Local Optimality and Generalization Guarantees for the Langevin Algorithm via Empirical Metastability

Belinda Tzen[*]   Tengyuan Liang[†]   Maxim Raginsky[‡]

Abstract

We study the detailed path-wise behavior of the discrete-time Langevin algorithm for non-convex Empirical Risk Minimization (ERM) through the lens of metastability, adopting some techniques from Berglund and Gentz (2003).

For a particular local optimum of the empirical risk, with an *arbitrary initialization*, we show that, with high probability, one of the two mutually exclusive events will occur: either the Langevin trajectory ends up somewhere outside the ε-neighborhood of this particular optimum within a short *recurrence time*; or it enters this ε-neighborhood by the recurrence time and stays there until an exponentially long *escape time*. We call this phenomenon *empirical metastability*.

This two-timescale characterization aligns nicely with the existing literature in the following two senses. First, the recurrence time is dimension-independent, and resembles the convergence time of deterministic Gradient Descent (GD). However unlike GD, the Langevin algorithm does not require strong conditions on local initialization, and has the possibility of eventually visiting all optima. Second, the scaling of the escape time is consistent with the Eyring-Kramers law, which states that the Langevin scheme will eventually visit all local minima, but it will take an exponentially long time to transit among them. We apply this path-wise concentration result in the context of statistical learning to examine local notions of generalization and optimality.

1   Introduction and informal summary of results

While it is a classical algorithm, gradient descent (along with variants, such as SGD) is now one of the most popular tools in large-scale optimization due to its simplicity and speed. Consider the following stochastic optimization problem:

$$\text{minimize} \qquad F(w) := \mathbf{E}_P[f(w, Z)]$$

over $w \in \mathbb{R}^d$, where Z is a random element of some space Z with probability law P, which is typically unknown. We have access to an n-tuple $\mathbf{Z} = (Z_1, \ldots, Z_n)$ of i.i.d. samples from P, and thus can attempt to minimize the empirical risk

$$F_{\mathbf{Z}}(w) := \frac{1}{n} \sum_{i=1}^{n} f(w, Z_i)$$

---

[*]University of Illinois; `btzen2@illinois.edu`.

[†]University of Chicago; `tengyuan.liang@chicagobooth.edu`.

[‡]University of Illinois; `maxim@illinois.edu`.

For example, we may use vanilla gradient descent, which has the form

$$W^{(k+1)} = W^{(k)} - \eta \nabla F_{\mathbf{Z}}(W^{(k)}), \qquad k = 0, 1, \dots$$

where $\eta > 0$ is the step size.

The behavior of gradient descent when the objective function F is convex is well understood. However, there has recently been much focus on non-convex optimization, motivated by the remarkable success of deep neural nets on problems across numerous disciplines. Gradient descent (or some stochastic variant) is often the algorithm of choice in these settings, but a theoretical characterization of the performance of these methods remains elusive in the non-convex case. A variant that has proven to be more amenable to analysis is the *Langevin algorithm*, i.e., gradient descent with appropriately scaled isotropic Gaussian noise added in each iteration:

$$W^{(k+1)} = W^{(k)} - \eta \nabla F_{\mathbf{Z}}(W^{(k)}) + \sqrt{2\beta^{-1}\eta}\,\xi^{(k)}, \tag{1.1}$$

where $\beta > 0$ is the *inverse temperature*, and the noise sequence $\xi^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ is independent of the initial point $W^{(0)}$. This modest modification preserves many of the desirable properties of gradient descent, while ensuring that it can aptly navigate a landscape containing multiple critical points. The analysis of (1.1) is facilitated by the fact that it can be viewed as a discrete-time approximation to a continuous-time Langevin diffusion, described by the Itô stochastic differential equation

$$dW_t = -\nabla F_{\mathbf{Z}}(W_t)dt + \sqrt{2\beta^{-1}}\,dB_t \tag{1.2}$$

where $(B_t)_{t \geq 0}$ is a standard d-dimensional Brownian motion (e.g., Borkar and Mitter (1999)).

The fact that the injection of Gaussian noise will cause the Langevin algorithm to asymptotically converge to a global minimum has long been known from the physics literature, and revisited more recently by, e.g., Welling and Teh (2011). The first non-asymptotic guarantee of global convergence of the Langevin algorithm was given by Raginsky, Rakhlin, and Telgarsky (2017), who showed that the Langevin algorithm with stochastic gradients converges to ε-approximate global minimizers after $\tilde{\mathcal{O}}\left(\text{poly}\left(\frac{1}{\varepsilon}, \beta, d, \frac{1}{\lambda^*}\right)\right)$ iterations, where λ^* is a spectral gap parameter that reflects the rate of convergence of the Langevin diffusion process to its stationary distribution, and is exponential in both d and β in general. A complementary result is that of Zhang, Liang, and Charikar (2017), who showed that a certain modified version of the stochastic gradient Langevin algorithm will hit an ε-approximate local minimum in time polynomial in all parameters.

The Gaussian perturbation inherent in the Langevin scheme is certainly necessary for ensuring its eventual convergence to a global optimum. However, even the vanilla gradient method can hit a local optimum (or at least an approximate stationary point) in polynomial time. To provide a more complete characterization of the behavior of the Langevin algorithm on non-convex problems, we analyze it over three timescales: colloquially, short, intermediate, and long. We show that, with high probability, the iterates of (1.1) will either be at least ε away from some local minimum within the *recurrence time* $k = \tilde{\mathcal{O}}\left(\frac{1}{\eta}\log\frac{1}{\varepsilon}\right)$, or fall within an ε-neighborhood of some local minimum and remain there through the *escape time* $\tilde{\mathcal{O}}\left(\frac{1}{\eta}\left(e^d + \log\frac{1}{\varepsilon}\right)\right)$. We refer to this behavior as *empirical metastability*. Moreover, we build on this result to provide generalization bounds under mild assumptions on the population risk function, e.g., those in recent work of Mei, Bai, and Montanari (2016) that demonstrates a close correspondence between the landscapes of the population and

empirical risk. Such a nuanced characterization of the Langevin algorithm is valuable for machine learning because it corroborates the empirical evidence that minimizers are approached quickly; provides good generalization guarantees; and ensures that a local minimum can be escaped and others reached in a non-asymptotic setting.

1.1 Method of analysis: an overview

Our analysis draws upon ideas and techniques from metastability theory, which is concerned with "the long-term behavior of dynamical systems under the influence of weak random perturbations" (Bovier and Den Hollander (2016)), where the system spends a large but random amount of time within a given region of its state space and subsequently transits to another due to noise. Metastability theory offers a view of the behavior of the Langevin algorithm over the risk landscape that is congruent with the interpretation of it as the discretization of the Langevin diffusion process.

As in Raginsky, Rakhlin, and Telgarsky (2017), the analysis proceeds by first characterizing the behavior of the continuous-time Langevin diffusion (1.2). In that context, the fast recurrence and slow escape phenomena can be stated precisely as follows: Starting at an arbitrary distance r from some local minimum of the empirical risk, with probability at least $1 - \delta$, the diffusion with sufficiently large β will either end up at distance at least $\varepsilon + re^{-t}$ from the minimum by time $t = \tilde{\mathcal{O}}\left(\log \frac{r}{\varepsilon}\right)$; or will approach it within distance $\varepsilon + re^{-t}$ at each t through $T = \tilde{\mathcal{O}}\left(e^d + \log \frac{r}{\varepsilon}\right)$. The exponential scaling of the eventual escape time is consistent with the Eyring-Kramers law, which guarantees exponentially long mean transition times between local optima (Bovier et al., 2004; Olivieri and Vares, 2004; Bovier and Den Hollander, 2016).

In order to prove this, we adopt the path-wise concentration approach of Berglund and Gentz (2003). We linearize the empirical gradient $\nabla F_{\mathbf{Z}}$ around some local minimum $\overline{w}_{\mathbf{Z}}$. This allows us to decompose the diffusion process, after an appropriate transformation, into a stable Gaussian process and a remainder term. The latter can be controlled using a delicate stopping-time analysis. The former lends itself to the construction of a martingale, whose behavior can be precisely described over time intervals of desired length. In particular, over small time intervals, we can obtain sharp Gaussian tail bounds on the the maximal deviation of that martingale, and extend the result to larger time intervals via the union bound. Using this decomposition, we establish the fast recurrence and slow escape result for the Langevin diffusion.

Next, we relate the continuous-time diffusion to the discrete-time Langevin algorithm via the usual linear interpolation and the Girsanov theorem, as in Raginsky, Rakhlin, and Telgarsky (2017). The Langevin algorithm iterates have the same joint distribution as that of the interpolation at the corresponding points, which allows us to construct a coupling of the discrete-time Langevin algorithm (1.1) and the continuous-time diffusion (1.2) sampled at integer multiples of η, such that $W^{(k)} = W_{k\eta}$ for all $k \leq K$ with high probability. However, in order to transfer the continuous-time metastability result to the discrete-time setting, we also need to take care to account for the behavior of the diffusion in the intervening intervals $(k\eta, (k + 1)\eta)$. This is done via a quantitative continuity estimate for the diffusion using Gronwall's lemma and the reflection principle for the Brownian motion (Mörters and Peres, 2010).

Finally, we use the metastability results to provide local optimality and generalization guarantees for the Langevin algorithm. The main quantity of interest is the difference

$$F(\overline{w}_{\mathbf{Z}}) - \min_{K_0 \leq k \leq K} F_{\mathbf{Z}}(W^{(k)}),$$

where K_0 is the recurrence time of the Langevin algorithm. Note that, while $F(\overline{w}_{\mathbf{Z}})$ involves the population risk, the term $\min_{K_0 \le k \le K} F_{\mathbf{Z}}(W^{(k)})$ can be readily computed from the trajectory of the Langevin algorithm. To control this difference, we use the techniques recently developed by Mei, Bai, and Montanari (2016) in order to relate the local geometry of the empirical risk to that of the population risk.

1.2 Notation

Any positive definite matrix $A \in \mathbb{R}^{d \times d}$ induces a norm $\|w\|_A := \sqrt{\langle w, Aw \rangle}$. We will denote by $\mathsf{B}_A^d(v, r)$ the ball of radius r with center $v \in \mathbb{R}^d$ in this norm: $\mathsf{B}_A^d(v, r) := \left\{ w \in \mathbb{R}^d : \|w - v\|_A \le r \right\}$. The usual Euclidean norm on $\mathbb{R}^d$ will be denoted by $\|\cdot\|$. We will denote by $\mathsf{B}^d(v, r)$ the Euclidean ball of radius r centered at v; when $v = 0$, we will simply write $\mathsf{B}^d(r)$. We will denote by $\lambda_1(A), \dots, \lambda_d(A)$ the eigenvalues of a symmetric matrix $A \in \mathbb{R}^{d \times d}$, and by $\|A\|$ the spectral norm of A.

2 Main Results

We impose the following assumptions:

(A.1) The functions $f(\cdot, z)$ are C^2, and there exist constants $A, B, C \ge 0$, such that

$$|f(0, z)| \le A, \qquad \|\nabla f(0, z)\| \le B, \qquad \|\nabla^2 f(0, z)\| \le C, \qquad \forall z \in \mathsf{Z}.$$

(A.2) The functions $f(\cdot, z)$ have Lipschitz-continuous gradients and Hessians, uniformly in $z \in \mathsf{Z}$: there exist constants $L, M > 0$, such that, for all $w, v \in \mathbb{R}^d$,

$$\|\nabla f(w, z) - \nabla f(v, z)\| \le M \|w - v\| \quad \text{and} \quad \|\nabla^2 f(w, z) - \nabla^2 f(v, z)\| \le L \|w - v\|.$$

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

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

The first two assumptions are standard. For the discussion of the dissipativity assumption, see Raginsky, Rakhlin, and Telgarsky (2017). It is not hard to see that, if it holds, then any critical point (i.e., the one where the gradient is equal to zero) of both the empirical risk $F_{\mathbf{Z}}$ and the population risk F is contained in the ball $\mathsf{B}^d(R)$ with $R = \sqrt{b/m}$.

2.1 Empirical metastability

Our first result concerns the behavior of the Langevin algorithm (1.1) with respect to an arbitrary nondegenerate local minimum $\overline{w}_{\mathbf{Z}}$ of the empirical risk $F_{\mathbf{Z}}$ – that is, the Hessian $H := \nabla^2 F_{\mathbf{Z}}(\overline{w}_{\mathbf{Z}})$ is positive definite. Without loss of generality, we may assume that the eigenvalues of H are lower-bounded by m.

Theorem 2.1. *Fix some $\delta \in (0, 1)$ and $r > 0$. There exist absolute constants $c_1, c_2 > 0$, such that the following holds. Assume $\varepsilon \in (0, \frac{c_1 m^2}{L \sqrt{M}} \wedge 8r)$ and consider the Langevin algorithm (1.1) with initial point $w \in \mathsf{B}_H^d(\overline{w}_{\mathbf{Z}}, r) \cap \mathsf{B}^d(R)$, where the number of iterations K, the step size η, and the inverse temperature β satisfy the following conditions:*

4

- $K\eta \le e^{c_1 d} + \dfrac{2}{m} \log \dfrac{8r}{\varepsilon}$;

- $\beta \ge \dfrac{c_2}{\varepsilon^2}\left(d + \log \dfrac{M}{\delta}\right) \vee \dfrac{c_2 M}{\varepsilon^2} \log \dfrac{Kd}{\delta}$;

- $\eta \le 1 \wedge \dfrac{c_1 \delta^2}{M^2(\beta G_0 + d)K} \wedge \dfrac{c_1 \varepsilon \delta}{\sqrt{M^3 G_1 K}}$,

where

$$G_0 := 2M^2\left(R^2 + 2\left(1 \vee \frac{1}{m}\right)\left(b + 2B^2 + \frac{d}{\beta}\right)\right) + 2B^2, \quad G_1 := R + \frac{b+d}{m}. \tag{2.1}$$

Then, for any realization of $\mathbf{Z}$, with probability at least $1 - \delta$ with respect to the Gaussian noise sequence $\xi^{(1)}, \xi^{(2)}, \ldots$, exactly one of the following two events will occur:

1. $\|W^{(k)} - \overline{w}_{\mathbf{Z}}\|_H \ge \frac{1}{2}\left(\varepsilon + re^{-mk\eta}\right)$ for some $k \le \dfrac{2}{m\eta} \log \dfrac{8r}{\varepsilon}$

2. $\|W^{(k)} - \overline{w}_{\mathbf{Z}}\|_H \le \varepsilon + re^{-mk\eta}$ for all iterations $\dfrac{2}{m\eta} \log \dfrac{8r}{\varepsilon} \le k \le K$.

2.2 Local optimality and generalization

Theorem 2.2. *Suppose that f satisfies Assumptions* **(A.1)–(A.3)**, *and that the population risk F is $(2\varepsilon_0, 2m)$-strongly Morse in the sense of Mei, Bai, and Montanari (2016):*

$$\|\nabla F(w)\| \le 2\varepsilon_0 \quad \text{implies} \quad \min_{j \in [d]} |\lambda_j(\nabla^2 F(w))| \ge 2m.$$

Consider running the discrete-time Langevin algorithm (1.1) initialized with $W^{(0)} = w$, where β, η, K are chosen as in Theorem 2.1. Suppose also that the sample size satisfies $n \ge cd \log d$ and $\dfrac{n}{\log n} \ge \dfrac{c\sigma^2 d}{(\varepsilon_0 \wedge m)^2}$, where

$$c = c_0\left(1 \vee \log((M \vee L \vee (B + MR))R\sigma/\delta)\right)$$

for some absolute constant c_0, and $\sigma := (A + (B + MR)R) \vee (B + MR) \vee (C + LR)$, and that $\varepsilon \le \frac{3m^{3/2}}{2L}$. Then with probability at least $1 - \delta$ (with respect to the training data $\mathbf{Z}$ and the isotropic Gaussian noise added by the Langevin algorithm), for any local minimum $\overline{w}_{\mathbf{Z}}$ of the empirical risk $F_{\mathbf{Z}}$, either $\|W^{(k)} - \overline{w}_{\mathbf{Z}}\|_H \ge 2\varepsilon$ for some $k \le K_0 := \dfrac{2}{m\eta} \log \dfrac{8r}{\varepsilon}$ or

$$F(\overline{w}_{\mathbf{Z}}) \le \min_{K_0 \le k \le K} F_{\mathbf{Z}}(W^{(k)}) + \sigma\sqrt{\frac{cd \log n}{n}}, \tag{2.2}$$

where $r = \|w - \overline{w}_{\mathbf{Z}}\|_H$ and $H = \nabla^2 F(\overline{w}_{\mathbf{Z}})$.

5

3 Proof of Theorem 2.1

3.1 Preliminaries: a martingale concentration lemma

We first define some auxiliary quantities and establish a useful result for the Langevin diffusion (1.2), along the lines of Berglund and Gentz (2003).

We begin by linearizing the gradient $\nabla F_{\mathbf{Z}}$ around $\overline{w}_{\mathbf{Z}}$: Recalling that $H = \nabla^2 F_{\mathbf{Z}}(\overline{w}_{\mathbf{Z}})$ and writing $W_t = Y_t + \overline{w}_{\mathbf{Z}}$, we have $\nabla F_{\mathbf{Z}}(W_t) = HY_t - \rho(Y_t)$. Since the Hessian of $F_{\mathbf{Z}}$ is L-Lipschitz, the remainder satisfies $\|\rho(Y_t)\| \leq \frac{L}{2}\|Y_t\|^2$ (Nesterov, 2013, Lemma 1.2.4). The diffusion process $\{Y_t\}_{t\geq 0}^\infty$ obeys the Itô SDE

$$\mathrm{d}Y_t = -(HY_t - \rho(Y_t))\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t \qquad (3.1)$$

with the initial condition $Y_0 = w - \overline{w}_{\mathbf{Z}}$. The solution of (3.1) is given by

$$Y_t = e^{-tH}Y_0 + \sqrt{2\beta^{-1}}\int_0^t e^{(s-t)H}\mathrm{d}B_s + \int_0^t e^{(s-t)H}\rho(Y_s)\mathrm{d}s,$$

where $e^A := \sum_{k=0}^\infty \frac{A^k}{k!}$, for $A \in \mathbb{R}^{d\times d}$, is the matrix exponential.

Given $0 \leq t_0 \leq t_1$, let us define the matrix flow $Q_{t_0}(t) := H^{1/2}e^{(t_0-t)H}$ and the Itô process

$$Z_t := e^{(t-t_0)H}Y_t$$
$$= e^{-t_0 H}Y_0 + \sqrt{2\beta^{-1}}\int_0^t e^{(s-t_0)H}\mathrm{d}B_s + \int_0^t e^{(s-t_0)H}\rho(Y_s)\mathrm{d}s,$$

for $t \in [t_0, t_1]$. These definitions are motivated by the observation that $\|Y_t\|_H = \|H^{1/2}Y_t\| = \|Q_{t_0}(t)Z_t\|$. Next, we decompose Z_t into the Ornstein–Uhlenbeck term

$$Z_t^0 := e^{-t_0 H}Y_0 + \sqrt{2/\beta}\int_0^t e^{(s-t_0)H}\mathrm{d}B_s$$

and the remainder term

$$Z_t^1 := \int_0^t e^{(s-t_0)H}\rho(Y_s)\mathrm{d}s.$$

Note that, for any $t \in [t_0, t_1]$,

$$Q_{t_0}(t_1)Z_t^0 = H^{1/2}e^{-t_1 H}Y_0 + \sqrt{2\beta^{-1}}\int_0^t H^{1/2}e^{(s-t_1)H}\mathrm{d}B_s,$$

so the process $\{Q_{t_0}(t_1)Z_t^0\}_{t\in[t_0,t_1]}$ is a martingale. This leads to the following tail bound (see Appendix A for the proof):

Lemma 3.1. *For any $\lambda \in (0, 1/2)$, $h > 0$, and $y_0 \in \mathbb{R}^d$,*

$$\mathbf{P}^{y_0}\left[\sup_{t_0\leq t\leq t_1}\|Q_{t_0}(t_1)Z_t^0\| \geq h\right] \leq \left(\frac{1}{1-\lambda}\right)^{d/2}\exp\left(-\frac{\beta\lambda}{2}\left[h^2 - \langle\mu, (I-\beta\lambda\Sigma)^{-1}\mu\rangle\right]\right), \qquad (3.2)$$

where $\mu := H^{1/2}e^{-t_1 H}y_0$ and $\Sigma_t := \beta^{-1}(I - e^{-2t_1 H})$.

3.2 Langevin diffusion: fast recurrence and slow escape

We now provide the quantitative path-wise metastability result for the Langevin diffusion.

Proposition 3.1. *Fix any $r > 0$ and $\varepsilon \in \left(0, \frac{(\sqrt{2}-1)m^2}{4L\sqrt{2M}} \wedge 8r\right)$. Define the* recurrence time $T_{\mathrm{rec}} = \frac{2}{m} \log \frac{8r}{\varepsilon}$
and the escape time $T_{\mathrm{esc}} = e^{Cd} + T_{\mathrm{rec}}$, *for an arbitrary constant $C > 0$, and consider the stopping time*

$$\tau := \inf\left\{t \geq 0 : \frac{\|W_t - \overline{w}_{\mathbf{z}}\|_H}{\varepsilon + re^{-mt}} \geq 1\right\}.$$

Then, for any initial point $w \in \mathsf{B}_H^d(\overline{w}_{\mathbf{z}}, r)$,

$$\mathbf{P}^w\left[\tau \in [T_{\mathrm{rec}}, T_{\mathrm{esc}}]\right] \leq \delta,$$

provided $\beta \geq \frac{128}{\varepsilon^2}\left[(C+1)d + \log \frac{2M+1}{\delta}\right]$.

Proof. Since $\|Y_t\|_H \leq r$, we know that $\tau > 0$. Fix some $T_{\mathrm{rec}} \leq t_0 \leq t_1$, such that $t_1 - t_0 \leq \frac{1}{2M}$. For every $t \in [t_0, t_1]$, $Y_t = W_t - \overline{w}_{\mathbf{z}}$ satisfies

$$\|Y_t\|_H = \|Q_{t_0}(t)Z_t\| = \|e^{(t_1-t)H}Q_{t_0}(t_1)Z_t\| \leq e^{1/2}\|Q_{t_0}(t_1)Z_t\|.$$

Therefore,

$$\begin{aligned}
&\mathbf{P}^w\left[\tau \in [t_0, t_1]\right] \\
&= \mathbf{P}^{y_0}\left[\sup_{t_0 \leq t \leq t_1} \frac{\|Y_t\|_H}{\varepsilon + re^{-mt}} \geq 1, \tau \geq t_0\right] \\
&\leq \mathbf{P}^{y_0}\left[\sup_{t_0 \leq t \leq t_1} \frac{\|Q_{t_0}(t_1)Z_t\|}{\varepsilon + re^{-mt}} \geq \frac{1}{2}, \tau \geq t_0\right] \\
&\leq \mathbf{P}^{y_0}\left[\sup_{t_0 \leq t \leq t_1} \frac{\|Q_{t_0}(t_1)Z_t^0\|}{\varepsilon + re^{-mt}} \geq c_0, \tau \geq t_0\right] + \mathbf{P}^{y_0}\left[\sup_{t_0 \leq t \leq t_1} \frac{\|Q_{t_0}(t_1)Z_t^1\|}{\varepsilon + re^{-mt}} \geq c_1, \tau \geq t_0\right] \\
&=: P_0 + P_1,
\end{aligned}$$

for any choice of $c_0, c_1 > 0$ satisfying $c_0 + c_1 = \frac{1}{2}$.

We first upper-bound P_1. On the event $\tau \in [t_0, t_1]$, for any $0 \leq s \leq t_1 \wedge \tau$, $\|\rho(Y_s)\| \leq \frac{L}{2}\|Y_s\|^2 \leq \frac{L}{2m}(\varepsilon + re^{-ms})^2$. Therefore, for any $t \in [t_0, t_1 \wedge \tau]$,

$$\begin{aligned}
\|Q_{t_0}(t_1)Z_t^1\| &\leq \|H^{1/2}\| \int_0^t \|e^{(s-t_1)H}\rho(Y_s)\|\mathrm{d}s \\
&\leq \|H^{1/2}\| \int_0^t \|e^{(s-t_1)H}\|\|\rho(Y_s)\|\mathrm{d}s \\
&\leq \frac{L\sqrt{M}}{2m} \int_0^t e^{(s-t_1)m}(\varepsilon + re^{-ms})^2 \mathrm{d}s \\
&\leq \frac{L\sqrt{M}}{m^2} \cdot (\varepsilon^2 + r^2 e^{-mt}) \\
&< \frac{2L\sqrt{M}}{m^2}\varepsilon^2
\end{aligned}$$

7

since $t \geq t_0 \geq T_{\text{rec}}$. Consequently, if we take $c_1 = \frac{2L\sqrt{M}}{m^2}\varepsilon$, then

$$\sup_{t_0 \leq t \leq t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1)Z_t\|}{\varepsilon + re^{-mt}} \leq \frac{1}{\varepsilon} \sup_{t_0 \leq t \leq t_1 \wedge \tau} \|Q_{t_0}(t_1)Z_t\| < c_1,$$

which implies $P_1 = 0$. Moreover, $c_0 = \frac{1}{2} - c_1 = \frac{1}{2} - \frac{2L\sqrt{M}}{m^2}\varepsilon > \frac{1}{4}$.

We next estimate P_0. Since $\|y_0\|_H^2 \leq r^2$,

$$\left\langle H^{1/2}y_0, e^{-t_1 H}\left(I - \frac{3}{4}(I - e^{-2t_1 H})\right)^{-1} e^{-t_1 H}H^{1/2}y_0 \right\rangle$$

$$\leq \left\| e^{-2t_1 H}\left(I - \frac{3}{4}(I - e^{-2t_1 H})\right)^{-1} \right\| \|y_0\|_H^2$$

$$\leq 4e^{-2mt_1}r^2 \leq \varepsilon^2/16.$$

Therefore, by Lemma 3.1 with $h = (\varepsilon + re^{-mt_1})c_0$ and $\lambda = \frac{3}{4}$,

$$P_0 \leq \mathbf{P}^{y_0}\left[\sup_{t_0 \leq t \leq t_1} \frac{\|Q_{t_0}(t_1)Z_t^0\|}{\varepsilon + re^{-mt}} \geq c_0 \right]$$

$$\leq \mathbf{P}^{y_0}\left[\sup_{t_0 \leq t \leq t_1} \|Q_{t_0}(t_1)Z_t^0\| \geq (\varepsilon + re^{-mt_1})c_0 \right]$$

$$\leq 2^d \exp\left(-\frac{3\beta\varepsilon^2}{8}\left(c_0^2 - \frac{1}{16}\right) \right)$$

$$\leq 2^d \exp\left(-\frac{3\beta\varepsilon^2}{128} \right).$$

Thus, for any $t_0 \geq T_{\text{rec}}$ and $t_0 \leq t_1 \leq t_0 + \frac{1}{2M}$,

$$\mathbf{P}^w\left[\tau \in [t_0, t_1] \right] \leq 2^d \exp\left(-\frac{3\beta\varepsilon^2}{128} \right). \tag{3.3}$$

Now fix an arbitrary $C > 0$ and recall the definition of the escape time $T_{\text{esc}} = e^{Cd} + T_{\text{rec}}$. Let $J := \lceil 2Me^{Cd} \rceil$ and partition the interval $[0, T]$ using the points $T_{\text{rec}} = t_0 < t_1 < \ldots < t_J = T_{\text{esc}}$ with $t_j = \frac{j}{2M}$ for $j = 0, 1, \ldots, J - 1$. Then, using the union bound and Eq. (3.3), we obtain

$$\mathbf{P}^w\left[\tau \in [T_{\text{rec}}, T_{\text{esc}}] \right] = \sum_{j=0}^{J-1} \mathbf{P}^w\left[\tau \in [t_j, t_{j+1}] \right]$$

$$\leq (2Me^{Cd} + 1)2^d \exp\left(-\frac{3\beta\varepsilon^2}{128} \right),$$

and this probability will be smaller than δ if we choose β accordingly. $\qquad\square$

3.3 A path coupling argument

Next, we relate the discrete Langevin algorithm (1.1) to the diffusion (1.2) sampled at $t = 0, \eta, 2\eta, \ldots$. To that end, we introduce the standard continuous-time interpolation

$$V_t := V_0 - \int_0^t \nabla F_{\mathbf{Z}}(V_{\lfloor s/\eta \rfloor \eta}) \mathrm{d}s + \sqrt{2\beta^{-1}} \int_0^t \mathrm{d}B_s,$$

with the deterministic initial condition $V_0 = W^{(0)} = w$. Note that, for any K, the probability law μ_K of the random vector $(W^{(1)}, \ldots, W^{(K)})$ is equal to the probability law ν_K of the random vector $(V_\eta, \ldots, V_{K\eta})$.

Let us denote by $\mathbf{P}_V^t$ and $\mathbf{P}_W^t$ the probability laws of $(V_s : 0 \le s \le t)$ and $(W_s : 0 \le s \le t)$, respectively, with the same deterministic initialization $V_0 = W_0 = w$. Then we can invoke Girsanov's theorem in the same way as in Dalalyan (2017) and Raginsky, Rakhlin, and Telgarsky (2017) to show that, for any $K \in \mathbb{N}$,

$$D\left(\mathbf{P}_V^{K\eta} \big\| \mathbf{P}_W^{K\eta}\right) \le \frac{\beta M^2 \eta^3}{2} \sum_{k=0}^{K-1} \mathbf{E}^w \|\nabla F_{\mathbf{Z}}(W^{(k)})\|^2 + KdM^2\eta^2, \tag{3.4}$$

where $D(\cdot\|\cdot)$ denotes the relative entropy. The expectations of the squared norms of the gradients can be upper-bounded using Lemmas 2 and 3 of Raginsky, Rakhlin, and Telgarsky (2017): under **(A.1)**–**(A.3)**, $\sup_{k \ge 0} \mathbf{E}^w \|\nabla F_{\mathbf{Z}}(W^{(k)})\|^2 \le G_0$, where G_0 is defined in (2.1). Therefore,

$$D(\mu_K \| \nu_K) \le D\left(\mathbf{P}_V^{K\eta} \big\| \mathbf{P}_W^{K\eta}\right) \le M^2 \left(\frac{\beta G_0}{2} + d\right) K\eta^2,$$

where the first step is by the data processing inequality for the relative entropy, and the second one follows upon substituting the gradient bound into (3.4). Using Pinsker's inequality, we obtain

$$\|\mu_K - \nu_K\|_{\mathrm{TV}}^2 \le \frac{M^2}{2} \left(\frac{\beta G_0}{2} + d\right) K\eta^2.$$

Now we recall the following result about an optimal coupling (Lindvall, 1992, Theorem 5.2): Given any two random elements X, Y of a common standard Borel space X, there exists a coupling $\mathbf{M}$ of X and Y, i.e., a probability measure $\mathbf{M}$ on the product space $\mathsf{X} \times \mathsf{X}$, such that $\mathbf{M}(\cdot \times \mathsf{X}) = \mathcal{L}(X)$, $\mathbf{M}(\mathsf{X} \times \cdot) = \mathcal{L}(Y)$, and

$$\mathbf{M}[X \ne Y] \le \|\mathcal{L}(X) - \mathcal{L}(Y)\|_{\mathrm{TV}}$$

Hence, given any $\beta > 0$ and any $K \in \mathbb{N}$, we can choose $\eta \le \frac{4\delta^2}{M^2(\beta G_0 + 2d)K}$ to ensure that there exists a coupling $\mathbf{M}$ of $(W^{(k)} : k \in [K])$ and $(W_{k\eta} : k \in [K])$, such that

$$\mathbf{M}\left((W^{(1)}, W^{(2)}, \ldots, W^{(K)}) \ne (W_\eta, W_{2\eta}, \ldots, W_{K\eta})\right) \le \delta.$$

In that case, we have

$$\begin{aligned}
\mathbf{P}&\left((W^{(1)}, \ldots, W^{(K)}) \in \cdot\right) \\
&= \mathbf{M}\left((W^{(1)}, \ldots, W^{(K)}) \in \cdot\right) \\
&\le \mathbf{M}\left((W_\eta, \ldots, W_{K\eta}) \in \cdot\right) + \mathbf{M}\left((W^{(1)}, \ldots, W^{(K)}) \ne (W_1, \ldots, W_{K\eta})\right) \\
&\le \mathbf{P}\left((W_\eta, \ldots, W_{K\eta}) \in \cdot\right) + \delta.
\end{aligned} \tag{3.5}$$

3.4 Completing the proof

Consider the discrete-time Langevin algorithm (1.1). We need to upper-bound the probability of the event $\tilde{A} = \tilde{A}_1 \cap \tilde{A}_2$, where

$$\tilde{A}_1 := \left\{ \max_{k \leq T_{\mathrm{rec}}\eta^{-1}} \frac{\|W^{(k)} - \overline{w}_{\mathbf{Z}}\|_H}{\varepsilon + re^{-mk\eta}} \leq \frac{1}{2} \right\} \quad \text{and} \quad \tilde{A}_2 := \left\{ \max_{\eta^{-1}T_{\mathrm{rec}} \leq k \leq K} \frac{\|W^{(k)} - \overline{w}_{\mathbf{Z}}\|_H}{\varepsilon + re^{-mk\eta}} \geq 1 \right\}.$$

We also define the analogous events for the continuous-time Langevin diffusion (1.2):

$$A_1 := \left\{ \max_{k \leq T_{\mathrm{rec}}\eta^{-1}} \frac{\|W_{k\eta} - \overline{w}_{\mathbf{Z}}\|_H}{\varepsilon + re^{-mk\eta}} \leq \frac{1}{2} \right\} \quad \text{and} \quad A_2 := \left\{ \max_{\eta^{-1}T_{\mathrm{rec}} \leq k \leq K} \frac{\|W_{k\eta} - \overline{w}_{\mathbf{Z}}\|_H}{\varepsilon + re^{-mk\eta}} \geq 1 \right\}.$$

By Proposition 3.1, we can choose $\beta > 0$ large enough, so that, with probability at least $1 - \delta/3$, for the continuous-time Langevin diffusion (1.2) we have either $\|W_t - \overline{w}_{\mathbf{Z}}\|_H \geq \varepsilon + re^{-mt}$ for some $t \leq T_{\mathrm{esc}}$ or $\|W_t - \overline{w}_{\mathbf{Z}}\|_H \leq \varepsilon + re^{-mt}$ for all $t \leq T_{\mathrm{esc}}$. Moreover, for any K, η, and β satisfying the conditions of the theorem, there exists a coupling $\mathbf{M}$ of $(W^{(1)}, \ldots, W^{(K)})$ and $(W_\eta, \ldots, W_{K\eta})$, such that, with probability $1 - \delta/3$, $W^{(k)} = W_{k\eta}$ for all $k \in [K]$. Then, using the path coupling estimate (3.5), we can write

$$\mathbf{P}^w[\tilde{A}] = \mathbf{P}^w[\tilde{A}_1 \cap \tilde{A}_2] \leq \mathbf{P}^w[A_1 \cap A_2] + \frac{\delta}{3}.$$

It remains to estimate the probability of $A_1 \cap A_2$. For $K_0 := \lceil T_{\mathrm{rec}}\eta^{-1} \rceil$, partition the interval $[0, T_{\mathrm{rec}}]$ using the points $0 = t_0 < t_1 < \ldots < t_{K_0} = T_{\mathrm{rec}}$ with $t_k = k\eta$ for $k = 0, 1, \ldots, K_0 - 1$, and consider the event

$$B := \left\{ \max_{0 \leq k \leq K_0 - 1} \max_{t \in [t_k, t_{k+1}]} \|W_t - W_{t_{k+1}}\| \leq \frac{\varepsilon}{2\sqrt{M}} \right\}.$$

On the event $A_1 \cap B$,

$$\sup_{t \in [0, T_{\mathrm{rec}}]} \frac{\|W_t - \overline{w}_{\mathbf{Z}}\|_H}{\varepsilon + re^{-ms}} = \max_{0 \leq k \leq K_0 - 1} \sup_{t \in [t_k, t_{k+1}]} \frac{\|W_t - \overline{w}_{\mathbf{Z}}\|_H}{\varepsilon + re^{-mt}}$$

$$\leq \frac{1}{2} + \max_{0 \leq k \leq K_0 - 1} \max_{t \in [t_k, t_{k+1}]} \frac{\sqrt{M}}{\varepsilon} \|W_t - W_{t_{k+1}}\| \leq 1,$$

hence

$$\mathbf{P}^w[A_1 \cap A_2] \leq \mathbf{P}^w[A_1 \cap A_2 \cap B] + \mathbf{P}^w[B^c]$$
$$\leq \mathbf{P}^w[\tau \in [T_{\mathrm{rec}}, T_{\mathrm{esc}}]] + \mathbf{P}^w[B^c]$$
$$\leq \frac{\delta}{3} + \mathbf{P}^w[B^c].$$

To complete the proof, we need to arrange things so that $\mathbf{P}^w[B^c] \leq \frac{\delta}{3}$.

For $t \in [t_k, t_{k+1}]$,

$$\|W_t - W_{t_{k+1}}\| \leq \int_t^{t_{k+1}} \|\nabla F_{\mathbf{Z}}(W_s)\| \mathrm{d}s + \sqrt{2\beta^{-1}} \|B_t - B_{t_{k+1}}\|$$
$$\leq M \int_t^{t_{k+1}} \|W_s - W_{t_{k+1}}\| \mathrm{d}s + M\eta \|W_{t_{k+1}}\| + \sqrt{2\beta^{-1}} \|B_t - B_{t_{k+1}}\|.$$

Therefore, using Gronwall's lemma, we obtain

$$\sup_{t \in [t_k, t_{k+1}]} \|W_t - W_{t_{k+1}}\| \le e^{M\eta} \left[M\eta \|W_{t_{k+1}}\| + \sqrt{2\beta^{-1}} \sup_{t \in [t_k, t_{k+1}]} \|B_t - B_{t_{k+1}}\| \right].$$

By Lemma 3 in Raginsky, Rakhlin, and Telgarsky (2017) and Markov's inequality, for any $u > 0$

$$\mathbf{P}^w \left[\|W_{t_{k+1}}\| \ge u \right] \le \frac{\sup_{t \ge 0} \mathbf{E}^w \|W_t\|^2}{u^2} \le \frac{G_1}{u^2}, \tag{3.6}$$

with G_1 given in (2.1). In addition, by the reflection principle for the Brownian motion (Mörters and Peres, 2010),

$$\mathbf{P} \left[\sup_{t \in [t_k, t_{k+1}]} \|B_t - B_{t_{k+1}}\| \ge u \right] \le 2d e^{-cu^2/\eta} \tag{3.7}$$

for some absolute constant $c > 0$. Therefore,

$$\mathbf{P}^w[B^c] \le \sum_{k=0}^{K_0-1} \left(\mathbf{P}^w \left[\|W_{t_{k+1}}\| \ge \frac{\varepsilon e^{-M\eta}}{4M^{3/2}\eta} \right] + \mathbf{P} \left[\sup_{t \in [t_k, t_{k+1}]} \|B_t - B_{t_{k+1}}\| \ge \varepsilon e^{-M\eta} \sqrt{\frac{\beta}{32M}} \right] \right)$$

$$\le K \left(\frac{16 G_1 M^3 \eta^2}{\varepsilon^2} e^{2M\eta} + 2d \exp\left(-\frac{c'\beta\varepsilon^2}{M\eta} e^{-2M\eta} \right) \right) \tag{3.8}$$

for some absolute constant $c' > 0$. We can first choose $\eta > 0$ small enough to make the first term in (3.8) smaller than $\delta/6$, and then $\beta > 0$ large enough to make the second term smaller than $\delta/6$.

4 Proof of Theorem 2.2

4.1 Preliminaries

The following lemma can be proved using the methodology of Mei, Bai, and Montanari (2016); we give a self-contained proof in Appendix B for completeness.

Lemma 4.1 (uniform deviation guarantees). *Under Assumptions* **(A.1)** *and* **(A.2)**, *there exists an absolute constant c_0, such that the following holds, for $c = c_0 \left(1 \vee \log((M \vee L \vee (B + MR))R\sigma/\delta) \right)$ and $\sigma := (A + (B + MR)R) \vee (B + MR) \vee (C + LR)$:*

- *If $n \ge cd \log d$, then, with probability at least $1 - \delta$,*

$$\sup_{w \in \mathsf{B}^d(R)} |F_{\mathbf{Z}}(w) - F(w)| \le \sigma \sqrt{\frac{cd \log n}{n}}. \tag{4.1}$$

- *If $n \ge cd \log d$, then, with probability at least $1 - \delta$,*

$$\sup_{w \in \mathsf{B}^d(R)} \|\nabla F_{\mathbf{Z}}(w) - \nabla F(w)\| \le \sigma \sqrt{\frac{cd \log n}{n}}. \tag{4.2}$$

- *If $n \ge cd \log d$, then, with probability at least $1 - \delta$,*

$$\sup_{w \in \mathsf{B}^d(R)} \|\nabla^2 F_{\mathbf{Z}}(w) - \nabla^2 F(w)\| \le \sigma \sqrt{\frac{cd \log n}{n}}. \tag{4.3}$$

11

4.2 Empirical risk is strongly Morse with high probability

The following result is due to Mei, Bai, and Montanari (2016). We give a short proof, to keep the presentation self-contained.

Proposition 4.1. *If the population risk $F(w)$ is $(2\varepsilon_0, 2m)$-strongly Morse, then, provided n satisfies the conditions of Lemma 4.1 as well as the condition $\dfrac{n}{d\log n} \geq \dfrac{c\sigma^2}{(\varepsilon_0 \wedge m)^2}$, the empirical risk $F_{\mathbf{Z}}(w)$ is (ε_0, m)-strongly Morse with probability at least $1 - \delta$.*

Proof. By Lemma 4.1, the following event occurs with probability at least $1 - \delta$:

$$\sup_{w \in B^d(R)} \|\nabla F_{\mathbf{Z}}(w) - \nabla F(w)\| \leq \varepsilon_0 \qquad \text{and} \qquad \sup_{w \in B^d(R)} \|\nabla^2 F_{\mathbf{Z}}(w) - \nabla^2 F(w)\| \leq m.$$

On this event, for any $w \in B^d(R)$ satisfying $\min_{j \leq d} |\lambda_j(\nabla^2 F(w))| \geq 2m$,

$$\min_{j \leq d} |\lambda_j(\nabla^2 F_{\mathbf{Z}}(w))| \geq \min_{j \leq d} |\lambda_j \nabla^2 F(w)| - \max_{j \leq d} |\lambda_j(\nabla^2 F_{\mathbf{Z}}(w)) - \lambda_j(\nabla^2 F(w))|$$

$$\geq \min_{j \leq d} |\lambda_j \nabla^2 F(w)| - \|\nabla^2 F_{\mathbf{Z}}(w) - \nabla^2 F(w)\|$$

$$\geq m,$$

by Weyl's perturbation theorem (Bhatia, 1997, Corollary III.2.6). Therefore, for any $w \in B^d(R)$ satisfying $\|\nabla F_{\mathbf{Z}}(w)\| \leq \varepsilon_0$, we have $\|\nabla F(w)\| \leq 2\varepsilon_0$, in which case the absolute values of all the eigenvalues of $\nabla^2 F_{\mathbf{Z}}(w)$ are at least m. Hence, the empirical risk $F_{\mathbf{Z}}$ is (ε_0, m)-strongly Morse. $\square$

4.3 Completing the proof: a posteriori risk bound

Now let $\overline{w}_{\mathbf{Z}}$ be a local minimum of the empirical risk $F_{\mathbf{Z}}$. By Proposition 4.1, all eigenvalues of the Hessian $H = \nabla^2 F_{\mathbf{Z}}(\overline{w}_{\mathbf{Z}})$ are at least m; therefore, the norm $\|\cdot\|_H$ is well defined, and satisfies $\|\cdot\|_H \geq \sqrt{m}\|\cdot\|$. We begin by decomposing

$$F(\overline{w}_{\mathbf{Z}}) - \min_{k \in [K]} F_{\mathbf{Z}}(W^{(k)}) = \left(F(\overline{w}_{\mathbf{Z}}) - F_{\mathbf{Z}}(\overline{w}_{\mathbf{Z}}) \right) + \left(F_{\mathbf{Z}}(\overline{w}_{\mathbf{Z}}) - \min_{k \in [K]} F_{\mathbf{Z}}(W^{(k)}) \right)$$

$$=: E_1 + E_2,$$

where $E_1 \leq \sigma\sqrt{(cd/n)\log n}$ with probability at least $1 - \delta$, by Lemma 4.1. Next we control the term E_2. Under Assumption **(A.2)**, for any $w \in \mathbb{R}^d$,

$$\left| F_{\mathbf{Z}}(w) - F_{\mathbf{Z}}(\overline{w}_{\mathbf{Z}}) - \frac{1}{2}\|w - \overline{w}_{\mathbf{Z}}\|_H^2 \right| \leq \frac{L}{6}\|w - \overline{w}_{\mathbf{Z}}\|^3.$$

(Nesterov, 2013, Lemma 1.20). Therefore,

$$E_1 = \max_{K_1 \leq k \leq K} \left(F_{\mathbf{Z}}(\overline{w}_{\mathbf{Z}}) - F_{\mathbf{Z}}(W^{(k)}) \right)$$

$$\leq \max_{K_1 \leq k \leq K} \left(\frac{L}{6m^{3/2}}\|W^{(k)} - \overline{w}_{\mathbf{Z}}\|_H^3 - \frac{1}{2}\|W^{(k)} - \overline{w}_{\mathbf{Z}}\|_H^2 \right).$$

Then, by Theorem 2.1, with probability $1 - \delta$, either $\|W^{(k)} - \overline{w}_{\mathbf{Z}}\|_H \geq 2\varepsilon$ for some $k \leq K_0$ or $\|W^{(k)} - \overline{w}_{\mathbf{Z}}\|_H \leq 2\varepsilon$ for all $K_0 \leq k \leq K$. If the latter occurs, then $E_1 \leq 0$ for $\varepsilon \leq \dfrac{3m^{3/2}}{2L}$.

Acknowledgments

References

Nils Berglund and Barbara Gentz. Geometric singular perturbation theory for stochastic differential equations. *J. Differential Equations*, 191:1–54, 2003.

Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

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

Anton Bovier and Frank Den Hollander. *Metastability: A Potential-Theoretic Approach*. Springer, 2016.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes I: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79 (3):651–676, 2017.

Torgny Lindvall. *Lectures on the Coupling Method*. Dover Publications, Mineola, NY, 1992.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *Annals of Statistics*, 2016. To appear.

Peter Mörters and Yuval Peres. *Brownian Motion*, volume 30. Cambridge University Press, 2010.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2013.

Enzo Olivieri and Maria Eulália Vares. *Large Deviations and Metastability*. Cambridge University Press, 2004.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *COLT*, 2017.

Max Welling and Yee Whye Teh. Bayesian learning via Stochastic Gradient Langevin Dynamics. In *ICML*, 2011.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of Stochastic Gradient Langevin Dynamics. In *COLT*, 2017.

A   Proof of Lemma 3.1

Since $\{Q_{t_0}(t_1)Z_t^0\}_{t\in[t_0,t_1]}$ is a martingale and the function $x \mapsto e^{\gamma x^2}$ is convex for $\gamma > 0$, $\{\exp[(\beta\lambda/2)\|Q_{t_0}(t_1)Z_t^0\|^2]\}_{t\in[t_0,t_1]}$ is a positive submartingale. Therefore, using Doob's maximal inequality,

$$\mathbf{P}^{y_0}\left[\sup_{t_0\leq t\leq t_1}\|Q_{t_0}(t_1)Z_t^0\| \geq h\right] \leq \mathbf{P}^{y_0}\left[\sup_{t_0\leq t\leq t_1} e^{(\beta\lambda/2)\|Q_{t_0}(t_1)Z_t^0\|^2} \geq e^{\beta\lambda h^2/2}\right]$$

$$\leq e^{-\beta\lambda h^2/2}\mathbf{E}^{y_0}\left[e^{(\beta\lambda/2)\|Q_{t_0}(t_1)Z_{t_1}^0\|^2}\right]. \tag{A.1}$$

Since $Q_{t_0}(t_1)Z_{t_1}$ is a d-dimensional Gaussian random vector with mean μ and covariance matrix Σ, Lemma C.1 in Appendix C gives

$$\mathbf{E}^{y_0}\left[e^{(\beta\lambda/2)\|Q_{t_0}(t_1)Z_{t_1}^0\|^2}\right] = \frac{1}{\sqrt{\det(I-\beta\lambda\Sigma)}}\exp\left(\frac{\beta\lambda}{2}\langle\mu,(I-\beta\lambda\Sigma)^{-1}\mu\rangle\right)$$

$$\leq \left(\frac{1}{1-\lambda}\right)^{d/2}\exp\left(\frac{\beta\lambda}{2}\langle\mu,(I-\beta\lambda\Sigma)^{-1}\mu\rangle\right), \tag{A.2}$$

where the inequality follows from the fact that, by the choice of λ, the eigenvalues of $I-\beta\lambda\Sigma = (1-\lambda)I + \lambda e^{-2tH}$ are lower-bounded by $1-\lambda$. Putting (A.1)–(A.2) together, we get (3.2).

B   Proof of Lemma 4.1

Subgaussian noise conditions.   For any $w \in B^d(R)$, we have

$$|f(w,Z)| \leq |f(0,Z)| + |f(w,Z)-f(0,Z)|$$

$$\leq |f(0,Z)| + \sup_{v\in B^d(R)}\|\nabla f(v,Z)\|\|w\|$$

$$\leq |f(0,Z)| + \left(\|\nabla f(0,Z)\| + MR\right)\|w\|$$

$$\leq A + (B+MR)R.$$

Thus, by the Hoeffding lemma, the random variable $f(w,Z)$ is subgaussian: for any $t > 0$

$$\mathbf{E}\exp\left(t\left(f(w,Z)-F(w)\right)\right) \leq \exp\left(\frac{\sigma_0^2 t^2}{2}\right) \tag{B.1}$$

with $\sigma_0^2 = [A+(B+MR)R]^2$. For any $w \in B^d(R)$ and $v \in B^d(1)$, we have

$$|\langle v,\nabla f(w,Z)\rangle| \leq \|v\|\|\nabla f(0,Z)\| + \|v\|\|\nabla f(w,Z)-\nabla f(0,Z)\|$$

$$\leq B + MR.$$

Thus, by the Hoeffding lemma, the random variable $\langle v,\nabla f(w,Z)\rangle$ is subgaussian: for any $t > 0$,

$$\mathbf{E}\exp\left(t\langle v,\nabla f(w,Z)-\nabla F(w)\rangle\right) \leq \exp\left(\frac{\sigma_1^2 t^2}{2}\right) \tag{B.2}$$

14

with $\sigma_1^2 = (B+MR)^2$. In the same vein,

$$
\begin{aligned}
|\langle v, \nabla^2 F(w,Z)v\rangle| &\leq \|v\|^2\|\nabla^2 F(0,Z)\| + \|v\|^2\|\nabla^2 F(w,Z) - \nabla^2 F(0,Z)\| \\
&\leq C + LR,
\end{aligned}
$$

hence, for any $t > 0$,

$$
\mathbf{E}\exp\left(t\left\langle v, \left(\nabla^2 f(w,Z) - \nabla^2 F(w)\right)v\right\rangle\right) \leq \exp\left(\frac{\sigma_2^2 t^2}{2}\right) \tag{B.3}
$$

with $\sigma_2^2 = (C+LR)^2$.

Uniform deviation bound for the risk. For any two $w, w' \in \mathsf{B}^d(1)$,

$$
\begin{aligned}
|f(w,Z) - f(w',Z)| &\leq \sup_{v \in \mathsf{B}^d(R)} \|\nabla f(v,Z)\|\|w - w'\| \\
&\leq (B+MR)\|w - w'\|.
\end{aligned}
$$

Therefore, if we let $\{w_1, \dots, w_N\}$ be an ε-cover of $\mathsf{B}^d(R)$, with $N \leq (3R/\varepsilon)^d$, then

$$
\sup_{w \in \mathsf{B}^d(R)}\left|\frac{1}{n}\sum_{i=1}^n f(w,Z_i) - F(w)\right| \leq \max_{1 \leq j \leq N}\left|\frac{1}{n}\sum_{i=1}^n f(w_j,Z_i) - F(w_j)\right| + 2(B+MR)\varepsilon.
$$

Hence, for any $t > 4(B+MR)\varepsilon$, we have

$$
\begin{aligned}
\mathbf{P}&\left[\sup_{w \in \mathsf{B}^d(R)}\left|\frac{1}{n}\sum_{i=1}^n f(w,Z_i) - F(w)\right| \geq t\right] \\
&\leq \mathbf{P}\left[\max_{1 \leq j \leq N}\left|\frac{1}{n}\sum_{i=1}^n f(w_j,Z_i) - F(w_j)\right| \geq \frac{t}{2}\right] \\
&\leq \left(\frac{3R}{\varepsilon}\right)^d \max_{1 \leq j \leq N}\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^n f(w_j,Z_i) - F(w_j)\right| \geq \frac{t}{2}\right] \\
&\leq \left(\frac{3R}{\varepsilon}\right)^d \exp\left(-\frac{nt^2}{32\sigma_0^2}\right),
\end{aligned}
$$

where we have used the fact that $f(w,Z)$ is σ_0^2-subgaussian [cf. (B.1)]. Choosing $\varepsilon = \frac{\sigma_0}{4(B+MR)dn}$ and

$$
t \geq \frac{\sigma_0}{dn} \vee \sqrt{\frac{32\sigma_0^2}{n}\left(d\log\frac{144(B+MR)Rdn}{\sigma_0} + \log\frac{1}{\delta}\right)},
$$

we see that (4.2) holds with probability at least $1 - \delta$ for a suitable choice of c_0.

15

Uniform deviation bound for the gradient. Let $\{v_1, \ldots, v_J\}$ be a $(1/2)$-cover of $\mathsf{B}^d(1)$, with $J \le 6^d$. Then, for any $w \in \mathsf{B}^d(R)$, using the fact that $\langle v, f(w, Z) \rangle$ is σ_1^2-subgaussian [cf. (B.2)],

$$
\mathbf{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(w, Z_i) - \nabla F(w)\right\| \ge t\right]
$$

$$
\le \mathbf{P}\left[\max_{1 \le j \le J}\left\langle v_j, \frac{1}{n}\sum_{i=1}^{n}\nabla f(w, Z_i) - \nabla F(w)\right\rangle \ge \frac{t}{2}\right]
$$

$$
\le 6^d \exp\left(-\frac{nt^2}{8\sigma_1^2}\right).
$$

Next, let $\{w_1, \ldots, w_N\}$ be an ε-cover of $\mathsf{B}^d(R)$, with $N \le (3R/\varepsilon)^d$. Then, since

$$
\sup_{w \in \mathsf{B}^d(R)}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(w, Z_i) - \nabla F(w)\right\| \le \max_{1 \le j \le N}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(w_j, Z_i) - \nabla F(w_j)\right\| + 2M\varepsilon,
$$

for any $t > 4M\varepsilon$ we have

$$
\mathbf{P}\left[\sup_{w \in \mathsf{B}^d(R)}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(w, Z_i) - \nabla F(w)\right\| \ge t\right]
$$

$$
\le \mathbf{P}\left[\max_{1 \le j \le N}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(w_j, Z_i) - \nabla F(w_j)\right\| \ge \frac{t}{2}\right]
$$

$$
\le \left(\frac{3R}{\varepsilon}\right)^d \max_{1 \le j \le N}\mathbf{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(w_j, Z_i) - \nabla F(w_j)\right\| \ge \frac{t}{2}\right]
$$

$$
\le \left(\frac{18R}{\varepsilon}\right)^d \exp\left(-\frac{nt^2}{32\sigma_1^2}\right).
$$

Choosing $\varepsilon = \frac{\sigma_1}{4Mdn}$ and

$$
t \ge \frac{\sigma_1}{dn} \vee \sqrt{\frac{32\sigma_1^2}{n}\left(d\log\frac{72MRdn}{\sigma_1} + \log\frac{1}{\delta}\right)},
$$

we see that (4.2) holds with probability at least $1 - \delta$ for a suitable choice of c_0.

Uniform deviation bound for the Hessian. Now, let $\{v_1, \ldots, v_J\}$ be an $(1/4)$-cover of $B^d(1)$, with $J \le 12^d$. Then, for any $w \in B^d(R)$, using the fact that $\langle v, \nabla^2 f(w, Z)v \rangle$ is σ_2^2-subgaussian [cf. (B.3)],

$$\mathbf{P}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f(w, Z_i) - \nabla^2 F(w) \right\| \ge t \right]$$

$$\le \mathbf{P}\left[\max_{1 \le j \le J} \left\langle v_j, \left(\frac{1}{n} \sum_{i=1}^{n} \nabla^2 f(w, Z_i) - \nabla^2 F(w) \right) v_j \right\rangle \ge \frac{t}{2} \right]$$

$$\le 12^d \exp\left(-\frac{nt^2}{8\sigma_2^2} \right).$$

Again, let $\{w_1, \ldots, w_N\}$ be an ε-cover of $B^d(R)$. Then, since

$$\sup_{w \in B^d(R)} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f(w, Z_i) - \nabla^2 F(w) \right\| \le \max_{1 \le j \le N} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f(w_j, Z_i) - \nabla^2 F(w_j) \right\| + 2L\varepsilon,$$

for any $t > 4L\varepsilon$ we have

$$\mathbf{P}\left[\sup_{w \in B^d(R)} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f(w, Z_i) - \nabla^2 F(w) \right\| \ge t \right]$$

$$\le \mathbf{P}\left[\max_{1 \le j \le N} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f(w_j, Z_i) - \nabla^2 F(w_j) \right\| \ge \frac{t}{2} \right]$$

$$\le \left(\frac{3R}{\varepsilon} \right)^d \max_{1 \le j \le N} \mathbf{P}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f(w_j, Z_i) - \nabla^2 F(w_j) \right\| \ge \frac{t}{2} \right]$$

$$\le \left(\frac{36R}{\varepsilon} \right)^d \exp\left(-\frac{nt^2}{32\sigma_1^2} \right).$$

Choosing $\varepsilon = \frac{\sigma_1}{4Ldn}$ and

$$t \ge \frac{\sigma_2}{dn} \vee \sqrt{ \frac{32\sigma_2^2}{n} \left(d \log \frac{144LRdn}{\sigma_2} + \log \frac{1}{\delta} \right) },$$

we see that (4.2) holds with probability at least $1 - \delta$ for a suitable choice of c_0.

C   A Gaussian integral

Lemma C.1. *Consider a random variable $V \sim \mathcal{N}(\mu, \Sigma)$. Then, for any $\gamma > 0$ such that $I - 2\gamma\Sigma$ is positive definite,*

$$\mathbf{E}[e^{\gamma \|V\|^2}] = \frac{1}{\sqrt{\det(I - 2\gamma\Sigma)}} \exp\left(\gamma \langle \mu, (I - 2\gamma\Sigma)^{-1} \mu \rangle \right). \tag{C.1}$$

17

Proof. We have

$$\mathbf{E}[e^{\gamma\|V\|^2}] = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int_{\mathbb{R}^d} e^{-\frac{1}{2}\langle(v-\mu),\Sigma^{-1}(v-\mu)\rangle} e^{\gamma\langle v,v\rangle} dv$$

$$= \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int_{\mathbb{R}^d} e^{-\frac{1}{2}[\langle(v-\mu),\Sigma^{-1}(v-\mu)\rangle-2\gamma\langle v,v\rangle]} dv. \tag{C.2}$$

We need to complete the square in the exponent:

$$\langle(v-\mu),\Sigma^{-1}(v-\mu)\rangle - 2\gamma\langle v,v\rangle$$
$$= \langle v,\underbrace{(\Sigma^{-1}-2\gamma I)}_{K}v\rangle - 2\langle\underbrace{\Sigma^{-1}\mu}_{Kp},v\rangle + \langle\mu,\Sigma^{-1}\mu\rangle$$
$$= \langle v-p,K(v-p)\rangle + \langle(I-K^{-1}\Sigma^{-1})\mu,\Sigma^{-1}\mu\rangle$$
$$= \langle v-p,K(v-p)\rangle + \langle\mu,(\Sigma^{-1}-\Sigma^{-1}K^{-1}\Sigma^{-1})\mu\rangle. \tag{C.3}$$

Now we recall the Woodbury matrix identity: If $B = A + UV$, such that A and $I + VA^{-1}U$ are invertible, then

$$B^{-1} = A^{-1} - A^{-1}U(I + VA^{-1}U)^{-1}VA^{-1}. \tag{C.4}$$

Applying (C.4) with $B = K$, $A = \Sigma^{-1}$, $U = -2\gamma I$, $V = I$, we can write

$$K^{-1} = (\Sigma^{-1}-2\gamma I)^{-1}$$
$$= \Sigma + 2\gamma\Sigma(I-2\gamma\Sigma)^{-1}\Sigma,$$

which gives

$$\Sigma^{-1} - \Sigma^{-1}K^{-1}\Sigma^{-1} = \Sigma^{-1} - \Sigma^{-1}\left(\Sigma + 2\gamma\Sigma(I-2\gamma\Sigma)^{-1}\Sigma\right)\Sigma^{-1}$$
$$= \Sigma^{-1} - \Sigma^{-1} - 2\gamma(I-2\gamma\Sigma)^{-1}$$
$$= -2\gamma(I-2\gamma\Sigma)^{-1}.$$

Substituting this into (C.3), we get

$$\langle(v-\mu),\Sigma^{-1}(v-\mu)\rangle - 2\gamma\langle v,v\rangle = \langle(v-p),K(v-p)\rangle - 2\gamma\langle\mu,(I-2\gamma\Sigma)^{-1}\mu\rangle.$$

We can now compute the Gaussian integral in (C.2):

$$\frac{1}{\sqrt{\det(2\pi\Sigma)}} \int_{\mathbb{R}^d} e^{-\frac{1}{2}[\langle(v-\mu),\Sigma^{-1}(v-\mu)\rangle-2\gamma\langle v,v\rangle]} dv$$
$$= \frac{1}{\sqrt{\det(K\Sigma)}} \exp\left(\gamma\langle\mu,(I-2\gamma\Sigma)^{-1}\mu\rangle\right)$$
$$= \frac{1}{\sqrt{\det(I-2\gamma\Sigma)}} \exp\left(\gamma\langle\mu,(I-2\gamma\Sigma)^{-1}\mu\rangle\right),$$

which gives (C.1). $\qquad\square$