Proceedings of Machine Learning Research vol 65:1–30, 2017

Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis

Maxim Raginsky MAXIM@ILLINOIS.EDU
University of Illinois

Alexander Rakhlin RAKHLIN@WHARTON.UPENN.EDU
University of Pennsylvania

Matus Telgarsky MJT@ILLINOIS.EDU
University of Illinois and Simons Institute

Abstract

Stochastic Gradient Langevin Dynamics (SGLD) is a popular variant of Stochastic Gradient Descent, where properly scaled isotropic Gaussian noise is added to an unbiased estimate of the gradient at each iteration. This modest change allows SGLD to escape local minima and suffices to guarantee asymptotic convergence to global minimizers for sufficiently regular non-convex objectives (Gelfand and Mitter, 1991).

The present work provides a nonasymptotic analysis in the context of non-convex learning problems, giving finite-time guarantees for SGLD to find approximate minimizers of both empirical and population risks.

As in the asymptotic setting, our analysis relates the discrete-time SGLD Markov chain to a continuous-time diffusion process. A new tool that drives the results is the use of weighted transportation cost inequalities to quantify the rate of convergence of SGLD to a stationary distribution in the Euclidean 2-Wasserstein distance.

1. Introduction and informal summary of results

Consider a stochastic optimization problem

$$\text{minimize} \qquad F(w) := \mathbf{E}_P[f(w,Z)] = \int_{\mathsf{Z}} f(w,z)P(\mathrm{d}z),$$

where w takes values in $\mathbb{R}^d$ and Z is a random element of some space Z with an unknown probability law P. We have access to an n-tuple $\mathbf{Z} = (Z_1, \ldots, Z_n)$ of i.i.d. samples drawn from P, and our goal is to generate a (possibly random) hypothesis $\widehat{W} \in \mathbb{R}^d$ with small expected excess risk

$$\mathbf{E}F(\widehat{W}) - F^*, \tag{1.1}$$

where $F^* := \inf_{w\in\mathbb{R}^d} F(w)$, and the expectation is with respect to the training data $\mathbf{Z}$ and any additional randomness used by the algorithm for generating $\widehat{W}$.

When the functions $w \mapsto f(w,z)$ are not convex, theoretical analysis of global convergence becomes largely intractable. On the other hand, non-convex optimization is currently witnessing an impressive string of empirical successes, most notably in the realm of deep neural networks. Towards the aim of bridging this gap between theory and practice, this paper provides a theoretical

justification for *Stochastic Gradient Langevin Dynamics* (SGLD), a popular variant of stochastic gradient descent, in which properly scaled isotropic Gaussian noise is added to an unbiased estimate of the gradient at each iteration (Gelfand and Mitter, 1991; Borkar and Mitter, 1999; Welling and Teh, 2011).

Since the population distribution P is unknown, we attempt to (approximately) minimize

$$F_{\mathbf{z}}(w) := \frac{1}{n}\sum_{i=1}^{n} f(w, z_i), \tag{1.2}$$

the empirical risk of a hypothesis $w \in \mathbb{R}^d$ on a dataset $\mathbf{z} = (z_1, \ldots, z_n) \in \mathsf{Z}^n$. The SGLD algorithm studied in this work is given by the recursion

$$W_{k+1} = W_k - \eta g_k + \sqrt{2\eta\beta^{-1}}\xi_k \tag{1.3}$$

where g_k is a conditionally unbiased estimate of the gradient $\nabla F_{\mathbf{z}}(W_k)$, ξ_k is a standard Gaussian random vector in $\mathbb{R}^d$, $\eta > 0$ is the step size, and $\beta > 0$ is the inverse temperature parameter. Our analysis begins with the standard observation (see, e.g., Borkar and Mitter (1999) for a rigorous treatment or Welling and Teh (2011) for a heuristic discussion) that the discrete-time Markov process (1.3) can be viewed as a discretization of the continuous-time Langevin diffusion described by the Itô stochastic differential equation

$$\mathrm{d}W(t) = -\nabla F_{\mathbf{z}}(W(t))\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B(t), \qquad t \geq 0 \tag{1.4}$$

where $\{B(t)\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^d$. Under suitable assumptions on f, it can be shown that the Gibbs measure $\pi_{\mathbf{z}}(\mathrm{d}w) \propto \exp(-\beta F_{\mathbf{z}}(w))$ is the unique invariant distribution of (1.4), and that the distributions of $W(t)$ converge rapidly to $\pi_{\mathbf{z}}$ as $t \to \infty$ (Chiang et al., 1987). Moreover, for all sufficiently large values of β, the Gibbs distribution concentrates around the minimizers of $F_{\mathbf{z}}$ (Hwang, 1980). Consequently, a draw from the Gibbs distribution is, with high probability, an almost-minimizer of the empirical risk (1.2), and, if one can show that the SGLD recursion tracks the Langevin diffusion in a suitable sense, then it follows that the distributions of W_k will be close to the Gibbs measure for all sufficiently large k. Hence, one can argue that, for large enough k, the output of SGLD is also an almost-minimizer of the empirical risk.

It is well-recognized, however, that minimization of the empirical risk $F_{\mathbf{z}}$ does not immediately translate into minimization of the population risk F. A standard approach for addressing the issue is to decompose the excess risk into a sum of two terms, $F(\widehat{W}) - F_{\mathbf{z}}(\widehat{W})$ (the generalization error of $\widehat{W}$) and $F_{\mathbf{z}}(\widehat{W}) - F^*$ (the gap between the empirical risk of $\widehat{W}$ and the minimum of the population risk), and then show that both of these terms are small (either in expectation or with high probability). Taking $\widehat{W} = W_k$ and letting $\widehat{W}^*$ be the output of the Gibbs algorithm under which the conditional distribution of $\widehat{W}^*$ given $\mathbf{Z} = \mathbf{z}$ is equal to $\pi_{\mathbf{z}}$, we decompose the excess risk (1.1) as follows:

$$\mathbf{E}F(\widehat{W}) - F^* = \big(\mathbf{E}F(\widehat{W}) - \mathbf{E}F(\widehat{W}^*)\big) + \big(\mathbf{E}F(\widehat{W}^*) - \mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*)\big) + \big(\mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*) - F^*\big), \tag{1.5}$$

where the first term is the difference of expected population risks of SGLD and the Gibbs algorithm, the second term is the generalization error of the Gibbs algorithm, and the third term is easily upper-bounded in terms of expected suboptimality $\mathbf{E}\big(F_{\mathbf{Z}}(\widehat{W}^*) - \min_w F_{\mathbf{Z}}(w)\big)$ of the Gibbs algorithm

for the empirical risk. Observe that only the first term pertains to SGLD, whereas the other two involve solely the Gibbs distribution. The main contribution of this work is in showing finite-time convergence of SGLD for a non-convex objective function. Informally, we can state our main result as follows:

1. For any $\varepsilon > 0$, the first term in (1.5) scales as

$$\varepsilon \cdot \mathrm{Poly}\left(\beta, d, \frac{1}{\lambda_*}\right) \quad \text{for } k \succeq \mathrm{Poly}\left(\beta, d, \frac{1}{\lambda_*}\right) \cdot \frac{1}{\varepsilon^4} \text{ and } \eta \leq \left(\frac{\varepsilon}{\log(1/\varepsilon)}\right)^4, \tag{1.6}$$

   where λ_* is a certain spectral gap parameter that governs the exponential rate of convergence of the Langevin diffusion to its stationary distribution. This spectral gap parameter itself might depend on β and d, but is independent of n.

2. The second and third terms in (1.5) scale, respectively, as

$$\frac{(\beta + d)^2}{\lambda_* n}, \quad \frac{d \log(\beta + 1)}{\beta}. \tag{1.7}$$

1.1. Method of analysis: an overview

Our analysis draws heavily on the theory of optimal transportation (Villani, 2003) and on the analysis of Markov diffusion operators (Bakry et al., 2014) (the necessary background on Markov semigroups and functional inequalities is given in Appendix A). In particular, we control the convergence of SGLD to the Gibbs distribution in terms of 2-*Wasserstein distance*

$$\mathcal{W}_2(\mu, \nu) := \inf\left\{(\mathbf{E}\|V - W\|^2)^{1/2} : \mu = \mathcal{L}(V), \nu = \mathcal{L}(W)\right\},$$

where $\|\cdot\|$ is the Euclidean (ℓ^2) norm on $\mathbb{R}^d$, μ and ν are Borel probability measures on $\mathbb{R}^d$ with finite second moments, and the infimum is taken over all random couples (V, W) taking values in $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $V \sim \mu$ and $W \sim \nu$.

To control the first term on the right-hand side of (1.5), we first upper-bound the 2-Wasserstein distance between the distributions of W_k (the kth iterate of SGLD) and $W(k\eta)$ (the point reached by the Langevin diffusion at time $t = k\eta$). This requires some heavy lifting: Existing bounds on the 2-Wasserstein distance between a diffusion process and its time-discretized version due to Alfonsi et al. (2015) scale like $\eta e^{k\eta}$, which is far too crude for our purposes. By contrast, we take an indirect route via a Girsanov-type change of measure and a weighted transportation-cost inequality of Bolley and Villani (2005) to obtain a bound that scales like $k\eta \cdot \eta^{1/4}$. This step relies crucially on a certain exponential integrability property of the Langevin diffusion. Next, we show that the Gibbs distribution satisfies a logarithmic Sobolev inequality, which allows us to conclude that the 2-Wasserstein distance between the distribution of $W(k\eta)$ and the Gibbs distribution decays exponentially as $e^{-k\eta}$. Since $\mathcal{W}_2$ satisfies the triangle inequality, we can produce an upper bound on the first term in (1.5) that scales as $k\eta \cdot \eta^{1/4} + e^{-k\eta}$. This immediately suggests that we can make this term as small as we wish by first choosing a large enough horizon $t = k\eta$ and then a small enough step size η. Overall, this leads to the bounds stated in (1.6).

To control the second term in (1.5), we show that the Gibbs algorithm is *stable* in 2-Wasserstein distance with respect to local perturbations of the training dataset. This step, again, relies on the

logarithmic Sobolev inequality for the Gibbs distribution. To control the third term, we use a nonasymptotic Laplace integral approximation to show that a single draw from the Gibbs distribution is an approximate minimizer of the empirical risk. We use a Wasserstein continuity result due to Polyanskiy and Wu (2016) and a well-known equivalence between stability of empirical minimization and generalization (Mukherjee et al., 2006; Rakhlin et al., 2005) to show that, in fact, the Gibbs algorithm samples from near-minimizers of the population risk.

We remark that our result readily extends to the case when the stochastic gradients g_k in (1.3) are formed with respect to independent draws from the data-generating distribution P – e.g., when taking a single pass through the dataset. In this case, the target of optimization is F itself rather than $F_{\mathbf{Z}}$, and we simply omit the second term in (1.5). If the main concern is not consistency (as in (1.1)) but rather the generalization performance of SGLD itself, then the same analysis applied to the decomposition

$$\begin{aligned}\mathbf{E}F_{\mathbf{Z}}(\widehat{W}) - \mathbf{E}F(\widehat{W}) = \big(\mathbf{E}F_{\mathbf{Z}}(\widehat{W}) - \mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*)\big) \\ + \big(\mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*) - \mathbf{E}F(\widehat{W}^*)\big) + \big(\mathbf{E}F(\widehat{W}^*) - \mathbf{E}F(\widehat{W})\big) \qquad (1.8)\end{aligned}$$

gives an upper bound of (1.6) plus the first term of (1.7). In other words, while the rate of (1.1) may be hampered by the slow convergence of $\frac{d \log \beta}{\beta}$, the rate of generalization is not. Finally, if each data point is used only once, the generalization performance is controlled by (1.6) alone.

1.2. Related work

The asymptotic study of convergence of discretized Langevin diffusions for non-convex objectives has a long history, starting with the work of Gelfand and Mitter (1991). Most of the work has focused on annealing-type schemes, where both the step size η and the temperature $1/\beta$ are decreased with time. Márquez (1997) and Pelletier (1998) studied the rates of weak convergence for both the Langevin diffusion and the discrete-time updates. However, when η and β are kept fixed, the updates do not converge to a global minimizer, but one can still aim for convergence to a stationary distribution. An asymptotic study of this convergence, in the sense of relative entropy, was initiated by Borkar and Mitter (1999).

Dalalyan and Tsybakov (2012) and Dalalyan (2016) analyzed rates of convergence of discrete-time Langevin updates (with exact gradients) in the case of convex functions, and provided nonasymptotic rates of convergence in the total variation distance for sampling from log-concave densities. Durmus and Moulines (2015) refined these results by establishing geometric convergence in total variation distance for convex and strongly convex objective functions, and provided some results for non-convex objectives that can be represented as a bounded perturbation of a convex or a strongly convex function. Bubeck et al. (2015) studied projected Langevin updates in the convex case.

Our work is motivated in part by recent papers on non-convex optimization and, in particular, on optimization problems related to neural networks. A heuristic analysis of SGLD was given by Welling and Teh (2011), and a modification of SGLD to improve generalization performance was recently proposed by Chaudhari et al. (2016). Deliberate addition of noise was also proposed by Ge et al. (2015) as a strategy for escaping from saddle points, and Belloni et al. (2015) analyzed a simulated anealing method based on Hit-and-Run for sampling from nearly log-concave distributions. While these methods aim at avoiding local minima through randomn perturbations, the line

of work on continuation methods and graduated optimization (Hazan et al., 2016) attempts to create sequences of smoothed approximations that can successively localize the optimum.

Hardt et al. (2015) studied uniform stability and generalization properties of stochastic gradient descent with both convex and non-convex objectives. For the non-convex case, their upper bound on stability degrades with the number of steps of the optimization procedure, which was taken by the authors as a prescription for early stopping. In contrast, we show that, under our assumptions, non-convexity does not imply loss of stability when the latter is measured in terms of 2-Wasserstein distance to the stationary distribution. In addition, we use the fact that Gibbs distribution concentrates on approximate empirical minimizers, implying convergence for the *population* risk via stability (Rakhlin et al., 2005; Mukherjee et al., 2006).

2. The main result

We begin by giving a precise description of the SGLD recursion. A *stochastic gradient oracle*, i.e., the mechanism for accessing the gradient of $F_{\mathbf{z}}$ at each iteration, consists of a collection $(Q_{\mathbf{z}})_{\mathbf{z}\in\mathsf{Z}^n}$ of probability measures on some space U and a mapping $g : \mathbb{R}^d \times \mathsf{U} \to \mathbb{R}^d$, such that, for every $\mathbf{z} \in \mathsf{Z}^n$,

$$\mathbf{E}g(w, U_{\mathbf{z}}) = \nabla F_{\mathbf{z}}(w), \qquad \forall w \in \mathbb{R}^d \tag{2.1}$$

where $U_{\mathbf{z}}$ is a random element of U with probability law $Q_{\mathbf{z}}$. Conditionally on $\mathbf{Z} = \mathbf{z}$, the SGLD update takes the form

$$W_{k+1} = W_k - \eta g(W_k, U_{\mathbf{z},k}) + \sqrt{2\eta\beta^{-1}}\xi_k, \qquad k = 0, 1, 2, \ldots \tag{2.2}$$

where $\{U_{\mathbf{z},k}\}_{k=0}^{\infty}$ is a sequence of i.i.d. random elements of U with probability law $Q_{\mathbf{z}}$ and $\{\xi_k\}_{k=0}^{\infty}$ is a sequence of i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$. We assume that W_0, $(\mathbf{Z}, \{U_{\mathbf{Z},k}\}_{k=0}^{\infty})$, and $\{\xi_k\}_{k=0}^{\infty}$ are mutually independent. We impose the following assumptions (see the discussion in Section 4 for additional details):

(A.1) The function f takes nonnegative real values, and there exist constants $A, B \geq 0$, such that

$$|f(0, z)| \leq A \qquad \text{and} \qquad \|\nabla f(0, z)\| \leq B \qquad \forall z \in \mathsf{Z}.$$

(A.2) For each $z \in \mathsf{Z}$, the function $f(\cdot, z)$ is M-*smooth*: for some $M > 0$,

$$\|\nabla f(w, z) - \nabla f(v, z)\| \leq M\|w - v\|, \qquad \forall w, v \in \mathbb{R}^d.$$

(A.3) For each $z \in \mathsf{Z}$, the function $f(\cdot, z)$ is (m, b)-*dissipative* (Hale, 1988): for some $m > 0$ and $b \geq 0$,

$$\langle w, \nabla f(w, z)\rangle \geq m\|w\|^2 - b, \qquad \forall w \in \mathbb{R}^d. \tag{2.3}$$

(A.4) There exists a constant $\delta \in [0, 1)$, such that, for each $\mathbf{z} \in \mathsf{Z}^n$,[1]

$$\mathbf{E}[\|g(w, U_{\mathbf{z}}) - \nabla F_{\mathbf{z}}(w)\|^2] \leq 2\delta\left(M^2\|w\|^2 + B^2\right), \qquad \forall w \in \mathbb{R}^d. \tag{2.4}$$

1. We are reusing the constants M and B from **(A.1)** and **(A.2)** in (2.4) mainly out of considerations of technical convenience; any other constants $M', B' > 0$ can be substituted in their place without affecting the results.

(A.5) The probability law μ_0 of the initial hypothesis W_0 has a bounded and strictly positive density p_0 with respect to the Lebesgue measure on $\mathbb{R}^d$, and

$$\kappa_0 := \log \int_{\mathbb{R}^d} e^{\|w\|^2} p_0(w) \mathrm{d}w < \infty.$$

We are now ready to state our main result. A crucial role will be played by the *uniform spectral gap*

$$\lambda_* := \inf_{\mathbf{z} \in \mathsf{Z}^n} \inf \left\{ \frac{\int_{\mathbb{R}^d} \|\nabla g\|^2 \mathrm{d}\pi_{\mathbf{z}}}{\int_{\mathbb{R}^d} g^2 \mathrm{d}\pi_{\mathbf{z}}} : g \in C^1(\mathbb{R}^d) \cap L^2(\pi_{\mathbf{z}}),\, g \neq 0,\, \int_{\mathbb{R}^d} g \mathrm{d}\pi_{\mathbf{z}} = 0 \right\}, \tag{2.5}$$

where $\pi_{\mathbf{z}}(\mathrm{d}w) \propto e^{-\beta F_{\mathbf{z}}(w)} \mathrm{d}w$ is the Gibbs distribution. As detailed in Section 4, Assumptions **(A.1)–(A.3)** suffice to ensure that $\lambda_* > 0$. In the statement of the theorem, the notation $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ gives explicit dependence on the parameters β, λ_*, and d, but hides factors that depend (at worst) polynomially on the parameters $A, B, 1/m, b, M, \kappa_0$. Explicit expressions for all constants are given in the proof.

Theorem 1 *Suppose that the regularity conditions* **(A.1)–(A.5)** *hold. Then, for any* $\beta \geq 1 \vee 2/m$ *and any* $\varepsilon \in (0, \frac{m}{4M^2} \wedge e^{-\tilde{\Omega}(\lambda_*/\beta(d+\beta))})$*, the expected excess risk of* W_k *is bounded by*

$$\mathbf{E}F(W_k) - F^* \leq \tilde{\mathcal{O}}\left(\frac{\beta(\beta+d)^2}{\lambda_*} \left(\delta^{1/4} \log\left(\frac{1}{\varepsilon}\right) + \varepsilon \right) + \frac{(\beta+d)^2}{\lambda_* n} + \frac{d \log(\beta+1)}{\beta} \right), \tag{2.6}$$

provided

$$k = \tilde{\Omega}\left(\frac{\beta(d+\beta)}{\lambda_* \varepsilon^4} \log^5\left(\frac{1}{\varepsilon}\right) \right) \quad \textit{and} \quad \eta \leq \left(\frac{\varepsilon}{\log(1/\varepsilon)} \right)^4. \tag{2.7}$$

3. Proof of Theorem 1

3.1. A high-level overview

Let $\mu_{\mathbf{z},k} := \mathcal{L}(W_k|\mathbf{Z} = \mathbf{z})$, $\nu_{\mathbf{z},t} := \mathcal{L}(W(t)|\mathbf{Z} = \mathbf{z})$, and $\mathbf{E}_{\mathbf{z}}[\cdot] := \mathbf{E}[\cdot|\mathbf{Z} = \mathbf{z}]$. In a nutshell, our proof of Theorem 1 consists of the following steps:

1. We first show that, for all sufficiently small $\eta > 0$, the SGLD recursion (2.2) tracks the continuous-time Langevin diffusion process (1.4) in 2-Wasserstein distance:

$$\mathcal{W}_2(\mu_{\mathbf{z},k}, \nu_{\mathbf{z},k\eta}) = \tilde{\mathcal{O}}\left((\beta+d)(\delta^{1/4} + \eta^{1/4}) k\eta \right) \tag{3.1}$$

   (the precise statement with explicit constants is given in Proposition 8).

2. Next, we show that the Langevin diffusion (1.4) converges exponentially fast to the Gibbs distribution $\pi_{\mathbf{z}}$:

$$\mathcal{W}_2(\nu_{\mathbf{z},k\eta}, \pi_{\mathbf{z}}) = \tilde{\mathcal{O}}\left(\frac{\beta+d}{\sqrt{\lambda_*}} \right) e^{-\tilde{\Omega}(\lambda_* k\eta/\beta(d+\beta))}.$$

This, together with (3.1) and the triangle inequality, yields the estimate

$$\mathcal{W}_2(\mu_{\mathbf{z},k},\pi_{\mathbf{z}}) = \tilde{\mathcal{O}}\left((\beta+d)(\delta^{1/4}+\eta^{1/4})k\eta\right) + \tilde{\mathcal{O}}\left(\frac{\beta+d}{\sqrt{\lambda_*}}\right)e^{-\tilde{\Omega}(\lambda_* k\eta/\beta(d+\beta))} \tag{3.2}$$

(see Proposition 10 for explicit constants). Observe that there are two terms on the right-hand side of (3.2), one of which grows linearly with $t = k\eta$, while the other one decays exponentially with t. Thus, we can first choose t large enough and then η small enough, so that

$$\mathcal{W}_2(\mu_{\mathbf{z},k},\pi_{\mathbf{z}}) = \tilde{\mathcal{O}}\left(\frac{\beta(d+\beta)^2}{\lambda_*}\left(\delta^{1/4}\log\left(\frac{1}{\varepsilon}\right)+\varepsilon\right)\right). \tag{3.3}$$

The resulting choices of $t = k\eta$ and η translate into the expressions for k and η given in (A.14).

3. The upshot of Eq. (3.3) is that, for large enough k, the conditional probability law of W_k given $\mathbf{Z} = \mathbf{z}$ is close, in 2-Wasserstein, to the Gibbs distribution $\pi_{\mathbf{z}}$. Thus, we are led to consider the *Gibbs algorithm* that generates a random draw from $\pi_{\mathbf{z}}$. We show that the resulting hypothesis is an almost-minimizer of the empirical risk, i.e.,

$$\int_{\mathbb{R}^d} F_{\mathbf{z}}(w)\pi_{\mathbf{z}}(\mathrm{d}w) - \min_{w\in\mathbb{R}^d} F_{\mathbf{z}}(w) = \tilde{\mathcal{O}}\left(\frac{d}{\beta}\log\frac{\beta+1}{d}\right) \tag{3.4}$$

(see Proposition 11 for the exact statement), and also that the Gibbs algorithm is stable in the 2-Wasserstein distance: for any two datasets $\mathbf{z}, \bar{\mathbf{z}}$ that differ in a single coordinate,

$$\mathcal{W}_2(\pi_{\mathbf{z}},\pi_{\bar{\mathbf{z}}}) = \tilde{\mathcal{O}}\left(\frac{\beta(d+\beta)\sqrt{1+d/\beta}}{\lambda_* n}\right).$$

This estimate, together with Lemma 6 below, implies that the Gibbs algorithm is uniformly stable (Bousquet and Elisseeff, 2002):

$$\sup_{z\in\mathsf{Z}}\left|\int_{\mathbb{R}^d} f(w,z)\pi_{\mathbf{z}}(\mathrm{d}w) - \int_{\mathbb{R}^d} f(w,z)\pi_{\bar{\mathbf{z}}}(\mathrm{d}w)\right| = \tilde{\mathcal{O}}\left(\frac{(\beta+d)^2}{\lambda_* n}\right) \tag{3.5}$$

(see Proposition 12). The almost-ERM property (3.4) and the uniform stability propery (3.5), together with (3.3), yield the statement of the theorem.

3.2. Technical lemmas

We first collect a few lemmas that will be used in the sequel; see Appendix C for the proofs.

Lemma 2 (quadratic bounds on f) *For all $w \in \mathbb{R}^d$ and $z \in \mathsf{Z}$,*

$$\|\nabla f(w,z)\| \le M\|w\| + B \tag{3.6}$$

and

$$\frac{m}{3}\|w\|^2 - \frac{b}{2}\log 3 \le f(w,z) \le \frac{M}{2}\|w\|^2 + B\|w\| + A. \tag{3.7}$$

Lemma 3 (uniform L^2 bounds on SGLD and Langevin diffusion) *For all $0 < \eta < 1 \wedge \frac{m}{4M^2}$ and all $\mathbf{z} \in \mathsf{Z}^n$,*

$$\sup_{k \geq 0} \mathbf{E}_{\mathbf{z}} \|W_k\|^2 \leq \kappa_0 + 2\left(1 \vee \frac{1}{m}\right)\left(b + 2B^2 + \frac{d}{\beta}\right). \tag{3.8}$$

and

$$\mathbf{E}_{\mathbf{z}} \|W(t)\|^2 \leq \kappa_0 e^{-2mt} + \frac{b + d/\beta}{m}\left(1 - e^{-2mt}\right) \tag{3.9}$$

$$\leq \kappa_0 + \frac{b + d/\beta}{m}. \tag{3.10}$$

Lemma 4 (exponential integrability of Langevin diffusion) *For all $\beta \geq 2/m$, we have*

$$\log \mathbf{E}_{\mathbf{z}}\left[e^{\|W(t)\|^2}\right] \leq \kappa_0 + 2\left(b + \frac{d}{\beta}\right) t. \tag{3.11}$$

Lemma 5 (relative entropy bound) *For any $w \in \mathbb{R}^d$ and any $\mathbf{z} \in \mathsf{Z}^n$,*

$$D(\mu_0 \| \pi_{\mathbf{z}}) \leq \log \|p_0\|_\infty + \frac{d}{2} \log \frac{3\pi}{m\beta} + \beta\left(\frac{M\kappa_0}{3} + B\sqrt{\kappa_0} + A + \frac{b}{2}\log 3\right). \tag{3.12}$$

Lemma 6 (2-Wasserstein continuity for functions of quadratic growth, Polyanskiy and Wu (2016)) *Let μ, ν be two probability measures on $\mathbb{R}^d$ with finite second moments, and let $g : \mathbb{R}^d \to \mathbb{R}$ be a C^1 function obeying*

$$\|\nabla g(w)\| \leq c_1 \|w\| + c_2, \qquad \forall w \in \mathbb{R}^d \tag{3.13}$$

for some constants $c_1 > 0$ and $c_2 \geq 0$. Then

$$\left| \int_{\mathbb{R}^d} g \, \mathrm{d}\mu - \int_{\mathbb{R}^d} g \, \mathrm{d}\nu \right| \leq (c_1 \sigma + c_2) \mathcal{W}_2(\mu, \nu) \tag{3.14}$$

where $\sigma^2 := \int_{\mathbb{R}^d} \mu(\mathrm{d}w)\|w\|^2 \vee \int_{\mathbb{R}^d} \nu(\mathrm{d}w)\|w\|^2$.

3.3. The diffusion approximation

Recall that $\mu_{\mathbf{z},k} = \mathcal{L}(W_k | \mathbf{Z} = \mathbf{z})$ and $\nu_{\mathbf{z},t} = \mathcal{L}(W(t) | \mathbf{Z} = \mathbf{z})$, and we take $\mu_{\mathbf{z},0} = \nu_{\mathbf{z},0} = \mu_0$. In this section, we derive an upper bound on the 2-Wasserstein distance $\mathcal{W}_2(\mu_{\mathbf{z},k}, \nu_{\mathbf{z},k\eta})$. The analysis consists of two steps. We first upper-bound the relative entropy $D(\mu_{\mathbf{z},k} \| \nu_{\mathbf{z},k\eta})$ via a change-of-measure argument following Dalalyan and Tsybakov (2012) (see also Dalalyan (2016)), except that we also have to deal with the error introduced by the stochastic gradient oracle. We then use a weighted transportation-cost inequality of Bolley and Villani (2005) to control the Wasserstein distance $\mathcal{W}_2(\mu_{\mathbf{z},k}, \nu_{\mathbf{z},k\eta})$ in terms of $D(\mu_{\mathbf{z},k} \| \nu_{\mathbf{z},k\eta})$.

The proof of the following lemma is somewhat lengthy, and is given in Appendix D:

Lemma 7 *For any $k \in \mathbb{N}$ and any $\eta \in (0, 1 \wedge \frac{m}{4M^2})$, we have*

$$D(\mu_{\mathbf{z},k} \| \nu_{\mathbf{z},k\eta}) \leq (C_0 \beta \delta + C_1 \eta)\, k\eta,$$

with

$$C_0 = \left(M^2 \left(\kappa_0 + 2 \left(1 \vee \frac{1}{m} \right) \left(b + 2B^2 + \frac{d}{\beta} \right) \right) + B^2 \right), \qquad C_1 = 6M^2 \left(\beta C_0 + d \right).$$

We now use the following result of Bolley and Villani (2005, Cor. 2.3): For any two Borel probability measures μ, ν on $\mathbb{R}^d$ with finite second moments,

$$\mathcal{W}_2(\mu, \nu) \leq C_\nu \left[\sqrt{D(\mu \| \nu)} + \left(\frac{D(\mu \| \nu)}{2} \right)^{1/4} \right],$$

where

$$C_\nu = 2 \inf_{\lambda > 0} \left(\frac{1}{\lambda} \left(\frac{3}{2} + \log \int_{\mathbb{R}^d} e^{\lambda \| w \|^2} \nu(\mathrm{d}w) \right) \right)^{1/2}.$$

Let $\mu = \mu_{\mathbf{z},k}$, $\nu = \nu_{\mathbf{z},k\eta}$, and take $\lambda = 1$. Suppose $k\eta \geq 1$. Since $\beta \geq \frac{2}{m}$, we can use Lemma 4 to write

$$\mathcal{W}_2^2(\mu_{\mathbf{z},k}, \nu_{\mathbf{z},k\eta}) \leq \left(12 + 8 \left(\kappa_0 + 2b + \frac{2d}{\beta} \right) k\eta \right) \cdot \left(D(\mu_{\mathbf{z},k} \| \nu_{\mathbf{z},k\eta}) + \sqrt{D(\mu_{\mathbf{z},k} \| \nu_{\mathbf{z},k\eta})} \right).$$

Moreover, for all k and η satisfying the conditions of Lemma 7, plus the additional requirement $k\eta \geq 1$, we can write

$$D(\mu_{\mathbf{z},k} \| \nu_{\mathbf{z},k\eta}) + \sqrt{D(\mu_{\mathbf{z},k} \| \nu_{\mathbf{z},k\eta})} \leq \left(C_1 + \sqrt{C_1} \right) k\eta^{3/2} + \left(\beta C_0 + \sqrt{\beta C_0} \right) \cdot k\eta \sqrt{\delta}.$$

Putting everything together, we obtain the following result:

Proposition 8 *For any $k \in \mathbb{N}$ and any $\eta \in (0, 1 \wedge \frac{m}{4M^2})$ obeying $k\eta \geq 1$, we have*

$$\mathcal{W}_2^2(\mu_{\mathbf{z},k}, \nu_{\mathbf{z},k\eta}) \leq \left(\tilde{C}_0^2 \sqrt{\delta} + \tilde{C}_1^2 \sqrt{\eta} \right) \cdot (k\eta)^2, \tag{3.15}$$

with

$$\tilde{C}_0^2 := \left(12 + 8 \left(\kappa_0 + 2b + \frac{2d}{\beta} \right) \right) \left(\beta C_0 + \sqrt{\beta C_0} \right)$$

and

$$\tilde{C}_1^2 := \left(12 + 8 \left(\kappa_0 + 2b + \frac{2d}{\beta} \right) \right) \left(C_1 + \sqrt{C_1} \right).$$

3.4. Wasserstein distance to the Gibbs distribution

We now fix a time $t \geq 0$ and examine the 2-Wasserstein distance $\mathcal{W}_2(\nu_{\mathbf{z},t}, \pi_{\mathbf{z}})$. At this point, we need to use a number of concepts from the analysis of Markov diffusion operators; see Appendix A for the requisite background. We start by showing the following:

Proposition 9 *For $\beta \geq 2/m$, all of the the Gibbs measures $\pi_{\mathbf{z}}$ satisfy a logarithmic Sobolev inequality with constant*

$$c_{\mathrm{LS}} \leq \frac{2m^2 + 8M^2}{m^2 M\beta} + \frac{1}{\lambda_*}\left(\frac{6M(d+\beta)}{m} + 2\right).$$

Therefore,

$$\mathcal{W}_2(\mu, \pi_{\mathbf{z}}) \leq \sqrt{2c_{\mathrm{LS}} D(\mu\|\pi_{\mathbf{z}})} \tag{3.16}$$

by the Otto–Villani theorem, and, since $D(\nu_{\mathbf{z},0}\|\pi_{\mathbf{z}}) = D(\mu_0\|\pi_{\mathbf{z}}) < \infty$ by Lemma 5, we also have

$$D(\nu_{\mathbf{z},t}\|\pi_{\mathbf{z}}) \leq D(\mu_0\|\pi_{\mathbf{z}}) e^{-2t/\beta c_{\mathrm{LS}}}. \tag{3.17}$$

by the theorem on exponential decay of entropy. Combining Eqs. (3.16) (with $\mu = \nu_{\mathbf{z},t}$) and (3.17) and using Lemma 5, we get

$$\begin{aligned}\mathcal{W}_2(\nu_{\mathbf{z},t}, \pi_{\mathbf{z}}) &\leq \sqrt{2c_{\mathrm{LS}}\left(\log\|p_0\|_\infty + \frac{d}{2}\log\frac{3\pi}{m\beta} + \beta\left(\frac{M\kappa_0}{3} + B\sqrt{\kappa_0} + A + \frac{b}{2}\log 3\right)\right)} e^{-t/\beta c_{\mathrm{LS}}} \\ &=: \tilde{C}_2 e^{-t/\beta c_{\mathrm{LS}}}.\end{aligned}$$

Letting $t = k\eta$ and invoking Proposition 8, we obtain the following:

Proposition 10 *For all k and η satisfying the conditions of Proposition 8, we have*

$$\mathcal{W}_2(\mu_{\mathbf{z},k}, \pi_{\mathbf{z}}) \leq \left(\tilde{C}_0\delta^{1/4} + \tilde{C}_1\eta^{1/4}\right) k\eta + \tilde{C}_2 e^{-k\eta/\beta c_{\mathrm{LS}}}.$$

3.5. Almost-ERM property of the Gibbs algorithm

In this section and the next one, we focus on the properties of the Gibbs algorithm that generates a random hypothesis $\widehat{W}^*$ with $\mathcal{L}(\widehat{W}^*|\mathbf{Z} = \mathbf{z}) = \pi_{\mathbf{z}}$. Let $p_{\mathbf{z}}(w) = e^{-\beta F_{\mathbf{z}}(w)}/\Lambda_{\mathbf{z}}$ denote the density of the Gibbs measure $\pi_{\mathbf{z}}$ with respect to the Lebesgue measure on $\mathbb{R}^d$, where $\Lambda_{\mathbf{z}} := \int_{\mathbb{R}^d} e^{-\beta F_{\mathbf{z}}(w)}\mathrm{d}w$ is the normalization constant known as the partition function. We start by writing

$$\int_{\mathbb{R}^d} F_{\mathbf{z}}(w)\pi_{\mathbf{z}}(\mathrm{d}w) = \frac{1}{\beta}\left(h(p_{\mathbf{z}}) - \log\Lambda_{\mathbf{z}}\right), \tag{3.18}$$

where

$$h(p_{\mathbf{z}}) = -\int_{\mathbb{R}^d} p_{\mathbf{z}}(w)\log p_{\mathbf{z}}(w)\mathrm{d}w = -\int_{\mathbb{R}^d} \frac{e^{-\beta F_{\mathbf{z}}(w)}}{\Lambda_{\mathbf{z}}}\log\frac{e^{-\beta F_{\mathbf{z}}(w)}}{\Lambda_{\mathbf{z}}}\mathrm{d}w$$

is the differential entropy of $p_{\mathbf{z}}$ (Cover and Thomas, 2006). To upper-bound $h(p_{\mathbf{z}})$, we estimate the second moment of $\pi_{\mathbf{z}}$. From (3.17), it follows that $\mathcal{W}_2(\nu_{\mathbf{z},t}, \pi_{\mathbf{z}}) \xrightarrow{t\to\infty} 0$. Since convergence of

probability measures in 2-Wasserstein distance is equivalent to weak convergence plus convergence of second moments (Villani, 2003, Theorem 7.12), we have by Theorem 3

$$\int_{\mathbb{R}^d} \|w\|^2 \pi_{\mathbf{z}}(\mathrm{d}w) = \lim_{t\to\infty} \int_{\mathbb{R}^d} \|w\|^2 \nu_{\mathbf{z},t}(\mathrm{d}w) \leq \frac{b + d/\beta}{m}. \tag{3.19}$$

The differential entropy of a probability density with a finite second moment is upper-bounded by that of a Gaussian density with the same second moment, so we immediately get

$$h(p_{\mathbf{z}}) \leq \frac{d}{2} \log\left(\frac{2\pi e(b + d/\beta)}{md}\right). \tag{3.20}$$

Moreover, let $w^*_{\mathbf{z}}$ be any point that minimizes $F_{\mathbf{z}}(w)$, i.e., $F^*_{\mathbf{z}} := \min_{w\in\mathbb{R}^d} F_{\mathbf{z}}(w) = F_{\mathbf{z}}(w^*_{\mathbf{z}})$. Then $\nabla F_{\mathbf{z}}(w^*_{\mathbf{z}}) = 0$, and, since $F_{\mathbf{z}}$ is M-smooth, we have $F_{\mathbf{z}}(w) - F^*_{\mathbf{z}} \leq \frac{M}{2}\|w - w^*\|^2$ by Lemma 1.2.3 in Nesterov (2004). As a consequence, we can lower-bound $\log \Lambda_{\mathbf{z}}$ using a Laplace integral approximation:

$$\begin{aligned}
\log \Lambda_{\mathbf{z}} &= \log \int_{\mathbb{R}^d} e^{-\beta F_{\mathbf{z}}(w)} \mathrm{d}w \\
&= -\beta F^*_{\mathbf{z}} + \log \int_{\mathbb{R}^d} e^{\beta(F^*_{\mathbf{z}} - F_{\mathbf{z}}(w))} \mathrm{d}w \\
&\geq -\beta F^*_{\mathbf{z}} + \log \int_{\mathbb{R}^d} e^{-\beta M \|w - w^*_z\|^2/2} \mathrm{d}w \\
&= -\beta F^*_{\mathbf{z}} + \frac{d}{2} \log\left(\frac{2\pi}{M\beta}\right).
\end{aligned} \tag{3.21}$$

Using Eqs. (3.20) and (3.21) in (3.18) and simplifying, we obtain the following result:

Proposition 11 *For any $\beta \geq 2/m$,*

$$\int_{\mathbb{R}^d} F_{\mathbf{z}}(w)\pi_{\mathbf{z}}(\mathrm{d}w) - \min_{w\in\mathbb{R}^d} F_{\mathbf{z}}(w) \leq \frac{d}{2\beta} \log\left(\frac{eM}{m}\left(\frac{b\beta}{d} + 1\right)\right).$$

3.6. Stability of the Gibbs algorithm

Our last step before the final analysis is to show that the Gibbs algorithm is *uniformly stable*. Fix two n-tuples $\mathbf{z} = (z_1, \ldots, z_n), \bar{\mathbf{z}} = (\bar{z}_1, \ldots, \bar{z}_n) \in \mathsf{Z}^n$ with $\mathrm{card}|\{i : z_i \neq \bar{z}_i\}| = 1$. Then the Radon–Nikodym derivative $p_{\mathbf{z},\bar{\mathbf{z}}} = \frac{\mathrm{d}\pi_{\mathbf{z}}}{\mathrm{d}\pi_{\bar{\mathbf{z}}}}$ can be expressed as

$$p_{\mathbf{z},\bar{\mathbf{z}}}(w) = \frac{\exp\left(-\frac{\beta}{n}\big(f(w, z_{i_0}) - f(w, \bar{z}_{i_0})\big)\right)}{\Lambda_{\mathbf{z}}/\Lambda_{\bar{\mathbf{z}}}},$$

where $i_0 \in [n]$ is the index of the coordinate where $\mathbf{z}$ and $\bar{\mathbf{z}}$ differ. In particular,

$$\nabla\sqrt{p_{\mathbf{z},\bar{\mathbf{z}}}(w)} = \frac{\beta}{2n}\Big(\nabla_w f(w, \bar{z}_{i_0}) - \nabla_w f(w, z_{i_0})\Big)\sqrt{p_{\mathbf{z},\bar{\mathbf{z}}}(w)}.$$

Therefore, since $\pi_{\bar{\mathbf{z}}}$ satisfies a logarithmic Sobolev inequality with constant c_{LS} given in Proposition 9, we can write

$$\begin{aligned} D(\pi_{\mathbf{z}}\|\pi_{\bar{\mathbf{z}}}) &\leq 2c_{\mathrm{LS}} \int \left\|\nabla\sqrt{p_{\mathbf{z},\bar{\mathbf{z}}}}\right\|^2 \mathrm{d}\pi_{\bar{\mathbf{z}}} \\ &= \frac{c_{\mathrm{LS}}\beta^2}{2n^2} \int_{\mathbb{R}^d} \left\|\nabla_w f(w,\bar{z}_{i_0}) - \nabla_w f(w,z_{i_0})\right\|^2 p_{\mathbf{z},\bar{\mathbf{z}}}(w)\pi_{\bar{\mathbf{z}}}(\mathrm{d}w) \\ &= \frac{c_{\mathrm{LS}}\beta^2}{2n^2} \int_{\mathbb{R}^d} \left\|\nabla_w f(w,\bar{z}_{i_0}) - \nabla_w f(w,z_{i_0})\right\|^2 \pi_{\mathbf{z}}(\mathrm{d}w) \\ &\leq \frac{2c_{\mathrm{LS}}\beta^2}{n^2}\left(M^2\int_{\mathbb{R}^d}\|w\|^2\pi_{\mathbf{z}}(\mathrm{d}w) + B^2\right), \end{aligned}$$

where the last line follows from the quadratic growth estimate (3.6). Taking $\mu = \pi_{\mathbf{z}}$ in (3.16) and using the above bound and the second-moment estimate (3.19), we obtain

$$\mathcal{W}_2(\pi_{\mathbf{z}},\pi_{\bar{\mathbf{z}}}) \leq \frac{2c_{\mathrm{LS}}\beta}{n}\sqrt{B^2 + \frac{M^2(b+d/\beta)}{m}}.$$

Finally, observe that, for each $z \in \mathsf{Z}$, the function $w \mapsto f(w,z)$ satisfies the conditions of Lemma 6 with $c_1 = M$ and $c_2 = B$, while $\pi_{\mathbf{z}}$ and $\pi_{\bar{\mathbf{z}}}$ satisfy the conditions of Lemma 6 with $\sigma^2 = \frac{b+d/\beta}{m}$. Thus, we obtain the following uniform stability estimate for the Gibbs algorithm:

Proposition 12 *For any two $\mathbf{z},\bar{\mathbf{z}} \in \mathsf{Z}^n$ that differ only in a single coordinate,*

$$\sup_{z\in\mathsf{Z}}\left|\int_{\mathbb{R}^d} f(w,z)\pi_{\mathbf{z}}(\mathrm{d}w) - \int_{\mathbb{R}^d} f(w,z)\pi_{\bar{\mathbf{z}}}(\mathrm{d}w)\right| \leq \frac{\tilde{C}_3}{n}$$

with

$$\tilde{C}_3 := 4\left(M^2\frac{b+d/\beta}{m} + B^2\right)\beta c_{\mathrm{LS}}.$$

3.7. Completing the proof

Now that we have all the ingredients in place, we can complete the proof of Theorem 1. Choose $k \in \mathbb{N}$ and $\eta \in (0, 1 \wedge \frac{m}{4M^2})$ to satisfy

$$k\eta = \beta c_{\mathrm{LS}} \log\left(\frac{1}{\varepsilon}\right) \qquad \text{and} \qquad \eta \leq \left(\frac{\varepsilon}{\log(1/\varepsilon)}\right)^4.$$

Then, by Proposition 10,

$$\mathcal{W}_2(\mu_{\mathbf{z},k},\pi_{\mathbf{z}}) \leq \tilde{C}_0\beta c_{\mathrm{LS}}\delta^{1/4}\log\left(\frac{1}{\varepsilon}\right) + \left(\tilde{C}_1\beta c_{\mathrm{LS}} + \tilde{C}_2\right)\varepsilon.$$

Now consider the random hypotheses $\widehat{W}$ and $\widehat{W}^*$ with $\mathcal{L}(\widehat{W}|\mathbf{Z}=\mathbf{z}) = \mu_{\mathbf{z},k}$ and $\mathcal{L}(\widehat{W}^*|\mathbf{Z}=\mathbf{z}) = \pi_{\mathbf{z}}$. Then

$$\begin{aligned} \mathbf{E}F(\widehat{W}) - F^* &= \mathbf{E}F(\widehat{W}) - \mathbf{E}F(\widehat{W}^*) + \mathbf{E}F(\widehat{W}^*) - F^* \\ &= \int_{\mathsf{Z}^n} \mathbf{P}^{\otimes n}(\mathrm{d}\mathbf{z})\left(\int_{\mathbb{R}^d} F(w)\mu_{k,\mathbf{z}}(\mathrm{d}w) - \int_{\mathbb{R}^d} F(w)\pi_{\mathbf{z}}(\mathrm{d}w)\right) + \mathbf{E}F(\widehat{W}^*) - F^*. \end{aligned}$$

The function F satisfies the conditions of Lemma 6 with $c_1 = M$ and $c_2 = B$, while the probability measures $\mu_{\mathbf{z},k}, \pi_{\mathbf{z}}$ satisfy the conditions of Lemma 6 with

$$\sigma^2 = \kappa_0 + 2\left(1 \vee \frac{1}{m}\right)\left(b + B^2(1+\delta) + \frac{d}{\beta}\right),$$

by Lemma 3. Therefore, for all $\mathbf{z} \in \mathsf{Z}^n$,

$$\int_{\mathbb{R}^d} F(w)\mu_{k,\mathbf{z}}(\mathrm{d}w) - \int_{\mathbb{R}^d} F(w)\pi_{\mathbf{z}}(\mathrm{d}w) \le K_0\delta^{1/4}\log\left(\frac{1}{\varepsilon}\right) + K_1\varepsilon \tag{3.22}$$

with

$$K_0 := \left(M\sqrt{\kappa_0 + 2\left(1 \vee \frac{1}{m}\right)\left(b + 2B^2 + \frac{d}{\beta}\right)} + B\right)\tilde{C}_0\beta c_{\mathrm{LS}}$$

and

$$K_1 := \left(M\sqrt{\kappa_0 + 2\left(1 \vee \frac{1}{m}\right)\left(b + 2B^2 + \frac{d}{\beta}\right)} + B\right)(\tilde{C}_1\beta c_{\mathrm{LS}} + \tilde{C}_2).$$

It remains to analyze the expected excess risk $\mathbf{E}F(\widehat{W}^*) - F^*$ of the Gibbs algorithm. To that end, we will use stability-based arguments along the lines of Bousquet and Elisseeff (2002) and Rakhlin et al. (2005). We begin by decomposing the excess risk as

$$\mathbf{E}F(\widehat{W}^*) - F^* = \underbrace{\mathbf{E}F(\widehat{W}^*) - \mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*)}_{T_1} + \underbrace{\mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*) - F^*}_{T_2}.$$

The term T_1 is the generalization error of the Gibbs algorithm. To upper-bound it, let $\mathbf{Z}' = (Z'_1, \ldots, Z'_n) \sim \mathbf{P}^{\otimes n}$ be independent of $\mathbf{Z}$ and $\widehat{W}^*$. Then

$$\begin{aligned}\mathbf{E}F(\widehat{W}^*) - \mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*) &= \mathbf{E}[F_{\mathbf{Z}'}(\widehat{W}^*) - F_{\mathbf{Z}}(\widehat{W}^*)] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}[f(\widehat{W}^*, Z'_i) - f(\widehat{W}^*, Z_i)].\end{aligned} \tag{3.23}$$

Using the fact that $Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n$ are i.i.d., as well as the fact that $\mathbf{Z}'$ is indepdenent of $\widehat{W}^*$, the ith term in the summation in (3.23) can be written out explicitly as follows:

$$\begin{aligned}&\mathbf{E}[f(\widehat{W}^*, Z'_i) - f(\widehat{W}^*, Z_i)] \\ &= \int_{\mathsf{Z}^n} \mathbf{P}^{\otimes n}(\mathrm{d}z) \int_{\mathsf{Z}} \mathbf{P}(\mathrm{d}z'_i) \int_{\mathbb{R}^d} \pi_{\mathbf{z}}(\mathrm{d}w)\left[f(w, z'_i) - f(w, z_i)\right] \\ &= \int_{\mathsf{Z}^n} \mathbf{P}^{\otimes n}(\mathrm{d}z_1, \ldots, \mathrm{d}z'_i, \ldots, \mathrm{d}z_n) \int_{\mathsf{Z}} \mathbf{P}(\mathrm{d}z_i) \int_{\mathbb{R}^d} \pi_{(z_1,\ldots,z'_i,\ldots,z_n)}(\mathrm{d}w) f(w, z_i) \\ &\qquad - \int_{\mathsf{Z}^n} \mathbf{P}^{\otimes n}(\mathrm{d}z_1, \ldots, \mathrm{d}z_i, \ldots, \mathrm{d}z_n) \int_{\mathsf{Z}} \mathbf{P}(\mathrm{d}z'_i) \int_{\mathbb{R}^d} \pi_{(z_1,\ldots,z_i,\ldots,z_n)}(\mathrm{d}w) f(w, z_i) \\ &= \int_{\mathsf{Z}^n} \mathbf{P}^{\otimes n}(\mathrm{d}\mathbf{z}) \int_{\mathsf{Z}} \mathbf{P}(\mathrm{d}z'_i) \left(\int_{\mathbb{R}^d} \pi_{\bar{\mathbf{z}}^{(i)}}(\mathrm{d}w) f(w, z_i) - \int_{\mathbb{R}^d} \pi_{\bar{\mathbf{z}}}(\mathrm{d}w) f(w, z_i)\right),\end{aligned} \tag{3.24}$$

where $\bar{\mathbf{z}}^{(i)} := (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)$. Noting that $\bar{\mathbf{z}}^{(i)}$ and $\mathbf{z}$ differ only in the ith coordinate, we can use Proposition 12 to upper-bound the integral in (3.24). Since the resulting estimate is uniform in i, from (3.23) we obtain

$$\mathbf{E}F(\widehat{W}^*) - \mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*) \leq \frac{\tilde{C}_3}{n}. \tag{3.25}$$

The term T_2 can be handled as follows: Let $w^* \in \mathbb{R}^d$ be any minimizer of $F(w)$, i.e., $F(w^*) = F^*$. Then

$$\begin{aligned}
\mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*) - F^* &= \mathbf{E}\left[F_{\mathbf{Z}}(\widehat{W}^*) - \min_{w\in\mathbb{R}^d} F_{\mathbf{Z}}(w)\right] + \mathbf{E}\left[\min_{w\in\mathbb{R}^d} F_{\mathbf{Z}}(w) - F_{\mathbf{Z}}(w^*)\right] \\
&\leq \mathbf{E}\left[F_{\mathbf{Z}}(\widehat{W}^*) - \min_{w\in\mathbb{R}^d} F_{\mathbf{Z}}(w)\right] \\
&\leq \frac{d}{2\beta}\log\left(\frac{eM}{m}\left(\frac{b\beta}{d}+1\right)\right),
\end{aligned} \tag{3.26}$$

where the last step is by Proposition 11. From (3.25) and (3.26), we get

$$\mathbf{E}F(\widehat{W}^*) - F^* = \mathbf{E}F(\widehat{W}^*) - \mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*) + \mathbf{E}F_{\mathbf{Z}}(\widehat{W}^*) - F^* \leq \frac{\tilde{C}_3}{n} + \frac{d}{2\beta}\log\left(\frac{eM}{m}\left(\frac{b\beta}{d}+1\right)\right). \tag{3.27}$$

Combining Eqs. (3.22) and (3.27), we obtain the claimed excess risk bound (2.6).

4. Discussion and directions for future research

Regularity assumptions. The first two assumptions are fairly standard in the literature on nonconvex optimization. The dissipativity assumption **(A.3)** merits some discussion. The term "dissipative" comes from the theory of dynamical systems (Hale, 1988; Stuart and Humphries, 1996), where it has the following interpretation: Consider the gradient flow described by the ordinary differential equation

$$\frac{\mathrm{d}w}{\mathrm{d}t} = -\nabla f(w, z), \qquad w(0) = w_0. \tag{4.1}$$

If f is (m, b)-dissipative, then a simple argument based on the Gronwall lemma shows that, for any $\varepsilon > 0$ and any initial condition w_0, the trajectory of (4.1) satisfies $\|w(t)\| \leq \sqrt{b/m + \varepsilon}$ for all $t \geq \frac{1}{2m}\log\frac{\|w_0\|^2}{\varepsilon}$. In other words, for any $\varepsilon > 0$, the Euclidean ball of radius $\sqrt{b/m+\varepsilon}$ centered at the origin is an *absorbing set* for the flow (4.1). If we think of $w(t)$ as the position of a particle moving in $\mathbb{R}^d$ in the presence of the potential $f(w, z)$, then the above property means that the particle rapidly loses (or dissipates) energy and stays confined in the absorbing set. However, the behavior of the flow inside this absorbing set may be arbitrarily complicated; in particular, even though (2.3) implies that all of the critical points of $w \mapsto f(w, z)$ are contained in the ball of radius $\sqrt{b/m}$ centered at the origin, there can be arbitrarily many such points. The dissipativity

assumption seems restrictive, but, in fact, it can be enforced using weight decay regularization (Krogh and Hertz, 1992). Indeed, consider the regularized objective

$$f(w,z) = f_0(w,z) + \frac{\gamma}{2}\|w\|^2.$$

Then it is not hard to show that, if the function $w \mapsto f_0(w,z)$ is L-Lipschitz, then f satisfies **(A.2)** with $m = \gamma/2$ and $b = L^2/2\gamma$. Thus, a byproduct of our analysis is a fine-grained characterization of the impact of weight decay on learning.

Assumption **(A.4)** provides control of the relative mean-square error of the stochastic gradient, viz., $\mathbf{E}\|g(w,U_{\mathbf{z}})\|^2 \preceq (1+\delta)\|\nabla F_{\mathbf{z}}(w)\|^2$, and is also easy to satisfy in practice. For example, consider the case where, at each iteration of SGLD, we sample (uniformly with replacement) a random minibatch of size ℓ. Then we can take $U_{\mathbf{z}} = (z_{I_1},\dots,z_{I_\ell})$, where $I_1,\dots,I_\ell \overset{\text{i.i.d.}}{\sim} \text{Uniform}(\{1,\dots,n\})$, and

$$g(w,U_{\mathbf{z}}) = \frac{1}{\ell}\sum_{j=1}^{\ell} \nabla f(w,z_{I_j}). \tag{4.2}$$

This gradient oracle is clearly unbiased, and a simple calculation shows that **(A.4)** holds with $\delta = 1/\ell$. On the other hand, using the full empirical gradient clearly gives $\delta = 0$.

Finally, the exponential integrability assumption **(A.5)** is satisfied, for example, by the Gaussian initialization $W_0 \sim N(0,\sigma^2 I_d)$ with $\sigma^2 < 1/2$.

Effect of gradient noise and minibatch size selection. Observe that the excess risk bound (2.6) contains a term that goes to zero as $\varepsilon \to 0$, as well as a term that grows as $\log \varepsilon^{-1}$, but goes to zero as the gradient noise level $\delta \to 0$. This suggests selecting the minibatch size

$$\ell \geq \frac{1}{\eta} \geq \left(\frac{\log(1/\varepsilon)}{\varepsilon}\right)^4.$$

to offset the $\log \varepsilon^{-1}$ term.

Uniform spectral gap. As shown in Appendix B, Assumptions **(A.1)–(A.3)** are enough to guarantee that the spectral gap λ_* is strictly positive. In particular, we give a very conservative estimate

$$\frac{1}{\lambda_*} = \tilde{\mathcal{O}}\left(\frac{1}{\beta(d+\beta)}\right) + \tilde{\mathcal{O}}\left(1+\frac{d}{\beta}\right) e^{\tilde{\mathcal{O}}(\beta+d)}. \tag{4.3}$$

Using this estimate in Eq. (2.6), we end up with a bound on the excess risk that has a dependence on $\exp(\tilde{\mathcal{O}}(\beta+d))$. This in turn suggests choosing $\varepsilon = 1/n$ and $\beta = \tilde{\mathcal{O}}(\log n)$; as a consequence, the excess risk will decay as $1/\log n$, and the number of iterations k will scale as $n^{\tilde{\mathcal{O}}(1)}\exp(\tilde{\mathcal{O}}(d))$. The alternative regime of conditionally independent stochastic gradients (e.g., using a fresh minibatch at each iteration) amounts to direct optimization of F rather than $F_{\mathbf{z}}$ and suggests the choice of $\beta \approx 1/\varepsilon$. The number of iterations k will then scale like $\exp(d+1/\varepsilon)$.

Therefore, in order to apply Theorem 1, one needs to fully exploit the structural properties of the problem at hand and produce an upper bound on $1/\lambda_*$ which is polynomial in d or even dimension-free. (By contrast, exponential dependence of $1/\lambda_*$ on β is unavoidable in the presence of multiple local minima and saddle points; this is a consequence of sharp upper and lower bounds on the

spectral gap due to Bovier et al. (2005).) For example, consider replacing the empirical risk (1.2) with a smoothed objective

$$\begin{aligned}\tilde{F}_{\mathbf{z}}(w) &= -\frac{1}{\beta}\log\int_{\{\|v\|\leq R\}} e^{-\beta\gamma\|v-w\|^2/2}e^{-\beta F_{\mathbf{z}}(v)}\mathrm{d}v \\ &= \frac{\gamma}{2}\|w\|^2 - \frac{1}{\beta}\log\int_{\{\|v\|\leq R\}} e^{\beta\gamma\langle v,w\rangle-\beta\gamma\|v\|^2/2}e^{-\beta F_{\mathbf{z}}(v)}\mathrm{d}v,\end{aligned}$$

and running SGLD with $\nabla\tilde{F}_{\mathbf{z}}$ instead of $\nabla F_{\mathbf{z}}$. Here, $\gamma > 0$ and $R > 0$ are tunable parameters. This modification is closely related to the Entropy-SGD method, recently proposed by Chaudhari et al. (2016). Observe that the modified Gibbs measures $\tilde{\pi}_{\mathbf{z}}(\mathrm{d}w) \propto e^{-\beta\tilde{F}_{\mathbf{z}}(w)}$ are convolutions of a Gaussian measure and a compactly supported probability measure. In this case, it follows from the results of Bardet et al. (2015) that

$$\frac{1}{\lambda_*} \leq \frac{1}{\beta\gamma}e^{4\beta\gamma R^2}.$$

Note that here, in contrast with (4.3), this bound is completely dimension-free. A tantalizing line of future work is, therefore, to find other settings where $1/\lambda_*$ is indeed small.

Acknowledgments

The authors would like to thank Arnak Dalalyan and Ramon van Handel for enlightening discussions. The work of M.R. was supported in part by the NSF under CAREER award CCF-1254041, and in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. The work of A.R. was supported in part by the NSF under grant no. CDS&E-MSS 1521529.

References

A. Alfonsi, B. Jourdain, and A. Kohatsu-Higa. Optimal transport bounds between the time-marginals of a multidimensional diffusion and its Euler scheme. *Electron. J. Probab.*, 20, 2015. paper no. 70.

D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case. *Electron. Comm. Probab.*, 13: 60–66, 2008.

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

J.-B. Bardet, N. Gozlan, F. Malrieu, and P.-A. Zitt. Functional inequalities for Gaussian convolutions of compactly supported measures: explicit bounds and dimension dependence, 2015. URL http://arxiv.org/abs/1507.02389. To appear in *Bernoulli*.

A. Belloni, T. Liang, H. Narayanan, and A. Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *COLT*, pages 240–265, 2015.

F. Bolley and C. Villani. Weighted Csiszár–Kullback–Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des Science de Toulouse*, XIV(3):331–352, 2005.

V. S. Borkar and S. K. Mitter. A strong approximation theorem for stochastic recursive algorithms. *Journal of Optimization Theory and Applications*, 100(3):499–513, 1999.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes II. Precise asymptotics for small eigenvalues. *J. Eur. Math. Soc.*, 7:69–99, 2005.

S. Bubeck, R. Eldan, and J. Lehec. Sampling from a log-concave distribution with Projected Langevin Monte Carlo. arXiv preprint 1507.02564, 2015. URL http://arxiv.org/abs/1507/02564.

P. Cattiaux, A. Guillin, and L. Wu. A note on Talagrand's transportation inequality and logarithmic Sobolev inequality. *Prob. Theory Rel. Fields*, 148:285–334, 2010.

P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. arXiv preprint 1611.01838, 2016. URL http://arxiv.org/abs/1611.01838.

T.-S. Chiang, C.-R. Hwang, and S.-J. Sheu. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 2nd edition, 2006.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. Roy. Stat. Soc. Ser. B*, 2016. To appear.

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte Carlo. *J. Comp. Sys. Sci.*, 78(1423-1443), 2012.

H. Djellout, A. Guillin, and L. Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Annals of Probability*, 32(3B):2702–2732, 2004.

A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the unadjusted langevin algorithm. arXiv preprint 1507.05021, 2015. URL http://arxiv.org/abs/1507.05021.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842, 2015.

S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

I. Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an Ito differential. *Prob. Theory Rel. Fields*, 71:501–516, 1986.

J. K. Hale. *Asymptotic Behavior of Dissipative Systems*. American Mathematical Society, 1988.

M Hardt, B Recht, and Y Singer. Train faster, generalize better: Stability of stochastic gradient descent. arXiv preprint 1509.01240, 2015. URL http://arxiv.org/abs/1509.01240.

E. Hazan, K. Levi, and S. Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *ICML*, 2016.

C.-R. Hwang. Laplace's method revisited: weak convergence of probability measures. *Annals of Probability*, 8(1177-1182), 1980.

A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 950–957. 1992.

R. S. Liptser and A. N. Shiryaev. *Statistics of Random Processes I: General Theory*. Springer, 2nd edition, 2001.

D. Márquez. Convergence rates for annealing diffusion processes. *The Annals of Applied Probability*, pages 1118–1139, 1997.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer, 2004.

M. Pelletier. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Annals of Applied Probability*, pages 10–44, 1998.

Y. Polyanskiy and Y. Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Trans. Inf. Theory*, 62(7):3992–4002, July 2016.

A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

A. M. Stuart and A. R. Humphries. *Dynamical Systems and Numerical Analysis*. Cambridge University Press, 1996.

C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. Amer. Math. Soc., Providence, RI, 2003.

M. Welling and Y. W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, pages 681–688, 2011.

Appendix A. Background on Markov semigroups and functional inequalities

Our analysis relies on the theory of Markov diffusion operators and associated functional inequalities. In this Appendix, we only summarize the key ideas and results; the book by Bakry et al. (2014) provides an in-depth exposition.

Let $\{W(t)\}_{t\geq 0}$ be a continuous-time homogeneous Markov process with values in $\mathbb{R}^d$, and let $P = \{P_t\}_{t\geq 0}$ be the corresponding Markov semigroup, i.e.,

$$P_s g(W(t)) = \mathbf{E}[g(W(s+t))|W(t)]$$

for all $s, t \geq 0$ and all bounded measurable functions $g : \mathbb{R}^d \to \mathbb{R}$. (The semigroup law $P_s \circ P_t = P_{s+t}$ is just another way to express the Markov property.) A Borel probability measure π is called *stationary* or *invariant* if $\int_{\mathbb{R}^d} P_t g \, \mathrm{d}\pi = \int_{\mathbb{R}^d} g \, \mathrm{d}\pi$ for all g and t. Each P_t can be extended to a bounded linear operator on $L^2(\pi)$, such that $P_t g \geq 0$ whenever $g \geq 0$ and $P_t 1 = 1$ for all t. The *generator* of the semigroup is a linear operator $\mathcal{L}$ defined on a dense subspace $\mathcal{D}(\mathcal{L})$ of $L^2(\pi)$ (the *domain* of $\mathcal{L}$), such that, for any $g \in \mathcal{D}(\mathcal{L})$,

$$\partial_t P_t g = \mathcal{L} P_t g = P_t \mathcal{L} g.$$

In particular, $\mathcal{L}1 = 0$, and π is an invariant probability measure of the semigroup if and only if $\int_{\mathbb{R}^d} \mathcal{L} g \, \mathrm{d}\pi = 0$ for all $g \in \mathcal{D}(\mathcal{L})$. The generator $\mathcal{L}$ defines the *Dirichlet form*

$$\mathcal{E}(g) := -\int_{\mathbb{R}^d} g \mathcal{L} g \, \mathrm{d}\pi. \tag{A.1}$$

It can be shown that $\mathcal{E}(g) \geq 0$, i.e., $-\mathcal{L}$ is a positive operator (since $\mathcal{L}1 = 0$, zero is an eigenvalue).

Let P be a Markov semigroup with the unique invariant distribution π and the Dirichlet form $\mathcal{E}$. We say that π satisfies a *Poincaré* (or *spectral gap*) *inequality* with constant c if, for all probability measures $\mu \ll \pi$,

$$\chi^2(\mu\|\pi) \leq c\,\mathcal{E}\left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\right), \tag{A.2}$$

where $\chi^2(\mu\|\pi) := \|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\|^2_{L^2(\pi)}$ is the χ^2 divergence between μ and π. The name "spectral gap" comes from the fact that, if (A.2) holds with some constant c, then $1/c \geq \lambda$, where

$$\begin{aligned}\lambda &:= \inf\left\{\frac{\mathcal{E}(g)}{\int_{\mathbb{R}^d} g^2 \mathrm{d}\pi} : g \in C^2,\, g \neq 0,\, \int_{\mathbb{R}^d} g = 0\right\} \\ &= \inf\left\{\frac{-\langle g, \mathcal{L} g\rangle_{L^2(\pi)}}{\|g\|^2_{L^2(\pi)}} : g \in C^2,\, g \neq 0,\, \int_{\mathbb{R}^d} g = 0\right\}.\end{aligned}$$

Hence, if $\lambda > 0$, then the spectrum of $-\mathcal{L}$ is contained in the set $\{0\} \cup [\lambda, \infty)$, so λ is the gap between the zero eigenvalue and the rest of the spectrum. We say that π satisfies a *logarithmic Sobolev inequality* with constant c if, for all $\mu \ll \pi$,

$$D(\mu\|\pi) \leq 2c\,\mathcal{E}\left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\right), \tag{A.3}$$

where $D(\mu\|\pi) = \int \mathrm{d}\mu \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}$ is the relative entropy (Kullback–Leibler divergence). We record a couple of key consequences of the logarithmic Sobolev inequality. Consider a Markov process $\{W(t)\}_{t\geq 0}$ with a unique invariant distribution π and a Dirichlet form $\mathcal{E}$, such that π satisfies a logarithmic Sobolev inequality with constant c. Then we have the following:

1. Exponential decay of entropy (Bakry et al., 2014, Th. 5.2.1): Let $\mu_t := \mathcal{L}(W(t))$. Then
$$D(\mu_t\|\pi) \leq D(\mu_0\|\pi)e^{-2t/c}. \tag{A.4}$$

2. Otto–Villani theorem (Bakry et al., 2014, Th. 9.6.1): If $\mathcal{E}(g) = \alpha \int \|\nabla g\|^2 \mathrm{d}\pi$ for some $\alpha > 0$, then, for any $\mu \ll \pi$,
$$\mathcal{W}_2(\mu, \pi) \leq \sqrt{2c\alpha D(\mu\|\pi)}. \tag{A.5}$$

Our analysis of SGLD revolves around Markov diffusion processes, so we particularize the above abstract framework to this concrete setting. Let $\{W(t)\}_{t\geq 0}$ be a Markov process evolving in $\mathbb{R}^d$ according to an Itô SDE
$$\mathrm{d}W(t) = -\nabla H(W(t))\mathrm{d}t + \sqrt{2}\,\mathrm{d}B(t), \qquad t \geq 0 \tag{A.6}$$
where H is a C^1 function and $\{B(t)\}$ is the standard d-dimensional Brownian motion. (Replacing the factor $\sqrt{2}$ by $\sqrt{2\beta^{-1}}$ is equivalent to the time rescaling $t \mapsto \beta t$.) The generator of this semigroup is the second-order differential operator
$$\mathcal{L}g := \Delta g - \langle \nabla H, \nabla g\rangle \tag{A.7}$$
for all C^2 functions g, where $\Delta := \nabla \cdot \nabla$ is the Laplace operator. If the map $w \mapsto \nabla H(w)$ is Lipschitz, then the Gibbs measure $\pi(\mathrm{d}w) \propto e^{-H(w)}\mathrm{d}w$ is the unique invariant measure of the underlying Markov semigroup, and a simple argument using integration by parts shows that the Dirichlet form is given by
$$\mathcal{E}(g) = \int_{\mathbb{R}^d} \|\nabla g\|^2 \mathrm{d}\pi. \tag{A.8}$$
Thus, the Gibbs measure π satisfies a Poincaré inequality with constant c if, for any $\mu \ll \pi$,
$$\chi^2(\mu\|\pi) \leq c \int_{\mathbb{R}^d} \left\|\nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\right\|^2 \mathrm{d}\pi \tag{A.9}$$
and a logarithmic Sobolev inequality with constant c if
$$D(\mu\|\pi) \leq 2c \int_{\mathbb{R}^d} \left\|\nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\right\|^2 \mathrm{d}\pi. \tag{A.10}$$
If H is C^2 and strongly convex, i.e., $\nabla^2 H \succeq K I_d$ for some $K > 0$, then π satisfies a logarithmic Sobolev inequality with constant $c = 1/K$. In the absence of convexity, it is in general difficult to obtain upper bounds on Poincaré or log-Sobolev constants. The following two propositions give sufficient conditions based on so-called Lyapunov function criteria:

Proposition 13 (Bakry et al. (2008)) *Suppose that there exist constants $\kappa_0, \lambda_0 > 0, R \geq 0$ and a C^2 function $V : \mathbb{R}^d \to [1, \infty)$ such that*

$$\frac{\mathcal{L}V(w)}{V(w)} \leq -\lambda_0 + \kappa_0 \mathbf{1}\{\|w\| \leq R\}. \tag{A.11}$$

Then π satisfies a Poincaré inequality with constant

$$c_{\mathrm{P}} \leq \frac{1}{\lambda_0}\left(1 + C\kappa_0 R^2 e^{\mathrm{Osc}_R(H)}\right), \tag{A.12}$$

where $C > 0$ is a universal constant and $\mathrm{Osc}_R(H) := \max_{\|w\| \leq R} H(w) - \min_{\|w\| \leq R} H(w)$.

Remark 14 The term involving $\mathrm{Osc}_R(H)$ in (A.12) arises from a (very crude) estimate of the Poincaré constant of the truncated Gibbs measure $\pi_R(\mathrm{d}w) \propto e^{-H(w)}\mathbf{1}\{\|w\| \leq R\}\mathrm{d}w$, cf. the discussion preceding the statement of Theorem 1.4 in Bakry et al. (2008).

Proposition 15 (Cattiaux et al. (2010)) *Suppose the following conditions hold:*

1. *There exist constants $\kappa, \gamma > 0$ and a C^2 function $V : \mathbb{R}^d \to [1, \infty)$ such that*

$$\frac{\mathcal{L}V(w)}{V(w)} \leq \kappa - \gamma\|w\|^2 \tag{A.13}$$

   for all $w \in \mathbb{R}^d$.

2. *π satisfies a Poincaré inequality with constant c_{P}.*

3. *There exists some constant $K \geq 0$, such that $\nabla^2 H \succeq -K I_d$.*

Let C_1 and C_2 be defined, for some $\varepsilon > 0$, by

$$C_1 = \frac{2}{\gamma}\left(\frac{1}{\varepsilon} + \frac{K}{2}\right) + \varepsilon \quad \textit{and} \quad C_2 = \frac{2}{\gamma}\left(\frac{1}{\varepsilon} + \frac{K}{2}\right)\left(\kappa + \gamma\int_{\mathbb{R}^d} \|w\|^2 \pi(\mathrm{d}w)\right).$$

Then π satisfies a logarithmic Sobolev inequality with constant $c_{\mathrm{LS}} = C_1 + (C_2 + 2)c_{\mathrm{P}}$.

Remark 16 *In particular, if $K \neq 0$, we can take $\varepsilon = 2/K$, in which case*

$$C_1 = \frac{2K}{\gamma} + \frac{2}{K} \quad \textit{and} \quad C_2 = \frac{2K}{\lambda}\left(\kappa + \gamma\int_{\mathbb{R}^d} \|w\|^2 \mathrm{d}\pi\right). \tag{A.14}$$

Appendix B. A lower bound on the uniform spectral gap

Our goal here is to prove the crude lower bound on λ_* given in Section 4. To that end, we will use the Lyapunov function criterion due to Bakry et al. (2008), which is reproduced as Proposition 13 in Appendix A.

We will apply this criterion to the Gibbs distribution $\pi_{\mathbf{z}}$ for some $\mathbf{z} \in \mathsf{Z}^n$. Thus, we have $H = \beta F_{\mathbf{z}}$ and

$$\mathcal{L}g = \Delta g - \beta\langle \nabla F_{\mathbf{z}}, \nabla g\rangle.$$

Consider the candidate Lyapunov function $V(w) = e^{m\beta\|w\|^2/4}$. From the fact that $V \geq 1$ and from the dissipativity assumption **(A.3)**, it follows that

$$\begin{aligned}\mathcal{L}V(w) &= \left(\frac{m\beta d}{2} + \frac{(m\beta)^2}{4}\|w\|^2 - \frac{m\beta^2}{2}\langle w, \nabla F_{\mathbf{z}}(w)\rangle\right) V(w) \\ &\leq \left(\frac{m\beta(d+b\beta)}{2} - \frac{(m\beta)^2}{4}\|w\|^2\right) V(w). \end{aligned} \tag{B.1}$$

Thus, V evidently satisfies (A.11) with $R^2 = \frac{2\kappa}{\gamma}$, $\kappa_0 = \kappa$ and $\lambda_0 = 2\kappa$, where

$$\kappa := \frac{m\beta(d+b\beta)}{2} \qquad \text{and} \qquad \gamma := \frac{(m\beta)^2}{4}. \tag{B.2}$$

Moreover, from Lemma 2 and from the fact that $F_{\mathbf{z}} \geq 0$, it follows that

$$\mathrm{Osc}_R(\beta F_{\mathbf{z}}) \leq \beta\left(\frac{MR^2}{2} + BR + A\right) \leq \beta\left(\frac{(M+B)R^2}{2} + A + B\right).$$

Thus, by Proposition 13, $\pi_{\mathbf{z}}$ satisfies a Poincaré inequality with constant

$$c_{\mathrm{P}} \leq \frac{1}{m\beta(d+b\beta)} + \frac{2C(d+b\beta)}{m\beta}\exp\left(\frac{2}{m}(M+B)(b\beta+d) + \beta(A+B)\right).$$

Observe that this bound holds for all $\mathbf{z} \in \mathsf{Z}^n$. Using this fact and the relation $1/\lambda \leq c_{\mathrm{P}}$ between the spectral gap and the Poincaré constant, we see that

$$\frac{1}{\lambda_*} \leq \frac{1}{m\beta(d+b\beta)} + \frac{2C(d+b\beta)}{m\beta}\exp\left(\frac{2}{m}(M+B)(b\beta+d) + \beta(A+B)\right),$$

which proves the claimed bound.

Appendix C. Proofs for Section 3.2

Proof [Proof of Lemma 2] The estimate (3.6) is an easy consequence of conditions **(A.1)** and **(A.2)**. Next, observe that, for any two $v, w \in \mathbb{R}^d$,

$$f(w,z) - f(v,z) = \int_0^1 \langle w - v, \nabla f(tw + (1-t)v, z)\rangle \mathrm{d}t. \tag{C.1}$$

In particular, taking $v = 0$, we obtain

$$\begin{aligned} f(w,z) &= f(0,z) + \int_0^1 \langle w, \nabla f(tw)\rangle \mathrm{d}t \\ &\overset{\text{(i)}}{\leq} A + \int_0^1 \|w\|\,\|\nabla f(tw,z)\|\,\mathrm{d}t \\ &\overset{\text{(ii)}}{\leq} A + \|w\|\int_0^1 \left(Mt\|w\| + B\right)\mathrm{d}t \\ &= A + \frac{M}{2}\|w\|^2 + B\|w\|, \end{aligned}$$

where (i) follows from **(A.1)** and from Cauchy–Schwarz, while (ii) follows from (3.6). This proves the upper bound on $f(w,z)$. Now take $v = cw$ for some $c \in (0,1]$ to be chosen later. With this choice, we proceed from Eq. (C.1) as follows:

$$\begin{aligned} f(w,z) &= f(cw,z) + \int_c^1 \langle w, \nabla f(tw,z)\rangle \mathrm{d}t \\ &\overset{\text{(i)}}{\geq} \int_c^1 \frac{1}{t}\langle tw, \nabla f(tw,z)\rangle \mathrm{d}t \\ &\overset{\text{(ii)}}{\geq} \int_c^1 \frac{1}{t}\left(mt^2\|w\|^2 - b\right)\mathrm{d}t \\ &= \frac{m(1-c^2)}{2}\|w\|^2 + b\log c, \end{aligned}$$

where (i) uses the fact that $f \geq 0$, while (ii) uses the dissipativity property (2.3). Taking $c = \frac{1}{\sqrt{3}}$, we get the lower bound in (3.7). ∎

Proof [Proof of Lemma 3] From (2.2), it follows that

$$\begin{aligned} \mathbf{E}_{\mathbf{z}}\|W_{k+1}\|^2 &= \mathbf{E}_{\mathbf{z}}\|W_k - \eta g(W_k, U_{\mathbf{z},k})\|^2 + \sqrt{\frac{8\eta}{\beta}}\mathbf{E}_{\mathbf{z}}\langle W_k - \eta g(W_k, U_{\mathbf{z},k}), \xi_k\rangle + \frac{2\eta}{\beta}\mathbf{E}_{\mathbf{z}}\|\xi_k\|^2 \\ &= \mathbf{E}_{\mathbf{z}}\|W_k - \eta g(W_k, U_{\mathbf{z},k})\|^2 + \frac{2\eta d}{\beta}, \end{aligned} \tag{C.2}$$

where the second step uses independence of $W_k - g(W_k, U_{\mathbf{z},k})$ and ξ_k and the unbiasedness property (2.1) of the gradient oracle. We can further expand the first term in (C.2):

$$\begin{aligned} &\mathbf{E}_{\mathbf{z}}\|W_k - \eta g(W_k, U_{\mathbf{z},k})\|^2 \\ &= \mathbf{E}_{\mathbf{z}}\left\|W_k - \eta\nabla F_{\mathbf{z}}(W_k)\right\|^2 + 2\eta\mathbf{E}_{\mathbf{z}}\langle W_k - \eta\nabla F_{\mathbf{z}}(W_k), \nabla F_{\mathbf{z}}(W_k) - g(W_k, U_{\mathbf{z},k})\rangle \\ &\qquad + \eta^2\mathbf{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(W_k) - g(W_k, U_{\mathbf{z},k})\|^2 \\ &= \mathbf{E}_{\mathbf{z}}\|W_k - \eta\nabla F_{\mathbf{z}}(W_k)\|^2 + \eta^2\mathbf{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(W_k) - g(W_k, U_{\mathbf{z},k})\|^2, \end{aligned} \tag{C.3}$$

where we have used (2.1) once again. By (2.4), the second term in (C.3) can be upper-bounded by

$$\mathbf{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(W_k) - g(W_k, U_{\mathbf{z},k})\|^2 \leq 2\delta(M^2\mathbf{E}_{\mathbf{z}}\|W_k\|^2 + B^2),$$

whereas the first term can be estimated as

$$\begin{aligned} \mathbf{E}_{\mathbf{z}}\|W_k - \eta\nabla F_{\mathbf{z}}(W_k)\|^2 &= \mathbf{E}_{\mathbf{z}}\|W_k\|^2 - 2\eta\mathbf{E}_{\mathbf{z}}\langle W_k, \nabla F_{\mathbf{z}}(W_k)\rangle + \eta^2\mathbf{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(W_k)\|^2 \\ &\leq \mathbf{E}_{\mathbf{z}}\|W_k\|^2 + 2\eta(b - m\mathbf{E}_{\mathbf{z}}\|W_k\|^2) + 2\eta^2(M^2\mathbf{E}_{\mathbf{z}}\|W_k\|^2 + B^2) \\ &= \left(1 - 2\eta m + 2\eta^2 M^2\right)\mathbf{E}_{\mathbf{z}}\|W_k\|^2 + 2\eta b + 2\eta^2 B^2, \end{aligned}$$

where the inequality follows from the dissipativity condition (2.3) and the bound (3.6) in Lemma 2. Combining all of the above, we arrive at the recursion

$$\mathbf{E}_{\mathbf{z}}\|W_{k+1}\|^2 \leq (1 - 2\eta m + 4\eta^2 M^2)\mathbf{E}_{\mathbf{z}}\|W_k\|^2 + 2\eta b + 4\eta^2 B^2 + \frac{2\eta d}{\beta}. \tag{C.4}$$

Fix some $\eta \in (0, 1 \wedge \frac{m}{2M^2})$. There are two cases to consider:

- If $1 - 2\eta m + 4\eta^2 M^2 \leq 0$, then from (C.4) it follows that

$$\begin{aligned}\mathbf{E}_{\mathbf{z}}\|W_{k+1}\|^2 &\leq 2\eta b + 4\eta^2 B^2 + \frac{2\eta d}{\beta} \\ &\leq \mathbf{E}_{\mathbf{z}}\|W_0\|^2 + 2\left(b + 2B^2 + \frac{d}{\beta}\right). \end{aligned} \tag{C.5}$$

- If $0 < 1 - 2\eta m + 4\eta^2 M^2 < 1$, then iterating (C.4) gives

$$\begin{aligned}\mathbf{E}_{\mathbf{z}}\|W_k\|^2 &\leq (1 - 2\eta m + 4\eta^2 M^2)^k \mathbf{E}_{\mathbf{z}}\|W_0\|^2 + \frac{\eta b + 2\eta^2 B^2 + \frac{\eta d}{\beta}}{\eta m - 2\eta^2 M^2} \\ &\leq \mathbf{E}_{\mathbf{z}}\|W_0\|^2 + \frac{2}{m}\left(b + 2B^2 + \frac{d}{\beta}\right). \end{aligned} \tag{C.6}$$

The bound (3.8) follows from Eqs. (C.5) and (C.6) and from the estimate

$$\mathbf{E}_{\mathbf{z}}\|W_0\|^2 = \mathbf{E}\|W_0\|^2 \leq \log \mathbf{E} e^{\|W_0\|^2} = \kappa_0, \tag{C.7}$$

which easily follows from the independence of $\mathbf{Z}$ and W_0 and from Jensen's inequality.

We now analyze the diffusion (1.4). Let $Y(t) := \|W(t)\|^2$. Then Itô's lemma gives

$$\mathrm{d}Y(t) = -2\langle W(t), \nabla F_{\mathbf{z}}(W(t))\rangle \mathrm{d}t + \frac{2d}{\beta}\mathrm{d}t + \sqrt{\frac{8}{\beta}} W(t)^* \mathrm{d}B(t),$$

where $W(t)^*\mathrm{d}B(t) := \sum_{i=1}^d W_i(t)\mathrm{d}B_i(t)$. This can be rewritten as

$$\begin{aligned}&2me^{2mt}Y(t)\mathrm{d}t + e^{2mt}\mathrm{d}Y(t) \\ &= -2e^{2mt}\langle W(t), \nabla F_{\mathbf{z}}(W(t))\rangle \mathrm{d}t + 2me^{2mt}Y(t)\mathrm{d}t + \frac{2d}{\beta}e^{2mt}\mathrm{d}t + \sqrt{\frac{8}{\beta}}e^{2mt}W(t)^*\mathrm{d}B(t). \end{aligned} \tag{C.8}$$

Recognizing the left-hand side of (C.8) as the total Itô derivative of $e^{2mt}Y(t)$, we arrive at

$$\begin{aligned}\mathrm{d}\big(e^{2mt}Y(t)\big) = &-2e^{2mt}\langle W(t), \nabla F_{\mathbf{z}}(W(t))\rangle \mathrm{d}t + 2me^{2mt}Y(t)\mathrm{d}t \\ &+ \frac{2d}{\beta}e^{2mt}\mathrm{d}t + \sqrt{\frac{8}{\beta}}e^{2mt}W(t)^*\mathrm{d}B(t), \end{aligned} \tag{C.9}$$

which, upon integrating and rearranging, becomes

$$\begin{aligned}Y(t) = e^{-2mt}Y(0) &- 2\int_0^t e^{2m(s-t)}\langle W(s), \nabla F_{\mathbf{z}}(W(s))\rangle \mathrm{d}s \\ &+ 2m\int_0^t e^{2m(s-t)}Y(s)\mathrm{d}s + \frac{d}{m\beta}\left(1 - e^{-2mt}\right) + \sqrt{\frac{8}{\beta}}\int_0^t e^{2m(s-t)}W(s)^*\mathrm{d}B(s). \end{aligned} \tag{C.10}$$

Now, using the dissipativity condition (2.3), we can write

$$\begin{aligned}-2\int_0^t e^{2m(s-t)}\langle W(s),\nabla F_{\mathbf{z}}(W(s))\rangle \mathrm{d}s &\leq 2\int_0^t e^{2m(s-t)}\left(b-mY(s)\right)\mathrm{d}s\\ &= 2b\int_0^t e^{2m(s-t)}\mathrm{d}s - 2m\int_0^t e^{2m(s-t)}Y(s)\mathrm{d}s\\ &= \frac{b}{m}\left(1-e^{-2mt}\right) - 2m\int_0^t e^{2m(s-t)}Y(s)\mathrm{d}s.\end{aligned}$$

Substituting this into (C.10), we end up with

$$\|W(t)\|^2 \leq e^{-2mt}\|W(0)\|^2 + \frac{b+d/\beta}{m}\left(1-e^{-2mt}\right) + \sqrt{\frac{8}{\beta}}\int_0^t e^{2m(s-t)}W(s)^*\mathrm{d}B(s).$$

Taking expectations and using the martingale property of the Itô integral together with (C.7), we get (3.9). Eq. (3.10) follows from maximizing the right-hand side of (3.9) over all $t \geq 0$. ■

Proof [Proof of Lemma 4] For $L(t) = e^{\|W(t)\|^2}$, Itô's lemma gives

$$\mathrm{d}L(t) = -2\langle W(t),\nabla F_{\mathbf{z}}(W(t))\rangle L(t)\mathrm{d}t + \frac{4}{\beta}L(t)\|W(t)\|^2\mathrm{d}t + \frac{2d}{\beta}L(t)\mathrm{d}t + \sqrt{\frac{8}{\beta}}L(t)W(t)^*\mathrm{d}B(t).$$

Integrating, we obtain

$$\begin{aligned}L(t) = L(0) &+ \int_0^t\left(\frac{4}{\beta}\|W(s)\|^2 - 2\langle W(s),\nabla F_{\mathbf{z}}(W(s))\rangle\right)L(s)\mathrm{d}s\\ &+ \frac{2d}{\beta}\int_0^t L(s)\mathrm{d}s + \sqrt{\frac{8}{\beta}}\int_0^t L(s)W(s)^*\mathrm{d}B(s).\end{aligned}$$

From the dissipativity condition (2.3) and from the assumption that $\beta \geq 2/m$, it follows that

$$\frac{4}{\beta}\|W(s)\|^2 - 2\langle W(s),\nabla F_{\mathbf{z}}(W(s))\rangle \leq 2b + \left(\frac{4}{\beta}-2m\right)\|W(s)\|^2 \leq 2b,$$

hence

$$L(t) \leq L(0) + 2\left(b+\frac{d}{\beta}\right)\int_0^t L(s)\mathrm{d}s + \sqrt{\frac{8}{\beta}}\int_0^t L(s)W(s)^*\mathrm{d}B(s).$$

It can be shown (see, e.g., the proof of Corollary 4.1 in Djellout et al. (2004)) that $\int_0^T \mathbf{E}\|L(t)W(t)\|^2\mathrm{d}t < \infty$ for all $T \geq 0$. Therefore, the Itô integral $\int L(s)W(s)^*\mathrm{d}B(s)$ is a zero-mean martingale, so, taking expecations, we get

$$\begin{aligned}\mathbf{E}[L(t)] &\leq \mathbf{E}[L(0)] + 2\left(b+\frac{d}{\beta}\right)\int_0^t \mathbf{E}[L(s]\mathrm{d}s\\ &= e^{\kappa_0} + 2\left(b+\frac{d}{\beta}\right)\int_0^t \mathbf{E}[L(s)]\mathrm{d}s.\end{aligned}$$

Eq. (3.11) then follows by an application of the Gronwall lemma. ■

Proof [Proof of Lemma 5] Let $p_{\mathbf{z}}$ denote the density of $\pi_{\mathbf{z}}$ with respect to the Lebesgue measure on $\mathbb{R}^d$:

$$p_{\mathbf{z}}(w) = \frac{e^{-\beta F_{\mathbf{z}}(w)}}{\Lambda_{\mathbf{z}}}, \qquad \text{where } \Lambda_{\mathbf{z}} = \int_{\mathbb{R}^d} e^{-\beta F_{\mathbf{z}}(w)} \mathrm{d}w.$$

Since $p_{\mathbf{z}} > 0$ everywhere, we can write

$$\begin{aligned} D(\mu_0 \| \pi_{\mathbf{z}}) &= \int_{\mathbb{R}^d} p_0(w) \log \frac{p_0(w)}{p_{\mathbf{z}}(w)} \mathrm{d}w \\ &= \int_{\mathbb{R}^d} p_0(w) \log p_0(w) \mathrm{d}w + \log \Lambda_{\mathbf{z}} + \beta \int_{\mathbb{R}^d} p_0(w) F_{\mathbf{z}}(w) \mathrm{d}w \\ &\leq \log \|p_0\|_\infty + \log \Lambda_{\mathbf{z}} + \beta \int_{\mathbb{R}^d} p_0(w) F_{\mathbf{z}}(w) \mathrm{d}w. \end{aligned} \tag{C.11}$$

We first upper-bound the partition function:

$$\begin{aligned} \Lambda_{\mathbf{z}} &= \int_{\mathbb{R}^d} e^{-\beta F_{\mathbf{z}}(w)} \mathrm{d}w \\ &= \int_{\mathbb{R}^d} \exp\left(-\frac{\beta}{n} \sum_{i=1}^n f(w, z_i) \right) \mathrm{d}w \\ &\leq e^{\frac{1}{2}\beta b \log 3} \int_{\mathbb{R}^d} e^{-\frac{m\beta \|w\|^2}{3}} \mathrm{d}w \\ &= 3^{\beta b/2} \left(\frac{3\pi}{m\beta} \right)^{d/2}, \end{aligned}$$

where the inequality follows from Lemma 2. Thus,

$$\log \Lambda_{\mathbf{z}} \leq \frac{d}{2} \log \frac{3\pi}{m\beta} + \frac{\beta b}{2} \log 3. \tag{C.12}$$

Moreover, invoking Lemma 2 once again, we have

$$F_{\mathbf{z}}(w) = \frac{1}{n} \sum_{i=1}^n f(w, z_i) \leq \frac{M}{3} \|w\|^2 + B\|w\| + A. \tag{C.13}$$

Therefore,

$$\begin{aligned} \int_{\mathbb{R}^d} F_{\mathbf{z}}(w) p_0(w) \mathrm{d}w &\leq \int_{\mathbb{R}^d} \mu_0(\mathrm{d}w) \left(\frac{M}{3} \|w\|^2 + B\|w\| + A \right) \\ &\leq \frac{M}{3} \kappa_0 + B\sqrt{\kappa_0} + A. \end{aligned} \tag{C.14}$$

Substituting (C.12), (C.13), and (C.14) into (C.11), we get (3.12). ■

Proof [Proof of Lemma 6] The proof is a minor tweak of the proof of Proposition 1 in Polyanskiy and Wu (2016); we reproduce it here to keep the presentation self-contained. Without loss of generality, we assume that $\sigma^2 < \infty$, otherwise the bound holds trivially. For any two $v, w \in \mathbb{R}^d$, we have

$$\begin{aligned} g(w) - g(v) &= \int_0^1 \langle w - v, \nabla g((1-t)v + tw)\rangle \mathrm{d}t \\ &\leq \int_0^1 \|\nabla g((1-t)v + tw)\| \, \|w - v\| \, \mathrm{d}t \\ &\leq \int_0^1 \left(c_1(1-t)\|v\| + c_1 t\|w\| + c_2\right) \|w - v\| \, \mathrm{d}t \\ &= \left(\frac{c_1}{2}\|v\| + \frac{c_1}{2}\|w\| + c_2\right) \|w - v\|, \end{aligned} \tag{C.15}$$

where we have used Cauchy–Schwarz and the growth condition (3.13). Now let $\mathbf{P}$ be the coupling of μ and ν that achieves $\mathcal{W}_2(\mu, \nu)$. That is, $\mathbf{P} = \mathcal{L}((W, V))$ with $\mu = \mathcal{L}(W)$, $\nu = \mathcal{L}(V)$, and

$$\mathcal{W}_2^2(\mu, \nu) = \mathbf{E}_{\mathbf{P}}\|W - V\|^2.$$

Taking expectations in (C.15), we have

$$\begin{aligned} \int_{\mathbb{R}^d} g\mathrm{d}\mu - \int_{\mathbb{R}^d} g\mathrm{d}\nu &= \mathbf{E}_{\mathbf{P}}[g(W) - g(V)] \\ &\leq \sqrt{\mathbf{E}_{\mathbf{P}}\left(\frac{c_1}{2}\|W\| + \frac{c_1}{2}\|V\| + c_2\right)^2} \cdot \sqrt{\mathbf{E}_{\mathbf{P}}[\|W - V\|^2]} \\ &\leq \left(\frac{c_1}{2}\sqrt{\mathbf{E}\|W\|^2} + \frac{c_1}{2}\sqrt{\mathbf{E}\|V\|^2} + c_2\right) \cdot \mathcal{W}_2(\mu, \nu) \\ &= (c_1\sigma + c_2)\, \mathcal{W}_2(\mu, \nu). \end{aligned}$$

Interchanging the roles of μ and ν, we complete the proof. ■

Appendix D. Proof of Lemma 7

Conditioned on $\mathbf{Z} = \mathbf{z}$, $\{W_k\}_{k=0}^{\infty}$ is a time-homogeneous Markov process. Consider the following continuous-time interpolation of this process:

$$\overline{W}(t) = W_0 - \int_0^t g(\overline{W}(\lfloor s/\eta \rfloor \eta), \overline{U}_{\mathbf{z}}(s))\mathrm{d}s + \sqrt{\frac{2}{\beta}}\int_0^t \mathrm{d}B(s), \qquad t \geq 0 \tag{D.1}$$

where $\overline{U}_{\mathbf{z}}(t) \equiv U_{\mathbf{z},k}$ for $t \in [k\eta, (k+1)\eta)$. Note that, for each k, $\overline{W}(k\eta)$ and W_k have the same probability law $\mu_{\mathbf{z},k}$. Moreover, by a result of Gyöngy (1986), the process $\overline{W}(t)$ has the same one-time marginals as the Itô process

$$V(t) = W_0 - \int_0^t g_{\mathbf{z},s}(V(s))\mathrm{d}s + \sqrt{\frac{2}{\beta}}\int_0^t \mathrm{d}B(s)$$

with

$$g_{\mathbf{z},t}(v) := \mathbf{E}_{\mathbf{z}}\Big[g(\overline{W}(\lfloor t/\eta \rfloor \eta), \overline{U}_{\mathbf{z}}(t))\Big|\overline{W}(t) = v\Big]. \tag{D.2}$$

Crucially, $V(t)$ is a Markov process, while $\overline{W}(t)$ is not. Let $\mathbf{P}_V^t := \mathcal{L}\big(V(s) : 0 \le s \le t\big|\mathbf{Z} = \mathbf{z}\big)$ and $\mathbf{P}_W^t := \mathcal{L}\big(W(s) : 0 \le s \le t\big|\mathbf{Z} = \mathbf{z}\big)$. The Radon–Nikodym derivative of $\mathbf{P}_W^t$ w.r.t. $\mathbf{P}_V^t$ is given by the Girsanov formula

$$\frac{\mathrm{d}\mathbf{P}_W^t}{\mathrm{d}\mathbf{P}_V^t}(V) = \exp\left\{\frac{\beta}{2}\int_0^t \big(\nabla F_{\mathbf{z}}(V(s)) - g_{\mathbf{z},s}(V(s))\big)^* \,\mathrm{d}B(s) - \frac{\beta}{4}\int_0^t \|\nabla F_{\mathbf{z}}(V(s)) - g_{\mathbf{z},s}(V(s))\|^2 \mathrm{d}s\right\} \tag{D.3}$$

(see, e.g., Sec. 7.6.4 in Liptser and Shiryaev (2001)). Using (D.3) and the martingale property of the Itô integral, we have

$$\begin{aligned} D(\mathbf{P}_V^t\|\mathbf{P}_W^t) &= -\int \mathrm{d}\mathbf{P}_V^t \log \frac{\mathrm{d}\mathbf{P}_W^t}{\mathrm{d}\mathbf{P}_V^t} \\ &= \frac{\beta}{4}\int_0^t \mathbf{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(V(s)) - g_{\mathbf{z},s}(V(s))\right\|^2 \mathrm{d}s \\ &= \frac{\beta}{4}\int_0^t \mathbf{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\overline{W}(s)) - g_{\mathbf{z},s}(\overline{W}(s))\right\|^2 \mathrm{d}s, \end{aligned}$$

where the last line follows from the fact that $\mathcal{L}(\overline{W}(s)) = \mathcal{L}(V(s))$ for each s.

Now let $t = k\eta$ for some $k \in \mathbb{N}$. Then, using the definition (D.2) of $g_{\mathbf{z},s}$, Jensen's inequality, and the M-smoothness of $F_{\mathbf{z}}$, we can write

$$\begin{aligned} D(\mathbf{P}_V^{k\eta}\|\mathbf{P}_W^{k\eta}) &= \frac{\beta}{4}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta} \mathbf{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\overline{W}(s)) - g_{\mathbf{z},s}(\overline{W}(s))\right\|^2 \mathrm{d}s \\ &\le \frac{\beta}{2}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta} \mathbf{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\overline{W}(s)) - \nabla F_{\mathbf{z}}(\overline{W}(\lfloor s/\eta \rfloor \eta))\right\|^2 \mathrm{d}s \\ &\qquad + \frac{\beta}{2}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta} \mathbf{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\overline{W}(\lfloor s/\eta \rfloor \eta)) - g(\overline{W}(\lfloor s/\eta \rfloor \eta), \overline{U}_{\mathbf{z}}(s))\right\|^2 \mathrm{d}s \\ &\le \frac{\beta M^2}{2}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta} \mathbf{E}_{\mathbf{z}}\left\|\overline{W}(s) - \overline{W}(\lfloor s/\eta \rfloor \eta)\right\|^2 \mathrm{d}s \\ &\qquad + \frac{\beta}{2}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta} \mathbf{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\overline{W}(\lfloor s/\eta \rfloor \eta)) - g(\overline{W}(\lfloor s/\eta \rfloor \eta), \overline{U}_{\mathbf{z}}(s))\right\|^2 \mathrm{d}s. \end{aligned} \tag{D.4}$$

We first estimate the first summation in (D.4). Consider some $s \in [j\eta, (j+1)\eta)$. From (D.1), we have

$$\begin{aligned} \overline{W}(s) - \overline{W}(j\eta) &= -(s - j\eta)g(W_j, U_{\mathbf{z},j}) + \sqrt{\frac{2}{\beta}}\big(B(s) - B(j\eta)\big) \\ &= -(s - j\eta)\nabla F_{\mathbf{z}}(W_j) + (s - j\eta)\big(\nabla F_{\mathbf{z}}(W_j) - g(W_j, U_{\mathbf{z},j})\big) + \sqrt{\frac{2}{\beta}}\big(B(s) - B(j\eta)\big). \end{aligned}$$

Therefore, using Lemmas 2 and 3 and the gradient noise assumption **(A.4)**, we arrive at

$$\begin{aligned}
&\mathbf{E}_{\mathbf{z}}\|\overline{W}(s)-\overline{W}(j\eta)\|^2\\
&\quad\le 3\eta^2\mathbf{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(W_j)\|^2+3\eta^2\mathbf{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(W_j)-g(W_j,U_{\mathbf{z},j})\|^2+\frac{6\eta d}{\beta}\\
&\quad\le 12\eta^2\left(M^2\mathbf{E}_{\mathbf{z}}\|W_j\|^2+B^2\right)+\frac{6\eta d}{\beta}\\
&\quad\le 12\eta^2\left(M^2\left(\kappa_0+2\left(1\vee\frac{1}{m}\right)\left(b+2B^2+\frac{d}{\beta}\right)\right)+B^2\right)+\frac{6\eta d}{\beta}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbf{E}_{\mathbf{z}}\left\|\overline{W}(s)-\overline{W}(\lfloor s/\eta\rfloor\eta)\right\|^2\mathrm{d}s\\
&\quad\le 12\left(M^2\left(\kappa_0+2\left(1\vee\frac{1}{m}\right)\left(b+2B^2+\frac{d}{\beta}\right)\right)+B^2\right)k\eta^3+\frac{6d}{\beta}k\eta^2\\
&\quad\le\left(12\left(M^2\left(\kappa_0+2\left(1\vee\frac{1}{m}\right)\left(b+2B^2+\frac{d}{\beta}\right)\right)+B^2\right)+\frac{6d}{\beta}\right)\cdot k\eta^2\\
&\quad=:6\left(2C_0+\frac{d}{\beta}\right)\cdot k\eta^2. \qquad \text{(D.5)}
\end{aligned}$$

Similarly, the second summation on the right-hand side of (D.4) can be estimated as follows:

$$\begin{aligned}
&\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbf{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\overline{W}(\lfloor s/\eta\rfloor\eta))-g(\overline{W}(\lfloor s/\eta\rfloor\eta),\overline{U}(s))\right\|^2\mathrm{d}s\\
&\quad=\eta\sum_{j=0}^{k-1}\mathbf{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(W_j)-g(W_j,U_{\mathbf{z},j})\right\|^2\\
&\quad\le\eta\delta\sum_{j=0}^{k-1}2\left(M^2\mathbf{E}_{\mathbf{z}}\|W_j\|^2+B^2\right)\\
&\quad\le 2M^2\left(\kappa_0+2\left(1\vee\frac{1}{m}\right)\left(b+2B^2+\frac{d}{\beta}\right)\right)k\eta\delta+2\delta B^2k\eta\\
&\quad=2\left(M^2\left(\kappa_0+2\left(1\vee\frac{1}{m}\right)\left(b+2B^2+\frac{d}{\beta}\right)\right)+B^2\right)\cdot k\eta\delta\\
&\quad=2C_0\cdot k\eta\delta. \qquad \text{(D.6)}
\end{aligned}$$

Substituting Eqs. (D.5) and (D.6) into (D.4), we obtain

$$D(\mathbf{P}_V^{k\eta}\|\mathbf{P}_W^{k\eta})\le 6\left(\beta M^2C_0+M^2d\right)\cdot k\eta^2+\beta C_0\cdot k\eta\delta.$$

Now, since $\mu_{\mathbf{z},k} = \mathcal{L}(\overline{W}(k\eta)|\mathbf{Z} = \mathbf{z})$ and $\nu_{\mathbf{z},k\eta} = \mathcal{L}(W(k\eta)|\mathbf{Z} = \mathbf{z})$, the data-processing inequality for the KL divergence gives

$$\begin{aligned} D(\mu_{\mathbf{z},k}\|\nu_{\mathbf{z},k\eta}) &\leq D(\mathbf{P}_V^{k\eta}\|\mathbf{P}_W^{k\eta}) \\ &\leq 6\left(\beta M^2 C_0 + M^2 d\right) \cdot k\eta^2 + \beta C_0 \cdot k\eta\delta \\ &=: C_1 k\eta^2 + \beta C_0 k\eta\delta. \end{aligned}$$

Appendix E. Proof of Proposition 9

To establish the log-Sobolev inequality, we will use the Lyapunov function criterion of Cattiaux et al. (2010), reproduced as Proposition 15 in Appendix A.

We will apply this proposition to the Gibbs distribution $\pi_{\mathbf{z}}$ for some $\mathbf{z} \in \mathsf{Z}^n$, so that $H = \beta F_{\mathbf{z}}$ and

$$\mathcal{L}g = \Delta g - \beta\langle\nabla F_{\mathbf{z}}, \nabla g\rangle.$$

We consider the same Lyapunov function $V(w) = e^{m\beta\|w\|^2/4}$ as in Appendix B. From Eq. (B.1), V evidently satisfies (A.13) with κ and γ given in (B.2), i.e., the first condition of Proposition 15 is satisfied. Moreover, $\pi_{\mathbf{z}}$ satisfies a Poincaré inequality with constant $c_{\mathrm{P}} \leq 1/\lambda_*$. Thus, the second condition is also satisfied. Finally, by the M-smoothness assumption **(A.2)**, $\nabla^2 F_{\mathbf{z}} \succeq -MI_d$, so the third condition of Proposition 15 is satisfied with $K = \beta M$. Consequently, the constants C_1 and C_2 in (A.14) are given by

$$C_1 = \frac{2m^2 + 8M^2}{m^2 M\beta} \qquad \text{and} \qquad C_2 \leq \frac{6M(d+\beta)}{m}, \tag{E.1}$$

where we have also used the estimate (3.19) to upper-bound C_2. Therefore, from Proposition 15 and from (E.1) it follows that $\pi_{\mathbf{z}}$ satisfies a logarithmic Sobolev inequality with

$$c_{\mathrm{LS}} \leq \frac{2m^2 + 8M^2}{m^2 M\beta} + \frac{1}{\lambda_*}\left(\frac{6M(d+\beta)}{m} + 2\right).$$