Gibbs Gradient Descent: A Langevin Approach to Optimization

Oussama Zekri
CREST, ENSAE
Institut Polytechnique de Paris
Paris, France
oussama.zekri@ensae.fr

Anna Korba
CREST, ENSAE
Institut Polytechnique de Paris
Paris, France
anna.korba@ensae.fr

Nicolas Boullé
Department of Mathematics
Imperial College London
London, SW7 2AZ, UK
n.boulle@imperial.ac.uk

Abstract

Training energy-based models and estimating parameters in latent-variable models often require evaluating gradients through sampling from Gibbs distributions, leading to costly nested sampling and optimization procedures. Gibbs gradient descent (GGD) addresses this difficulty by coupling Langevin sampling and parameter updates in a single loop. We analyze its continuous-time dynamics for nonconvex objectives, establishing convergence bounds for the ideal infinite-particle flow and its finite-particle approximation. These bounds separate a time-dependent optimization term from a finite-particle contribution of order $1/n$, where n is the number of particles. Under additional regularity assumptions, a central limit theorem identifies the asymptotic covariance of the gradient error, revealing how fluctuations in the empirical expectations propagate through the gradient computation. This characterization motivates a control-variate version of GGD that corrects these expectations before evaluating the gradient. For gradients depending on finitely many expectations, we establish variance reduction for the corrected moment estimates at equilibrium. Experiments on a synthetic problem and a neural energy-based model demonstrate reduced estimator variance and improved optimization performance.

1 Introduction

Minimizing a functional on a space of parametric probability distributions defined in Gibbs form is a fundamental task underpinning many modern machine learning methods. This includes for instance maximum likelihood estimation for latent variable models (Kuntz et al., 2023), or variational inference with energy-based models (Tomczak, 2022). In this paper, we consider solving a (potentially) *non-convex* optimization problem of the form:

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := \mathcal{F}(\pi(\theta)) \tag{1}$$

where $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$ is a functional defined over the set of probability distributions over $\mathbb{R}^d$, denoted $\mathcal{P}(\mathbb{R}^d)$, and $\pi(\theta)$ is the Gibbs distribution associated to a potential $V(\cdot, \theta) \in C^1(\mathbb{R}^d; \mathbb{R})$. Namely, the density of $\pi(\theta)$ (also denoted $\pi(\theta)$) is given by $\pi(\theta) = \mathrm{e}^{-V(\cdot,\theta)}/Z_\theta$, where $Z_\theta = \int_{\mathbb{R}^d} \mathrm{e}^{-V(x,\theta)} \, \mathrm{d}x$ is a normalization constant.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Gibbs gradient descent (GGD). One usually solves Eq. (1) using a first-order optimization algorithm (e.g., gradient descent), where the computation of $\nabla \ell(\theta)$ requires the knowledge of $\pi(\theta)$. We assume that the gradient of $\ell : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ satisfies the *Gibbs gradient* condition: there exists $\Gamma : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^p \to \mathbb{R}^p$ such that $\nabla \ell(\theta) = \Gamma(\pi(\theta), \theta)$. This condition expresses the gradient of the loss as a function of both the current distribution and the model parameters. In many applications, i.e., for many functionals, the objective and its gradient typically involves expectations with respect to the intractable current Gibbs distribution $\pi(\theta)$, leading to a tight coupling between sampling and optimization. Indeed, if the latter gradient writes as an expectation over $\pi(\theta)$, this amounts to solving a sampling problem (i.e., acquiring fresh samples from $\pi(\theta)$) at each step. However, this may be costly if Z_θ is intractable and sampling itself requires running a Langevin Monte Carlo or another Markov chain Monte Carlo (MCMC) procedure, leading to a nested loop procedure.

The single-loop approach proposed by Marion et al. (2025), inspired by bilevel optimization methods, avoids the nested-loop computational bottleneck by performing sampling (from the current Gibbs distribution) and optimization (of the parameter) at the same time. We will refer to this approach as *Gibbs gradient descent* (GGD). In continuous time, it writes as the following scheme:

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad X_t \sim p_t, \tag{2a}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\,\mathrm{d}t, \tag{2b}$$

where $\theta_0 \in \mathbb{R}^p$, $t \geq 0$, and X_0 is distributed as $p_0 \in \mathcal{P}(\mathbb{R}^d)$. Eq. (2) is a coupled system with two different time scales: Eq. (2a) can be referred to as an inner loop (the fast system used to sample from the current target Gibbs distribution) and Eq. (2b) as an outer loop (the slow system used to update the parameters to optimize the loss). The parameter $\varepsilon_t > 0$ determines how fast the inner loop must be run with respect to t and is typically smaller than 1. Obtaining convergence results of the law p_t of the process X_t of Eq. (1) in this setting is challenging, due to the coupling between the sampling and the optimization processes.

In practice, GGD cannot be implemented exactly as the law p_t, required in the evolution (2b) through the drift $\Gamma(p_t, \theta_t)$, is typically unknown. It can only be approximated by a system of n interacting particles following Eq. (2a), and using its empirical measure as a proxy for Eq. (2b). Hence, the practical finite-particle counterpart of GDD writes

$$\mathrm{d}X_t^i = -\nabla_1 V(X_t^i, \theta_t^{(n)})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^i, \tag{3a}$$

$$\mathrm{d}\theta_t^{(n)} = -\varepsilon_t \Gamma(p_t^{(n)}, \theta_t^{(n)})\,\mathrm{d}t, \quad p_t^{(n)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^i} \tag{3b}$$

where $X_0^1, \ldots, X_0^n$ are i.i.d. as $p_0 \in \mathcal{P}(\mathbb{R}^d)$ and $\theta_0^{(n)} = \theta_0 \in \mathbb{R}^p$. Note that X_t^i is marginally distributed according to the same law $q_{t,n}$ for $1 \leq i \leq n$[1]. However, since the particles interact through the empirical measure $p_t^{(n)}$, they do not evolve independently, and $q_{t,n} \neq p_t$.

Contributions. We analyze the convergence of Gibbs gradient descent and study the effect of finite-particle approximations, both theoretically and empirically. Our main contributions are as follows:

1) We prove, in the non convex setting, convergence rates in continuous time, in terms of the averaged gradient norm of the loss with respect to time, for the ideal infinite-particle and practical finite-particle systems defined in Eqs. (2) and (3).
2) We provide a central limit theorem describing the fluctuations of the empirical distribution corresponding to the practical, finite-particle Gibbs descent, with respect to its infinite particle counterpart.
3) We consider Gibbs gradients that depend on finitely many expectations, a representation shared by our motivating examples. We use this structure to construct a control-variate version of GGD and establish variance reduction for the estimated expectations under independent sampling from a fixed Gibbs distribution.
4) Experiments on a synthetic problem and a neural energy-based model show that CV-GGD reduces estimator variance and improves optimization performance compared with GGD.

[1] The dependence in n in the notation $q_{t,n}$ emphasizes that the law of each particle depends on the n-particle system, yet it should not be understood as a discrete measure supported on n atoms (in contrast with $p_t^{(n)}$).

Organization of the paper. Section 2 introduces the framework and the regularity assumptions on the Gibbs gradient and the target distributions. Section 3 establishes convergence bounds for the ideal and finite-particle dynamics, then characterizes the particle fluctuations through a central limit theorem. Section 4 introduces CV-GGD and explains its variance-reduction principle. Section 5 presents numerical experiments on a one-dimensional model and a neural energy-based model. Additional examples, proofs, and experimental details are provided in the appendix.

2 Preliminaries

2.1 Related work

There exists a growing literature on the theoretical study of methods that couple sampling and optimization, that are referred here as single-loop but are sometimes also referred to as slow-fast systems. The idealized version presented in Eq. (2) of GGD (i.e., continuous time and infinite particle case) was analyzed by Marion et al. (2025). Closely related, Akyıldız et al. (2024) studied a system similar to Eq. (3), but that also incorporates a diffusion term in Eq. (3b) (controlled by a temperature parameter β). Also closely related, several works (Kuntz et al., 2023; Caprio et al., 2025; Caprio and Johansen, 2025; Akyıldız et al., 2025) considered similar schemes in the setting of maximum likelihood estimation (MLE) but that do not incorporate a parameter ε_t (hence assuming that the processes $(\theta_t)_{t\geq 0}$ and $(X_t)_{t\geq 0}$ are controlled by the same step-size in practice). In these works, the theoretical analysis of the coupling system is generally conducted in the case where the Gibbs distribution are log-concave, which is a strong assumption as it excludes multimodal distributions.

Coupled sampling and optimization methods also arise in reward-based training of Langevin samplers (Marion et al., 2025), entropy-regularized Wasserstein distributionally robust optimization (Mohajerin Esfahani and Kuhn, 2018; Wang et al., 2025; Azizian et al., 2023b,a), and interacting-particle training of latent diffusion models (Wang et al., 2026). A related Gibbs-gradient formulation underlies contrastive maximum-likelihood training of energy-based models (Gutmann and Hyvärinen, 2012; Du and Mordatch, 2019; Tomczak, 2022). In these examples, the gradient depends on finitely many expectations under the Gibbs law, which motivates the representation introduced below.

2.2 Regularity assumptions

In the following, we assume that the Gibbs gradient Γ admits the finite-observable representation:

$$\Gamma(p,\theta) = H_\theta\big(\mathbb{E}_p[\phi_1^\theta],\dots,\mathbb{E}_p[\phi_r^\theta]\big), \quad \mathbb{E}_p[\phi] \coloneqq \langle \phi, p\rangle, \tag{4}$$

where $r < \infty$, we denote $\langle \phi, p\rangle = \int_{\mathbb{R}^d} \phi(x)\, p(\mathrm{d}x)$, each observable $\phi_j^\theta : \mathbb{R}^d \to \mathbb{R}^{q_j}$ and $H_\theta : \mathbb{R}^{q_1}\times\cdots\times\mathbb{R}^{q_r} \to \mathbb{R}^p$ is referred to as the observable map. An interesting case arises by taking $r = 1$, $\phi_1^\theta(x) = f(x,\theta)$, and $H_\theta(z) = z$, and covers examples such as $\ell(\theta) = \mathrm{KL}(p_{\text{ref}} \mid\mid \pi(\theta))$ for some reference probability distribution $p_{\text{ref}} \in \mathcal{P}(\mathbb{R}^d)$ (Marion et al., 2025, Sec. 3.2), where KL denotes the Kullback-Leibler divergence. However, Eq. (4) includes more general instances, such as the reward-maximization problem written as the minimization of $\ell(\theta) = -\mathbb{E}_{\pi(\theta)}[R]$ for some reward function $R : \mathbb{R}^d \to \mathbb{R}$, which is obtained by selecting $\phi_1^\theta(x) = R(x)$, $\phi_2^\theta(x) = \nabla_2 V(x,\theta)$, $\phi_3^\theta(x) = R(x)\nabla_2 V(x,\theta)$, $H_\theta(\rho,u,w) = w - \rho u$. There, $r = 3$ with $q_1 = 1$ and $q_2 = q_3 = p$. The reader may refer to Section B.2 for detailed computations.

For $q = \sum_{j=1}^r q_j$, we identify $\mathbb{R}^{q_1}\times\cdots\times\mathbb{R}^{q_r}$ with $\mathbb{R}^q$ and write

$$\Phi^\theta(x) \coloneqq \big(\phi_1^\theta(x),\dots,\phi_r^\theta(x)\big), \qquad m_\theta(p) \coloneqq \mathbb{E}_p[\Phi^\theta] \in \mathbb{R}^q.$$

With this notation, Eq. (4) simply reads $\Gamma(p,\theta) = H_\theta(m_\theta(p))$.

Throughout this paper, we assume that the map $\theta \mapsto \pi(\theta)$ is differentiable and that $x \mapsto \nabla_2 V(x,\theta)$ is L_X-Lipschitz, uniformly in θ. We also impose Lipschitz continuity on Γ with respect to its measure argument, as defined below.

Definition 2.1 (g-Lipschitz continuity)**.** *Let $g : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$, we say that Γ is g-Lipschitz continuous if there exists $L_\Gamma > 0$ such that for $p, q \in \mathcal{P}(\mathbb{R}^d)$ and $\theta \in \mathbb{R}^p$,*

$$\|\Gamma(p, \theta) - \Gamma(q, \theta)\|_{\mathbb{R}^p} \leq L_\Gamma g(p, q).$$

When g is a norm, Definition 2.1 is equivalent to imposing Lipschitz continuity with respect to this norm. In this work, we will choose different discrepancies g depending on the regularity needed, such as $g(p, q) = \sqrt{\mathrm{KL}(p \parallel q)}$ or $g(p, q) = \|p - q\|$ for a certain norm on $\mathcal{P}(\mathbb{R}^d)$; e.g., Total Variation (TV) or Bounded Lipschitz (BL). Note that one has $\|p - q\|_{\mathrm{BL}} \leq \|p - q\|_{\mathrm{TV}} \leq \sqrt{2\,\mathrm{KL}(p||q)}$ by Pinsker's inequality (Lemma C.4), hence BL-Lipschitz implies $\sqrt{\mathrm{KL}}$-Lipschitz and is a stronger assumption. If Γ is of the form of Eq. (4), this condition is satisfied when the observables verify the corresponding regularity conditions and that the observable map is Lipschitz; see Remark C.1.

Finally, we assume that the parametric distributions iterate $\pi(\theta_t)$ and $\pi(\theta_t^{(n)})$ (related to Eqs. (2) and (3)) satisfy a log-Sobolev inequality (LSI) with constant $\mu > 0$ for $t > 0$ and $n \geq 1$.

Definition 2.2 (Log-Sobolev Inequality (LSI))**.** *A probability measure $q \in \mathcal{P}(\mathbb{R}^d)$ satisfies the Log-Sobolev inequality (LSI) with constant $\mu > 0$ if for every $p \in \mathcal{P}(\mathbb{R}^d)$,*

$$\mathrm{KL}(p \parallel q) \leq \frac{1}{2\mu} \|\nabla \log(p/q)\|^2_{L^2(p)}.$$

The LSI assumption generalizes the μ-strong convexity of the potentials $\{V(\cdot, \theta_t)\}_{t \geq 0}$, including distributions π whose potentials are bounded perturbations of a strongly convex potential (Bakry et al., 2014; Vempala and Wibisono, 2019).

3 Analysis of Gibbs Gradient Descent

In this section, we provide our results on the convergence of the coupled systems defined in Eqs. (2) and (3). We start by the infinite-particle case (Eq. (2)). Then for the finite-particle system (Eq. (3)) we first provide a non-asymptotic guarantee of convergence (with respect to the number of particles), then an asymptotic counterpart relying on a Central Limit Theorem.

3.1 Infinite-particle case

Note first that the infinite-particle case introduced in Eq. (2) correspond to a deterministic parameter dynamic. Our convergence result for the latter is the following.

Theorem 3.1 (Ideal Gibbs gradient flow)**.** *Let $\varepsilon > 0$, $0 \leq \delta < 1$, and select $\varepsilon_t = \min(\varepsilon, 1/t^\delta)$ in Eq. (2) for $t \geq 0$. Suppose that $\pi(\theta_t)$ satisfies the log-Sobolev inequality with constant $\mu > 0$ for all $t \geq 0$, with $\varepsilon < 2\mu^2/((L_\Gamma\sqrt{\mu/2} + L_X)^2 + 7L_X^2)$ if $\delta = 0$, and that Γ is $\sqrt{\mathrm{KL}}$-Lipschitz with constant L_Γ. Then, there exists $C > 0$, depending on $\mu, L_\Gamma, L_X, \varepsilon, \delta, \theta_0$, and p_0, such that*

$$\frac{1}{T}\int_0^T \|\nabla \ell(\theta_t)\|^2_{\mathbb{R}^p}\,\mathrm{d}t \leq \frac{C}{T^{1-\delta}}, \quad \textit{for all } T > 0.$$

Theorem 3.1 extends and sharpens the results of (Marion et al., 2025) to more general parameters ε_t (instead of $\delta = 1/2$) and Gibbs distributions that satisfy the log-Sobolev inequality. For $\delta = 1/2$, we obtain an asymptotic bound of $\mathcal{O}(T^{-1/2})$, which improves upon the $\mathcal{O}(\ln(T)^2 T^{-1/2})$-rate of (Marion et al., 2025, Thm. 4.5). Note that any (possibly stochastic) first-order method requires at least $T_\varepsilon = \Omega(\varepsilon^{-2})$ gradient evaluations to output a point $\theta_\star \in \mathbb{R}^p$ such that $\|\nabla \ell(\theta_\star)\|_{\mathbb{R}^p} \leq \varepsilon$ (Carmon et al., 2020, Thm. 2). Equivalently, after T gradient evaluations one necessarily has

$$\frac{1}{T}\int_0^T \|\nabla \ell(\theta_t)\|^2_{\mathbb{R}^p}\,\mathrm{d}t \geq \min_{0 \leq t < T} \|\nabla \ell(\theta_t)\|^2_{\mathbb{R}^p} = \Omega\left(T^{-1}\right).$$

Hence, the $O(T^{-1})$-rate of Theorem 3.1 matches that lower-bound for a constant ϵ (when $\delta = 0$). The lower bound of Carmon et al. (2020) is proved for discrete-time first-order methods, whereas our analysis evolves in continuous time. However, we believe that the proof techniques could extend to discrete-time systems.

3.2 Finite-particle case

We now study the finite n-particle dynamics in Eq. (3), which introduces stochasticity in the parameter dynamic. Throughout this section and Section 3.3, we consider the fully coupled, interacting system $(X_t^1, \dots, X_t^n)$, where the particles interact but remain exchangeable through the empirical–measure coupling. Given $s > d/2$, we recall that a probability measure $p \in \mathcal{P}(\mathbb{R}^d)$ belongs to the negative Sobolev space $H^{-s}(\mathbb{R}^d)$ if $\|p\|_{H^{-s}} := \sup_{\|\phi\|_{H^s} \leq 1} |\int_{\mathbb{R}^d} \phi(x)\, p(\mathrm{d}x)| < \infty$. Below, we require that Γ is H^{-s}-Lipschitz, which also implies the $\sqrt{\mathrm{KL}}$ control used in Theorem 3.1 by Lemma C.4.

Theorem 3.2 (Interacting n-particle system). *Let $n \geq 1$, $0 \leq \delta < 1$, and select $\varepsilon_t = \min(\varepsilon, 1/t^\delta)$ in Eq. (3) for $t \geq 0$. Suppose that $\pi(\theta_t^{(n)})$ satisfies the log-Sobolev inequality with constant μ for all $t \geq 0$, with $\varepsilon < 2\mu^2/((L_\Gamma\sqrt{\mu/2} + L_X)^2 + 7L_X^2)$ if $\delta = 0$, and that Γ is H^{-s}-Lipschitz for $s > d/2$, where L_Γ denotes the induced $\sqrt{\mathrm{KL}}$-Lipschitz constant. Then, for $T > 0$*

$$\frac{1}{T}\int_0^T \mathbb{E}\left[\|\nabla \ell(\theta_t^{(n)})\|_{\mathbb{R}^p}^2\right] \mathrm{d}t \leq C\left(\frac{1}{n} + \frac{1}{T^{1-\delta}}\right),$$

where $C > 0$ is a constant independent of n and T.

The non-asymptotic bound given by Theorem 3.2 guarantees that, for any fixed number of particles n in Eq. (2), the algorithm inherits the same optimization decay as the ideal flow (see Theorem 3.1), up to an additional $\mathcal{O}(1/n)$ term. This term reflects an irreducible variance plateau that cannot be eliminated by decreasing the stepsize. Indeed, for a simple one-dimensional Gaussian setup for the Gibbs distributions, Proposition 3.1 below shows that the stochastic gradient estimator $\Gamma(p_t^{(n)}, \theta_t^{(n)})$ retains a non-vanishing variance of order $\Omega(1/n)$ once the inner Langevin process reaches stationary.

Proposition 3.1 (Lower bound). *Let $\ell \in C^1(\mathbb{R}; \mathbb{R}^+)$ and $\theta_0 \in \mathbb{R}$ such that $\ell'(\theta_0) \neq 0$, and $n \geq 1$. There exist a potential $V : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ and $\Gamma : \mathcal{P}(\mathbb{R}) \times \mathbb{R} \to \mathbb{R}$, verifying the regularity assumptions of Theorem 3.2, such that*

$$\liminf_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{E}\left[\|\nabla \ell(\theta_t^{(n)})\|^2\right] \mathrm{d}t \geq \frac{c}{n},$$

for some constant $c > 0$, independent of n and T.

3.3 Asymptotic number of particles regime

The non-asymptotic bound in Theorem 3.2 contains an additional $\mathcal{O}(1/n)$ term arising from the finite-particle approximation. We now study the fluctuations responsible for this term and identify their limiting covariance. The particles are exchangeable but not independent, since they interact through the common parameter path $\theta_t^{(n)}$. The usual i.i.d. central limit theorem therefore does not apply directly. Instead, the joint process $(X_t^1, \dots, X_t^n, \theta_t^{(n)})$ fits the weakly interacting/common-factor framework of Budhiraja et al. (2017), which yields a Gaussian fluctuation limit on each fixed time interval $[0, T]$. We denote by $\mathrm{BL}(\mathbb{R}^d)$ the space of bounded Lipschitz functions from $\mathbb{R}^d$ to $\mathbb{R}$, and by $\Rightarrow$ convergence in distribution.

Proposition 3.2 (CLT for Gibbs gradient descent). *Fix $T > 0$. For $\phi \in \mathrm{BL}(\mathbb{R}^d)$ and $t \in [0, T]$, define*

$$\xi_t^{(n)}[\phi] := \sqrt{n}\,\langle \phi, p_t^{(n)} - p_t \rangle.$$

Under the regularity assumptions stated in Appendix D (Assumption D.1), for every $\ell \geq 1$, $t_1, \ldots, t_\ell \in [0,T]$, and $\phi_1, \ldots, \phi_\ell \in \mathrm{BL}(\mathbb{R}^d)$,

$$\big(\xi_{t_1}^{(n)}[\phi_1], \ldots, \xi_{t_\ell}^{(n)}[\phi_\ell]\big) \xRightarrow[n\to\infty]{} \big(\xi_{t_1}[\phi_1], \ldots, \xi_{t_\ell}[\phi_\ell]\big),$$

where the limit is a centered Gaussian vector.

Proposition 3.2 describes the fluctuations of empirical expectations around their mean-field limits. In the finite-observable setting of Eq. (4), the update $\Gamma(p,\theta) = H_\theta(m_\theta(p))$ depends on finitely many such expectations. Applying the CLT to the observables entering m_{θ_t}, then linearizing H_{θ_t}, gives a centered Gaussian limit Δ_t for the update error:

$$\sqrt{n}\big(\Gamma(p_t^{(n)},\theta_t) - \Gamma(p_t,\theta_t)\big) \xRightarrow[n\to\infty]{} \Delta_t \coloneqq DH_{\theta_t}(m_{\theta_t}(p_t))\big(\xi_t[\Phi_j^{\theta_t}]\big)_{j=1}^q.$$

Its covariance, denoted by Σ_t, describes how fluctuations in the empirical expectations propagate into the parameter update; see Proposition D.1.

To refine the finite-particle term in Theorem 3.2 using the covariance Σ_t, we must relate the fluctuation described by the CLT to the sampling error appearing in the optimization proof. Recalling that $\nabla\ell(\theta_t^{(n)}) = \Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)})$, the proof decomposes the gradient error as

$$\Gamma(p_t^{(n)},\theta_t^{(n)}) - \nabla\ell(\theta_t^{(n)}) = \underbrace{\Gamma(p_t^{(n)},\theta_t^{(n)}) - \Gamma(q_{t,n},\theta_t^{(n)})}_{\text{finite-particle sampling error}} + \underbrace{\Gamma(q_{t,n},\theta_t^{(n)}) - \nabla\ell(\theta_t^{(n)})}_{\text{error from not sampling at equilibrium}}.$$

Here, $q_{t,n}$ is the common marginal law of the interacting particles. We seek to characterize the first term using Σ_t. However, the CLT compares $p_t^{(n)}$ with p_t at the mean-field parameter θ_t, whereas this term compares $p_t^{(n)}$ with $q_{t,n}$ at $\theta_t^{(n)}$. Accounting for these differences introduces an additional corrective term. The estimates in Appendix D.4 show that its contribution to the time-integrated second moment vanishes as $n \to \infty$, for fixed T. The leading finite-particle contribution can therefore be expressed through Σ_t, with the remaining discrepancy included in $u_n(T)$ below.

Theorem 3.3 (Continuous-time, finite-particle case, asymptotic in n)**.** *Fix $T > 0$. Let $0 \leq \delta < 1$, and select $\varepsilon_t = \min(\varepsilon, 1/t^\delta)$ in Eq.* (3)*. Assume the hypotheses of Theorem 3.2 and the regularity assumptions stated in Appendix D (Assumption D.1) on $[0,T]$. Let Σ_t be the covariance defined above. Then there exists a nonnegative deterministic sequence $u_n(T)$, with $nu_n(T) \to 0$ as $n \to \infty$ for each fixed T, such that, for all $n \geq 1$,*

$$\frac{1}{T}\int_0^T \mathbb{E}\left[\|\nabla\ell(\theta_t^{(n)})\|_{\mathbb{R}^p}^2\right] \mathrm{d}t \leq \frac{66}{10nT^{1-\delta}}\int_0^T \varepsilon_t \,\mathrm{Tr}(\Sigma_t)\,\mathrm{d}t + \frac{C}{T^{1-\delta}} + u_n(T),$$

where $C > 0$ is independent of n and T.

Theorem 3.3 retains the optimization term of Theorem 3.2 and identifies the coefficient of the leading $1/n$ contribution through the covariance Σ_t. This characterization motivates reducing the fluctuations of the empirical expectations entering Γ. In the next section, we introduce a control-variate construction for these expectations.

4 Control-variate GGD

Section 3.3 identifies how fluctuations in empirical expectations contribute to the finite-particle error. In the finite-observable setting of Eq. (4), the update takes the form $\Gamma(p,\theta) = H_\theta(m_\theta(p))$, where $m_\theta(p)$ collects the expectations of the observables Φ^θ. This suggests reducing the variance of these estimates before applying H_θ.

We use the standard linear control-variate construction (see, e.g., Owen, 2013). Fix θ, let $X \sim \pi(\theta)$, and introduce a vector of control variates $\Psi^\theta : \mathbb{R}^d \to \mathbb{R}^d$ with known zero mean, $\mathbb{E}_{\pi(\theta)}[\Psi^\theta] = 0$. For a matrix $B \in \mathbb{R}^{q\times d}$, we define the corrected observable as

$$Y_B^\theta(X) = \Phi^\theta(X) - B\Psi^\theta(X).$$

The correction preserves the expectation of Φ^θ under $\pi(\theta)$. The following proposition gives a choice of B that also reduces its covariance. This guarantee holds at equilibrium, with θ fixed and $X \sim \pi(\theta)$.

Proposition 4.1 (Observable variance reduction)**.** *Fix $\theta \in \mathbb{R}^p$. Assume that Φ^θ and Ψ^θ have finite second moments under $\pi(\theta)$, with $\mathbb{E}_{\pi(\theta)}[\Psi^\theta] = 0$. Define*

$$B_\theta^{\pi(\theta)} = \operatorname*{Cov}_{\pi(\theta)}(\Phi^\theta, \Psi^\theta) \operatorname*{Cov}_{\pi(\theta)}(\Psi^\theta, \Psi^\theta)^\dagger,$$

where $\dagger$ denotes the Moore–Penrose pseudoinverse. Then the corrected observable $Y^\theta_{B_\theta^{\pi(\theta)}}$ has the same expectation as Φ^θ and satisfies

$$\operatorname{Tr}\left(\operatorname*{Cov}_{\pi(\theta)}(Y^\theta_{B_\theta^{\pi(\theta)}})\right) \le \operatorname{Tr}\left(\operatorname*{Cov}_{\pi(\theta)}(\Phi^\theta)\right).$$

Here, $\operatorname{Cov}_{\pi(\theta)}(\Phi^\theta, \Psi^\theta) \in \mathbb{R}^{q\times d}$ and $\operatorname{Cov}_{\pi(\theta)}(\Psi^\theta, \Psi^\theta) \in \mathbb{R}^{d\times d}$, so $B_\theta^{\pi(\theta)} \in \mathbb{R}^{q\times d}$.

Remark 4.1. *Let $m_\theta^\pi = m_\theta(\pi(\theta))$, and assume that H_θ is differentiable at m_θ^π. At equilibrium, for a fixed θ, the delta method transfers the observable fluctuations to the update through $DH_\theta(m_\theta^\pi)$. With the coefficient $B_\theta^{\pi(\theta)}$, the control-variate correction changes the limiting update covariance as follows:*

$$DH_\theta(m_\theta^\pi) \operatorname*{Cov}_{\pi(\theta)}(\Phi^\theta)\, DH_\theta(m_\theta^\pi)^\intercal \quad \longrightarrow \quad DH_\theta(m_\theta^\pi) \operatorname*{Cov}_{\pi(\theta)}\left(\Phi^\theta - B_\theta^{\pi(\theta)}\Psi^\theta\right) DH_\theta(m_\theta^\pi)^\intercal.$$

The covariance reduction established in the proof of Proposition 4.1 therefore also reduces the total variance of the limiting update fluctuation in this equilibrium setting.

Choosing control variates. One can construct control variates with zero mean under $\pi(\theta)$ using Stein's identity. Since $\pi(\theta)(\mathrm{d}x) \propto \exp(-V(x,\theta))\,\mathrm{d}x$, Stein's identity gives, for $1 \le j \le d$ and sufficiently regular vector field $h_j : \mathbb{R}^d \to \mathbb{R}^d$, the control variate $\Psi_j^\theta(x) = \nabla \cdot h_j(x) - h_j(x)^\intercal \nabla_1 V(x,\theta) \in \mathbb{R}$, with zero mean $\mathbb{E}_{\pi(\theta)}[\Psi_j^\theta] = 0$. In particular, taking constant vector fields $h_j = e_j \in \mathbb{R}^d$ gives the vector of score control variates $\Psi^\theta(x) = \nabla_1 V(x,\theta) \in \mathbb{R}^d$. The gradient $\nabla_1 V(x,\theta)$ is already computed for the Langevin particle updates, so these control variates require no additional gradient evaluations.

Algorithm. We implement the control-variate correction by estimating its coefficient from the particles at each iteration. At iteration k, after the Langevin particle update, write $\Phi_i = \Phi^{\theta_k}(X_{k+1}^i)$ and $\Psi_i = \Psi^{\theta_k}(X_{k+1}^i)$, suppressing the dependence on k for readability. Writing $\bar\Phi$ and $\bar\Psi$ for their empirical means, we compute

$$\widehat{C}_{\Phi\Psi,k} = \frac{1}{n}\sum_{i=1}^n (\Phi_i - \bar\Phi)(\Psi_i - \bar\Psi)^\intercal, \qquad \widehat{C}_{\Psi\Psi,k} = \frac{1}{n}\sum_{i=1}^n (\Psi_i - \bar\Psi)(\Psi_i - \bar\Psi)^\intercal.$$

Denote the empirical measure after the particle update by $p_{k+1}^{(n)} = \frac{1}{n}\sum_{i=1}^n \delta_{X_{k+1}^i}$. Using these empirical covariances, we construct a regularized approximation of the equilibrium coefficient $B_\theta^{\pi(\theta)}$ from Proposition 4.1:

$$B_{\theta_k}^{p_{k+1}^{(n)}} = \widehat{C}_{\Phi\Psi,k}\big(\widehat{C}_{\Psi\Psi,k} + \lambda_{\mathrm{cv}} I_d\big)^{-1},$$

where $\lambda_{\mathrm{cv}} > 0$ is a ridge parameter.

We then form the corrected moment estimate

$$\widehat{m}_k = \frac{1}{n}\sum_{i=1}^n \left(\Phi_i - B_{\theta_k}^{p_{k+1}^{(n)}}\Psi_i\right)$$

and use $\widehat{\Gamma}_k = H_{\theta_k}(\widehat{m}_k)$ in the parameter update

$$\theta_{k+1} = \theta_k - h\varepsilon_k \widehat{\Gamma}_k.$$

Setting $B_{\theta_k}^{p_{k+1}^{(n)}} = 0$ recovers standard GGD. The complete procedure is summarized in Algorithm 1.

Algorithm 1 Discrete-time (CV)-GGD in the finite-observable setting

Input: stepsize h, outer stepsizes $\{\varepsilon_k\}$, initial θ_0, particles $\{X_0^i\}_{i=1}^n$, controls Ψ^θ
for $k \leftarrow 0$ to $K-1$ **do**
  $X_{k+1}^i \leftarrow X_k^i - h\,\nabla_1 V(X_k^i, \theta_k) + \sqrt{2h}\,\xi_k^i$, where $\xi_k^i \sim \mathcal{N}(0, I_d)$ for $1 \le i \le n$
  $B_{\theta_k}^{p_{k+1}^{(n)}} \leftarrow 0$; estimate $B_{\theta_k}^{p_{k+1}^{(n)}} \in \mathbb{R}^{q\times d}$ by ridge covariance regression on $(\Phi^{\theta_k}, \Psi^{\theta_k})$
  $\widehat{m}_k \leftarrow \frac{1}{n}\sum_{i=1}^n \Big(\Phi^{\theta_k}(X_{k+1}^i) - B_{\theta_k}^{p_{k+1}^{(n)}}\Psi^{\theta_k}(X_{k+1}^i)\Big)$
  $\widehat{\Gamma}_k \leftarrow H_{\theta_k}(\widehat{m}_k)$
  $\theta_{k+1} \leftarrow \theta_k - h\,\varepsilon_k\,\widehat{\Gamma}_k$
end for
return θ_K

Scope of the guarantee. Proposition 4.1 establishes variance reduction at equilibrium, for a fixed parameter and the population coefficient $B_\theta^{\pi(\theta)}$. In Algorithm 1, however, the parameter changes at each iteration, while the particles only take one Langevin step toward the corresponding equilibrium distribution. The regularized coefficient is also estimated from the same particles used to compute the update. The proposition therefore does not directly guarantee variance reduction for the implemented estimator. Establishing such a guarantee remains open. Our experiments in the next section will demonstrate variance reduction in the settings considered.

5 Numerical experiments

Both experiments compare GGD with CV-GGD using the score control variates $\Psi^\theta = \nabla_1 V(\cdot, \theta)$ introduced in Section 4.

5.1 Finite-particle error in a one-dimensional example

Theorem 3.2 and Proposition 3.1 identify a contribution of order $1/n$ to the optimization error due to the finite number of particles. Our first experiment illustrates this dependence and examines whether control variates reduce the resulting error. We consider a one-dimensional example based on the lower-bound construction in Proposition 3.1, with the finite-observable form of Eq. (4). In this setting, the normalizing constant and expectations under $\pi(\theta)$ can be computed accurately by numerical quadrature, so evaluating the population gradient does not require a Langevin sampling subroutine. This allows us to evaluate the population gradient and estimate

$$\mathrm{LHS}_K^{(n)} := \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\|\nabla \ell(\theta_k^{(n)})\|^2\right],$$

the discrete-time counterpart of the quantity bounded in Theorem 3.2. CV-GGD uses the same Langevin particles for the parameter update and for the periodic refit of the control-variate coefficient. After a refit, the updated coefficient is first used at the next iteration. The model and numerical details are given in Appendix F.1.

Figure 1 compares standard GGD (left) with CV-GGD (right). For standard GGD, the averaged squared gradient initially decreases as the number of iterations increases, before reaching a plateau whose level decreases with the number of particles. The curves rescaled by the number of particles, $n \cdot \mathrm{LHS}_K^{(n)}$, shown in Fig. 3 of Appendix F.1, approach approximately the same nonzero plateau across the tested particle counts, consistent with the $1/n$ dependence predicted by Theorem 3.2 and Proposition 3.1. CV-GGD substantially lowers the plateau without eliminating it. Its rescaled curves also approach a common level, but with a smaller constant. These observations suggest that the empirical correction reduces the finite-particle error while preserving its $1/n$ scaling.

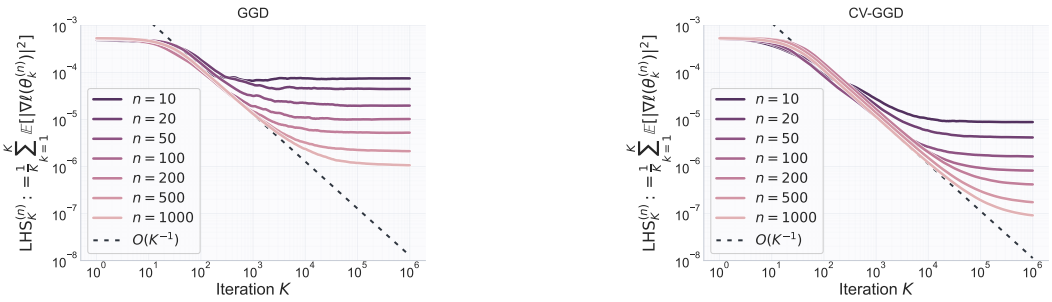


Figure 1: Averaged squared population gradient $\mathrm{LHS}_K^{(n)}$ versus the number of iterations K, for different particle counts n, in the one-dimensional example. Left: GGD. Right: CV-GGD. The dashed line indicates a K^{-1} reference rate. Both methods exhibit an initial decrease followed by a finite-particle plateau, which is substantially lower for CV-GGD.

5.2 Neural energy-based model

We next train a neural energy-based model on the UCI Digits dataset (Alpaydin and Kaynak, 1998). Unlike the one-dimensional example, the normalizing constant and model expectations under $\pi(\theta)$ are intractable in this setting. We therefore approximate the model expectation in the maximum-likelihood gradient using Langevin particles. For this experiment, we use a variant of Algorithm 1 in which the control-variate coefficient is estimated from a separate calibration particle cloud, rather than from the particles used to estimate the gradient. The derivation of the Gibbs-gradient formulation is given in Appendix B.5, and the model, calibration procedure, and hyperparameters are detailed in Appendix F.2.

We assess training performance by comparing the learned distribution with held-out digit images using the maximum mean discrepancy (MMD) (Gretton et al., 2012). Lower MMD values indicate a closer match between model samples and the validation data. Figure 2 (left) reports this metric after $K = 500$ training iterations: CV-GGD achieves a lower MMD than GGD for every tested particle count. Both methods use the same number of training iterations, although CV-GGD additionally uses the calibration particles described above.

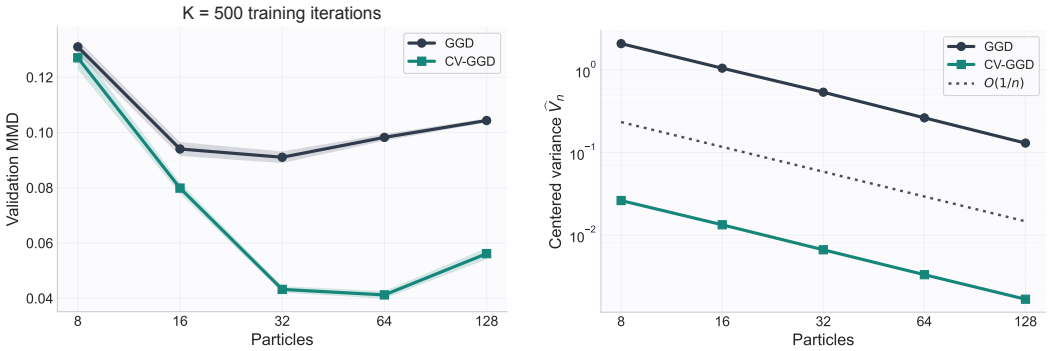


Figure 2: Neural EBM on UCI Digits. Left: validation MMD after $K = 500$ training iterations, for different particle counts n; lower values indicate a closer match to the validation data. CV-GGD uses a separate calibration particle cloud. Right: centered variance of the gradient estimator at fixed network parameters, estimated from 256 independent sampling replications. Both methods exhibit approximately $1/n$ variance scaling, with substantially lower variance for CV-GGD.

Validation MMD is non-monotonic in the particle count after $K = 500$ iterations. In longer runs with $K = 1000$, it decreases across the tested particle counts for CV-GGD, whereas GGD still exhibits an increase at the largest count. These observations concern validation performance; our theoretical bounds instead control the averaged squared population gradient.

With $K = 1000$ training iterations and $n = 128$ particles, CV-GGD also achieves lower validation MMD and Fréchet distance, as reported in Table 1. We compute the pixel-space Fréchet distance between Gaussian approximations to the generated and validation samples using their empirical means and covariances. This follows the construction of Heusel et al. (2017), applied directly to the 64-dimensional pixel vectors rather than Inception features.

Table 1: Longer-run UCI Digits EBM metrics ($K = 1000$, $n = 128$; **lower** is better). The Fréchet metric is computed in the native 64-dimensional image space.

Metric	GGD	CV-GGD
Val. MMD (10^{-2})	3.40 ± 0.20	$\mathbf{1.972} \pm 0.005$
Fréchet_{64}	8.90 ± 0.53	$\mathbf{4.88} \pm 0.14$

To examine sampling variability separately from training performance, we hold the network parameters fixed and repeat the sampling experiment independently 256 times. We estimate the centered variance of each gradient estimator from its fluctuations around the mean across replications, without requiring the population gradient. Figure 2 (right) shows that this variance decreases approximately as $1/n$ for both methods, with substantially smaller values for CV-GGD. The numerical values are reported in Table 2 of Appendix F.2. The observed $1/n$ variance scaling is consistent with the sampling fluctuations underlying the finite-particle term in our optimization bounds. However, this fixed-parameter experiment does not directly measure the optimization plateau or account for estimator bias.

These results show improved training performance and reduced estimator variance, but do not directly verify Theorem 3.3: the variance diagnostic fixes the network, and the EBM's quadratic confinement gives an unbounded drift, contrary to Assumption D.1(ii). Theorem 3.2 and Proposition 3.1 do not require Assumption D.1, but their own hypotheses would need to be checked for this model.

6 Conclusion

We analyzed Gibbs Gradient Descent (GGD), a single-loop Langevin-based method for optimization through Gibbs gradients. Our continuous-time analysis establishes bounds on the averaged squared gradient norm in both infinite- and finite-particle regimes, without assuming convexity of the outer objective, under regularity assumptions on the Gibbs gradient. We characterized finite-particle fluctuations and their contribution to the optimization bounds, and proposed a control-variate correction for the finite-observable setting. Our experiments illustrate the finite-particle plateau and show that control variates can improve optimization performance and reduce estimator variance.

Limitations. Our analysis is in continuous time. Extending the guarantees to discrete-time GGD and CV-GGD, including discretization effects and coefficients estimated from particles, remains open. Further directions include understanding the bias–variance trade-off of learned or approximate control variates and extending the framework to non-Euclidean parameter spaces.

Acknowledgments and Disclosure of Funding

A.K. thanks Google for their academic support through a gift in aid of her research. A.K. wishes to thank Ömer Deniz Akyıldız, Francesca Crucinio, Paula Cordero Encinar for discussion during the preparation of this manuscript. O.Z. gratefully acknowledges the support of the ARPE program of ENS Paris-Saclay for this work. O.Z. wishes to thank Ian Oliver McPherson for discussion about variance reduction techniques. The authors also wish to particularly thank Pierre Marion for fruitful discussions about WDRO and for suggesting the terminology *Gibbs Gradient Descent*.

References

Robert A Adams and John JF Fournier. *Sobolev spaces*. Elsevier, 2003.

Ö. Deniz Akyıldız, Michela Ottobre, and Iain Souttar. A multiscale perspective on maximum marginal likelihood estimation. *arXiv preprint arXiv:2406.04187*, 2024.

Ö. Deniz Akyıldız, Francesca Romana Crucinio, Mark Girolami, Tim Johnston, and Sotirios Sabanis. Interacting particle Langevin algorithm for maximum marginal likelihood estimation. *ESAIM Probab. Stat.*, 29:243–280, 2025.

Ethem Alpaydin and Cenk Kaynak. Optical Recognition of Handwritten Digits. UCI Machine Learning Repository, 1998. URL `https://doi.org/10.24432/C50P49`.

Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Exact generalization guarantees for (regularized) Wasserstein distributionally robust models. In *Adv. Neural Inf. Process. Syst.*, volume 36, pages 14584–14596, 2023a.

Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Regularization for Wasserstein distributionally robust optimization. *ESAIM Control. Optim. Calc. Var.*, 29:33, 2023b.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probab. XIX 1983/84: Proc.*, pages 177–206. Springer, 1985.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. Springer, 2014.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2004.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018.

A. Budhiraja, E. Kira, and Subhamay Saha. Central limit results for jump diffusions with mean field interaction and a common factor. *Stoch. Anal. Appl.*, 35(5):767–802, 2017.

R Caprio and A M Johansen. Fast convergence of the expectation-maximization algorithm under a logarithmic Sobolev inequality. *Biometrika*, 112(4):asaf061, 2025.

Rocco Caprio, Juan Kuntz, Samuel Power, and Adam M Johansen. Error bounds for particle gradient descent, and extensions of the log-Sobolev and Talagrand inequalities. *J. Mach. Learn. Res.*, 26 (103):1–38, 2025.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Math. Program.*, 184(1):71–120, 2020.

Rujian Chen. *Approximate Bayesian Modeling with Embedded Gaussian Processes*. PhD thesis, Massachusetts Institute of Technology, 2023.

Dan Crisan, Paul Dobson, and Michela Ottobre. Uniform in time estimates for the weak error of the Euler method for SDEs and a pathwise approach to derivative estimates for diffusion semigroups. *Trans. Amer. Math. Soc.*, 374(5):3289–3330, 2021.

Valentin De Bortoli, Alain Durmus, Marcelo Pereyra, and Ana F Vidal. Efficient stochastic optimisation by unadjusted Langevin Monte Carlo: Application to maximum marginal likelihood and empirical Bayesian estimation. *Stat. Comput.*, 31(3):29, 2021.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Method.*, 39(1):1–22, 1977.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Commun. Pure Appl. Math.*, 28(1):1–47, 1975.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Adv. Neural Inf. Process. Syst.*, volume 32, 2019.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *J. Mach. Learn. Res.*, 24(56):1–39, 2023.

Paula Cordero Encinar, Francesca Romana Crucinio, and Ömer Deniz Akyıldız. Proximal Interacting Particle Langevin Algorithms. In *Conf. Uncertain. Artif. Intell.*, volume 286, pages 1220–1265, 2025.

Lawrence C Evans. *Partial Differential Equations*. American Mathematical Society, 1998.

Zhou Fan, Leying Guan, Yandi Shen, and Yihong Wu. Gradient flows for empirical Bayes in high-dimensional linear models. *arXiv preprint arXiv:2312.12708*, 2023.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *J. Mach. Learn. Res.*, 13(25):723–773, 2012. URL `https://jmlr.org/papers/v13/gretton12a.html`.

Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13(11):307–361, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Adv. Neural Inf. Process. Syst.*, volume 30, 2017. URL `https://arxiv.org/abs/1706.08500`.

Richard Holley and Daniel Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *J. Stat. Phys.*, 46(5):1159–1194, 1987.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition. In *Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases*, pages 795–811. Springer, 2016.

Juan Kuntz, Jen Ning Lim, and Adam M Johansen. Particle algorithms for maximum likelihood training of latent variable models. In *Int. Conf. Artif. Intell. Stat.*, pages 5134–5180, 2023.

Jen Ning Lim, Juan Kuntz, Samuel Power, and Adam Michael Johansen. Momentum Particle Maximum Likelihood. In *Int. Conf. Mach. Learn.*, pages 29816–29871, 2024.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Adv. Neural Inf. Process. Syst.*, volume 30, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Adv. Neural Inf. Process. Syst.*, volume 29, 2016.

Pierre Marion, Anna Korba, Peter Bartlett, Mathieu Blondel, Valentin De Bortoli, Arnaud Doucet, Felipe Llinares-López, Courtney Paquette, and Quentin Berthet. Implicit Diffusion: Efficient optimization through stochastic sampling. In *Int. Conf. Artif. Intell. Stat.*, pages 1999–2007, 2025.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.*, 171(1):115–166, 2018.

Radford M Neal and Geoffrey E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learn. Graph. Model.*, pages 355–368. Springer, 1998.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400, 2000.

Art B Owen. *Monte Carlo theory, methods and examples*. Art Owen, 2013.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

Herbert E Robbins. An empirical Bayes approach to statistics. In *Break. Stat. Found. basic theory*, pages 388–394. Springer, 1992.

Louis Sharrock, Daniel Dodd, and Christopher Nemeth. Tuning-free maximum likelihood training of latent variable models via coin betting. In *Int. Conf. Artif. Intell. Stat.*, pages 1810–1818, 2024.

Tuomas Sivula. Distributed Bayesian Inference Using Expectation Propagation. Master's thesis, Aalto University, 2015.

Michel Talagrand. Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.*, 6(3):587–600, 1996.

Jakub M Tomczak. Energy-based models. In *Deep. Gener. Model.*, pages 143–158. Springer, 2022.

Paul Valsecchi Oliva and Ö. Deniz Akyıldız. Kinetic Interacting Particle Langevin Monte Carlo. *Stat. Comput.*, 36(5):221, 2026. doi: 10.1007/s11222-026-10980-z.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996. doi: 10.1007/978-1-4757-2545-2.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Adv. Neural Inf. Process. Syst.*, volume 32, 2019.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *Oper. Res.*, 2025.

Tim Y. J. Wang, Juan Kuntz, and Ömer Deniz Akyıldız. Training Latent Diffusion Models with Interacting Particle Algorithms. In *Int. Conf. Artif. Intell. Stat.*, volume 300 of *Proceedings of Machine Learning Research*, pages 2791–2799, 2026.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3):229–256, 1992.

Appendix

A Notations **14**

B Related work and applications **14**

B.1 Marginal maximum likelihood estimation 14

B.2 Reward training of Langevin processes 17

B.3 Wasserstein Distributionally Robust Optimization 18

B.4 Training of latent diffusion models 19

B.5 Training energy-based models 20

C Technical lemmas for proofs **20**

D Proofs of the GGD convergence and fluctuation results **22**

D.1 Proof of Theorem 3.1: infinite-particle case 22

D.2 Proof of Theorem 3.2: finite-particle case 24

D.3 Proof of Proposition 3.1: lower bound 28

D.4 Proofs of the asymptotic results 29

E Proof of Proposition 4.1: variance reduction for CV-GGD **40**

F Details on numerical experiments **41**

F.1 Finite-observable lower-bound witness 41

F.2 Neural energy-based model 41

A Notations

Throughout this work, $\mathcal{M}(\mathbb{R}^d)$ denotes the vector space of finite signed measures on $\mathbb{R}^d$, and $\mathcal{P}(\mathbb{R}^d) = \{\mu \in \mathcal{M}_+(\mathbb{R}^d) \mid \mu(\mathbb{R}^d) = 1\}$ is the space of probability measures over $\mathbb{R}^d$. For any $f : \mathbb{R}^d \to \mathbb{R}$, $\|f\|_\infty$ denotes the supremum norm and $\|f\|_{\mathrm{L}}$ the Lipschitz semi-norm. For any probability measure $p \in \mathcal{P}(\mathbb{R}^d)$, we define $\|f\|^2_{L^2(p)} = \int_{\mathbb{R}^d} f(x)^2 \, \mathrm{d}p(x)$, while $\|f\|_{L^2}$ denotes the same quantity for the Lebesgue measure. For any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the bounded-Lipschitz metric is defined as $\|\mu - \nu\|_{\mathrm{BL}} = \sup_{\|f\|_\infty \le 1,\, \|f\|_{\mathrm{L}} \le 1} \left| \int_{\mathbb{R}^d} f \, \mathrm{d}\mu - \int_{\mathbb{R}^d} f \, \mathrm{d}\nu \right|$, the Wassertein-1 metric as $\|\mu - \nu\|_{W_1} = \sup_{\|f\|_{\mathrm{L}} \le 1} \left| \int_{\mathbb{R}^d} f \, \mathrm{d}\mu - \int_{\mathbb{R}^d} f \, \mathrm{d}\nu \right|$, and for $s > d/2$ the negative Sobolev norm as $\|\mu - \nu\|_{H^{-s}(\mathbb{R}^d)} \coloneqq \sup_{\|\varphi\|_{H^s(\mathbb{R}^d)} \le 1} \left| \int_{\mathbb{R}^d} \varphi \, \mathrm{d}\mu - \int_{\mathbb{R}^d} \varphi \, \mathrm{d}\nu \right|$. For $p, q \in \mathcal{P}(\mathbb{R}^d)$, we denote $p \ll q$ if p is absolutely continuous relative to q, and p/q its Radon-Nikodym density. If $p \ll q$, $\mathrm{KL}(p \parallel q) = \int \log(p/q) \, \mathrm{d}p$, and $+\infty$ otherwise.

B Related work and applications

B.1 Marginal maximum likelihood estimation

In this section we discuss a line of work dedicated to algorithms, and their theoretical guarantees, closely related to Gibbs Gradient Descent in the setting of MMLE.

MMLE: alternative schemes to EM. Our work is closely related to several papers on particle-based algorithms for marginal maximum likelihood estimation (MMLE) of latent variable models. Indeed, in a statistical context, consider a probabilistic model with Lebesgue density $p_\theta(x, y)$, defined in terms of a vector of parameters, θ, to infer some unknown latent quantities, $x \in \mathcal{X}$, from some

observation y. Then, following an Empirical Bayes approach (Robbins, 1992), we may want to find a θ^* maximizing the marginal likelihood, $p_\theta[y] := \int p_\theta[x, y]dx$; and obtain the corresponding posterior distribution, $p_{\theta^*}[x|y] := p_{\theta^*}[x, y]/p_{\theta^*}[y]$. A standard approach to tackle this problem is the one of Expectation-Maximization (EM) (Dempster et al., 1977), which aims at minimizing the functional

$$L(p, \theta) = \mathrm{KL}\left(p||p_\theta(\cdot, y)\right) = \int \log\Big(\frac{p[x]}{p_\theta[x, y]}\Big)p[x]\,\mathrm{d}x \tag{5}$$

for $p, \theta \in \mathcal{P}(\mathcal{X}) \times \Theta$. In the above expression, $\bar{\ell}(\theta, x) := \log(p_\theta[x, y])$ denotes the log-likelihood. Notice that (p^*, θ^*) minimize L if and only if θ^* maximizes the marginal likelihood and $p^* = p_{\theta^*}[\cdot|y]$ ($\propto p_{\theta^*}[\cdot, y]$). In the notation of this paper, this corresponds to the Gibbs law $\pi(\theta)(x) \propto p_\theta(x, y)$, or equivalently to the potential $V(x, \theta) = -\log p_\theta(x, y)$ up to additive constants independent of x. The associated Gibbs gradient is

$$\Gamma(p, \theta) = -\int \nabla_\theta \log p_\theta(x, y)\, p(\mathrm{d}x).$$

In that setting, EM is known to compute a coordinate descent on (p, θ), i.e. the expectation step (E) solves for $p_k := \mathrm{argmin}_{p\in\mathcal{P}(\mathcal{X})} L(p, \theta_k)$, and the maximization step (M) solves for $\theta_{k+1} := \mathrm{argmin}_{\theta\in\Theta} L(p_k, \theta_k)$ (Neal and Hinton, 1998). In recent years, a number of methods have been proposed for MMLE problem that are approximating the E-step of the EM with an unadjusted Langevin chain. In this context, (De Bortoli et al., 2021) studied an algorithm, abbreviated as SOUL, which consists of an E-step based on an approximation provided by a ULA chain and an M-step based on gradient descent. They showed the convergence of the algorithm (in discrete-time) under some strict conditions, and furthermore pro- vided empirical evidence of the performance of this method. Closer to our setting, there have also been algorithms developed based on *interacting particle systems* - which instead of a single ULA chain for the E-step, use a system of interacting n particles run in parallel with parameter updates. These methods are termed particle gradient descent (PGD) (Kuntz et al., 2023) and interacting particle Langevin algorithm (IPLA) (Akyıldız et al., 2025). In spirit, compared to EM, all these algorithms rely on alternative optimization minimizing L, analog to gradient descent (instead of coordinate descent), and implementable when Θ and $\mathcal{X}$ are Euclidean. Below, we detail state-of-the art theoretical guarantees and recent literature related to these approaches.

PGD and its theoretical guarantees. (Kuntz et al., 2023) proposed a scheme called *Particle gradient Descent* (PGD), whose discrete-time updates match the Euler discretization of Eq. (3) with $\epsilon_k = 1$, but without deriving any theoretical guarantees. (Caprio et al., 2025) provide several theoretical results validating the use of PGD. They first provide the following result *in continuous time, for an infinite number of particles*, i.e. the system (2a)-(2b). They show that the functional $(L(p_t, \theta_t))_{t\geq 0}$ evaluated along the flow converges at an exponential rate to its infimum, provided that the model satisfies the following condition (referred to as xLSI): there exists a constant $\lambda > 0$ such that

$$2\lambda[L(p, \theta) - L_*] \leq I(p, \theta), \quad L_* := \inf_{(p,\theta)\in\mathcal{P}(\mathcal{X})\times\Theta} L(q, \theta),$$

for all parameters θ and distributions p, where the functional I is defined by

$$I(p, \theta) := \left\|\int \nabla_2\bar{\ell}(\theta, x)p(dx)\right\|^2 + \int \left\|\nabla_x \log\left(\frac{p(x)}{p_\theta(x \mid y)}\right)\right\|^2 p(\mathrm{d}x). \tag{6}$$

Note that this condition generalizes the log-Sobolev inequality (LSI) and includes a PL-type structure in the parameter direction. This is stronger than the assumptions used here in the outer variable: it yields convergence of the objective value, whereas our non-convex result controls the averaged squared gradient norm. The condition is satisfied for models with strongly concave likelihoods $\bar{\ell}$ as shown in (Caprio et al., 2025) by generalizing a Bakry-Emery criterion (Bakry and Émery, 1985) (see also (Caprio and Johansen, 2025, Section 2.3) for more examples of models). Then, they extend a result of (Otto and Villani, 2000) and show that models which satisfy (6) also satisfy a generalization of both the Talagrand inequality (Talagrand, 1996) and the *quadratic growth* condition used in optimization (Karimi et al., 2016): for all parameters θ and distributions p, it holds that

$$2[L(p, \theta) - L_*] \geq \lambda d((p, \theta), \mathcal{M}_*)^2, \tag{7}$$

where $\mathcal{M}_* := \arg\min_{(p,\theta)\in\mathcal{P}(\mathcal{X})\times\Theta} L(p,\theta)$ and $d((\theta,q),\mathcal{M}_*) := \inf_{(\theta_*,q_*)\in\mathcal{M}_*} d((\theta,q),(\theta_*,q_*))$, where d the following metric acting on $\Theta\times\mathcal{P}(\mathcal{X})$-valued random variables:

$$d((\theta,p),(\theta',p')) := \sqrt{\mathbb{E}\left[||\theta-\theta'||^2\right] + \mathbb{E}\left[W_2(p,p')^2\right]}, \tag{8}$$

and W_2 denotes the Wasserstein-2 metric. Using (7), they then show that the flow $(\theta_t,q_t)_{t\geq 0}$ also converges to $\mathcal{M}_*$ exponentially fast in d. Under the further assumption that $\bar{\ell}$'s gradients are Lipschitz, they provide convergence rates for the *discrete time, finite particle system*: the algorithm's output, $(\theta_K^{n,h}, p_K^{n,h})$, satisfies

$$d((\theta_K^{n,h},p_K^{n,h}),(\theta_*,p_*^n)) = \mathcal{O}\left(h^{\frac{1}{2}} + n^{-\frac{1}{2}} + e^{-h\lambda K}\right), \tag{9}$$

assuming that the algorithm's step size h is no greater than $1/(\lambda+L)$, for n particles and K (discrete) time-steps. In the above, p_*^n denotes the empirical measure of n i.i.d. particles drawn from $\pi(\theta^*) = p_{\theta_*}(\cdot \mid y)$. These smoothness assumptions are different from ours: our convergence proof mainly requires spatial regularity of $x\mapsto\nabla_2 V(x,\theta)$ and Lipschitz control of the implicit Gibbs gradient Γ with respect to its measure argument.

(Akyıldız et al., 2025) consider an alternative approximation to the flow of L than PGD, by including an extra additive Gaussian noise term $(2h/n)\eta_k$ in the θ_k parameter update of that same discrete-time GGD scheme, resulting in an algorithm called *Interacting Particle Langevin Algorithm* (IPLA). This differs from our finite-particle GGD dynamics: IPLA injects an additional Brownian perturbation in θ, whereas in our analysis the stochasticity in the parameter evolution comes only through the empirical particle system. The authors provide a bound on the L^2 error of the parameter estimates produced by IPLA of the same order as (9) for models satisfying Lipschitz gradient and strong concavity assumptions. This theoretical analysis came out previously to (Caprio et al., 2025) but bounds are of the same order.

Algorithmic variants of PGD and IPLA. There was generally a great interest recently in proposing different variations of the PGD and IPLA scheme. First note that the original PGD paper (Kuntz et al., 2023) came out with two variants for the PGD scheme: the first one being *Particle Quasi-Newton* (PQN) that replaces the standard gradient descent update on θ with a Newton's update; the second one being *Particle Marginal Gradient Descent* (PMGD) that replaces the update on θ with its optimal value at step k (with respect to the current particle distribution) when the (M) step is tractable in closed-form, which is true for a large class of models. (Lim et al., 2024; Chen, 2023; Fan et al., 2023) proposed related momentum or gradient-flow-inspired variants. (Sharrock et al., 2024) introduced two algorithms, one of which, named SVGD-EM is a discrete time version of a gradient flow of L in a Stein geometry (Liu, 2017; Duncan et al., 2023), in analogy with the famous Stein Variational Gradient Descent (SVGD) algorithm (Liu and Wang, 2016). (Encinar et al., 2025) proposed a variant of IPLA, named Proximal IPLA, adapted to non-differentiable statistical models p_θ. Finally, (Valsecchi Oliva and Akyıldız, 2026) propose a discretized underdamped Langevin diffusion regarding the particles update in the inner loop, in contrast with the original overdamped version in IPLA or PGD.

Two time-scale schemes. Similarly to Marion et al. 2025, Akyıldız et al. 2024 develop a two-time-scale analysis for a related coupled Langevin system. Their dynamics include a parameter controlling the relative speed of the particle and parameter equations, as well as an additional diffusion term in the parameter dynamics controlled by a temperature parameter. This two-scale system covers the most practical implementations of these joint inference methods, that use multiscale dynamics ideas in their implementation, while PGD was assuming that the processes $(\theta_t)_{t\geq 0}$ and $(X_t)_{t\geq 0}$ were controlled by the same step-size. Akyıldız et al. 2024 show that the averaged parameter dynamics obtained in the limit of scale separation can be used to estimate the optimal parameter, within the strongly convex setting. Their assumptions are then stronger than the ones in (Marion et al., 2025) who work under log-Sobolev (which is more general than strong convexity). The reason why they are working exclusively in the strongly convex setting, is because they use recent uniform-in-time non-asymptotic averaging bounds that rely on the semigroup (associated to the SDE) (space-)derivative estimates (Crisan et al., 2021); and it is not known that these hold without strong- convexity-type assumptions. Yet, they have more explicit assumptions since they directly focus on the MMLE problem and make regularity assumptions on the joint log-likelihood. For example, (Marion et al., 2025, Assump. 4.3) assumes regularity of $\Gamma(p,\theta)$, which in the MMLE setting would be making assumptions on the marginal likelihood (which is unavailable and implicit)

- rather than the joint likelihood (available and explicit). A second important difference is that the continuous-time rate given in (Marion et al., 2025, Thm. 4.5) necessitates a decaying time-scale parameter $\varepsilon_t = \min(1, 1/t^\delta)$ with $1/2 \leq \delta < 1$ ($\delta = 1/2$ in (Marion et al., 2025)); similarly the discrete-time rate given in (Marion et al., 2025, Thm. 4.7) requires $\epsilon_k = 1/\sqrt{k}$ and step size $\gamma = c_1/\sqrt{k}$. Our continuous-time bounds allow both decaying schedules and the constant-step-size case $\delta = 0$, under the smallness condition on ε stated in the theorems. Several related strong-error or global optimization guarantees rely on convexity, strong-convexity, or PL/xLSI-type structure in the parameter objective, or in the joint latent-variable/parameter objective. Our result instead gives non-convex stationarity guarantees for the outer objective ℓ through averaged squared-gradient bounds. Finally, compared to (Caprio et al., 2025; Akyıldız et al., 2025), (Marion et al., 2025; Akyıldız et al., 2024) obtain weak-error estimates, i.e., convergence of expectations of bounded continuous test functions of θ_t rather than L^2-type strong error bounds.

B.2 Reward training of Langevin processes

We include the Langevin reward-training example presented in Marion et al. (2025) to explain how it fits the notation of Eqs. (1) and (2) and the finite-observable form in Eq. (4). Let $\pi(\theta)(\mathrm{d}x) = \mathrm{e}^{-V(x,\theta)} Z_\theta^{-1}\,\mathrm{d}x$ be the stationary law of the Langevin diffusion:

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t.$$

Using the identity $\nabla_\theta \log \pi(\theta)(x) = \mathbb{E}_{\pi(\theta)}[\nabla_2 V(\cdot, \theta)] - \nabla_2 V(x, \theta)$, and introducing the pure reward objective $\ell_R(\theta) = -\mathbb{E}_{\pi(\theta)}[R]$, we have

$$\begin{aligned}\nabla \ell_R(\theta) &= -\mathbb{E}_{\pi(\theta)}[R \nabla_\theta \log \pi(\theta)] \\ &= \mathbb{E}_{\pi(\theta)}[R \nabla_2 V(\cdot, \theta)] - \mathbb{E}_{\pi(\theta)}[R]\,\mathbb{E}_{\pi(\theta)}[\nabla_2 V(\cdot, \theta)] \\ &= \operatorname*{Cov}_{\pi(\theta)}(R(X), \nabla_2 V(X, \theta)).\end{aligned}$$

This is the REINFORCE / score-function computation (Williams, 1992); it does not require differentiability of R. Similarly, for the reference-matching objective $\ell_{\text{ref}}(\theta) = \mathrm{KL}(p_{\text{ref}} \parallel \pi(\theta))$,

$$\nabla \ell_{\text{ref}}(\theta) = -\mathbb{E}_{p_{\text{ref}}}[\nabla_\theta \log \pi(\theta)] = \mathbb{E}_{p_{\text{ref}}}[\nabla_2 V(\cdot, \theta)] - \mathbb{E}_{\pi(\theta)}[\nabla_2 V(\cdot, \theta)].$$

Combining these two terms, we consider the minimization problem

$$\ell_{\text{rew}}(\theta) = -\lambda\,\mathbb{E}_{\pi(\theta)}[R] + \beta\ \mathrm{KL}(p_{\text{ref}} \parallel \pi(\theta)),$$

where $\lambda, \beta \geq 0$ and the reference distribution p_{ref} is either known or available through samples. The first term maximizes the reward inside the Gibbs family $\{\pi(\theta)\}$; the second term is the sample-access reference-matching term used in contrastive EBM training (Gutmann and Hyvärinen, 2012). Differentiating the Gibbs density gives

$$\nabla \ell_{\text{rew}}(\theta) = \lambda \operatorname*{Cov}_{\pi(\theta)}(R(X), \nabla_2 V(X, \theta)) + \beta\Big(\mathbb{E}_{p_{\text{ref}}}[\nabla_2 V(\cdot, \theta)] - \mathbb{E}_{\pi(\theta)}[\nabla_2 V(\cdot, \theta)]\Big).$$

Thus, this is a Gibbs-gradient problem with

$$\Gamma_{\text{rew}}(p, \theta) = \lambda\Big(\mathbb{E}_p[R \nabla_2 V(\cdot, \theta)] - \mathbb{E}_p[R]\,\mathbb{E}_p[\nabla_2 V(\cdot, \theta)]\Big) + \beta\Big(\mathbb{E}_{p_{\text{ref}}}[\nabla_2 V(\cdot, \theta)] - \mathbb{E}_p[\nabla_2 V(\cdot, \theta)]\Big).$$

In particular, taking

$$\Phi^\theta(x) = \big(R(x), \nabla_2 V(x, \theta), R(x) \nabla_2 V(x, \theta)\big),$$

and absorbing the reference expectation $\mathbb{E}_{p_{\text{ref}}}[\nabla_2 V(\cdot, \theta)]$ into H_θ places the example in Eq. (4). The GGD dynamics are then simply the Langevin sampler above, run with the current parameter θ_t, coupled with

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma_{\text{rew}}(p_t, \theta_t)\,\mathrm{d}t.$$

With n particles, p_t is replaced by $p_t^{(n)} = n^{-1} \sum_{i=1}^n \delta_{X_t^i}$, which yields the empirical covariance estimator for the reward term and the empirical model expectation for the reference-matching term.

In a discrete Euler implementation, with reference samples $\tilde{X}_k^{(1)}, \dots, \tilde{X}_k^{(m)} \sim p_{\text{ref}}$, this gives

$$\begin{aligned} X_{k+1}^{(i)} &= X_k^{(i)} - \gamma_X \nabla_1 V(X_k^{(i)}, \theta_k) + \sqrt{2\gamma_X}\Delta B_k^{(i)}, \\ \theta_{k+1} &= \theta_k - \gamma_\theta \Bigg[\lambda \, \hat{\mathrm{Cov}}_n\big(R(X_k), \nabla_2 V(X_k, \theta_k)\big) \\ &\qquad + \beta \left(\frac{1}{m}\sum_{j=1}^{m} \nabla_2 V(\tilde{X}_k^{(j)}, \theta_k) - \frac{1}{n}\sum_{i=1}^{n} \nabla_2 V(X_k^{(i)}, \theta_k) \right) \Bigg]. \end{aligned}$$

Guidance methods that instead modify the sampling potential to $V - \lambda R$ should be distinguished from this setup: they change the target sampler directly, whereas GGD updates θ so that the final distribution remains in the original Gibbs family.

B.3 Wasserstein Distributionally Robust Optimization

This application is the new example in this appendix, so we spell out the computation in the notation of the main text. Let $\{\xi_1, \dots, \xi_M\} \subset \mathbb{R}^{d_\xi}$ be the training data and $P_M = M^{-1}\sum_{m=1}^{M} \delta_{\xi_m}$. Wasserstein distributionally robust optimization (Mohajerin Esfahani and Kuhn, 2018) considers

$$\min_{\eta} \sup_{W_c(P_M, Q) \le \rho} \mathbb{E}_{\zeta \sim Q}[f_\eta(\zeta)].$$

The usual dual form is

$$\min_{\eta, \lambda \ge 0} \lambda\rho + \frac{1}{M}\sum_{m=1}^{M} \sup_{\zeta \in \mathbb{R}^{d_\xi}} \{f_\eta(\zeta) - \lambda c(\xi_m, \zeta)\}.$$

To obtain a smooth sampling-based objective, we replace the supremum by a log-sum-exp under the Gaussian reference $\mathcal{N}(\xi_m, \sigma^2 I)$ with temperature $\tau > 0$. With $\theta = (\eta, \lambda)$,

$$g(\theta) = \lambda\rho, \quad h_\theta(\xi, \zeta) = f_\eta(\zeta) - \lambda c(\xi, \zeta),$$

define

$$\ell_{\text{WDRO}}(\theta) = g(\theta) + \frac{\tau}{M}\sum_{m=1}^{M} \log \int \exp\left(\frac{h_\theta(\xi_m, \zeta)}{\tau}\right) \varphi_\sigma(\zeta - \xi_m)\, \mathrm{d}\zeta, \tag{10}$$

where φ_σ is the density of $\mathcal{N}(0, \sigma^2 I)$. This is the entropy-regularized WDRO dual (Wang et al., 2025; Azizian et al., 2023b), for which exact generalization guarantees are available (Azizian et al., 2023a). For each datum ξ_m, introduce the conditional Gibbs distribution

$$\pi_m(\theta)(\mathrm{d}\zeta) = \frac{\mathrm{e}^{-V_m(\zeta,\theta)}}{Z_m(\theta)}\,\mathrm{d}\zeta, \quad V_m(\zeta, \theta) = \frac{\|\zeta - \xi_m\|^2}{2\sigma^2} - \frac{h_\theta(\xi_m, \zeta)}{\tau}.$$

Constants independent of ζ are omitted from V_m because they are absorbed by $Z_m(\theta)$. Equivalently, define an augmented state $X = (m, \zeta)$ with base measure $M^{-1}\sum_{m=1}^{M} \delta_m \otimes \mathrm{d}\zeta$ and Gibbs potential $V((m, \zeta), \theta) = V_m(\zeta, \theta)$. The corresponding augmented Gibbs law is

$$\pi(\theta)(\mathrm{d}m, \mathrm{d}\zeta) = \frac{1}{M}\delta_m(\mathrm{d}m)\pi_m(\theta)(\mathrm{d}\zeta).$$

If one wants to stay literally in the Euclidean state-space notation of the main text, this is the same as running M conditional Langevin systems in parallel, one for each fixed datum.

Differentiating Eq. (10) gives the Gibbs-gradient identity

$$\nabla \ell_{\text{WDRO}}(\theta) = \nabla g(\theta) + \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\zeta \sim \pi_m(\theta)}[\nabla_\theta h_\theta(\xi_m, \zeta)].$$

Therefore, for any augmented probability measure $p(\mathrm{d}m, \mathrm{d}\zeta) = M^{-1}\sum_{m=1}^{M} \delta_m(\mathrm{d}m)p_m(\mathrm{d}\zeta)$, the Gibbs gradient is

$$\Gamma_{\text{WDRO}}(p, \theta) = \nabla g(\theta) + \int \nabla_\theta h_\theta(\xi_m, \zeta)\, p(\mathrm{d}m, \mathrm{d}\zeta). \tag{11}$$

In components, if $\theta = (\eta, \lambda)$ then $\nabla_\eta h_\theta(\xi, \zeta) = \nabla_\eta f_\eta(\zeta)$ and $\nabla_\lambda h_\theta(\xi, \zeta) = -c(\xi, \zeta)$, so the λ part in Eq. (11) is the familiar constraint residual $\rho - \int c(\xi_m, \zeta) p(\mathrm{d}m, \mathrm{d}\zeta)$.

The finite-observable structure is immediate:

$$\phi^\theta(m, \zeta) = \nabla_\theta h_\theta(\xi_m, \zeta), \quad H_\theta(u) = \nabla g(\theta) + u, \quad \Gamma_{\text{WDRO}}(p, \theta) = H_\theta(\mathbb{E}_p[\phi^\theta]).$$

Thus WDRO enters Eq. (4) with $r = 1$ and H_θ affine. The associated ideal GGD system is

$$\mathrm{d}Z_t^m = -\nabla_\zeta V_m(Z_t^m, \theta_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^m, \quad \mathrm{d}\theta_t = -\varepsilon_t \Gamma_{\text{WDRO}}(p_t, \theta_t)\,\mathrm{d}t,$$

where $p_t = M^{-1}\sum_m \delta_m \otimes \mathrm{Law}(Z_t^m)$. The implementable finite-particle version uses particles $Z_t^{m,i}$ and replaces p_t by $p_t^{(n)} = (Mn)^{-1}\sum_{m=1}^M \sum_{i=1}^n \delta_{(m, Z_t^{m,i})}$. If h_θ is bounded, or more generally if $V_m(\cdot, \theta)$ is a bounded perturbation of the Gaussian quadratic potential, the conditional targets inherit a log-Sobolev inequality by the Holley–Stroock perturbation principle (Holley and Stroock, 1987; Bakry et al., 2014).

Finally, the Gibbs law also appears directly in the objective, not only in the gradient. By the Donsker–Varadhan variational formula (Donsker and Varadhan, 1975), for each m,

$$\tau \log \mathbb{E}_{\zeta \sim \mathcal{N}(\xi_m, \sigma^2 I)}\left[\mathrm{e}^{h_\theta(\xi_m, \zeta)/\tau}\right] = \sup_{q \in \mathcal{P}(\mathbb{R}^{d_\xi})} \left\{\mathbb{E}_q[h_\theta(\xi_m, \cdot)] - \tau\,\mathrm{KL}(q \parallel \mathcal{N}(\xi_m, \sigma^2 I))\right\},$$

and the optimizer is exactly $q = \pi_m(\theta)$. Hence

$$\ell_{\text{WDRO}}(\theta) = g(\theta) + \frac{1}{M}\sum_{m=1}^M \left\{\mathbb{E}_{\pi_m(\theta)}[h_\theta(\xi_m, \cdot)] - \tau\,\mathrm{KL}(\pi_m(\theta) \parallel \mathcal{N}(\xi_m, \sigma^2 I))\right\}.$$

B.4 Training of latent diffusion models

The particle formulation and the analysis of this example is due to Wang et al. (2026). We include it because it makes clear that their interacting-particle latent diffusion dynamics have exactly the same Gibbs-gradient shape as the main text, up to carrying a data index as a fixed coordinate. Write the optimization parameter of the latent diffusion model as $\vartheta = (\theta, \phi)$, where θ parameterizes the latent diffusion prior and ϕ parameterizes the decoder. For a data point x^m, define the diffusion loss

$$\mathcal{L}_D(\theta, z) = \mathbb{E}_{q(z_{1:K}|z)}\left[\log \frac{q(z_{1:K} \mid z)}{p_\theta(z_{0:K})}\right],$$

and the tilted log-density

$$\mathcal{L}^m(\vartheta, z) = \log p_\phi(x^m \mid z) - \mathcal{L}_D(\theta, z).$$

The negative tilted log-likelihood can be written

$$\ell_{\text{LDM}}(\vartheta) = -\frac{1}{M}\sum_{m=1}^M \log \int \mathrm{e}^{\mathcal{L}^m(\vartheta, z)}\,\mathrm{d}z.$$

Therefore each datum defines a Gibbs posterior on the latent variable

$$\pi_m(\vartheta)(\mathrm{d}z) = \frac{\mathrm{e}^{-V_m(z,\vartheta)}}{Z_m(\vartheta)}\,\mathrm{d}z, \quad V_m(z, \vartheta) = \mathcal{L}_D(\theta, z) - \log p_\phi(x^m \mid z).$$

Differentiating the log-partition gives

$$\nabla_\theta \ell_{\text{LDM}}(\vartheta) = \frac{1}{M}\sum_{m=1}^M \mathbb{E}_{\pi_m(\vartheta)}[\nabla_\theta \mathcal{L}_D(\theta, Z)],$$

$$\nabla_\phi \ell_{\text{LDM}}(\vartheta) = -\frac{1}{M}\sum_{m=1}^M \mathbb{E}_{\pi_m(\vartheta)}[\nabla_\phi \log p_\phi(x^m \mid Z)].$$

Equivalently, with augmented state $X = (m, z)$ and $p = M^{-1}\sum_m \delta_m \otimes p_m$, define the observable

$$\Phi^\vartheta(m, z) = \begin{pmatrix} \nabla_\theta \mathcal{L}_D(\theta, z) \\ -\nabla_\phi \log p_\phi(x^m \mid z) \end{pmatrix}.$$

Then

$$\Gamma_{\mathrm{LDM}}(p, \vartheta) = \mathbb{E}_p[\Phi^\vartheta],$$

so the example lies in the finite-observable class with $r = 1$ and $H_\vartheta = \mathrm{Id}$. The ideal Gibbs-gradient flow is

$$\begin{aligned} \mathrm{d}Z_t^m &= -\nabla_z V_m(Z_t^m, \vartheta_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^m \\ &= \nabla_z \mathcal{L}^m(\vartheta_t, Z_t^m)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^m, \\ \mathrm{d}\vartheta_t &= -\varepsilon_t \Gamma_{\mathrm{LDM}}(p_t, \vartheta_t)\,\mathrm{d}t. \end{aligned}$$

The practical algorithm of Wang et al. (2026) replaces each conditional law by N latent particles, uses mini-batches of data indices, and adds standard deep-learning stabilizers such as separate step sizes, preconditioning, momentum, and annealing. In the notation of this paper, those are implementation choices around the finite-particle approximation of the Gibbs-gradient system above.

B.5 Training energy-based models

The neural EBM experiment in Section F.2 is another direct instance of the framework, and is included here to fix the notation used by the numerical section. This is the standard maximum-likelihood training objective for energy-based models (Tomczak, 2022; Du and Mordatch, 2019). Let ν denote the data distribution, or the empirical training measure in practice, and let the model be the Gibbs family

$$\pi(\theta)(\mathrm{d}x) = \frac{\mathrm{e}^{-V(x,\theta)}}{Z_\theta}\,\mathrm{d}x, \qquad Z_\theta = \int \mathrm{e}^{-V(x,\theta)}\,\mathrm{d}x.$$

Maximum likelihood is equivalent, up to the entropy of ν, to minimizing the forward KL divergence

$$\ell_{\mathrm{EBM}}(\theta) = \mathrm{KL}(\nu \parallel \pi(\theta)) = \mathbb{E}_\nu[V(\cdot, \theta)] + \log Z_\theta + \text{const.}$$

Differentiating the log-partition function gives $\nabla_\theta \log Z_\theta = -\mathbb{E}_{\pi(\theta)}[\nabla_2 V(\cdot, \theta)]$, hence

$$\nabla \ell_{\mathrm{EBM}}(\theta) = \mathbb{E}_\nu[\nabla_2 V(\cdot, \theta)] - \mathbb{E}_{\pi(\theta)}[\nabla_2 V(\cdot, \theta)].$$

Thus the Gibbs gradient is

$$\Gamma_{\mathrm{EBM}}(p, \theta) = \mathbb{E}_\nu[\nabla_2 V(\cdot, \theta)] - \mathbb{E}_p[\nabla_2 V(\cdot, \theta)].$$

This has the finite-observable form of Eq. (4) with

$$\phi^\theta(x) = \nabla_2 V(x, \theta), \quad H_\theta(u) = \mathbb{E}_\nu[\nabla_2 V(\cdot, \theta)] - u.$$

The ideal GGD dynamics are therefore

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad \mathrm{d}\theta_t = -\varepsilon_t \Gamma_{\mathrm{EBM}}(p_t, \theta_t)\,\mathrm{d}t.$$

With particles $X_k^1, \ldots, X_k^n$ and training samples $Y_1, \ldots, Y_M \sim \nu$, the discrete-time estimator used in the experiment is the contrastive maximum-likelihood gradient

$$\widehat{\Gamma}_{\mathrm{EBM}}(\theta_k) = \frac{1}{M}\sum_{j=1}^{M} \nabla_2 V(Y_j, \theta_k) - \frac{1}{n}\sum_{i=1}^{n} \nabla_2 V(X_k^i, \theta_k),$$

combined with persistent Langevin particles. In the neural experiment, $V(x, \theta) = \|x\|^2/(2\sigma_0^2) + f_\theta(x)$, so $\nabla_2 V(x, \theta) = \nabla_\theta f_\theta(x)$ and the formula above is exactly the update reported in Section F.2. The CV-GGD variant used there does not change the Gibbs gradient; it only replaces the empirical model expectation by a lower-variance estimator based on a zero-mean score control variate.

C Technical lemmas for proofs

We first state some preliminary lemmas necessary for the proof of the main text results in Section D.

Lemma C.1 (Grönwall's inequality; Evans 1998). *Let $t \mapsto u(t)$ be a nonnegative absolutely continuous function on $[0, T]$ satisfying the differential inequality*

$$\frac{du(t)}{dt} \leq a(t)u(t) + b(t),$$

where a and b are nonnegative, summable functions on $[0, T]$. Then,

$$u(t) \leq \mathrm{e}^{\int_0^t a(s)\,\mathrm{d}s} \left[u(0) + \int_0^t b(s)\,\mathrm{d}s\right], \quad 0 \leq t \leq T.$$

Lemma C.2 (Log-Sobolev implies Talagrand; Otto and Villani 2000). *Assume that $q \in \mathcal{P}(\mathbb{R}^d)$ satisfies the Log-Sobolev inequality with constant $\mu > 0$. Then, q satisfies a Talagrand inequality: for every $p \in \mathcal{P}(\mathbb{R}^d)$ such that $p \ll q$, $\|p - q\|_{W_2} \leq \sqrt{\frac{2}{\mu}\,\mathrm{KL}(p||q)}$.*

Lemma C.3 (Bound on the derivative of the log Gibbs density). *Let $\theta \in \mathbb{R}^p$ and consider the Gibbs distribution $\pi(\theta) \in \mathcal{P}(\mathbb{R}^d)$ associated to a potential $V(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R}$, with normalization constant Z_θ. Let $\theta \in \mathbb{R}^p$ and assume that the function $x \mapsto \nabla_2 V(x, \theta)$ is L_X-Lipschitz. Then, the function $x \mapsto \nabla_\theta \log \pi(\theta)(x)$ is L_X-Lipschitz.*

Proof. Recall that for $x \in \mathbb{R}^d$, $\log(\pi(\theta))(x) = -V(x, \theta) - \log(Z_\theta)$, where $Z_\theta = \int_{\mathbb{R}^d} \mathrm{e}^{-V(y,\theta)}\,\mathrm{d}y$. Applying the chain rule yields

$$\nabla_\theta \log \pi(\theta)(x) = -\nabla_2 V(x, \theta) - \frac{\nabla_\theta Z_\theta}{Z_\theta}, \quad x \in \mathbb{R}^d.$$

We focus on the second term, which satisfies

$$\nabla_\theta Z_\theta = -\int_{\mathbb{R}^d} \mathrm{e}^{-V(y,\theta)} \nabla_2 V(y, \theta)\,\mathrm{d}y.$$

Dividing by $Z_\theta = \mathrm{e}^{-V(\cdot,\theta)}/\pi(\theta)$ gives us $\frac{1}{Z_\theta}\nabla_\theta Z_\theta = -\mathbb{E}_{y\sim\pi(\theta)}\left[\nabla_2 V(y, \theta)\right]$, i.e.,

$$\nabla_\theta \log \pi(\theta)(x) = \mathbb{E}_{y\sim\pi(\theta)}\left[\nabla_2 V(y, \theta)\right] - \nabla_2 V(x, \theta), \quad x \in \mathbb{R}^d.$$

This implies that $x \mapsto \nabla_\theta \log \pi(\theta)(x)$ is L_X-Lipschitz since $x \mapsto \nabla_2 V(x, \theta)$ is L_X-Lipschitz. □

Lemma C.4. *Let $p, q \in \mathcal{P}(\mathbb{R}^d)$, then*

1) If $s > d/2$, then there exists a constant $C_{s,d} > 0$ such that $\|p - q\|_{\mathrm{H}^{-s}} \leq C_{s,d}\,\|p - q\|_{\mathrm{TV}}$.

2) $\|p - q\|_{\mathrm{BL}} \leq \|p - q\|_{\mathrm{TV}} \leq \sqrt{2\,\mathrm{KL}(p||q)}$.

Proof. We begin with the proof of 1). By the Sobolev embedding of $H^s(\mathbb{R}^d)$ in $L^\infty(\mathbb{R}^d)$ (Adams and Fournier, 2003, Thm. 4.12), there exists a constant $C_{s,d} > 0$ such that $\|\varphi\|_\infty \leq C_{s,d}\|\varphi\|_{H^s(\mathbb{R}^d)}$ for all $\varphi \in H^s(\mathbb{R}^d)$. Hence, $\|\varphi\|_\infty \leq C_{s,d}$ for every φ such that $\|\varphi\|_{H^s(\mathbb{R}^d)} \leq 1$. Therefore,

$$\left|\int_{\mathbb{R}^d} \varphi(x)\,\mathrm{d}(p - q)(x)\right| = C_{s,d}\left|\int_{\mathbb{R}^d} \frac{\varphi(x)}{C_{s,d}}\,\mathrm{d}(p - q)(x)\right| \leq C_{s,d}\,\|p - q\|_{\mathrm{TV}},$$

since $\|\varphi/C_{s,d}\|_\infty \leq 1$. Taking the supremum over $\|\varphi\|_{H^s(\mathbb{R}^d)} \leq 1$ yields

$$\|p - q\|_{\mathrm{H}^{-s}} \leq C_{s,d}\,\|p - q\|_{\mathrm{TV}}.$$

Comcerning 2), the left inequality comes from the fact that for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,

$$\begin{aligned}\|p-q\|_{\mathrm{BL}} &= \sup_{\|f\|_\infty \le 1, \|f\|_{\mathrm{L}} \le 1} \left| \int_{\mathbb{R}^d} f \,\mathrm{d}p(x) - \int_{\mathbb{R}^d} f \,\mathrm{d}q(x) \right| \\ &\le \sup_{\|f\|_\infty \le 1} \left| \int_{\mathbb{R}^d} f \,\mathrm{d}p(x) - \int_{\mathbb{R}^d} f \,\mathrm{d}q(x) \right| = \|p-q\|_{\mathrm{TV}}\,,\end{aligned}$$

while the right inequality follows from Pinsker's inequality. □

Note that if $s > d/2 + 1$, one also has $\|p-q\|_{\mathrm{H}^{-s}} \le C_{s,d} \|p-q\|_{\mathrm{BL}}$ for some constant $C_{s,d} > 0$ by the same argument as 1).

Remark C.1 (Regularity of finite-observable Gibbs gradients)**.** *Assume that Γ has the finite-observable form*

$$\Gamma(p,\theta) = H_\theta(m_\theta(p)), \qquad m_\theta(p) = \big(\mathbb{E}_p[\phi_1^\theta], \dots, \mathbb{E}_p[\phi_r^\theta]\big),$$

and that H_θ is Lipschitz on the range of m_θ, uniformly in θ. Then regularity of the observables transfers directly to Γ. Indeed,

$$\|\Gamma(p,\theta) - \Gamma(q,\theta)\|_{\mathbb{R}^p} \le L_H \|m_\theta(p) - m_\theta(q)\|.$$

Each coordinate of $m_\theta(p) - m_\theta(q)$ is of the form $\int \phi_j^\theta \,\mathrm{d}(p-q)$. Hence bounded observables give $\sqrt{\mathrm{KL}}$-Lipschitz continuity by total variation and Pinsker's inequality, bounded-Lipschitz observables give BL-Lipschitz continuity, and observables in $H^s(\mathbb{R}^d)$ give H^{-s}-Lipschitz continuity by duality:

$$\left| \int \phi_j^\theta \,\mathrm{d}(p-q) \right| \le \|\phi_j^\theta\|_{H^s} \|p-q\|_{H^{-s}}.$$

D Proofs of the GGD convergence and fluctuation results

D.1 Proof of Theorem 3.1: infinite-particle case

In this section, we provide the proof of Theorem 3.1, which strengthens the bounds of (Marion et al., 2025), and decompose it into different steps.

Proof of Theorem 3.1. **Evolution of the loss.** Since $\nabla \ell(\theta_t) = \Gamma(\pi(\theta_t), \theta_t)$, using the evolution equation (2b) for θ_t yields

$$\begin{aligned}\frac{\mathrm{d}\ell(\theta_t)}{\mathrm{d}t} &= \left\langle \nabla \ell(\theta_t), \frac{\mathrm{d}}{\mathrm{d}t}\theta_t \right\rangle = -\varepsilon_t \left\langle \nabla \ell(\theta_t), \Gamma(p_t, \theta_t) \right\rangle \\ &= -\varepsilon_t \left\langle \nabla \ell(\theta_t), \Gamma(\pi(\theta_t), \theta_t) \right\rangle + \varepsilon_t \left\langle \nabla \ell(\theta_t), \Gamma(\pi(\theta_t), \theta_t) - \Gamma(p_t, \theta_t) \right\rangle \\ &\le -\varepsilon_t \|\nabla \ell(\theta_t)\|_{\mathbb{R}^p}^2 + \varepsilon_t \|\nabla \ell(\theta_t)\|_{\mathbb{R}^p} \|\Gamma(\pi(\theta_t), \theta_t) - \Gamma(p_t, \theta_t)\|_{\mathbb{R}^p}\,,\end{aligned}$$

by Cauchy–Schwarz inequality. We then use the $\sqrt{\mathrm{KL}}$-Lipschitz assumption on Γ to obtain

$$\begin{aligned}\frac{\mathrm{d}\ell(\theta_t)}{\mathrm{d}t} &\le -\varepsilon_t \|\nabla \ell(\theta_t)\|_{\mathbb{R}^p}^2 + \varepsilon_t L_\Gamma \|\nabla \ell(\theta_t)\|_{\mathbb{R}^p} \sqrt{\mathrm{KL}(p_t \,||\, \pi(\theta_t))} \\ &\le -\frac{\varepsilon_t}{2} \|\nabla \ell(\theta_t)\|_{\mathbb{R}^p}^2 + \frac{\varepsilon_t L_\Gamma^2}{2} \mathrm{KL}(p_t \,||\, \pi(\theta_t)),\end{aligned} \tag{12}$$

where we used Young's inequality $ab \le (a^2 + b^2)/2$ in the last step.

Evolution of the KL divergence of p_t from $\pi(\theta_t)$. We begin by studing the KL divergence term in Eq. (12). Since $\mathrm{KL}(p_t \,||\, \pi(\theta_t)) = \int_{\mathbb{R}^d} \log(p_t/\pi(\theta_t)) \,\mathrm{d}p_t(x)$, its first variations with respect to the first and second argument write

$$\frac{\delta \,\mathrm{KL}(p_t \,||\, \pi(\theta_t))}{\delta p_t} = \log\left(\frac{p_t}{\pi(\theta_t)}\right) + 1, \quad \text{and} \quad \frac{\delta \,\mathrm{KL}(p_t \,||\, \pi(\theta_t))}{\delta \pi(\theta_t)} = -\frac{p_t}{\pi(\theta_t)}.$$

Then, by the chain rule we have

$$\frac{\mathrm{d}\,\mathrm{KL}(p_t \,||\, \pi(\theta_t))}{\mathrm{d}t} = \underbrace{\int_{\mathbb{R}^d} \log\left(\frac{p_t}{\pi(\theta_t)}\right) \frac{\partial p_t}{\partial t}\,\mathrm{d}x}_{\text{a}} - \underbrace{\int_{\mathbb{R}^d} \frac{p_t}{\pi(\theta_t)} \frac{\partial \pi(\theta_t)}{\partial t}\,\mathrm{d}x}_{\text{b}}, \tag{13}$$

since $\int_{\mathbb{R}^d} \partial_t p_t \,\mathrm{d}x = \partial_t \int_{\mathbb{R}^d} p_t(x)\,\mathrm{d}x = \partial_t 1 = 0$. We analyze terms a) and b) separately below.

a) First, the law of the process defined in Eq. (2a) follows a Fokker-Planck equation

$$\frac{\partial p_t}{\partial t} = \nabla \cdot \left(p_t \nabla \log(p_t/\pi(\theta_t))\right).$$

Using an integration by parts and the log-Sobolev inequality assumption at all times $t > 0$, we have

$$\begin{aligned}
\text{a} &= \int_{\mathbb{R}^d} \log\left(\frac{p_t}{\pi(\theta_t)}\right) \frac{\partial p_t}{\partial t}\,\mathrm{d}x = \int_{\mathbb{R}^d} \log\left(\frac{p_t}{\pi(\theta_t)}\right) \nabla \cdot \left(p_t \nabla \log(p_t/\pi(\theta_t))\right)\,\mathrm{d}x \\
&= \langle \nabla \log(p_t/\pi(\theta_t)), -\nabla \log(p_t/\pi(\theta_t)) \rangle_{L^2(p_t)} = -\left\|\nabla \log(p_t/\pi(\theta_t))\right\|^2_{L^2(p_t)} \\
&\leq -2\mu\,\mathrm{KL}(p_t \,||\, \pi(\theta_t)).
\end{aligned}$$

b) Using the chain rule, $\partial_t \pi(\theta_t) = \langle \nabla_\theta \pi(\theta_t),\, \mathrm{d}\theta_t/\,\mathrm{d}t \rangle_{\mathbb{R}^p}$, and plugging in Eq. (2b), we can rewrite b) as

$$\begin{aligned}
\text{b} &= -\varepsilon_t \left\langle \Psi(p_t, \theta_t), \Gamma(p_t, \theta_t) \right\rangle_{\mathbb{R}^p} \\
&= -\varepsilon_t \left\langle \Psi(p_t, \theta_t), \Gamma(p_t, \theta_t) - \Gamma(\pi(\theta_t), \theta_t) \right\rangle_{\mathbb{R}^p} - \varepsilon_t \left\langle \Psi(p_t, \theta_t), \Gamma(\pi(\theta_t), \theta_t) \right\rangle_{\mathbb{R}^p},
\end{aligned}$$

where $\Psi(p_t, \theta_t) = \int_{\mathbb{R}^d} \nabla_\theta \log \pi(\theta_t)\,\mathrm{d}p_t(x)$. Note that

$$\begin{aligned}
\Psi(\pi(\theta_t), \theta_t) &= \int_{\mathbb{R}^d} \nabla_\theta \log \pi(\theta_t)(x)\,\mathrm{d}\pi(\theta_t)(x) = \int_{\mathbb{R}^d} \nabla_\theta \pi(\theta_t)(x)\,\mathrm{d}x = \nabla_\theta \int_{\mathbb{R}^d} \mathrm{d}\pi(\theta_t)(x) \\
&= \nabla_\theta 1 = 0.
\end{aligned}$$

Then,

$$\begin{aligned}
\left\|\Psi(p_t, \theta_t)\right\|_{\mathbb{R}^p} = \left\|\Psi(p_t, \theta_t) - \Psi(\pi(\theta_t), \theta_t)\right\|_{\mathbb{R}^p} &= \left\| \int_{\mathbb{R}^d} \nabla_\theta \log \pi(\theta_t)\,\mathrm{d}(p_t - \pi(\theta_t))(x) \right\|_{\mathbb{R}^p} \\
&\leq L_X \left\|p_t - \pi(\theta_t)\right\|_{W_1} \leq L_X \sqrt{\frac{2}{\mu}\,\mathrm{KL}(p_t \,||\, \pi(\theta_t))},
\end{aligned}$$

where we respectively used Lemmas C.2 and C.3 (along with $\|\cdot\|_{W_1} \leq \|\cdot\|_{W_2}$) to obtain the second and first inequalities since $x \mapsto \nabla_2 V(x, \theta)$ is L_X-Lipschitz and the LSI holds. By Cauchy-Schwarz inequality and the $\sqrt{\mathrm{KL}}$-Lipschitz assumption on Γ, we have that

$$\begin{aligned}
|\text{b}| &\leq \varepsilon_t \left\|\Psi(p_t, \theta_t)\right\|_{\mathbb{R}^p} \left(\left\|\Gamma(p_t, \theta_t) - \Gamma(\pi(\theta_t), \theta_t)\right\|_{\mathbb{R}^p} + \left\|\Gamma(\pi(\theta_t), \theta_t)\right\|_{\mathbb{R}^p}\right) \\
&\leq \varepsilon_t L_X \sqrt{\frac{2}{\mu}\,\mathrm{KL}(p_t \,||\, \pi(\theta_t))} \left(L_\Gamma \sqrt{\mathrm{KL}(p_t \,||\, \pi(\theta_t))} + \left\|\nabla \ell(\theta_t)\right\|_{\mathbb{R}^p}\right) \\
&\leq \sqrt{2}\varepsilon_t \frac{L_X L_\Gamma}{\sqrt{\mu}}\,\mathrm{KL}(p_t \,||\, \pi(\theta_t)) + \frac{8\varepsilon_t L_X^2}{\mu}\,\mathrm{KL}(p_t \,||\, \pi(\theta_t)) + \frac{\varepsilon_t}{4}\left\|\nabla \ell(\theta_t)\right\|^2_{\mathbb{R}^p},
\end{aligned}$$

where we used the identity $ab \leq (a^2 + b^2)/2$, with $a = 2L_X\sqrt{\frac{2}{\mu}\,\mathrm{KL}(p_t \,||\, \pi(\theta_t))}$ and $b = \left\|\nabla \ell(\theta_t)\right\|_{\mathbb{R}^p}/2$, to derive the last inequality.

Bounding the KL divergence of p_t from $\pi(\theta_t)$. Define the function $y : t \mapsto \mathrm{KL}(p_t \,||\, \pi(\theta_t))$. After combining the bounds on a) and b) with Eq. (13), we obtain the following coupled differential inequalities with Eq. (12):

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} \leq \left[L_X\left(\frac{\sqrt{2}L_\Gamma}{\sqrt{\mu}} + \frac{8L_X}{\mu}\right)\varepsilon_t - 2\mu\right] y(t) + \frac{\varepsilon_t}{4}\left\|\nabla \ell(\theta_t)\right\|^2_{\mathbb{R}^p}, \tag{14a}$$

$$\frac{\mathrm{d}\ell(\theta_t)}{\mathrm{d}t} \leq -\frac{\varepsilon_t}{2}\left\|\nabla \ell(\theta_t)\right\|^2_{\mathbb{R}^p} + \frac{\varepsilon_t L_\Gamma^2}{2} y(t). \tag{14b}$$

We sum Eqs. (14a) and (14b) to obtain

$$\frac{\mathrm{d}(y(t)+\ell(\theta_t))}{\mathrm{d}t} \leq \Bigg[\underbrace{\left(\left(\frac{L_\Gamma}{\sqrt{2}}+\frac{L_X}{\sqrt{\mu}}\right)^2+\frac{7L_X^2}{\mu}\right)}_{A}\varepsilon_t - 2\mu\Bigg] y(t) - \frac{\varepsilon_t}{4}\left\|\nabla\ell(\theta_t)\right\|_{\mathbb{R}^p}^2,$$

Let $T>0$, we begin by integrating the last equation over the interval $[0,T]$:

$$\int_0^T \frac{\varepsilon_t}{4}\left\|\nabla\ell(\theta_t)\right\|_{\mathbb{R}^p}^2 \,\mathrm{d}t \leq y(0)+\ell(\theta_0)+\int_0^T \left[A\varepsilon_t-2\mu\right] y(t)\,\mathrm{d}t, \tag{15}$$

where we dropped the term $-(y(T)+\ell(\theta_T))\leq 0$ on the right-hand side.

Case $\delta>0$**.** Since $A\varepsilon_t=\mathcal{O}(t^{-\delta})$, there exists $t_0\geq 0$ such that for all $t\geq t_0$, $A\varepsilon_t-2\mu\leq 0$, and $\int_{t_0}^T \left[A\varepsilon_t-2\mu\right] y(t)\,\mathrm{d}t\leq 0$.

Case $\delta=0$**.** If $\delta=0$, we observe that $-2\mu+\varepsilon A\leq 0$ following the stepsize assumption. This also makes the integral $\int_{t_0}^T \left[A\varepsilon_t-2\mu\right] y(t)\,\mathrm{d}t\leq 0$.

Therefore, after cutting the integral on the right-hand side of Eq. (15) at t_0, we obtain:

$$\int_0^T \varepsilon_t\|\nabla\ell(\theta_t)\|_{\mathbb{R}^p}^2\,\mathrm{d}t \leq 4(y(0)+\ell(\theta_0))+4\int_0^{t_0}\left[A\varepsilon_t-2\mu\right]y(t)\,\mathrm{d}t =: \tilde{C}.$$

Using the fact that $t\to\varepsilon_t$ is decreasing and dividing by $T\varepsilon_T=T^{1-\delta}$ yields

$$\frac{1}{T}\int_0^T \left\|\nabla\ell(\theta_t)\right\|_{\mathbb{R}^p}^2\,\mathrm{d}t \leq \frac{\tilde{C}}{T^{1-\delta}},$$

which concludes the proof. □

D.2 Proof of Theorem 3.2: finite-particle case

Before proving Theorem 3.2, we record the empirical-process estimate used to control the particle fluctuation term. The following lemma shows that the expectation of the norm of $\zeta_t^{(n)}:=\sqrt{n}(p_t^{(n)}-q_{t,n})$ is uniformly bounded for $n\geq 1$ and $t\geq 0$. Note that the proof does not require the functional central limit theorem proved later in the Appendix and hence does not need Assumption D.1.

Lemma D.1. *Let $t\geq 0$, $n\geq 1$, and $\zeta_t^{(n)}:=\sqrt{n}(p_t^{(n)}-q_{t,n})\in\mathcal{M}_0(\mathbb{R}^d)$. There exists a constant $C>0$, independent of n and t, such that*

$$\mathbb{E}\left[\|\zeta_t^{(n)}\|_{H^{-s}(\mathbb{R}^d)}^2\right]\leq C,$$

where $\|\zeta_t^{(n)}\|_{H^{-s}(\mathbb{R}^d)}:=\sup_{\|\varphi\|_{H^s(\mathbb{R}^d)}\leq 1}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n\varphi(X_t^i)-\mathbb{E}\varphi(X_t^i)\right|$.

Proof. We decompose the proof in the following two steps: 1) ghost symmetrization and 2) bounding with a RKHS. First, since the $(X_t^1,\ldots,X_t^n)$ are not independent, one cannot use a simple symmetrization lemma, e.g. (van der Vaart and Wellner, 1996, Lem. 2.3.6). To overcome this, we use a technique called ghost symmetrization, which does not rely on the independence of the $(X_t^1,\ldots,X_t^n)$. This is usually done by drawing $(Y_t^1,\ldots,Y_t^n)$ to be an independent copy of $(X_t^1,\ldots,X_t^n)$. Applying (van der Vaart and Wellner, 1996, Lem. 2.3.1) to the convex function $\Phi(u)=(\sqrt{n}u)^2$, and noting that Φ commutes with the supremum, implies that

$$\mathbb{E}\left[\|\zeta_t^{(n)}\|_{H^{-s}}^2\right]\leq 4\,\mathbb{E}_\epsilon\left[\mathbb{E}\left[\sup_{\|\varphi\|_{H^s}\leq 1}Z_\varphi^2\right]\right],\quad Z_\varphi:=\frac{1}{\sqrt{n}}\sum_{i=1}^n\epsilon_i\varphi(X_t^i), \tag{16}$$

where the ϵ_i are Rademacher random variables. Since $s>d/2$, the Sobolev space $H^s(\mathbb{R}^d)$ is continuously embedded into $C_b^0(\mathbb{R}^d)$, so point evaluation is a continuous linear functional on

$H^s(\mathbb{R}^d)$. Hence $H^s(\mathbb{R}^d)$ is a reproducing kernel Hilbert space (Berlinet and Thomas-Agnan, 2004, Thm. 1.3.2), with reproducing kernel

$$K_s(x,y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\xi\cdot(x-y)} (1+|\xi|^2)^{-s} \,\mathrm{d}\xi.$$

In particular, for every $x \in \mathbb{R}^d$ and every $\varphi \in H^s(\mathbb{R}^d)$, $\varphi(x) = \langle \varphi, K_s(x,\cdot)\rangle_{H^s(\mathbb{R}^d)}$. Therefore, conditionally on $X = (X_t^1, \dots, X_t^n)$,

$$Z_\varphi = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i \varphi(X_t^i) = \langle \varphi,\, \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i K_s(X_t^i,\cdot)\rangle_{H^s(\mathbb{R}^d)}.$$

Recall that $(\sup_{\|\varphi\|_{H^s(\mathbb{R}^d)}\le 1} |Z_\varphi|)^2 = \sup_{\|\varphi\|_{H^s(\mathbb{R}^d)}\le 1} Z_\varphi^2$, by Cauchy–Schwarz inequality,

$$\sup_{\|\varphi\|_{H^s(\mathbb{R}^d)}\le 1} Z_\varphi^2 \le \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i K_s(X_t^i,\cdot) \right\|^2_{H^s(\mathbb{R}^d)}.$$

Taking expectation with respect to ϵ, conditionally on X, yields

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\sup_{\|\varphi\|_{H^s(\mathbb{R}^d)}\le 1} Z_\varphi^2 \right] &\le \mathbb{E}_\epsilon\left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i K_s(X_t^i,\cdot) \right\|^2_{H^s(\mathbb{R}^d)} \right] \\
&= \frac{1}{n} \sum_{i,j=1}^n \mathbb{E}_\epsilon[\epsilon_i \epsilon_j] \, \langle K_s(X_t^i,\cdot), K_s(X_t^j,\cdot)\rangle_{H^s(\mathbb{R}^d)} \\
&= \frac{1}{n} \sum_{i=1}^n \|K_s(X_t^i,\cdot)\|^2_{H^s(\mathbb{R}^d)},
\end{aligned}$$

since $\mathbb{E}_\epsilon[\epsilon_i\epsilon_j] = \delta_{ij}$. By the reproducing property, $\|K_s(X_t^i,\cdot)\|^2_{H^s(\mathbb{R}^d)} = K_s(X_t^i, X_t^i) = K_s(0,0)$, since $K_s(x,y)$ is translation-invariant. Hence,

$$\mathbb{E}_\epsilon\left[\sup_{\|\varphi\|_{H^s(\mathbb{R}^d)}\le 1} Z_\varphi^2 \right] \le K_s(0,0) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} (1+|\xi|^2)^{-s} \,\mathrm{d}\xi,$$

which is finite since $s > d/2$. Therefore, there exists a constant $C_s > 0$, depending only on s and d, such that $\mathbb{E}_\epsilon[\sup_{\|\varphi\|_{H^s(\mathbb{R}^d)}\le 1} Z_\varphi^2] \le C_s$. Taking expectation with respect to X and combining with Eq. (16), we conclude that $\mathbb{E}\left[\|\zeta_t^{(n)}\|^2_{H^{-s}(\mathbb{R}^d)} \right] \le 4\, C_s$. □

We are now ready to prove Theorem 3.2.

Roadmap of the proof. The main quantity to compare is the algorithmic drift $\Gamma(p_t^{(n)}, \theta_t^{(n)})$ with the exact gradient $\Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)}) = \nabla \ell(\theta_t^{(n)})$. We insert the intermediate law $q_{t,n}$, namely the law of one particle driven by the random path $\theta_t^{(n)}$, and decompose the drift error as

$$\Gamma(p_t^{(n)}, \theta_t^{(n)}) - \Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)}) = \underbrace{\Gamma(p_t^{(n)}, \theta_t^{(n)}) - \Gamma(q_{t,n}, \theta_t^{(n)})}_{\text{finite-particle fluctuation}} + \underbrace{\Gamma(q_{t,n}, \theta_t^{(n)}) - \Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)})}_{\text{inner-loop bias}}.$$

The bias term is controlled through $y^{(n)}(t) = \mathrm{KL}(q_{t,n} \,||\, \pi(\theta_t^{(n)}))$, while the fluctuation term is encoded in $\Lambda_t^{(n)} = \sqrt{n}\big(\Gamma(p_t^{(n)}, \theta_t^{(n)}) - \Gamma(q_{t,n}, \theta_t^{(n)})\big)$.

Proof of Theorem 3.2. We begin by characterizing the evolution equation of the loss function.

Evolution of the loss. Since $\nabla \ell(\theta_t^{(n)}) = \Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)})$, we have,

$$\begin{aligned}
\frac{\mathrm{d}\ell(\theta_t^{(n)})}{\mathrm{d}t} &= \left\langle \nabla \ell(\theta_t^{(n)}), \frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(n)} \right\rangle = -\varepsilon_t \left\langle \nabla \ell(\theta_t^{(n)}), \Gamma(p_t^{(n)}, \theta_t^{(n)}) \right\rangle \\
&= -\varepsilon_t \langle \nabla \ell(\theta_t^{(n)}), \Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)}) \rangle + \varepsilon_t \langle \nabla \ell(\theta_t^{(n)}), \Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)}) - \Gamma(q_{t,n}, \theta_t^{(n)}) \rangle \\
&\quad + \varepsilon_t \langle \nabla \ell(\theta_t^{(n)}), \Gamma(q_{t,n}, \theta_t^{(n)}) - \Gamma(p_t^{(n)}, \theta_t^{(n)}) \rangle \\
&\leq -\varepsilon_t \|\nabla \ell(\theta_t^{(n)})\|_{\mathbb{R}^p}^2 + \varepsilon_t \|\nabla \ell(\theta_t^{(n)})\|_{\mathbb{R}^p} \|\Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)}) - \Gamma(q_{t,n}, \theta_t^{(n)})\|_{\mathbb{R}^p} \\
&\quad + \varepsilon_t \|\nabla \ell(\theta_t^{(n)})\|_{\mathbb{R}^p} \|\Gamma(q_{t,n}, \theta_t^{(n)}) - \Gamma(p_t^{(n)}, \theta_t^{(n)})\|_{\mathbb{R}^p}.
\end{aligned}$$

The first term above corresponds to the decay if the inner-loop solution was achieved instantaneously. The second and third term, respectively, take into account the fact that the inner loop is not solved exactly (so the dynamic is lagging behind) and the error in the drift Γ arising from the finite-particle system empirical measure instead of the true distribution of the sampling process. Define $\Lambda_t^{(n)} \coloneqq \sqrt{n}(\Gamma(p_t^{(n)}, \theta_t^{(n)}) - \Gamma(q_{t,n}, \theta_t^{(n)}))$ and $y^{(n)}(t) \coloneqq \mathrm{KL}(q_{t,n} \,||\, \pi(\theta_t^{(n)}))$. By the H^{-s}-Lipschitz continuity of Γ and Lemma C.4, Γ is $\sqrt{\mathrm{KL}}$-Lipschitz; we denote the induced constant by L_Γ. Hence,

$$\begin{aligned}
\frac{\mathrm{d}\ell(\theta_t^{(n)})}{\mathrm{d}t} &\leq -\varepsilon_t \|\nabla \ell(\theta_t^{(n)})\|_{\mathbb{R}^p}^2 + \varepsilon_t \|\nabla \ell(\theta_t^{(n)})\|_{\mathbb{R}^p} \left(L_\Gamma \sqrt{y^{(n)}(t)} + \frac{1}{\sqrt{n}} \|\Lambda_t^{(n)}\|_{\mathbb{R}^p} \right) \\
&\leq -\frac{1}{2}\varepsilon_t \|\nabla \ell(\theta_t^{(n)})\|_{\mathbb{R}^p}^2 + \frac{\varepsilon_t}{2} \left(L_\Gamma \sqrt{y^{(n)}(t)} + \frac{1}{\sqrt{n}} \|\Lambda_t^{(n)}\|_{\mathbb{R}^p} \right)^2 \\
&\leq -\frac{1}{2}\varepsilon_t \|\nabla \ell(\theta_t^{(n)})\|_{\mathbb{R}^p}^2 + L_\Gamma^2 \varepsilon_t y^{(n)}(t) + \varepsilon_t \frac{\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}^2}{n},
\end{aligned} \tag{17}$$

where we used the identity $ab \leq (a^2 + b^2)/2$ for the first inequality and $(a+b)^2 \leq 2(a^2+b^2)$ for the last inequality.

Evolution of $y^{(n)}(t)$. Similarly to Eq. (13) in Section D.1, we obtain

$$\frac{\mathrm{d}y^{(n)}(t)}{\mathrm{d}t} = \frac{\mathrm{d}\,\mathrm{KL}(q_{t,n} \,||\, \pi(\theta_t^{(n)}))}{\mathrm{d}t} = \underbrace{\int_{\mathbb{R}^d} \log\left(\frac{q_{t,n}}{\pi(\theta_t^{(n)})} \right) \frac{\partial q_{t,n}}{\partial t} \,\mathrm{d}x}_{\text{a}} - \underbrace{\int_{\mathbb{R}^d} \frac{q_{t,n}}{\pi(\theta_t^{(n)})} \frac{\partial \pi(\theta_t^{(n)})}{\partial t} \,\mathrm{d}x}_{\text{b}}. \tag{18}$$

We analyze the two terms on the right-hand side of Eq. (18) independently below.

a) Recall that the law of the process defined in Eq. (3a) follows a Fokker-Planck equation:

$$\frac{\partial q_{t,n}}{\partial t} = \nabla \cdot \left[q_{t,n} \nabla \log(q_{t,n}/\pi(\theta_t^{(n)})) \right].$$

Using an integration by parts and the log-Sobolev inequality assumption at $t > 0$, we have

$$\begin{aligned}
\text{a} &= \int_{\mathbb{R}^d} \log(q_{t,n}/\pi(\theta_t^{(n)})) \nabla \cdot \left[q_{t,n} \nabla \log(q_{t,n}/\pi(\theta_t^{(n)})) \right] \,\mathrm{d}x \\
&= -\langle \nabla \log(q_{t,n}/\pi(\theta_t^{(n)})), \nabla \log(q_{t,n}/\pi(\theta_t^{(n)})) \rangle_{L^2(q_{t,n})} \\
&= -\|\nabla \log(q_{t,n}/\pi(\theta_t^{(n)}))\|_{L^2(q_{t,n})}^2 \leq -2\mu y^{(n)}(t).
\end{aligned}$$

b) By the chain rule, $\partial_t \pi(\theta_t^{(n)}) = \langle \nabla_\theta \pi(\theta_t^{(n)}), \, \mathrm{d}\theta_t^{(n)}/\,\mathrm{d}t \rangle_{\mathbb{R}^p}$. After plugging in Eq. (3b), we can rewrite b) as

$$\text{b} = -\varepsilon_t \left\langle \Psi(q_{t,n}, \theta_t^{(n)}), \frac{1}{\sqrt{n}} \Lambda_t^{(n)} + [\Gamma(q_{t,n}, \theta_t^{(n)}) - \Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)})] + \Gamma(\pi(\theta_t^{(n)}), \theta_t^{(n)}) \right\rangle_{\mathbb{R}^p},$$

where $\Psi(q_{t,n}, \theta_t^{(n)}) = \int_{\mathbb{R}^d} \nabla_\theta \log \pi(\theta_t^{(n)}) \,\mathrm{d}q_{t,n}(x)$. With the same technique as in Section D.1, where we respectively used Lemmas C.2 and C.3 since the assumption on $\nabla_2 V$ and the LSI hold, we have

$$\|\Psi(q_{t,n}, \theta_t^{(n)}) - \Psi(\pi(\theta_t^{(n)}), \theta_t^{(n)})\|_{\mathbb{R}^p} \leq L_X \sqrt{\frac{2}{\mu} y^{(n)}(t)},$$

and $\Psi(\pi(\theta_t^{(n)}), \theta_t^{(n)}) = 0$. This, combined with the last inequality gives $\|\Psi(q_{t,n}, \theta_t^{(n)})\|_{\mathbb{R}^p} \leq L_X\sqrt{2y^{(n)}(t)/\mu}$. In the same vein as in Section D.1, by Cauchy-Schwarz inequality and the $\sqrt{\mathrm{KL}}$-Lipschitz continuity of Γ, we have

$$\begin{aligned}|\mathrm{b}| &\leq \varepsilon_t L_X\sqrt{\frac{2}{\mu}y^{(n)}(t)}\left(L_\Gamma\sqrt{y^{(n)}(t)} + \|\nabla\ell(\theta_t)\|_{\mathbb{R}^p} + \frac{1}{\sqrt{n}}\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}\right)\\ &\leq \varepsilon_t L_X\left(\frac{\sqrt{2}L_\Gamma}{\sqrt{\mu}} + \frac{8L_X}{\mu}\right)y^{(n)}(t) + \frac{\varepsilon_t}{3}\|\nabla\ell(\theta_t^{(n)})\|_{\mathbb{R}^p}^2 + \frac{\varepsilon_t}{10n}\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}^2,\end{aligned}$$

where we used two times the identity $ab \leq (a^2+b^2)/2$, to derive the last inequality. The first time with $a = \sqrt{3}L_X\sqrt{\frac{2}{\mu}y^{(n)}(t)}$ and $b = \|\nabla\ell(\theta_t)\|_{\mathbb{R}^p}^2/\sqrt{3}$, and the second time with $a = \sqrt{5}L_X\sqrt{\frac{2}{\mu}y^{(n)}(t)}$ and $b = \frac{1}{\sqrt{5n}}\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}$.

After combining the bounds on a) and b) with Eq. (18) and taking the expectation by dominated convergence theorem, we obtain the following coupled differential inequalities with Eq. (17):

$$\frac{\mathrm{d}\mathbb{E}[y^{(n)}(t)]}{\mathrm{d}t} \leq \left[L_X\left(\frac{\sqrt{2}L_\Gamma}{\sqrt{\mu}} + \frac{8L_X}{\mu}\right)\varepsilon_t - 2\mu\right]\mathbb{E}[y^{(n)}(t)] + \frac{\varepsilon_t}{3}\mathbb{E}[\|\nabla\ell(\theta_t^{(n)})\|^2] + \frac{\varepsilon_t}{10n}\mathbb{E}[\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}^2], \tag{19a}$$

$$\frac{\mathrm{d}\mathbb{E}[\ell(\theta_t^{(n)})]}{\mathrm{d}t} \leq -\frac{\varepsilon_t}{2}\mathbb{E}[\|\nabla\ell(\theta_t^{(n)})\|^2] + L_\Gamma^2\varepsilon_t\mathbb{E}[y^{(n)}(t)] + \frac{\varepsilon_t}{n}\mathbb{E}[\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}^2]. \tag{19b}$$

Bounding $\mathbb{E}[\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}^2]$ non-asymptotically. We focus on controlling the last term in Eq. (19b). Under the H^{-s}-Lipschitz assumption on Γ,

$$\|\Lambda_t^{(n)}\|_{\mathbb{R}^p} = \sqrt{n}\|\Gamma(q_{t,n}, \theta_t^{(n)}) - \Gamma(p_t^{(n)}, \theta_t^{(n)})\|_{\mathbb{R}^p} \leq L_{\Gamma,s}\|\zeta_t^{(n)}\|_{H^{-s}(\mathbb{R}^d)},$$

where $\zeta_t^{(n)} \coloneqq \sqrt{n}(p_t^{(n)} - q_{t,n})$. Squaring and taking the expectation give

$$\mathbb{E}[\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}^2] \leq L_{\Gamma,s}^2\mathbb{E}[\|\zeta_t^{(n)}\|_{H^{-s}(\mathbb{R}^d)}^2].$$

Following Lemma D.1, the right-hand side can be uniformly bounded and there exists a constant $C > 0$, independent of n and t, such that $\mathbb{E}[\|\zeta_t^{(n)}\|_{H^{-s}(\mathbb{R}^d)}^2] \leq C$. Therefore,

$$\mathbb{E}[\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}^2] \leq L_{\Gamma,s}^2 C =: D. \tag{20}$$

Bounding the loss function. Plugging Eq. (20) in Eq. (19) gives

$$\frac{\mathrm{d}\mathbb{E}[y^{(n)}(t)]}{\mathrm{d}t} \leq \left[(A - L_\Gamma^2)\varepsilon_t - 2\mu\right]\mathbb{E}[y^{(n)}(t)] + \frac{\varepsilon_t}{3}\mathbb{E}[\|\nabla\ell(\theta_t^{(n)})\|^2] + \frac{D\varepsilon_t}{10n}, \tag{21a}$$

$$\frac{\mathrm{d}\mathbb{E}[\ell(\theta_t^{(n)})]}{\mathrm{d}t} \leq L_\Gamma^2\varepsilon_t\mathbb{E}[y^{(n)}(t)] - \frac{\varepsilon_t}{2}\mathbb{E}[\|\nabla\ell(\theta_t^{(n)})\|^2] + \frac{D\varepsilon_t}{n}, \tag{21b}$$

where A is the same as in Eq. (15). We sum Eqs. (21a) and (21b) to obtain

$$\frac{\mathrm{d}(\mathbb{E}[y^{(n)}(t)] + \mathbb{E}[\ell(\theta_t^{(n)})])}{\mathrm{d}t} \leq (A\varepsilon_t - 2\mu)\mathbb{E}[y^{(n)}(t)] - \frac{\varepsilon_t}{6}\mathbb{E}[\|\nabla\ell(\theta_t^{(n)})\|^2] + \frac{11D\varepsilon_t}{10n}, \tag{22}$$

Let $T > 0$, we begin by integrating the last equation over the interval $[0, T]$:

$$\frac{1}{6}\int_0^T \varepsilon_t\mathbb{E}[\|\nabla\ell(\theta_t^{(n)})\|^2]\,\mathrm{d}t \leq \mathbb{E}[y^{(n)}(0) + \ell(\theta_0^{(n)})] + \int_0^T(-2\mu + A\varepsilon_t)\mathbb{E}[y^{(n)}(t)]\,\mathrm{d}t + \frac{11D}{10n}\int_0^T \varepsilon_t\,\mathrm{d}t, \tag{23}$$

where we dropped the term $-\mathbb{E}[y^{(n)}(T) + \ell(\theta_T^{(n)})] \leq 0$ on the right-hand side. We analyze each term of Eq. (23) independently. First, $\mathbb{E}[y^{(n)}(0) + \ell(\theta_0^{(n)})]$ is independent of n since $\theta_0^{(n)} = \theta_0$ and $q_{0,n} = p_0$. Then, following the proof of Section D.1, there exists $t_0 \geq 0$ such that $A\varepsilon_t - 2\mu \leq 0$ for

all $t \geq t_0$. Next, we remark that $\int_0^T \varepsilon_t \, \mathrm{d}t \sim (1-\delta)^{-1} T^{1-\delta}$ as $T \to \infty$. Therefore, there exist two constants $c_1, c_2 > 0$, independent of n and T such that Eq. (23) becomes:

$$\frac{1}{T} \int_0^T \mathbb{E}[\|\nabla \ell(\theta_t^{(n)})\|^2] \, \mathrm{d}t \leq \frac{c_1}{T^{1-\delta}} + \frac{c_2}{n} + \frac{6}{T^{1-\delta}} \int_0^{t_0} |A\varepsilon_t - 2\mu| \mathbb{E}[y^{(n)}(t)] \, \mathrm{d}t, \tag{24}$$

since $y^{(n)}(t) \geq 0$. We bound the integral on the right-hand side of Eq. (24) using Eq. (22) as

$$\frac{\mathrm{d}(\mathbb{E}[y^{(n)}(t)] + \mathbb{E}[\ell(\theta_t^{(n)})])}{\mathrm{d}t} \leq |A\varepsilon_t - 2\mu| \mathbb{E}[y^{(n)}(t) + \ell(\theta_t^{(n)})] + \frac{c_3}{n},$$

since $y^{(n)}(t) \leq y^{(n)}(t) + \ell(\theta_t^{(n)})$, and where we introduced the constant $c_3 > 0$ independent of n and t. Applying Gronwall's inequality (Lemma C.1) over $[0, t]$ for $t \leq t_0$ yields[2]

$$\mathbb{E}[y^{(n)}(t)] \leq \left[\mathbb{E}[y^{(n)}(0) + \ell(\theta_0^{(n)})] + \frac{c_3}{n}\right] \exp\left(\int_0^t |A\varepsilon_s - 2\mu| \, \mathrm{d}s\right) \leq d_1 + \frac{d_2}{n},$$

where we dropped the term $-\mathbb{E}[\ell(\theta_t^{(n)})] \leq 0$ on the right-hand side of the first inequality and introduced the constants $d_1, d_2 > 0$, independent of n and t, by bounding uniformly over $[0, t_0]$. Multiplying the last inequality by $|\varepsilon_t A - 2\mu|$ and integrating over $[0, t_0]$ yield

$$\int_0^{t_0} |\varepsilon_t A - 2\mu| \mathbb{E}[y^{(n)}(t)] \, \mathrm{d}t \leq \left(d_1 + \frac{d_2}{n}\right) \int_0^{t_0} |A\varepsilon_t - 2\mu| \, \mathrm{d}t.$$

Combining this with Eq. (24), we obtain the existence of constants $c_1, c_2, c_3 > 0$, independent of n and T, such that

$$\frac{1}{T} \int_0^T \mathbb{E}[\|\nabla \ell(\theta_t^{(n)})\|^2] \, \mathrm{d}t \leq \frac{c_1}{T^{1-\delta}} + \frac{c_2}{n} + \frac{c_3}{nT^{1-\delta}},$$

which concludes the proof. □

D.3 Proof of Proposition 3.1: lower bound

Proof of Proposition 3.1. Let $\chi_R \in C_c^\infty(\mathbb{R})$ be even, with $0 \leq \chi_R \leq 1$, $\chi_R = 1$ on $[-R, R]$, and $\chi_R = 0$ outside $[-2R, 2R]$, and set $a_R(x) = \chi_R(x) \tanh(x)$. We consider the potential $V(x, \theta) = x^2/2 + \tanh(\theta) a_R(x)$ for $(x, \theta) \in \mathbb{R} \times \mathbb{R}$, leading to

$$\pi(\theta)(x) \propto e^{-x^2/2} e^{-\tanh(\theta) a_R(x)}, \quad (x, \theta) \in \mathbb{R} \times \mathbb{R}.$$

For $(p, \theta) \in \mathcal{P}(\mathbb{R}) \times \mathbb{R}$, define $\Gamma(p, \theta) = \int_{\mathbb{R}} R(x) \, \mathrm{d}p(x)$ with $R(x) = u a_R(x)$ for some $u > 0$.

Verification of the regularity assumptions. The function $\theta \mapsto \pi(\theta)$ is differentiable and $\nabla_2 V(x, \theta) = a_R(x)/\cosh^2 \theta$ is L_X-Lipschitz. Also, the function $R : x \mapsto u a_R(x)$ is Lipschitz and bounded, hence Γ is $\sqrt{\mathrm{KL}}$-Lipschitz by Lemma C.4. Moreover, since $a_R \in C_c^\infty(\mathbb{R}) \subset H^s(\mathbb{R})$ for every s, we have

$$|\Gamma(p, \theta) - \Gamma(q, \theta)| \leq |u| \, \|a_R\|_{H^s} \|p - q\|_{H^{-s}},$$

so Γ is H^{-s}-Lipschitz. For the LSI, the potential of the Gibbs distribution $\pi(\theta)(x) \propto e^{-x^2/2} e^{-\tanh(\theta) a_R(x)}$ is the sum of a 1-strongly convex term and a bounded perturbation. A classical result (Holley and Stroock, 1987; Bakry et al., 2014) states that $\pi(\theta_t^{(n)})$ satisfies the LSI with constant $\mu \geq \exp(-2\|a_R\|_\infty)$. This corresponds to the uniform LSI constant profile.

Lower bound. Since $\theta_0 \in \mathbb{R}$ satisfies $\ell'(\theta_0) \neq 0$, there exist $r, m > 0$ such that $|\ell'(\theta)| \geq m$ for all $\theta \in (\theta_0 - r, \theta_0 + r)$ by continuity of ℓ'. From Eq. (3b), one has that $\theta_t^{(n)}$ tends to $\theta_0^{(n)} = \theta_0$ as t tends to 0. Therefore, by continuity at $t = 0$, there exists some $\Delta > 0$ such that $\theta_t^{(n)} \in (\theta_0, r)$ for all $t \in [0, \Delta]$. On this window, $|\ell'(\theta_t^{(n)})| \geq m$. For each $t \geq \Delta$, define $\tilde{\theta}_s := \theta_t^{(n)}$ for $s \in [t - \Delta, t]$, and run the inner stochastic differential equation on this window, with the same Brownian motions as in Eq. (3a) and the same initial states at $t - \Delta$:

$$\mathrm{d}Y_s^i = -\nabla_1 V(Y_s^i, \tilde{\theta}_s) \, \mathrm{d}s + \sqrt{2} \, \mathrm{d}B_s^i, \quad Y_{t-\Delta}^i = X_{t-\Delta}^i.$$

[2] Note that this is necessary to control the dependence of the constant arising from the integral in Eq. (24) with respect to n.

Given the path $\tilde{\theta}$, the Y_t^i are i.i.d. of law $\tilde{q}_t$ (same drift, independent noises, i.i.d. initialization at $t-\Delta$). Let $\tilde{G}_t^{(n)} = \Gamma(\frac{1}{n}\sum_{i=1}^n \delta_{Y_t^i}, \cdot) = \frac{1}{n}\sum_{i=1}^n R(Y_t^i)$. By i.i.d. at time t,

$$\mathbb{E}\left[(\tilde{G}_t^{(n)})^2 \mid \tilde{\theta}\right] = \frac{1}{n}\mathbb{V}_{X\sim\tilde{q}_t}(R(X)) + (\mathbb{E}_{X\sim\tilde{q}_t}R(X))^2 \geq \frac{1}{n}\mathbb{V}_{X\sim\tilde{q}_t}(R(X)),$$

and $\mathbb{V}_{X\sim\tilde{q}_t}(R(X)) = c_0 u^2$ for some $c_0 > 0$. Therefore, $\mathbb{E}[(\tilde{G}_t^{(n)})^2 \mid \tilde{\theta}] \geq c_0 u^2/n$ for all $t \geq \Delta$. Then, let $G_t^{(n)} = \Gamma(p_t^{(n)}, \cdot) = \frac{1}{n}\sum_{i=1}^n R(X_t^i)$ be the real empirical average. Following Minkowski's inequality, we have

$$\mathbb{E}\left[(G_t^{(n)})^2\right]^{1/2} \geq \mathbb{E}\left[(\tilde{G}_t^{(n)})^2\right]^{1/2} - \mathbb{E}\left[(G_t^{(n)} - \tilde{G}_t^{(n)})^2\right]^{1/2}.$$

However, $\mathbb{E}\left[(G_t^{(n)} - \tilde{G}_t^{(n)})^2\right] \leq \frac{C_R u^2}{n^2}\sum_{i=1}^n \mathbb{E}\left[(X_t^i - Y_t^i)^2\right] = \frac{C_R u^2}{n}\mathbb{E}\left[(X_t^1 - Y_t^1)^2\right]$ since R is Lipschitz and each $(X_t^i - Y_t^i)^2$ has the same law. Let $Z_t = X_t^1 - Y_t^1$. One has that $\mathrm{d}Z_s = -(\nabla_1 V(X_s^1, \theta_s^{(n)}) - \nabla_1 V(Y_s^i, \tilde{\theta}_t))$ for $t-\Delta \leq s \leq t$. Since $\nabla_1 V$ is Lipschitz in both arguments,

$$\mathrm{d}\mathbb{E}\left[Z_s^2\right] \leq \mathbb{E}\left[Z_s^2\right]\mathrm{d}s + \mathbb{E}\left[|\theta_s^{(n)} - \theta_t^{(n)}|^2\right]\mathrm{d}s.$$

Since $Z_{t-\Delta} = 0$, applying Lemma C.1 from $t-\Delta$ to t yields

$$\mathrm{d}\mathbb{E}\left[Z_t^2\right] \leq \int_{t-\Delta}^t \mathbb{E}\left[|\theta_s^{(n)} - \theta_t^{(n)}|^2\right]\mathrm{d}s.$$

However, $|\theta_s^{(n)} - \theta_t^{(n)}|^2 \leq C_R\,|u|\,(t-s)$. Hence, $\mathbb{E}[(G_t^{(n)} - \tilde{G}_t^{(n)})^2] \leq C_R u^2 \Delta^3/n$ so

$$\mathbb{E}\left[(G_t^{(n)})^2\right] \geq \frac{u^2}{n}(\sqrt{c_0} - \sqrt{C_R\Delta^3})^2.$$

Finally, remark that $G_t^{(n)} = \nabla\ell(\theta_t^{(n)}) + \eta_t^{(n)}$ with $\mathbb{E}[(\eta_t^{(n)})^2] \leq c_1/n$ for some $c_1 > 0$. Hence, one can lower bound $\mathbb{E}\left[|\nabla\ell(\theta_t^{(n)})|^2\right]$ by $u^2(\sqrt{c_0} - \sqrt{C_R\Delta^3})^2/n - c_1/n$, and choose a value for Δ so that the lower bound $c = u^2(\sqrt{c_0} - \sqrt{C_R\Delta^3})^2 - c_1$ is such that $c > 0$. This leads to the desired results, i.e.

$$\liminf_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{E}\left[|\ell'(\theta_t^{(n)})|^2\right]\mathrm{d}t \geq \frac{c}{n},$$

□

D.4 Proofs of the asymptotic results

D.4.1 Proof of Proposition 3.2: central limit theorem

We begin by stating the extra regularity assumptions needed to apply (Budhiraja et al., 2017, Thm. 2.4).

Assumption D.1 (CLT regularity on the time horizon $0 \leq t \leq T$)**.**

(i) The schedule $t \mapsto \varepsilon_t$ belongs to $C_b^2([0,T])$.

(ii) The particle drift is regular: $\nabla_1 V$ is bounded on $\mathbb{R}^d \times \mathbb{R}^p$, globally Lipschitz in (x,θ), and for each $x \in \mathbb{R}^d$, the map $\theta \mapsto \nabla_1 V(x,\theta)$ is twice continuously differentiable, with first and second θ-derivatives bounded uniformly in x.

(iii) The Gibbs gradient has the finite-observable form of Eq. (4)*, and this representation is uniformly regular on the time horizon. For each $1 \leq j \leq r$, ϕ_j^θ and its first two θ-derivatives are bounded Lipschitz functions of x, uniformly in θ. The map $(\theta, z_1, \ldots, z_r) \mapsto H_\theta(z_1, \ldots, z_r)$ and its first two derivatives are bounded on the range of the observable moments reached on $[0,T]$. The map is C^2 on a neighborhood of the line segments used in the Taylor expansions below, with the same bounds along these segments.*

First, Assumption D.1(iii) implies that Γ is bounded and globally Lipschitz in (ν,θ) when $\mathcal{P}(\mathbb{R}^d)$ is equipped with d_{BL}. Indeed, the observable maps are uniformly bounded Lipschitz and the finite-dimensional map H_θ has bounded first derivative on the relevant range.

Proof of Proposition 3.2. Fix $T > 0$. To fit the time-dependent common-factor drift

$$b_0(t, \theta, \nu) = -\varepsilon_t \Gamma(\nu, \theta)$$

into the autonomous framework of Budhiraja et al. (2017), we augment the common factor by deterministic time and set

$$\bar{U}_t^n \coloneqq (t, \theta_t^{(n)}) \in \mathbb{R}^{1+p}, \qquad \bar{U}_t \coloneqq (t, \theta_t) \in \mathbb{R}^{1+p}.$$

On $[0, T]$, the particle system can then be written as

$$\mathrm{d}X_t^i = \bar{b}(X_t^i, \bar{U}_t^n, p_t^{(n)})\,\mathrm{d}t + \bar{\sigma}\,\mathrm{d}W_t^i, \qquad \mathrm{d}\bar{U}_t^n = \bar{b}_0(\bar{U}_t^n, p_t^{(n)})\,\mathrm{d}t,$$

with

$$\bar{b}(x, (s, \theta), \nu) \coloneqq -\nabla_1 V(x, \theta), \qquad \bar{b}_0((s, \theta), \nu) \coloneqq \big(1, -\varepsilon_s \Gamma(\nu, \theta)\big).$$

The initial condition is the one fixed in the model formulation of Section 2: θ_0 is deterministic and the particles start i.i.d. with common law p_0. The particle diffusion coefficient is constant. To match Budhiraja's structural form, we view it as

$$\bar{\sigma} = [I_d\ I_d] \in \mathbb{R}^{d \times 2d},$$

driven by a $2d$-dimensional Brownian motion, so that $\bar{\sigma}\bar{\sigma}^\intercal = 2I_d$. The common factor has no diffusion term: $\bar{\sigma}_0 \equiv 0$. The jump coefficients d and d_0 are absent.

We now verify the assumptions used in (Budhiraja et al., 2017, Thm. 2.4).

1. **Condition 2.1.** This condition asks for boundedness and Lipschitz continuity of the coefficients. For the particle drift, this follows from Assumption D.1(ii). For the common-factor drift, Assumption D.1(i) controls the schedule on $[0, T]$, while the consequence of Assumption D.1(iii) stated above gives boundedness and Lipschitz continuity of Γ. The diffusion coefficients $\bar{\sigma}$ and $\bar{\sigma}_0$ are constant.

2. **Condition 2.3.** This condition fixes the diffusion structure. Our representation

   $$\bar{\sigma} = [I_d\ I_d] \in \mathbb{R}^{d \times 2d}$$

   is of the required form and generates the covariance $\bar{\sigma}\bar{\sigma}^\intercal = 2I_d$ appearing in the particle dynamics.

3. **Condition 6.2.** This condition requires a first-order expansion of the common-factor drift with a quadratic remainder. We verify it for

   $$\bar{b}_0((s, \theta), \nu) = \big(1, -\varepsilon_s \Gamma(\nu, \theta)\big).$$

   Write

   $$m_\theta(\nu) \coloneqq \big(\mathbb{E}_\nu[\phi_1^\theta], \ldots, \mathbb{E}_\nu[\phi_r^\theta]\big).$$

   Taylor expansion of the finite-dimensional map H_θ together with the θ-smoothness of the observables gives bounded continuous maps $\Gamma^{(1)}$ and $\Gamma^{(2)}$ defined by

   $$\Gamma^{(1)}(\theta, \nu)h = \partial_\theta H_\theta(m_\theta(\nu))h + \sum_{j=1}^{r} D_j H_\theta(m_\theta(\nu))\,\mathbb{E}_\nu[\partial_\theta \phi_j^\theta h],$$

   and

   $$\Gamma^{(2)}(\theta, \nu, x) = \sum_{j=1}^{r} D_j H_\theta(m_\theta(\nu))\,\phi_j^\theta(x).$$

   Hence, for some constant $c_\Gamma > 0$,

   $$\Gamma(\nu', \theta') - \Gamma(\nu, \theta) = \Gamma^{(1)}(\theta, \nu)(\theta' - \theta) + \langle \Gamma^{(2)}(\theta, \nu, \cdot), \nu' - \nu \rangle + r_\Gamma(\theta, \theta', \nu, \nu'),$$

   with

   $$\|r_\Gamma(\theta, \theta', \nu, \nu')\| \le c_\Gamma \left(\|\theta' - \theta\|^2 + d_{\mathrm{BL}}(\nu, \nu')^2\right).$$

   Moreover, Assumption D.1(i) gives a bounded continuous derivative $\dot{\varepsilon}$ and a constant $c_\varepsilon > 0$ such that

   $$\varepsilon_{s'} - \varepsilon_s = \dot{\varepsilon}_s(s' - s) + r_\varepsilon(s, s'), \qquad |r_\varepsilon(s, s')| \le c_\varepsilon |s' - s|^2.$$

Writing $\Delta s = s' - s$, $\Delta\theta = \theta' - \theta$, and $\Delta\nu = \nu' - \nu$, define

$$\bar{b}_0^{(1)}((s,\theta),\nu)(\alpha,h) \coloneqq \Big(0, -\dot{\varepsilon}_s\Gamma(\nu,\theta)\,\alpha - \varepsilon_s\Gamma^{(1)}(\theta,\nu)h\Big),$$

and

$$\bar{b}_0^{(2)}((s,\theta),\nu,x) \coloneqq \Big(0, -\varepsilon_s\Gamma^{(2)}(\theta,\nu,x)\Big).$$

Since the first component of $\bar{b}_0$ is constant,

$$\bar{b}_0((s',\theta'),\nu') - \bar{b}_0((s,\theta),\nu) = \bar{b}_0^{(1)}((s,\theta),\nu)(\Delta s,\Delta\theta) + \langle \bar{b}_0^{(2)}((s,\theta),\nu,\cdot),\Delta\nu\rangle + \bar{r}_\Gamma((s,\theta),(s',\theta'),\nu,\nu'),$$

where

$$\bar{r}_\Gamma((s,\theta),(s',\theta'),\nu,\nu') \coloneqq \Big(0,\ -\varepsilon_s r_\Gamma(\theta,\theta',\nu,\nu') - \dot{\varepsilon}_s\Delta s\big[\Gamma^{(1)}(\theta,\nu)\Delta\theta + \langle\Gamma^{(2)}(\theta,\nu,\cdot),\Delta\nu\rangle + r_\Gamma(\theta,\theta',\nu,\nu')\big] - r_\varepsilon(s,s')\Gamma(\nu',\theta')\Big).$$

Because all coefficients are bounded on $[0,T]$, the bounds on r_Γ and r_ε imply that, for some $\bar{c}_\Gamma > 0$,

$$\|\bar{r}_\Gamma((s,\theta),(s',\theta'),\nu,\nu')\| \le \bar{c}_\Gamma\left(|s'-s|^2 + \|\theta'-\theta\|^2 + d_{\mathrm{BL}}(\nu,\nu')^2\right).$$

This is the required S_1 expansion for $\bar{b}_0$. Since $\bar{\sigma}_0 \equiv 0$, there is no additional diffusion expansion. The special linear case is recovered from Eq. (4) by taking $r = 1$ and $H_\theta = \mathrm{Id}$.

4. **Condition 7.1.** This condition requires the analogous first-order expansion for the particle drift. By Taylor's theorem and Assumption D.1(ii), there exist a bounded continuous map

$$b_2^V : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^{d\times p}$$

and a constant $c_V > 0$ such that, for all $x \in \mathbb{R}^d$ and $\theta,\theta' \in \mathbb{R}^p$,

$$-\nabla_1 V(x,\theta') + \nabla_1 V(x,\theta) = b_2^V(x,\theta)(\theta'-\theta) + r_V(x,\theta,\theta'),$$

with

$$\|r_V(x,\theta,\theta')\| \le c_V\|\theta'-\theta\|^2.$$

Since the particle drift $b(x,\theta,\nu) = -\nabla_1 V(x,\theta)$ does not depend on ν, Budhiraja's measure-derivative term is identically zero.

5. **Condition 7.4.** This condition is vacuous because there are no jumps, i.e., $d = d_0 = 0$.

Let $\ell \ge 1$, $t_1,\dots,t_\ell \in [0,T]$, and $\psi_1,\dots,\psi_\ell \in \mathrm{BL}(\mathbb{R}^d)$. For each j, define the path functional

$$\varphi_j(\omega) \coloneqq \psi_j(\omega_{t_j}), \qquad \omega \in D([0,T];\mathbb{R}^d).$$

Since ψ_j is bounded, the integrability requirement for the class $\mathcal{A}$ of Budhiraja et al. (2017) is immediate. Moreover,

$$m_{\varphi_j}(V^0) = \mathbb{E}[\varphi_j(X)] = \mathbb{E}[\psi_j(X_{t_j})] = \langle p_{t_j},\psi_j\rangle.$$

Applying (Budhiraja et al., 2017, Thm. 2.4) to arbitrary linear combinations $\sum_{j=1}^\ell a_j\varphi_j$ and using the Cramér–Wold device yields the claimed convergence of

$$\big(\xi_{t_1}^{(n)}[\psi_1],\dots,\xi_{t_\ell}^{(n)}[\psi_\ell]\big)$$

to a centered Gaussian vector.

Finally, the Gaussian-mixture feature of Budhiraja et al. (2017) degenerates in our setting. Indeed, θ_0 is deterministic and $\bar{\sigma}_0 = d_0 = 0$, so the common-factor law is a Dirac mass at the deterministic limit path θ. Hence the limiting law is an ordinary centered Gaussian law, not a non-trivial Gaussian mixture. □

D.4.2 Proof of Proposition D.1: projected fluctuations

We now state the fixed-time projection result that identifies the Gaussian fluctuation suggested by Theorem 3.3. In the finite-observable case, no distribution-valued CLT is needed: one only projects the empirical fluctuation onto the finite list of observables seen by Γ.

For $q = \sum_{j=1}^{r} q_j$, identify the product observable space in Eq. (4) with $\mathbb{R}^q$ and write

$$\Phi^\theta(x) \coloneqq \big(\phi_1^\theta(x), \dots, \phi_r^\theta(x)\big) \in \mathbb{R}^q, \qquad m_\theta(p) \coloneqq \mathbb{E}_p[\Phi^\theta], \qquad m_t \coloneqq m_{\theta_t}(p_t).$$

Let Z_t denote the centered Gaussian limit of

$$Z_t^{(n)} \coloneqq \sqrt{n}\left(m_{\theta_t}(p_t^{(n)}) - m_{\theta_t}(p_t)\right) \in \mathbb{R}^q.$$

This limit exists by Proposition 3.2, applied componentwise to the bounded Lipschitz components of Φ^{θ_t}.

Definition D.1 (Projected fluctuation at fixed time)**.** *Let $t \in [0,T]$. The* projected fluctuation *of the Gibbs gradient is*

$$\Delta_t \coloneqq DH_{\theta_t}(m_t) Z_t \in \mathbb{R}^p,$$

where $DH_{\theta_t}(m_t) : \mathbb{R}^q \to \mathbb{R}^p$ is the finite-dimensional derivative of H_{θ_t} at m_t.

Proposition D.1 (Projected fluctuation CLT at fixed time)**.** *Under Assumption D.1, for each $t \in [0,T]$,*

$$\Delta_t^{(n)} \coloneqq \sqrt{n}\big(\Gamma(p_t^{(n)}, \theta_t) - \Gamma(p_t, \theta_t)\big) \Longrightarrow_{n\to\infty} \Delta_t.$$

Moreover, Δ_t is a centered Gaussian vector with covariance matrix $\Sigma_t \coloneqq \mathrm{Cov}(\Delta_t)$, and $\mathbb{E}\left[\|\Delta_t^{(n)}\|^2\right] \to \mathrm{Tr}(\Sigma_t)$ as $n \to \infty$.

Proof of Proposition D.1. Fix $t \in [0,T]$. By Assumption D.1(iii), the components of Φ^{θ_t} are bounded Lipschitz functions. Hence Proposition 3.2, applied to these finitely many components, gives

$$Z_t^{(n)} = \sqrt{n}\left(m_{\theta_t}(p_t^{(n)}) - m_{\theta_t}(p_t)\right) \Longrightarrow_{n\to\infty} Z_t$$

in $\mathbb{R}^q$, where Z_t is centered Gaussian. Since

$$\Gamma(p_t^{(n)}, \theta_t) = H_{\theta_t}\left(m_{\theta_t}(p_t^{(n)})\right), \qquad \Gamma(p_t, \theta_t) = H_{\theta_t}(m_t),$$

the ordinary finite-dimensional delta method applied to the C^1 map H_{θ_t} yields

$$\Delta_t^{(n)} \Longrightarrow_{n\to\infty} DH_{\theta_t}(m_t) Z_t = \Delta_t.$$

Since Δ_t is a linear image of the centered Gaussian random vector Z_t, it is itself centered Gaussian. We denote its covariance matrix by

$$\Sigma_t \coloneqq \mathrm{Cov}(\Delta_t).$$

To justify convergence of second moments, we first prove a uniform fourth-moment bound and then deduce uniform integrability. Couple X^i with independent copies $\bar{X}^i$ of the ideal flow in Eq. (2a), using the same initial values and Brownian motions as in Eq. (3a):

$$\mathrm{d}\bar{X}_t^i = -\nabla_1 V(\bar{X}_t^i, \theta_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^i, \qquad \bar{X}_0^i = X_0^i.$$

Since θ_t is deterministic, Assumption D.1(ii) and the i.i.d. initialization following Eq. (3) ensure that these copies are independent with law p_t. Throughout this proof and the comparison below, $C_T > 0$ denotes a deterministic constant that may change from line to line, depending only on T, the fixed dimensions, and the uniform bounds and Lipschitz constants in Assumption D.1(i)–(iii), and independent of n, $t \in [0,T]$, and i.

Fourth moments for the independent copies. For each scalar component f_t of Φ^{θ_t} or $\partial_\theta \Phi^{\theta_t}$, bounded by Assumption D.1(iii), put $Y_i = f_t(\bar{X}_t^i) - \langle f_t, p_t \rangle$. Since $\bar{X}_t^i \sim p_t$, $\mathbb{E}[Y_i] = \mathbb{E}\left[f_t(\bar{X}_t^i)\right] - \langle f_t, p_t \rangle = 0$. When expanding the fourth power of $\sum_i Y_i$, independence and centering eliminate every term containing an index that occurs only once. Only four equal indices or two pairs remain:

$$\begin{aligned} \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n Y_i\right|^4\right] &= \frac{1}{n^4}\left(\sum_{i=1}^n \mathbb{E}\left[Y_i^4\right] + 6\sum_{i<j} \mathbb{E}\left[Y_i^2\right]\mathbb{E}\left[Y_j^2\right]\right) \\ &= \frac{\mathbb{E}\left[Y_1^4\right]}{n^3} + \frac{3(n-1)}{n^3}(\mathbb{E}\left[Y_1^2\right])^2 \le \frac{C\|f_t\|_\infty^4}{n^2}. \end{aligned} \tag{25}$$

The last inequality follows directly from boundedness, without any moment assumption on $\bar{X}_t^i$:

$$\begin{aligned} |Y_i| \le |f_t(\bar{X}_t^i)| + |\langle f_t, p_t \rangle| \le 2\|f_t\|_\infty, \qquad \mathbb{E}\left[Y_i^4\right] \le 16\|f_t\|_\infty^4, \\ \mathbb{E}\left[Y_i^2\right] = \mathbb{E}\left[f_t(\bar{X}_t^i)^2\right] - \langle f_t, p_t \rangle^2 \le \|f_t\|_\infty^2. \end{aligned}$$

Here $C > 0$ is a universal constant, independent of f_t, n, and t. For $v = \frac{1}{n}\sum_{i=1}^n \Phi^{\theta_t}(\bar{X}_t^i) - m_t \in \mathbb{R}^q$, the j-th scalar coordinate is

$$v_j = \frac{1}{n}\sum_{i=1}^n \big(\Phi_j^{\theta_t}(\bar{X}_t^i) - (m_t)_j\big) = \frac{1}{n}\sum_{i=1}^n Y_i \qquad \text{with } f_t = \Phi_j^{\theta_t}.$$

Since $(m_t)_j = \langle \Phi_j^{\theta_t}, p_t \rangle$, Eq. (25) bounds $\mathbb{E}\left[|v_j|^4\right]$ by $C\|\Phi_j^{\theta_t}\|_\infty^4/n^2$. By Cauchy–Schwarz, $\|v\|^4 = (\sum_{j=1}^q |v_j|^2)^2 \le q\sum_{j=1}^q |v_j|^4$. Taking expectations and using the uniform bounds in Assumption D.1(iii) gives

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \Phi^{\theta_t}(\bar{X}_t^i) - m_t\right\|^4\right] \le \frac{C_T}{n^2}. \tag{26}$$

The same argument applies to $\partial_\theta \Phi^{\theta_t}$, viewed componentwise as a finite-dimensional matrix.

Coupling the particle and parameter errors. Subtracting Eqs. (2) and (3), with the shared initialization, cancels the Brownian terms and gives

$$\begin{aligned} X_t^i - \bar{X}_t^i &= -\int_0^t \big(\nabla_1 V(X_s^i, \theta_s^{(n)}) - \nabla_1 V(\bar{X}_s^i, \theta_s)\big)\,\mathrm{d}s, \\ \theta_t^{(n)} - \theta_t &= -\int_0^t \varepsilon_s\big(\Gamma(p_s^{(n)}, \theta_s^{(n)}) - \Gamma(p_s, \theta_s)\big)\,\mathrm{d}s. \end{aligned}$$

By Assumption D.1(ii), the norm of the particle drift difference is bounded by $C_T(\|X_s^i - \bar{X}_s^i\| + \|\theta_s^{(n)} - \theta_s\|)$.

The empirical moment in Eq. (4) is $m_{\theta_s^{(n)}}(p_s^{(n)}) = n^{-1}\sum_{i=1}^n \Phi^{\theta_s^{(n)}}(X_s^i)$ by the definition of $p_s^{(n)}$. Adding and subtracting $n^{-1}\sum_{i=1}^n \Phi^{\theta_s}(X_s^i)$ and $n^{-1}\sum_{i=1}^n \Phi^{\theta_s}(\bar{X}_s^i)$ gives

$$\begin{aligned} m_{\theta_s^{(n)}}(p_s^{(n)}) - m_s = &\frac{1}{n}\sum_{i=1}^n \big(\Phi^{\theta_s^{(n)}}(X_s^i) - \Phi^{\theta_s}(X_s^i)\big) \\ &+ \frac{1}{n}\sum_{i=1}^n \big(\Phi^{\theta_s}(X_s^i) - \Phi^{\theta_s}(\bar{X}_s^i)\big) \\ &+ \frac{1}{n}\sum_{i=1}^n \Phi^{\theta_s}(\bar{X}_s^i) - m_s. \end{aligned} \tag{27}$$

The bounded first θ-derivatives of Φ^θ in Assumption D.1(iii) control the first sum, and its uniform Lipschitz bound in x controls the second. Thus the triangle inequality in Eq. (27) gives

$$\|m_{\theta_s^{(n)}}(p_s^{(n)}) - m_s\| \le C_T\|\theta_s^{(n)} - \theta_s\| + \frac{C_T}{n}\sum_{i=1}^n \|X_s^i - \bar{X}_s^i\| + \left\|\frac{1}{n}\sum_{i=1}^n \Phi^{\theta_s}(\bar{X}_s^i) - m_s\right\|.$$

The bounded first derivatives of the joint map $(\theta, z) \mapsto H_\theta(z)$, on the segments specified in Assumption D.1(iii), give the first inequality below. Substituting the moment bound and absorbing both parameter-error contributions into C_T gives the second.

$$\begin{aligned}
&\|\Gamma(p_s^{(n)}, \theta_s^{(n)}) - \Gamma(p_s, \theta_s)\| \\
&\quad \le C_T \left(\|\theta_s^{(n)} - \theta_s\| + \|m_{\theta_s^{(n)}}(p_s^{(n)}) - m_s\| \right) \\
&\quad \le C_T \left(\|\theta_s^{(n)} - \theta_s\| + \frac{1}{n}\sum_{i=1}^n \|X_s^i - \bar{X}_s^i\| + \left\| \frac{1}{n}\sum_{i=1}^n \Phi^{\theta_s}(\bar{X}_s^i) - m_s \right\| \right).
\end{aligned} \tag{28}$$

To apply Gronwall's inequality, set

$$e_t^{(n)} \coloneqq \|\theta_t^{(n)} - \theta_t\| + \max_{i \le n} \|X_t^i - \bar{X}_t^i\|.$$

The differences $X_t^i - \bar{X}_t^i$ and $\theta_t^{(n)} - \theta_t$ are absolutely continuous in time, as are their norms and finite maximum. Applying $\frac{\mathrm{d}}{\mathrm{d}t}\|u_t\| \le \|\dot{u}_t\|$ almost everywhere to the subtracted parameter equations gives the gradient term below. By Assumption D.1(ii), each particle error norm has derivative at most $C_T e_t^{(n)}$, which also bounds the derivative of their finite maximum. Adding these bounds gives

$$\frac{\mathrm{d}e_t^{(n)}}{\mathrm{d}t} \le \varepsilon_t \|\Gamma(p_t^{(n)}, \theta_t^{(n)}) - \Gamma(p_t, \theta_t)\| + C_T e_t^{(n)} \qquad \text{a.e.}$$

Since $n^{-1}\sum_i \|X_t^i - \bar{X}_t^i\| \le \max_i \|X_t^i - \bar{X}_t^i\|$, Eq. (28) and the bound on ε_t in Assumption D.1(i) imply, almost everywhere,

$$\frac{\mathrm{d}e_t^{(n)}}{\mathrm{d}t} \le C_T e_t^{(n)} + C_T \left\| \frac{1}{n}\sum_{i=1}^n \Phi^{\theta_t}(\bar{X}_t^i) - m_t \right\|, \qquad e_0^{(n)} = 0.$$

Here $e_0^{(n)} = 0$ uses both initializations following Eq. (3). Applying Lemma C.1 and absorbing its exponential factor on $[0, T]$ into C_T yields

$$\begin{aligned}
&\sup_{t \le T} \|\theta_t^{(n)} - \theta_t\| + \max_{i \le n} \sup_{t \le T} \|X_t^i - \bar{X}_t^i\| \\
&\quad \le 2 \sup_{t \le T} e_t^{(n)} \le C_T \int_0^T \left\| \frac{1}{n}\sum_{i=1}^n \Phi^{\theta_s}(\bar{X}_s^i) - m_s \right\| \mathrm{d}s.
\end{aligned} \tag{29}$$

For each fixed i, Eq. (29) and $a^4 + b^4 \le (a+b)^4$ for $a, b \ge 0$ give

$$\sup_{t \le T} \|\theta_t^{(n)} - \theta_t\|^4 + \sup_{t \le T} \|X_t^i - \bar{X}_t^i\|^4 \le C_T \left(\int_0^T \left\| \frac{1}{n}\sum_{j=1}^n \Phi^{\theta_s}(\bar{X}_s^j) - m_s \right\| \mathrm{d}s \right)^4.$$

Taking expectations and then maximizing over i gives the left-hand side below, since the bound is the same for every i. Hölder's inequality gives $(\int_0^T a_s \,\mathrm{d}s)^4 \le T^3 \int_0^T a_s^4 \,\mathrm{d}s$ for nonnegative a_s. Using this inequality, Tonelli's theorem, and Eq. (26), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t \le T} \|\theta_t^{(n)} - \theta_t\|^4 \right] + \max_{i \le n} \mathbb{E}\left[\sup_{t \le T} \|X_t^i - \bar{X}_t^i\|^4 \right] \\
&\quad \le C_T T^3 \int_0^T \mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \Phi^{\theta_s}(\bar{X}_s^i) - m_s \right\|^4 \right] \mathrm{d}s \le \frac{C_T}{n^2}.
\end{aligned} \tag{30}$$

Fourth moments of the interacting empirical averages. For $f_t = \Phi^{\theta_t}$ or $f_t = \partial_\theta \Phi^{\theta_t}$, split

$$\sqrt{n}\langle f_t, p_t^{(n)} - p_t \rangle = \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f_t(X_t^i) - f_t(\bar{X}_t^i) \big) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f_t(\bar{X}_t^i) - \langle f_t, p_t \rangle \big).$$

The second sum has a bounded fourth moment by Eq. (25) applied componentwise. For the first sum, Jensen's inequality for the convex map $v \mapsto \|v\|^4$ does not require independence. Combining these facts with $\|a+b\|^4 \le 8(\|a\|^4 + \|b\|^4)$ gives

$$\begin{aligned}\mathbb{E}\left[\left\|\sqrt{n}\langle f_t, p_t^{(n)} - p_t\rangle\right\|^4\right] &\le \frac{8n^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|f_t(X_t^i) - f_t(\bar{X}_t^i)\|^4\right] + C_T \\ &\le 8n^2\|f_t\|_{\mathrm{L}}^4 \max_{i\le n}\mathbb{E}\left[\sup_{s\le T}\|X_s^i - \bar{X}_s^i\|^4\right] + C_T \\ &\le C_T, \end{aligned} \tag{31}$$

where Assumption D.1(iii) supplies the uniform Lipschitz bound, with the seminorm convention in Section A, and the last step uses Eq. (30).

Uniform integrability and second moments. Applying Eq. (31) with $f_t = \Phi^{\theta_t}$ bounds the fourth moment of $\sqrt{n}\langle \Phi^{\theta_t}, p_t^{(n)} - p_t\rangle = \sqrt{n}(m_{\theta_t}(p_t^{(n)}) - m_t) = Z_t^{(n)}$. Combining this bound with the uniform Lipschitz constant L_H from Assumption D.1(iii), using Eq. (4) and Remark C.1, gives

$$\|\Delta_t^{(n)}\| \le L_H\|Z_t^{(n)}\|, \qquad \sup_{n\ge 1,\, t\le T}\mathbb{E}\left[\|\Delta_t^{(n)}\|^4\right] \le C_T.$$

Consequently, for $M > 0$,

$$\sup_{n\ge 1,\, t\le T}\mathbb{E}\left[\|\Delta_t^{(n)}\|^2 \mathbf{1}_{\{\|\Delta_t^{(n)}\|^2 > M\}}\right] \le \frac{1}{M}\sup_{n\ge 1,\, t\le T}\mathbb{E}\left[\|\Delta_t^{(n)}\|^4\right] \le \frac{C_T}{M} \longrightarrow 0.$$

This is uniform integrability of the squared norms. Weak convergence gives convergence of the expectations of the bounded continuous truncations $\min(\|x\|^2, M)$, and the preceding tail bound allows $M \to \infty$ (van der Vaart and Wellner, 1996, Thm. 1.11.3(i)). Hence

$$\mathbb{E}\left[\|\Delta_t^{(n)}\|^2\right] \longrightarrow \mathbb{E}\left[\|\Delta_t\|^2\right] = \sum_{j=1}^{p}(\Sigma_t)_{jj} = \mathrm{Tr}(\Sigma_t),$$

where the equality uses $\mathbb{E}[\Delta_t] = 0$. Finally, Cauchy–Schwarz and Assumption D.1(i) give the integrable bound

$$0 \le \varepsilon_t\mathbb{E}\left[\|\Delta_t^{(n)}\|^2\right] \le \|\varepsilon\|_\infty\big(\mathbb{E}\left[\|\Delta_t^{(n)}\|^4\right]\big)^{1/2} \le C_T,$$

where $\|\varepsilon\|_\infty = \sup_{0\le s\le T}\varepsilon_s$. Dominated convergence on $[0,T]$ therefore yields

$$\int_0^T \varepsilon_t\mathbb{E}\left[\|\Delta_t^{(n)}\|^2\right]\,\mathrm{d}t \longrightarrow \int_0^T \varepsilon_t\,\mathrm{Tr}(\Sigma_t)\,\mathrm{d}t. \tag{32}$$

□

D.4.3 Proof of Theorem 3.3: asymptotic convergence

The projected CLT in Proposition D.1 identifies the observable-level Gaussian covariance obtained at the deterministic limit (p_t, θ_t). The finite-particle entropy estimate, however, contains the moving-reference fluctuation

$$\Lambda_t^{(n)} := \sqrt{n}\big(\Gamma(p_t^{(n)}, \theta_t^{(n)}) - \Gamma(q_{t,n}, \theta_t^{(n)})\big).$$

This is the actual random quantity that appears in Eq. (19). To connect it with the deterministic-limit fluctuation $\Delta_t^{(n)}$ from Proposition D.1, write

$$\Lambda_t^{(n)} = \underbrace{\Delta_t^{(n)}}_{\text{projected CLT fluctuation}} + \underbrace{\big(\Lambda_t^{(n)} - \Delta_t^{(n)}\big)}_{\text{moving-reference correction}},$$

where the correction accounts for replacing the deterministic pair (p_t, θ_t) in $\Delta_t^{(n)}$ by the moving pair $(q_{t,n}, \theta_t^{(n)})$ in $\Lambda_t^{(n)}$.

We define the asymptotic time-averaged variance level of the moving fluctuation by

$$\sigma_{\mathrm{mov}}^2(T) := \limsup_{n\to\infty} S_n(T), \qquad S_n(T) := \frac{1}{T^{1-\delta}}\int_0^T \varepsilon_t\mathbb{E}[\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}^2]\,\mathrm{d}t. \tag{33}$$

This quantity is finite under the hypotheses of Theorem 3.2. Indeed, the uniform empirical-process estimate used in the proof of Theorem 3.2 gives $\mathbb{E}[\|\Lambda_t^{(n)}\|_{\mathbb{R}^p}^2] \le D$ for a constant D independent of n and t. Hence $S_n(T) \le DT^{-(1-\delta)} \int_0^T \varepsilon_t \, \mathrm{d}t < \infty$ uniformly in n. In the finite-observable case, writing

$$m_t^{(n)} \coloneqq m_{\theta_t^{(n)}}(p_t^{(n)}), \qquad \bar{m}_t^{(n)} \coloneqq m_{\theta_t^{(n)}}(q_{t,n}),$$

gives

$$\Lambda_t^{(n)} = \sqrt{n}\big(H_{\theta_t^{(n)}}(m_t^{(n)}) - H_{\theta_t^{(n)}}(\bar{m}_t^{(n)})\big).$$

The preceding decomposition yields the elementary comparison

$$S_n(T) \le \frac{2}{T^{1-\delta}} \int_0^T \varepsilon_t \mathbb{E}[\|\Delta_t^{(n)}\|_{\mathbb{R}^p}^2] \, \mathrm{d}t + 2\mathcal{R}_n(T),$$

where

$$\mathcal{R}_n(T) \coloneqq \frac{1}{T^{1-\delta}} \int_0^T \varepsilon_t \mathbb{E}[\|\Lambda_t^{(n)} - \Delta_t^{(n)}\|_{\mathbb{R}^p}^2] \, \mathrm{d}t.$$

Hence the first term in $S_n(T)$ is the CLT fluctuation contribution, whose deterministic-limit covariance is Σ_t, while $\mathcal{R}_n(T)$ collects the moving-reference correction. In particular, at the CLT level

$$\frac{1}{T^{1-\delta}} \int_0^T \varepsilon_t \, \mathbb{E}[\|\Delta_t\|_{\mathbb{R}^p}^2] \, \mathrm{d}t = \frac{1}{T^{1-\delta}} \int_0^T \varepsilon_t \, \mathrm{Tr}(\Sigma_t) \, \mathrm{d}t.$$

Thus $\sigma_{\mathrm{mov}}^2(T)$ should be read as the moving finite-particle variance entering the theorem: it contains the projected CLT covariance contribution identified by Σ_t, plus the residual caused by replacing the deterministic pair (p_t, θ_t) in $\Delta_t^{(n)}$ by the moving pair $(q_{t,n}, \theta_t^{(n)})$ in $\Lambda_t^{(n)}$. We now show that $\mathcal{R}_n(T) \to 0$ on the fixed horizon $[0, T]$, which will identify the limit of $S_n(T)$. We use the convention for C_T from the proof of Proposition D.1. Let $\mathbb{P}$ be the probability measure underlying the full particle system and $\mathbb{E}$ the corresponding expectation. The law of each X_t^i under $\mathbb{P}$ is $q_{t,n}$, as defined after Eq. (3). For random vectors or matrices, set $\|U\|_{L^r(\mathbb{P})} \coloneqq (\mathbb{E}\,[\|U\|^r])^{1/r}$ for $r = 2, 4$.

Moment bounds for the two reference measures. For each deterministic bounded observable f_t, linearity of expectation gives

$$\mathbb{E}\left[\langle f_t, p_t^{(n)} \rangle\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[f_t(X_t^i)\right] = \langle f_t, q_{t,n} \rangle.$$

For $f_t = \Phi^{\theta_t}$, the marginal identity gives the first equality below by taking expectations in the definition of $Z_t^{(n)}$. The second follows by subtracting this mean from $Z_t^{(n)}$.

$$m_{\theta_t}(q_{t,n}) = m_t + n^{-1/2} \mathbb{E}\left[Z_t^{(n)}\right], \qquad \sqrt{n}\langle \Phi^{\theta_t}, p_t^{(n)} - q_{t,n} \rangle = Z_t^{(n)} - \mathbb{E}\left[Z_t^{(n)}\right]. \tag{34}$$

Set $h = \theta_t^{(n)} - \theta_t$ and let ν denote either $p_t^{(n)}$ or $q_{t,n}$. For $f_t = \Phi^{\theta_t}$ or $f_t = \partial_\theta \Phi^{\theta_t}$, Eq. (30) and Eq. (31) (after dividing the latter by $(\sqrt{n})^4 = n^2$) give

$$\mathbb{E}\left[\|h\|^4\right] \le \mathbb{E}\left[\sup_{s \le T} \|\theta_s^{(n)} - \theta_s\|^4\right] \le \frac{C_T}{n^2}, \qquad \mathbb{E}\left[\|\langle f_t, p_t^{(n)} - p_t \rangle\|^4\right] \le \frac{C_T}{n^2}.$$

For $\nu = q_{t,n}$, the marginal identity above and Jensen's inequality give

$$\|\langle f_t, q_{t,n} - p_t \rangle\|^4 = \|\mathbb{E}\left[\langle f_t, p_t^{(n)} - p_t \rangle\right]\|^4 \le \mathbb{E}\left[\|\langle f_t, p_t^{(n)} - p_t \rangle\|^4\right] \le \frac{C_T}{n^2}.$$

Taking fourth roots and using that the marginal term is deterministic gives

$$\|h\|_{L^4(\mathbb{P})} \le \frac{C_T}{\sqrt{n}}, \qquad \|\langle f_t, \nu - p_t \rangle\|_{L^4(\mathbb{P})} \le \frac{C_T}{\sqrt{n}}, \qquad f_t = \Phi^{\theta_t}, \partial_\theta \Phi^{\theta_t}. \tag{35}$$

These bounds are uniform in $t \in [0, T]$. Inserting $m_{\theta_t}(\nu)$ separates the change in parameter from the change in measure and gives

$$m_{\theta_t + h}(\nu) - m_t = \langle \Phi^{\theta_t + h} - \Phi^{\theta_t}, \nu \rangle + \langle \Phi^{\theta_t}, \nu - p_t \rangle,$$
$$\|m_{\theta_t + h}(\nu) - m_t\| \le C_T \|h\| + \|\langle \Phi^{\theta_t}, \nu - p_t \rangle\|.$$

The first term on the right is bounded by $C_T\|h\|$ using the first θ-derivatives in Assumption D.1(iii) and the fact that ν has mass one. Taking $L^4(\mathbb{P})$ norms and using Eq. (35) bounds the differences from m_t for $\nu = p_t^{(n)}$ and $\nu = q_{t,n}$, respectively. The triangle inequality then bounds their difference, so

$$\|m_t^{(n)} - m_t\|_{L^4(\mathbb{P})} + \|\bar{m}_t^{(n)} - m_t\|_{L^4(\mathbb{P})} + \|m_t^{(n)} - \bar{m}_t^{(n)}\|_{L^4(\mathbb{P})} \le \frac{C_T}{\sqrt{n}}. \tag{36}$$

We will use Cauchy–Schwarz to control products, in the form

$$\left(\mathbb{E}\left[\|U\|^2\|V\|^2\right]\right)^{1/2} \le \left(\mathbb{E}\left[\|U\|^4\right]\right)^{1/4}\left(\mathbb{E}\left[\|V\|^4\right]\right)^{1/4} = \|U\|_{L^4(\mathbb{P})}\|V\|_{L^4(\mathbb{P})}.$$

Changing the parameter in the centered moments. At the $\sqrt{n}$ scale, Eq. (35) only bounds $\sqrt{n}\|h\|_{L^4(\mathbb{P})}$ by a constant. To gain a small factor, expand Φ against $p_t^{(n)} - q_{t,n}$ so that the coefficient of h is itself a difference of averages. For each x, integrate the derivative of $a \mapsto \Phi^{\theta_t + ah}(x)$ from 0 to 1 and separate the derivative at $a = 0$. Integrating against the measure difference gives

$$\begin{aligned}
&\langle \Phi^{\theta_t + h} - \Phi^{\theta_t}, p_t^{(n)} - q_{t,n}\rangle \\
&\quad = \langle \partial_\theta \Phi^{\theta_t}, p_t^{(n)} - q_{t,n}\rangle h \\
&\qquad + \int_0^1 \langle \partial_\theta \Phi^{\theta_t + ah} - \partial_\theta \Phi^{\theta_t}, p_t^{(n)} - q_{t,n}\rangle h \,\mathrm{d}a.
\end{aligned} \tag{37}$$

By Assumption D.1(iii), the second θ-derivatives bound the change in the first derivative by $C_T a\|h\|$ uniformly in x. Each probability measure has mass one, so integrating this change against their difference costs at most $2C_T a\|h\|$. Multiplication by h and integration over a therefore bound the remainder in Eq. (37) by $\int_0^1 2C_T a\|h\|^2 \,\mathrm{d}a = C_T\|h\|^2$. For the coefficient of the linear term, insert p_t between the two measures. Applying Eq. (35) with $f_t = \partial_\theta \Phi^{\theta_t}$ gives

$$\begin{aligned}
&\|\langle \partial_\theta \Phi^{\theta_t}, p_t^{(n)} - q_{t,n}\rangle\|_{L^4(\mathbb{P})} \\
&\quad \le \|\langle \partial_\theta \Phi^{\theta_t}, p_t^{(n)} - p_t\rangle\|_{L^4(\mathbb{P})} + \|\langle \partial_\theta \Phi^{\theta_t}, q_{t,n} - p_t\rangle\|_{L^4(\mathbb{P})} \le \frac{C_T}{\sqrt{n}}.
\end{aligned}$$

After multiplication by $\sqrt{n}$, the two terms on the left of Eq. (37) are $\sqrt{n}(m_t^{(n)} - \bar{m}_t^{(n)})$ and $Z_t^{(n)} - \mathbb{E}\left[Z_t^{(n)}\right]$, by Eq. (34). Taking $L^2(\mathbb{P})$ norms, Cauchy–Schwarz bounds the product in the linear term, and $\left\|\|h\|^2\right\|_{L^2(\mathbb{P})} = \|h\|_{L^4(\mathbb{P})}^2$ bounds the quadratic remainder. Thus

$$\begin{aligned}
&\left\|\sqrt{n}(m_t^{(n)} - \bar{m}_t^{(n)}) - Z_t^{(n)} + \mathbb{E}\left[Z_t^{(n)}\right]\right\|_{L^2(\mathbb{P})} \\
&\quad \le \sqrt{n}\left(\|\langle \partial_\theta \Phi^{\theta_t}, p_t^{(n)} - q_{t,n}\rangle\|_{L^4(\mathbb{P})}\|h\|_{L^4(\mathbb{P})} + C_T\|h\|_{L^4(\mathbb{P})}^2\right) \le \frac{C_T}{\sqrt{n}}.
\end{aligned} \tag{38}$$

Both terms in parentheses are bounded by C_T/n before multiplication by $\sqrt{n}$.

Passing the centered moments through H. The two values of H in $\Lambda_t^{(n)}$ have the same parameter $\theta_t^{(n)}$. Differentiating along the moment segment from $\bar{m}_t^{(n)}$ to $m_t^{(n)}$ gives $DH_{\theta_t^{(n)}}$ applied to $m_t^{(n)} - \bar{m}_t^{(n)}$. Integrating this derivative gives

$$\Lambda_t^{(n)} = \sqrt{n}\int_0^1 DH_{\theta_t^{(n)}}\left(\bar{m}_t^{(n)} + a(m_t^{(n)} - \bar{m}_t^{(n)})\right)(m_t^{(n)} - \bar{m}_t^{(n)})\,\mathrm{d}a. \tag{39}$$

To replace the derivative in the integral by $DH_{\theta_t}(m_t)$, first move to $DH_{\theta_t^{(n)}}(\bar{m}_t^{(n)})$, at a cost of at most $C_T a\|m_t^{(n)} - \bar{m}_t^{(n)}\|$. Then move along the joint segment from $(\theta_t^{(n)}, \bar{m}_t^{(n)})$ to (θ_t, m_t), at a cost of at most $C_T(\|h\| + \|\bar{m}_t^{(n)} - m_t\|)$. Both bounds use the second derivatives along these segments in Assumption D.1(iii). Adding them gives, uniformly in $0 \le a \le 1$,

$$\begin{aligned}
&\left\|DH_{\theta_t^{(n)}}\left(\bar{m}_t^{(n)} + a(m_t^{(n)} - \bar{m}_t^{(n)})\right) - DH_{\theta_t}(m_t)\right\| \\
&\quad \le C_T\left(\|m_t^{(n)} - \bar{m}_t^{(n)}\| + \|h\| + \|\bar{m}_t^{(n)} - m_t\|\right).
\end{aligned}$$

Subtracting the term with $DH_{\theta_t}(m_t)$ in Eq. (39) leaves the difference of derivatives inside the integral, multiplied by the moment increment. Since the interval of integration has length one, the pathwise error is at most

$$C_T\sqrt{n}\big(\|m_t^{(n)} - \bar{m}_t^{(n)}\| + \|h\| + \|\bar{m}_t^{(n)} - m_t\|\big)\|m_t^{(n)} - \bar{m}_t^{(n)}\|.$$

Cauchy–Schwarz bounds the $L^2(\mathbb{P})$ norm of this product by the two $L^4(\mathbb{P})$ norms, each of order $n^{-1/2}$ by Eqs. (35) and (36). Therefore

$$\begin{aligned}
&\Big\|\Lambda_t^{(n)} - DH_{\theta_t}(m_t)\sqrt{n}(m_t^{(n)} - \bar{m}_t^{(n)})\Big\|_{L^2(\mathbb{P})} \\
&\quad\le C_T\sqrt{n}\left(\|m_t^{(n)} - \bar{m}_t^{(n)}\|_{L^4(\mathbb{P})} + \|h\|_{L^4(\mathbb{P})} + \|\bar{m}_t^{(n)} - m_t\|_{L^4(\mathbb{P})}\right)\|m_t^{(n)} - \bar{m}_t^{(n)}\|_{L^4(\mathbb{P})} \quad (40)\\
&\quad\le \frac{C_T}{\sqrt{n}}.
\end{aligned}$$

For $\Delta_t^{(n)}$, the moment increment at the fixed parameter θ_t is $m_{\theta_t}(p_t^{(n)}) - m_t = n^{-1/2}Z_t^{(n)}$. Integrating along this segment and subtracting the constant derivative gives

$$\begin{aligned}
&\Delta_t^{(n)} - DH_{\theta_t}(m_t)Z_t^{(n)} \\
&\quad= \int_0^1 \big(DH_{\theta_t}(m_t + an^{-1/2}Z_t^{(n)}) - DH_{\theta_t}(m_t)\big)Z_t^{(n)}\,\mathrm{d}a.
\end{aligned}$$

The derivative difference has norm at most $C_T a\|m_{\theta_t}(p_t^{(n)}) - m_t\|$ by Assumption D.1(iii). Multiplying by $\|Z_t^{(n)}\|$ thus gives a quadratic bound in the moment increment. Taking $L^2(\mathbb{P})$ norms and using Eq. (35) yields

$$\|\Delta_t^{(n)} - DH_{\theta_t}(m_t)Z_t^{(n)}\|_{L^2(\mathbb{P})} \le C_T\sqrt{n}\|m_{\theta_t}(p_t^{(n)}) - m_t\|_{L^4(\mathbb{P})}^2 \le \frac{C_T}{\sqrt{n}}.$$

Inserting the two linear approximations gives the exact decomposition

$$\begin{aligned}
&\Lambda_t^{(n)} - \Delta_t^{(n)} + DH_{\theta_t}(m_t)\mathbb{E}\left[Z_t^{(n)}\right] \\
&\quad= \big(\Lambda_t^{(n)} - DH_{\theta_t}(m_t)\sqrt{n}(m_t^{(n)} - \bar{m}_t^{(n)})\big) \\
&\qquad+ DH_{\theta_t}(m_t)\big(\sqrt{n}(m_t^{(n)} - \bar{m}_t^{(n)}) - Z_t^{(n)} + \mathbb{E}\left[Z_t^{(n)}\right]\big) \\
&\qquad- \big(\Delta_t^{(n)} - DH_{\theta_t}(m_t)Z_t^{(n)}\big).
\end{aligned}$$

The first term is bounded in $L^2(\mathbb{P})$ by Eq. (40), the second by Eq. (38) and the bound on DH in Assumption D.1(iii), and the third by the fixed-parameter estimate above. Each bound is $C_T/\sqrt{n}$ uniformly in t. The triangle inequality followed by squaring gives

$$\sup_{t\le T}\mathbb{E}\left[\left\|\Lambda_t^{(n)} - \Delta_t^{(n)} + DH_{\theta_t}(m_t)\mathbb{E}\left[Z_t^{(n)}\right]\right\|^2\right] \le \frac{C_T}{n}. \quad (41)$$

Vanishing of the moving-reference correction. It remains to control the centering term $\mathbb{E}\left[Z_t^{(n)}\right]$. Applying Eq. (31) with $f_t = \Phi^{\theta_t}$ and using Cauchy–Schwarz gives $\sup_{n,t\le T}\mathbb{E}\left[\|Z_t^{(n)}\|^2\right] \le C_T$. On $\{\|Z_t^{(n)}\| > M\}$ we have $\|Z_t^{(n)}\| \le \|Z_t^{(n)}\|^2/M$, so

$$\sup_{n,t\le T}\mathbb{E}\left[\|Z_t^{(n)}\|\mathbf{1}_{\{\|Z_t^{(n)}\|>M\}}\right] \le \frac{C_T}{M} \longrightarrow 0.$$

Thus each coordinate of $Z_t^{(n)}$ is uniformly integrable. Together with the centered Gaussian limit from Proposition 3.2, convergence of expectations for each coordinate (van der Vaart and Wellner, 1996, Thm. 1.11.3(i)) and Jensen's inequality give, for each fixed t,

$$\mathbb{E}\left[Z_t^{(n)}\right] \longrightarrow \mathbb{E}\left[Z_t\right] = 0, \qquad \|\mathbb{E}\left[Z_t^{(n)}\right]\|^2 \le \mathbb{E}\left[\|Z_t^{(n)}\|^2\right] \le C_T.$$

Subtract $DH_{\theta_t}(m_t)\mathbb{E}\left[Z_t^{(n)}\right]$ from the vector in Eq. (41). Using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and the bound on DH in Assumption D.1(iii) gives

$$\mathbb{E}\left[\|\Lambda_t^{(n)} - \Delta_t^{(n)}\|^2\right] \le \frac{C_T}{n} + C_T\|\mathbb{E}\left[Z_t^{(n)}\right]\|^2.$$

Integrating this inequality in the definition of $\mathcal{R}_n(T)$ gives

$$\begin{aligned}\mathcal{R}_n(T) &\le \frac{C_T}{nT^{1-\delta}}\int_0^T \varepsilon_t \,\mathrm{d}t + \frac{C_T}{T^{1-\delta}}\int_0^T \varepsilon_t\|\mathbb{E}\left[Z_t^{(n)}\right]\|^2 \,\mathrm{d}t \\ &\le \frac{C_T}{n} + \frac{C_T}{T^{1-\delta}}\int_0^T \varepsilon_t\|\mathbb{E}\left[Z_t^{(n)}\right]\|^2 \,\mathrm{d}t \longrightarrow 0.\end{aligned} \tag{42}$$

Here T is fixed, and Assumption D.1(i) bounds ε_t. The second integrand tends to zero for every t and is bounded by the integrable constant $\|\varepsilon\|_\infty C_T$, so dominated convergence justifies the last limit. In L^2 over $\mathbb{P}$ and time with weight $\varepsilon_t/T^{1-\delta}$, $\sqrt{S_n(T)}$ is the norm of $\Lambda^{(n)}$ and $\sqrt{\mathcal{R}_n(T)}$ is the norm of $\Lambda^{(n)} - \Delta^{(n)}$. The reverse triangle inequality and Eq. (42) give

$$\left|\sqrt{S_n(T)} - \left(\frac{1}{T^{1-\delta}}\int_0^T \varepsilon_t\mathbb{E}\left[\|\Delta_t^{(n)}\|^2\right] \,\mathrm{d}t\right)^{1/2}\right| \le \sqrt{\mathcal{R}_n(T)} \longrightarrow 0.$$

By Eq. (32), the two square roots therefore have the same limit. Squaring shows that the limsup in Eq. (33) is a limit, and

$$\lim_{n\to\infty} S_n(T) = \sigma^2_{\mathrm{mov}}(T) = \frac{1}{T^{1-\delta}}\int_0^T \varepsilon_t \,\mathrm{Tr}(\Sigma_t)\,\mathrm{d}t.$$

Proof of Theorem 3.3. Under the hypotheses of Theorem 3.2, the proof of Theorem 3.2 gives exactly the same differential inequalities as Eq. (19), keeping the fluctuation term $\Lambda_t^{(n)}$ instead of replacing it by the crude non-asymptotic bound from Eq. (20). Integrating those inequalities as in the proof of Theorem 3.2 gives

$$\begin{aligned}\frac{1}{T}\int_0^T \mathbb{E}[\|\nabla\ell(\theta_t^{(n)})\|^2]\,\mathrm{d}t \le \frac{6\mathbb{E}[y^{(n)}(0) + \ell(\theta_0^{(n)})]}{T^{1-\delta}} + \frac{6}{T^{1-\delta}}\int_0^T (-2\mu + \varepsilon_t A)\mathbb{E}[y^{(n)}(t)]\,\mathrm{d}t \\ + \frac{66}{10nT^{1-\delta}}\int_0^T \varepsilon_t\mathbb{E}\left[\|\Lambda_t^{(n)}\|^2_{\mathbb{R}^p}\right]\,\mathrm{d}t.\end{aligned}$$

The same finite-horizon entropy estimate used in Theorem 3.2 controls the first two terms on the right by $C/T^{1-\delta}$, up to a residual term of order $n^{-1}T^{-(1-\delta)}$, which is absorbed into $C/T^{1-\delta}$ since $n \ge 1$. Therefore,

$$\frac{1}{T}\int_0^T \mathbb{E}[\|\nabla\ell(\theta_t^{(n)})\|^2]\,\mathrm{d}t \le \frac{C}{T^{1-\delta}} + \frac{66}{10n}S_n(T).$$

The preceding uniform bound implies $\sigma^2_{\mathrm{mov}}(T) < \infty$. Since $\sigma^2_{\mathrm{mov}}(T) = \limsup_{n\to\infty} S_n(T)$, we may choose, for instance,

$$r_n(T) \coloneqq \sup_{k\ge n} S_k(T) - \sigma^2_{\mathrm{mov}}(T),$$

after increasing finitely many initial values if needed. Then $r_n(T) \ge 0$, $r_n(T) \to 0$, and $S_n(T) \le \sigma^2_{\mathrm{mov}}(T) + r_n(T)$ for all n. Setting

$$u_n(T) \coloneqq \frac{66}{10n}r_n(T)$$

gives $nu_n(T) \to 0$ and yields the announced bound. □

The projected fluctuation in Proposition D.1 should therefore be read as the CLT-level component of the variance mechanism entering $\sigma^2_{\mathrm{mov}}(T)$. In particular, the relevant finite-dimensional object is the empirical observable vector before it is passed through H_θ. This is precisely the object modified

in Section 4: control variates replace Φ^θ by a corrected observable with smaller covariance, and thus target the same observable-level variance mechanism.

Comparison with stochastic gradient descent. It is instructive to compare the variance behavior of stochastic gradient descent (SGD) with finite-particle Gibbs gradient descent. In SGD, the stochastic gradient at each step has variance of order $\mathcal{O}(\sigma^2/N)$, where σ^2 is the variance of the per-sample gradient noise and N is the batch size (Garrigos and Gower, 2023; Bottou et al., 2018). However, new minibatches are resampled at each step so that the noise is refreshed rather than being carried along the trajectory. With a decaying stepsize, the effect of the noise can be averaged out over time, resulting in the classical convergence rate $\mathcal{O}(T^{-1/2})$. The stochasticity in SGD appears as a *per-iteration noise* term that decreases with the number of iterations when stepsizes are decayed appropriately. In GGD with n particles, the gradient estimate $\Gamma(p_t^{(n)}, \theta_t)$ is subject to fluctuations of order $\mathcal{O}(n^{-1/2})$ (see Proposition 3.2). Unlike SGD, these fluctuations are not re-sampled at each iteration: they stem from the fixed empirical approximation of p_t by its empirical law $p_t^{(n)}$ and persist along the trajectory. As a result, even with decaying stepsizes, the averaged stationary error admits an irreducible plateau of order $\mathcal{O}(n^{-1})$.

E Proof of Proposition 4.1: variance reduction for CV-GGD

We prove the control-variate statements used in Section 4.

Proof of Proposition 4.1. Fix θ and write

$$C_{\Phi\Phi} = \operatorname*{Cov}_{\pi(\theta)}(\Phi^\theta, \Phi^\theta), \quad C_{\Psi\Psi} = \operatorname*{Cov}_{\pi(\theta)}(\Psi^\theta, \Psi^\theta), \quad C_{\Phi\Psi} = \operatorname*{Cov}_{\pi(\theta)}(\Phi^\theta, \Psi^\theta).$$

Then $C_{\Phi\Phi} \in \mathbb{R}^{q\times q}$, $C_{\Psi\Psi} \in \mathbb{R}^{d\times d}$, and $C_{\Phi\Psi} \in \mathbb{R}^{q\times d}$; we also write $C_{\Psi\Phi} = C_{\Phi\Psi}^\intercal$. The standard linear control-variate calculation gives

$$\operatorname{Tr}\left(\operatorname*{Cov}_{\pi(\theta)}(\Phi^\theta - B\Psi^\theta)\right) = \operatorname{Tr}(C_{\Phi\Phi}) - 2\operatorname{Tr}(BC_{\Psi\Phi}) + \operatorname{Tr}(BC_{\Psi\Psi}B^\intercal).$$

Let $F(B) = \operatorname{Tr}\big(\operatorname{Cov}_{\pi(\theta)}(\Phi^\theta - B\Psi^\theta)\big)$ denote this covariance trace. The minimization of $F(B)$ is the standard multiresponse control-variate problem. In (Sivula, 2015, App. B), the vector estimator $f - C^\intercal(h - \mathbb{E}[h])$ is minimized by the matrix regression coefficient $C^\star = \mathbb{V}(h)^{-1}\operatorname{Cov}(h, f)$. Rewriting this result in our notation, with $f = \Phi^\theta$, $h = \Psi^\theta$, and $\mathbb{E}[h] = 0$, gives

$$B_\theta^{\pi(\theta)} = C_{\Phi\Psi}C_{\Psi\Psi}^\dagger,$$

where $\dagger$ denotes the Moore–Penrose inverse, equal to the ordinary inverse when $C_{\Psi\Psi}$ is nonsingular. Since $B = 0$ is an admissible choice in the same minimization, optimality gives $F(B_\theta^{\pi(\theta)}) \leq F(0) = \operatorname{Tr}(C_{\Phi\Phi})$, so the covariance trace is no larger than without the control variate. In fact, $C_{\Phi\Psi}$ vanishes on the kernel of $C_{\Psi\Psi}$, since any zero-variance linear combination of Ψ^θ has zero covariance with Φ^θ. Thus the same calculation gives

$$\operatorname*{Cov}_{\pi(\theta)}(\Phi^\theta - B_\theta^{\pi(\theta)}\Psi^\theta) = C_{\Phi\Phi} - C_{\Phi\Psi}C_{\Psi\Psi}^\dagger C_{\Psi\Phi} \preceq C_{\Phi\Phi},$$

where $\preceq$ denotes the positive semidefinite order. This inequality is preserved under multiplication by a fixed matrix and its transpose, which justifies the update-covariance reduction in the remark following Proposition 4.1. □

Finally, the Stein controls used in Section 4 are zero-mean by a direct integration by parts. Let $\pi(\theta)(dx) = Z(\theta)^{-1}\exp(-V(x,\theta))\,\mathrm{d}x$. For any vector field $h \in C_b^1(\mathbb{R}^d;\mathbb{R}^d)$ for which the usual Gibbs integration by parts is valid, we have

$$\int_{\mathbb{R}^d} \nabla\cdot h(x)\,\pi(\theta)(\mathrm{d}x) = \int_{\mathbb{R}^d} h(x)^\intercal \nabla_1 V(x,\theta)\,\pi(\theta)(\mathrm{d}x),$$

which gives $\mathbb{E}_{\pi(\theta)}[\nabla\cdot h - h^\intercal\nabla_1 V(\cdot,\theta)] = 0$. In particular, this applies to the constant fields $h_j = -e_j$ used to obtain the score controls $\Psi^\theta(x) = \nabla_1 V(x,\theta)$.

F Details on numerical experiments

F.1 Finite-observable lower-bound witness

The lower-bound experiment in Section 5 uses $d = p = 1$ with

$$a(x) = \tanh(x), \quad \alpha(\theta) = \tanh(\theta),$$

and Gibbs potential $V(x, \theta) = \frac{x^2}{2} - \alpha(\theta)a(x)$. The loss is $\ell(\theta) = \rho_\tau(m(\theta))$, where $m(\theta) = \mathbb{E}_{\pi(\theta)}[a]$, $\tau = 0.8$, and

$$\rho_\tau(z) = \frac{1}{2}\big(\tau(z_+)^2 + (1-\tau)(z_-)^2\big).$$

The finite-particle GGD estimator is

$$\Gamma(p, \theta) = \operatorname{sech}^2(\theta)\rho'_\tau(\mathbb{E}_p[a])\big(\mathbb{E}_p[a^2] - \mathbb{E}_p[a]^2\big),$$

so the model has the finite-observable form with observables $\phi_1(x) = \tanh(x)$ and $\phi_2(x) = \tanh^2(x)$.

For the reported metric, we evaluate the population gradient rather than the empirical update. In this one-dimensional example, this is done by quadrature: for each θ, we normalize the density proportional to $\exp[-V(x, \theta)]$, compute

$$m(\theta) = \mathbb{E}_{\pi(\theta)}[a], \quad v(\theta) = \mathbb{E}_{\pi(\theta)}[a^2] - \mathbb{E}_{\pi(\theta)}[a]^2,$$

and use

$$\nabla\ell(\theta) = \operatorname{sech}^2(\theta)\rho'_\tau(m(\theta))v(\theta).$$

The implementation precomputes this quantity on a θ-grid and then uses linear interpolation along the simulated trajectories. The simulations use Euler–Maruyama for the particles:

$$X^i_{k+1} = X^i_k - h\,\partial_x V(X^i_k, \theta_k) + \sqrt{2h}\,\xi^i_{k+1},$$

with $h = 0.2$, constant outer update $\eta = 0.2$, initialization $\theta_0 = 0.2$, horizon $K = 10^6$, and

$$n \in \{10, 20, 50, 100, 200, 500, 1000\}.$$

The figures report Monte Carlo averages over four independent seeds. The late-tail constants are computed over the last 10^5 iterations, and the collapse plots start at $K = 10^3$.

For Ridge CV-GGD, we correct the empirical observable vector before applying the nonlinear map H_θ. Writing $\Phi(x) = (\tanh(x), \tanh^2(x))$, we use the scalar score control $\psi_\theta(x) = \partial_x V(x, \theta)$. This is the one-dimensional instance of the Stein construction in Section 4, obtained by taking the constant vector field $h = -1$, and it is available at essentially no extra cost because the Langevin step already computes $\partial_x V$. After the Langevin step, the corrected moment is

$$\widehat{m}^{\mathrm{cv}}_k = \frac{1}{n}\sum_{i=1}^n \Phi(X^i_{k+1}) - \widetilde{B}^{p^{(n)}_{j+1}}_{\theta_j} \frac{1}{n}\sum_{i=1}^n \psi_{\theta_k}(X^i_{k+1}),$$

where $\widetilde{B}^{p^{(n)}_{j+1}}_{\theta_j}$ is the smoothed coefficient from the most recent refit $j < k$, with $p^{(n)}_{j+1}$ the empirical measure after the particle update as in Section 4. The correction is zero before the first refit. Starting at the 200th iteration, we refit the coefficient every 25 iterations by ridge regression on the same particles X^i_{k+1} used in $\widehat{m}^{\mathrm{cv}}_k$, with relative ridge parameter 0.1 and EMA weight 0.25. The smoothed coefficient is first used at the next iteration.

F.2 Neural energy-based model

The EBM experiment in Fig. 2 uses the UCI Digits dataset (Alpaydin and Kaynak, 1998), loaded through scikit-learn (Pedregosa et al., 2011), and represented as vectors $x \in [-1, 1]^{64}$. The model is the neural Gibbs family

$$\pi(\theta)(x) \propto \exp[-V(x, \theta)], \quad V(x, \theta) = \frac{\|x\|^2}{2\sigma_0^2} + f_\theta(x),$$

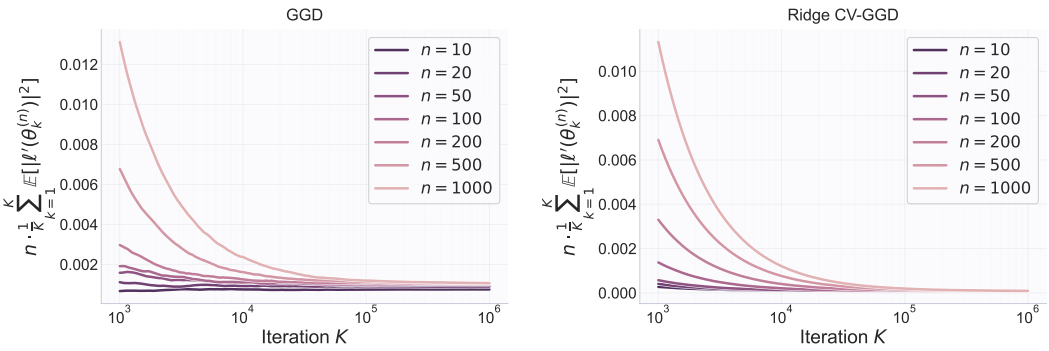


Figure 3: Late-time floor-collapse plots for the finite-observable lower-bound witness. Left: standard GGD. Right: Ridge CV-GGD. We plot $n \cdot \mathrm{LHS}_K^{(n)}$ from $K = 10^3$ onward; both panels are consistent with a C/n floor, with a much smaller constant for the control-variate estimator.

where f_θ is a one-hidden-layer $\tanh$ network with hidden width 64 and $\sigma_0 = 1$. The parameter $\theta \in \mathbb{R}^p$ contains all weights and biases of the network. Let ν_{train} denote the empirical measure on the training subset. For the negative log-likelihood objective, the gradient has the Gibbs form

$$\nabla \ell(\theta) = \mathbb{E}_{X \sim \nu_{\text{train}}}[\nabla_\theta f_\theta(X)] - \mathbb{E}_{X \sim \pi(\theta)}[\nabla_\theta f_\theta(X)] \,.$$

The data expectation is computed exactly over the training subset, while the model expectation is estimated by the particles. With $\Phi^\theta(x) = \nabla_\theta f_\theta(x)$, standard GGD uses

$$\widehat{g}_k = \frac{1}{N_{\text{train}}} \sum_{j=1}^{N_{\text{train}}} \Phi^{\theta_k}(x_j) - \frac{1}{n} \sum_{i=1}^{n} \Phi^{\theta_k}(X_k^i).$$

The particles are persistent and are updated by unadjusted Langevin steps

$$X_{k+1}^i = X_k^i - h \nabla_1 V(X_k^i, \theta_k) + \sqrt{2h}\, \xi_{k+1}^i, \qquad \xi_{k+1}^i \sim \mathcal{N}(0, I_{64}).$$

CV-GGD uses the score control

$$\Psi^\theta(x) = \nabla_1 V(x, \theta).$$

This control is zero-mean under $\pi(\theta)$ by Stein's identity, and it is available at essentially no additional cost since it is already computed to move the Langevin particles. The corrected model expectation is

$$\widehat{m}_k^{\text{cv}} = \frac{1}{n} \sum_{i=1}^{n} \Phi^{\theta_k}(X_k^i) - \widetilde{B}_{\theta_j}^{p_j^Y} \frac{1}{n} \sum_{i=1}^{n} \Psi^{\theta_k}(X_k^i),$$

where $\widetilde{B}_{\theta_j}^{p_j^Y}$ is the smoothed coefficient from the most recent refit $j \le k$. The correction is zero before the first refit. The CV-GGD gradient estimate replaces the empirical model expectation in $\widehat{g}_k$ by $\widehat{m}_k^{\text{cv}}$. At each refit, we evaluate $(\Phi^{\theta_k}, \Psi^{\theta_k})$ on a separate persistent Langevin cloud $Y_k^1, \ldots, Y_k^{512}$ with empirical measure p_k^Y. The centered empirical covariances on that cloud give the ridge fit

$$B_{\theta_k}^{p_k^Y} = \widehat{C}_{\Phi\Psi,k}^Y \big(\widehat{C}_{\Psi\Psi,k}^Y + \lambda_k I_{64}\big)^{-1},$$

where $\lambda_k = 10^{-2} \max\{\mathrm{Tr}(\widehat{C}_{\Psi\Psi,k}^Y)/64, 10^{-8}\}$. The smoothed coefficient is applied to the main cloud X_k, so the reported particle count n refers to the particles used in the gradient estimate. Algorithm 2 gives the training update with the two clouds and stores the smoothed coefficient in $\widetilde{B}$.

For the fixed-network diagnostic in Fig. 2, we keep the initial network parameters fixed and estimate the centered variance of the finite-particle gradient estimator,

$$\widehat{V}_n = \mathbb{E}\Big[\|\widehat{\Gamma}^{(n)} - \mathbb{E}\widehat{\Gamma}^{(n)}\|^2\Big] \,.$$

This is an empirical centered-variance diagnostic. We use $n \in \{8, 16, 32, 64, 128\}$ particles, 256 independent replications, and 100 Langevin burn-in steps.

Algorithm 2 CV-GGD training with a separate calibration cloud for the neural EBM

Input: initial θ_0, training images $\{x_\ell\}_{\ell=1}^{N_{\mathrm{train}}}$, stepsizes h, η, particle count n, and budget K
Draw the $n + 512$ initial particles i.i.d. from $\mathcal{N}(0, \sigma_0^2 I_{64})$ and set $\widetilde{B} \leftarrow 0$
for $k \leftarrow 0$ to $K - 1$ **do**
  if $k + 1 \in \{50, 60, 70, \ldots\}$ **then**
    Fit $B_{\theta_k}^{p_k^Y}$ by ridge covariance regression on Y_k and set $\widetilde{B} \leftarrow (\widetilde{B} + B_{\theta_k}^{p_k^Y})/2$
  end if
  $\widehat{m}_k^{\mathrm{cv}} \leftarrow \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \big(\Phi^{\theta_k}(X_k^i) - \widetilde{B}\Psi^{\theta_k}(X_k^i)\big)$
  $\theta_{k+1} \leftarrow \theta_k - \eta \left(\dfrac{1}{N_{\mathrm{train}}} \displaystyle\sum_{\ell=1}^{N_{\mathrm{train}}} \Phi^{\theta_k}(x_\ell) - \widehat{m}_k^{\mathrm{cv}} \right)$
  Draw independent standard Gaussian vectors ξ_{k+1}^i and ζ_{k+1}^j for the two clouds
  $X_{k+1}^i \leftarrow X_k^i - h\nabla_1 V(X_k^i, \theta_k) + \sqrt{2h}\,\xi_{k+1}^i$ for $1 \le i \le n$
  $Y_{k+1}^j \leftarrow Y_k^j - h\nabla_1 V(Y_k^j, \theta_{k+1}) + \sqrt{2h}\,\zeta_{k+1}^j$ for $1 \le j \le 512$
end for
return θ_K

Table 2: Fixed-network neural EBM centered-variance diagnostic. Standard errors are over 256 independent replications. The approximately constant $n\widehat{V}_n$ columns estimate the raw and corrected empirical covariance constants.

Particles n	GGD $\widehat{V}_n$	CV-GGD $\widehat{V}_n$	GGD $n\widehat{V}_n$	CV-GGD $n\widehat{V}_n$
8	2.099 ± 0.024	$(2.619 \pm 0.035) \cdot 10^{-2}$	16.79	0.209
16	1.056 ± 0.012	$(1.330 \pm 0.013) \cdot 10^{-2}$	16.90	0.213
32	$(5.397 \pm 0.062) \cdot 10^{-1}$	$(6.617 \pm 0.050) \cdot 10^{-3}$	17.27	0.212
64	$(2.639 \pm 0.032) \cdot 10^{-1}$	$(3.299 \pm 0.018) \cdot 10^{-3}$	16.89	0.211
128	$(1.303 \pm 0.014) \cdot 10^{-1}$	$(1.663 \pm 0.008) \cdot 10^{-3}$	16.68	0.213

The control coefficient is fitted once on a separate sample of 8192 Langevin particles and reused across replications.

Table 3 reports an additional fixed-network EBM diagnostic against a many-particle reference, decomposing the CV-GGD mean squared error (MSE) into centered variance and squared bias. CV-GGD retains a squared bias of about $4.5 \cdot 10^{-3}$, while its total MSE is 13–72 times smaller than GGD's.

Table 3: Additional fixed-network EBM bias/MSE diagnostic. The $\pm$ values are standard errors over replications; squared bias is a point estimate.

n	GGD MSE	CV-GGD MSE	CV variance	CV bias^2
8	2.184 ± 0.052	0.03022 ± 0.00052	0.02531 ± 0.00053	0.00491
16	1.008 ± 0.025	0.01808 ± 0.00027	0.01326 ± 0.00025	0.00481
32	0.5308 ± 0.0115	0.01124 ± 0.00009	0.006708 ± 0.000072	0.00454
64	0.2652 ± 0.0062	0.007845 ± 0.000055	0.003346 ± 0.000032	0.00450
128	0.13025 ± 0.00293	0.006152 ± 0.000034	0.001657 ± 0.000015	0.00449
256	0.06743 ± 0.00155	0.005361 ± 0.000021	0.000821 ± 0.000006	0.00454

For the training diagnostics in Fig. 2, we train on 1200 digit images and evaluate a validation MMD on a held-out set of 400 images. The left panel uses $K = 500$ outer iterations, 8 random seeds, particle counts $n \in \{8, 16, 32, 64, 128\}$, Langevin step $h = 2 \cdot 10^{-3}$, and outer step $2 \cdot 10^{-2}$. The middle and right panels use the same hyperparameters with a longer budget $K = 1000$, particle count $n = 128$, and 2 random seeds. In all training diagnostics, the CV coefficient is updated every

10 iterations after a warmup of 50 iterations, using 512 calibration particles, relative ridge parameter 10^{-2}, and EMA weight 0.5. Error bars in Fig. 2 are standard errors over replications or seeds.

The Fréchet metric reported in Table 1 is computed directly in the native 64-dimensional coordinate space of the 8×8 UCI Digits images.

The validation MMD should not be expected to decrease monotonically with n at a fixed training horizon. Increasing n reduces Monte Carlo variance but also changes the transient dynamics, the amount of particle-induced exploration, and the interaction with the persistent Langevin bias. For this reason, the particle-count curve is interpreted as a fixed-budget comparison between GGD and CV-GGD at the same n, rather than as a monotonicity statement in the number of particles.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the continuous-time GGD analysis, finite-particle bounds, fluctuation CLT, and control-variate construction. These claims are stated and supported in Sections 3 to 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion includes a dedicated limitations paragraph discussing the continuous-time scope of the analysis, discretization effects, fitted control variates, and possible geometric extensions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are stated in the corresponding theorem and proposition statements in Sections 3 and 4. Complete proofs and auxiliary regularity results are provided in Sections D and E.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experiments are specified in Section 5, with implementation details, parameter choices, seeds or replications, and dataset information given in Section F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper describes the experimental setup and uses a public dataset, but it does not currently provide open access to code with reproduction instructions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main experimental results are described in Section 5, and detailed settings for the lower-bound witness and neural EBM experiments are provided in Section F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments report Monte Carlo averages over independent seeds or replications, and the EBM diagnostics report standard errors; see Section F.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper specifies experimental hyperparameters and budgets, but it does not currently report hardware, memory, runtime, or total compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and experimental research on optimization and sampling, uses standard public benchmark data, and does not involve human subjects or high-risk data/model release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The work is foundational research on sampling and optimization, and is not tied to a specific deployment or application with direct societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release data or models with high misuse risk, such as pre-trained generative models or scraped datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The UCI Digits dataset and scikit-learn are credited in Section F.2, but the manuscript does not currently state their licenses or terms of use explicitly.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset, model, benchmark, or software asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve crowdsourcing or research with human subjects, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development and experiments do not involve LLMs as an important, original, or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.