, if there exists a constant A such that $\tilde{W}_{n_j}(\mu_j) \leq A$ for every j then (μ_j) is relatively compact.
3. *(Lower limit)* For any sequence of probability measures (μ_n) on $\mathbb{R}^d$ converging to μ, we have $W(\mu) \leq \underline{\lim}\, \tilde{W}_n(\mu_n)$.
4. *(Regularity)* For any probability measure μ on $\mathbb{R}^d$ such that $W(\mu) < +\infty$, we can find a sequence of probability measures (μ_n) in $\mathcal{N}$ such that $\mu_n \to \mu$ and $\overline{\lim}\, W(\mu_n) \leq W(\mu)$, where

$$\mathcal{N} = \{\xi \in \mathcal{P}(\mathbb{R}^d) \, : \, \mathrm{KL}(\xi, \nu) < +\infty, \, \overline{\lim}\, \mathbb{E}_{\xi^{\otimes n}}\left[W_n\right] \leq W(\xi)\}, \tag{5.52}$$

   and $\mathrm{KL}(\xi, \nu)$ is the relative entropy (2.45).

Under Conditions 1, 2, 3, and 4, further assume that $Z_n = \gamma_n(\left(\mathbb{R}^d\right)^n)$ is positive for n large enough. Define the normalized Gibbs measure

$$\mathrm{d}\mathbb{P}_n = \frac{1}{Z_n}\gamma_n, \tag{5.53}$$

as well as $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ the associated empirical measure with $y_1, \ldots, y_n \sim \mathbb{P}_n$.

Then μ_n satisfies a LDP (Definition 3.1.6) as $n \to +\infty$ with speed β_n and good rate function

$$\mu \mapsto W(\mu) - \inf_{\eta \in \mathcal{P}(\mathbb{R}^d)} W(\eta). \tag{5.54}$$

When Conditions 1 to 4 are met, it is said that W is the macroscopic zero-temperature limit of W_n. Our objective is to apply Theorem 5.5.1 to the discrete energy H_n^{Q}, where we recall from Definition 5.3.5 that

$$H_n^{\mathrm{Q}}(y_1, \dots, y_n) = \frac{1}{2n^2} \sum_{i \neq j} K(y_i, y_j) + \frac{1}{n} \sum_{i=1}^n v(y_i) - \frac{1}{n} \sum_{i=1}^n U_K^{\nu_n}(y_i), \tag{5.55}$$

with $v = 0$ inside $B(0, R)$ and $v = \Phi$ outside and where ν_n is a random measure. The role of the macroscopic zero-temperature limit W in Theorem 5.5.1 would naturally be played by the energy functional in Definition 3.3.2, that is

$$I_K^{V^\pi}(\mu) = \frac{1}{2} \iint K(x, y) \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y) + \int v \, \mathrm{d}\mu - \int U_K^\pi \, \mathrm{d}\mu = I_K^v(\mu) - \int U_K^\pi \, \mathrm{d}\mu \tag{5.56}$$

when $I_K^{V^\pi}(\mu) < \infty$.

The main difficulty in applying Theorem 5.5.1 is the dependence of H_n^{Q} to the random measure ν_n. García-Zelada (2019) does use Theorem 5.5.1 to prove large deviations for conditioned Gibbs measures (their Theorem 4.1) and particles in varying environment (their Theorem 4.2), but their framework cannot be directly applied to our case since they either restrict to particles supported on a compact subset, or use a different sign convention for the background distributions. Neither of these settings covers our case, since we wish for a Gibbs measure supported on $\left(\mathbb{R}^d\right)^n$, and because the conditioned Gibbs measure considered in (García-Zelada, 2019) does not converge to an easily predefinable target due to their sign convention. We will however rely on some standard arguments from (García-Zelada, 2019) in the upcoming proofs. Relating their proof to our framework turned out to be relatively easy for non-singular interactions, but tricky in the Coulomb case, where our Proposition 5.3.15 will give the missing key result.

We start with a result that only considers the deterministic part of H_n^{Q}, and can be considered a corollary of Theorem 5.5.1.

Proposition 5.5.2 Define

$$W_n^0(y_1, \dots, y_n) = \frac{1}{2n^2} \sum_{i \neq j} K(y_i, y_j) + \frac{1}{n} \sum_{i=1}^n v(y_i) \tag{5.57}$$

and

$$I_K^v(\mu) = \frac{1}{2}\iint K(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) + \int v\,\mathrm{d}\mu, \tag{5.58}$$

where K is either the Coulomb kernel g or a non-singular interaction kernel that satisfies Assumption 5.3.3, and v satisfies Assumption 5.3.4 with constant $c' > 0$.

Then one can build a reference probability measure ν and a functional W_n such that

- $\gamma_n = \exp\left(-\beta_n W_n\right)\,\mathrm{d}\nu^{\otimes n}$ is exactly (up to normalization) the Gibbs measure associated to $H_n = W_n^0$ and confinement v in 3.45;
- (W_n, I_K^v) satisfies Conditions 1, 2, 3 and 4 of Theorem 5.5.1. In particular, the empirical measure associated to $\mathbb{P}^v_{n,\beta_n}$ with $\beta_n \gg n$ satisfies a LDP with speed β_n and rate function

$$\mu \mapsto I_K^v(\mu) - I_K^v(\mu_v). \tag{5.59}$$

Proof. We follow the proof of (García-Zelada, 2019, Theorem 4.11). Consider the reference probability measure

$$\nu = \frac{\exp\left(-c'v(z)\right)}{\int \exp\left(-c'v(z)\right)\,\mathrm{d}z}\,\mathrm{d}z$$

and set, for any fixed $\epsilon \in (0,1)$,

$$G_1(x,y) = K(x,y) + \epsilon v(x) + \epsilon v(y)\,,\; G_2(x,y) = (1-\epsilon)v(x) + (1-\epsilon)v(y),$$

$$W_n^1(y_1,\dots,y_n) = \frac{1}{n^2}\sum_{i\neq j} G_1(y_i,y_j)\,,\; W_n^2(y_1,\dots,y_n) = \frac{1}{n^2}\sum_{i\neq j} G_2(y_i,y_j),$$

$$a_n = \frac{1}{1-\epsilon}\left(\frac{1}{n-1}\left(n - \frac{n^2}{\beta_n}c'\right) - \epsilon\right),$$

and $W_n = W_n^1 + a_n W_n^2$.

The assumptions to apply Theorem 5.5.1 are straightforward given our restrictions on g and v, see (Chafaï et al., 2014, Proposition 2.8), and the second point is just a rewriting of the different terms in the Gibbs measure. ■

For the specific choice of $v = 0$ in $B(0,R)$ and $v = \Phi$ outside, Assumption 5.3.4 is clearly satisfied, so that we can already apply Proposition 5.5.2 with K being either Coulomb or satisfying Assumption 5.3.3. Defining W_n and ν as in Proposition 5.5.2, I_K^v is the zero-temperature macroscopic limit of W_n, and we can rewrite the unnormalized

quenched Gibbs measure we are interested in as

$$\exp\left(-\beta_n H_n^{\mathrm{Q}}(y_1,\dots,y_n)\right)\,\mathrm{d}y_1\,\dots\,\mathrm{d}y_n = \exp\left(-\beta_n\left[W_n(y_1,\dots,y_n)-\frac{1}{n}\sum_{i=1}^n U_K^{\nu_n}(y_i)\right]\right)\,\mathrm{d}\nu(y_1)\,\dots\,\mathrm{d}\nu(y_n). \tag{5.60}$$

If we denote

$$W_n^{\mathrm{Q}}(y_1,\dots,y_n) = W_n(y_1,\dots,y_n)-\frac{1}{n}\sum_{i=1}^n U_K^{\nu_n}(y_i), \tag{5.61}$$

which is symmetric and measurable, we want to show that $I_K^{V^\pi}$ is the zero-temperature macroscopic limit of W_n^{Q} $\mathbb{P}$-almost surely. Indeed, since the partition function of $\mathbb{P}_{n,\beta_n}^{\mathrm{Q}}$ is well-defined $\mathbb{P}$-almost surely for n large enough (i.e. finite and positive, see Definition 5.3.5), we will then be able to apply Theorem 5.5.1, concluding the proof of Theorem 5.3.7 and Theorem 5.3.13 respectively. It remains to check that $(W_n^{\mathrm{Q}}, I_K^{V^\pi})$ satisfies Condition 1 to 4 $\mathbb{P}$-almost surely and for n large enough. This part of the proof will differ in the non-singular case and in the Coulomb case.

5.5.2 Proof of Theorem 5.3.7 (non-singular kernels)

Let K satisfy Assumption 5.3.3. We examine the conditions of Theorem 5.5.1 one by one.

First, $0 \le K \le C$ by assumption, so that $-U_K^{\nu_n} \ge -C$ and W_n^{Q} satisfies Condition 1 since W_n satisfies it also.

Second, assume that $\tilde{W}_n^{\mathrm{Q}}(\mu_j) \le A$ for some constant A. Then the $(\mu_j)_j$ have to be empirical measures $\mu_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\delta_{y_{i,j}}$. Using again $U_K^{\nu_j} \le C$, this yields $W_{n_j}(\mu_j) \le A+C$. Since W_n satisfies Condition 2, this implies that $(\mu_j)_j$ is relatively compact, and thus W_n^{Q} satisfies the same condition.

Third, assume that $\mu_n \to \mu$ with μ_n empirical measures. Since K is bounded and continuous and $\nu_n \to \pi$ $\mathbb{P}$-almost surely,

$$\iint K\,\mathrm{d}\nu_n\otimes\mu_n \to \int U_K^\pi\,\mathrm{d}\mu \tag{5.62}$$

$\mathbb{P}$-almost surely by Fubini. Since W_n satisfies Condition 3, W_n^{Q} also satisfies it $\mathbb{P}$-almost surely.

Fourth, let μ be such that $I_K^{V^\pi}(\mu) < +\infty$. Then, since U_K^π is bounded, $I_K^v(\mu) < +\infty$. By Proposition 5.5.2, W_n satisfies Condition 4, so that there exists a sequence $(\mu_n) \in \mathcal{N}$ such that $\mu_n \to \mu$ and $\overline{\lim}\, I_K^v(\mu_n) \le W(\mu)$ with

$$\mathcal{N} = \{\xi \in \mathcal{P}(\mathbb{R}^d) \,:\, \mathrm{KL}(\xi,\nu) < +\infty,\ \overline{\lim}\,\mathbb{E}_{\xi^{\otimes n}}[W_n] \le I_K^v(\xi)\}. \tag{5.63}$$

Since $-U_K^\pi$ is bounded and continuous by dominated convergence, $\mu \mapsto -\int U_K^\pi \,\mathrm{d}\mu$ is continuous on $\mathcal{P}(\mathbb{R}^d)$, so that $\overline{\lim} I_K^{V^\pi}(\mu_n) \leq I_K^{V^\pi}(\mu)$. Finally, for any $\xi \in \mathcal{P}(\mathbb{R}^d)$,

$$\mathbb{E}_{\xi^{\otimes n}}\left[-\frac{1}{n}\sum_{i=1}^n U_K^{\nu_n}(y_i)\right] = -\int U_K^{\nu_n}\,\mathrm{d}\xi = -\int U_K^{\xi}\,\mathrm{d}\nu_n \tag{5.64}$$

by Fubini. Again, this yields

$$\mathbb{E}_{\xi^{\otimes n}}\left[-\frac{1}{n}\sum_{i=1}^n U_K^{\nu_n}(y_i)\right] \to -\int U_K^{\xi}\,\mathrm{d}\pi = -\int U_K^{\pi}\,\mathrm{d}\xi \tag{5.65}$$

$\mathbb{P}$-almost surely, so that $\overline{\lim}\,\mathbb{E}_{\xi^{\otimes n}}\left[W_n^{\mathrm{Q}}\right] \leq I_K^{V^\pi}(\xi)$ for any $\xi \in \mathcal{N}$, and in particular for the $(\mu_n)_n$. Thus Condition 4 is satisfied for W_n^{Q}, $\mathbb{P}$-almost surely. This concludes the proof of Theorem 5.3.7.

5.5.3 Proof of Theorem 5.3.13 (Coulomb kernel)

Similarly, when $K = g$ is the Coulomb kernel and for our specific choice of v, we check that all the conditions to apply Theorem 5.5.1 are satisfied.

Since the interaction is neither bounded nor continuous, we lose the continuity and boundedness properties for the energies compared to Section 5.5.2, so our approach must differ a bit. We define

$$B_n := \sup_{z\in\mathbb{R}^d}\left|U_{K_\zeta}^{\nu_n}(z) - U_g^\pi(z)\right|. \tag{5.66}$$

We show Theorem 5.3.13 given Proposition 5.3.15, whose proof is the technical crux and is postponed to Section 5.6.2.

We first consider Condition 3. Assume that the sequence of empirical measures $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ converges to μ. Due to the maximum principle, if $U_g^\pi \leq M$ on the support of π then this holds everywhere (Landkof, 1972, Theorem 1.10, page 71). Under the assumption that π has compact support and continuous density, $z \mapsto U_g^\pi(z)$ is continuous (Chafaï et al., 2014, Lemma 4.3), so that U_g^π is bounded on the support of π and thus everywhere. Thus, $\mu \longmapsto -\int U_g^\pi \,\mathrm{d}\mu$ is lower semi-continuous and bounded from below. In particular, we have

$$C \leq -\int U_g^\pi \,\mathrm{d}\mu \leq \underline{\lim}\left\{-\frac{1}{n}\sum_{i=1}^n U_g^\pi(z_i)\right\} \leq \underline{\lim}\left\{-\frac{1}{n}\sum_{i=1}^n U_{K_\zeta}^{\nu_n}(z_i) + B_n\right\}. \tag{5.67}$$

Furthermore,

$$\underline{\lim}\, W_n(z_1,\ldots,z_n) + \underline{\lim}\left\{-\frac{1}{n}\sum_{i=1}^n U_{K_\zeta}^{\nu_n}(z_i)\right\} \leq \underline{\lim}\, W_n^Q(z_1,\ldots,z_n). \tag{5.68}$$

Because W_n satisfies Condition 3 by Proposition 5.5.2, and $B_n \to 0$ $\mathbb{P}$-almost surely by Proposition 5.3.15, Equation (5.67) and Equation (5.68) imply that W_n^{Q} also satisfies it, $\mathbb{P}$-almost surely.

We now consider Condition 1. Using again Proposition 5.3.15 along with

$$-\frac{1}{n}\sum_{i=1}^{n} U_{K_\zeta}^{\nu_n}(z_i) \geq -\frac{1}{n}\sum_{i=1}^{n} U_g^{\pi}(z_i) - B_n, \tag{5.69}$$

and the fact that U_g^π is bounded, $-\frac{1}{n}\sum_{i=1}^n U_{K_\zeta}^{\nu_n}(z_i)$ is bounded from below by a constant C uniformly in n for n large enough and $\mathbb{P}$-almost surely. Since W_n is also uniformly lower bounded, this stays true for W_n^{Q}.

We now consider Condition 2. We proceed as for non-singular interactions: if $\tilde{W}_{n_j}^{\mathrm{E}}(\mu_j) \leq A$, then $W_{n_j}(\mu_j) \leq A + C$ since we just showed that $-\frac{1}{n}\sum_{i=1}^n U_{K_\zeta}^{\nu_n}(z_i)$ is bounded below for n large enough. So (μ_j) is relatively compact and W_n^{Q} satisfies the assumption $\mathbb{P}$-almost surely for n large enough.

We finish by Condition 4. We already know that U_g^π is bounded and continuous so $\overline{\lim}\, I_g^{V^\pi}(\mu_n) \leq I_g^{V^\pi}(\mu)$ by the same argument as for non-singular interactions. Then, for $\xi \in \mathcal{P}(\mathbb{R}^d)$,

$$\mathbb{E}_{\xi^{\otimes n}}\left[-\frac{1}{n}\sum_{i=1}^{n} U_{K_\zeta}^{\nu_n}(y_i)\right] = -\int U_{K_\zeta}^{\nu_n}\,\mathrm{d}\xi. \tag{5.70}$$

Using Proposition 5.3.15, we get

$$\overline{\lim}\,\mathbb{E}_{\xi^{\otimes n}}\left[-\frac{1}{n}\sum_{i=1}^{n} U_{K_\zeta}^{\nu_n}(y_i)\right] \leq -\int U_g^{\pi}\,\mathrm{d}\xi, \tag{5.71}$$

$\mathbb{P}$-almost surely. Again, the inequality for the sum of superior limits is in the good direction and we obtain $\overline{\lim}\,\mathbb{E}_{\xi^{\otimes n}}\left[W_n^{\mathrm{Q}}\right] \leq I_g^{V^\pi}(\xi)$ for any $\xi \in \mathcal{N}$, the set in the Condition 4 satisfied by W_n. This concludes the proof of Theorem 5.3.13.

5.6 Proofs: uniform convergence for the Coulomb potential field

We use here the notations of Section 5.3, in particular the definitions of the Coulomb kernel g from (5.13) and its regularized version K_ζ in (5.15). We give in this section the proofs of Proposition 5.3.14 and Proposition 5.3.15 along with the statements of the auxiliary lemmas we use. To ease the reading, we delay the proofs of all lemmas to the next section.

5.6.1 Proof of Proposition 5.3.14

Recall the proof strategy that we sketched after stating Proposition 5.3.14. We first bound the differences $U_g^\pi - U_{K_\zeta}^\pi$ and $U_{K_\zeta}^{\mu_n} - U_{K_\zeta}^{\mu_n^{(\epsilon)}}$ uniformly in z, in Lemma 5.6.1 and Lemma 5.6.2. Then, we use Lemma 5.6.3 and Lemma 5.6.4 to conclude by the concentration properties of the Coulomb gas and with a suitable choice of regularization parameters. These intermediate results are then assembled into the proof of Proposition 5.3.14.

Lemma 5.6.1 Let μ be a probability measure with compact support and continuous density. Then for any $\zeta > 0$ and $0 < s < 1$,

$$\begin{aligned}\sup_{z\in\mathbb{R}^d} \left|U_{K_\zeta}^\mu(z) - U_g^\mu(z)\right| &= \sup_{z\in\mathbb{R}^d} \left|\int K_\zeta(z,y)\,\mathrm{d}\mu(y) - \int g(z,y)\,\mathrm{d}\mu(y)\right| \\ &= \mathcal{O}\left(n^{-4s\zeta/(d+2s)}\right). \end{aligned} \tag{5.72}$$

Lemma 5.6.2 For any $x_1, \dots, x_n \in \mathbb{R}^d$, let μ_n be the empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Let $\epsilon > 0$ and $\mu_n^{(\epsilon)} = \mu_n \star \lambda_{B(0,n^{-\epsilon})}$, where $\star$ denotes the convolution, and $\lambda_{B(0,R)}$ denotes the uniform probability measure on $B(0,R)$. In other words,

$$\mu_n^{(\epsilon)} = \frac{1}{n}\sum_{i=1}^n \lambda_{B(x_i,n^{-\epsilon})}. \tag{5.73}$$

Then, $\mathbb{P}$-almost surely,

$$\sup_{z\in\mathbb{R}^d} \left|U_{K_\zeta}^{\mu_n}(z) - U_{K_\zeta}^{\mu_n^{(\epsilon)}}(z)\right| = \mathcal{O}\left(n^{-2\epsilon+d\zeta}\right). \tag{5.74}$$

Lemma 5.6.3 For any finite Borel measures μ, ν in $\mathbb{R}^d$,

$$I_{K_\zeta}(\mu-\nu) \le I_g(\mu-\nu). \tag{5.75}$$

Lemma 5.6.4 — Chafaï et al., 2018. Assume that $\beta_n \gg n$, $n \ge 2$, and that Assumptions 5.3.4 and 5.3.9 hold. Then

$$I_g(\mu_n^{(\epsilon)} - \mu_V) \le 2\left(I_g^V(\mu_n^{(\epsilon)}) - I_g^V(\mu_V)\right). \tag{5.76}$$

Moreover,

$$Z_{n,\beta_n}^V \ge \exp\left\{-\beta_n I_g^V(\mu_V) + n\left(\frac{\beta_n}{2n^2} I_g(\mu_V) + S(\mu_V)\right)\right\}, \tag{5.77}$$

where $S(\mu_V) = \mathrm{KL}(\mu_V, \lambda)$ is the entropy (2.45) of μ_V with respect to the Lebesgue measure. Finally, setting $\eta = c'n/\beta_n$ where c' is the constant of Assumption 5.3.4, and $\epsilon_n = n^{-\epsilon}$ for $\epsilon > 0$, we have

$$H_n(x_1,\dots,x_n) \ge \eta\frac{1}{n}\sum_{i=1}^n V(x_i) + (1-\eta)\left(I_g^V(\mu_n^{(\epsilon)}) - C\frac{1}{n\epsilon_n^{d-2}} - C\epsilon_n^2\right). \tag{5.78}$$

As a consequence, for any Borel set $A \subset \left(\mathbb{R}^d\right)^n$,

$$\begin{aligned}&\log \mathbb{P}^V_{n,\beta_n}(A) \leq \\ &-(\beta_n - c'n) \inf_A \left(I^V_g(\mu_n^{(\epsilon)}) - I^V_g(\mu_V)\right) + Cn + C\beta_n \left(\frac{1}{n} + \frac{1}{n\epsilon_n^{d-2}} + \epsilon_n^2\right),\end{aligned} \tag{5.79}$$

and, using Equation (5.76), we also get

$$\log \mathbb{P}^V_{n,\beta_n}(A) \leq -\frac{1}{2}(\beta_n - c'n) \inf_A I_g(\mu_n^{(\epsilon)} - \mu_V) + Cn + C\beta_n \left(\frac{1}{n} + \frac{1}{n\epsilon_n^{d-2}} + \epsilon_n^2\right). \tag{5.80}$$

Proof of Proposition 5.3.14. We can apply Lemma 5.6.1 with $\mu = \mu_V$ and Lemma 5.6.2 to get

$$A_n \leq \sup_{z\in\mathbb{R}^d} \left|U^{\mu_n^{(\epsilon)}}_{K_\zeta}(z) - U^{\mu_V}_{K_\zeta}(z)\right| + Cn^{-2\epsilon+d\zeta} + Cn^{-4s\zeta/(d+2s)}, \tag{5.81}$$

for any $0 < s < 1$ and $\epsilon > 0$.

Then, recall that since K_ζ satisfies Assumption 5.3.3, we can build a RKHS (Definition 2.1.5) with reproducing kernel K_ζ. To wit, the construction starts with

$$\mathcal{H}_0 = \operatorname{span}\left\{\sum_{j=1}^p a_j K_\zeta(z_j, .)\,, a_1, \dots, a_p \in \mathbb{R}\,, z_1, \dots, z_p \in \mathbb{R}^d\right\}, \tag{5.82}$$

endowed with the inner product induced by the reproducing property

$$\langle f\,, K_\zeta(z, .)\rangle = f(z) \tag{5.83}$$

for any $f \in \mathcal{H}_0$ and $z \in \mathbb{R}^d$. Then $\mathcal{H}_{K_\zeta}$ is obtained by completion of $\mathcal{H}_0$ with respect to the latter inner product, and the reproducing property is true for any $f \in \mathcal{H}_{K_\zeta}$. Clearly, $K_\zeta(z, .)$ is an element of $\mathcal{H}_{K_\zeta}$ for any fixed $z \in \mathbb{R}^d$, and

$$\|K_\zeta(z, .)\|_{\mathcal{H}_{K_\zeta}} = \sqrt{\langle K_\zeta(z, .)\,,\, K_\zeta(z, .)\rangle} = \sqrt{K_\zeta(z, z)} = \sqrt{n^{\zeta(d-2)}}. \tag{5.84}$$

Hence, using the duality formula from Proposition 4.3.3 with $f = K_\zeta(z, .)$, we have

$$\begin{aligned}\sup_{z\in\mathbb{R}^d} \left|\int K_\zeta(z, y)\, \mathrm{d}(\mu_n^{(\epsilon)} - \mu_V)\right| &\leq \sqrt{n^{\zeta(d-2)}} \sup_{\|f\|_{\mathcal{H}_{K_\zeta}} \leq 1} \int f\, \mathrm{d}(\mu_n^{(\epsilon)} - \mu_V) \\ &= \sqrt{n^{\zeta(d-2)} I_{K_\zeta}(\mu_n^{(\epsilon)} - \mu_V)} \\ &\leq \sqrt{n^{\zeta(d-2)} I_g(\mu_n^{(\epsilon)} - \mu_V)}.\end{aligned} \tag{5.85}$$

where we used Lemma 5.6.3 in the last inequality. Thus,

$$A_n \leq \sqrt{n^{\zeta(d-2)} I_g(\mu_n^{(\epsilon)} - \mu_V)} + Cn^{-2\epsilon+d\zeta} + Cn^{-4s\zeta/(d+2s)}. \tag{5.86}$$

It remains to make a good choice of $\epsilon > \frac{d\zeta}{2}$ so that all the terms go to zero $\mathbb{P}$-almost surely, using the concentration inequality of Lemma 5.6.4. In particular, for any fixed $r > 0$, we set

$$A = \left\{(x_1, \dots, x_n) \,:\, n^{\zeta(d-2)} I_g\left(\mu_n^{(\epsilon)} - \mu_V\right) > r^2\right\} \tag{5.87}$$

in Lemma 5.6.4 to get

$$\log \mathbb{P}^V_{n,\beta_n}(A) \leq -c\beta_n n^{-\zeta(d-2)} r^2 + Cn + C\beta_n \left(\frac{1}{n} + \frac{1}{n\epsilon_n^{d-2}} + \epsilon_n^2\right), \tag{5.88}$$

where we used the fact that for any constant $0 < c < 1/2$, $\beta_n - c'n \geq 2c\beta_n$ for n large enough since $\beta_n \gg n$. As a consequence, for any fixed $r > 0$ and for n large enough,

$$\log \mathbb{P}^V_{n,\beta_n}\left(n^{\zeta(d-2)} I_g(\mu_n^{(\epsilon)} - \mu_V) > r^2\right) \leq -C\beta_n n^{-\zeta(d-2)} r^2, \tag{5.89}$$

as soon as

$$\zeta(d-2) < \delta \wedge 1 \wedge (1 - \epsilon(d-2)) \wedge 2\epsilon.$$

Together with $\epsilon > \frac{d\zeta}{2}$, this gives

$$\zeta(d-2) < \delta \wedge 1 \wedge (1 - \epsilon(d-2)) \wedge 2\frac{d-2}{d}\epsilon. \tag{5.90}$$

Choosing $\epsilon = \frac{d}{(d-2)(d+2)}$, the value for which all the constraints on ϵ match, we see that the constraints (5.90) are always satisfied for

$$\zeta < \frac{\delta \wedge \frac{2}{d+2}}{d-2}, \tag{5.91}$$

which is the assumption we made in Proposition 5.3.14. The Borel-Cantelli lemma then yields

$$n^{\zeta(d-2)} I_g(\mu_n^{(\epsilon)} - \mu_V) \underset{n\to+\infty}{\to} 0, \tag{5.92}$$

$\mathbb{P}$-almost surely, which concludes the proof. ■

5.6.2 Proof of Proposition 5.3.15

In this section, we will always work under Assumptions 5.3.1, 5.3.4, 5.3.8, and 5.3.9, even though most of the following lemmas do not require all of them to be true at once.

We begin in Lemma 5.6.5 by a Taylor expansion similar to Lemma 5.6.2, but rather dealing with the weights W from (5.27), the motivation of this will become clear afterward in the proof. Since W is only assumed to be continuous, we smooth W by convolving it with a Gaussian.

Lemma 5.6.5 For $\eta > 0$, consider

$$\rho_\eta(x) = \frac{1}{(2\pi\eta^2)^{d/2}} \exp\left(-\frac{|x|^2}{2\eta^2}\right), \quad x \in \mathbb{R}^d. \tag{5.93}$$

For any $x_1, \dots, x_n \in \mathbb{R}^d$ and $\epsilon > 0$, let

$$\nu_n^{\epsilon,\eta}(A) = \frac{1}{n} \sum_{i=1}^n \int_A (W \star \rho_\eta)(y) \, \mathrm{d}\lambda_{B(x_i, n^{-\epsilon})}(y). \tag{5.94}$$

Then,

$$\sup_{z \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n (W \star \rho_\eta)(x_i) \int K_\zeta(z,y) \lambda_{B(x_i,n^{-\epsilon})}(y) - U_{K_\zeta}^{\nu_n^{\epsilon,\eta}}(z) \right| \le \eta^{-1} \mathcal{O}\left(n^{-\epsilon+\zeta(d-2)}\right), \tag{5.95}$$

where the constant in the $\mathcal{O}$ does not depend on η.

In the following lemma, we bound the difference of potentials for weighted measures. To do so, we rely on Cauchy–Schwarz to get bounds in terms of the L_2 norm of the difference of the characteristic functions of the measures. We then isolate the weights from the L_2 norm of the difference of characteristic functions for unweighted measures, which is the interaction energy between the unweighted measures (see Lemma 5.7.1). Recall that we know how to bound the latter through concentration inequalities (Lemma 5.6.4).

Lemma 5.6.6 Recall the definition of $\nu_n^{\epsilon,\eta}$ from Lemma 5.6.5, and that μ_V' is the density of μ_V. For any $0 < s < 1$ and $1 < \alpha < 1/s$, we have

$$\begin{aligned} \sup_{z \in \mathbb{R}^d} \left| U_{K_\zeta}^{\nu_n^{\epsilon,\eta}}(z) - U_{K_\zeta}^{(W\star\rho_\eta)\cdot\mu_V'}(z) \right| &\le \frac{C}{\eta^{d/2}} \|W\|_{L^2(\mathbb{R}^d)} n^{\zeta(d-2+1/s)} d_{\mathrm{BL}}(\mu_n^{(\epsilon)}, \mu_V) \\ &+ \frac{C\|W\|_{L^1(\mathbb{R}^d)}}{\eta^{d/2}} \exp\left(-\frac{1}{4} n^{\zeta(\alpha-1)}\right) \left(n^{\zeta(\alpha+1)(d/2-1)} + n^{\zeta(d-2)}\right) \\ &+ \frac{C\|W\|_{L^1(\mathbb{R}^d)}}{\eta^{d/2}} n^{\zeta(d-2)} \exp\left(-\eta^2 \left(n^{\zeta/s} - n^{\alpha\zeta}\right)^2/2\right), \end{aligned} \tag{5.96}$$

where d_{BL} is the bounded Lipschitz distance (2.11) given by

$$d_{\mathrm{BL}}(\mu,\nu) := \sup_{\|f\|_\infty \le 1, \|f\|_{Lip} \le 1} \left| \int f \, \mathrm{d}(\mu - \nu) \right|. \tag{5.97}$$

In particular, setting for instance $s = 1/2$ and any $1 < \alpha < 2$,

$$\limsup_{n\to+\infty} \sup_{z \in \mathbb{R}^d} \left| U_{K_\zeta}^{\nu_n^{\epsilon,\eta}}(z) - U_{K_\zeta}^{(W\star\rho_\eta)\cdot\mu_V'}(z) \right| \le \limsup_{n\to+\infty} \frac{C\|W\|_{L^1(\mathbb{R}^d)}}{\eta^{d/2}} n^{\zeta d} d_{\mathrm{BL}}(\mu_n^{(\epsilon)}, \mu_V) \tag{5.98}$$

We will now conclude using the concentration as before and trying to optimize the different parameters to make all the error terms go to 0.

Proof of Proposition 5.3.15. Recall that we want to show the convergence of

$$B_n := \sup_{z\in\mathbb{R}^d} \left| U_{K_\zeta}^{\nu_n}(z) - U_g^{\pi}(z) \right| \tag{5.99}$$

where $\nu_n = \sum_{i=1}^n w(x_i)\delta_{x_i}$. Since $\sum_{i=1}^n w(x_i) = 1$, the exact same computations as in Lemma 5.6.2 yield

$$\left| U_{K_\zeta}^{\nu_n}(z) - \sum_{i=1}^n w(x_i) \int K_\zeta(z,y)\,\mathrm{d}\lambda_{B(x_i,n^{-\epsilon})}(y) \right| = \mathcal{O}\left(n^{-2\epsilon+d\zeta}\right), \tag{5.100}$$

uniformly in z, for any $\epsilon > 0$. Together with Lemma 5.6.1, this yields, for any $0 < s < 1$,

$$\begin{aligned} B_n \le \sup_{z\in\mathbb{R}^d} &\left| \sum_{i=1}^n w(x_i) \int K_\zeta(z,y)\,\mathrm{d}\lambda_{B(x_i,n^{-\epsilon})}(y) - \int K_\zeta(z,y)W(y)\,\mathrm{d}\mu_V(y) \right| \\ &+ Cn^{-2\epsilon+d\zeta} + Cn^{-4s\zeta/(d+2s)}. \end{aligned} \tag{5.101}$$

Moreover, by definition of w and since $0 \le K_\zeta \le g$ and $U_g^\pi(z) \le C$ uniformly in z (Lemma 5.7.3),

$$\begin{aligned} &\sup_{z\in\mathbb{R}^d} \left| \sum_{i=1}^n w(x_i) \int K_\zeta(z,y)\,\mathrm{d}\lambda_{B(x_i,n^{-\epsilon})}(y) - \int K_\zeta(z,y)W(y)\,\mathrm{d}\mu_V(y) \right| \\ &\quad\le \frac{n}{\sum_{i=1}^n W(x_i)} \sup_{z\in\mathbb{R}^d} \left| \frac{1}{n}\sum_{i=1}^n W(x_i) \int K_\zeta(z,y)\,\mathrm{d}\lambda_{B(x_i,n^{-\epsilon})}(y) - \int K_\zeta(z,y)W(y)\,\mathrm{d}\mu_V(y) \right| \\ &\qquad + \left| 1 - \frac{n}{\sum_{i=1}^n W(x_i)} \right| C. \end{aligned} \tag{5.102}$$

Recall that $\frac{1}{n}\sum_{i=1}^n W(x_i) \to 1$, $\mathbb{P}$-a.s. Moreover, for any $\eta > 0$,

$$\begin{aligned} &\sup_{z\in\mathbb{R}^d} \left| \frac{1}{n}\sum_{i=1}^n W(x_i) \int K_\zeta(z,y)\,\mathrm{d}\lambda_{B(x_i,n^{-\epsilon})}(y) - \int K_\zeta(z,y)W(y)\,\mathrm{d}\mu_V(y) \right| \\ &\quad\le \sup_{z\in\mathbb{R}^d} \left| \frac{1}{n}\sum_{i=1}^n (W\star\rho_\eta)(x_i) \int K_\zeta(z,y)\,\mathrm{d}\lambda_{B(x_i,n^{-\epsilon})}(y) - \int K_\zeta(z,y)(W\star\rho_\eta)(y)\,\mathrm{d}\mu_V(y) \right| \\ &\qquad + \|W - W\star\rho_\eta\|_\infty \sup_{z\in\mathbb{R}^d} \left(\left| U_{K_\zeta}^{\mu_n^{(\epsilon)}}(z) \right| + \left| U_{K_\zeta}^{\mu_V}(z) \right| \right). \end{aligned} \tag{5.103}$$

But $U_{K_\zeta}^{\mu_V}(z) \le U_g^{\mu_V}(z) \le C$ for the same reasons as $U_g^\pi \le C$, so that using Lemma 5.6.5, the left-hand side of (5.103) is bounded by

$$\begin{aligned} &\sup_{z\in\mathbb{R}^d} \left| U_{K_\zeta}^{\nu_n^{\epsilon,\eta}}(z) - U_{K_\zeta}^{(W\star\rho_\eta)\cdot\mu_V'}(z) \right| \\ &\qquad + \eta^{-1}Cn^{-\epsilon+\zeta(d-2)} + \|W - W\star\rho_\eta\|_\infty \left(\sup_{z\in\mathbb{R}^d} \left| U_{K_\zeta}^{\mu_n^{(\epsilon)}}(z) - U_{K_\zeta}^{\mu_V}(z) \right| + 2C \right). \end{aligned} \tag{5.104}$$

Recall now from the proof of Proposition 5.3.14 in Section 5.6.1 that

$$\sup_{z\in\mathbb{R}^d}\left|U_{K_\zeta}^{\mu_n^{(\epsilon)}}(z)-U_{K_\zeta}^{\mu_V}(z)\right|\leq\sqrt{n^{\zeta(d-2)}I_{K_\zeta}(\mu_n^{(\epsilon)}-\mu_V)}. \tag{5.105}$$

Using Lemma 5.6.6 and Lemma 5.6.3, we thus get that, for any $\epsilon>(\frac{d}{2}\vee(d-2))\zeta$,

$$\begin{aligned}\limsup_{n\to+\infty}B_n\leq\limsup_{n\to+\infty}C_{\eta,W}&\left[n^{\zeta d}d_{\mathrm{BL}}(\mu_n^{(\epsilon)},\mu_V)+\sqrt{n^{\zeta(d-2)}I_g(\mu_n^{(\epsilon)}-\mu_V)}\right]\\&+2C\|W-W\star\rho_\eta\|_\infty.\end{aligned} \tag{5.106}$$

It now remains to use the concentration inequality from Lemma 5.6.4 to show that under the right constraint on ζ and ϵ, the first term goes to 0 $\mathbb{P}-$a.s. Using (Chafaï et al., 2018, Theorem 1.2) and (5.76), it is enough to show that

$$n^{2\zeta d}(I_g^V(\mu_n^{(\epsilon)})-I_g^V(\mu_V))\to 0. \tag{5.107}$$

Using Lemma 5.6.4, we know that for any $r>0$, setting

$$A=\{(x_1,\dots,x_n):n^{2\zeta d}\left(I_g^V(\mu_n^{(\epsilon)})-I_g^V(\mu_V)\right)>r^2\}, \tag{5.108}$$

we have

$$\log\mathbb{P}_{n,\beta_n}^V(A)\leq-c\beta_n n^{-2\zeta d}r^2+Cn+C\beta_n\left(\frac{1}{n}+\frac{1}{n\epsilon_n^{d-2}}+\epsilon_n^2\right) \tag{5.109}$$

with $\epsilon_n=n^{-\epsilon}$. Thus,

$$n^{2\zeta d}\left(I_g^V(\mu_n^{(\epsilon)})-I_g^V(\mu_V)\right)\to 0 \tag{5.110}$$

upon choosing $0<\epsilon<\frac{1}{d-2}$ and

$$2\zeta d<\delta\wedge 1\wedge(1-(d-2)\epsilon)\wedge 2\epsilon. \tag{5.111}$$

Incorporating the constraint $\epsilon>(d/2\vee(d-2))\zeta$ yields

$$\begin{aligned}\zeta<&\frac{\delta\wedge 1}{2d}\wedge\frac{1}{2d}(1-(d-2)\epsilon)\wedge\left(\frac{1}{d}\wedge\frac{2}{d}\wedge\frac{1}{d-2}\right)\epsilon\\=&\frac{\delta}{2d}\wedge\frac{1}{2d}(1-(d-2)\epsilon)\wedge\frac{1}{d}\epsilon.\end{aligned} \tag{5.112}$$

Setting $\epsilon=\frac{1}{d}<\frac{1}{d-2}$, to make the two constraints on ϵ coincide, it is thus enough to choose

$$0<\zeta<\frac{\delta\wedge 2/d}{2d}, \tag{5.113}$$

which is the assumption we made in Proposition 5.3.15. With this choice, we thus get

$$\limsup_{n\to+\infty}B_n\leq 2C\|W-W\star\rho_\eta\|_\infty. \tag{5.114}$$

Letting $\eta\to 0$, we finally get the result, since W is continuous and compactly supported under Assumption 5.3.10. ■

5.7 Proof of auxiliary lemmas

5.7.1 Preliminaries

We will repeatedly use a particular integral representation for g and K_ζ. First, a quick computation shows that for any $c > 0$ and $\alpha > 0$,

$$\Gamma(\alpha) := \int_0^{+\infty} t^{\alpha-1} e^{-t} \, \mathrm{d}t = c^\alpha \int_0^{+\infty} t^{\alpha-1} e^{-ct} \, \mathrm{d}t.$$

Choosing $\alpha = \frac{d-2}{2}$, this yields

$$g(z,y) = \frac{1}{\Gamma\left(\frac{d-2}{2}\right)} \int_0^{+\infty} t^{\frac{d-2}{2}-1} \exp\left(-t|z-y|^2\right) \, \mathrm{d}t, \tag{5.115}$$

$$K_\zeta(z,y) = \frac{1}{\Gamma\left(\frac{d-2}{2}\right)} \int_0^{+\infty} t^{\frac{d-2}{2}-1} \exp\left(-n^{-2\zeta}t\right) \exp\left(-t|z-y|^2\right) \, \mathrm{d}t. \tag{5.116}$$

It will be often convenient to work in Fourier space to bound the energies. In fact (Sriperumbudur et al., 2009, Corollary 4) states that for bounded translation invariant kernels $K(x,y) = \Psi(\|x-y\|)$, the energy I_K of a signed measure can be seen as the L^2-norm of the characteristic function of that measure, w.r.t. to a spectral measure which Ψ is the Fourier transform of. We give explicit expressions for this spectral measure in the following lemma for K_ζ and g, which proves the formula at the same time even though g is not bounded.

Lemma 5.7.1 Set

$$\Lambda_c(\omega) := \int_0^{+\infty} t^{-2} \exp\left(-\frac{1}{4t}|\omega|^2\right) \, \mathrm{d}t = C' \frac{1}{|\omega|^2}, \tag{5.117}$$

where $C' > 0$ is a constant, and

$$\Lambda_\zeta(\omega) := \int_0^{+\infty} t^{-2} \exp(-n^{-2\zeta}t) \exp\left(-\frac{1}{4t}|\omega|^2\right) \, \mathrm{d}t \leq \Lambda_c(\omega). \tag{5.118}$$

Then for any signed measure μ with finite energy I_{K_ζ}, we have

$$K_\zeta(z,y) = \int e^{-i\langle z-y\,,\,\omega\rangle} \Lambda_\zeta(\omega) \, \mathrm{d}\omega, \tag{5.119}$$

$$U^\mu_{K_\zeta}(z) = \int e^{-i\langle z\,,\,\omega\rangle} \Phi_\mu(\omega) \Lambda_\zeta(\omega) \, \mathrm{d}\omega, \tag{5.120}$$

$$I_{K_\zeta}(\mu) = \int |\Phi_\mu|^2(\omega) \Lambda_\zeta(\omega) \, \mathrm{d}\omega, \tag{5.121}$$

where Φ_μ denotes the characteristic function of μ. Similarly, if μ is a signed measure such that $I_g(\mu) < \infty$,

$$g(z,y) = \int e^{-i\langle z-y\,,\,\omega\rangle}\Lambda_c(\omega)\,\mathrm{d}\omega, \tag{5.122}$$

$$U_g^\mu(z) = \int e^{-i\langle z\,,\,\omega\rangle}\Phi_\mu(\omega)\Lambda_c(\omega)\,\mathrm{d}\omega, \tag{5.123}$$

$$I_g(\mu) = \int |\Phi_\mu|^2(\omega)\Lambda_c(\omega)\,\mathrm{d}\omega. \tag{5.124}$$

Proof. Since the Gaussian is its own Fourier transform, we can write

$$\exp\left(-t|x-y|^2\right) = \frac{1}{(4\pi t)^{d/2}}\int_{\mathbb{R}^d}\exp\left(i\langle x-y\,,\,\omega\rangle\right)\exp\left(-\frac{1}{4t}|\omega|^2\right)\,\mathrm{d}\omega. \tag{5.125}$$

Using Fubini and plugging Equation (5.125) into Equations (5.116) and (5.115), we directly obtain the expressions for $U^\mu_{K_\zeta}$ and U^μ_g. Next, using the fact that for any kernel K,

$$I_K(\mu) = \int U_K^\mu(z)\,\mathrm{d}\mu(z) \tag{5.126}$$

and using Fubini again, we get the expressions for $I_{K_\zeta}(\mu)$ and $I_g(\mu)$. ■

We give estimates on Λ_ζ in the next lemma, which we will use in the proof of Proposition 5.3.15.

Lemma 5.7.2

$$\int \Lambda_\zeta(\omega)\,\mathrm{d}\omega = n^{\zeta(d-2)}, \tag{5.127}$$

$$\int_{|\omega|>n^{\alpha\zeta}}\Lambda_\zeta(\omega)\,\mathrm{d}\omega \le C\exp\left(-\frac{1}{4}n^{\zeta(\alpha-1)}\right)\left(n^{\zeta(\alpha+1)(d/2-1)}+n^{\zeta(d-2)}\right) \tag{5.128}$$

for any $\alpha > 1$.

Proof. Equation (5.127) is easy to see upon noting that $K_\zeta(z,z) = n^{\zeta(d-2)}$ is the Fourier transform of Λ_ζ at 0 i.e. $\int \Lambda_\zeta(\omega)\,\mathrm{d}\omega$.

We turn to Equation (5.128), we have

$$\begin{aligned}\int_{|\omega|>n^{\alpha\zeta}}\Lambda_\zeta(\omega)\,\mathrm{d}\omega &= C\int_0^{+\infty} t^{-2}e^{-n^{-2\zeta}t}\int \mathbf{1}_{|\omega|>n^{\alpha\zeta}}\exp\left(-\frac{1}{4t}|\omega|^2\right)\,\mathrm{d}\omega\,\mathrm{d}t \\ &\le C\int t^{d/2-2}e^{-tn^{-2\zeta}}e^{-\frac{1}{4t}n^{2\alpha\zeta}}\,\mathrm{d}t.\end{aligned} \tag{5.129}$$

We split the integral in two parts, namely $\{t > n^{\zeta(\alpha+1)}\}$ and $\{t < n^{\zeta(\alpha+1)}\}$. In the first case $tn^{-2\zeta} > \frac{1}{4t}n^{2\alpha\zeta}$ so we keep the first exponential term and bound the second one by 1, and do the opposite in the second case. We thus get

$$\int_{|\omega|>n^{\alpha\zeta}} \Lambda_\zeta(\omega)\,\mathrm{d}\omega \leq C(I_1 + I_2) \tag{5.130}$$

with

$$I_1 = \int_{n^{\zeta(\alpha+1)}}^{+\infty} t^{d/2-2}e^{-tn^{-2\zeta}}\,\mathrm{d}t, \tag{5.131}$$

and

$$I_2 = \int_0^{n^{\zeta(\alpha+1)}} t^{d/2-2}e^{-\frac{1}{4t}n^{2\alpha\zeta}}\,\mathrm{d}t. \tag{5.132}$$

We first bound I_2. Setting $t = un^{\zeta(\alpha+1)}$, we get

$$\begin{aligned} I_2 &\leq n^{\zeta(\alpha+1)(d/2-1)} \int_0^1 u^{d/2-2} \exp\left(-\frac{1}{4u}n^{\zeta(\alpha-1)}\right)\,\mathrm{d}u \\ &\leq n^{\zeta(\alpha+1)(d/2-1)} \exp\left(-\frac{1}{4}n^{\zeta(\alpha-1)}\right) \int_0^1 u^{d/2-2}\,\mathrm{d}u \\ &\leq Cn^{\zeta(\alpha+1)(d/2-1)} \exp\left(-\frac{1}{4}n^{\zeta(\alpha-1)}\right). \end{aligned} \tag{5.133}$$

We now turn to I_1. Setting $t = un^{2\zeta}$, we get

$$I_1 \leq n^{\zeta(d-2)} \int_{n^{\zeta(\alpha-1)}}^{+\infty} u^{d/2-2}e^{-u}\,\mathrm{d}u. \tag{5.134}$$

The second term is, up to a finite constant, the tail probability $\mathbb{P}\left(X > n^{\zeta(\alpha-1)}\right)$ when X follows a Gamma distribution with parameter $d/2-1$. Recall that its Laplace transform is $\Psi_X(s) = \frac{1}{(1-s)^{d/2-1}}$ when $s < 1$. Using Markov's inequality, we get

$$\begin{aligned} I_1 &\leq Cn^{\zeta(d-2)}\mathbb{P}\left(e^{\frac{1}{2}X} > e^{\frac{1}{2}n^{\zeta(\alpha-1)}}\right) \\ &\leq 2^{d/2-1}Cn^{\zeta(d-2)} \exp\left(-\frac{1}{2}n^{\zeta(\alpha-1)}\right). \end{aligned} \tag{5.135}$$

■

We will also use the following boundedness property of Riesz potentials.

Lemma 5.7.3 Let μ be a probability measure with compact support and continuous density. Let $0 < \alpha \leq 2$, then the Riesz potential

$$A_\alpha^\mu(z) := \int \frac{1}{|y-z|^{d-\alpha}} \mathrm{d}\mu(y) \tag{5.136}$$

is uniformly bounded on $\mathbb{R}^d$, i.e. there exists a constant C_α such that, for all z, $A_\alpha^\mu(z) \leq C_\alpha$.

Proof. We first invoke the maximum principle from (Landkof, 1972, Theorem 1.10, page 71) stating that under the assumption $0 < \alpha \leq 2$, if $A_\alpha^\mu(z) \leq M$ holds for any z in the support of μ then this holds for all $z \in \mathbb{R}^d$. Under the assumption on μ (i.e compact support and continuous density), $z \mapsto A_\alpha^\mu(z)$ is continuous (Chafaï et al., 2014, Lemma 4.3) so that A_α^μ is bounded on the support of μ. Hence, using the maximum principle, there exists a constant $C_\alpha > 0$ such that $A_\alpha^\mu(z) \leq C_\alpha$ for all z. ■

5.7.2 Auxiliary lemmas for Proposition 5.3.14

Proof of Lemma 5.6.1. K_ζ is a bit too rough of a regularization to compare directly the two integral representations, so we introduce yet another auxiliary kernel,

$$K_\Lambda^s(z,y) := \frac{1}{\Gamma\left(\frac{d-2}{2}\right)} \int_0^{+\infty} t^{\frac{d-2}{2}-1} \exp\left(-n^{-2\Lambda}t^s\right) \exp\left(-t|z-y|^2\right) \tag{5.137}$$

with $\Lambda > 0$ and $0 < s < 1$. Note that K_ζ can be recovered as a limiting case, choosing $\Lambda = \zeta$ and $s = 1$. We first compare K_Λ^s to g and then K_ζ to K_Λ^s.

$$\begin{aligned}
&\left| \int g(z,y)\,\mathrm{d}\mu(y) - \int K_\Lambda^s(z,y)\,\mathrm{d}\mu(y) \right| \\
&= \frac{1}{\Gamma\left(\frac{d-2}{2}\right)} \left| \iint_0^{+\infty} \left(1 - \exp\left(-n^{-2\Lambda}t^s\right)\right) t^{\frac{d-2}{2}-1} \exp\left(-t|z-y|^2\right)\,\mathrm{d}t\,\mathrm{d}\mu(y) \right| \\
&\leq \frac{n^{-2\Lambda}}{\Gamma\left(\frac{d-2}{2}\right)} \iint_0^{+\infty} t^s t^{\frac{d-2}{2}-1} \exp\left(-t|z-y|^2\right)\,\mathrm{d}t\,\mathrm{d}\mu(y) \\
&= Cn^{-2\Lambda} A_{2(1-s)}^\mu(z),
\end{aligned} \tag{5.138}$$

with

$$A_{2(1-s)}^\mu(z) = \int \frac{1}{|z-y|^{d-2+2s}}\,\mathrm{d}\mu(y), \tag{5.139}$$

which is the Riesz potential of μ introduced in Lemma 5.7.3 with Riesz parameter $\alpha = 2(1-s)$. Here we see that it would have been indeed insufficient to take $s = 1$, since we would have lost integrability at 0. Using Lemma 5.7.3, we thus get

$$\sup_{z \in \mathbb{R}^d} \left| \int g(z,y)\,\mathrm{d}\mu(y) - \int K_\Lambda^s(z,y)\,\mathrm{d}\mu(y) \right| = \mathcal{O}\left(n^{-2\Lambda}\right). \tag{5.140}$$

In the same way, we have

$$\begin{aligned}&\Big|\int K_\zeta(z,y)\,\mathrm{d}\mu(y)-\int K_\Lambda^s(z,y)\,\mathrm{d}\mu(y)\Big|\\&\leq\frac{1}{\Gamma\left(\frac{d-2}{2}\right)}\iint_0^{+\infty}\left|\exp\left(-n^{-2\zeta}t\right)-\exp\left(-n^{-2\Lambda}t^s\right)\right|t^{\frac{d-2}{2}-1}\exp\left(-t|z-y|^2\right)\,\mathrm{d}t\,\mathrm{d}\mu(y).\end{aligned}\tag{5.141}$$

We see that the term inside the absolute value is nonnegative iff $t^{1-s}\leq n^{2(\zeta-\Lambda)}$ and we thus split the integral in two parts. First,

$$\begin{aligned}&\iint_0^{n^{\frac{2}{1-s}(\zeta-\Lambda)}}\exp\left(-n^{-2\zeta}t\right)\left(1-\exp\left(-(n^{-2\Lambda}t^s-n^{-2\zeta}t)\right)\right)t^{\frac{d-2}{2}-1}\exp\left(-t|z-y|^2\right)\,\mathrm{d}t\,\mathrm{d}\mu(y)\\&\qquad\leq n^{-2\Lambda}\int\int_0^{n^{\frac{2}{1-s}(\zeta-\Lambda)}}\exp\left(-n^{-2\zeta}t\right)t^{\frac{d-2}{2}+s-1}\exp\left(-t|z-y|^2\right)\,\mathrm{d}t\,\mathrm{d}\mu(y)\\&\qquad\leq Cn^{-2\Lambda}A^\mu_{2(1-s)}(z)\\&\qquad=\mathcal{O}(n^{-2\Lambda})\end{aligned}\tag{5.142}$$

uniformly in z, where we bounded $\exp\left(-n^{-2\zeta}t\right)\leq 1$ and the integral over t by the integral over $\mathbb{R}_+$, and we used Lemma 5.7.3 again. The remaining part of (5.141) is

$$\begin{aligned}&\iint_{n^{\frac{2}{1-s}(\zeta-\Lambda)}}^{+\infty}\exp\left(-n^{-2\Lambda}t^s\right)\left(1-\exp\left(-(n^{-2\zeta}t-n^{-2\Lambda}t^s)\right)\right)t^{\frac{d-2}{2}-1}\exp\left(-t|z-y|^2\right)\,\mathrm{d}t\,\mathrm{d}\mu(y)\\&\qquad\leq n^{-2\zeta}\int\int_{n^{\frac{2}{1-s}(\zeta-\Lambda)}}^{+\infty}\exp\left(-n^{-2\Lambda}t^s\right)t^{\frac{d-2}{2}}\exp\left(-t|z-y|^2\right)\,\mathrm{d}t\,\mathrm{d}\mu(y).\end{aligned}\tag{5.143}$$

We cannot rely on the same argument again since we do not have integrability in the associated Riesz potential anymore. However, we bound $\exp\left(-t|z-y|^2\right)\leq 1$ and we perform the change of variable $h=t^sn^{-2\Lambda}$ i.e. $t=h^{1/s}n^{2\Lambda/s}$, with Jacobian $h^{1/s-1}n^{2\Lambda/s}$. Bounding the integral in t by the integral over $\mathbb{R}_+$, we obtain

$$n^{-2\zeta+\frac{2\Lambda}{s}+\frac{(d-2)\Lambda}{s}}\int_0^{+\infty}\exp\left(-h\right)h^{\frac{d-2}{2s}+\frac{1}{s}-1}\,\mathrm{d}h=n^{-2\zeta+\frac{d\Lambda}{s}}\,\Gamma\left(\frac{d}{2s}\right).\tag{5.144}$$

Choosing for instance $\Lambda=\frac{2s\zeta}{d+2s}$, we finally obtain

$$\sup_{z\in\mathbb{R}^d}\left|\int K_\zeta(z,y)\,\mathrm{d}\mu(y)-\int K_\Lambda^s(z,y)\,\mathrm{d}\mu(y)\right|=\mathcal{O}(n^{-4s\zeta/(d+2s)}).\tag{5.145}$$

Our choice of Λ along with Equation (5.140) and Equation (5.145) then yields the result. ■

Proof of Lemma 5.6.2. We start with

$$\left|U_{K_\zeta}^{\mu_n}(z) - U_{K_\zeta}^{\mu_n^{(\epsilon)}}(z)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|K_\zeta(z,x_i) - \int K_\zeta(z,y)\,\mathrm{d}\lambda_{B(x_i,n^{-\epsilon})}(y)\right|$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left|\int \left\{K_\zeta(z,x_i) - K_\zeta(z,x_i+n^{-\epsilon}u)\right\}\,\mathrm{d}\lambda_{B(0,1)}(u)\right|. \quad (5.146)$$

We now Taylor expand $K_\zeta(z,.)$ around x_i at order 2. Respectively denoting by $\nabla_2, \mathrm{Hess}_2, \partial_2^{j,k}$ the gradient, Hessian and second order partial derivatives w.r.t the second coordinate of K_ζ, for each z and u, there exists $c_u \in [0,1]$ such that

$$\left|\int \left\{K_\zeta(z,x_i) - K_\zeta(z,x_i+n^{-\epsilon}u)\right\}\,\mathrm{d}\lambda_{B(0,1)}(u)\right|$$
$$= \left|\int \left\{n^{-\epsilon}\left\langle \nabla_2 K_\zeta(z,x_i)\,,\,u\right\rangle + \frac{1}{2}n^{-2\epsilon}\left\langle \mathrm{Hess}_2 K_\zeta(z,x_i+c_u n^{-\epsilon}u)u\,,\,u\right\rangle\right\}\,\mathrm{d}\lambda_{B(0,1)}(u)\right|. \quad (5.147)$$

We easily see with the change of variable $u \mapsto -u$ that the first term is always zero. Setting $x_u = x_i + c_u n^{-\epsilon}u$, it remains to bound the Hessian term

$$\frac{n^{-2\epsilon}}{2}\left|\sum_{1\leq j,k\leq d}\int \partial_2^{j,k}K_\zeta(z,x_u)u_j u_k\,\mathrm{d}\lambda_{B(0,1)}(u)\right|$$
$$\leq \frac{n^{-2\epsilon}}{2}\sum_{1\leq j,k\leq d}\int \left|\partial_2^{j,k}K_\zeta(z,x_u)\right||u_j u_k|\,\mathrm{d}\lambda_{B(0,1)}(u). \quad (5.148)$$

A quick computation for the second order derivatives yields

$$\partial_2^{j,j}K_\zeta(z,y) = \left(|z-y|^2+n^{-2\zeta}\right)^{-d/2}(d-2)\left\{d(z_j-y_j)^2\left(|z-y|^2+n^{-2\zeta}\right)^{-1} - 1\right\} \quad (5.149)$$

and

$$\partial_2^{j,k}K_\zeta(z,y) = (d-2)d(z_j-y_j)(z_k-y_k)\left(|z-y|^2+n^{-2\zeta}\right)^{-1-d/2} \quad (5.150)$$

for $j \neq k$. Bounding

$$|z_j-y_j||z_k-y_k| \leq \max\left(|z_j-y_j|^2, |z_k-y_k|^2\right), \quad (5.151)$$

we get that

$$|\partial_2^{j,k}K_\zeta(z,y)| \leq (d-2)(d+1)n^{d\zeta} \quad (5.152)$$

for all j,k. Hence,

$$\left|U_{K_\zeta}^{\mu_n}(z) - U_{K_\zeta}^{\mu_n^{(\epsilon)}}(z)\right| \leq \frac{1}{2}n^{-2\epsilon+d\zeta}(d-2)(d+1)\sum_{1\leq j,k\leq d}\int |u_j u_k|\,\mathrm{d}\lambda_{B(0,1)}(u). \quad (5.153)$$

■

Proof of Lemma 5.6.3. This is straightforward using Lemma 5.7.1. ■

Proof of Lemma 5.6.4. Equation (5.76) is exactly Equation (3.3), page 17 of Chafaï et al. (2018). Equation (5.77) is an easy generalization of Lemma 4.1, page 20 of Chafaï et al. (2018) with general β_n and the exact same proof. Note that the $\frac{1}{2}$ factor difference in front of the first term only comes from a different choice of normalization in the energies.

For Equation (5.78), recall that

$$H_n(x_1, \dots, x_n) = \frac{1}{2n^2} \sum_{i \neq j} g(x_i, x_j) + \frac{1}{n} \sum_{i=1}^{n} V(x_i)$$

and split the energy $H_n = \eta H_n + (1-\eta) H_n$. Then, we lower bound the first term using $g \geq 0$; this term is kept to ensure integrability under $\mathbb{P}^V_{n,\beta_n}$. The second term is lower bounded as follows. Use the superharmonicity of g to write

$$\frac{1}{n^2} \sum_{i \neq j} g(x_i, x_j) \geq \frac{1}{n^2} \sum_{i,j} \iint g(x,y) \, \mathrm{d}\lambda_{B(x_i,\epsilon_n)}(x) \, \mathrm{d}\lambda_{B(x_j,\epsilon_n)}(y) - I_g(\lambda_{B(0,1)}) \frac{1}{n \epsilon_n^{d-2}} \tag{5.154}$$

where the second term comes from the n diagonal terms that are exceeding in the first term, and the fact that

$$I_g(\lambda_{B(0,n^{-\epsilon})}) = \frac{1}{\epsilon_n^{d-2}} I_g(\lambda_{B(0,1)}). \tag{5.155}$$

Then, we bound

$$V(x_i) - \int V \, \mathrm{d}\lambda_{B(x_i,\epsilon_n)} \tag{5.156}$$

by Taylor expanding V at order 2. This gives a uniform bound in $\mathcal{O}(\epsilon_n^2)$ under the assumptions. Grouping everything together and writing down the definition of $I_g(\mu_n^{(\epsilon)})$, we obtain Equation (5.78). All of this is done in detail in (Chafaï et al., 2018), Lemma 4.2 page 21 and steps 1 and 2 of the proof of Theorem 1.9 pages 21-23, and all the computations stay the same.

Finally, Equation (5.80) is obtained by combining the previous ones with our choice of parameter $\eta = c' n / \beta_n$. Note that in (Chafaï et al., 2018), $\eta = n^{-1}$ since the authors work in the scaling $\beta_n = \beta n^2$. Again, the computations stay the same (see step 3 of the proof of Theorem 1.9 of (Chafaï et al., 2018), pages 23-24) but we give a glimpse of it

for the sake of completeness.

$$\begin{aligned}\mathbb{P}^V_{n,\beta_n}(A) &= \frac{1}{Z^V_{n,\beta_n}} \int \exp\left(-\beta_n H_n(x_1,\dots,x_n)\right) \mathrm{d}x_1 \dots \mathrm{d}x_n \\ &\leq \frac{1}{Z^V_{n,\beta_n}} \exp\left(-(1-\eta)\beta_n \inf_A I^V_g(\mu_n^{(\epsilon)}) + C\beta_n(1-\eta)\left(\frac{1}{n\epsilon_n^{d-2}} + \epsilon_n^2\right)\right) \\ &\int \exp\left(-\beta_n \eta \frac{1}{n}\sum_{i=1}^n V(x_i)\right) \mathrm{d}x_1 \dots \mathrm{d}x_n. \end{aligned} \tag{5.157}$$

Due to the choice $\eta = c'n/\beta_n$, the remaining integral factorizes as C^n with

$$C = \int \exp\left(-c'V\right) < +\infty. \tag{5.158}$$

Using Equation (5.77) then yields the result after splitting the leading order term of (5.77) into $(1-\eta)I^V_g(\mu_V) + \eta I^V_g(\mu_V)$, replacing η by its expression and using (5.76). Note also that $\epsilon = 1/d$ in (Chafaï et al., 2018) but it is useful for us to keep freedom on this choice at this stage. ■

5.7.3 Auxiliary proofs for Proposition 5.3.15

Proof of Lemma 5.6.5. We begin as in Lemma 5.6.2, writing

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n (W\star\rho_\eta)(x_i)\int K_\zeta(z,y)\,\mathrm{d}\lambda_{B(x_i,n^{-\epsilon})}(y) - \frac{1}{n}\sum_{i=1}^n \int (W\star\rho_\eta)(y) K_\zeta(z,y)\lambda_{B(x_i,n^{-\epsilon})}(y) \\ &= \frac{1}{n}\sum_{i=1}^n \int \left\{(W\star\rho_\eta)(x_i) - (W\star\rho_\eta)(x_i+n^{-\epsilon}u)\right\} K_\zeta(z, x_i+n^{-\epsilon}u)\,\mathrm{d}\lambda_{B(0,1)}(u).\end{aligned} \tag{5.159}$$

A quick computation yields

$$|\partial_j(W\star\rho_\eta)(x)| \leq \frac{1}{(2\pi\eta)^{d/2}\eta^2}\int W(y)|x_j - y_j| \exp\left(-\frac{1}{2\eta^2}|x-y|^2\right) \mathrm{d}y. \tag{5.160}$$

Using the fact that W is bounded and performing the change of variable $y \mapsto x - \eta u$, we get

$$|\partial_j(W\star\rho_\eta)(x)| \leq \|W\|_\infty \frac{1}{\eta}\frac{1}{(2\pi)^{d/2}}\int |y| \exp\left(-|y|^2/2\right) \mathrm{d}y. \tag{5.161}$$

Taylor expanding $W\star\rho$ as in Lemma 5.6.2 at order one around x_i, and using the bound on the first-order derivative, yields

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n \int \left\{(W\star\rho_\eta)(x_i) - (W\star\rho_\eta)(x_i+n^{-\epsilon}u)\right\} K_\zeta(z,x_i+n^{-\epsilon}u)\,\mathrm{d}\lambda_{B(0,1)}(u)\right| \\ &\leq \frac{1}{\eta}\mathcal{O}\left(n^{-\epsilon+\zeta(d-2)}\right)\end{aligned} \tag{5.162}$$

where the $\mathcal{O}$ does not depend on η. Equation (5.159) then gives the result. ■

Proof of Lemma 5.6.6. Using Lemma 5.7.1, we first write

$$\sup_{z\in\mathbb{R}^d}\left|U_{K_\zeta}^{\nu_n^{\epsilon,\eta}}(z)-U_{K_\zeta}^{(W\star\rho_\eta).\mu_V'}(z)\right|=\left|\int e^{-i\langle z,\omega\rangle}\Phi_{\nu_n^{\epsilon,\eta}-(W\star\rho_\eta).\mu_V'}(\omega)\Lambda_\zeta(\omega)\,\mathrm{d}\omega\right| \tag{5.163}$$

Since $\nu_n^{\epsilon,\eta}$ has density

$$x\mapsto\frac{1}{n}\sum_{i=1}^n\frac{1}{|B(x_i,n^{-\epsilon})|}(W\star\rho_\eta)(x)\mathbf{1}_{x\in B(x_i,n^{-\epsilon})}, \tag{5.164}$$

we have

$$\left|\Phi_{\nu_n^{\epsilon,\eta}}-\Phi_{(W\star\rho_\eta).\mu_V'}\right|=\left|\overline{\mathcal{F}(W\star\rho_\eta)}\star\left(\Phi_{\mu_n^{(\epsilon)}}-\Phi_{\mu_V}\right)\right|, \tag{5.165}$$

so that

$$\sup_{z\in\mathbb{R}^d}\left|U_{K_\zeta}^{\nu_n^{\epsilon,\eta}}(z)-U_{K_\zeta}^{(W\star\rho_\eta).\mu_V'}(z)\right|\le A_1+A_2 \tag{5.166}$$

with

$$A_1:=\int\int_{|x|<n^{\zeta/s}}|\mathcal{F}(W\star\rho_\eta)(\omega-x)|\left|\Phi_{\mu_n^{(\epsilon)}}-\Phi_{\mu_V}\right|(x)\,\mathrm{d}x\,\Lambda_\zeta(\omega)\,\mathrm{d}\omega, \tag{5.167}$$

$$A_2:=\int\int_{|x|>n^{\zeta/s}}|\mathcal{F}(W\star\rho_\eta)(\omega-x)|\left|\Phi_{\mu_n^{(\epsilon)}}-\Phi_{\mu_V}\right|(x)\,\mathrm{d}x\,\Lambda_\zeta(\omega)\,\mathrm{d}\omega. \tag{5.168}$$

for some $0<s<1$.

We first bound A_2. We bound $\left|\Phi_{\mu_n^{(\epsilon)}}-\Phi_{\mu_V}\right|(x)\le 2$. Using

$$\begin{aligned}|\mathcal{F}(W\star\rho_\eta)(\omega-x)|&=|\mathcal{F}(W)(\omega-x)\mathcal{F}(\rho_\eta)(\omega-x)|\\&\le C\|W\|_{L^1(\mathbb{R}^d)}\exp(-\eta^2|\omega-x|^2/2),\end{aligned} \tag{5.169}$$

we obtain

$$A_2\le 2C\|W\|_{L^1(\mathbb{R}^d)}\int\int_{|x|>n^{\zeta/s}}\exp\left(-\eta^2|\omega-x|^2/2\right)\,\mathrm{d}x\,\Lambda_\zeta(\omega)\,\mathrm{d}\omega. \tag{5.170}$$

For $1<\alpha<1/s$, we split again the first integral in $|\omega|<n^{\alpha\zeta}$ and $|\omega|>n^{\alpha\zeta}$. When $|\omega|>n^{\alpha\zeta}$, we bound the exponential integral by the integral over the whole domain which is bounded by $\frac{C}{\eta^{d/2}}$. We then use the estimate 5.128 to bound the remaining term. In the case $|\omega|<n^{\alpha\zeta}$, we then have $|\omega-x|>n^{\zeta/s}-n^{\alpha\zeta}$, and we can use the decay of

the Gaussian tail since $\alpha < 1/s$, along with 5.127. Doing so, we get

$$\begin{aligned} A_2 &\leq C\|W\|_{L^1(\mathbb{R}^d)} \frac{1}{\eta^{d/2}} \int_{|\omega|>n^{\alpha\zeta}} \Lambda_\zeta(\omega)\,\mathrm{d}\omega \\ &+ \int \int_{|\omega-x|>n^{\zeta/s}-n^{\alpha\zeta}} \exp\left(-\eta^2|\omega-x|^2/2\right)\,\mathrm{d}x\,\Lambda_\zeta(\omega)\,\mathrm{d}\omega \\ &\leq \frac{C\|W\|_{L^1(\mathbb{R}^d)}}{\eta^{d/2}} \exp\left(-\frac{1}{4}n^{\zeta(\alpha-1)}\right)\left(n^{\zeta(\alpha+1)(d/2-1)}+n^{\zeta(d-2)}\right) \\ &+ \frac{C\|W\|_{L^1(\mathbb{R}^d)}}{\eta^{d/2}} n^{\zeta(d-2)} \exp\left(-\eta^2\left(n^{\zeta/s}-n^{\alpha\zeta}\right)^2/2\right). \end{aligned} \tag{5.171}$$

We now turn to bounding A_1. Recall that

$$\Phi_{\mu_n^{(\epsilon)}-\mu_V}(x) = \int e^{i\langle y\,,\,x\rangle}\,\mathrm{d}(\mu_n^{(\epsilon)}-\mu_V)(y). \tag{5.172}$$

Since $y \mapsto e^{i\langle y\,,\,x\rangle}$ is $|x|$-Lipschitz and bounded by 1,

$$\sup_{|x|<n^{\zeta/s}} \left|\Phi_{\mu_n^{(\epsilon)}-\mu_V}(x)\right| \leq n^{\zeta/s} \sup_{\|f\|_{Lip}\leq 1,\,\|f\|_\infty\leq 1} \left|\int f\,\mathrm{d}(\mu_n^{(\epsilon)}-\mu_V)\right| = n^{\zeta/s} d_{\mathrm{BL}}(\mu_n^{(\epsilon)},\mu_V). \tag{5.173}$$

Moreover, we have

$$\begin{aligned} \|\mathcal{F}(W\star\rho_\eta)\|_{L^1(\mathbb{R}^d)} &= \|\mathcal{F}(W)\mathcal{F}(\rho_\eta)\|_{L^1(\mathbb{R}^d)} \\ &\leq \|\mathcal{F}(W)\|_{L^2(\mathbb{R}^d)}\|\mathcal{F}(\rho_\eta)\|_{L^2(\mathbb{R}^d)} \\ &= \|W\|_{L^2(\mathbb{R}^d)}\|\rho_\eta\|_{L^2(\mathbb{R}^d)} \\ &= \frac{C}{\eta^{d/2}}\|W\|_{L^2(\mathbb{R}^d)} \end{aligned} \tag{5.174}$$

where we used Cauchy–Schwarz and Plancherel theorem. Thus, using Equation (5.127),

$$\begin{aligned} A_1 &\leq \sup_{|x|<n^{\zeta/s}} \left|\Phi_{\mu_n^{(\epsilon)}-\mu_V}(x)\right| \|\mathcal{F}(W\star\rho_\eta)\|_{L^1(\mathbb{R}^d)} \int \Lambda_\zeta(\omega)\,\mathrm{d}\omega \\ &\leq \frac{C}{\eta^{d/2}}\|W\|_{L^2(\mathbb{R}^d)} n^{\zeta(d-2+1/s)} d_{\mathrm{BL}}(\mu_n^{(\epsilon)},\mu_V), \end{aligned} \tag{5.175}$$

which concludes the proof of the lemma. ■

5.8 Conclusion

In this chapter, we have shown that a tractable random approximation of the kernel embedding can be incorporated into large deviations results. This support the use of

Gibbs measures to get tighter coverage in Monte Carlo integration as in Chapter 4, even when the kernel embedding is approximated using e.g. MCMC. Some questions deserve further investigation.

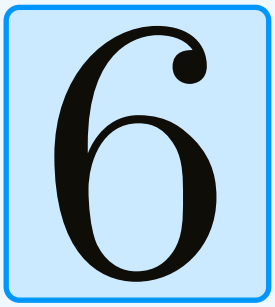
First, for exhaustiveness, generalizing Theorem 5.3.13 to dimensions 1 and 2 could be of interest. The difficulty is that the natural replacement for the regularized Coulomb kernel (5.15) in dimensions 1 and 2 is not positive semidefinite.

A second interesting question lies within the choice of the size M_n of the background ν_n. We delay this discussion to the next chapter, in which we conclude this thesis and discuss further work.

Conclusion and further work

" — What be our heading, captain?
— We shall follow the stars, master Gibbs!"

Some pirate in the Caribbean



6.1 Further work 160

6.1.1 Investigating background quality 160
6.1.2 A fast CLT for bounded kernels 163
6.1.3 Sampling from Gibbs distributions 167

6.2 Openings 169

6.2.1 Smale's 7th problem 169
6.2.2 Integration in RKHSs 170
6.2.3 Gibbs measures and parallel MCMC 170
6.2.4 Mean-field theory for neural networks 172

In this thesis, we introduced numerical integration based on Gibbs distributions.

This quadrature rule can be seen as a probabilistic relaxation of existing minimization algorithms for the MMD: our Gibbs distribution favors configurations of points with small MMD, with a strength given by the inverse temperature β_n.

In Chapter 4, we gave a concentration inequality for the MMD with points drawn from our Gibbs measure. While not improving the rate on the MMD, this bound comes with faster tail decay than MCMC with the same number of points, depending on the inverse temperature β_n. This implies that for a fixed integration error, our quadrature comes with tighter confidence regions, which is desirable when the goal is to keep the number of integrand evaluations low. Yet, any existing algorithm to draw this joint distribution on the nodes $x_1, \dots, x_n$ relies on an intractable integral with respect to the target π. We have shown in Chapter 5 that one can approximate this integral without loss on the statistical guarantees for Monte Carlo integration, at least asymptotically. Finally, we illustrated the improvement obtained with points approximately drawn from such Gibbs distributions over MCMC with the same number of points, for instance in toy Bayesian classification examples in Chapter 5.

There are limitations for the whole quadrature pipeline that remain to be addressed. We discuss them in the following section, along with some ideas on how we could tackle them.

6.1 Further work

6.1.1 Investigating background quality

Quenched concentration

In Chapter 5, we provided a large deviation principle for the quenched Gibbs measure $\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}$ defined conditionally on a background measure ν_n. A first important follow-up question is to quantify how the choice of this background affects the final quadrature.

In fact, our Theorem 5.3.7 for bounded kernels holds for any background ν_n converging to π. Yet, Theorem 5.3.7 is only asymptotic, and does not bring to light the impact of the quality of ν_n. In practice, we saw in Section 5.4.3 that the choice of the ν_n has an effect. For instance backgrounds ν_n relying on M_n MCMC samples provide better results when M_n is higher.

Quantifying this impact at fixed n would require a concentration inequality in the spirit of Theorem 5.3.7. Under the assumptions of Corollary 4.3.11, meaning essentially that π has compact support and K is bounded, we can actually show the following crude bound.

Corollary 6.1.1 Let $\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}$ be the quenched Gibbs measure of Definition 5.3.5, with K satisfying Assumption 5.3.3. Let π satisfy Assumption 4.3.9.

Let $n \geq 2$ and assume that $\beta_n \gg n$ and that

$$r \geq 4 \max\left(\sqrt{c_2} n^{-1/2}, \sqrt{c_1}(\beta_n/n)^{-1/2}\right), \tag{6.1}$$

where c_1, c_2 are defined in Corollary 4.3.11. For $\mathbb{P}$-almost all ω,

$$\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}\left(\mathrm{MMD}_K\left(\mu_n^{\mathrm{Q}}, \pi\right) > r; \nu_n(\omega)\right) \leq \exp\left(-\frac{1}{8}\beta_n r^2 + 2\beta_n\sqrt{C}\,\mathrm{MMD}_K(\nu_n(\omega), \pi)\right), \tag{6.2}$$

where C is the constant of Assumption 4.3.1.

Proof. For the ease of reading, we recall the definitions of V^π and V_n for bounded kernels.

$$V^\pi(z) = \begin{cases} -U_K^\pi(z) \text{ when } z \in B(0,R) \\ -U_K^\pi(z) + \Phi(z) \text{ otherwise} \end{cases}, \tag{6.3}$$

$$V_n(z) = \begin{cases} -U_K^{\nu_n}(z) \text{ when } z \in B(0,R) \\ -U_K^{\nu_n}(z) + \Phi(z) \text{ otherwise} \end{cases}, \tag{6.4}$$

with

$$U_K^\mu(z) = \int K(z,y)\,\mathrm{d}\mu(y) \tag{6.5}$$

for any probability measure μ.

We then just use the fact that for non-singular kernels,

$$\begin{aligned} \sup_{z\in\mathbb{R}^d} |V_n(z) - V^\pi(z)| &= \sup_{z\in\mathbb{R}^d} |U_K^{\nu_n}(z) - U_K^\pi(z)| \\ &\leq \sqrt{C} \sup_{\|f\|_{\mathcal{H}_K}\leq 1} \left| \int f \mathrm{d}(\nu_n - \pi) \right| \\ &= \sqrt{C}\,\mathrm{MMD}_K(\nu_n, \pi), \end{aligned} \tag{6.6}$$

since $K(z,\cdot)$ is in $\mathcal{H}_K$ with norm $\sqrt{K(z,z)} \leq \sqrt{C}$ for all z. Bounding

$$\left| \frac{1}{n}\sum_{i=1}^n V^\pi(x_i) - \frac{1}{n}\sum_{i=1}^n V_n(x_i) \right| \leq \sqrt{C}\,\mathrm{MMD}_K(\nu_n(\omega), \pi) \tag{6.7}$$

both in the normalization constant $Z^{V_n}_{n,\beta_n}$ and in the energy H_n^{Q} of $\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}$, we thus have

$$\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}(A; \nu_n(\omega)) \leq \exp\left(2\beta_n\sqrt{C}\,\mathrm{MMD}_K(\nu_n(\omega), \pi)\right) \mathbb{P}^{V^\pi}_{n,\beta_n}(A), \tag{6.8}$$

for any Borel set A. Using Corollary 4.3.11 then gives the result. ∎

If ν_n is the empirical measure of the history of a Markov chain of size M_n, we know that

$$\mathrm{MMD}_K(\nu_n, \pi) \in \mathcal{O}(M_n^{-1/2}) \tag{6.9}$$

with high probability, see (2.72). Therefore, assuming for instance that $\beta_n \geq cn^2$ for some constant $c > 0$, there exists constants u_0, u_1 such that for any $r > u_0(n^{-1/2} \vee M_n^{-1/4})$, with high probability,

$$\mathbb{P}^{\mathrm{Q}}_{n,\beta_n}\left(\mathrm{MMD}_K\left(\mu_n^{\mathrm{Q}}, \pi\right) > r\right) \leq \exp\left(-u_1 \beta_n r^2\right). \tag{6.10}$$

When $M_n = n$ for instance, this suggests a downgrade in the constraint on r compared to the concentration result from Corollary 4.3.11 with exact knowledge of the potential U_K^π, while there is no downgrade when using a larger initial Markov chain $M_n \geq n^2$. In line with the literature on thinning algorithms (Section 2.6.4), this seems to point out that the background should be a significantly longer Markov chain than the cardinality of the final quadrature, e.g. n^2 for kernel thinning.

However, from a practical point of view, each evaluation of the log-density of the quenched Gibbs measure (5.3.5) with $M_n = n^2$ would then require $\mathcal{O}(n^2 + nM_n) = \mathcal{O}(n^3)$ operations. Thus, sampling with e.g. MALA would then require $\mathcal{O}(n^3)$ operations per iteration. Note again that the issue would be the same for any MMD minimization algorithm using MCMC to approximate the potential U_K^π with n^2 points (Section 2.6). To circumvent this issue, we could investigate different backgrounds than the ones we used in Section 5.4, with variance reduction in order to lower M_n and the cost per iteration.

Still, these recommendations are to be taken with a pinch of salt as our bound (6.2) might be crude and calls for refinements. In particular, obtaining a lower bound in (6.2), would be interesting to understand whether the dependence in the background we just mentioned could be improved. In order to do so, we can for instance take ideas from the lower bounds techniques used in large deviations, for instance in (Chafaï et al., 2014) for Gibbs measures.

Obtaining non-asymptotic results like (6.2) for the singular Coulomb interaction as in Theorem 5.3.13 would also be of interest, but much more difficult as always. Moreover, recall that in this case we need to choose ν_n to be the realization of a Gibbs measure with Coulomb interaction.

Annealed large deviations

Alternately, the impact of the quality of ν_n could also be quantified through *annealed* large deviations. Roughly, the annealed Gibbs measure is the quenched Gibbs measure (5.3.5) averaged over the law of the background ν_n. Annealed large deviations then correspond to large deviations (Definition 3.1.6) for the annealed Gibbs measure. In that case, there should be an interplay between the speed β_n in the LDP for the quenched measure and the speed of convergence of ν_n to π. In fact, assuming that ν_n satisfies a LDP with rate

function J and speed a_n, we have the heuristics

$$\left(\mathbb{P}^{\mathrm{Q}}_{n,\beta_n} \otimes \mathbb{P}\right)\left(\mu^{\mathrm{Q}}_n \approx \mu, \nu_n \approx \nu\right) \approx \exp\left(-\beta_n\{I_K^{V^\pi}(\mu) - I_K^{V^\pi}(\pi)\}\right)\exp\left(-a_n J(\nu)\right) \quad (6.11)$$

as $n \to +\infty$, following Theorem 5.3.7. This suggests a LDP with speed β_n, only when $a_n = \mathcal{O}(\beta_n)$. For instance, if ν_n is the empirical measure of a Markov chain of size M_n, $a_n = M_n$ (Dembo et al., 2010), which would suggest the choice $M_n = \mathcal{O}(\beta_n)$. When $\beta_n \geq n^2$, this is worse than the choice $M_n = n^2$ suggested by (6.10), maybe because (6.11) would also be a stronger result.

Again, choosing a ν_n that already satisfies a LDP at speed β_n should help in theory to reduce the background size M_n. To my knowledge, there are however no results in the literature concerning such annealed large deviations for Gibbs measures in general dimensions, so I think this would be harder to prove than a sharp quenched concentration inequality.

6.1.2 A fast CLT for bounded kernels

This section is the object of a paper in preparation.

Motivations

In Chapter 4, we provided a concentration inequality for the MMD, i.e. the worst-case integration error in the RKHS $\mathcal{H}_K$. Giving guarantees on the MMD is the main objective of all the methods presented in Section 2.6. Yet, we saw in Chapter 2 that, in the case of randomized algorithms, we can rather try to bound the root mean squared error

$$\sqrt{\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f \mathrm{d}\pi\right|^2\right]}, \quad (6.12)$$

for all $f \in \mathcal{F}$. Recall from Chapter 2 that we have the decomposition

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f \mathrm{d}\pi\right|^2\right] =$$
$$\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(x_i)\right] - \int f \,\mathrm{d}\pi\right)^2 + \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n f(x_i)\right], \quad (6.13)$$

and that this objective function is always weaker than minimizing the worst-case error in $\mathcal{F}$, which is why the optimal rates are faster.

For instance, we saw that the bias term in (6.13) is $\mathcal{O}(n^{-2})$ for MCMC with an ergodicity assumption and for bounded functions (2.69), while the variance term is $\mathcal{O}(n^{-1})$ from the CLT

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f \,\mathrm{d}\pi\right) \to \mathcal{N}(0, \sigma_f^2) \quad (6.14)$$

on $L^{2+\epsilon}(\pi)$ functions (Theorem 2.3.4). Thus, MCMC reaches the optimal $\mathcal{O}(n^{-1/2})$ mean squared error (6.12) on L^2 functions, assuming for instance that π has compact support. CLTs like (6.14) are very desirable for inference, since they provide asymptotic confidence regions.

We saw in Section 2.4.2 that this $\mathcal{O}(n^{-1/2})$ rate can be improved using control variates to achieve the optimal $n^{-1/2-k/d}$ for the mean squared error (6.12) on C^k functions. When $k = 1$, determinantal point processes provide zero bias in (6.13) and come up with CLTs like (6.14) at speed $n^{-1-1/d}$ for C^1 functions, or the optimal $n^{-1-2/d}$ if we assume that the integration domain is a complex compact manifold (Lemoine et al., 2024).

On the other hand, when f is assumed to belong to the RKHS $\mathcal{H}_K$, no optimal bounds are known to my knowledge. Still, we can hope to improve on the $n^{-1/2}$ rate from MCMC. I believe the only result in this side of the literature is obtained for kernel thinning, coming under the form

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int f \mathrm{d}\pi \right| = \mathcal{O}\left(n^{-1} \sqrt{\log n} \right) \tag{6.15}$$

with high probability and for any $f \in \mathcal{H}_K$, see Theorem 2.6.4. This is very satisfactory, since this is a quadratic improvement over the MCMC rate, which we observed on experiments in Section 5.4.

We saw in Section 4.4 that our Gibbs measure also yields variance reduction. Our goal is thus to obtain a CLT for the Gibbs measure $\mathbb{P}_{n,\beta_n}^{V^\pi}$ (5.1) with bounded kernel K, taking the form

$$\sqrt{\beta_n} \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int f \, \mathrm{d}\pi \right) \to \mathcal{N}(0, \sigma_f^2), \tag{6.16}$$

for any $f \in \mathcal{H}_K$, and up to the inverse temperature regime $\beta_n = \mathcal{O}(n^2)$. I think this would be an important result to support uncertainty quantification with Gibbs measures.

Let me give some heuristics and discussion that make me conjecture that $\mathbb{P}_{n,\beta_n}^{V^\pi}$ is tailored to such CLTs. First the large deviations principle (5.8) or the concentration inequality from Corollary 4.3.11 imply that, for $r > 0$ and n big enough,

$$\mathbb{P}_{n,\beta_n}^{V^\pi} \left(\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int f \, \mathrm{d}\pi \right| > r \right) \leq \exp\left(-\frac{1}{2\|f\|_{\mathcal{H}_K}^2} \beta_n r^2 \right), \tag{6.17}$$

for any $f \in \mathcal{H}_K$, due to the definition of the MMD in (2.19). The right-hand side in (6.17) is the tail of a Gaussian random variable with variance

$$v_{f,n} = \beta_n^{-1} \|f\|_{\mathcal{H}_K}^2. \tag{6.18}$$

The CLT (6.16) just corresponds to the matching lower bound as $n \to \infty$. While the constants in (6.18) may not be sharp, I still expect the rate $\sqrt{\beta_n}$ to hold due to the LDP lower bound.

Second, we already saw instances of CLTs like (6.16) for the one and two-dimensional logarithmic interaction in (3.34) and (3.59) at inverse temperature $\beta_n = \beta n^2$. Of course, this interaction is singular and does not define a RKHS. Instead, this requires f to be e.g. $\mathcal{C}^1$ for $d = 1$, while f is assumed to be at least $\mathcal{C}^2$ for $d = 2$. Note that, for this smoothness, we know from (2.40) that the $\beta_n^{-1/2} = n^{-1}$ rate is actually optimal. For higher dimensions $d \geq 3$, Gaussian fluctuations have been studied in (Serfaty, 2023) for the Coulomb interaction (3.43), but no complete CLT has been obtained for now. Still, we expect the CLT to be valid for sufficiently smooth functions such as $f \in \mathcal{C}^4$ and with $\beta_n = \beta n^{1+2/d}$ (Serfaty, 2023, Theorem 5 and Corollary 2.4). In that case, we know from (2.40) that this would be the optimal rate for $\mathcal{C}^1$ functions on a compact support in dimension d.

While the Coulomb interaction is well tailored to control integration errors for C^1 functions (see Theorem 3.3.5 and Equation (5.9)), Gibbs measures with bounded kernel K are suited to bound integration errors for functions in $\mathcal{H}_K$, hence the assumption that $f \in \mathcal{H}_K$ in (6.16).

Another important feature is in the choice of β_n in (6.16) that can be justified again both from the literature in statistical physics and numerical integration. On the one hand, while the Coulomb interaction already yields optimal worst-case integration errors for Lipschitz functions (Theorem 3.3.5), we saw in Corollary 4.3.11 that bounded interactions yield suboptimal but *universal* $n^{-1/2}$ rates for the MMD. By *universal*, we mean that the rate does not depend on the bounded kernel K we choose (or on the external confinement V). For CLTs like (6.16), we also expect the rate to be again suboptimal but universal for our Gibbs measures with bounded interactions: this was already the case with kernel thinning in (6.15).

On the other hand, since we know from (2.40) that the optimal rate for Sobolev spaces is $\mathcal{O}(n^{-1/2-k/d})$, and that this is a RKHS for $k > d/2$, this means that we cannot hope to achieve smaller than $\mathcal{O}(n^{-1})$ universal rates for the mean squared error (6.12) in RKHSs. Thus, that we should not hope for (6.16) to hold for $\beta_n > n^2$. The best we can hope for is then to match the rate of (6.15) up to a log term in the CLT: this is the fastest possible universal rate for the mean squared error in RKHSs, and already a quadratic improvement over MCMC. From the statistical physics literature, this can be understood from the fact that we expect *crystallization* at low temperature, in which case the Gaussian fluctuations disappear, see for instance (Serfaty, 2023, Theorem 6) and (Leblé and Serfaty, 2017) for the case of singular interactions.

One way to achieve faster kernel-dependent rates would be to break universality. For instance, instead of uniform weights $w_i = 1/n$ on each point, we could consider optimal weights as in continuous volume sampling (Section 2.6.3). In that case, there is however

little hope for a CLT, but we could already try to improve our bound on the MMD coming from the concentration inequality of Corollary 4.3.11. In fact, the same discussion applies: since the optimal worst-case error is $\mathcal{O}(n^{-k/d})$ for Sobolev spaces which are RKHSs for $k > d/2$, there is no hope for a faster universal rate than the MCMC $n^{-1/2}$ rate for the MMD. In my opinion, this is why we are unable to prove faster rates for most methods presented in Section 2.6, while methods relying on optimal weights like volume sampling provably achieve optimal rates.

As a last note, we should also investigate the bias term in (6.13). The CLTs we mentioned above for singular interactions usually come with quantification of the bias, and we can hope that this will still be the case with bounded kernels.

Tools

We quickly give a glimpse of our approach. Usually, CLTs for Gibbs measures are proved by studying the Laplace transform of the left-hand side of (6.16) and proving that it converges to that of a Gaussian distribution with suitable variance. In particular, this Laplace transform is expressed as a ratio of the normalization constant Z^V_{n,β_n} for varying V as pioneered by (Johansson, 1998).

To analyze the convergence, we then need to give fine asymptotic expansions for this quantity, called the partition function. Those are very subtle mathematical results, which require understanding the system at microscopic scales. This has been achieved for instance in (Leblé et al., 2018; Serfaty, 2023) by relying on an "electric formulation" for the energy, which is very specific to the Coulomb or Riesz interactions. In our case with general bounded interactions in any dimension, we do not hope to be able to work out such expansions, and we rather turn to more macroscopic tools.

As a refinement to existing CLTs, G. Lambert et al. (2019) and Angst et al. (2024) have obtained quantitative CLTs in (6.16) for β-ensembles (3.32), coming up with bounds of the form

$$\mathrm{d}\left(X_n, \mathcal{N}(0, \sigma_f^2)\right) \leq C n^{-\alpha} \tag{6.19}$$

for some $\alpha > 0$, and where

$$X_n = \sqrt{\beta_n}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int f \,\mathrm{d}\pi\right), \tag{6.20}$$

and d is for instance the Wassertein or total variation distance. Actually, (6.19) is strictly stronger than the CLT regardless of $\alpha > 0$, and our line of attack will thus be to give a quantitative result like (6.19) for bounded kernels to conclude on the CLT.

Obtaining results like (6.19) is done using Stein's method. While the latter is valid for any Gibbs measure and any dimension, an important mathematical step to establish (6.19) is to invert the operator

$$L_{n,\beta_n} = \Delta - \beta_n \nabla H_n \cdot \nabla \tag{6.21}$$

which is the generator of the Langevin diffusion with stationary measure $\mathbb{P}^V_{n,\beta_n}$. Since this operator depends on n, one must be very careful to isolate the dependencies in n for the convergence. This has been done in (G. Lambert et al., 2019) for β-ensembles, i.e. in dimension 1 and with singular logarithmic interaction, at the price of heavy technicalities.

We rather follow the methodology introduced in (Angst et al., 2024). In the latter, we still rely on Stein's method but avoid inversion of L_{n,β_n}. Let us start from X_n in (6.20), and assume that $K(x,y) = \Psi(x-y)$ is translation invariant. Assuming that there exists a function h such that

$$f = \Theta_V h, \tag{6.22}$$

where

$$(\Theta_V h)(x) = -\nabla V(x)\cdot\nabla h(x) - \int \nabla\Psi(x-y)\cdot(\nabla h(x) - \nabla h(y))\,\mathrm{d}\mu_V(y), \tag{6.23}$$

Angst et al. (2024) show that we can decompose

$$L_{n,\beta_n} F = X_n + Z \tag{6.24}$$

where F is the linear statistic

$$F(x_1,\dots,x_n) = \frac{1}{n}\sum_{i=1}^n h(x_i), \tag{6.25}$$

and Z is a remainder.

Then, the goal is to bound the remainder Z, and perform Stein method on the term $L_{n,\beta_n}F$ for which we readily know the pre-image by L_{n,β_n}. Using (6.24) however requires building an h in (6.25) by inverting the operator Θ_V in (6.23) e.g. on $\mathcal{H}_K$. Yet, Θ_V is this time independent of n, and so this is much easier to do than inverting L_{n,β_n} while keeping track of the dependence in n. If we are able to do so and show that Z goes to zero, we thus get the CLT (6.16).

This has been implemented in (Angst et al., 2024) in the one-dimensional case for β-ensembles, yet the general methodology is again valid for all dimensions and all interactions. Bounding Z is usually done via rigidity estimates like concentration inequalities, and we believe that our Corollary 4.3.11 is enough for that purpose. Inverting Θ_V is usually more involved, since it has been studied only for singular interactions. Yet, there seems to be no issue in our case with bounded kernels, upon assuming for instance that f belongs to $\mathcal{H}_K$, and that it is supported on the interior of the support of the equilibrium measure μ_V, or π in our case.

6.1.3 Sampling from Gibbs distributions

In my opinion, the most stringent issue with our integration rule is our lack of ability to sample efficiently from the Gibbs measure (5.1) outside toy settings. We discuss some leads and related problems in this section.

Practical sampling algorithms

In this thesis, we mainly used random walk Metropolis–Hastings (2.64) and MALA (2.66) to approximately sample from our Gibbs measure. While we saw that this was enough to show an improvement over MCMC in terms of confidence regions on toy examples, we also saw in Section 4.4 and Section 5.4 that Metropolis–Hastings with such proposals can have a hard time producing approximate samples from the Gibbs measure, for instance when we increase the number of particles n or the inverse temperature β_n. This is also the case when we are far from a warm-start for the target distribution π, in which case MALA may require many iterations T to reach regions of large mass under π.

I believe that sampling algorithms other than MALA, tailored to our Gibbs measures, could achieve better performance. For instance, we could try to concatenate the MALA proposal (2.66) with large jumps, to enhance exploration in the integration domain. Moreover, choosing a Metropolis–Hastings proposal inspired by the kernel thinning algorithm (Section 2.6.4) would also be interesting. We could as well think of drawing points from a discrete Gibbs measure supported on a pre-drawn background ν_n, following the two-step rationale of coresets or kernel thinning (Dwivedi et al., 2021).

Another promising approach is to rely on piecewise-deterministic Markov processes (Peters and With, 2012; Bouchard-Côté et al., 2018) that are non-reversible versions of MCMC (2.62), see also (Monemvassitis et al., 2023) for an application to the simulation of particle systems. Actually, sampling Gibbs states like $\mathbb{P}_{n,\beta_n}$ at low temperature, meaning high β_n, has decades of history in the statistical physics literature, see for instance (Leimkuhler and Matthews, 2015). Yet we would already be happy with efficient sampling at $\beta_n = n^2$ if what we seek is uncertainty quantification with CLTs like (6.16). Finally, at reasonable temperature schedules $\beta_n = \mathcal{O}(n^2)$, I think sampling methods involving tempered distributions like sequential Monte Carlo samplers (Chopin, 2002; Del Moral et al., 2006; Chopin et al., 2020) could help. It would be interesting to compare all these methods with fixed computational budget.

Another interesting question is whether we can include an online approximation of the potential U_K^π instead of having to draw points using e.g. MCMC first.

Mixing times

Recall that we only produce points that are approximately drawn from $\mathbb{P}_{n,\beta_n}^{V^\pi}$ after T iterations of e.g. MCMC. It is therefore important to be able to quantify how big T needs to be in terms of n, β_n and d to use the statistical properties of $\mathbb{P}_{n,\beta_n}^{V^\pi}$ in our Monte Carlo estimator. This is quantified by *mixing times*, see Corollary 4.3.14.

In Chapter 4, we used heuristics inspired by (Dwivedi et al., 2019), who give results in the log-concave setting for MALA. Our Gibbs measure $\mathbb{P}_{n,\beta_n}^{V^\pi}$ is usually not log-concave, for the same reasons that the MMD (2.97) is usually not convex.

Very few results are available in the MCMC literature in this non-convex setting. For instance, Ma et al. (2019) and Chak (2025) give polynomial convergence in terms of the

dimension d', under a condition linking d' and the radius R of a ball outside which the target is strongly log-concave. Moreover, Ma et al. (2019) show that this is a setting where we can actually even hope for sampling to be easier than optimization as d' grows. While assuming that $\mathbb{P}^{V^\pi}_{n,\beta_n}$ is log-concave outside a ball of radius R is not an issue with a bounded interaction K, the condition linking d' and R is still too rough at the moment to be applied to our Gibbs measure which has dimension $d' = dn$. Still, I believe that we could refine these results for the specific case of $\mathbb{P}^{V^\pi}_{n,\beta_n}$ and with a MCMC sampler tailored to our Gibbs distribution.

On a different line of work, sampling repulsive Gibbs point processes using random graphs instead of MCMC has been investigated in (Friedrich et al., 2025) with mixing times guarantees.

6.2 Openings

In this ultimate section, we turn to other interesting questions at the crossroads between probabilistic machine learning and Gibbs measures.

6.2.1 Smale's 7th problem

Efficient sampling from Gibbs measures like $\mathbb{P}^{V}_{n,\beta_n}$ is also linked to Smale's 7th problem (Smale, 1998; Beltrán, 2012). In the latter problem, the goal is to compute an approximated solution to

$$\min_{x_1,\dots,x_n \in \mathbb{S}} \tilde{H}_n(x_1,\dots,x_n), \tag{6.26}$$

in polynomial time in n, where $\mathbb{S}$ is the 2-sphere

$$\mathbb{S} = \{x \in \mathbb{R}^3 : |x| = 1\}, \tag{6.27}$$

and $\tilde{H}_n$ is the logarithmic energy

$$\tilde{H}_n(x_1,\dots,x_n) = \sum_{i \neq j} \log \frac{1}{|x_i - x_j|}. \tag{6.28}$$

By approximate solution, we mean that the logarithmic energy of the point configuration $x_1,\dots,x_n$ should be close to the minimal one in (6.26) up to a $\log n$ error.

Actually, Beltrán and Hardy (2019, Corollary 1.2) have shown that if

$$x_1,\dots,x_n \sim \frac{1}{Z_{n,\beta_n}} \exp\left(-\beta_n H_n(x_1,\dots,x_n)\right) \mathrm{d}\sigma(x_1)\dots\mathrm{d}\sigma(x_n), \tag{6.29}$$

where σ is the uniform measure on $\mathbb{S}$ and

$$H_n(x_1,\dots,x_n) = \frac{1}{2n^2} \sum_{i \neq j} \log \frac{1}{|x_i - x_j|}, \tag{6.30}$$

then, when $\beta_n = 2n^3$,

$$\tilde{H}_n(x_1, \dots, x_n) - \min_{x_1,\dots,x_n \in \mathbb{S}} \tilde{H}_n \leq 10 \log n, \tag{6.31}$$

with probability at least $1 - e^{-n \log n}$, and a similar statement holds in expectation.

Thus, if one is ready to accept probabilistic solutions, approximately sampling from the Gibbs measure (6.29) at low temperature $\beta_n = 2n^3$ and with polynomial mixing time in n would solve Smale's 7th problem. While the inherent dimension of the particles is only $d = 2$, I believe solving Smale's problem will still be much harder than giving polynomial mixing times for bounded kernels: we work here with the singular logarithmic interaction, and at very low temperature n^3 where sampling is expected to be hard when n grows.

6.2.2 Integration in RKHSs

In Chapter 2, we saw different methods to produce points $x_1, \dots, x_n$ with low MMD. This can be practically motivated in Bayesian quadrature (Pronzato et al., 2020), where we assume the integrand f of the numerical integration task to be the realization of a Gaussian process with kernel K. Then, minimizing the posterior variance amounts to minimizing the energy (2.97), and building designs with low MMD is thus already natural.

Still, we may wonder whether it is natural to assume that the integrand f belongs itself to the RKHS $\mathcal{H}_K$ as initially presented in Chapter 1. As discussed in Chapter 2, this is a standard assumption in the Monte Carlo literature in order to improve upon MCMC guarantees. Yet, how can we motivate a Bayesian scientist reading Chapter 1 that he should use numerical integration in RKHSs other than Sobolev spaces?

One way to do this would be to ask the Bayesian scientist for a set of relevant integrands $f_1, \dots, f_M$ that he wants to integrate against a target distribution π, and to build a meaningful RKHS based on these integrands. In that case, what would be the natural kernel K? How could we then interpret the resulting quadrature?

6.2.3 Gibbs measures and parallel MCMC

One different use of Gibbs measures would be to try to reduce bias for parallel MCMC. Recall from Section 2.3 that a standard way to build an estimator for

$$\int f \, \mathrm{d}\pi, \tag{6.32}$$

is to average f over the history of a suitable MCMC chain $x^{(1)}, \dots, x^{(T)}$, with T a large number of iterations. In that case, we saw that the root mean squared error is

$$\sqrt{\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T} f\left(x^{(t)}\right) - \int f \, \mathrm{d}\pi\right|^2\right]} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \tag{6.33}$$

as $T \to +\infty$.

In modern hardware, it is often the case that we have a potentially large number n of computing units. One natural way to leverage those n units is to run n independent MCMC chains $(x_i^{(t)})$ in parallel for T iterations each, and to consider the estimator

$$\int f \, \mathrm{d}\pi \approx \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} f\left(x_i^{(t)}\right). \tag{6.34}$$

This is computationally attractive for instance when we are limited in computing time on each machine, or when we want to speed up expensive simulations in Bayesian inference, see for instance (Angelino et al., 2016; Bouton et al., 2025). This is the case for instance in our motivating examples of Chapter 1 in computational biology where the integrand f is expensive to evaluate (Niederer et al., 2011). In that setup, we would rather run T iterations simultaneously on n different units than run nT MCMC steps on a single device.

Yet, naive parallel MCMC lacks theoretical support. In fact, the quadratic risk for the estimator (6.34) has the bias-variance decomposition

$$\frac{1}{n}\mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T} f\left(x_i^{(t)}\right)\right] + \left(\frac{1}{T}\sum_{t=1}^{T} \xi P^t f - \int f \, \mathrm{d}\pi\right)^2 \tag{6.35}$$

by independence, where P denotes the Markov kernel used on each unit and ξ is the starting distribution.

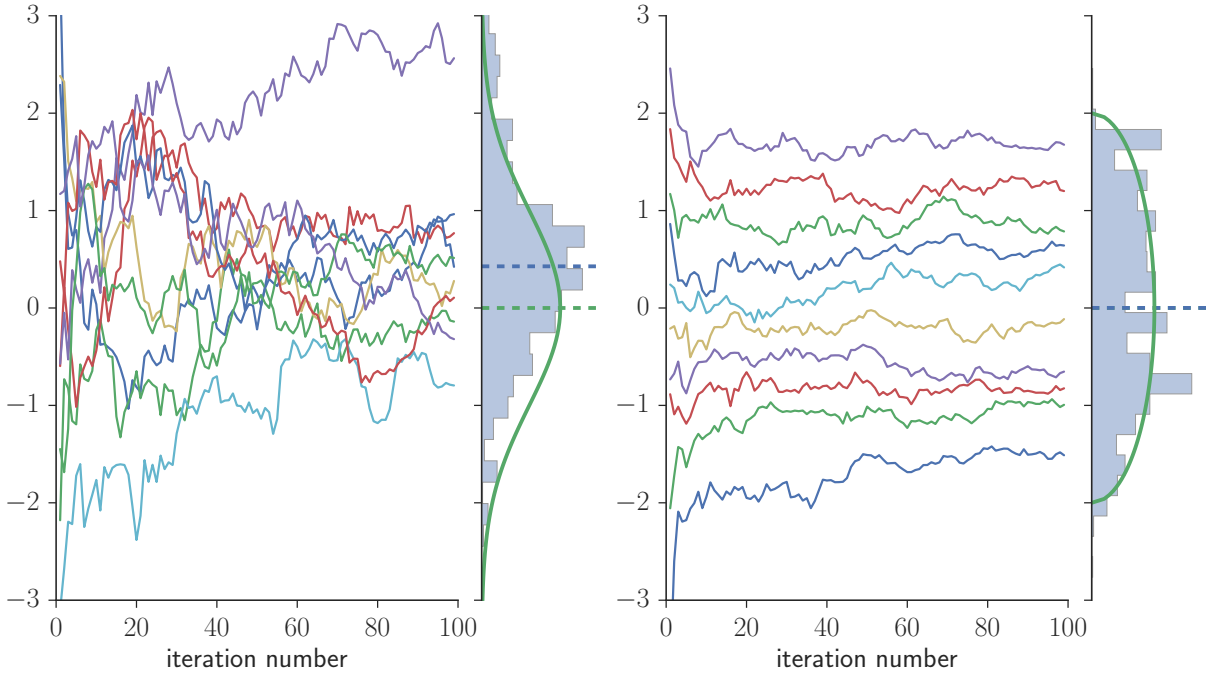
While the variance term in (6.35) decays with the number of computing units n, the bias term does not depend on n. This suggests that averaging estimators over a large number of parallel MCMC chains with T iterations each yields a bias that is as large as that of a single chain of size T. This observation dates back to (Rosenthal, 2000).

Although non-trivial parallel versions of MCMC have been considered (Bornn et al., 2013; Fort et al., 2014; Calderhead, 2014; Terenin et al., 2020), none has provably beaten the naive convergence rate in (6.35). As a side note, there has also been important advances for parallel algorithms in sequential Monte Carlo setups, see (Murray et al., 2016) and references therein.

One exciting line of research has however emerged with *unbiased MCMC* (Jacob et al., 2020; Atchadé and Jacob, 2024). The goal is to use couplings to completely remove the bias term in the MCMC quadratic risk (6.33), so that the resulting estimator can be parallelized with a $n^{-1/2}$ gain in (6.35), if n is the number of computing units.

As another take on parallel MCMC, following the rationale that with repulsiveness comes better exploration of the domain of integration and smaller Monte Carlo error, we think that considering a Markov process $X_i^{(t)}$ that leaves a Gibbs distribution like (3.45) invariant can help provably reduce the quadratic risk (6.35) as n grows. Our guess is

motivated by trajectorial results for Dyson's Brownian motion, an example of repulsive stochastic process from random matrix theory where n Brownian motions are forced to stay close to each other while never intersecting (Tao, 2012, Chapter 3), see Figure 6.1. Moreover, sampling interacting stochastic processes on parallel architectures has been for instance investigated in (Vergé et al., 2015), so that there would be hope for a practical implementation.



(a) $\beta_n = n^{3/2}$ **(b)** $\beta_n = n^2$

Fig. 6.1: (6.1a) The trajectories of $n = 50$ rescaled independent Brownian motions for $T = 50$ time steps. (6.1b) The trajectories of a rescaled Dyson Brownian Motion with $n = 50$ and $T = 50$ time steps. In both panels, the rightmost marginal plot is a histogram of the samples at the last time step, superimposed with the theoretical limiting pdf.

6.2.4 Mean-field theory for neural networks

As a final opening, energies like our H_n in (3.40) has also been used to study convergence of SGD dynamics in the training of neural networks (Rotskoff and Vanden-Eijnden, 2022), with choices of K and V depending on the parameters of the neural net. We could then wonder whether the theory of Gibbs measures could help give guarantees on such learning algorithms with e.g. central limit theorems.

Bibliography

Alquier, Pierre and Ridgway, James (June 2020). "Concentration of tempered posteriors and of their variational approximations". In: *The Annals of Statistics* 48.3. Publisher: Institute of Mathematical Statistics, pp. 1475–1497. ISSN: 0090-5364, 2168-8966.

Alquier, Pierre, Ridgway, James, and Chopin, Nicolas (2016). "On the properties of variational approximations of Gibbs posteriors". In: *Journal of Machine Learning Research* 17.236, pp. 1–41. ISSN: 1533-7928.

Anastasiou, Andreas, Barp, Alessandro, Briol, François-Xavier, Ebner, Bruno, Gaunt, Robert E., Ghaderinezhad, Fatemeh, Gorham, Jackson, Gretton, Arthur, Ley, Christophe, Liu, Qiang, Mackey, Lester, Oates, Chris J., Reinert, Gesine, and Swan, Yvik (June 22, 2022). *Stein's Method Meets Computational Statistics: A Review of Some Recent Developments*. arXiv: 2105.03481.

Anderson, Greg W., Guionnet, Alice, and Zeitouni, Ofer (2010). *An introduction to random matrices*. Vol. 118. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge. xiv+492. ISBN: 978-0-521-19452-5.

Angelino, Elaine, Johnson, Matthew James, and Adams, Ryan P. (Nov. 17, 2016). "Patterns of Scalable Bayesian Inference". In: *Foundations and Trends in Machine Learning* 9.2, pp. 119–247. ISSN: 1935-8237.

Angst, Jürgen, Herry, Ronan, Malicet, Dominique, and Poly, Guillaume (Mar. 25, 2024). *Sharp total variation rates of convergence for fluctuations of linear statistics of β-ensembles*. arXiv: 2403.17211.

Arbel, Michael, Korba, Anna, SALIM, Adil, and Gretton, Arthur (2019). "Maximum Mean Discrepancy Gradient Flow". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Atchadé, Yves F. and Jacob, Pierre E. (June 10, 2024). *Unbiased Markov Chain Monte Carlo: what, why, and how*. arXiv: 2406.06851[stat].

Augustin, Christoph M., Neic, Aurel, Liebmann, Manfred, Prassl, Anton J., Niederer, Steven A., Haase, Gundolf, and Plank, Gernot (Jan. 15, 2016). "Anatomically accurate high resolution modeling of human whole heart electromechanics: A strongly scalable algebraic multigrid solver method for nonlinear deformation". In: *Journal of Computational Physics* 305, pp. 622–646. ISSN: 0021-9991.

Ba, Jimmy, Erdogdu, Murat A., Ghassemi, Marzyeh, Sun, Shengyang, Suzuki, Taiji, Wu, Denny, and Zhang, Tianzong (Oct. 6, 2021). "Understanding the Variance Collapse of SVGD in High Dimensions". In: International Conference on Learning Representations.

Bach, Francis, Lacoste-Julien, Simon, and Obozinski, Guillaume (Sept. 11, 2012). *On the Equivalence between Herding and Conditional Gradient Algorithms*. arXiv: 1203.4523.

Balasubramanian, Krishnakumar, Banerjee, Sayan, and Ghosal, Promit (Sept. 30, 2024). *Improved Finite-Particle Convergence Rates for Stein Variational Gradient Descent*. version: 2. arXiv: 2409.08469.

Bardenet, Rémi and Hardy, Adrien (Feb. 2020). "Monte Carlo with determinantal point processes". In: *The Annals of Applied Probability* 30.1. Publisher: Institute of Mathematical Statistics, pp. 368–417. ISSN: 1050-5164, 2168-8737.

Bardenet, Rémi, Maïda, Mylène, and Rouault, Martin (Aug. 2, 2025). *Quenched large deviations for Monte Carlo integration with Coulomb gases*. arXiv: 2508.01392[cs].

Beck, József (Dec. 1, 1987). "Irregularities of distribution. I". In: *Acta Mathematica* 159.1, pp. 1–49. ISSN: 1871-2509.

Bekerman, Florent, Leblé, Thomas, and Serfaty, Sylvia (Jan. 1, 2018). "CLT for Fluctuations of $\beta $-ensembles with general potential". In: *Electronic Journal of Probability* 23 (none). ISSN: 1083-6489.

Belhadji, A., Bardenet, R., and Chainais, P. (2019). "Kernel quadrature with determinantal point processes". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

– (2020). "Kernel interpolation with continuous volume sampling". In: *International Conference on Machine Learning (ICML)*.

Beltrán, C. (2012). "The State of the Art in Smale's 7th Problem". In: *Foundations of Computational Mathematics, Budapest 2011*. Ed. by Agnes Szanto, Allan Pinkus, Felipe Cucker, and Teresa Krick. London Mathematical Society Lecture Note Series. Cambridge: Cambridge University Press, pp. 1–15. ISBN: 978-1-107-60407-0.

Beltrán, Carlos and Hardy, Adrien (Mar. 1, 2019). "Energy of the Coulomb Gas on the Sphere at Low Temperature". In: *Archive for Rational Mechanics and Analysis* 231.3, pp. 2007–2017. ISSN: 1432-0673.

Ben Arous, G. and Guionnet, A. (1997). "Large deviations for Wigner's law and Voiculescu's non-commutative entropy". In: *Probability Theory and Related Fields* 108.4, pp. 517–542. ISSN: 0178-8051,1432-2064.

Ben Arous, Gérard and Zeitouni, Ofer (1998). "Large deviations from the circular law". In: *European Series in Applied and Industrial Mathematics. Probability and Statistics* 2, pp. 123–134. ISSN: 1292-8100,1262-3318.

Bénard, Clément, Staber, Brian, and Veiga, Sébastien Da (Oct. 26, 2023). *Kernel Stein Discrepancy thinning: a theoretical perspective of pathologies and a practical fix with regularization*. arXiv: 2301.13528.

Berlinet, Alain and Thomas-Agnan, Christine (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston, MA: Springer US. ISBN: 978-1-4613-4792-7 978-1-4419-9096-9.

Blei, David M., Kucukelbir, Alp, and McAuliffe, Jon D. (Apr. 3, 2017). "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518, pp. 859–877. ISSN: 0162-1459.

Bolley, François, Guillin, Arnaud, and Villani, Cédric (Mar. 1, 2007). "Quantitative Concentration Inequalities for Empirical Measures on Non-compact Spaces". In: *Probability Theory and Related Fields* 137.3, pp. 541–593. ISSN: 1432-2064.

Bordenave, Charles and Chafaï, Djalil (Jan. 2012). "Around the circular law". In: *Probability Surveys* 9 (none). Publisher: Institute of Mathematical Statistics and Bernoulli Society, pp. 1–89. ISSN: 1549-5787, 1549-5787.

Bornn, Luke, Jacob, Pierre E., Del Moral, Pierre, and Doucet, Arnaud (July 1, 2013). "An Adaptive Interacting Wang–Landau Algorithm for Automatic Density Exploration". In: *Journal of Computational and Graphical Statistics* 22.3, pp. 749–773. ISSN: 1061-8600.

Borot, G. and Guionnet, A. (Jan. 2013). "Asymptotic Expansion of β Matrix Models in the One-cut Regime". In: *Communications in Mathematical Physics* 317.2, pp. 447–483. ISSN: 0010-3616, 1432-0916.

Bouchard-Côté, Alexandre, Vollmer, Sebastian J., and Doucet, Arnaud (Apr. 3, 2018). "The Bouncy Particle Sampler: A Nonreversible Rejection-Free Markov Chain Monte Carlo Method". In: *Journal of the American Statistical Association* 113.522, pp. 855–867. ISSN: 0162-1459.

Bouton, Maxime, Thouvenin, Pierre-Antoine, Repetti, Audrey, and Chainais, Pierre (2025). "A Distributed Plug-and-Play MCMC Algorithm for High-Dimensional Inverse Problems". In: Publisher: arXiv Version Number: 1.

Bradbury, James, Frostig, Roy, Hawkins, Peter, Johnson, Matthew James, Leary, Chris, Maclaurin, Dougal, Necula, George, Paszke, Adam, VanderPlas, Jake, Wanderman-Milne, Skye, and Zhang, Qiao (2018). *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13.

Breyer, L. A. and Roberts, G. O. (Dec. 1, 2000). "From metropolis to diffusions: Gibbs states and optimal scaling". In: *Stochastic Processes and their Applications* 90.2, pp. 181–206. ISSN: 0304-4149.

Brezis, Haim (2011). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York, NY: Springer. ISBN: 978-0-387-70913-0 978-0-387-70914-7.

Briol, François-Xavier, Oates, Chris J., Girolami, Mark, Osborne, Michael A., and Sejdinovic, Dino (Feb. 2019). "Probabilistic Integration: A Role in Statistical Computation?" In: *Statistical Science* 34.1. Publisher: Institute of Mathematical Statistics, pp. 1–22. ISSN: 0883-4237, 2168-8745.

Butez, Raphael and García-Zelada, David (Feb. 2022). "Extremal particles of two-dimensional Coulomb gases and random polynomials on a positive background". In: *The Annals of Applied Probability* 32.1. Publisher: Institute of Mathematical Statistics, pp. 392–425. ISSN: 1050-5164, 2168-8737.

Butez, Raphael, García-Zelada, David, Nishry, Alon, and Wennman, Aron (May 7, 2025). "Universality of Outliers in Weakly Confined Coulomb-Type Systems". In: *Communications in Mathematical Physics* 406.6, p. 127. ISSN: 1432-0916.

Calderhead, Ben (Dec. 9, 2014). "A general construction for parallelizing Metropolis-Hastings algorithms". In: *Proceedings of the National Academy of Sciences* 111.49. Publisher: Proceedings of the National Academy of Sciences, pp. 17408–17413.

Chafaï, Djalil and Ferré, Grégoire (Feb. 1, 2019). "Simulating Coulomb and Log-Gases with Hybrid Monte Carlo Algorithms". In: *Journal of Statistical Physics* 174.3, pp. 692–714. ISSN: 1572-9613.

Chafaï, Djalil, García-Zelada, David, and Jung, Paul (Mar. 27, 2020). "Macroscopic and edge behavior of a planar jellium". In: *Journal of Mathematical Physics* 61.3, p. 033304. ISSN: 0022-2488.

Chafaï, Djalil, Gozlan, Nathael, and Zitt, Pierre-André (2014). "First-order global asymptotics for confined particles with singular pair repulsion". In: *The Annals of Applied Probability* 24.6, pp. 2371–2413. ISSN: 1050-5164,2168-8737.

Chafaï, Djalil, Hardy, Adrien, and Maïda, Mylène (2018). "Concentration for Coulomb gases and Coulomb transport inequalities". In: *Journal of Functional Analysis* 275.6, pp. 1447–1483. ISSN: 0022-1236,1096-0783.

Chak, Martin (May 26, 2025). *On theoretical guarantees and a blessing of dimensionality for nonconvex sampling*. arXiv: 2411.07776[stat].

Chatalic, Antoine, Schreuder, Nicolas, Vito, Ernesto De, and Rosasco, Lorenzo (June 16, 2025). *Efficient Numerical Integration in Reproducing Kernel Hilbert Spaces via Leverage Scores Sampling*. arXiv: 2311.13548[stat].

Chen, Wilson Ye, Barp, Alessandro, Briol, François-Xavier, Gorham, Jackson, Girolami, Mark, Mackey, Lester, and Oates, Chris J. (Sept. 14, 2020). *Stein Point Markov Chain Monte Carlo*. arXiv: 1905.03673.

Chen, Wilson Ye, Mackey, Lester, Gorham, Jackson, Briol, François-Xavier, and Oates, Chris J. (June 19, 2018). *Stein Points*. arXiv: 1803.10161.

Chen, Yutian and Welling, Max (Mar. 31, 2010). "Parametric Herding". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. ISSN: 1938-7228. JMLR Workshop and Conference Proceedings, pp. 97–104.

Chen, Yutian, Welling, Max, and Smola, Alex (Mar. 15, 2012). *Super-Samples from Kernel Herding*. arXiv: 1203.3472.

Chewi, Sinho, Niles-Weed, Jonathan, and Rigollet, Philippe (Nov. 7, 2024). *Statistical optimal transport*. arXiv: 2407.18163[math].

Chopin, Nicolas (Aug. 1, 2002). "A sequential particle filter method for static models". In: *Biometrika* 89.3, pp. 539–552. ISSN: 0006-3444.

Chopin, Nicolas and Ducrocq, Gabriel (Aug. 2021). "Fast Compression of MCMC Output". In: *Entropy* 23.8. Publisher: Multidisciplinary Digital Publishing Institute, p. 1017. ISSN: 1099-4300.

Chopin, Nicolas and Gerber, Mathieu (Jan. 24, 2024). "Higher-Order Monte Carlo through Cubic Stratification". In: *SIAM Journal on Numerical Analysis*. Publisher: Society for Industrial and Applied Mathematics University City, Philadelphia.

Chopin, Nicolas and Papaspiliopoulos, Omiros (2020). *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Cham: Springer International Publishing. ISBN: 978-3-030-47844-5 978-3-030-47845-2.

Csiszár, Imre (2006). "A simple proof of Sanov's theorem". In: *Bull. Braz. Math. Soc. (N.S.)* 37.4, pp. 453–459. ISSN: 1678-7544,1678-7714.

Del Moral, Pierre, Doucet, Arnaud, and Jasra, Ajay (2006). "Sequential Monte Carlo samplers". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.3, pp. 411–436. ISSN: 1467-9868.

Dembo, Amir and Zeitouni, Ofer (2010). *Large Deviations Techniques and Applications*. Vol. 38. Stochastic Modelling and Applied Probability. Berlin, Heidelberg: Springer. ISBN: 978-3-642-03310-0 978-3-642-03311-7.

Dereudre, David, Flimmel, Daniela, Huesmann, Martin, and Leblé, Thomas (July 9, 2025). *(Non)-hyperuniformity of perturbed lattices*. arXiv: 2405.19881[math].

Dick, Josef (June 2011). "Higher order scrambled digital nets achieve the optimal rate of the root mean square error for smooth integrands". In: *The Annals of Statistics* 39.3. Publisher: Institute of Mathematical Statistics, pp. 1372–1398. ISSN: 0090-5364, 2168-8966.

Dick, Josef, Kuo, Frances Y., and Sloan, Ian H. (May 2013). "High-dimensional integration: The quasi-Monte Carlo way". In: *Acta Numerica* 22, pp. 133–288. ISSN: 0962-4929, 1474-0508.

Dick, Josef and Pillichshammer, Friedrich (2010). *Digital Nets and Sequences: Discrepancy Theory and Quasi–Monte Carlo Integration*. Cambridge: Cambridge University Press. ISBN: 978-0-521-19159-3.

Douc, Randal, Moulines, Éric, and Stoffer, David (2014). *Nonlinear time series : theory, methods and applications with R examples*. Texts in statistical science. Chapman et Hall - CRC Press. . Pp.

Dudley, R. M. (2002). *Real Analysis and Probability*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press. ISBN: 978-0-521-80972-6.

Dwivedi, Raaz, Chen, Yuansi, Wainwright, Martin J., and Yu, Bin (Dec. 11, 2019). *Log-concave sampling: Metropolis-Hastings algorithms are fast*. arXiv: 1801.02309.

Dwivedi, Raaz and Mackey, Lester (July 21, 2021). "Kernel Thinning". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Conference on Learning Theory. ISSN: 2640-3498. PMLR, pp. 1753–1753.

– (Aug. 1, 2024). *Generalized Kernel Thinning*. arXiv: 2110.01593.

Flegal, James M. and Jones, Galin L. (2010). "Batch Means and Spectral Variance Estimators in Markov Chain Monte Carlo". In: *The Annals of Statistics* 38.2. Publisher: Institute of Mathematical Statistics, pp. 1034–1070. ISSN: 0090-5364.

Fort, G., Moulines, E., Priouret, P., and Vandekerkhove, P. (May 2014). "A central limit theorem for adaptive and interacting Markov chains". In: *Bernoulli* 20.2. Publisher: Bernoulli Society for Mathematical Statistics and Probability, pp. 457–485. ISSN: 1350-7265.

Fournier, Nicolas and Guillin, Arnaud (Aug. 1, 2015). "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability Theory and Related Fields* 162.3, pp. 707–738. ISSN: 1432-2064.

Friedrich, Tobias, Göbel, Andreas, Katzmann, Maximilian, Krejca, Martin S., and Pappik, Marcus (Jan. 2025). "Sampling repulsive Gibbs point processes using random graphs". In: *Combinatorics, Probability and Computing* 34.1, pp. 63–89. ISSN: 0963-5483, 1469-2163.

Fyodorov, Yan V., Hiary, Ghaith A., and Keating, Jonathan P. (Apr. 26, 2012). "Freezing Transition, Characteristic Polynomials of Random Matrices, and the Riemann Zeta Function". In: *Physical Review Letters* 108.17. Publisher: American Physical Society, p. 170601.

García-Zelada, David (2019). "A large deviation principle for empirical measures on Polish spaces: application to singular Gibbs measures on manifolds". In: *Annales de l'Institut Henri Poincaré Probabilités et Statistiques* 55.3, pp. 1377–1401. ISSN: 0246-0203,1778-7017.

García-Zelada, David and Padilla-Garza, David (Jan. 23, 2023). *Generalized transport inequalities and concentration bounds for Riesz-type gases*. arXiv: 2209.00587.

Gauss, Carl Friedrich (1815). *Methodus nova integralium valores per approximationem inveniendi*. OCLC: 457651515. Gottingae: H. Dieterich. 30 pp.

Gautier, G., Bardenet, R., and Valko, M. (2019a). "On two ways to use determinantal point processes for Monte Carlo integration". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Gautier, Guillaume, Polito, Guillermo, Bardenet, Rémi, and Valko, Michal (2019b). "DPPy: DPP sampling with Python". In: *Journal of Machine Learning Research* 20.180, pp. 1–7.

Gautschi, Walter (2006). "Orthogonal Polynomials, Quadrature, and Approximation: Computational Methods and Software (in Matlab)". In: *Orthogonal Polynomials and Special Functions: Computation and Applications*. Ed. by Francisco Marcellán and Walter Van Assche. Berlin, Heidelberg: Springer, pp. 1–77. ISBN: 978-3-540-36716-1.

Gelman, A, Roberts, G.O., and Gilks, W. R. (May 9, 1996). "Efficient Metropolis Jumping Rules". In: *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*. Ed. by J M Bernardo, J O Berger, A P Dawid, and A F M Smith. Oxford University Press, p. 0. ISBN: 978-0-19-852356-7.

Ghosh, Subhro and Lebowitz, Joel L. (Dec. 6, 2016). *Fluctuations, large deviations and rigidity in hyperuniform systems: a brief survey*. arXiv: 1608.07496[math].

Glynn, Peter W. and Szechtman, Roberto (2002). "Some New Perspectives on the Method of Control Variates". In: *Monte Carlo and Quasi-Monte Carlo Methods 2000*. Ed. by Kai-Tai Fang, Harald Niederreiter, and Fred J. Hickernell. Berlin, Heidelberg: Springer, pp. 27–49. ISBN: 978-3-642-56046-0.

Gnewuch, Michael, Srivastav, Anand, and Winzen, Carola (Apr. 1, 2009). "Finding optimal volume subintervals with k points and calculating the star discrepancy are NP-hard problems". In: *Journal of Complexity* 25.2, pp. 115–127. ISSN: 0885-064X.

Gorham, Jackson and Mackey, Lester (Oct. 15, 2020). *Measuring Sample Quality with Kernels*. arXiv: 1703.01717.

Gretton, Arthur, Borgwardt, Karsten, Rasch, Malte, Schölkopf, Bernhard, and Smola, Alex (2006). "A Kernel Method for the Two-Sample-Problem". In: *Advances in Neural Information Processing Systems*. Vol. 19. MIT Press.

Hastings, W Keith (1970). "Monte Carlo sampling methods using Markov chains and their applications". In.

Hawat, Diala (Nov. 27, 2023). "Point processes for numerical integration". PhD thesis. Université de Lille.

Hoffman, Matt, Blei, David M., Wang, Chong, and Paisley, John (Apr. 22, 2013). *Stochastic Variational Inference*. arXiv: 1206.7051.

Hough, J. B., Krishnapur, M., Peres, Y., and Virág, B. (2006). "Determinantal processes and independence". In: *Probability surveys*.

– (2009). *Zeros of Gaussian analytic functions and determinantal point processes*. Vol. 51. American Mathematical Society.

Jacob, Pierre E., O'Leary, John, and Atchadé, Yves F. (2020). "Unbiased Markov chain Monte Carlo methods with couplings". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 82.3. Publisher: [Royal Statistical Society, Oxford University Press], pp. 543–600. ISSN: 1369-7412.

Johansson, Kurt (Jan. 1998). "On fluctuations of eigenvalues of random Hermitian matrices". In: *Duke Mathematical Journal* 91.1. Publisher: Duke University Press, pp. 151–204. ISSN: 0012-7094, 1547-7398.

Johnstone, Ross H., Chang, Eugene T. Y., Bardenet, Rémi, Boer, Teun P. de, Gavaghan, David J., Pathmanathan, Pras, Clayton, Richard H., and Mirams, Gary R. (July 2016). "Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models?" In: *Journal of Molecular and Cellular Cardiology* 96, pp. 49–62. ISSN: 1095-8584.

Kaipio, Jari P. and Somersalo, Erkki (2005). *Statistical and Computational Inverse Problems*. Vol. 160. Applied Mathematical Sciences. New York, NY: Springer. ISBN: 978-0-387-22073-4 978-0-387-27132-3.

Kanagawa, Heishiro, Barp, Alessandro, Gretton, Arthur, and Mackey, Lester (Dec. 2025). "Controlling moments with kernel Stein discrepancies". In: *The Annals of Applied Probability* 35.6. Publisher: Institute of Mathematical Statistics, pp. 3818–3843. ISSN: 1050-5164, 2168-8737.

Korba, Anna, Aubin-Frankowski, Pierre-Cyril, Majewski, Szymon, and Ablin, Pierre (May 20, 2021). *Kernel Stein Discrepancy Descent*. arXiv: 2105.09994.

Korba, Anna, Salim, Adil, Arbel, Michael, Luise, Giulia, and Gretton, Arthur (2020). "A Non-Asymptotic Analysis for Stein Variational Gradient Descent". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 4672–4682.

L'Ecuyer, Pierre (Aug. 14, 2016). "Randomized Quasi-Monte Carlo: An Introduction for Practitioners". In: 12th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing (MCQMC 2016).

Lachièze-Rey, Raphaël and Yogeshwaran, D. (Apr. 17, 2025). *Hyperuniformity and optimal transport of point processes*. version: 3. arXiv: 2402.13705[math].

Lambert, Gaultier (Sept. 1, 2020). "Maximum of the Characteristic Polynomial of the Ginibre Ensemble". In: *Communications in Mathematical Physics* 378.2, pp. 943–985. ISSN: 1432-0916.

Lambert, Gaultier, Leblé, Thomas, and Zeitouni, Ofer (Jan. 2024). "Law of large numbers for the maximum of the two-dimensional Coulomb gas potential". In: *Electronic Journal of Probability* 29. Publisher: Institute of Mathematical Statistics and Bernoulli Society, pp. 1–36. ISSN: 1083-6489, 1083-6489.

Lambert, Gaultier, Ledoux, Michel, and Webb, Christian (Sept. 2019). "Quantitative normal approximation of linear statistics of β-ensembles". In: *The Annals of*

Probability 47.5. Publisher: Institute of Mathematical Statistics, pp. 2619–2685. ISSN: 0091-1798, 2168-894X.

Lambert, Marc, Chewi, Sinho, Bach, Francis, Bonnabel, Silvère, and Rigollet, Philippe (Apr. 21, 2023). *Variational inference via Wasserstein gradient flows*. arXiv: 2205.15902.

Landkof, Naum Samoilovitch (1972). *Foundations of Modern Potential Theory*. Springer Berlin, Heidelberg.

Last, Günter and Penrose, Mathew (2018). *Lectures on the Poisson process*. Vol. 7. Cambridge University Press.

Leblé, Thomas and Serfaty, Sylvia (2017). "Large deviation principle for empirical fields of log and Riesz gases". In: *Inventiones Mathematicae* 210.3, pp. 645–757. ISSN: 0020-9910,1432-1297.

– (2018). "Fluctuations of two dimensional Coulomb gases". In: *Geometric and Functional Analysis* 28.2, pp. 443–508. ISSN: 1016-443X,1420-8970.

Leimkuhler, Ben and Matthews, Charles (2015). *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Vol. 39. Interdisciplinary Applied Mathematics. Cham: Springer International Publishing. ISBN: 978-3-319-16374-1 978-3-319-16375-8.

Leluc, Rémi, Portier, François, Segers, Johan, and Zhuman, Aigerim (Apr. 4, 2024). *Speeding up Monte Carlo Integration: Control Neighbors for Optimal Convergence*. arXiv: 2305.06151.

Lemoine, T. and Bardenet, R. (2024). "Monte Carlo methods on compact complex manifolds using Bergman kernels". In: *arXiv preprint arXiv:2405.09203*.

Lewin, Mathieu (2022). "Coulomb and Riesz gases: the known and the unknown". In: *Journal of Mathematical Physics* 63.6, Paper No. 061101, 77. ISSN: 0022-2488,1089-7658.

Li, Lingxiao, Dwivedi, Raaz, and Mackey, Lester (July 31, 2024). *Debiased Distribution Compression*. arXiv: 2404.12290.

Liu, Qiang and Wang, Dilin (Sept. 9, 2019). *Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm*. arXiv: 1608.04471.

Ma, Yi-An, Chen, Yuansi, Jin, Chi, Flammarion, Nicolas, and Jordan, Michael I. (Oct. 15, 2019). "Sampling Can Be Faster Than Optimization". In: *Proceedings of the National Academy of Sciences* 116.42, pp. 20881–20885. ISSN: 0027-8424, 1091-6490. arXiv: 1811.08413[stat].

Macchi, O. (1972). "Processus ponctuels et coincidences – Contributions à l'étude théorique des processus ponctuels, avec applications à l'optique statistique et aux communications optiques". PhD thesis. Université Paris-Sud.

Macchi, Odile (Mar. 1975). "The coincidence approach to stochastic point processes". In: *Advances in Applied Probability* 7.1, pp. 83–122. ISSN: 0001-8678, 1475-6064.

Mak, Simon and Joseph, V. Roshan (Dec. 2018). "Support points". In: *The Annals of Statistics* 46.6. Publisher: Institute of Mathematical Statistics, pp. 2562–2592. ISSN: 0090-5364, 2168-8966.

Metropolis, Nicholas, Rosenbluth, Arianna W., Rosenbluth, Marshall N., Teller, Augusta H., and Teller, Edward (June 1, 1953). "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092. ISSN: 0021-9606.

Moller, Jesper and Waagepetersen, Rasmus Plenge (Sept. 25, 2003). *Statistical Inference and Simulation for Spatial Point Processes*. Google-Books-ID: dBNOHvElXZ4C. CRC Press. 320 pp. ISBN: 978-0-203-49693-0.

Monemvassitis, Athina, Guillin, Arnaud, and Michel, Manon (Feb. 14, 2023). "PDMP Characterisation of Event-Chain Monte Carlo Algorithms for Particle Systems". In: *Journal of Statistical Physics* 190.3, p. 66. ISSN: 1572-9613.

Murphy, Kevin P. (Mar. 1, 2022). *Probabilistic Machine Learning: An Introduction*. Google-Books-ID: wrZNEAAAQBAJ. MIT Press. 858 pp. ISBN: 978-0-262-04682-4.

Murray, Lawrence M., Lee, Anthony, and Jacob, Pierre E. (July 2, 2016). "Parallel Resampling in the Particle Filter". In: *Journal of Computational and Graphical Statistics* 25.3, pp. 789–805. ISSN: 1061-8600.

Niederer, Steven, Mitchell, Lawrence, Smith, Nicolas, and Plank, Gernot (2011). "Simulating human cardiac electrophysiology on clinical time-scales". In: *Frontiers in Physiology* 2, p. 14. ISSN: 1664-042X.

Novak, Erich (2016). "Some Results on the Complexity of Numerical Integration". In: *Monte Carlo and Quasi-Monte Carlo Methods*. Ed. by Ronald Cools and Dirk Nuyens. Cham: Springer International Publishing, pp. 161–183. ISBN: 978-3-319-33507-0.

Novak, Erich, Ullrich, Mario, Woźniakowski, Henryk, and Zhang, Shun (Sept. 2018). "Reproducing kernels of Sobolev spaces on Rd and applications to embedding constants and tractability". In: *Analysis and Applications* 16.5. Publisher: World Scientific Publishing Co., pp. 693–715. ISSN: 0219-5305.

Oates, Chris J., Cockayne, Jon, Briol, François-Xavier, and Girolami, Mark (Dec. 27, 2017a). *Convergence Rates for a Class of Estimators Based on Stein's Method*. arXiv: 1603.03220.

Oates, Chris J., Girolami, Mark, and Chopin, Nicolas (2017b). "Control functionals for Monte Carlo integration". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 79.3. Publisher: [Royal Statistical Society, Wiley], pp. 695–718. ISSN: 1369-7412.

Owen, Art B. (Aug. 1997). "Scrambled net variance for integrals of smooth functions". In: *The Annals of Statistics* 25.4. Publisher: Institute of Mathematical Statistics, pp. 1541–1562. ISSN: 0090-5364, 2168-8966.

Peilen, Luke (Mar. 1, 2024). *On the Maximum of the Potential of a General Two-Dimensional Coulomb Gas*. version: 1. arXiv: 2403.00670.

Peters, E. A. J. F. and With, G. de (Feb. 15, 2012). "Rejection-free Monte Carlo sampling for general potentials". In: *Physical Review E* 85.2. Publisher: American Physical Society, p. 026703.

Phillips, Jeff M. and Tai, Wai Ming (June 1, 2020). "Near-Optimal Coresets of Kernel Density Estimates". In: *Discrete & Computational Geometry* 63.4, pp. 867–887. ISSN: 1432-0444.

Pinkus, Allan (1985). *n-Widths in Approximation Theory*. Berlin, Heidelberg: Springer. ISBN: 978-3-642-69896-5 978-3-642-69894-1.

Pronzato, Luc and Zhigljavsky, Anatoly (Jan. 2020). "Bayesian Quadrature, Energy Minimization, and Space-Filling Design". In: *SIAM/ASA Journal on Uncertainty Quantification* 8.3. Publisher: Society for Industrial and Applied Mathematics, pp. 959–1011.

Purves, Drew, Scharlemann, Jörn P. W., Harfoot, Mike, Newbold, Tim, Tittensor, Derek P., Hutton, Jon, and Emmott, Stephen (Jan. 2013). "Time to model all life on Earth". In: *Nature* 493.7432. Publisher: Nature Publishing Group, pp. 295–297. ISSN: 1476-4687.

Putinar, M. (Aug. 1997). "A note on Tchakaloff's Theorem". In: *ResearchGate*.

Rachev, S.T. (1991). *Probability Metrics and the Stability of Stochastic Models*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley. ISBN: 9780471928775.

Rasmussen, Carl Edward and Williams, Christopher K. I. (Nov. 23, 2005). *Gaussian Processes for Machine Learning*. The MIT Press. ISBN: 978-0-262-25683-4.

Riabiz, Marina, Chen, Wilson Ye, Cockayne, Jon, Swietach, Pawel, Niederer, Steven A., Mackey, Lester, and Oates, Chris. J. (Sept. 1, 2022). "Optimal Thinning of MCMC Output". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.4, pp. 1059–1081. ISSN: 1369-7412.

Rider, Brian and Virág, Bálint (2007). "The noise in the circular law and the Gaussian free field". In: *International Mathematics Research Notices. IMRN* 2, Art. ID rnm006, 33. ISSN: 1073-7928,1687-0247.

Robert, Christian P. (2007). *The Bayesian Choice*. Springer Texts in Statistics. New York, NY: Springer. ISBN: 978-0-387-71598-8.

Robert, Christian P. and Casella, George (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York, NY: Springer. ISBN: 978-1-4419-1939-7 978-1-4757-4145-2.

Roberts, G. O. and Tweedie, R. L. (1996a). "Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms". In: *Biometrika* 83.1. Publisher: [Oxford University Press, Biometrika Trust], pp. 95–110. ISSN: 0006-3444.

Roberts, Gareth O. and Tweedie, Richard L. (Dec. 1996b). "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* 2.4. Publisher: Bernoulli Society for Mathematical Statistics and Probability, pp. 341–363. ISSN: 1350-7265.

Rosenthal, Jeffrey S. (2000). "Parallel computing and Monte Carlo algorithms". In: *Far east journal of theoretical statistics* 4.2, pp. 207–236.

Rotskoff, Grant and Vanden-Eijnden, Eric (2022). "Trainability and Accuracy of Artificial Neural Networks: An Interacting Particle System Approach". In: *Communications on Pure and Applied Mathematics* 75.9, pp. 1889–1935. ISSN: 1097-0312.

Rouault, Martin, Bardenet, Rémi, and Maïda, Mylène (Feb. 18, 2024). *Monte Carlo with kernel-based Gibbs measures: Guarantees for probabilistic herding*. arXiv: 2402.11736.

– (Aug. 25, 2025). “Intégrer plusieurs fonctions simultanément avec des mesures de Gibbs”. In: *30° Colloque sur le traitement du signal et des images*. Issue: 2025-1751. Strasbourg: GRETSI - Groupe de Recherche en Traitement du Signal et des Images, p. 1157–1160.

Sanov, I. N. (1957). “On the probability of large deviations of random magnitudes”. In: *Mat. Sb. (N.S.)* 42(84), pp. 11–44. ISSN: 0368-8666.

Serfaty, Sylvia (2018). “Systems of points with Coulomb interactions”. In: *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. I. Plenary lectures*. World Sci. Publ., Hackensack, NJ, pp. 935–977. ISBN: 978-981-327-290-3 978-981-327-287-3.

– (2023). “Gaussian fluctuations and free energy expansion for Coulomb gases at any temperature”. In: *Annales de l’Institut Henri Poincaré Probabilités et Statistiques* 59.2, pp. 1074–1142. ISSN: 0246-0203,1778-7017.

– (July 30, 2024). *Lectures on Coulomb and Riesz gases*. arXiv: 2407.21194.

Shetty, Abhishek, Dwivedi, Raaz, and Mackey, Lester (Oct. 18, 2022). *Distribution Compression in Near-linear Time*. arXiv: 2111.07941.

Shi, Jiaxin and Mackey, Lester (Nov. 2, 2023). *A Finite-Particle Convergence Rate for Stein Variational Gradient Descent*. arXiv: 2211.09721.

Smale, Steve (Mar. 1, 1998). “Mathematical problems for the next century”. In: *The Mathematical Intelligencer* 20.2, pp. 7–15. ISSN: 0343-6993.

Soshnikov, A. (2000). “Determinantal random point fields”. In: *Russian Mathematical Surveys* 55, pp. 923–975.

Sriperumbudur, Bharath K., Gretton, A., Fukumizu, K., Scholkopf, B., and Lanckriet, Gert R. G. (2009). “Hilbert Space Embeddings and Metrics on Probability Measures”. In: *Journal of machine learning research*.

Steinwart, Ingo and Scovel, Clint (June 1, 2012). “Mercer’s Theorem on General Domains: On the Interaction between Measures, Kernels, and RKHSs”. In: *Constructive Approximation* 35.3, pp. 363–417. ISSN: 1432-0940.

Stoyan, Dietrich, Kendall, Wilfrid S., and Mecke, Joseph (1995). *Stochastic Geometry and Its Applications*. Google-Books-ID: oFDvAAAAMAAJ. Wiley. 480 pp. ISBN: 978-0-471-95099-8.

Strocchi, Marina, Gsell, Matthias A. F., Augustin, Christoph M., Razeghi, Orod, Roney, Caroline H., Prassl, Anton J., Vigmond, Edward J., Behar, Jonathan M., Gould, Justin S., Rinaldi, Christopher A., Bishop, Martin J., Plank, Gernot, and Niederer, Steven A. (Mar. 5, 2020). “Simulating ventricular systolic motion in a four-chamber heart model with spatially varying robin boundary conditions to model the effect of the pericardium”. In: *Journal of Biomechanics* 101, p. 109645. ISSN: 0021-9290.

Tao, Terence (2012). *Topics in random matrix theory*. Vol. 132. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI. x+282. ISBN: 978-0-8218-7430-1.

Terenin, Alexander, Simpson, Daniel, and Draper, David (June 3, 2020). “Asynchronous Gibbs Sampling”. In: *Proceedings of the Twenty Third International Conference on*

Artificial Intelligence and Statistics. International Conference on Artificial Intelligence and Statistics. ISSN: 2640-3498. PMLR, pp. 144–154.

Tolstikhin, Ilya O, Sriperumbudur, Bharath K., and Schölkopf, Bernhard (2016). "Minimax Estimation of Maximum Mean Discrepancy with Radial Kernels". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc.

Torquato, Salvatore (June 6, 2018). "Hyperuniform states of matter". In: *Physics Reports*. Hyperuniform States of Matter 745, pp. 1–95. ISSN: 0370-1573.

Ullrich, Mario (May 11, 2017). "A Monte Carlo Method for Integration of Multivariate Smooth Functions". In: *SIAM Journal on Numerical Analysis*. Publisher: Society for Industrial and Applied Mathematics.

Vergé, Christelle, Dubarry, Cyrille, Del Moral, Pierre, and Moulines, Eric (Mar. 1, 2015). "On parallel implementation of sequential Monte Carlo methods: the island particle model". In: *Statistics and Computing* 25.2, pp. 243–260. ISSN: 1573-1375.

Villani, Cédric (2009). *Optimal Transport, old and new*. Vol. 338. Grundlehren der mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-71049-3 978-3-540-71050-9.

Welling, Max (June 14, 2009). "Herding dynamical weights to learn". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09: The 26th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming. Montreal Quebec Canada: ACM, pp. 1121–1128. ISBN: 978-1-60558-516-1.

Wendland, Holger (Dec. 13, 2004). *Scattered Data Approximation*. Google-Books-ID: qy4cbWUmSyYC. Cambridge University Press. 346 pp. ISBN: 978-1-139-45665-4.

Widom, Harold (1963). "Asymptotic Behavior of the Eigenvalues of Certain Integral Equations". In: *Transactions of the American Mathematical Society* 109.2. Publisher: American Mathematical Society, pp. 278–295. ISSN: 0002-9947.

Wigner, Eugene P. (1958). "On the Distribution of the Roots of Certain Symmetric Matrices". In: *Annals of Mathematics* 67.2. Publisher: [Annals of Mathematics, Trustees of Princeton University on Behalf of the Annals of Mathematics, Mathematics Department, Princeton University], pp. 325–327. ISSN: 0003-486X.