Monte Carlo with kernel-based Gibbs measures: Guarantees for probabilistic herding

Martin Rouault, Rémi Bardenet, Mylène Maïda

Monte Carlo with kernel-based Gibbs measures: Guarantees for probabilistic herding

Martin Rouault*[1], Rémi Bardenet[1], and Mylène Maïda[2]

[1]Univ. Lille, CNRS, Centrale Lille, UMR 9189 – CRIStAL, 59651 Villeneuve d'Ascq, France
[2]Univ. Lille, CNRS, UMR 8524 – Laboratoire Paul Painlevé, F-59000 Lille, France

February 20, 2024

Abstract

Kernel herding belongs to a family of deterministic quadratures that seek to minimize the worst-case integration error over a reproducing kernel Hilbert space (RKHS). In spite of strong experimental support, it has revealed difficult to prove that this worst-case error decreases at a faster rate than the standard square root of the number of quadrature nodes, at least in the usual case where the RKHS is infinite-dimensional. In this theoretical paper, we study a joint probability distribution over quadrature nodes, whose support tends to minimize the same worst-case error as kernel herding. We prove that it does outperform i.i.d. Monte Carlo, in the sense of coming with a tighter concentration inequality on the worst-case integration error. While not improving the rate yet, this demonstrates that the mathematical tools of the study of Gibbs measures can help understand to what extent kernel herding and its variants improve on computationally cheaper methods. Moreover, we provide early experimental evidence that a faster rate of convergence, though not worst-case, is likely.

1 Introduction

Numerical integration with respect to a possibly unnormalized target distribution π on $\mathbb{R}^d$ has become routine in computational statistics (Robert, 2007) and probabilistic machine learning (Murphy, 2023). Monte Carlo algorithms (Robert & Casella, 2004) are randomized algorithms that tackle this task, defining estimators that rely on n evaluations of the target integrand at suitably chosen random points in $\mathbb{R}^d$, called *nodes*. Classical Monte Carlo algorithms, such as Markov Chain Monte Carlo (MCMC), come with a set of probabilistic error controls, such as central limit theorems, that involve errors of magnitude $n^{-1/2}$. The popularity of MCMC in practice has justified a continuous research effort to improve on that rate, which is considered too slow when evaluating the integrand is computationally costly. Quasi-Monte Carlo methods (QMC), for instance, rely on smoothness assumptions to obtain worst-case error controls of order $1/n$. A common such smoothness assumption is that the target integrands belong to the unit ball of a particular reproducing kernel Hilbert space (RKHS); see e.g. (Dick et al., 2013, Section 3).

At another end of the algorithmic spectrum, variational Bayesian methods (VB; Blei et al., 2017; Liu & Wang, 2016) sacrifice some of the error controls to gain in scalability. At its core, VB is the minimization of a dissimilarity measure between a candidate approximation and the target distribution π. Minimizing a relative entropy, for instance, yields algorithms amenable to stochastic gradient techniques (Hoffman et al., 2013), yet that usually come with no theoretical guarantee on how well integrals w.r.t. π are approximated.

An intermediate method between Monte Carlo and relative entropy-based VB is the minimization of an integral probability metric (IPM) of the form

$$\nu \mapsto I_K(\nu - \pi) = \iint K(x,y)\,\mathrm{d}(\nu-\pi)^{\otimes 2}(x,y), \tag{1}$$

where K is a positive definite kernel, known as the *interaction kernel*. It is known (Sriperumbudur et al., 2010) that the square root of $I_K(\nu-\pi)$ in (1) is the worst-case integration error for integrands in the unit

*Corresponding author: martin.rouault@univ-lille.fr

ball of the RKHS defined by K, when approximating π by ν; see (Pronzato & Zhigljavsky, 2020) for a recent survey. Loosely speaking, minimizing (1) is thus an attempt at designing efficient algorithms in the vein of VB, yet that come with a control on the integration error like Monte Carlo and QMC.

Kernel herding, for instance, which rose to attention in the context of learning Markov random fields (Welling, 2009, 2012; Chen & Welling, 2010), has been shown to actually be a conditional gradient descent that greedily minimizes (1) (Chen et al., 2010; Bach et al., 2012). One practical limitation of kernel herding is the requirement to evaluate the kernel embedding $\int K(\cdot, x)\,\mathrm{d}\pi(x)$. When the support of the target measure π is all of $\mathbb{R}^d$, this requirement can be circumvented by a specific choice of K, namely the *Stein kernel* (Anastasiou et al., 2023). In that case, the IPM (1) coincides with the *kernel Stein discrepancy* (KSD), and simple gradient descent schemes can yield efficient minimization algorithms (Korba et al., 2021). To our knowledge, besides the need to evaluate the kernel embedding, the main theoretical limitation of kernel herding and its variants is that, while there is experimental support in favor of an n^{-1} convergence rate of the worst-case integration error in the RKHS induced by K (Chen et al., 2010; Pronzato, 2023), there is no result that shows an improvement over the Monte Carlo rate $n^{-1/2}$ when the RKHS is infinite-dimensional (Bach et al., 2012), see Section 2.2. Our paper provides a step in this direction, for a randomized relaxation[1] of herding.

Gibbs measures are probability distributions that describe systems of interacting particles. By choosing the interaction carefully, one can arrange the corresponding Gibbs measure to favor configurations of points that tend to minimize I_K in (1), when a suitable inverse temperature parameter goes to infinity. Gibbs measures have been studied for decades in probability and mathematical physics, with a focus on models that relate to electromagnetism (Serfaty, 2018). Classical results include large deviation principles (LDPs; Chafaï et al., 2014) and concentration inequalities in some cases (Chafaï et al., 2018). Our main contribution is to prove that the Gibbs measure whose points *repel* each other by an amount given by a *bounded* kernel $K(x, y)$ satisfies a concentration inequality for the worst-case integration error (1); see Theorem 3.5. Our Corollary 3.6 shows a faster sub-Gaussian decay than the i.i.d or MCMC case, coming from the kernel-dependent repulsion. In other words, our probabilistic relaxation of herding provably outperforms classical Monte Carlo methods, in the limited sense that it requires fewer nodes to reach a given worst-case integration error in any given RKHS. We believe this is important, in the sense that this is a first step in establishing the faster convergence of IPM minimization algorithms. There are many limitations to be studied in future work, however. In particular, we have no exact sampling algorithm for our Gibbs measure, which forces us to use MCMC in our experiments. Moreover, we assume that the target measure has compact support, which prevents using the Stein kernel trick to avoid evaluating the kernel embedding.

The rest of the paper is organized as follows. In Section 2, we quickly survey worst-case controls on the integration error. In Section 3, we introduce a family of Gibbs measures and state our theoretical results. In Section 4, we explain how to approximately sample from such Gibbs measures, and experimentally validate our claims. We discuss perspectives in Section 5. Proofs are deferred to the appendix.

2 Related work

We survey worst-case guarantees for Monte Carlo and IPM minimization algorithms.

2.1 Uniform concentration for Monte Carlo

The introduction of a new Monte Carlo method is typically backed up by a central limit theorem (Robert & Casella, 2004). In practice, where the number n of quadrature nodes is fixed, one prefers a concentration inequality, to derive a confidence interval for $\int f\mathrm{d}\pi$. While rarely put forward, many applications further require a uniform control over several integrands at a time. For instance, in multi-class classification with 0/1 loss and M classes, determining the Bayesian predictor involves giving a joint confidence region over $M-1$ integrals. This motivates studying the simultaneous approximation of several integrals by a single set of n Monte Carlo nodes. One way to formalize this problem is by upper bounding the Wasserstein distance

$$W_1(\mu_n, \pi) = \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \left| \int f \,\mathrm{d}(\mu_n - \pi) \right|, \tag{2}$$

where μ_n is the Monte Carlo empirical approximation of the target measure π, and the supremum is taken over all Lipschitz functions of Lipschitz constant less than 1. Sanov's theorem (Dembo & Zeitouni, 2009,

[1] Actually, the original herding algorithm was inspired by the zero-temperature limit of a physical particle system (Welling, 2009), so our relaxation is a return to the roots of sorts.

Theorem 6.2.10) gives a large deviation principle (LDP), i.e. an asymptotic control on the tails of the random variable (2) when μ_n is the empirical measure of i.i.d. draws from π. Non-asymptotic counterparts have been obtained through sub-Gaussian concentration inequalities with speed n, see (Bolley et al., 2007; Fournier & Guillin, 2015) for more details.

Theorem 2.1 (Bolley, Guillin, and Villani, 2007)**.** *Assume that $x_1, \dots, x_n$ are drawn i.i.d from π. Under a suitable moment assumption on π, for any $d' > d$, there exists n_0 such that, for any $n \geq n_0$ and $r > n^{-1/(d'+2)}$,*

$$\mathbb{P}\left[W_1(\mu_n, \pi) > r\right] \leq \exp\left(-\alpha n r^2\right), \tag{3}$$

where α is a constant depending on π.

Note the regime restriction $r > n^{-1/(d'+2)}$. Analogous LDPs and concentration bounds with the same rate exist for the empirical measure of Markov chains, see (Dembo & Zeitouni, 2009, Chapter 6.5) and (Fournier and Guillin, 2015). From a quadrature point of view, results like Theorem 2.1 give simultaneous confidence intervals.

Corollary 2.2. *Fix $\epsilon > 0$, and $\delta \in (0, 1)$. For any n big enough so that $n^{-1/(d'+2)} \leq \epsilon$ and $\exp\left(-\alpha n \epsilon^2\right) \leq \delta$, with probability higher than $1 - \delta$,*

$$\sup_{\|f\|_{\text{Lip}} \leq 1} \left| \int f \, \mathrm{d}(\mu_n - \pi) \right| \leq \epsilon. \tag{4}$$

where $x_1, \dots, x_n \sim \pi$ are i.i.d. and $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$.

If we take the supremum over smoother functions, like the functions in the RKHS $\mathcal{H}_K$ (Berlinet & Thomas-Agnan, 2011) of positive definite kernel K, one can further hope to reach inequality (4) for smaller values of n. In that case, i.i.d. sampling from the so-called *kernel leverage score* yields a concentration like (4) with a faster rate, though not fully explicit (Bach, 2017, Proposition 2). Belhadji et al. (2020) give a more explicit rate, but depart from the i.i.d. setting by studying a kernel-dependent joint distribution on the quadrature nodes called *volume sampling*. In short, under generic assumptions on the kernel, Markov's inequality applied to (Belhadji et al., 2020, Theorem 4) shows that there is a constant $C > 0$ such that under volume sampling, with probability $1 - C\sigma_{n+1}/\epsilon$,

$$\sup_{f \in \mathcal{H}_K} \left| \int f \mathrm{d}\pi - \sum_{i=1}^n w_i f(x_i) \right|^2 \leq \epsilon, \tag{5}$$

where σ_n is the n-th eigenvalue of the operator on $L^2(\pi)$ with kernel K, and the weights (w_i) are suitably chosen. Because σ_n can go to zero arbitrarily fast with n (e.g., exponentially for the Gaussian kernel), (5) attains a given confidence and error levels at smaller n than under i.i.d. sampling. Downsides are that (i) there is no exact algorithm yet for volume sampling that does not require to evaluate the eigenvalues and eigenfunctions of the integral operator with kernel K, and (ii) the dependence of w_i on all nodes makes it hard to derive, e.g., a central limit theorem.

2.2 Variational Bayes and Kernel herding

Convergence guarantees for VB are often formulated in terms of the minimized dissimilarity measure; see (Alquier et al., 2016; Lambert et al., 2022) and references therein. For instance, under strong assumptions on the target π and the allowed variational approximation, Lambert et al. (2022) have given rates for the convergence to the minimal achievable relative entropy $\mathrm{KL}(\rho_t||\pi)$ between π and the $t-$th iterate ρ_t of an idealized (continuous-time) VB algorithm. For Stein variational gradient descent, Liu & Wang (2016) and Korba et al. (2020) proved a decay of $\mathrm{KL}(\cdot||\pi)$ along with non-asymptotic bounds at rate n^{-1} for the kernel Stein discrepancy (KSD) between an (at most) $n-$point empirical measure based on the algorithm and the target measure π. Since the KSD is a particular case of (1), this implies a control on a worst-case integration error. Indeed, if K is positive definite, so that it defines an RKHS $\mathcal{H}_K$, it can be shown that $I_K(\nu - \pi)$ in (1) is the square of the worst-case integration error over the unit ball of $\mathcal{H}_K$ when replacing π by ν; see e.g. (Sriperumbudur et al., 2010). For KSD, the RKHS corresponds to the Stein kernel.

Under some assumptions on $\mathcal{H}_K$, kernel herding algorithms have been proved to achieve $I_K(\mu_n - \pi) \leq cn^{-2}$, so that the worst-case quadrature error decreases at rate n^{-1} (Chen et al., 2010). Other variants of conditional gradient algorithms led to further improvement, up to convergences of the type

$I_K(\mu_n - \pi) \leq \exp(-cn)$ (Bach et al., 2012). While those assumptions are reasonable when the dimension of $\mathcal{H}_K$ is finite, Bach et al. (2012) have shown that they are *never* fulfilled in the infinite-dimensional setting. In that case, the only general result is the "slow" rate $I_K(\mu_n - \pi) \leq cn^{-1}$ (Bach et al., 2012). Proving a faster rate for a variant of herding remains an open problem.

2.3 LDPs and concentration for Gibbs measures

We informally define a (Gibbs) measure on $\left(\mathbb{R}^d\right)^n$ by

$$\mathbb{P}^V_{n,\beta_n}(\mathrm{d}x_1 \ldots \mathrm{d}x_n) \propto \mathrm{e}^{-\beta_n H_n(x_1,\ldots,x_n)} \mathrm{d}x_1 \ldots \mathrm{d}x_n, \tag{6}$$

where $\beta_n > 0$ is called *inverse temperature*, and where

$$H_n(x_1, \ldots, x_n) = \frac{1}{2n^2} \sum_{i \neq j} K(x_i, x_j) + \frac{1}{n} \sum_{i=1}^{n} V(x_i), \tag{7}$$

with $V : \mathbb{R}^d \to \mathbb{R}$. There are assumptions to be made on K and V to guarantee that (6) defines a *bona fide* probability distribution; see Section 3. H_n in (7) can be recognized to be a discrete analogous to I_K in (1). Intuitively, points distributed according to (6) tend to correspond to low pairwise kernel values $K(x_i, x_j)$ (we say that they *repel* by a force given by the kernel), yet stay confined in regions where V is not too large. We also emphasize that the zero temperature limit – informally taking $\beta_n = +\infty$ in (6) – corresponds to finding the deterministic minimizers of H_n, which in turn intuitively correspond to minimizer of I_K; see (Serfaty, 2018) for precise results. We focus in this paper on the so-called *low-temperature regime* $\beta_n/n \to +\infty$ (denoted in the sequel by $\beta_n \gg n$), in which one can hope to observe properties of the Gibbs measure (6) that depart from those of i.i.d. sets of n points.

Asymptotic properties of (6) as $n \to \infty$ have been studied by Chafaï, Gozlan, and Zitt (2014), who prove an LDP like Sanov's classical result, but with the rate $1/n$ replaced by $1/\beta_n$, which goes faster to 0 in the regime $\beta_n \gg n$. Moreover, the convergence of the empirical measure is now towards the so-called *equilibrium measure* μ_V, which depends in a non-trivial way on V and K. As for non-asymptotic counterparts to this LDP, concentration inequalities have been obtained for some singular[2] kernels, known as the Coulomb and Riesz kernels (Chafaï, Hardy, and Maïda, 2018; García-Zelada and Padilla-Garza, 2022). For instance, Chafaï et al. (2018) prove for the Coulomb kernel that whenever $r > n^{-1/d}$,

$$\mathbb{P}^V_{n,\beta_n}(W_1(\mu_n, \mu_V) > r) \leq \exp\left(-c\beta_n r^2\right). \tag{8}$$

The concentration result (8) improves on the i.i.d. concentration in (3). Besides being valid for values of r down to $n^{-1/d}$, the speed n in the exponential is replaced by β_n, which can increase arbitrarily fast, at the price of replacing the target measure by the equilibrium measure of the system[3]. After choosing a suitable potential V such that $\mu_V = \pi$, we rephrase this bound as a uniform quadrature guarantee.

Corollary 2.3 (Chafaï et al., 2018)**.** *Fix $\epsilon > 0$, and $\delta \in (0, 1)$. For any n big enough so that $n^{-1/d} \leq \epsilon$ and $\exp\left(-c\beta_n \epsilon^2\right) \leq \delta$, with probability higher than $1 - \delta$,*

$$\sup_{\|f\|_{\mathrm{Lip}} \leq 1} \left| \int f \, \mathrm{d}(\mu_n - \pi) \right| \leq \epsilon, \tag{9}$$

where $x_1, \ldots, x_n \sim \mathbb{P}^V_{n,\beta_n}$ in (6) and $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$.

As long as $\beta_n \gg n$, for a fixed confidence level δ and uniform worst-case error ϵ, as soon as n is big enough so that $n^{-1/d} \leq \epsilon$, the constraints in Corollary 2.3 are achieved with a smaller number n of points than for i.i.d. samples in Corollary 2.2. Fewer quadrature nodes are required by the Gibbs measure to achieve the same guarantee.

Results like (Chafaï et al., 2018) are motivated by statistical physics and focus on a particular family of singular kernels. The price of singularity is quite long and technical proofs. On the other hand, in machine learning, we typically consider bounded kernels like the Gaussian or Matern kernel. Our main result is a version of (8) that is valid for very general *bounded* kernels, bringing an improvement over i.i.d. sampling similar to Corollary 2.3, with $n^{-1/d}$ even replaced by $n^{-1/2}$. Maybe surprisingly, while our

[2] By *singular*, we mean that $K(x, x) = +\infty$ for all $x \in \mathbb{R}^d$.

[3] Throughout the section, we neglect sampling costs. While a fast-growing β_n implies better theoretical guarantees, the price of (approximately) sampling from $\mathbb{P}^V_{n,\beta_n}$ intuitively increases with β_n, introducing a trade-off in practice; see Section 4.

proof follows the lines of (Chafaï et al., 2018), we were able to considerably simplify the more technical arguments. We hope our work helps transfer tools and concepts from the theory of Gibbs measures to the study of IPM-based quadrature.

As a final note on existing work, and to prepare for the discussion in Section 5, we remark that central limit theorems for Gibbs measures like (6) (with speed depending on β_n) are very subtle mathematical results. Leblé & Serfaty (2018) and Bauerschmidt et al. (2016) have obtained a CLT only in dimension two so far, and yet only for the (singular) Coulomb kernel. While important steps have been made towards larger dimensions for the Coulomb kernel (Serfaty, 2023), this remains an important and difficult open problem in statistical physics. As a consequence, direct comparison with the $n^{-1/2}$ rate appearing in the CLTs of MCMC chains is currently out of reach.

3 Main results

We first rigorously introduce some key notions to understand the limiting behavior of Gibbs measures, like the *equilibrium measure*. We then introduce our Gibbs measure on quadrature nodes, and state our main result, which features the equilibrium measure. In the last paragraph, we explain how to choose the parameters of the Gibbs measure so that the equilibrium measure is a given target distribution π.

3.1 Energies and the equilibrium measure

Let $d \geq 1$, $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ and $V : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. For reasons that shall become clear shortly, we call K the *interaction kernel*, and V the *external potential*. Assumptions on K and V will be given to make the following definition meaningful.

Definition 3.1 (Energies)**.** 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

$$U_K^\mu(z) = \int K(z,y)\,\mathrm{d}\mu(y), \quad z \in \mathbb{R}^d.$$

The *interaction energy* between μ and ν is defined as

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

When $\mu = \nu$, we simply write $I_K(\mu) = I_K(\mu,\mu)$. Finally, we let[4]

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

The physics-inspired energy vocabulary is useful to the intuition: in a world where the Coulomb interaction is given by K, $U_K^\mu(z)$ would be the electric potential created at point z by charges distributed according to μ. In the same way, $I_K^V(\mu)$ is the energy of points distributed according to μ, repelling each other according to K, and confined by some external potential V. 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 will work under the following assumptions on K and V. The first one restricts our class of interaction kernels, in particular insisting that points should repel, but that the interaction cannot be singular.

Assumption 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 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 2. K is integrally strictly positive definite (ISDP), i.e. $I_K(\mu) > 0$ for any non-zero finite signed measure Borel μ.

[4]Note that with our convention $I_K^0 = \frac{1}{2} I_K$.

Assumption 1 and 2 allow most kernels used in machine learning (Rasmussen & Williams, 2006), like the Gaussian or isotropic Matern kernels, as well as truncated singular kernels like

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

for $\epsilon > 0$ and $s > 0$ (Pronzato & Zhigljavsky, 2020).

Under Assumptions 1 and 2, K is finite on the diagonal and ISDP, so that it is in particular positive definite. We can then consider the RKHS $\mathcal{H}_K$ induced by the kernel K (Berlinet & Thomas-Agnan, 2011). 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 3.1 are well-defined and finite for all finite signed Borel measures; see (Pronzato & Zhigljavsky, 2020) for more details. The following known duality formula then links energy minimization and quadrature guarantees for integrands in the unit ball of $\mathcal{H}_K$.

Proposition 3.2 (Sriperumbudur et al. (2010))**.** *Under Assumptions 1 and 2, for probabilities μ, ν in $\mathcal{E}_K$, let*

$$\gamma_K(\mu,\nu) = \sup_{\|f\|_{\mathcal{H}_K} \leq 1} \left| \int f \,\mathrm{d}(\mu-\nu) \right|. \tag{10}$$

Then $\gamma_K(\mu,\nu) = (I_K(\mu-\nu))^{1/2}$.

We now add an assumption to make sure that V is strong enough a confining term. Together with Assumption 1, this ensures that I_K^V is well-defined for any probability measure[5], again with possibly infinite value.

Assumption 3. V is lower semi-continuous, finite everywhere and $V(x) \longrightarrow +\infty$ when $|x| \longrightarrow +\infty$. Moreover, there exists a constant $c > 0$ such that $\int \exp(-cV(x)) \,\mathrm{d}x < \infty$.

We are now ready to consider the minimizers of I_K^V.

Proposition 3.3. *Let K satisfy Assumptions 1 and 2, and V satisfy Assumption 3. Then*

1. *I_K^V is lower semi-continuous, has compact level sets and $I_K^V(\mu) > -\infty$ 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 $I_K^V(\mu) = \frac{1}{2} I_K(\mu) + \int V \,\mathrm{d}\mu$.*
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 given in Appendix A. It is a careful assembly of arguments from (Chafaï et al., 2014) and (Pronzato & Zhigljavsky, 2020).

3.2 Concentration for the Gibbs measure

We saw in Section 1 that herding-like algorithms rely on finding a configuration of points $\{x_1, \ldots, x_n\}$ that minimizes the interaction energy $I_K(\frac{1}{n}\sum_{i=1}^n \delta_{x_i} - \pi)$. In this paper, we rather consider points that are drawn from a distribution that favors small values for an empirical proxy of this interaction energy. To properly define our distribution, consider first, for $x_1, \ldots, x_n \in \mathbb{R}^d$, the discrete energy

$$H_n(x_1, \ldots, x_n) = \frac{1}{2n^2} \sum_{i \neq j} K(x_i, x_j) + \frac{1}{n} \sum_{i=1}^n V(x_i), \tag{11}$$

which we copy here from (7) for ease of reference. Note the similarity, up to diagonal terms, with $I_K^V(\frac{1}{n}\sum \delta_{x_i})$, where I_K^V is defined in (3.1).

[5] Unlike I_K, we shall only evaluate I_K^V on probability measures.

Definition 3.4. Let K satisfy Assumptions 1 and 2, and V satisfy Assumption 3. Let $\beta_n \geq 2cn$, where c is the constant of Assumption 3. The Gibbs measure $\mathbb{P}^V_{n,\beta_n}$ is the probability measure on $\left(\mathbb{R}^d\right)^n$ defined by

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

where β_n is called the inverse temperature, X_n is short for $(x_1, \ldots, x_n)$, and Z^V_{n,β_n} is the normalization constant, which, under our assumptions, is finite and positive (Chafaï et al., 2014, 2018).

When $x_1, \ldots, x_n \sim \mathrm{d}\mathbb{P}^V_{n,\beta_n}$, we will henceforth denote by $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ the associated empirical measure. We saw in Section 2 that μ_n converges to the equilibrium measure μ_V, with a large deviations principle at speed β_n. We are able to give non-asymptotic guarantees on this convergence via a concentration inequality, which is the main result of the paper.

Theorem 3.5 (Concentration inequality)**.** *Let K satisfy Assumptions 1 and 2, and V satisfy Assumption 3. Further assume that the associated equilibrium measure μ_V has finite entropy. Let $\beta_n \geq 2cn$, where c is the constant of Assumption 3. Then there exist constants $c_0, c_1, c_2 > 0$, depending on K and V, such that for any $n \geq 2$ and for any $r > 0$,*

$$\begin{aligned}&\mathbb{P}^V_{n,\beta_n}\left(I_K\left(\mu_n - \mu_V\right) > r^2\right)\\ &\qquad \leq \exp\left(-\frac{c_0}{2}\beta_n r^2 + n(c_1 + \frac{\beta_n}{n^2}c_2)\right).\end{aligned} \tag{12}$$

We emphasize again that by Proposition 3.2, (12) 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 interesting 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 $\beta > 0$. We can rephrase a bit to get a more explicit sub-Gaussian decay in the bound.

Corollary 3.6. *Under the assumptions of Theorem 3.5, let further $\beta_n \gg n$. Then there exist constants $u_0, u_1 > 0$ such that for any $n \geq 2$ and for any*

$$r \geq u_0 \max\left(n^{-1/2}, (\beta_n/n)^{-1/2}\right), \tag{13}$$

$$\mathbb{P}^V_{n,\beta_n}\left(I_K\left(\mu_n - \mu_V\right) > r^2\right) \leq \exp\left(-u_1\beta_n r^2\right). \tag{14}$$

In particular, when $\beta_n \geq vn^2$ for some constant $v > 0$, Condition (13) simply becomes $r > u_0 n^{-1/2}$. The proof of Corollary 3.6 is straightforward from Theorem 3.5, itself proved in details in Appendix A.

We thus recover the known *dimension-independent* decay in $n^{-1/2}$ of the worst-case quadrature error as proved for deterministic herding (Bach et al., 2012), though for our probabilistic relaxation only with *very large* probability, and towards the equilibrium measure. The meaning of *very large* is that under i.i.d. sampling, analogous results such as Equation (3) feature n instead of β_n in the right-hand side of (14), and $\beta_n/n \to +\infty$. This fast-increasing coverage probability of our confidence interval is a trace of the repulsion introduced in the Gibbs measure.

We now explain in concrete terms how Corollary 3.6 implies that Monte Carlo integration with $\mathbb{P}^V_{n,\beta_n}$ and with respect to a target distribution π outperforms crude Monte Carlo.

3.3 Application to guarantees for probabilistic herding

Let $d \geq 1$ and π be a probability measure on $\mathbb{R}^d$, which we assume to be our target. To apply Theorem 3.5, we shall work under the following assumption.

Assumption 4. 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 $-\int \log \pi'(x) \mathrm{d}\pi(x) < \infty$.

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 $U^\pi_K(z)$ for all $z \in \mathbb{R}^d$.

Proposition 3.7. *Let K satisfying Assumptions 1 and 2 be fixed. Let π satisfy Assumption 4. 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, nonnegative function such that* $\Phi = 0$ *on* $\partial B(0,R)$, $\Phi(z) \to +\infty$ *as* $|z| \to +\infty$ *and*

$$\int_{\{|x|>R\}} e^{-\Phi(x)}\,\mathrm{d}x < \infty.$$

Then, setting $V^\pi(z) = -U_K^\pi(z)$ *when* $z \in B(0,R)$ *and* $V^\pi(z) = -U_K^\pi(z) + \Phi(z)$ *otherwise,* V^π *satisfies Assumption 3 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 Appendix A, 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 3.7 says that a suitably penalized kernel embedding is a good choice of confining potential. 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 (Chen et al., 2010; Bach et al., 2012).

With π now our equilibrium measure, Corollary 3.6 implies a uniform quadrature guarantees with respect to π.

Corollary 3.8. *Let* K *satisfying Assumptions 1 and 2 be fixed. Let* π *satisfy Assumption 4. Set* $V = V^\pi$ *as in Proposition 3.7, assume that* $\beta_n \gg n$, *and let* $x_1, \ldots, x_n \sim \mathbb{P}^V_{n,\beta_n}$. *Fix further* $\epsilon > 0$, *and* $\delta \in (0,1)$. *For any* n *big enough so that* $n^{-1/2} \leq \epsilon$ *and* $\exp\left(-c\beta_n\epsilon^2\right) \leq \delta$, *with probability larger than* $1-\delta$,

$$\sup_{\|f\|_{\mathcal{H}_K}\leq 1} \left| \int f \,\mathrm{d}(\mu_n - \pi) \right| \leq \epsilon. \tag{15}$$

.

The proof of Corollary 3.8 is a direct application of Corollary 3.6 and Proposition 3.7. Compared to Corollary 2.2, for a fixed uniform worst-case integration error ϵ and confidence level δ, fewer points are required under $\mathbb{P}^V_{n,\beta_n}$ than under i.i.d samples from π. In particular, β_n replaces n in the constraint that links δ and ϵ.

In the (admittedly limited) sense of Corollary 3.8, our probabilistic relaxation of kernel herding provably outperforms i.i.d. sampling, under generic assumptions on the underlying RKHS. In contrast, for deterministic IPM minimization algorithms, we recall that better guarantees than i.i.d sampling were only proved so far in the case where $\mathcal{H}_K$ is finite-dimensional, which is quite restrictive in that it forbids most classical kernels, such as the Gaussian.

4 Experiments

We explain how we approximately sample from $\mathbb{P}^V_{n,\beta_n}$, and then perform two toy experiments: one to illustrate Theorem 3.5, and one to assess whether our Gibbs measure might come with a better convergence rate for single integrals.

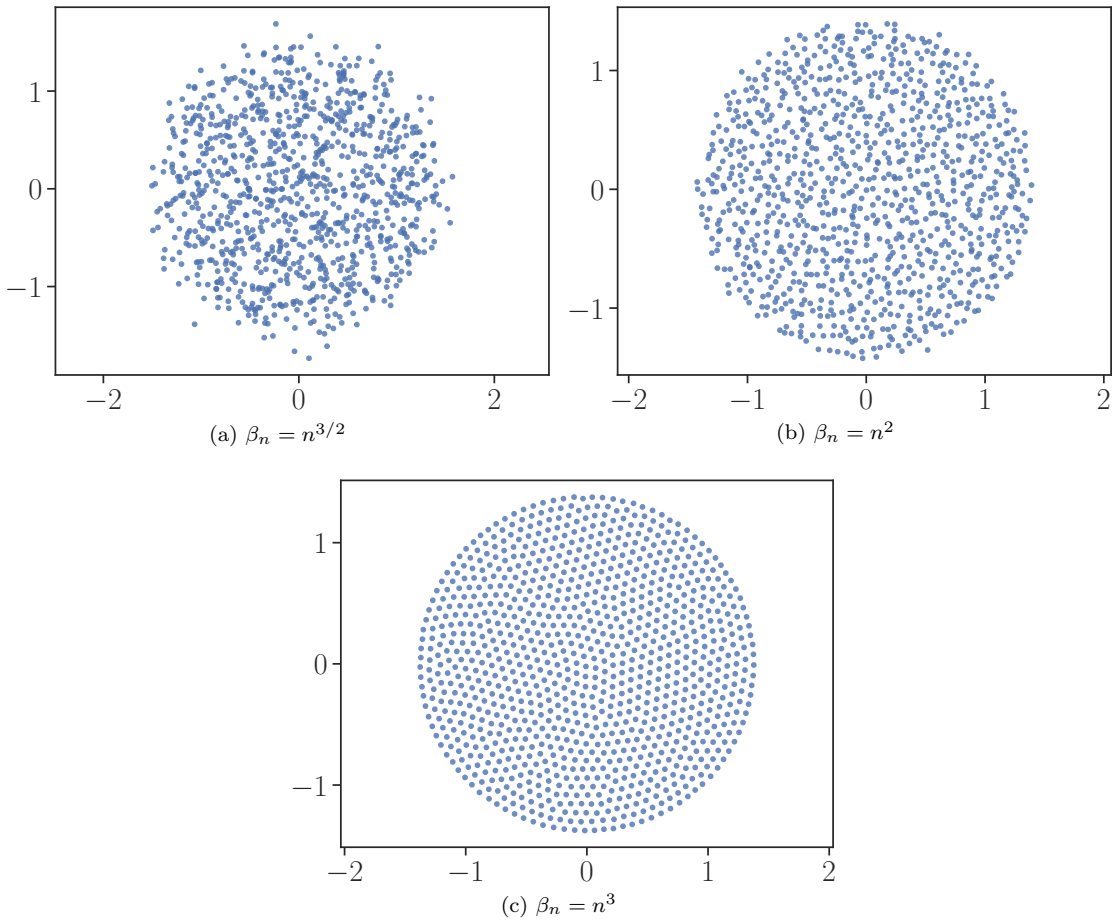
4.1 Approximately sampling from $\mathbb{P}^V_{n,\beta_n}$

There is no known algorithm to sample from (6) for a generic choice of K and V, so we resort to MCMC, namely the Metropolis-adjusted Langevin algorithm (MALA; (Robert & Casella, 2004)). While Chafaï & Ferré (2018) report having to tame the gradients in their experiments on singular kernels, vanilla MALA has been in our experience enough to get a good approximation to $\mathbb{P}^V_{n,\beta_n}$ for (smooth) bounded kernels such as the Gaussian and the truncated Riesz kernel. To wit, MALA uses a Metropolis–Hastings Markov kernel with proposal

$$y|y_t \sim \mathcal{N}\left(y_t - \alpha\beta_n \nabla H_n(y_t), 2\alpha I_{dn}\right) \tag{16}$$

where $y \in (\mathbb{R}^d)^n$, and α is a user-tuned step size parameter.

We first show 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



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

Figure 1: Three independent approximate samples of $\mathbb{P}^V_{n,\beta_n}$ corresponding to different temperature schedules.

each panel of Figure 1, we show the state of the MALA chain after $T = 5000$ iterations, with step size[6] $\alpha = \alpha_0 \beta_n^{-1}$, and α_0 is manually tuned at the beginning of each run so that acceptance reaches 50%.

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 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 1c is a manifestation of what physicists call *crystallization* (Serfaty, 2023): the Gibbs measure is concentrated around minimizers of the energy.

4.2 Comparing worst-case 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 $\mathbb{P}^V_{n,\beta_n}$. For the latter, we run MALA for $T = 5\,000$ iterations, with step size tuned as in Section 4.1.

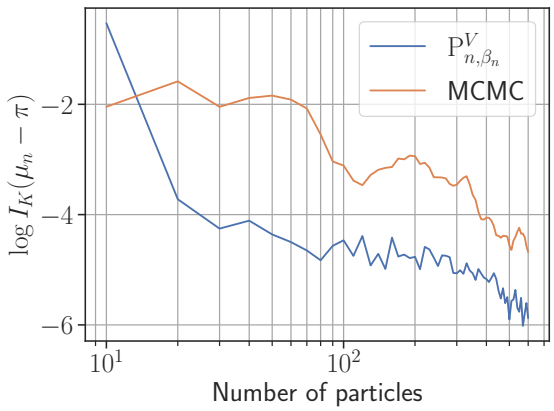
We consider two interaction kernels, the Gaussian kernel

$$K_1(x, y) = \exp\left(-|x - y|^2/2\right)$$

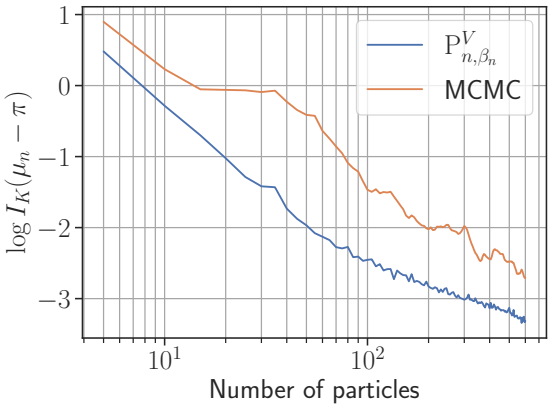
and the truncated Riesz kernel

$$K_2(x, y) = \left(|x - y|^2 + 0.1^2\right)^{-(d-2)/2}.$$

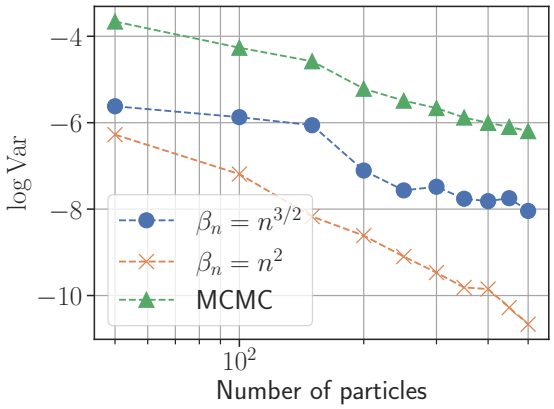
[6] Having α decrease at least as β_n^{-1} intuitively avoids the distance between two consecutive MALA states to grow with n; see the MALA proposal (16).



(a) Energy for the Gaussian kernel



(b) Energy for the truncated Riesz kernel



(c) Variance for the truncated Riesz kernel

Figure 2: Energy and variance comparisons between approximate samples from $\mathbb{P}^V_{n,\beta_n}$ and MCMC samples.

For $K \in \{K_1, K_2\}$, we set V to V^π as in Proposition 3.7, with $U^\pi_K(\cdot)$ replaced by a Monte Carlo approximation $M^{-1}\sum_{i=1}^M K(\cdot, z_i)$, 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. This approximation is usual in kernel herding experiments (Pronzato & Zhigljavsky, 2020).

Figure 2b and 2a 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 of the T-th iterate of our MALA chain targeting $\mathbb{P}^V_{n,\beta_n}$. The approximation results from writing

$$I_K(\mu - \pi) = I_K(\mu) - 2I_K(\mu, \pi) + I_K(\pi), \tag{17}$$

in which the first term of (17) can be computed exactly when μ has finite support, but the other terms require independent MH samples targeting π, here of length $10,000$.

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 (3.6) recovers the n^{-1} rate for the energy $I_K(\mu_n - \pi)$ under $\mathbb{P}^V_{n,\beta_n}$, the improvement rather being on the sub-Gaussian decay. We see nonetheless that the approximated energy (and hence, the worst-case integration error) is always smaller by about 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.

4.3 Comparing variances for a single integrand

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} as n grows, under appropriate assumptions on f. While this is not at all implied by our Corollary 3.6, analogies with the statistical physics literature make us expect a CLT to hold for $\mathbb{P}^V_{n,\beta_n}$, 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 the same setting as in Section 4.2: 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 2c, we run 100 independent MH chains of length n targeting π, and plot the empirical variance of the Monte Carlo estimator of $\int f \mathrm{d}\pi$. Similarly, we run 100 independent MALA chains targeting $\mathbb{P}^V_{n,\beta_n}$ with α tuned as in Section 4.2, for $T = 5000$ iterations each, and plot the 100 empirical variances obtained from each T-th sample. Again, $U^\pi_{K_2}$ is approximated through long MH chains.

We observe in Figure 2c 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 for $\mathbb{P}^V_{n,\beta_n}$.

5 Discussion

Using a Gibbs measure that favors nodes that repel according to a kernel K, we improved on the non-asymptotic worst-case integration guarantees of crude Monte Carlo, through a concentration inequality with a fairly easy proof, at least compared to the classical results in statistical physics that inspired us. When the RKHS of K is infinite-dimensional, such an improvement has yet to be proved for deterministic IPM minimization. A strong argument in favor of a Gibbs measure would be a CLT with a fast rate. We show experimental evidence that supports this expectation.

Limitations of our approach that deserve further inquiry are the impact of using an approximation of the kernel embedding U^π_K and an MCMC sampler, here MALA. Integrating a tractable approximation of U^π_K without loss on the convergence speed would be an important improvement. Simultaneously, 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.

Finally, compared to the bound (5) for volume sampling, our concentration bound features β_n, but not the eigenvalues of the kernel operator. In that sense, our confidence intervals are likely to be looser in an RKHS with fast-decaying kernel eigenvalues. Yet, establishing a CLT for the estimator in (5), where the weights in the estimator depend on all quadrature nodes, promises to be particularly hard. Moreover, if both distributions are approximately sampled through MCMC, one evaluation of our Gibbs density is quadratic in n, while it is cubic for volume sampling. A careful experimental comparison at fixed budget would thus be interesting.

Acknowledgments

We acknowledge support from ERC grant Blackjack ERC-2019-STG-851866, ANR grant Baccarat ANR-20-CHIA-0002 and Labex CEMPI (ANR-11-LABX-0007-01).

References

Alquier, P., Ridgway, J., and Chopin, N. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., Mackey, L., Oates, C. J., Reinert, G., and Swan, Y. Stein's Method Meets Computational Statistics: A Review of Some Recent Developments. *Statistical Science*, 38(1):120 – 139, 2023. doi: 10.1214/22-STS863.

Bach, F. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

Bach, F., Lacoste-Julien, S., and Obozinski, G. On the Equivalence between Herding and Conditional Gradient Algorithms. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1355–1362, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Bauerschmidt, R., Bourgade, P., Nikula, M., and Yau, H.-T. The two-dimensional Coulomb plasma: quasi-free approximation and central limit theorem. *arXiv preprint arXiv:1609.08582*, 2016.

Belhadji, A., Bardenet, R., and Chainais, P. Kernel interpolation with continuous volume sampling. In *International Conference on Machine Learning (ICML)*, 2020.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Bolley, F., Guillin, A., and Villani, C. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137:541–593, 2007.

Chafaï, D., Gozlan, N., and Zitt, P.-A. First-order global asymptotics for confined particles with singular pair repulsion. *Ann. Appl. Probab.*, 24(6):2371–2413, 2014. ISSN 1050-5164. doi: 10.1214/13-AAP980.

Chafaï, D., Hardy, A., and Maïda, M. Concentration for Coulomb gases and Coulomb transport inequalities. *J. Funct. Anal.*, 275(6):1447–1483, 2018. ISSN 0022-1236. doi: 10.1016/j.jfa.2018.06.004.

Chafaï, D. and Ferré, G. Simulating Coulomb gases and log-gases with hybrid Monte Carlo algorithms. 06 2018.

Chen, Y. and Welling, M. Parametric Herding. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 97–104, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Chen, Y., Welling, M., and Smola, A. Super-Samples from Kernel Herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pp. 109–116, Arlington, Virginia, USA, 2010. AUAI Press. ISBN 9780974903965.

Dembo, A. and Zeitouni, O. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.

Dick, J., Kuo, F. Y., and Sloan, I. H. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

García-Zelada, D. and Padilla-Garza, D. Generalized transport inequalities and concentration bounds for Riesz-type gases. *arXiv preprint arXiv:2209.00587*, 2022.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:4672–4682, 2020.

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. Kernel stein discrepancy descent. In *International Conference on Machine Learning*, pp. 5719–5730. PMLR, 2021.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447, 2022.

Leblé, T. and Serfaty, S. Fluctuations of two dimensional Coulomb gases. *Geometric and Functional Analysis*, 28:443–508, 2018.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Murphy, K. P. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

Pronzato, L. Performance analysis of greedy algorithms for minimising a Maximum Mean Discrepancy. *Statistics and Computing*, 33(1):14, 2023.

Pronzato, L. and Zhigljavsky, A. Bayesian quadrature, energy minimization, and space-filling design. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):959–1011, 2020.

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

Robert, C. P. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Robert, C. P. and Casella, G. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004. ISBN 0-387-21239-6. doi: 10.1007/978-1-4757-4145-2.

Serfaty, S. Systems with Coulomb interaction. *Notices of the American Mathematical Society*, 65(7): 787–788, August 2018. ISSN 0002-9920. doi: 10.1090/noti1697.

Serfaty, S. Gaussian fluctuations and free energy expansion for Coulomb gases at any temperature. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 59, pp. 1074–1142. Institut Henri Poincaré, 2023.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010. ISSN 1532-4435.

Welling, M. Herding Dynamical Weights to Learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1121–1128, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553517.

Welling, M. Herding dynamic weights for partially observed random field models. *arXiv preprint arXiv:1205.2605*, 2012.

A Proofs

A.1 Proof of Proposition 3.3

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 ISDP assumption on the kernel, using Lemma 3.1 in (Pronzato & Zhigljavsky, 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).

A.2 Proof of Proposition 3.7

The proof of Proposition 3.7 relies on the so-called Euler-Lagrange equations, which we recall here.

Lemma A.1. *Let K be a kernel satisfying Assumptions 1 and 2, and V be an external potential satisfying Assumption 3. Set $C_V = I_K(\mu_V) + \int V \, \mathrm{d}\mu_V$. Then 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 3.3 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 $\nu \in \mathcal{E}_K^V$,

$$\int (V + U_K^{\mu_V} - C_V) \, \mathrm{d}\nu \geq 0.$$

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 3.3. □

We are now ready to prove Proposition 3.7, by checking that the Euler-Lagrange equations are satisfied for π and V^π, and that V^π satisfies the assumptions of Proposition 3.3.

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 $V^\pi(z) \longrightarrow +\infty$ when $|z| \longrightarrow +\infty$. Finally, the integrability assumption of Assumption 3 is satisfied by our assumption on Φ and because U_K^π is bounded. Hence, V^π satisfies Assumption 3 and the equilibrium measure μ_{V^π} is well-defined. We conclude upon noting that, since $\Phi \geq 0$, Lemma A.1 yields that $\mu_{V^\pi} = \pi$.

A.3 Proof of Theorem 3.5

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_{n,β_n}^V, which generalizes the one of (Chafaï et al., 2018) to bounded kernels.

Proposition A.2. *Let K be a kernel satisfying Assumptions Assumption 1 and Assumption 2, and V be an external potential satisfying Assumption 3. Assume that the associated equilibrium measure μ_V has finite entropy, i.e. $S(\mu_V) = -\int \log \mu_V' \mathrm{d}\mu_V < \infty$, where μ_V' is the density of μ_V w.r.t. the Lebesgue measure. Then for $n \geq 2$, we have*

$$Z^V_{n,\beta_n} \geq \exp\left\{-\beta_n I^V_K(\mu_V) + n\left(\frac{\beta_n}{2n^2} I_K(\mu_V) + S(\mu_V)\right)\right\}.$$

Proof. The idea is to rephrase a bit the discrete energy H_n and use Jensen's inequality.

We let $X_n = (x_1, \dots, x_n)$ for brevity, and we start by writing

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

Then

$$\begin{aligned}\log Z^V_{n,\beta_n} &= \log \int_{(\mathbb{R}^d)^n} \exp\left(-\beta_n H_n(X_n)\right) \mathrm{d}x_1 \dots \mathrm{d}x_n \\ &\geq \log \int_{E_V^n} \exp\left(-\frac{\beta_n}{2n^2}\sum_{i\neq j}\{K(x_i,x_j)+V(x_i)+V(x_j)\} - \sum_{i=1}^n\left(\frac{\beta_n}{n^2}V(x_i) + \log\mu_V'(x_i)\right)\right)\mathrm{d}\mu_V(x_1)\dots\mathrm{d}\mu_V(x_n),\end{aligned}$$

where $E_V^n = \{(x_1,\dots,x_n) \in (\mathbb{R}^d)^n : \prod_{i=1}^n \mu_V'(x_i) > 0\}$. Using Jensen's inequality, we get

$$\begin{aligned}\log Z^V_{n,\beta_n} &\geq -\frac{\beta_n}{n^2}\sum_{i\neq j}\frac{1}{2}\int_{E_V^n}(K(x_i,x_j)+V(x_i)+V(x_j))\,\mathrm{d}\mu_V(x_1)\dots\mathrm{d}\mu_V(x_n) \\ &\quad - \sum_{i=1}^n \int_{E_V^n}\left(\frac{\beta_n}{n^2}V(x_i) + \log\mu_V'(x_i)\right)\mathrm{d}\mu_V(x_1)\dots\mathrm{d}\mu_V(x_n) \\ &= -\frac{\beta_n}{n^2}n(n-1)I^V_K(\mu_V) - \frac{\beta_n}{n}\int V\mathrm{d}\mu_V + nS(\mu_V).\end{aligned}$$

Using the definition $I^V_K(\mu_V) = \frac{1}{2}I_K(\mu_V) + \int V\mathrm{d}\mu_V$ 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 A.3. *Let K and V be a kernel and an external potential satisfying Assumptions 1 to 3. Let μ be any probability measure of finite energy, $I^V_K(\mu) < +\infty$. Then*

$$I_K(\mu - \mu_V) \leq 2\left(I^V_K(\mu) - I^V_K(\mu_V)\right).$$

Proof. We write

$$I_K(\mu-\mu_V) = \iint K(x,y)\,\mathrm{d}(\mu-\mu_V)^{\otimes 2} = I_K(\mu) - 2I_K(\mu,\mu_V) + I_K(\mu_V). \tag{18}$$

With the notation of Lemma A.1, we know that $U^{\mu_V}_K(z) + V(z) = C_V$ for all z in the support of μ_V, and that $U^{\mu_V}_K(z) + V(z) \geq C_V$ in general. Using Fubini, we thus see that

$$\begin{aligned}I_K(\mu,\mu_V) + \int V\,\mathrm{d}\mu &= \int\{U^{\mu_V}_K(z) + V(z)\}\,\mathrm{d}\mu(z) \\ &\geq C_V = I_K(\mu_V) + \int V\,\mathrm{d}\mu_V.\end{aligned}$$

Plugging this into (18) yields the result. □

We are now ready to prove Theorem 3.5. Recall that we work under the assumption $\beta_n \geq 2cn$ where c is the constant of Assumption 3. 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 $X_n = (x_1,\dots,x_n)$.

Let $\eta > 0$ and $X_n \in A$. The key idea is to split $H_n(X_n) = (1-\eta)H_n(X_n) + \eta H_n(X_n)$. 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

$$n^2 H_n(X_n) = n^2 I_K^V(\mu_n) - \frac{1}{2}\sum_{i=1}^n K(x_i, x_i),$$

we write, by definition,

$$\begin{aligned}
\mathbb{P}^V_{n,\beta_n}(A) &= \frac{1}{Z^V_{n,\beta_n}} \int_A \exp\left(-\beta_n H_n(X_n)\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n \qquad (19)\\
&= \frac{1}{Z^V_{n,\beta_n}} \exp\left(-\beta_n(1-\eta) I_K^V(\mu_V)\right) \int_A \exp\left(-\beta_n(1-\eta)\left(I_K^V(\mu_n) - I_K^V(\mu_V)\right)\right)\\
&\quad \times \exp\left(\frac{\beta_n}{2n^2}(1-\eta)\sum_{i=1}^n K(x_i,x_i) - \beta_n \eta H_n(X_n)\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n. \qquad (20)
\end{aligned}$$

Using Proposition A.2, we continue

$$\begin{aligned}
\mathbb{P}^V_{n,\beta_n}(A) &\le \exp\left(-n\left(S(\mu_V) + \frac{\beta_n}{2n^2} I_K(\mu_V)\right) + \beta_n \eta I_K^V(\mu_V)\right) \qquad (21)\\
&\quad \times \exp\left(-\beta_n(1-\eta)\inf_A\left(I_K^V(\mu_n) - I_K^V(\mu_V)\right)\right) \qquad (22)\\
&\quad \times \int_{\mathbb{R}^d} \exp\left(\frac{\beta_n}{2n^2}(1-\eta)\sum_{i=1}^n K(x_i,x_i) - \beta_n\eta H_n(X_n)\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n. \qquad (23)
\end{aligned}$$

As noted in the proof of Lemma A.1, $0 \le K \le C$, so that the last integral in (23) is easily bounded,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \exp\left(\frac{\beta_n}{2n^2}(1-\eta)\sum_{i=1}^n K(x_i,x_i) - \beta_n\eta H_n(X_n)\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n\\
&\quad \le \exp\left(\frac{\beta_n}{2n}(1-\eta)C\right)\left(\int_{\mathbb{R}^d} \exp\left(-\frac{\beta_n}{n}\eta V(x)\right) \mathrm{d}x\right)^n.
\end{aligned}$$

Now we choose a particular value for η, namely $\eta = cn/\beta_n$ where c is the constant of Assumption 3. Further let $C_2 = \log \int_{\mathbb{R}^d} \exp\left(-cV(x)\right) \mathrm{d}x$, which is finite by assumption. Setting $c_1 = cI_K^V(\mu_V) + C_2 - S(\mu_V)$ and $c_2 = \frac{1}{2}C - \frac{1}{2}I_K(\mu_V)$, (23) yields

$$\mathbb{P}^V_{n,\beta_n}(A) \le \exp\left(-(\beta_n - cn)\inf_A\left(I_K^V(\mu_n) - I_K^V(\mu_V)\right) + nc_1 + \frac{\beta_n}{n}c_2\right). \qquad (24)$$

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 A.3 applied to (24), with

$$A = \{I_K(\mu_n - \mu_V) > r^2; \mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}\},$$

yields Theorem 3.5, upon noting that $\beta_n - cn \ge \beta_n/2$ by assumption.