A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics

Yuchen Zhang   ZHANGYUC@CS.STANFORD.EDU
Computer Science Department, Stanford University

Percy Liang   PLIANG@CS.STANFORD.EDU
Computer Science Department, Stanford University

Moses Charikar   MOSES@CS.STANFORD.EDU
Computer Science Department, Stanford University

Abstract

We study the Stochastic Gradient Langevin Dynamics (SGLD) algorithm for non-convex optimization. The algorithm performs stochastic gradient descent, where in each step it injects appropriately scaled Gaussian noise to the update. We analyze the algorithm's hitting time to an arbitrary subset of the parameter space. Two results follow from our general theory: First, we prove that for empirical risk minimization, if the empirical risk is pointwise close to the (smooth) population risk, then the algorithm finds an approximate local minimum of the population risk in polynomial time, escaping suboptimal local minima that only exist in the empirical risk. Second, we show that SGLD improves on one of the best known learnability results for learning linear classifiers under the zero-one loss.

1. Introduction

A central challenge of non-convex optimization is avoiding sub-optimal local minima. Although escaping all local minima is NP-hard in general (e.g. Blum and Rivest, 1992), one might expect that it should be possible to escape "appropriately shallow" local minima, whose basins of attraction have relatively low barriers. As an illustrative example, consider minimizing an empirical risk function in Figure 1. As the figure shows, although the empirical risk is uniformly close to the population risk, it contains many poor local minima that don't exist in the population risk. Gradient descent is unable to escape such local minima.

A natural workaround is to inject random noise to the gradient. Empirically, adding gradient noise has been found to improve learning for deep neural networks and other non-convex models (Neelakantan et al., 2015a,b; Kurach et al., 2015; Kaiser and Sutskever, 2015; Zeyer et al., 2016). However, theoretical understanding of the value of gradient noise is still incomplete. For example, Ge et al. (2015) show that by adding isotropic noise w and by choosing a sufficiently small stepsize η, the iterative update:

$$x \leftarrow x - \eta \left(\nabla f(x) + w \right) \tag{1}$$

is able to escape strict saddle points. Unfortunately, this approach, as well as the subsequent line of work on escaping saddle points (Lee et al., 2016; Anandkumar and Ge, 2016; Agarwal et al., 2016), doesn't guarantee escaping even shallow local minima.

Another line of work in Bayesian statistics studies the Langevin Monte Carlo (LMC) method (Roberts and Tweedie, 1996), which employs an alternative noise term. Given a function f, LMC
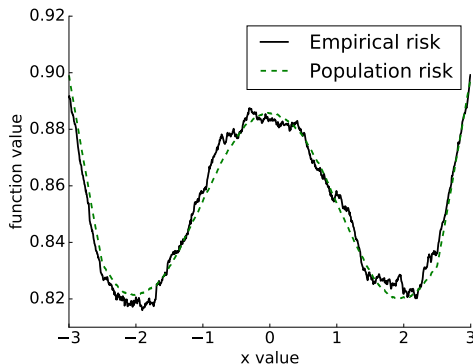
Figure 1: Empirical risk (sample size $=$ 5000) versus population risk (sample size $\to \infty$) on one-dimensional zero-one losses. The two functions are uniformly close, but the empirical risk contains local minima that that are far worse than the population local minima.

performs the iterative update:

$$x \leftarrow x - \eta \left(\nabla f(x) + \sqrt{2/(\eta\xi)}\, w \right) \quad \text{where} \quad w \sim N(0, I), \tag{2}$$

where $\xi > 0$ is a "temperature" hyperparameter. Unlike the bounded noise added in formula (1), LMC adds a large noise term that scales with $\sqrt{1/\eta}$. With a small enough η, the noise dominates the gradient, enabling the algorithm to escape any local minimum. For empirical risk minimization, one might substitute the exact gradient $\nabla f(x)$ with a stochastic gradient, which gives the Stochastic Gradient Langevin Dynamics (SGLD) algorithm (Welling and Teh, 2011). It can be shown that both LMC and SGLD asymptotically converge to a stationary distribution $\mu(x) \propto e^{-\xi f(x)}$ (Roberts and Tweedie, 1996; Teh et al., 2016). As $\xi \to \infty$, the probability mass of μ concentrates on the global minimum of the function f, and the algorithm asymptotically converges to a neighborhood of the global minimum.

Despite asymptotic consistency, there is no theoretical guarantee that LMC is able to find the global minimum of a general non-convex function, or even a local minimum of it, in polynomial time. Recent works focus on bounding the mixing time (i.e. the time for converging to μ) of LMC and SGLD. Bubeck et al. (2015), Dalalyan (2016) and Bonis (2016) prove that on convex functions, LMC converges to the stationary distribution in polynomial time. On non-convex functions, however, an exponentially long mixing time is unavoidable in general. According to Bovier et al. (2004), it takes the Langevin diffusion at least $e^{\Omega(\xi h)}$ time to escape a depth-h basin of attraction. Thus, if the function contains multiple "deep" basins with $h = \Omega(1)$, then the mixing time is lower bounded by $e^{\Omega(\xi)}$.

In parallel work to this paper, Raginsky et al. (2017) upper bound the time of SGLD converging to an approximate global minimum of non-convex functions. They show that the upper bound is polynomial in the inverse of a quantity they call the *uniform spectral gap*. Similar to the mixing time bound, in the presence of multiple local minima, the convergence time to an approximate global minimum can be exponential in dimension d and the temperature parameter ξ.

Contributions   In this paper, we present an alternative analysis of SGLD algorithm.[1] Instead of bounding its mixing time, we bound the algorithm's *hitting time* to an arbitrary set U on a general non-convex function. The hitting time captures the algorithm's optimization efficiency, and more importantly, it enjoys polynomial rates for hitting appropriately chosen sets regardless of the mixing time, which could be exponential. We highlight two consequences of the generic bound: First, under suitable conditions, SGLD hits an approximate local minimum of f, with a hitting time that is polynomial in dimension d and all hyperparameters; this extends the polynomial-time guarantees proved for convex functions (Bubeck et al., 2015; Dalalyan, 2016; Bonis, 2016). Second, the time complexity bound is stable, in the sense that any $\mathcal{O}(1/\xi)$ perturbation in ℓ_∞-norm of the function f doesn't significantly change the hitting time. This second property is the main strength of SGLD: For any function f, if there exists another function F such that $\|f - F\|_\infty = \mathcal{O}(1/\xi)$, then we define the set U to be the approximate local minima of F. The two properties together imply that even if we execute SGLD on function f, it hits an approximate local minimum of F in polynomial time. In other words, SGLD is able to escape "shallow" local minima of f that can be eliminated by slightly perturbing the function.

This stability property is useful in studying empirical risk minimization (ERM) in situations where the empirical risk f is pointwise close to the population risk F, but has poor local minima that don't exist in the population risk. This phenomenon has been observed in statistical estimation with non-convex penalty functions (Wang et al., 2014; Loh and Wainwright, 2015), as well as in minimizing the zero-one loss (see Figure 1). Under this setting, our result implies that SGLD achieves an approximate local minimum of the (smooth) *population risk* in polynomial time, ruling out local minima that only exist in the empirical risk. It improves over recent results on non-convex optimization (Ge et al., 2015; Lee et al., 2016; Anandkumar and Ge, 2016; Agarwal et al., 2016), which compute approximate local minima only for the empirical risk.

As a concrete application, we prove a stronger learnability result for the problem of learning linear classifiers under the zero-one loss (Arora et al., 1993), which involves non-convex and non-smooth empirical risk minimization. Our result improves over the recent result of Awasthi et al. (2015): the method of Awasthi et al. (2015) handles noisy data corrupted by a very small Massart noise (at most 1.8×10^{-6}), while our algorithm handles Massart noise up to any constant less than 0.5. As a Massart noise of 0.5 represents completely random observations, we see that SGLD is capable of learning from very noisy data.

Techniques   The key step of our analysis is to define a positive quantity called the *restricted Cheeger constant*. This quantity connects the hitting time of SGLD, the geometric properties of the objective function, and the stability of the time complexity bound. For an arbitrary function $f : K \to \mathbb{R}$ and an arbitrary set $V \subset K$, the restricted Cheeger constant is defined as the minimal ratio between the surface area of a subset $A \subset V$ and its volume with respect to a measure $\mu(x) \propto e^{-f(x)}$. We prove that the hitting time is polynomial in the inverse of the restricted Cheeger constant (Section 2.3). The stability of the time complexity bound follows as a natural consequence of the definition of this quantity (Section 2.2). We then develop techniques to lower bound the restricted Cheeger constant based on geometric properties of the objective function (Section 2.4).

Notation   For any positive integer n, we use $[n]$ as a shorthand for the discrete set $\{1, 2, \ldots, n\}$. For a rectangular matrix A, let $\|A\|_*$ be its nuclear norm (i.e., the sum of singular values), and $\|A\|_2$

---

1. The theory holds for the standard LMC algorithm as well.

Algorithm 1 Stochastic Gradient Langevin Dynamics

Input: Objective function $f : K \to \mathbb{R}$; hyperparameters $(\xi, \eta, k_{\max}, D)$.

1. Initialize $x_0 \in K$ by uniformly sampling from the parameter space K.

2. For each $k \in \{1, 2, \ldots, k_{\max}\}$: Sample $w \sim N(0, I_{d \times d})$. Compute a stochastic gradient $g(x_{k-1})$ such that $\mathbb{E}[g(x_{k-1})|x_{k-1}] = \nabla f(x_{k-1})$. Then update:

$$y_k = x_{k-1} - \eta \, g(x_{k-1}) + \sqrt{2\eta/\xi} \, w; \tag{3a}$$

$$x_k = \begin{cases} y_k & \text{if } y_k \in K \cap \mathcal{B}(x_{k-1}; D), \\ x_{k-1} & \text{otherwise.} \end{cases} \tag{3b}$$

Output: $\widehat{x} = x_{k^*}$ where $k^* := \operatorname{argmin}_k\{f(x_k)\}$.

---

be its spectral norm (i.e., the maximal singular value). For any point $x \in \mathbb{R}^d$ and an arbitrary set $V \subset \mathbb{R}^d$, we denote their Euclidean distance by $d(x, V) := \inf_{y \in V} \|x - y\|_2$. We use $\mathcal{B}(x; r)$ to denote the Euclidean ball of radius r that centers at point x.

2. Algorithm and main results

In this section, we define the algorithm and the basic concepts, then present the main theoretical results of this paper.

2.1. The SGLD algorithm

Our goal is to minimize a function f in a compact parameter space $K \subset \mathbb{R}^d$. The SGLD algorithm (Welling and Teh, 2011) is summarized in Algorithm 1. In step (3a), the algorithm performs SGD on the function f, then adds Gaussian noise to the update. Step (3b) ensures that the vector x_k always belong to the parameter space, and is not too far from x_{k-1} of the previous iteration.[2] After $k_{\max}$ iterations, the algorithm returns a vector $\widehat{x}$. Although standard SGLD returns the last iteration's output, we study a variant of the algorithm which returns the best vector across all iterations. This choice is important for our analysis of hitting time. We note that evaluating $f(x_k)$ can be computationally more expensive than computing the stochastic gradient g_k, because the objective function is defined on the entire dataset, while the stochastic gradient can be computed via a single instance. Returning the best x_k merely facilitates theoretical analysis and might not be necessary in practice.

Because of the noisy update, the sequence $(x_0, x_1, x_2, \ldots)$ asymptotically converges to a stationary distribution rather than a stationary point (Teh et al., 2016). Although this fact introduces challenges to the analysis, we show that its non-asymptotic efficiency can be characterized by a positive quantity called the *restricted Cheeger constant*.

---

2. The hyperparameter D can be chosen large enough so that the constraint $y_k \in \mathcal{B}(x_{k-1}; D)$ is satisfied with high probability, see Theorem 1.

2.2. Restricted Cheeger constant

For any measurable function f, we define a probability measure μ_f whose density function is:

$$\mu_f(x) := \frac{e^{-f(x)}}{\int_K e^{-f(x)}dx} \propto e^{-f(x)} \quad \text{for all} \quad x \in K. \tag{4}$$

For any function f and any subset $V \subset K$, we define the restricted Cheeger constant as:

$$\mathcal{C}_f(V) := \liminf_{\epsilon \searrow 0} \inf_{A \subset V} \frac{\mu_f(A_\epsilon) - \mu_f(A)}{\epsilon \, \mu_f(A)}, \quad \text{where} \quad A_\epsilon := \{x \in K : d(x, A) \leq \epsilon\}. \tag{5}$$

The restricted Cheeger constant generalizes the notion of the Cheeger isoperimetric constant (Cheeger, 1969), quantifying how well a subset of V can be made as least connected as possible to the rest of the parameter space. The connectivity is measured by the ratio of the surface measure $\liminf_{\epsilon \searrow 0} \frac{\mu_f(A_\epsilon) - \mu_f(A)}{\epsilon}$ to the set measure $\mu_f(A)$. Intuitively, this quantifies the chance of escaping the set A under the probability measure μ_f.

Stability of restricted Cheeger constant A property that will be important in the sequal is that the restricted Cheeger constant is stable under perturbations: if we perturb f by a small amount, then the values of μ_f won't change much, so that the variation on $\mathcal{C}_f(V)$ will also be small. More precisely, for functions f_1 and f_2 satisfying $\sup_{x \in K} |f_1(x) - f_2(x)| = \nu$, we have

$$\mathcal{C}_{f_1}(V) = \liminf_{\epsilon \searrow 0} \inf_{A \subset V} \frac{\int_{A_\epsilon \setminus A} e^{-f_1(x)}dx}{\epsilon \int_A e^{-f_1(x)}dx} \geq \liminf_{\epsilon \searrow 0} \inf_{A \subset V} \frac{\int_{A_\epsilon \setminus A} e^{-f_2(x)-\nu}dx}{\epsilon \int_A e^{-f_2(x)+\nu}dx} = e^{-2\nu}\mathcal{C}_{f_2}(V), \tag{6}$$

and similarly $\mathcal{C}_{f_2}(V) \geq e^{-2\nu}\mathcal{C}_{f_1}(V)$. As a result, if two functions f_1 and f_2 are uniformly close, then we have $\mathcal{C}_{f_1}(V) \approx \mathcal{C}_{f_2}(V)$ for a constant ν. This property enables us to lower bound $\mathcal{C}_{f_1}(V)$ by lower bounding the restricted Cheeger constant of an alternative function $f_2 \approx f_1$, which might be easier to analyze.

2.3. Generic non-asymptotic bounds

We make several assumptions on the parameter space and on the objective function.

Assumption A (parameter space and objective function)

- *The parameter space K satisfies: there exists $h_{max} > 0$, such that for any $x \in K$ and any $h \leq h_{max}$, the random variable $y \sim N(x, 2hI)$ satisfies $P(y \in K) \geq \frac{1}{3}$.*

- *The function $f : K \to [0, B]$ is bounded, differentiable and L-smooth in K, meaning that for any $x, y \in K$, we have $|f(y) - f(x) - \langle y - x, \nabla f(x)\rangle| \leq \frac{L}{2}\|y - x\|_2^2$.*

- *The stochastic gradient vector $g(x)$ has sub-exponential tails: there exists $b_{max} > 0$, $G > 0$, such that given any $x \in K$ and any vector $u \in \mathbb{R}^d$ satisfying $\|u\|_2 \leq b_{max}$, the vector $g(x)$ satisfies $\mathbb{E}\left[e^{\langle u, g(x)\rangle^2} \mid x\right] \leq \exp(G^2\|u\|_2^2)$.*

The first assumption states that the parameter space doesn't contain sharp corners, so that the update (3b) won't be stuck at the same point for too many iterations. It can be satisfied, for example, by defining the parameter space to be an Euclidean ball and choosing $h_{\max} = o(d^{-2})$. The probability $1/3$ is arbitrary and can be replaced by any constant in $(0, 1/2)$. The second assumption

requires the function f to be smooth. We show how to handle non-smooth functions in Section 3 by appealing to the stability property of the restricted Cheeger constant discussed earlier. The third assumption requires the stochastic gradient to have sub-exponential tails, which is a standard assumption in stochstic optimization.

Theorem 1 *Assume that Assumption A holds. For any subset $U \subset K$ and any $\xi, \rho, \delta > 0$, there exist $\eta_0 > 0$ and $k_{max} \in \mathbb{Z}^+$, such that if we choose any stepsize $\eta \in (0, \eta_0]$ and hyperparameter $D := 4\sqrt{2\eta d/\xi}$, then with probability at least $1 - \delta$, SGLD after $k_{\max}$ iterations returns a solution $\widehat{x}$ satisfying:*

$$f(\widehat{x}) \leq \sup_{x:\, d(x,U) \leq \rho} f(x). \tag{7}$$

The iteration number k_{max} is bounded by

$$k_{max} \leq \frac{M}{\left(\mathcal{C}_{(\xi f)}(K \backslash U)\right)^4} \tag{8}$$

where the numerator M is polynomial in $(B, L, G, \log(1/\delta), d, \xi, \eta_0/\eta, h_{max}^{-1}, b_{max}^{-1}, \rho^{-1})$. See Appendix B.2 for the explicit polynomial dependence.

Theorem 1 is a generic result that applies to all optimization problems satisfying Assumption A. The right-hand side of the bound (7) is determined by the choice of U. If we choose U to be the set of (approximate) local minima, and let $\rho > 0$ be sufficiently small, then $f(\widehat{x})$ will roughly be bounded by the worst local minimum. The theorem permits ξ to be arbitrary provided the stepsize η is small enough. Choosing a larger ξ means adding less noise to the SLGD update, which means that the algorithm will be more efficient at finding a stationary point, but less efficient at escaping local minima. Such a trade-off is captured by the restricted Cheeger constant in inequality (8) and will be rigorously studied in the next subsection.

The iteration complexity bound is governed by the restricted Cheeger constant. For any function f and any target set U with a positive Borel measure, the restricted Cheeger constant is strictly positive (see Appendix A), so that with a small enough η, the algorithm always converges to the global minimum asymptotically. We remark that the SGD doesn't enjoy the same asymptotic optimality guarantee, because it uses a $O(\eta)$ gradient noise in contrast to SGLD's $O(\sqrt{\eta})$ one. Since the convergence theory requires a small enough η, we often have $\eta \ll \sqrt{\eta}$. the SGD noise is too conservative to allow the algorithm to escape local minima.

Proof sketch The proof of Theorem 1 is fairly technical. We defer the full proof to Appendix B, only sketching the basic proof ideas here. At a high level, we establish the theorem by bounding the hitting time of the Markov chain $(x_0, x_1, x_2, \dots)$ to the set $U_\rho := \{x : d(x, U) \leq \rho\}$. Indeed, if some x_k hits the set, then:

$$f(\widehat{x}) \leq f(x_k) \leq \sup_{x \in U_\rho} f(x),$$

which establishes the risk bound (7).

In order to bound the hitting time, we construct a time-reversible Markov chain, and prove that its hitting time to U_ρ is on a par with the original hitting time. To analyze this second Markov chain, we define a notion called the *restricted conductance*, which measures how easily the Markov chain can transition between states within $K \backslash U_\rho$. This quantity is related to the notion of *conductance* in the analysis of time-reversible Markov processes (Lovász and Simonovits, 1993), but the ratio

6

between these two quantities can be exponentially large for non-convex f. We prove that the hitting time of the second Markov chain depends inversely on the restricted conductance, so that the problem reduces to lower bounding the restricted conductance.

Finally, we lower bound the restricted conductance by the restricted Cheeger constant. The former quantity characterizes the Markov chain, while the later captures the geometric properties of the function f. Thus, we must analyze the SGLD algorithm in depth to establish a connection between them. Once we prove this lower bound, putting all pieces together completes the proof. ∎

2.4. Lower bounding the restricted Cheeger constant

In this subsection, we prove lower bounds on the restricted Cheeger constant $\mathcal{C}_{(\xi f)}(K \backslash U)$ in order to flesh out the iteration complexity bound of Theorem 1. We start with a lower bound for the class of convex functions:

Proposition 2 *Let K be a d-dimensional unit ball. For any convex G-Lipschitz continuous function f and any $\epsilon > 0$, let the set of ϵ-optimal solutions be defined by:*

$$U := \{x \in K : f(x) \leq \inf_{y \in K} f(y) + \epsilon\}.$$

Then for any $\xi \geq \frac{2d \log(4G/\epsilon)}{\epsilon}$, we have $\mathcal{C}_{(\xi f)}(K \backslash U) \geq 1$.

The proposition shows that if we choose a big enough ξ, then $\mathcal{C}_{(\xi f)}(K \backslash U)$ will be lower bounded by a universal constant. The lower bound is proved based on an isoperimetric inequality for log-concave distributions. See Appendix C for the proof.

For non-convex functions, directly proving the lower bound is difficult, because the definition of $\mathcal{C}_{(\xi f)}(K \backslash U)$ involves verifying the properties of all subsets $A \subset K \backslash U$. We start with a generic lemma that reduces the problem to checking properties of all points in $K \backslash U$.

Lemma 3 *Consider an arbitrary continuously differentiable vector field $\phi : K \to \mathbb{R}^d$ and a positive number $\epsilon_0 > 0$ such that:*

$$\|\phi(x)\|_2 \leq 1 \quad and \quad x - \epsilon \phi(x) \in K \quad for \ any \quad \epsilon \in [0, \epsilon_0], \ x \in K. \tag{9}$$

For any continuously differentiable function $f : K \to \mathbb{R}$ and any subset $V \subset K$, the restricted Cheeger constant $\mathcal{C}_f(V)$ is lower bounded by

$$\mathcal{C}_f(V) \geq \inf_{x \in V} \left\{ \langle \phi(x), \nabla f(x) \rangle - div \, \phi(x) \right\} \quad where \quad div \, \phi(x) := \sum_{i=1}^{d} \frac{\partial \phi_i(x)}{\partial x_i}.$$

Lemma 3 reduces the problem of lower bounding $\mathcal{C}_f(V)$ to the problem of finding a proper vector field ϕ and verifying its properties for all points $x \in V$. Informally, the quantity $\mathcal{C}_f(V)$ measures the chance of escaping the set V. The lemma shows that if we can construct an "oracle" vector field ϕ, such that at every point $x \in V$ it gives the correct direction (i.e. $-\phi(x)$) to escape V, but always stay in K, then we obtain a strong lower bound on $\mathcal{C}_f(V)$. This construction is merely for the theoretical analysis and doesn't affect the execution of the algorithm.
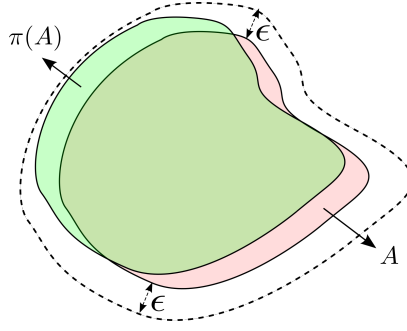
Figure 2: Consider a mapping $\pi : x \mapsto x - \epsilon\phi(x)$. If the conditions of Lemma 3 hold, then we have $\pi(A) \subset A_\epsilon$ and consequentely $\mu_f(\pi(A)) \leq \mu_f(A_\epsilon)$. We use inequality (10) to lower bound the restricted Cheeger constant.

Proof sketch The proof idea is illustrated in Figure 2: by constructing a mapping $\pi : x \mapsto x - \epsilon\phi(x)$ that satisfies the conditions of the lemma, we obtain $\pi(A) \subset A_\epsilon$ for all $A \subset V$, and consequently $\mu_f(\pi(A)) \leq \mu_f(A_\epsilon)$. Then we are able to lower bound the restricted Cheeger constant by:

$$\mathcal{C}_f(V) \geq \liminf_{\epsilon \searrow 0} \inf_{A \subset V} \frac{\mu_f(\pi(A)) - \mu_f(A)}{\epsilon\, \mu_f(A)} = \liminf_{\epsilon \searrow 0} \inf_{dA \subset V} \frac{1}{\epsilon}\Big(\frac{\mu_f(\pi(dA))}{\mu_f(dA)} - 1\Big), \qquad (10)$$

where dA is an infinitesimal of the set V. It can be shown that the right-hand side of inequality (10) is equal to $\inf_{x \in V}\{\langle\phi(x), \nabla f(x)\rangle - \operatorname{div}\phi(x)\}$, which establishes the lemma. See Appendix D for a rigorous proof. ∎

Before demonstrating the applications of Lemma 3, we make several additional mild assumptions on the parameter space and on the function f.

Assumption B (boundary condition and smoothness)

- *The parameter space K is a d-dimensional ball of radius $r > 0$ centered at the origin. There exists $r_0 > 0$ such that for every point x satisfying $\|x\|_2 \in [r - r_0, r]$, we have $\langle x, \nabla f(x)\rangle \geq \|x\|_2$.*

- *For some $G, L, H > 0$, the function f is third-order differentiable with $\|\nabla f(x)\|_2 \leq G$, $\|\nabla^2 f(x)\|_* \leq L$ and $\|\nabla^2 f(x) - \nabla^2 f(y)\|_* \leq H\|x - y\|_2$ for any $x, y \in K$.*

The first assumption requires the parameter space to be an Euclidean ball and imposes a gradient condition on its boundary. This is made mainly for the convenience of theoretical analysis. We remark that for any function f, the condition on the boundary can be satisfied by adding a smooth barrier function $\rho(\|x\|_2)$ to it, where the function $\rho(t) = 0$ for any $t < r - 2r_0$, but sharply increases on the interval $[r - r_0, r]$ to produce large enough gradients. The second assumption requires the function f to be third-order differentiable. We shall relax the second assumption in Section 3.

The following proposition describes a lower bound on $\mathcal{C}_{(\xi f)}(K \backslash U)$ when f is a smooth function and the set U consists of approximate stationary points. Although we shall prove a stronger result, the proof of this proposition is a good example for demonstrating the power of Lemma 3.

Proposition 4 *Assume that Assumption B holds. For any $\epsilon > 0$, define the set of ϵ-approximate stationary points $U := \{x \in K : \|\nabla f(x)\|_2 < \epsilon\}$. For any $\xi \geq 2L/\epsilon^2$, we have $\mathcal{C}_{(\xi f)}(K \backslash U) \geq \frac{\xi \epsilon^2}{2G}$.*

Proof Recall that G is the Lipschitz constant of function f. Let the vector field be defined by $\phi(x) := \frac{1}{G}\nabla f(x)$, then we have $\|\phi(x)\|_2 \leq 1$. By Assumption B, it is easy to verify that the conditions of Lemma 3 hold. For any $x \in K \backslash U$, the fact that $\|\nabla f(x)\|_2 \geq \epsilon$ implies:

$$\langle \phi(x), \xi \nabla f(x) \rangle = \frac{\xi}{G}\|\nabla f(x)\|_2^2 \geq \frac{\xi \epsilon^2}{G}.$$

Recall that L is the smoothness parameter. By Assumption B, the divergence of $\phi(x)$ is upper bounded by $\text{div}\,\phi(x) = \frac{1}{G}\text{tr}(\nabla^2 f(x)) \leq \frac{1}{G}\|\nabla^2 f(x)\|_* \leq \frac{L}{G}$. Consequently, if we choose $\xi \geq 2L/\epsilon^2$ as assumed, then we have:

$$\langle \phi(x), \xi \nabla f(x) \rangle - \text{div}\,\phi(x) \geq \frac{\xi \epsilon^2}{G} - \frac{L}{G} \geq \frac{\xi \epsilon^2}{2G}.$$

Lemma 3 then establishes the claimed lower bound. ∎

Next, we consider approximate local minima (Nesterov and Polyak, 2006; Agarwal et al., 2016), which rules out local maxima and strict saddle points. For an arbitrary $\epsilon > 0$, the set of ϵ-approximate local minima is defined by:

$$U := \{x \in K : \|\nabla f(x)\|_2 < \epsilon \text{ and } \nabla^2 f(x) \succeq -\sqrt{\epsilon}I\}. \tag{11}$$

We note that an approximate local minimum is not necessarily close to any local minimum of f. However, if we assume in addition the the function satisfies the (robust) strict-saddle property (Ge et al., 2015; Lee et al., 2016), then any point $x \in U$ is guaranteed to be close to a local minimum. Based on definition (11), we prove a lower bound for the set of approximate local minima.

Proposition 5 *Assume that Assumption B holds. For any $\epsilon > 0$, let U be the set of ϵ-approximate local minima. For any ξ satisfying*

$$\xi \geq \widetilde{\mathcal{O}}(1) \cdot \frac{\max\{1, G^{5/2}L^5, H^{5/2}\}}{\epsilon^2 G^{1/2}}, \tag{12}$$

we have $\mathcal{C}_{(\xi f)}(K \backslash U) \geq \frac{\sqrt{\epsilon}}{8(2G+1)G}$. The notation $\widetilde{\mathcal{O}}(1)$ hides a poly-logarithmic function of $(L, 1/\epsilon)$.

Proof sketch Proving Proposition 5 is significantly more challenging than proving Proposition 4. From a high-level point of view, we still construct a vector field ϕ, then lower bound the expression $\langle \phi(x), \xi \nabla f(x) \rangle - \text{div}\,\phi(x)$ for every point $x \in K \backslash U$ in order to apply Lemma 3. However, there exist saddle points in the set $K \backslash U$, such that the inner product $\langle \phi(x), \xi \nabla f(x) \rangle$ can be very close to zero. For these points, we need to carefully design the vector field so that the term $\text{div}\,\phi(x)$ is strictly negative and bounded away from zero. To this end, we define $\phi(x)$ to be the sum of two components. The first component aligns with the gradient $\nabla f(x)$. The second component aligns with a projected vector $\Pi_x(\nabla f(x))$, which projects $\nabla f(x)$ to the linear subspace spanned by the eigenvectors of $\nabla^2 f(x)$ with negative eigenvalues. It can be shown that the second component produces a strictly negative divergence in the neighborhood of strict saddle points. See Appendix E for the complete proof. ∎

2.5. Polynomial-time bound for finding an approximate local minimum

Combining Proposition 5 with Theorem 1, we conclude that SGLD finds an approximate local minimum of the function f in polynomial time, assuming that f is smooth enough to satisfy Assumption B.

Corollary 6 *Assume that Assumptions A,B hold. For an arbitrary $\epsilon > 0$, let U be the set of ϵ-approximate local minima. For any $\rho, \delta > 0$, there exists a large enough ξ and hyperparameters (η, k_{max}, D) such that with probability at least $1 - \delta$, SGLD returns a solution $\widehat{x}$ satisfying*

$$f(\widehat{x}) \leq \sup_{x:\, d(x,U) \leq \rho} f(x).$$

The iteration number k_{max} is bounded by a polynomial function of all hyperparameters in the assumptions as well as $(\epsilon^{-1}, \rho^{-1}, \log(1/\delta))$.

Similarly, we can combine Proposition 2 or Proposition 4 with Theorem 1, to obtain complexity bounds for finding the global minimum of a convex function, or finding an approximate stationary point of a smooth function.

Corollary 6 doesn't specify any upper limit on the temperature parameter ξ. As a result, SGLD can be stuck at the worst approximate local minima. It is important to note that the algorithm's capability of escaping certain local minima relies on a more delicate choice of ξ. Given objective function f, we consider an arbitrary smooth function F such that $\|f - F\|_\infty \leq 1/\xi$. By Theorem 1, for any target subset U, the hitting time of SGLD can be controlled by lower bounding the restricted Cheeger constant $\mathcal{C}_{\xi f}(K \backslash U)$. By the stability property (6), it is equivalent to lower bounding $\mathcal{C}_{\xi F}(K \backslash U)$ because f and F are uniformly close. If $\xi > 0$ is chosen large enough (w.r.t. smoothness parameters of F), then the lower bound established by Proposition 5 guarantees a polynomial hitting time to the set U_F of approximate local minima of F. Thus, SGLD can efficiently escape all local minimum of f that lie outside of U_F. Since the function F is arbitrary, it can be thought as a favorable perturbation of f such that the set U_F eliminates as many local minima of f as possible. The power of such perturbations are determined by their maximum scale, namely the quantity $1/\xi$. Therefore, it motivates choosing the smallest possible ξ whenever it satisfies the lower bound in Proposition 5.

The above analysis doesn't specify any concrete form of the function F. In Section 3, we present a concrete analysis where the function F is assumed to be the population risk of empirical risk minimization (ERM). We establish sufficient conditions under which SGLD efficiently finds an approximate local minima of the population risk.

3. Applications to empirical risk minimization

In this section, we apply SGLD to a specific family of functions, taking the form:

$$f(x) := \frac{1}{n} \sum_{i=1}^{n} \ell(x; a_i) \quad \text{for} \quad x \in K.$$

These functions are generally referred as the *empirical risk* in the statistical learning literature. Here, every instance $a_i \in \mathcal{A}$ is i.i.d. sampled from a distribution $\mathcal{P}$, and the function $\ell : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ measures the loss on individual samples. We define *population risk* to be the function $F(x) := \mathbb{E}_{x \sim \mathcal{P}}[\ell(x, a)]$.

We shall prove that under certain conditions, SGLD finds an approximate local minimum of the (presumably smooth) population risk in polynomial time, even if it is executed on a non-smooth empirical risk. More concretely, we run SGLD on a smoothed approximation of the empirical risk that satisfies Assumption A. With large enough sample size, the empirical risk f and its smoothed approximation will be close enough to the population risk F, so that combining the stability property with Theorem 1 and Proposition 5 establishes the hitting time bound. First, let's formalize the assumptions.

Assumption C (parameter space, loss function and population risk)

- *The parameter space K satisfies: there exists $h_{max} > 0$, such that for any $x \in K$ and any $h \leq h_{max}$, the random variable $y \sim N(x, 2hI)$ satisfies $P(y \in K) \geq \frac{1}{3}$.*

- *There exist $\rho_K, \nu > 0$ such that in the set $\overline{K} := \{x : d(x, K) \leq \rho_K\}$, the population risk F is G-Lipschitz continuous, and $\sup_{x \in \overline{K}} |f(x) - F(x)| \leq \nu$.*

- *For some $B > 0$, the loss $\ell(x; a)$ is uniformly bounded in $[0, B]$ for any $(x, a) \in \mathbb{R}^d \times \mathcal{A}$.*

The first assumption is identical to that of Assumption A. The second assumption requires the population risk to be Lipschitz continuous, and it bounds the ℓ_∞-norm distance between f and F. The third assumption requires the loss to be uniformly bounded. Note that Assumption C allows the empirical risk to be non-smooth or even discontinuous.

Since the function f can be non-differentiable, the stochastic gradient may not be well defined. We consider a smooth approximation of it following the idea of Duchi et al. (2015):

$$\tilde{f}_\sigma(x) := \mathbb{E}_z[f(x + z)] \quad \text{where} \quad z \sim N(0, \sigma^2 I_{d\times d}), \tag{13}$$

where $\sigma > 0$ is a smoothing parameter. We can easily compute a stochastic gradient g of $\tilde{f}_\sigma$ as follows:

$$\nabla \tilde{f}_\sigma(x) = \mathbb{E}[g(x) \mid x] \quad \text{where} \quad g(x) := \frac{z}{\sigma^2}(\ell(x + z; a) - \ell(x; a)), \tag{14}$$

Here, z is sampled from $N(0, \sigma^2 I_{d\times d})$ and a is uniformly sampled from $\{a_1, \ldots, a_n\}$. This stochastic gradient formulation is useful when the loss function ℓ is non-differentiable, or when its gradient norms are unbounded. The former happens for minimizing the zero-one loss, and the later can arise in training deep neural networks (Pascanu et al., 2013; Bengio et al., 2013). Since the loss function is uniformly bounded, formula (14) guarantees that the squared-norm $\|g(x)\|_2^2$ is sub-exponential.

We run SGLD on the function $\tilde{f}_\sigma$. Theorem 1 implies that the time complexity inversely depends on the restricted Cheeger constant $\mathcal{C}_{(\xi \tilde{f}_\sigma)}(K \backslash U)$. We can lower bound this quantity using $\mathcal{C}_{(\xi F)}(K \backslash U)$ — the restricted Cheeger constant of the population risk. Indeed, by choosing a small enough σ, it can be shown that $\sup_{x \in K} |\tilde{f}_\sigma(x) - F(x)| \leq 2\nu$. The stability property (6) then implies

$$\mathcal{C}_{(\xi \tilde{f}_\sigma)}(K \backslash U) \geq e^{-4\xi\nu} \mathcal{C}_{(\xi F)}(K \backslash U). \tag{15}$$

For any $\xi \in (0, 1/\nu]$, we have $e^{-4\xi\nu} \geq e^{-4}$, thus the term $\mathcal{C}_{(\xi \tilde{f}_\sigma)}(K \backslash U)$ is lower bounded by $e^{-4} \mathcal{C}_{(\xi F)}(K \backslash U)$. As a consequence, we obtain the following special case of Theorem 1.

Theorem 7 *Assume that Assumptions C holds. For any subset $U \subset K$, any $\delta > 0$ and any $\xi \in (0, 1/\nu]$, there exist hyperparameters $(\eta, \sigma, k_{max}, D)$ such that with probability at least $1 - \delta$,*

running SGLD on $\tilde{f}_\sigma$ returns a solution $\widehat{x}$ satisfying:

$$F(\widehat{x}) \leq \sup_{x \in U} F(x) + 5\nu. \tag{16}$$

The iteration number k_{max} is polynomial in $(B, \log(1/\delta), d, h_{max}^{-1}, \nu^{-1}, \rho_\kappa^{-1}, \mathcal{C}_{(\xi F)}^{-1}(K \backslash U))$.

See Appendix F for the proof.

In order to lower bound the restricted Cheeger constant $\mathcal{C}_{(\xi F)}(K \backslash U)$, we resort to the general lower bounds in Section 2.4. Consider population risks that satisfy the conditions of Assumption B. By combining Theorem 7 with Proposition 5, we conclude that SGLD finds an approximate local minimum of the population risk in polynomial time.

Corollary 8 *Assume that Assumption C holds. Also assume that Assumption B holds for the population risk F with smoothness parameters (G, L, H). For any $\epsilon > 0$, let U be the set of ϵ-approximate local minima of F. If*

$$\sup_{x \in K} |f(x) - F(x)| \leq \widetilde{\mathcal{O}}(1) \cdot \frac{\epsilon^2 G^{1/2}}{\max\{1, G^{5/2}L^5, H^{5/2}\}}, \tag{17}$$

then there exist hyperparameters $(\xi, \eta, \sigma, k_{max}, D)$ such that with probability at least $1 - \delta$, running SGLD on $\tilde{f}_\sigma$ returns a solution $\widehat{x}$ satisfying $F(\widehat{x}) \leq \sup_{x \in U} F(x) + 5\nu$. The time complexity will be bounded by a polynomial function of all hyperparameters in the assumptions as well as $(\epsilon^{-1}, \log(1/\delta))$. The notation $\widetilde{\mathcal{O}}(1)$ hides a poly-logarithmic function of $(L, 1/\epsilon)$.

Assumption B requires the population risk to be sufficiently smooth. Nonetheless, assuming smoothness of the population risk is relatively mild, because even if the loss function is discontinuous, the population risk can be smooth given that the data is drawn from a smooth density. The generalization bound (17) is a necessary condition, because the constraint $\xi \leq 1/\nu$ for Theorem 7 and the constraint (12) for Proposition 5 must simultaneously hold. With a large sample size n, the empirical risk can usually be made sufficiently close to the population risk. There are multiple ways to bound the ℓ_∞-distance between the empirical risk and the population risk, either by bounding the VC-dimension (Vapnik, 1999), or by bounding the metric entropy (Haussler, 1992) or the Rademacher complexity (Bartlett and Mendelson, 2003) of the function class. We note that for many problems, the function gap uniformly converges to zero in a rate $\mathcal{O}(n^{-c})$ for some constant $c > 0$. For such problems, the condition (17) can be satisfied with a polynomial sample complexity.

4. Learning linear classifiers with zero-one loss

As a concrete application, we study the problem of learning linear classifiers with zero-one loss. The learner observes i.i.d. training instances (a, b) where $(a, b) \in \mathbb{R}^d \times \{-1, 1\}$ are feature-label pairs. The goal is to learn a linear classifier $a \mapsto \langle x, a \rangle$ in order to minimize the zero-one loss:

$$F(x) := \mathbb{E}_{(a,b) \sim \mathcal{P}}[\ell(x; (a, b))] \quad \text{where} \quad \ell(x; (a, b)) := \begin{cases} 0 & \text{if } b \times \langle x, a \rangle > 0, \\ 1 & \text{if } b \times \langle x, a \rangle < 0, \\ 1/2 & \text{if } \langle x, a \rangle = 0, \end{cases}$$

For a finite dataset $\{(a_i, b_i)\}_{i=1}^n$, the empirical risk is $f(x) := \frac{1}{n} \sum_{i=1}^n \ell(x; (a_i, b_i))$. Clearly, the function f is non-convex and discontinous, and has zero gradients almost everywhere. Thus the optimization cannot be accomplished by gradient descent.

For a general data distribution, finding a global minimizer of the population risk is NP-hard (Arora et al., 1993). We follow Awasthi et al. (2015) to assume that the feature vectors are drawn uniformly from the unit sphere, and the observed labels b are corrupted by the Massart noise. More precisely, we assume that there is an unknown unit vector x^* such that for every feature $a \in \mathbb{R}^d$, the observed label b satisfies:

$$b = \begin{cases} \text{sign}(\langle x^*, a \rangle) & \text{with probability } \frac{1+q(a)}{2}; \\ -\text{sign}(\langle x^*, a \rangle) & \text{with probability } \frac{1-q(a)}{2}; \end{cases} \tag{18}$$

where $\frac{1-q(a)}{2} \in [0, 0.5]$ is the Massart noise level. We assume that the noise level is strictly smaller than 0.5 when the feature vector a is separated apart from the decision boundary. Formally, there is a constant $0 < q_0 \le 1$ such that

$$q(a) \ge q_0 \, |\langle x^*, a \rangle|. \tag{19}$$

The value of $q(a)$ can be adversarially perturbed as long as it satisfies the constraint (19). Awasthi et al. (2015) studied the same Massart noise model, but they impose a stronger constraint $q(a) \ge 1 - 3.6 \times 10^{-6}$ for all $a \in \mathbb{R}^d$, so that almost all observed labels are accurate. In contrast, our model (19) captures arbitrary Massart noises (because q_0 can be arbitrarily small), and allows for completely random observations at the decision boundary. Our model is thus more general than that of Awasthi et al. (2015).

Given function f, we use SGLD to optimize its smoothed approximation (13) in a compact parameter space $K := \{x \in \mathbb{R}^d : 1/2 \le \|x\|_2 \le 1\}$. The following theorem shows that the algorithm finds an approximate global optimum in polynomial time, with a polynomial sample complexity.

Theorem 9 *Assume that $d \ge 2$. For any $q_0 \in (0, 1]$ and $\epsilon, \delta > 0$, if the sample size n satisfies:*

$$n \ge \widetilde{O}(1) \cdot \frac{d^4}{q_0^2 \epsilon^4},$$

then there exist hyperparameters $(\xi, \eta, \sigma, k_{max}, D)$ such that SGLD on the smoothed function (13) returns a solution $\widehat{x}$ satisfying $F(\widehat{x}) \le F(x^*) + \epsilon$ with probability at least $1 - \delta$. The notation $\widetilde{O}(1)$ hides a poly-logarithmic function of $(d, 1/q_0, 1/\epsilon, 1/\delta)$. The time complexity of the algorithm is polynomial in $(d, 1/q_0, 1/\epsilon, \log(1/\delta))$.

Proof sketch The proof consists of two parts. For the first part, we prove that the population risk is Lipschitz continuous and the empirical risk uniformly converges to the population risk, so that Assumption C hold. For the second part, we lower bound the restricted Cheeger constant by Lemma 3. The proof is spiritually similar to that of Proposition 4 or Proposition 5. We define U to be the set of approximately optimal solutions, and construct a vector field ϕ such that:

$$\phi(x) \propto \langle x, x^* \rangle x - \|x\|_2^2 x^*.$$

By lower bounding the expression $\langle \phi(x), \nabla f(x) \rangle - \text{div}\, \phi(x)$ for all $x \in K \backslash U$, Lemma 3 establishes a lower bound on the restricted Cheeger constant. The theorem is established by combining the two parts together and by Theorem 7. We defer the full proof to Appendix G. ∎

5. Conclusion

In this paper, we analyzed the hitting time of the SGLD algorithm on non-convex functions. Our approach is different from existing analyses on Langevin dynamics (Bubeck et al., 2015; Dalalyan, 2016; Bonis, 2016; Teh et al., 2016; Raginsky et al., 2017), which connect LMC to a continuous-time Langevin diffusion process, then study the mixing time of the latter process. In contrast, we are able to establish polynomial-time guarantees for achieving certain optimality sets, regardless of the exponential mixing time.

For future work, we hope to establish stronger results on non-convex optimization using the techniques developed in this paper. Our current analysis doesn't apply to training over-specified models. For these models, the empirical risk can be minimized far below the population risk (Safran and Shamir, 2015), thus the assumption of Corollary 8 is violated. In practice, over-specification often makes the optimization easier, thus it could be interesting to show that this heuristic actually improves the restricted Cheeger constant. Another open problem is avoiding poor population local minima. Jin et al. (2016) proved that there are many poor population local minima in training Gaussian mixture models. It would be interesting to investigate whether a careful initialization could prevent SGLD from hitting such bad solutions.

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *arXiv preprint arXiv:1602.05908*, 2016.

Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 724–733. IEEE, 1993.

P. Awasthi, M.-F. Balcan, N. Haghtalab, and R. Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8624–8628. IEEE, 2013.

Avrim L Blum and Ronald L Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5(1):117–127, 1992.

Thomas Bonis. Guarantees in wasserstein distance for the langevin monte carlo algorithm. *arXiv preprint arXiv:1602.02616*, 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.

Sebastien Bubeck, Ronen Eldan, and Joseph Lehec. Finite-time analysis of projected langevin monte carlo. In *Advances in Neural Information Processing Systems*, pages 1243–1251, 2015.

Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. 1969.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: the power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. On local maxima in the population likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances In Neural Information Processing Systems*, pages 4116–4124, 2016.

Łukasz Kaiser and Ilya Sutskever. Neural gpus learn algorithms. *arXiv preprint arXiv:1511.08228*, 2015.

Karol Kurach, Marcin Andrychowicz, and Ilya Sutskever. Neural random-access machines. *arXiv preprint arXiv:1511.06392*, 2015.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

Arvind Neelakantan, Quoc V Le, and Ilya Sutskever. Neural programmer: Inducing latent programs with gradient descent. *arXiv preprint arXiv:1511.04834*, 2015a.

Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015b.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318, 2013.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. 2017.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. *arXiv preprint arXiv:1511.04210*, 2015.

Yee Whye Teh, Alexandre H. Thiery, and Sebastian J. Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 17(7):1–33, 2016.

Yung Liang Tong. *The multivariate normal distribution*. Springer Science & Business Media, 2012.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition. *arXiv preprint arXiv:1606.06871*, 2016.

Appendix A. Restricted Cheeger constant is strictly positive

In this appendix, we prove that under mild conditions, the restricted Cheeger constant for a convex parameter space is always strictly positive. Let K be an arbitrary convex parameter space with diameter $D < +\infty$. Lovász and Simonovits (1993, Theorem 2.6) proved the following isoperimetric inequality: for any subset $A \subset K$ and any $\epsilon > 0$, the following lower bound holds:

$$\frac{\text{vol}(A_\epsilon) - \text{vol}(A)}{\epsilon \min\{\text{vol}(A), \text{vol}(K \backslash A_\epsilon)\}} \geq \frac{2}{D}, \tag{20}$$

where $\text{vol}(A)$ represents the Borel measure of set A. Let $f_0(x) := 0$ be a constant zero function. By the definition of the function-induced probability measure, we have

$$\mu_{f_0}(A) = \frac{\text{vol}(A)}{\text{vol}(K)} \quad \text{for all} \quad A \subset K. \tag{21}$$

Combining the inequality (20) with equation (21), we obtain:

$$\frac{\mu_{f_0}(A_\epsilon) - \mu_{f_0}(A)}{\epsilon\,\mu_{f_0}(A)\,(1 - \mu_{f_0}(A_\epsilon))} \geq \frac{\mu_{f_0}(A_\epsilon) - \mu_{f_0}(A)}{\epsilon\,\min\{\mu_{f_0}(A), 1 - \mu_{f_0}(A_\epsilon)\}} \geq \frac{2}{D}.$$

If the set A satisfies $A \subset V \subset K$, then $1 - \mu_{f_0}(A_\epsilon) \geq 1 - \mu_{f_0}(V_\epsilon)$. Combining it with the above inequality, we obtain:

$$\frac{\mu_{f_0}(A_\epsilon) - \mu_{f_0}(A)}{\epsilon\,\mu_{f_0}(A)} \geq \frac{2(1 - \mu_{f_0}(V_\epsilon))}{D} = \frac{2(\mathrm{vol}(K) - \mathrm{vol}(V_\epsilon))}{D\,\mathrm{vol}(K)}.$$

According to the definition of the restricted Cheeger constant, the above lower bound implies:

$$\mathcal{C}_{f_0}(V) \geq \frac{2(\mathrm{vol}(K) - \lim_{\epsilon \to 0} \mathrm{vol}(V_\epsilon))}{D\,\mathrm{vol}(K)}. \tag{22}$$

Consider an arbitrary bounded function f satisfying $\sup_{x \in K} |f(x)| \leq B < +\infty$, combining the stability property (6) and inequality (22), we obtain:

$$\mathcal{C}_f(V) \geq e^{-2B} \times \frac{2(\mathrm{vol}(K) - \lim_{\epsilon \to 0} \mathrm{vol}(V_\epsilon))}{D\,\mathrm{vol}(K)}.$$

We summarize the result as the following proposition.

Proposition 10 *Assume that K is a convex parameter space with finite diameter. Also assume that $V \subset K$ is a measurable set satisfying $\lim_{\epsilon \to 0} vol(V_\epsilon) < vol(K)$. For any bounded function $f : K \to \mathbb{R}$, the restricted Cheeger constant $\mathcal{C}_f(V)$ is strictly positive.*

Appendix B. Proof of Theorem 1

The proof consists of two parts. We first establish a general bound on the hitting time of Markov chains to a certain subset $U \subset K$, based on the notion of *restricted conductance*. Then we prove that the hitting time of SGLD can be bounded by the hitting time of a carefully constructed time-reversible Markov chain. This Markov chain runs a Metropolis-Hastings algorithm that converges to the stationary distribution $\mu_{\xi f}$. We prove that this Markov chain has a bounded restricted conductance, whose value is characterized by the restricted Cheeger constant that we introduced in Section 2.2. Combining the two parts establishes the general theorem.

B.1. Hitting time of Markov chains

For an arbitrary Markov chain defined on the parameter space K, we represent the Markov chain by its *transition kernel* $\pi(x, A)$, which gives the conditional probability that the next state satisfies $x_{k+1} \in A$ given the current state $x_k = x$. Similarly, we use $\pi(x, x')$ to represent the conditional probability $P(x_{k+1} = x' | x_k = x)$. If π has a stationary distribution, then we denote it by Q_π.

A Markov chain is call *lazy* if $\pi(x, x) \geq 1/2$ for every $x \in K$, and is called *time-reversible* if it satisfies

$$\int_A \pi(x, B) Q_\pi(x) = \int_B \pi(x, A) Q_\pi(x) \quad \text{for any } A, B \subset K.$$

If $(x_0, x_1, x_2, \dots)$ is a realization of the Markov chain π, then the *hitting time* to some set $U \subset K$ is denoted by:

$$\tau_\pi(U) := \min\{k : x_k \in U\}.$$

For arbitrary subset $V \subset K$, we define the *restricted conductance*, denoted by $\Phi_\pi(V)$, to be the following infimum ratio:

$$\Phi_\pi(V) := \inf_{A \subset V} \frac{\int_A \pi(x, K \backslash A) Q_\pi(x) dx}{Q_\pi(A)}. \tag{23}$$

Based on the notion of restricted conductance, we present a general upper bound on the hitting time. For arbitrary subset $U \subset K$, suppose that $\widetilde{\pi}$ is an arbitrary Markov chain whose transition kernel is stationary inside U, namely it satisfies $\widetilde{\pi}(x, x) = 1$ for any $x \in U$. Let $(\widetilde{x}_0, \widetilde{x}_1, \widetilde{x}_2, \dots)$ be a realization of the Markov chain $\widetilde{\pi}$. We denote by Q_k the probability distribution of $\widetilde{x}_k$ at iteration k. In addition, we define a measure of closeness between any two Markov chains.

Definition *For two Markov chains π and $\widetilde{\pi}$, we say that $\widetilde{\pi}$ is ϵ-close to π w.r.t. a set U if the following condition holds for any $x \in K \backslash U$ and any $A \subset K \backslash \{x\}$:*

$$\pi(x, A) \leq \widetilde{\pi}(x, A) \leq (1 + \epsilon) \pi(x, A). \tag{24}$$

Then we are able to prove the following lemma.

Lemma 11 *Let π be a time-reversible lazy Markov chain with atom-free stationary distribution Q_π. Assume that $\widetilde{\pi}$ is ϵ-close to π w.r.t. U where $\epsilon \leq \frac{1}{4} \Phi_\pi(K \backslash U)$. If there is a constant M such that the distribution Q_0 satisfies $Q_0(A) \leq M Q_\pi(A)$ for any $A \subset K \backslash U$, then for any $\delta > 0$, the hitting time of the Markov chain is bounded by:*

$$\tau_{\widetilde{\pi}}(U) \leq \frac{4 \log(M/\delta)}{\Phi_\pi^2(K \backslash U)}, \tag{25}$$

with probability at least $1 - \delta$.

See Appendix B.3.1 for the proof of Lemma 11. The lemma shows that if the two chains π and $\widetilde{\pi}$ are sufficiently close, then the hitting time of the Markov chain $\widetilde{\pi}$ will be inversely proportional to the square of the restricted conductance of the Markov chain π, namely $\Phi_\pi(K \backslash U)$. Note that if the density function of distribution Q_π is bounded, then by choosing Q_0 to be the uniform distribution over K, there exists a finite constant M such that $Q_0(A) \leq M Q_\pi(A)$, satisfying the last condition of Lemma 11.

B.2. Proof of the theorem

The SGLD algorithm initializes x_0 by the uniform distribution μ_{f_0} (with $f_0(x) \equiv 0$). Then at iteration $k \geq 1$, it performs the following update:

$$y_k = x_{k-1} - \eta \cdot g(x_{k-1}) + \sqrt{2\eta/\xi} \cdot w; \quad x_k = \begin{cases} y_k & \text{if } y_k \in K \cap \mathcal{B}(x_{k-1}; 4\sqrt{2\eta d/\xi}), \\ x_{k-1} & \text{otherwise.} \end{cases} \tag{26}$$

We refer the particular setting $\xi = 1$ as the "standard setting". For the "non-standard" setting of $\xi \neq 1$, we rewrite the first equation as:

$$y_k = x_{k-1} - (\eta/\xi) \cdot (\xi g(x_{k-1})) + \sqrt{2(\eta/\xi)} \cdot w.$$

This re-formulation reduces to the problem to the standard setting, with stepsize η/ξ and objective function ξf. Thus it suffices to prove the theorem in the standard setting, then plug in the stepsize

η/ξ and the objective function ξf to obtain the general theorem. Therefore, we assume $\xi = 1$ and consider the sequence of points $(x_0, x_1, \dots)$ generated by:

$$y_k = x_{k-1} - \eta \cdot g(x_{k-1}) + \sqrt{2\eta} \cdot w; \quad x_k = \begin{cases} y_k & \text{if } y_k \in K \cap \mathcal{B}(x_{k-1}; 4\sqrt{2\eta d}), \\ x_{k-1} & \text{otherwise.} \end{cases} \quad (27)$$

We introduce two additional notations: for arbitrary functions f_1, f_2, we denote the maximal gap $\sup_{x \in K} |f_1(x) - f_2(x)|$ by the shorthand $\|f_1 - f_2\|_\infty$. For arbitrary set $V \subset K$ and $\rho > 0$, we denote the super-set $\{x \in K : d(x, V) \leq \rho\}$ by the shorthand V_ρ. Then we prove the following theorem for the standard setting.

Theorem 12 *Assume that Assumption A holds. Let x_0 be sampled from μ_{f_0} and let the Markov chain $(x_0, x_1, x_2, \cdots)$ be generated by update (27). Let $U \subset K$ be an arbitrary subset and let $\rho > 0$ be an arbitrary positive number. Let $\mathcal{C} := \mathcal{C}_f(K \backslash U)$ be a shorthand notation. Then for any $\delta > 0$ and any stepsize η satisfying*

$$\eta \leq c \min\left\{ d\rho^2, h_{max}, \frac{b_{max}^2}{d}, \frac{\mathcal{C}^2}{d^3(G^2 + L)^2} \right\}, \quad (28)$$

the hitting time to set U_ρ is bounded by

$$\min\{k : x_k \in U_\rho\} \leq \frac{c' d \left(\|f - f_0\|_\infty + \log(1/\delta) \right)}{\eta \mathcal{C}^2}, \quad (29)$$

with probability at least $1 - \delta$. Here, $c, c' > 0$ are universal constants.

Theorem 12 shows that if we choose $\eta \in (0, \eta_0]$, where η_0 is the right-hand side of inequality (28), then with probability at least $1 - \delta$, the hitting time to the set U_ρ is bounded by

$$\min\{k : x_k \in U_\rho\} \leq \frac{c' d \left(B + \log(1/\delta) \right)}{\eta \mathcal{C}^2} = \frac{c' d \left(B + \log(1/\delta) \right)}{(\eta/\eta_0) \eta_0 \mathcal{C}^2}.$$

Combining it with the definition of η_0, we conclude that $\min\{k : x_k \in U_\rho\} \leq \frac{M}{\mathcal{C}^4}$ where M is polynomial in $(B, L, G, \log(1/\delta), d, \eta_0/\eta, h_{max}^{-1}, b_{max}^{-1}, \rho^{-1})$. This establishes the iteration complexity bound. Whenever x_k hits U_ρ, we have

$$f(\widehat{x}) \leq f(x_k) \leq \sup_{x : d(x,U) \leq \rho} f(x),$$

which establishes the risk bound. Thus, Theorem 12 establishes Theorem 1 for the special case of $\xi = 1$.

In the non-standard setting ($\xi \neq 1$), we follow the reduction described above to substitute (η, f) in Theorem 12 with the pair $(\eta/\xi, \xi f)$. As a consequence, the quantity $\mathcal{C}$ is substituted with $\mathcal{C}_{(\xi f)}(K \backslash U)$, and $(B, L, G, \eta_0, b_{max})$ are substituted with $(\xi B, \xi L, \xi G, \eta_0/\xi, b_{max}/\xi)$. Both the iteration complexity bound and the risk bound hold as in the standard setting, except that after the substitution, the numerator M in the iteration complexity bound has an additional polynomial dependence on ξ. Thus we have proved the general conclusion of Theorem 1.

Proof of Theorem 12 For the function $f : K \to \mathbb{R}^d$ satisfying Assumption A, we define a time-reversible Markov chain represented by the following transition kernel π_f. Given any current state $x_k = x \in K$, the Markov chain draws a "candidate state" $y \in \mathbb{R}^d$ from the following the density

function:

$$q_x(y) := \frac{1}{2}\delta_x(y) + \frac{1}{2} \cdot \frac{1}{(4\pi\eta)^{d/2}} \mathbb{E}\left[e^{-\frac{\|y-x+\eta \cdot g(x)\|_2^2}{4\eta}} \mid x \right] \tag{30}$$

where δ_x is the Dirac delta function at point x. The expectation is taken over the stochastic gradient g defined in equation (27), conditioning on the current state x. Then for any candidate state $y \in K \cap \mathcal{B}(x; 4\sqrt{2\eta d})$, we accept the candidate state (i.e., $x_{k+1} = y$) with probability:

$$\alpha_x(y) := \min\left\{ 1, \frac{q_y(x)}{q_x(y)} e^{f(x)-f(y)} \right\}, \tag{31}$$

or reject the candidate state (i.e., $x_{k+1} = x$) with probability $1 - \alpha_x(y)$. All candidate states $y \notin K \cap \mathcal{B}(x; 4\sqrt{2\eta d})$ are rejected (i.e., $x_{k+1} = x$). It is easy to verify that π_f executes a Metropolis-Hastings algorithm. Therefore, it induces a time-reversible Markov chain, and its stationary distribution is equal to $\mu_f(x) \propto e^{-f(x)}$.

Given the subset $U_\rho \subset K$, we define an auxiliary Markov chain and its transition kernel $\widetilde{\pi}_f$ as follow. Given any current state $\widetilde{x}_k = x \in K$, the Markov chain proposes a candidate state $y \in \mathbb{R}^d$ through the density function $q_x(y)$ defined by equation (30), then accepts the candidate state if and only if $x \notin U_\rho$ and $y \in K \cap \mathcal{B}(x; 4\sqrt{2\eta d})$. Upon acceptance, the next state of $\widetilde{\pi}_f$ is defined to be $\widetilde{x}_{k+1} = y$, otherwise $\widetilde{x}_{x+1} = x$. The Markov chains $\widetilde{\pi}_f$ differs from π_f only in their different probabilities for acccepting the candidate state y. If $x \in U_\rho$, then π_f may accept y with probability $\alpha_x(y)$, but $\widetilde{\pi}_f$ always rejects y. If $x \notin U_\rho$ and $y \in K \cap \mathcal{B}(x; 4\sqrt{2\eta d})$, then π_f accepts y with probability $\alpha_x(y)$, while $\widetilde{\pi}_f$ accepts with probability 1.

Despite the difference in their definitions, we are able to show that the two Markov chains are ϵ-close, where ϵ depends on the stepsize η and the properties of the objective function.

Lemma 13 *Assume that $0 < \eta \le \frac{b_{max}^2}{8d}$ and Assumption A hold. Then the Markov chain $\widetilde{\pi}_f$ is ϵ-close to π_f w.r.t. U_ρ with $\epsilon = e^{16\eta d(G^2+L)} - 1$.*

See Appendix B.3.2 for the proof.

Lemma 13 shows that if we choose η small enough, then ϵ will be sufficiently small. Recall from Lemma 11 that we need $\epsilon \le \frac{1}{4}\Phi_{\pi_f}(K \backslash U_\rho)$ to bound the Markov chain π_f's hitting time to the set U_ρ. It means that η has to be chosen based on the restricted conductance of the Markov chain π_f. Although calculating the restricted conductance of a Markov chain might be difficult, the following lemma shows that the restricted conductance can be lower bounded by the restricted Cheeger constant.

Lemma 14 *Assume that $\eta \le \min\{h_{max}, 16d\rho^2, \frac{b_{max}^2}{8d}, \frac{1}{100d(G^2+L)}\}$ and Assumption A hold. Then for any $V \subset K$, we have:*

$$\Phi_{\pi_f}(V) \ge \frac{1}{192}(1 - e^{-\frac{1}{4}\sqrt{\eta/d}\mathcal{C}_f(V_\rho)}).$$

See Appendix B.3.3 for the proof.

By Lemma 13 and Lemma 14, we are able to choose a sufficiently small η such that the Markov chains π_f and $\widetilde{\pi}_f$ are close enough to satisfy the conditions of Lemma 11. Formally, the following condition on η is sufficient.

Lemma 15 *There exists a universal constant $c > 0$ such that for any stepsize η satisfying:*

$$\eta \leq c \min \left\{ h_{max}, d\rho^2, \frac{b_{max}^2}{d}, \frac{\mathcal{C}^2}{d^3(G^2+L)^2} \right\}, \tag{32}$$

the Markov chains π_f and $\widetilde{\pi}_f$ are ϵ-close with $\epsilon \leq \frac{1}{4}\Phi_{\pi_f}(K\backslash U_\rho)$. In addition, the restricted conductance satisfies the lower bound $\Phi_{\pi_f}(K\backslash U_\rho) \geq \frac{1}{1536}\sqrt{\eta/d}\,\mathcal{C}$.

See Appendix B.3.4 for the proof.

Under condition (32), the Markov chains π_f and $\widetilde{\pi}_f$ are ϵ-close with $\epsilon \leq \frac{1}{4}\Phi_{\pi_f}(K\backslash U_\rho)$. Recall that the Markov chain π_f is time-reversible and lazy. Since f is bounded, the stationary distribution $Q_{\pi_f} = \mu_f$ is atom-free, and sampling x_0 from $Q_0 := \mu_{f_0}$ implies:

$$Q_0(A) = \frac{\int_A e^{-f_0(x)}dx}{\int_K e^{-f_0(x)}dx} \leq \frac{e^{\sup_{x\in K} f(x)-f_0(x)}\int_A e^{-f(x)}dx}{e^{\inf_{x\in K} f(x)-f_0(x)}\int_K e^{-f(x)}dx} \leq e^{2\|f-f_0\|_\infty}Q_{\pi_f}(A). \tag{33}$$

Thus the last condition of Lemma 11 is satisfied. Combining Lemma 11 with the lower bound $\Phi_{\pi_f}(K\backslash U_\rho) \geq \frac{1}{1536}\sqrt{\eta/d}\,\mathcal{C}$ in Lemma 15, it implies that with probability at least $1 - \delta > 0$, we have

$$\tau_{\widetilde{\pi}_f}(U) \leq \frac{c'\,d\,(\|f-f_0\|_\infty + \log(1/\delta))}{\eta\,\mathcal{C}^2}, \tag{34}$$

where $c' > 0$ is a universal constant.

Finally, we upper bound the hitting time of SGLD (i.e., the Markov chain induced by formula (27)) using the hitting time upper bound (34). We denote by π_{sgld} the transition kernel of SGLD, and claim that the Markov chain induced by it can be generated as a sub-sequence of the Markov chain induced by $\widetilde{\pi}_f$. To see why the claim holds, we consider a Markov chain $(\widetilde{x}_0, \widetilde{x}_1, \widetilde{x}_2, \dots)$ generated by $\widetilde{\pi}_f$, and construct a sub-sequence $(x'_0, x'_1, x'_2, \dots)$ of this Markov chain as follows:

1. Assign $x'_0 = \widetilde{x}_0$ and initialize an index variable $\ell \leftarrow 0$.

2. Examine the states $\widetilde{x}_k$ in the order $k = 1, 2, \dots, \tau$, where $\tau = \min\{k : \widetilde{x}_k \in U\}$:

   - For any state $\widetilde{x}_k$, in order to sample its next state $\widetilde{x}_{k+1}$, the candidate state y is either drawn from a delta distribution $\delta_{\widetilde{x}_k}$, or drawn from a normal distribution with stochastic mean vector $x - \eta g(x)$. The probability of these two cases are equal, according to equation (30).

   - If y is drawn from the normal distribution, then generate a state $x'_{\ell+1} = \widetilde{x}_{k+1}$ and add it to the sub-sequence $(x'_0, x'_1, x'_2, \dots)$. Update the index variable $\ell \leftarrow \ell + 1$.

By this construction, it is easy to verify that $(x'_0, x'_1, x'_2, \dots)$ is a Markov chain and its transition kernel exactly matches formula (27). Since the sub-sequence $(x'_0, x'_1, x'_2, \dots)$ hits U in at most τ steps, we have

$$\tau_{\pi_{\text{sgld}}}(U) \leq \tau = \tau_{\widetilde{\pi}_f}(U).$$

Combining this upper bound with (34) completes the proof of Theorem 12.

B.3. Proof of technical lemmas

B.3.1. PROOF OF LEMMA 11

Let $q := Q_\pi(K\backslash U)$ be a shorthand notation. Let $\mathcal{G}_p$ be the class of functions $g : K\backslash U \to [0, 1]$ such that $\int_{K\backslash U} g(x)Q_\pi(x)dx = p$. We define a sequence of functions $h_k : [0, q] \to \mathbb{R}$ $(k = 1, 2, \ldots)$ such that

$$h_k(p) := \sup_{g \in \mathcal{G}_p} \int_{K\backslash U} g(x)Q_k(x)dx. \tag{35}$$

By its definition, the function h_k is a concave function on $[0, q]$. In addition, (Lovász and Simonovits, 1993, Lemma 1.2) proved the following properties for the function h_k: if Q_π is atom-free, then for any $p \in [0, q]$ there exists a function $g(x) := \mathbb{I}(x \in A)$ that attains the supremum in the definition of h_k. We claim the following property of the function h_k.

Claim 1 *If there is a constant C such that the inequality $h_0(p) \leq C\sqrt{p}$ holds for any $p \in [0, q]$, then the inequality*

$$h_k(p) \leq C\sqrt{p}(1 - \frac{1}{4}\Phi_\pi^2(K\backslash U))^k \tag{36}$$

holds for any $k \in \mathbb{N}$ and any $p \in [0, q]$.

According to the claim, it suffices to upper bound $h_0(p)$ for $p \in [0, q]$. Indeed, since $Q_0(A) \leq M\,Q_\pi(A)$ for any $A \subset K\backslash U$, we immediately have:

$$h_0(p) = \sup_{A \subset K\backslash U:\, Q_\pi(A)=p} Q_0(A) \leq Mp \leq M\sqrt{p}.$$

Thus, we have

$$Q_k(K\backslash U) \leq h_k(q) \leq M(1 - \frac{1}{4}\Phi_\pi^2(K\backslash U))^k.$$

Choosing $k := \frac{4\log(M/\delta)}{\Phi_\pi^2(K\backslash U)}$ implies $Q_k(K\backslash U) \leq \delta$. As a consequence, the hitting time is bounded by k with probability at least $1 - \delta$.

Proof of Claim 1 Recall the properties of the function h_k. For any $p \in [0, q]$, we can find a set $A \subset K\backslash U$ such that $Q_\pi(A) = p$ and $h_k(p) = Q_k(A)$. Define, for $x \in K$, two functions:

$$g_1(x) = \begin{cases} 2\widetilde{\pi}(x, A) - 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases},$$

$$g_2(x) = \begin{cases} 1 & \text{if } x \in A \\ 2\widetilde{\pi}(x, A) & \text{if } x \notin A \end{cases},$$

By the laziness of the Markov chain $\widetilde{\pi}$, we obtain $0 \leq g_i \leq 1$, so that they are functions mapping from $K\backslash U$ to $[0, 1]$. Using the relation $2\widetilde{\pi}(x, A) - 1 = 1 - 2\widetilde{\pi}(x, K\backslash A)$, the definition of g_1 implies that:

$$\int_{K\backslash U} g_1(x)Q_\pi(x)dx = Q_\pi(A) - 2\int_A \widetilde{\pi}(x, K\backslash A)Q_\pi(x)dx = p - 2\int_A \widetilde{\pi}(x, K\backslash A)Q_\pi(x)dx$$

$$\leq p - 2\int_A \pi(x, K\backslash A)Q_\pi(x)dx. \tag{37}$$

where the last inequality follows since the δ-closeness ensures $\pi(x, K\backslash A) \leq \widetilde{\pi}(x, K\backslash A)$. Similarly, using the definition of g_2 and the relation $\widetilde{\pi}(x, A) \leq (1 + \epsilon)\pi(x, A)$, we obtain:

$$\int_{K\backslash U} g_2(x)Q_\pi(x)dx = Q_\pi(A) + 2\int_{K\backslash(U\cup A)} \widetilde{\pi}(x, A)Q_\pi(x)dx$$

$$\leq p + 2\int_{K\backslash A} \widetilde{\pi}(x, A)Q_\pi(x)dx$$

$$\leq p + 2\int_{K\backslash A} (1 + \epsilon)\pi(x, A)Q_\pi(x)dx \qquad (38)$$

Since Q_π is the stationary distribution of the time-reversible Markov chain π, the right-hand side of (38) is equal to:

$$p + 2\int_{K\backslash A} (1 + \epsilon)\pi(x, A)Q_\pi(x)dx = p + 2(1 + \epsilon)\int_A \pi(x, K\backslash A)Q_\pi(x)dx \qquad (39)$$

Let p_1 and p_2 be the left-hand side of inequality (37) and (38) respectively, and define a short-hand notation:

$$r := \frac{1}{p}\int_A \pi(x, K\backslash A)Q_\pi(x)dx.$$

Then by definition of restricted conductance and the laziness of π, we have $\Phi_\pi(K\backslash U) \leq r \leq 1/2$. Combining inequalities (37), (38) and (39) and by simple algebra, we obtain:

$$\sqrt{p_1} + \sqrt{p_2} \leq \sqrt{p}\Big(\sqrt{1 - 2r} + \sqrt{1 + 2\epsilon r + 2r}\Big).$$

By the condition $\epsilon \leq \frac{1}{4}\Phi_\pi(K\backslash U) \leq \frac{r}{4}$, the above inequality implies

$$\sqrt{p_1} + \sqrt{p_2} \leq \sqrt{p}\Big(\sqrt{1 - 2r} + \sqrt{1 + 2r + r^2/2}\Big)$$

It is straightforward to verify that for any $0 \leq r \leq 1$, the right-hand side is upper bounded by $2(1 - r^2/4)\sqrt{p}$. Thus we obtain:

$$\sqrt{p_1} + \sqrt{p_2} \leq 2(1 - r^2/4)\sqrt{p}. \qquad (40)$$

On the other hand, the definition of g_1 and g_2 implies that $\widetilde{\pi}(x, A) = \frac{g_1(x)+g_2(x)}{2}$ for any $x \in K\backslash U$. For all $x \in U$, the transition kernel $\widetilde{\pi}$ is stationary, so that we have $\widetilde{\pi}(x, A) = 0$. Combining these two facts implies

$$h_k(p) = Q_k(A) = \int_{K\backslash U} \widetilde{\pi}(x, A)Q_{k-1}(x)dx$$

$$= \frac{1}{2}\Big(\int_{K\backslash U} g_1(x)Q_{k-1}(x)dx + \int_{K\backslash U} g_2(x)Q_{k-1}(x)dx\Big) \leq \frac{1}{2}(h_{k-1}(p_1) + h_{k-1}(p_2)). \qquad (41)$$

The last inequality uses the definition of function h_{k-1}.

Finally, we prove inequality (36) by induction. The inequality holds for $k = 0$ by the assumption. We assume by induction that it holds for an aribtrary integer $k - 1$, and prove that it holds for k. Combining the inductive hypothesis with inequalities (40) and (41), we have

$$h_k(p) \leq \frac{C}{2}(\sqrt{p_1} + \sqrt{p_2})(1 - \frac{1}{4}\Phi_\pi^2(K\backslash U))^{k-1} \leq C\sqrt{p}(1 - r^2/4)(1 - \frac{1}{4}\Phi_\pi^2(K\backslash U))^{k-1}$$

$$\leq C\sqrt{p}(1 - \frac{1}{4}\Phi_\pi^2(K\backslash U))^k,$$

Thus, inequality (36) holds for k, which completes the proof.

B.3.2. PROOF OF LEMMA 13

By the definition of the ϵ-closeness, it suffices to consider an arbitrary $x \notin U_\rho$ and verify the inequality (24). We focus on cases when the acceptance ratio of π_f and $\widetilde{\pi}_f$ are different, that is, when the candidate state y satisfies $y \neq x$ and $y \in K \cap \mathcal{B}(x; 4\sqrt{2\eta d})$. We make the following claim on the acceptance ratio.

Claim 2 *For any $0 < \eta \leq \frac{b_{max}^2}{8d}$, if we assume $x \notin U_\rho$, $y \notin x$, and $y \in K \cap \mathcal{B}(x; 4\sqrt{2\eta d})$, then the acceptance ratio is lower bounded by $\alpha_x(y) \geq e^{-16\eta d(G^2+L)}$.*

Consider an arbitrary point $x \in K \backslash U_\rho$ and an arbitrary subset $A \subset K \backslash \{x\}$. The definitions of π_f and $\widetilde{\pi}_f$ imply that $\pi_f(x, A) \leq \widetilde{\pi}_f(x, A)$ always hold. In order to prove the opposite, we notice that:

$$\widetilde{\pi}_f(x, A) = \int_{A \cap \mathcal{B}(x; 4\sqrt{2\eta d})} q_x(y) dy. \tag{42}$$

The definition of π_f and Claim 2 implies

$$\int_{A \cap \mathcal{B}(x; 4\sqrt{2\eta d})} q_x(y) dy \leq e^{16\eta d(G^2+L)} \int_{A \cap \mathcal{B}(x; 4\sqrt{2\eta d})} q_x(y) \alpha_x(y) dy$$
$$= e^{16\eta d(G^2+L)} \pi(x, A),$$

which completes the proof.

Proof of Claim 2 By plugging in the definition of $\alpha_x(y)$ and $\alpha_y(x)$ and the fact that $x \neq y$, we obtain

$$\frac{q_y(x)}{q_x(y)} e^{f(x)-f(y)} = \frac{\mathbb{E}\left[e^{-\frac{\|x-y+\eta \cdot g(y)\|_2^2}{4\eta}} \mid y\right]}{\mathbb{E}\left[e^{-\frac{\|y-x+\eta \cdot g(x)\|_2^2}{4\eta}} \mid x\right]} \cdot e^{f(x)-f(y)}. \tag{43}$$

In order to prove the claim, we need to lower bound the numerator and upper bound the denominator of equation (43). For the numerator, Jensen's inequality implies:

$$\mathbb{E}\left[e^{-\frac{\|x-y+\eta \cdot g(y)\|_2^2}{4\eta}} \mid y\right] \geq e^{-\mathbb{E}\left[\frac{\|x-y+\eta \cdot g(y)\|_2^2}{4\eta}\mid y\right]} = e^{-\frac{\|x-y\|_2^2}{4\eta} - \mathbb{E}\left[\frac{\langle x-y, g(y)\rangle}{2} + \frac{\eta\|g(y)\|_2^2}{4}\mid y\right]}$$
$$\geq e^{-\frac{\|x-y\|_2^2}{4\eta} - \frac{\langle x-y, \nabla f(y)\rangle}{2} - \frac{\eta dG^2}{4}} \tag{44}$$

where the last inequality uses the upper bound

$$\mathbb{E}[\|g(y)\|_2^2|y] = \frac{1}{b_{\max}^2} \sum_{i=1}^d \mathbb{E}[(b_{\max}g_i(y))^2|y] \leq \frac{1}{b_{\max}^2} \sum_{i=1}^d \log\left(\mathbb{E}[e^{(b_{\max}g_i(y))^2}|y]\right)$$
$$\leq \frac{1}{b_{\max}^2} \sum_{i=1}^d \log\left(e^{G^2 b_{\max}^2}\right) = dG^2. \tag{45}$$

For the above deduction, we have used the Jensen's inequality as well as Assumption A.

For the denominator, we notice that the term inside the expectation satisfies:

$$e^{-\frac{\|y-x+\eta \cdot g(x)\|_2^2}{4\eta}} \leq e^{-\frac{\|x-y\|_2^2}{4\eta} - \frac{\langle y-x, g(x)\rangle}{2}}$$

24

Using the relation that $e^a \leq a + e^{a^2}$ for any $a \in \mathbb{R}$, we have

$$e^{-\frac{\|y-x+\eta \cdot g(x)\|_2^2}{4\eta}} \leq e^{-\frac{\|x-y\|_2^2}{4\eta}} \left(-\frac{\langle y - x, g(x) \rangle}{2} + e^{(\langle \frac{x-y}{2}, g(x) \rangle)^2} \right). \tag{46}$$

Since $\|\frac{x-y}{2}\|_2 \leq 2\sqrt{2\eta d} \leq b_{\max}$, Assumption A implies

$$\mathbb{E}\big[e^{(\langle \frac{x-y}{2}, g(x) \rangle)^2}\big] \leq e^{G^2 \|\frac{x-y}{2}\|_2^2} \leq e^{8\eta d G^2}.$$

Thus, by taking expectation on both sides of inequality (46) (conditional on x) and using the relation that $a + e^b \leq e^{a+b}$ for any $a \in \mathbb{R}$, $b \geq 0$, we obtain

$$\mathbb{E}\Big[e^{-\frac{\|y-x+\eta \cdot g(x)\|_2^2}{4\eta}} \mid x\Big] \leq e^{-\frac{\|x-y\|_2^2}{4\eta}} \left(-\frac{\langle y - x, \nabla f(x) \rangle}{2} + e^{8\eta d G^2} \right)$$

$$\leq e^{-\frac{\|x-y\|_2^2}{4\eta} - \frac{\langle y-x, \nabla f(x) \rangle}{2} + 8\eta d G^2}. \tag{47}$$

Combining equation (43) with inequalities (44), (47), we obtain

$$\frac{q_y(x)}{q_x(y)} e^{f(x)-f(y)} \geq e^{f(x)-f(y)-\langle x-y, \frac{\nabla f(x)+\nabla f(y)}{2} \rangle - \frac{33\eta d G^2}{4}}. \tag{48}$$

The L-smoothness of function f implies that

$$f(x) - f(y) - \langle x - y, \frac{\nabla f(x) + \nabla f(y)}{2} \rangle \geq -\frac{L\|x-y\|_2^2}{2} \geq -16\eta d L.$$

Combining this inequality with the lower bound (48) completes the proof.

B.3.3. PROOF OF LEMMA 14

Recall that μ_f is the stationary distribution of the Markov chain π_f. We consider an arbitrary subset $A \subset V$, and define $B := K \backslash A$. Let A_1 and B_1 be defined as

$$A_1 := \{x \in A : \pi_f(x, B) < 1/96\} \quad \text{and} \quad B_1 := \{x \in B : \pi_f(x, A) < 1/96\},$$

In other words, the points in A_1 and B_1 have low probability to move across the broader between A and B. We claim that the distance between points in A_1 and B_1 must be bounded away from a positive number.

Claim 3 *Assume that $\eta \leq \min\{h_{max}, \frac{b_{max}^2}{8d}, \frac{1}{100d(G^2+L)}\}$. If $x \in A_1$ and $y \in B_1$, then $\|x - y\|_2 > \frac{1}{4}\sqrt{\eta/d}$.*

For any point $x \in K \backslash (A_1 \cup B_1)$, we either have $x \in A$ and $\pi_f(x, B) \geq 1/96$, or we have $x \in B$ and $\pi_f(x, A) \geq 1/96$. It implies:

$$\mu_f(K \backslash (A_1 \cup B_1)) \geq \int_A \mu_f(x)dx + \int_B \mu_f(x)dx$$

$$\geq \int_A 96\pi_f(x, B)\mu_f(x)dx + \int_B 96\pi_f(x, A)\mu_f(x)dx \tag{49}$$

Since μ_f is the stationary distribution of the time-reversible Markov chain π_f, inequality (49) implies:

$$\int_A \pi_f(x, B)\mu_f(x)dx = \frac{1}{2}\int_A \pi_f(x, B)\mu_f(x)dx + \frac{1}{2}\int_B \pi_f(x, A)\mu_f(x)dx$$
$$\geq \frac{1}{192}\mu_f(K\backslash(A_1 \cup B_1)) = \frac{1}{192}\Big(\mu_f(K\backslash B_1) - \mu_f(A_1)\Big). \qquad (50)$$

Notice that $A \subset K\backslash B_1$, so that $\mu_f(K\backslash B_1) \geq \mu_f(A)$. According to Claim 3, by defining an auxiliary quantity:

$$\rho_\eta := \frac{1}{4}\sqrt{\eta/d},$$

we find that the set $(A_1)_{\rho_\eta}$ belongs to $K\backslash B_1$, so that $\mu_f(K\backslash B_1) \geq \mu_f((A_1)_{\rho_\eta})$. The following property is a direct consequence of the definition of restricted Cheeger constant.

Claim 4 *For any $A \subset V$ and any $\nu > 0$, we have $\mu_f(A_\nu) \geq e^{\nu \cdot \mathcal{C}_f(V_\nu)}\mu_f(A)$.*

Letting $A := A_1$ and $\nu := \rho_\eta$ in Claim 4, we have $\mu_f((A_1)_{\rho_\eta}) \geq e^{\rho_\eta \cdot \mathcal{C}_f(V_{\rho_\eta})}\mu_f(A_1)$. Combining these inequalities, we obtain

$$\mu_f(K\backslash B_1) - \mu_f(A_1) \geq \max\Big\{\mu_f(A) - \mu_f(A_1), \Big(e^{\rho_\eta \cdot \mathcal{C}_f(V_{\rho_\eta})} - 1\Big)\mu_f(A_1)\Big\}$$
$$\geq \Big(1 - e^{-\rho_\eta \cdot \mathcal{C}_f(V_{\rho_\eta})}\Big)\mu_f(A),$$

where the last inequality uses the relation $\max\{a - b, (\alpha - 1)b\} \geq \frac{\alpha-1}{\alpha}(a-b) + \frac{1}{\alpha}(\alpha-1)b = \frac{\alpha-1}{\alpha}a$ with $\alpha := e^{\rho_\eta \cdot \mathcal{C}_f(V_{\rho_\eta})}$. Combining it with inequality (50), we obtain

$$\Phi_{\pi_f}(V) \geq \frac{1}{192}(1 - e^{-\rho_\eta \cdot \mathcal{C}_f(V_{\rho_\eta})}).$$

The lemma's assumption gives $\rho_\eta = \frac{1}{4}\sqrt{\eta/d} \leq \rho$. Plugging in this relation completes the proof.

Proof of Claim 3 Consider any two points $x \in A$ and $y \in B$. Let s be a number such that $2s\sqrt{2\eta d} = \|x - y\|_2$. If $s > 1$, then the claim already holds for the pair (x, y). Otherwise, we assume that $s \leq 1$, and as a consequence assume $\|x - y\|_2 \leq 2\sqrt{2\eta d}$.

We consider the set of points

$$Z := \Big\{z \in \mathbb{R}^d\backslash\{x, y\} : \|z - \frac{x+y}{2}\|_2 \leq 3\sqrt{2\eta d}\Big\}.$$

Denote by $q(z)$ the density function of distribution $N(\frac{x+y}{2}; 2\eta I)$. The integral $\int_Z q(z)dz$ is equal to $P(X \leq 9d)$, where X is a random variable satisfying the chi-square distribution with d degrees of freedom. The following tail bound for the chi-square distribution was proved by Laurent and Massart (2000).

Lemma 16 *If X is a random variable satisfying the Chi-square distribution with d degrees of freedom, then for any $x > 0$,*

$$P(X \geq d(1 + 2\sqrt{x} + 2x)) \leq e^{-xd} \quad and \quad P(X \leq d(1 - 2\sqrt{x})) \leq e^{-xd}.$$

By choosing $x = 9/5$ in Lemma 16, the probability $P(X \leq 9d)$ is lower bounded by $1 - e^{-(9/5)d} > 5/6$. Since $\eta \leq h_{\max}$, the first assumption of Assumption A implies $\int_K q(z)dz \geq 1/3$. Combining

these two bounds, we obtain

$$\int_{K \cap Z} q(z)dz \geq \int_{K} q(z)dz - \int_{Z^c} q(z)dz > 1/6. \tag{51}$$

For any point $z \in Z$, the distances $\|z - x\|_2$ and $\|z - y\|_2$ are bounded by $4\sqrt{2\eta d}$. It implies

$$Z \subset \mathcal{B}(x; 4\sqrt{2\eta d}) \cap \mathcal{B}(y; 4\sqrt{2\eta d}). \tag{52}$$

Claim 2 in the proof of Lemma 13 demonstrates that the acceptance ratio $\alpha_x(z)$ and $\alpha_y(z)$ for any $z \in K \cap Z$ are both lower bounded by $e^{-16\eta d(G^2 + L)}$ given the assumption $0 < \eta \leq \frac{b_{\max}^2}{8d}$. This lower bound is at least equal to $1/2$ because of the assumption $\eta \leq \frac{1}{100d(G^2 + L)}$, so that we have

$$\alpha_x(z) \geq \frac{1}{2} \quad \text{and} \quad \alpha_y(z) \geq \frac{1}{2} \quad \text{for all} \quad z \in K \cap Z. \tag{53}$$

Next, we lower bound the ratio $q_x(z)/q(z)$ and $q_y(z)/q(z)$. For $z \in Z$ but $z \neq x$, the function $q_x(z)$ is defined by

$$q_x(z) = \frac{1}{2} \cdot \frac{1}{(4\pi\eta)^{d/2}} \mathbb{E}\left[e^{-\frac{\|z - x + \eta \cdot g(x)\|_2^2}{4\eta}} \mid x\right],$$

so that we have

$$\frac{q_x(z)}{q(z)} = \frac{1}{2}\mathbb{E}\left[\exp\left(-\frac{\|z - x + \eta \cdot g(x)\|_2^2 - \|z - \frac{x+y}{2}\|_2^2}{4\eta}\right) \Big| x\right]$$

$$= \frac{1}{2}\mathbb{E}\left[\exp\left(-\frac{\langle \frac{y-x}{2} + \eta \cdot g(x), 2(z - \frac{x+y}{2}) - \frac{x-y}{2} + \eta \cdot g(x)\rangle}{4\eta}\right) \Big| x\right].$$

$$= \left(\frac{1}{2}e^{-\frac{1}{4\eta}\langle \frac{y-x}{2}, 2(z - \frac{x+y}{2}) + \frac{y-x}{2}\rangle}\right)\mathbb{E}\left[e^{-\frac{1}{4}\langle y-x+2(z-\frac{x+y}{2}), g(x)\rangle - \frac{\eta}{4}\|g(x)\|_2^2} | x\right]$$

$$\geq \left(\frac{1}{2}e^{-\frac{s(6+s)d}{2}}\right)e^{-\mathbb{E}\left[\frac{1}{4}\langle y-x+2(z-\frac{x+y}{2}), g(x)\rangle + \frac{\eta}{4}\|g(x)\|_2^2 | x\right]}, \tag{54}$$

where the last inequality uses Jensen's inequality; It also uses the fact $\|\frac{y-x}{2}\|_2 = s\sqrt{2\eta d}$ and $\|z - \frac{x+y}{2}\|_2 \leq 3\sqrt{2\eta d}$.

For any unit vector $u \in \mathbb{R}^d$, Jensen's inequality and Assumption A imply:

$$\mathbb{E}[(\langle u, g(x)\rangle)^2 | x] \leq \frac{1}{b_{\max}^2} \log\left(\mathbb{E}[e^{(\langle b_{\max}u, g(x)\rangle)^2} | x]\right) \leq \frac{1}{b_{\max}^2} \log(e^{b_{\max}^2 G^2}) = G^2.$$

As a consequence of this upper bound and using Jensen's inequality, we have:

$$\mathbb{E}[\langle y - x + 2(z - \frac{x+y}{2}), g(x)\rangle | x] \leq \|y - x + 2(z - \frac{x+y}{2})\|_2 G \leq (2s + 6)\sqrt{2\eta d}\, G. \tag{55}$$

Combining inequalities (45), (54) and (55), we obtain:

$$\frac{q_x(z)}{q(z)} \geq \frac{1}{2}e^{-\frac{s(6+s)d}{2} - \frac{(3+s)\sqrt{2\eta d}\, G}{2} - \frac{\eta d G^2}{4}}. \tag{56}$$

The assumption $\eta \leq \frac{1}{100d(G^2+L)}$ implies $G \leq \frac{1}{10\sqrt{\eta d}}$. Plugging in this inequality to (56), a sufficient condition for $q_x(z)/q(z) > 1/4$ is

$$s \leq \frac{1}{10d}. \tag{57}$$

Following identical steps, we can prove that inequality (57) is a sufficient condition for $q_y(z)/q(z) > 1/4$ as well.

Assume that condition (57) holds. Combining inequalities (51), (53) with the fact $q_x(z) > q(z)/4$ and $q_y(z) > q(z)/4$, we obtain:

$$\int_{K \cap Z} \min\{q_x(z)\alpha_x(z), q_y(z)\alpha_y(z)\}dz \geq \frac{1}{8} \int_{K \cap Z} q(z)dz \geq \frac{1}{48}. \tag{58}$$

Notice that the set Z satisfies $Z \subset \mathcal{B}(x; 4\sqrt{2\eta d}) \cap \mathcal{B}(y; 4\sqrt{2\eta d})$, thus the following lower bound holds:

$$\begin{aligned}
\pi_f(x, B) + \pi_f(y, A) &= \int_{B \cap \mathcal{B}(x; 4\sqrt{2\eta d})} q_x(z)\alpha_z(z)dz + \int_{A \cap \mathcal{B}(y; 4\sqrt{2\eta d})} q_y(z)\alpha_y(z)dz \\
&\geq \int_{K \cap Z} \min\{q_x(z)\alpha_x(z), q_y(z)\alpha_y(z)\}dz \geq \frac{1}{48}.
\end{aligned}$$

It implies that either $\pi_f(x, B) \geq \frac{1}{96}$ or $\pi_f(y, A) \geq \frac{1}{96}$. In other words, if $x \in A_1$ and $y \in B_1$, then inequality (57) *must not* hold. As a consequence, we obtain the lower bound:

$$\|x - y\|_2 = 2s\sqrt{2\eta d} \geq \frac{\sqrt{2}}{5}\sqrt{\eta/d} > \frac{1}{4}\sqrt{\eta/d}.$$

Proof of Claim 4 Let n be an arbitrary integer and let $i \in \{1, \ldots, n\}$. By the definition of the restricted Cheeger constant (see equation (5)), we have

$$\log(\mu_f(A_{i\nu/n})) - \log(\mu_f(A_{(i-1)\nu/n})) \geq (\nu/n)(\mathcal{C}_f(V_\nu) - \epsilon_n) \quad \text{for } i = 1, \ldots, n$$

where ϵ_n is an indexed variable satisfying $\lim_{n \to \infty} \epsilon_n = 0$. Suming over $i = 1, \ldots, n$, we obtain

$$\log(\mu_f(A_\nu) - \log(\mu_f(A)) \geq \nu \cdot (\mathcal{C}_f(V_\nu) - \epsilon_n).$$

Taking the limit $n \to \infty$ on both sides of the inequality completes the proof.

B.3.4. PROOF OF LEMMA 15

First, we impose the following constraints on the choice of η:

$$\eta \leq \min\left\{h_{\max}, 16d\rho^2, \frac{b_{\max}^2}{8d}, \frac{1}{100d(G^2 + L)}\right\}, \tag{59}$$

so that the preconditions of both Lemma 13 and Lemma 14 are satisfied. By plugging $V := K \backslash U_\rho$ to Lemma 14, the restricted conductance is lower bounded by:

$$\Phi_{\pi_f}(K \backslash U_\rho) \geq \frac{1}{192}(1 - e^{-\frac{1}{4}\sqrt{\eta/d} \cdot \mathcal{C}_f((K \backslash U_\rho)_\rho)}) \geq \frac{1}{1536}\sqrt{\eta/d}\,\mathcal{C}_f((K \backslash U_\rho)_\rho). \tag{60}$$

The last inequality holds because $1 - e^{-t} \geq t/2$ for any $t \in [0, 1]$, and on the other hand, we have $\sqrt{\eta/d}\,\mathcal{C}_f((K \backslash U_\rho)_\rho) \in [0, 1]$. It is easy to verify that $(K \backslash U_\rho)_\rho \subset K \backslash U$, so that we have the lower bound $\mathcal{C}_f((K \backslash U_\rho)_\rho) \geq \mathcal{C}_f(K \backslash U) = \mathcal{C}$. Plugging this lower bound to inequality (60), we obtain

$$\Phi_{\pi_f}(K \backslash U_\rho) \geq \frac{1}{1536}\sqrt{\eta/d}\,\mathcal{C}. \tag{61}$$

Inequality (61) establishes the restricted conductance lower bound for the lemma.

Combining inequality (61) with Lemma 13, it remains to choose a small enough η such that $\widetilde{\pi}_f$ is ϵ-close to π_f with $\epsilon \leq \frac{1}{4}\Phi_\pi(K \backslash U_\rho)$. More precisely, it suffices to make the following inequality hold:

$$e^{16\eta d(G^2 + L)} - 1 \leq \frac{1}{4} \cdot \frac{1}{1536}\sqrt{\eta/d}\,\mathcal{C}.$$

In order to satisfy this inequality, it suffices to choose $\eta \lesssim \frac{C^2}{d^3(G^2+L)^2}$. Combining this result with (59) completes the proof.

Appendix C. Proof of Proposition 2

Lovász and Simonovits (1993, Theorem 2.6) proved the following isoperimetric inequality: Let K be an arbitrary convex set with diameter 2. For any convex function f and any subset $V \subset K$ satisfying $\mu_f(V) \leq 1/2$, the following lower bound holds:

$$\frac{\mu_f(A_\epsilon) - \mu_f(A)}{\epsilon \, \mu_f(A)} \geq 1 \quad \text{for all} \quad A \subset V, \ \epsilon > 0. \tag{62}$$

The lower bound (62) implies $C_f(V) \geq 1$. In order to establish the proposition, it suffices to choose $V := K \backslash U$ and $f := \xi f$, then prove the pre-condition $\mu_{\xi f}(K \backslash U) \leq 1/2$.

Let x^* be one of the global minimum of function f and let $\mathcal{B}(x^*; r)$ be the ball of radius r centering at point x^*. If we choose $r = \frac{\epsilon}{2G}$, then for any point $x \in \mathcal{B}(x^*; r) \cap K$, we have

$$f(x) \leq f(x^*) + G\|x - x^*\|_2 \leq f(x^*) + \epsilon/2.$$

Moreover, for any $y \in K \backslash U$ we have:

$$f(y) \geq f(x^*) + \epsilon.$$

It means for the probability measure $\mu_{\xi f}$, the density function inside $\mathcal{B}(x^*; r) \cap K$ is at least $e^{\xi\epsilon/2}$ times greater than the density inside $K \backslash U$. It implies

$$\frac{\mu_{\xi f}(U)}{\mu_{\xi f}(K \backslash U)} \geq e^{\xi\epsilon/2} \frac{\text{vol}(\mathcal{B}(x^*; r) \cap K)}{\text{vol}(K \backslash U)} \geq e^{\xi\epsilon/2} \frac{\text{vol}(\mathcal{B}(x^*; r) \cap K)}{\text{vol}(K)}. \tag{63}$$

Without loss of generality, we assume that K is the unit ball centered at the origin. Consider the Euclidean ball $\mathcal{B}(x'; r/2)$ where $x' = \max\{0, 1 - r/(2\|x^*\|_2)\}x^*$. It is easy to verify that $\|x'\|_2 \leq 1 - r/2$ and $\|x' - x^*\|_2 \leq r/2$, which implies $\mathcal{B}(x'; r/2) \subset \mathcal{B}(x^*; r) \cap K$. Combining this relation with inequality (63), we have

$$\frac{\mu_{\xi f}(U)}{\mu_{\xi f}(K \backslash U)} \geq e^{\xi\epsilon/2} \frac{\text{vol}(\mathcal{B}(x'; r/2))}{\text{vol}(K)} = e^{\xi\epsilon/2 - d\log(2/r)} = e^{\xi\epsilon/2 - d\log(4G/\epsilon)}.$$

The right-hand side is greater than or equal to 1, because we have assumed $\xi \geq \frac{2d\log(4G/\epsilon)}{\epsilon}$. As a consequence, we have $\mu_{\xi f}(K \backslash U) \leq 1/2$.

Appendix D. Proof of Lemma 3

Consider a sufficiently small ϵ and a continuous mapping $\pi(x) := x - \epsilon\phi(x)$. Since ϕ is continuously differentiable in the compact set K, there exists a constant G such that $\|\phi(x) - \phi(y)\|_2 \leq G\|x - y\|_2$ for any $x, y \in K$. Assuming $\epsilon < 1/G$, it implies

$$\|\pi(x) - \pi(y)\|_2 \geq \|x - y\|_2 - \epsilon\|\phi(x) - \phi(y)\|_2 > 0 \quad \text{for any } x \neq y.$$

Thus, the mapping π is a continuous one-to-one mapping. For any set $A \subset K$, we define $\pi(A) := \{\pi(s) : x \in A\}$.

Since the parameter set K is compact, we can partition K into a finite number of small compact subsets, such that each subset has diameter at most $\delta := \epsilon^2$. Let S be the collection of these subsets

that intersect with A. The definition implies $A \subset \cup_{B \in S} B \subset A_\delta$. The fact that $\|\phi(x)\|_2 \leq 1$ implies

$$\mu_f(\pi(\cup_{B \in S} B)) \leq \mu_f(\pi(A_\delta)) \leq \mu_f(A_{\delta+\epsilon}).$$

As a consequence, we have:

$$\frac{\mu(A_{\delta+\epsilon})}{\mu(A)} \geq \frac{\mu_f(\pi(\cup_{B \in S} B))}{\mu_f(\cup_{B \in S} B)} = \frac{\sum_{B \in S} \mu_f(\pi(B))}{\sum_{B \in S} \mu_f(B)} \geq \min_{B \in S} \frac{\mu_f(\pi(B))}{\mu_f(B)}. \tag{64}$$

For arbitrary $B \in S$, we consider a point $x \in B \cap A$, and remark that every point in B is δ-close to the point x. Since ϕ is continuously differentiable, the Jacobian matrix of the transformation π has the following expansion:

$$J(y) = I - \epsilon H(x) + r_1(x, y) \quad \text{where } J_{ij}(x) = \frac{\partial \pi_i(x)}{\partial x_j}, \tag{65}$$

where H is the Jacobian matrix of ϕ satisfying $H_{ij}(x) = \frac{\partial \phi_i(x)}{\partial x_j}$. The remainder term $r_1(x, y)$, as a consequence of the continuous differentiability of ϕ and the fact $\|y - x\|_2 \leq \delta = \epsilon^2$, satisfies $\|r_1(x, y)\|_2 \leq C_1 \epsilon^2$ for some constant C_1.

On the other hand, using the relation $\nabla \mu_f(y) = -\mu_f(y) \nabla f(y)$ and the continuous differentiability of μ_f, the density function at $\pi(y)$ can be approximated by

$$\mu_f(\pi(y)) = \mu_f(y) + \nabla \mu_f(y)(\pi(y) - y) + r_2(y) = \left(1 + \epsilon \langle \phi(y), \nabla f(y) \rangle\right) \mu_f(y) + r_2(y),$$

where the remainder term $r_2(y)$ satisfies $|r_2(y)| \leq C_2 \epsilon^2$ for some constant C_2. Further using the continuity of ϕ, ∇f and the fact $\|y - x\|_2 \leq \epsilon^2$, we obtain:

$$\mu_f(\pi(y)) = \left(1 + \epsilon \langle \phi(x), \nabla f(x) \rangle\right) \mu_f(y) + r_3(x, y), \tag{66}$$

where the remainder term $r_3(x, y)$ satisfies $|r_3(x, y)| \leq C_3 \epsilon^2$ for some constant C_3.

Combining equation (65) and equation (66), we can quantify the measure of the set $\pi(B)$ using that of the set B. In particular, we have

$$\begin{aligned}
\mu_f(\pi(B)) &= \int_B \mu_f(\pi(y)) d\pi(y) = \int_B \mu_f(\pi(y)) \det(J(y)) dy \\
&= \int_B \left\{\left(1 + \epsilon \langle \phi(x), \nabla f(x) \rangle\right) \mu_f(y) + r_3(x, y)\right\} \det(I - \epsilon H(x) + r_1(x, y)) dy \\
&= \left(1 + \epsilon \langle \phi(x), \nabla f(x) \rangle - \epsilon \operatorname{tr}(H(x))\right) \int_B \mu_f(y) dy + \mathcal{O}(\epsilon^2). \\
&= \left(1 + \epsilon \langle \phi(x), \nabla f(x) \rangle - \epsilon \operatorname{tr}(H(x))\right) \mu_f(B) + \mathcal{O}(\epsilon^2). \tag{67}
\end{aligned}$$

Plugging equation (67) to the lower bound (64) and using the relation $\operatorname{tr}(H(x)) = \operatorname{div} \phi(x)$, implies

$$\begin{aligned}
\frac{\mu(A_{\delta+\epsilon})}{\mu(A)} - 1 &\geq \min_{B \in S} \left(\langle \phi(x), \nabla f(x) \rangle - \operatorname{div} \phi(x)\right) \epsilon + \mathcal{O}(\epsilon^2) \\
&\geq \inf_{x \in A} \left(\langle \phi(x), \nabla f(x) \rangle - \operatorname{div} \phi(x)\right) \epsilon + \mathcal{O}(\epsilon^2),
\end{aligned}$$

Finally, plugging in the definition of the restricted Cheeger constant and taking the limit $\epsilon \to 0$ completes the proof.

Appendix E. Proof of Proposition 5

Notations   Let Φ denote the CDF of the standard normal distribution. The function Φ satisfies the following tail bounds:

$$0 \le \Phi(t) \le e^{-t^2/2} \quad \text{for any} \quad t \le 0. \tag{68}$$

We define an auxiliary variable σ based on the value of ϵ:

$$\sigma := \frac{1}{2\sqrt{\log(4L/\sqrt{\epsilon})}}. \tag{69}$$

Since $e^{-1/(2\sigma^2)} = \frac{\sqrt{\epsilon}}{4L}$, the tail bound (68) implies $\Phi(t) \le \frac{\sqrt{\epsilon}}{4L}$ for all $t \le -\frac{1}{\sigma}$.

Define a vector field   Let $g(x) := \|\nabla f(x)\|_2$ be a shorthand notation. We define a vector field:

$$\phi(x) := \frac{1}{(2G+1)G} \underbrace{\Big(2\sqrt{G\,g(x)}\,I + \Phi\Big(\frac{-\sqrt{\epsilon}I - \nabla^2 f(x)}{\sigma\sqrt{\epsilon}}\Big) \Big)}_{:=A(x)} \nabla f(x) \tag{70}$$

Note that the function Φ admits a polynomial expansion:

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{(-1)^j x^{2j+1}}{j!(2j+1)}$$

Therefore, for any symmetric matrix $A \in \mathbb{R}^{d\times d}$, the matrix $\Phi(A)$ is well-defined by:

$$\Phi(A) := \frac{I}{2} + \frac{1}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{(-1)^j A^{2j+1}}{j!(2j+1)}, \tag{71}$$

We remark that the matrix definition (71) implies $\Phi(A + dA) = \Phi(A) + \Phi'(A)dA$ where Φ' is the derivative of function Φ.

Verify the condition of Lemma 3   The matrix $A(x)$ satisfies $0 \preceq A(x) \preceq (2G+1)I$, so that $\|\phi(x)\|_2 \le 1$ holds. For points that are r_0-close to the boundary, we have $\langle x, \nabla f(x)\rangle \ge \|x\|_2$. By these lower bounds and definition (70), we obtain:

$$\|x - \epsilon\,\phi(x)\|_2^2 \le \|x\|_2^2 + \epsilon^2 - \frac{\epsilon}{(2G+1)G}\Big(2\sqrt{G\,g(x)}\langle x, \nabla f(x)\rangle - \|x\|_2 \cdot \|\nabla f(x)\|_2 \Big)$$

$$\le \|x\|_2^2 + \epsilon^2 - \frac{\epsilon\|x\|_2}{(2G+1)G}\Big(2\sqrt{G\,g(x)} - g(x) \Big).$$

$$\le \|x\|_2^2 + \epsilon^2 - \frac{\epsilon\|x\|_2\sqrt{G\,g(x)}}{(2G+1)G} \le \|x\|_2^2 + \epsilon^2 - \frac{\epsilon\|x\|_2}{(2G+1)\sqrt{G}}, \tag{72}$$

where the last inequality holds because $g(x) \ge \langle x, \nabla f(x)\rangle/\|x\|_2 \ge 1$. For any $\epsilon < \frac{r-r_0}{(2G+1)\sqrt{G}}$, the right-hand side is smaller than $\|x\|_2^2$, so that $x - \epsilon\,\phi(x) \in K$. For points that are not r_0-close to the boundary, we have $x - \epsilon\,\phi(x) \in K$ given $\epsilon < r_0$. Combining results for the two cases, we conclude that ϕ satisfies the conditions of Lemma 3

Prove the Lower bound   By applying Lemma 3, we obtain the following lower bound:

$$\mathcal{C}_{(\xi f)}(K\backslash U) \ge \frac{1}{(2G+1)G} \inf_{x\in K\backslash U} \Big\{ \xi\,(\nabla f(x))^\top A(x)\nabla f(x) - \operatorname{div} A(x)\nabla f(x) \Big\}. \tag{73}$$

Since $A(x) \succeq 2\sqrt{G\,g(x)}I$, the term $(\nabla f(x))^\top A(x)\nabla f(x)$ is lower bounded by $2\sqrt{G}(g(x))^{5/2}$. For the second term, we claim the following bound:

$$\text{div}\, A(x)\nabla f(x) \le 3\sqrt{G\,g(x)}L + \frac{g(x)H}{\sigma\sqrt{\epsilon}} + \frac{\sqrt{\epsilon}}{4} - \mathbb{I}[g(x) < \epsilon]\,\frac{\sqrt{\epsilon}}{2} \tag{74}$$

We defer the proof to Appendix E.1 and focus on its consequence. Combining inequalities (73) and (74), we obtain

$$\mathcal{C}_{(\xi f)}(K\backslash U) \ge \frac{1}{(2G+1)G} \inf_{x \in K\backslash U}$$
$$\left\{ 2\sqrt{G}\xi(g(x))^{5/2} + \mathbb{I}[g(x) < \epsilon]\,\frac{\sqrt{\epsilon}}{2} - 3\sqrt{G\,g(x)}L - \frac{g(x)H}{\sigma\sqrt{\epsilon}} - \frac{\sqrt{\epsilon}}{4} \right\}. \tag{75}$$

The right-hand side of inequality (75) can be made strictly positive if we choose a large enough ξ. In particular, we choose:

$$\xi \ge \frac{1}{\epsilon^2 G^{1/2}} \cdot \max\left\{ \frac{6G^{1/2}L}{h^2}, \frac{2H}{\sigma h^{3/2}}, \frac{1}{2h^{5/2}} \right\} \quad \text{where} \quad h := \min\left\{ 1, \frac{1}{G(48L)^2}, \frac{\sigma}{16H} \right\}. \tag{76}$$

To proceed, we do a case study based on the value of $g(x)$. For all x satisfying $g(x) < h\epsilon$, we plug in the upper bound $g(x) < h\epsilon$ for $g(x)$, then plug in the definition of h. It implies:

$$\mathbb{I}[g(x) < \epsilon]\,\frac{\sqrt{\epsilon}}{2} - 3\sqrt{G\,g(x)}L - \frac{g(x)H}{\sigma\sqrt{\epsilon}} - \frac{\sqrt{\epsilon}}{4}$$
$$\ge \frac{\sqrt{\epsilon}}{4} - 3\sqrt{G}L \cdot \sqrt{\frac{\epsilon}{G(48L)^2}} - \frac{H}{\sigma\sqrt{\epsilon}} \cdot \frac{\epsilon\sigma}{16H} = \frac{\sqrt{\epsilon}}{8}. \tag{77}$$

For all x satisfying $g(x) \ge h\epsilon$, we ignore the non-negative term $\mathbb{I}[g(x) < \epsilon]\,\frac{\sqrt{\epsilon}}{2}$ on the right-hand side of (75), then re-arrange the remaining terms. It gives:

$$2\sqrt{G}\xi(g(x))^{5/2} - 3\sqrt{G\,g(x)}L - \frac{g(x)H}{\sigma\sqrt{\epsilon}} - \frac{\sqrt{\epsilon}}{4} \ge \frac{\xi\sqrt{G}(g(x))^{5/2}}{2} + \sqrt{G\,g(x)}\left(\frac{\xi(g(x))^2}{2} - 3L \right)$$
$$+ g(x)\left(\frac{\xi\sqrt{G}(g(x))^{3/2}}{2} - \frac{H}{\sigma\sqrt{\epsilon}} \right) + \left(\frac{\xi\sqrt{G}(g(x))^{5/2}}{2} - \frac{\sqrt{\epsilon}}{4} \right).$$

Using lower bound (76) for ξ and the lower bound $g(x) \ge h\epsilon$ for $g(x)$, it is easy to verify that the last three terms on the right-hand side are non-negative. Furthermore, plugging in the lower bound $\xi \ge \frac{1}{\epsilon^2 G^{1/2}} \cdot \frac{1}{2h^{5/2}}$ from (76), it implies:

$$2\sqrt{G}\xi(g(x))^{5/2} - 3\sqrt{g(x)}L - \frac{g(x)H}{\sigma\sqrt{\epsilon}} - \frac{\sqrt{\epsilon}}{4} \ge \sqrt{G}\xi \cdot \frac{(h\epsilon)^{5/2}}{2} \ge \frac{1}{2\epsilon^2 h^{5/2}} \cdot \frac{(h\epsilon)^{5/2}}{2} = \frac{\sqrt{\epsilon}}{4} \tag{78}$$

Combining inequalities (75), (77), (78) proves that the restricted Cheeger constant is lower bounded by $\frac{\sqrt{\epsilon}}{8(2G+1)G}$. Since $1/\sigma = \widetilde{\mathcal{O}}(1)$, it is easy to verify that the constraint (76) can be satisfied if we choose:

$$\xi \ge \widetilde{\mathcal{O}}(1) \cdot \frac{\max\{1, G^{5/2}L^5, H^{5/2}\}}{\epsilon^2 G^{1/2}},$$

which completes the proof.

E.1. Proof of inequality (74)

Let $T(x)$ be the third order tensor such that $T_{ijk}(x) = \frac{\partial^3 f(x)}{\partial x_i \partial x_j \partial x_k}$. Consider an arbitrary unit vector $u \in \mathbb{R}^d$. By the definition of $A(x)\nabla f(x)$, we have:

$$\lim_{t \to 0} \frac{A(x+ut)\nabla f(x+ut) - A(x)\nabla f(x)}{t} = A(x)\nabla^2 f(x)\, u$$
$$+ \frac{\sqrt{G}\nabla f(x)(\nabla g(x))^\top}{\sqrt{g(x)}}u - \frac{1}{\sigma\sqrt{\epsilon}}\Phi'\Big(\frac{-\sqrt{\epsilon}I - \nabla^2 f(x)}{\sigma\sqrt{\epsilon}}\Big)T(x)[u]\nabla f(x),$$

where the matrix $T(x)[u] \in \mathbb{R}^{d \times d}$ is defined by $(T(x)[u])_{ij} = \sum_{k=1}^d T_{ijk}(x)u_k$. By simple algebra, we obtain $T(x)[u]\nabla f(x) = T(x)[\nabla f(x)]u$. Thus, the derivative of the vector field $A(x)\nabla f(x)$ can be represented by $D_{ij} = \frac{\partial(A(x)\nabla f(x))_i}{\partial x_j}$ where D is the following d-by-d matrix:

$$D := A(x)\nabla^2 f(x) + \frac{\sqrt{G}\nabla f(x)(\nabla g(x))^\top}{\sqrt{g(x)}} - \frac{1}{\sigma\sqrt{\epsilon}}\Phi'\Big(\frac{-\sqrt{\epsilon}I - \nabla^2 f(x)}{\sigma\sqrt{\epsilon}}\Big)T(x)[\nabla f(x)]. \quad (79)$$

Note that $\operatorname{div}(A(x)\nabla f(x))$ is equal to the trace of matrix D. In order to proceed, we perform a case study on the value of $g(x)$.

Case $g(x) < \epsilon$: We first upper bound the trace of $A(x)\nabla f^2(x)$, which can be written as:

$$A(x)\nabla f^2(x) = 2\sqrt{G\,g(x)}\nabla f^2(x) + \Phi\Big(\frac{-\sqrt{\epsilon}I - \nabla^2 f(x)}{\sigma\sqrt{\epsilon}}\Big)\nabla f^2(x). \quad (80)$$

The trace of the first term on the right-hand side is bounded by $2\sqrt{G\,g(x)}L$. For the second term, we assume that the matrix $\nabla^2 f(x)$ has eigenvalues $\lambda_1 \le \lambda_2 \le \dots \lambda_d$ with associated eigenvectors $u_1, \dots, u_d$. As a consequence, the matrix $\Phi(\frac{-\sqrt{\epsilon}I - \nabla^2 f(x)}{\sigma\sqrt{\epsilon}})$ has the same set of eigenvectors, but with eigenvalues $\Phi(\frac{-\lambda_1/\sqrt{\epsilon}-1}{\sigma}), \dots, \Phi(\frac{-\lambda_d/\sqrt{\epsilon}-1}{\sigma})$. Thus, the trace of this term is equal to

$$\operatorname{tr}\Big(\Phi\Big(\frac{-\sqrt{\epsilon}I - \nabla^2 f(x)}{\sigma\sqrt{\epsilon}}\Big)\nabla f^2(x)\Big) = \sum_{i=1}^d \lambda_i\, \Phi\Big(\frac{-\lambda_i/\sqrt{\epsilon}-1}{\sigma}\Big).$$

By the assumptions $x \in K\backslash U$ and $g(x) < \epsilon$, and using the definition of ϵ-approximate local minima, we obtain $\lambda_1 \le -\sqrt{\epsilon}$. As a consequence

$$\lambda_1\, \Phi\Big(\frac{-\lambda_1/\sqrt{\epsilon}-1}{\sigma}\Big) \le \lambda_1\, \Phi(0) \le -\sqrt{\epsilon}/2.$$

For other eigenvalues, if λ_i is negative, then we use the upper bound $\lambda_i\Phi(\frac{-\lambda_i/\sqrt{\epsilon}-i}{\sigma}) < 0$; If λ_i is positive, then we have $\lambda_i\Phi(\frac{-\lambda_i/\sqrt{\epsilon}-1}{\sigma}) \le \lambda_i\Phi(-\frac{1}{\sigma}) \le \lambda_i\frac{\sqrt{\epsilon}}{4L}$. Combining these relations, we have

$$\operatorname{tr}\Big(\Phi\Big(\frac{-\sqrt{\epsilon}I - \nabla^2 f(x)}{\sigma\sqrt{\epsilon}}\Big)\nabla f^2(x)\Big) \le -\frac{\sqrt{\epsilon}}{2} + \frac{\sqrt{\epsilon}}{4L}\sum_{i=1}^d [\lambda_i]_+ \le -\frac{\sqrt{\epsilon}}{2} + \frac{\sqrt{\epsilon}}{4L}\|\nabla^2 f(x)\|_* \le -\frac{\sqrt{\epsilon}}{4}.$$

Combining this inquality with the upper bound on the first term of (80), we obtain

$$\operatorname{tr}(A(x)\nabla f^2(x)) \le 2\sqrt{G\,g(x)}L - \sqrt{\epsilon}/4. \quad (81)$$

Thus, we have upper bounded the trace of first term on the right-hand side of (79).

For the second term on the right-hand side of (79) , we have

$$\text{tr}\left(\frac{\sqrt{G}\nabla f(x)(\nabla g(x))^\top}{\sqrt{g(x)}}\right) = \frac{\sqrt{G}\langle\nabla f(x), \nabla g(x)\rangle}{\sqrt{g(x)}} \leq \sqrt{G\,g(x)}\,\|\nabla g(x)\|_2 \leq \sqrt{G\,g(x)}L, \quad (82)$$

where the last inequality uses the relation $\nabla g(x) = \frac{(\nabla^2 f(x))\nabla f(x)}{g(x)}$, so that $\|\nabla g(x)\|_2 \leq \|\nabla^2 f(x)\|_2 \leq \|\nabla^2 f(x)\|_* \leq L$.

For the third term on the right-hand side of (79), since $0 \preceq \Phi'(\frac{-\sqrt{\epsilon}I - \nabla^2 f(x)}{\sigma\sqrt{\epsilon}}) \preceq I$, we have

$$\text{tr(the third term)} \leq \|\text{the third term}\|_* \leq \frac{1}{\sigma\sqrt{\epsilon}}\|T(x)[\nabla f(x)]\|_*.$$

By Assumption B, the function f satisfies $\frac{\|\nabla^2 f(x) - \nabla^2 f(y)\|_*}{\|x-y\|_2} \leq H$, which implies $\|T(x)[u]\|_* \leq H\|u\|_2$ for any $x \in K$ and $u \in \mathbb{R}^d$. As a consequence, the term $\|T(x)[\nabla f(x)]\|_*$ is bounded by

$$\|T(x)[\nabla f(x)]\|_* \leq H\|\nabla f(x)\|_2 = g(x)H,$$

which further implies

$$\text{tr(the third term)} \leq \frac{g(x)H}{\sigma\sqrt{\epsilon}}. \tag{83}$$

Combining upper bounds (81), (82), (83) implies

$$\text{div } A(x)\nabla f(x) \leq 3\sqrt{G\,g(x)}L + \frac{g(x)H}{\sigma\sqrt{\epsilon}} - \sqrt{\epsilon}/4.$$

Case $g(x) \geq \epsilon$: The proof is similar to the previous case. For the first term on the right-hand side of equation (79), we follow the same arguments for establishing the upper bound (81), but without using the relation $\lambda_1 \leq -\sqrt{\epsilon}$ (because conditioning on $g(x) \geq \epsilon$, the definition of approximate local minima won't give $\lambda_1 \leq -\sqrt{\epsilon}$). Then the trace of $A(x)\nabla f^2(x)$ is bounded by:

$$\text{tr}(A(x)\nabla f^2(x)) \leq 2\sqrt{G\,g(x)}L + \frac{\sqrt{\epsilon}}{4L}\|\nabla^2 f(x)\|_* \leq 2\sqrt{g(x)}L + \sqrt{\epsilon}/4. \tag{84}$$

For the second and the third term, the upper bounds (82) and (83) still hold, so that

$$\text{div } A(x)\nabla f(x) \leq 3\sqrt{G\,g(x)}L + \frac{g(x)H}{\sigma\sqrt{\epsilon}} + \sqrt{\epsilon}/4.$$

Combining the two cases completes the proof.

Appendix F. Proof of Theorem 7

We apply the general Theorem 1 to prove this theorem. In order to apply Theorem 1, the first step is to show that the function $\tilde{f}_\sigma$ satisfies Assumption A. Recall that the function is defined by:

$$\tilde{f}_\sigma(x) = \mathbb{E}[f(x+z)] \quad \text{where} \quad z \sim N(0, \sigma^2 I), \tag{85}$$

and its stochastic gradient is computed by:

$$g(x) := \frac{z}{\sigma^2}(\ell(x+z; a) - \ell(x; a)).$$

By Assumption C, the function $\tilde{f}_\sigma$ is uniformly bounded in $[0, B]$. The following lemma captures additional properties of functions $\tilde{f}_\sigma$ and g. See Appendix F.1 for the proof.

Lemma 17 *The following properties hold:*

1. *For any $x \in K$, the stochastic function g satisfies $\mathbb{E}[g(x)|x] = \nabla \tilde{f}_\sigma(x)$. For any vector $u \in \mathbb{R}^d$ with $\|u\|_2 \leq \frac{\sigma}{2B}$, it satisfies*

$$\mathbb{E}[e^{\langle u, g(x) \rangle^2}|x] \leq e^{\|u\|_2^2 (2B/\sigma)^2}.$$

2. *The function $\tilde{f}_\sigma$ is $(2B/\sigma^2)$-smooth.*

Lemma 17 shows that $\tilde{f}_\sigma$ is an L-smooth function, with $L = (2B/\sigma^2)$. In addition, the stochastic gradient g satisfies the third condition of Assumption A with $b_{\max} = \frac{\sigma}{2B}$ and $G = \frac{2B}{\sigma}$. As a consequence, Theorem 1 implies the risk bound:

$$\tilde{f}_\sigma(\widehat{x}) \leq \sup_{x:\, d(x,U) \leq \rho} \tilde{f}_\sigma(x), \tag{86}$$

We claim the following inequality:

$$\text{If } \sigma = \frac{\nu}{\max\{G, B/\rho_K\}} \quad \text{then} \quad \sup_{x \in K} |\tilde{f}_\sigma(x) - F(x)| \leq 2\nu. \tag{87}$$

We defer the proof of claim (87) to the end of this section, focusing on its consequence. Let σ take the value in claim (87). The conseuqence of (87) and the G-Lipschitz continuity of the function F imply:

$$F(\widehat{x}) \leq \tilde{f}_\sigma(\widehat{x}) + 2\nu \leq \sup_{x:\, d(x,U) \leq \rho} \tilde{f}_\sigma(x) + 2\nu \leq \sup_{x:\, d(x,U) \leq \rho} F(x) + 4\nu \leq \sup_{x \in U} F(x) + 4\nu + G\rho,$$

By choosing $\rho := \nu/G$, we establish the risk bound $F(\widehat{x}) \leq \sup_{x \in U} F(x) + 5\nu$. It remains to establish the iteration complexity bound.

According to Theorem 1, by choosing stepsize $\eta := \eta_0$, SGLD achieves the risk bound (86) with iteration number polynomial in $(B, L, G, \log(1/\delta), d, \xi, h_{\max}^{-1}, b_{\max}^{-1}, \rho^{-1} \mathcal{C}_{(\xi \tilde{f}_\sigma)}^{-1}(K \backslash U))$, where $(L, G, b_{\max})$ depend on σ. Therefore, it remains to lower bound the restricted Cheeger constant $\mathcal{C}_{\xi \tilde{f}_\sigma}(K \backslash U)$. By combining the claim (87) with inequality (6), we obtain

$$\mathcal{C}_{(\xi \tilde{f}_\sigma)}(K \backslash U) \geq e^{-4\xi\nu} \mathcal{C}_{(\xi F)}(K \backslash U) \geq e^{-4} \mathcal{C}_{(\xi F)}(K \backslash U).$$

It means that $\mathcal{C}_{(\xi \tilde{f}_\sigma)}(K \backslash U)$ and $\mathcal{C}_{(\xi F)}(K \backslash U)$ differs by a constant multiplicative factor. Finally, plugging in the values of ρ and σ completes the proof.

Proof of Claim (87) We define an auxiliary function $\tilde{F}_\sigma$ as follow:

$$\tilde{F}_\sigma(x) = \mathbb{E}_z[F(x + z)] \quad \text{where} \quad z \sim N(0, \sigma^2 I), \tag{88}$$

Since $f(x) \in [F(x) - \nu, F(x) + \nu]$, the definitions (85) and (88) imply $\tilde{f}_\sigma(x) \in [\tilde{F}_\sigma(x) - \nu, \tilde{F}_\sigma(x) + \nu]$. The G-Lipschitz continuity of function F implies that for any $x \in K$ and $y \in \overline{K}$, there is $|F(y) - F(x)| \leq G\|y - x\|_2$. For any $x \in K$ and $y \notin \overline{K}$, we have $F(x), F(y) \in [0, B]$ and that the distance between x and y is at least ρ_K, thus $|F(y) - F(x)| \leq B \leq (B/\rho_K)\|y - x\|_2$. As a consequence, for any $x \in K$ we have:

$$|F(x + z) - F(x)| \leq \mathbb{I}[x + z \in \overline{K}] G\|z\|_2 + \mathbb{I}[x + z \notin \overline{K}] (B/\rho_K)\|z\|_2 \leq \max\{G, B/\rho_K\} \|z\|_2.$$

Taking expectation over z on both sides and using Jensen's inequality, we obtain

$$|\tilde{F}_\sigma(x) - F(x)| \leq \mathbb{E}[|F(x + z) - F(x)|] \leq \max\{G, B/\rho_K\} \mathbb{E}[\|z\|_2] \leq \max\{G, B/\rho_K\} \sigma.$$

Thus, by choosing $\sigma := \frac{\nu}{\max\{G, B/\rho_K\}}$, it ensures that for any $x \in K$:

$$|\tilde{f}_\sigma(x) - F(x)| \le |\tilde{f}_\sigma(x) - \tilde{F}_\sigma(x)| + |\tilde{F}_\sigma(x) - F(x)| \le 2\nu.$$

F.1. Proof of Lemma 17

(1) The function $\tilde{f}_\sigma$ is a differentiable function, because it is the convolution of a bounded function f and a Gaussian density function (which is infinite-order differentiable). We can write the gradient vector $\nabla \tilde{f}_\sigma(x)$ as:

$$\nabla \tilde{f}_\sigma(x) = \frac{1}{(2\pi)^{d/2}} \frac{\partial}{\partial x} \int e^{-\frac{\|z\|_2^2}{2\sigma^2}} f(x+z) dz.$$

Let $z' := x + z$. By change of variables, the above equation implies:

$$\nabla \tilde{f}_\sigma(x) = \frac{1}{(2\pi)^{d/2}} \frac{\partial}{\partial x} \int e^{-\frac{\|z'-x\|_2^2}{2\sigma^2}} f(z') dz' = \frac{1}{(2\pi)^{d/2}} \int \left(\frac{\partial}{\partial x} e^{-\frac{\|z'-x\|_2^2}{2\sigma^2}} \right) f(z') dz'.$$

$$= \frac{1}{(2\pi)^{d/2}} \int \frac{z'-x}{\sigma^2} e^{-\frac{\|x-z'\|_2^2}{2\sigma^2}} f(z') dz' = \frac{1}{(2\pi)^{d/2}} \int \frac{z}{\sigma^2} e^{-\frac{\|z\|_2^2}{2\sigma^2}} f(x+z) dz$$

$$= \mathbb{E}\left[\frac{z}{\sigma^2} f(x+z) \right] \overset{(i)}{=} \mathbb{E}\left[\frac{z}{\sigma^2} (f(x+z) - f(x)) \right] \overset{(ii)}{=} \mathbb{E}\left[\frac{z}{\sigma^2} (\ell(x+z; a) - \ell(x; a)) \right]. \quad (89)$$

For the above deduction, equation (i) holds because $\mathbb{E}[zf(x)] = \mathbb{E}[z]\mathbb{E}[f(x)] = 0$; equation (ii) holds because $\mathbb{E}[\ell(y; a)|y] = f(y)$ for any $y \in K$. It shows that $g(x)$ is an unbiased estimate of $\nabla \tilde{f}_\sigma(x)$. Since $\ell(\cdot; a) \in [0, B]$, any vector $u \in \mathbb{R}^d$ satisfies

$$(\langle u, g(x) \rangle)^2 = (\ell(x+z; a) - \ell(x; a))^2 (\langle u, \frac{z}{\sigma^2} \rangle)^2 \le B^2 (\langle u, \frac{z}{\sigma^2} \rangle)^2 = \frac{B^2 \|u\|_2^2}{\sigma^2} (\langle \frac{u}{\|u\|_2}, \frac{z}{\sigma} \rangle)^2$$

Thus the following bound holds:

$$\mathbb{E}[e^{(\langle u, g(x) \rangle)^2} | x] \le \mathbb{E}[e^{\frac{B^2 \|u\|_2^2}{\sigma^2} (\langle u/\|u\|_2, z/\sigma \rangle)^2}].$$

Notice that $\langle u/\|u\|_2, z/\sigma \rangle$ satisfies the standard normal distribution. Thus the right-hand side of the above inequality is bounded by

$$\mathbb{E}[e^{\frac{B^2 \|u\|_2^2}{\sigma^2} (\langle u/\|u\|_2, z/\sigma \rangle)^2}] = \frac{1}{\sqrt{1 - 2B^2\|u\|_2^2/\sigma^2}} \le e^{B^2\|u\|_2^2(2/\sigma)^2} \quad \text{if} \quad \|u\|_2 \le \frac{\sigma}{2B}.$$

Combining the two inequalities above completes the proof.

(2) In order to bound the smoothness of the function $\tilde{f}_\sigma(x)$, we derive the second derivative of $\tilde{f}_\sigma(x)$ using equation (89):

$$\nabla^2 \tilde{f}_\sigma(x) = \frac{\partial}{\partial x}(\nabla \tilde{f}_\sigma(x)) = \frac{1}{(2\pi)^{d/2}} \frac{\partial}{\partial x} \int \frac{z}{\sigma^2} e^{-\frac{\|z\|_2^2}{2\sigma^2}} f(x+z) dz$$

$$= \frac{1}{(2\pi)^{d/2}} \frac{\partial}{\partial x} \int \frac{z'-x}{\sigma^2} e^{-\frac{\|z'-x\|_2^2}{2\sigma^2}} f(z') dz' = \frac{1}{(2\pi)^{d/2}} \int \left(\frac{\partial}{\partial x} \frac{z'-x}{\sigma^2} e^{-\frac{\|z'-x\|_2^2}{2\sigma^2}} \right) f(z') dz'$$

$$= \frac{1}{(2\pi)^{d/2}} \int \left(\frac{(z'-x)(z'-x)^\top}{\sigma^4} - \frac{I}{\sigma^2} \right) e^{-\frac{\|z'-x\|_2^2}{2\sigma^2}} f(z') dz' = \mathbb{E}\left[\frac{zz^\top - \sigma^2 I}{\sigma^4} f(x+z) \right].$$

$$(90)$$

Using the fact $f(x + z) \in [0, B]$, equation (90) implies:

$$\|\nabla^2 \tilde{f}_\sigma(x)\|_2 \leq \frac{1}{\sigma^4} \|\mathbb{E}[zz^\top f(x + z)]\|_2 + \frac{1}{\sigma^2} \|\mathbb{E}[f(x + z)I]\|_2 \leq \frac{B}{\sigma^4} \|\mathbb{E}[zz^\top]\|_2 + \frac{B}{\sigma^2} = \frac{2B}{\sigma^2},$$

which establishes that the function $\tilde{f}_\sigma$ is $(2B/\sigma^2)$-smooth.

Appendix G. Proof of Theorem 9

We use Theorem 7 to upper bound the population risk as well as the time complexity. To apply the theorem, we need to verify Assumption C. Recall that the parameter space is defined by $K := \{x \in \mathbb{R}^d : 1/2 \leq \|x\|_2 \leq 1\}$. Let $\overline{K} := \{x \in \mathbb{R}^d : 1/4 \leq \|x\|_2 \leq 5/4\}$ be an auxiliary super-set. The following lemma shows that these assumptions hold under our problem set-up.

Lemma 18 *The following properties hold:*

(1) There exists $h_{max} = \Omega(d^{-2})$ such that for any $x \in K$, $h \leq h_{max}$ and $y \sim N(x, 2hI)$, we have $P(y \in K) \geq 1/3$.

(2) The function F is 3-Lipschitz continuous in $\overline{K}$.

(3) For any $\nu, \delta > 0$, if the sample size n satisfies $n \gtrsim \frac{d}{\nu^2}$, then with probability at least $1 - \delta$ we have $\sup_{x \in \overline{K}} |f(x) - F(x)| \leq \nu$. The notation "$\lesssim$" hides a poly-logarithmic function of $(d, 1/\nu, 1/\delta)$.

See Appendix G.1 for the proof.

Let $\alpha_0 \in (0, \pi/4]$ be an arbitrary angle. We define $U \subset K$ to be the set of points such that the angle between the point and x^* is bounded by α_0, or equivalently:

$$U := \{x \in K : \langle x/\|x\|_2, x^* \rangle \geq \cos(\alpha_0)\}.$$

For any $x \in K$, the 3-Lipschitz continuity of function F implies:

$$F(x) = F(x/\|x\|_2) \leq F(x^*) + 3 \left\| \frac{x}{\|x\|_2} - x^* \right\|_2.$$

By simple geometry, it is easy to see that

$$\left\| \frac{x}{\|x\|_2} - x^* \right\|_2 = 2 \sin(\alpha/2) \leq 2 \sin(\alpha_0/2) \leq 2 \sin(\alpha_0).$$

Thus, we have

$$F(x) \leq F(x^*) + 6 \sin(\alpha_0). \tag{91}$$

Inequality (91) implies that for small enough α_0, any point in U is a nearly optimal solutions. Thus we can use U as a target optimality set. The following lemma lower bounds the restricted Cheeger constant for the set U.

Lemma 19 *Assume that $d \geq 2$. For any $\alpha_0 \in (0, \pi/4]$, there are universal constant $c_1, c_2 > 0$ such that if we choose $\xi \geq \frac{c_1 d^{3/2}}{q_0 \sin^2(\alpha_0)}$, then the restricted Cheeger constant is lower bounded by $\mathcal{C}_{(\xi F)}(K \backslash U) \geq c_2 d$.*

See Appendix G.2 for the proof.

Given a target optimality $\epsilon > 0$, we choose $\alpha_0 := \arcsin(\epsilon/12)$. The risk bound (91) implies

$$F(x) \leq F(x^*) + \epsilon/2 \quad \text{for all} \quad x \in U. \tag{92}$$

Lemma 18 ensures that the pre-conditions of Theorem 7 hold with a small enough quantity ν. Combining Theorem 7 with inequality (92), with probability at least $1 - \delta$, SGLD achieves the risk bound:

$$F(\widehat{x}) \leq \sup_{x \in U} F(x) + 5\nu \leq F(x^*) + \epsilon/2 + 5\nu. \tag{93}$$

In order to have a small enough ν, we want the functions f and F to be uniformly close. More precisely, we want the gap between them to satisfy:

$$\sup_{x \in \overline{K}} |f(x) - F(x)| \leq \nu := \min \left\{ \frac{q_0 \sin^2(\alpha_0)}{c_1 d^{3/2}}, \epsilon/10 \right\}, \tag{94}$$

By Lemma 18, this can be achieved by assuming a large enough sample size n. In particular, if the sample size satisfies $n \gtrsim \frac{d^4}{q_0^2 \epsilon^4}$, then inequality (94) is guaranteed to be true. The notation "$\gtrsim$" hides a poly-logarithmic function.

If inequity (94) holds, then $\nu \leq \epsilon/10$ holds, so that we can rewrite the risk bound (93) as $F(\widehat{x}) \leq F(x^*) + \epsilon$. By combining the choice of ν in (94) with the choice of $\xi := \frac{c_1 d^{3/2}}{q_0 \sin^2(\alpha_0)}$ in Lemma 19, we find that the relation $\xi \in (0, 1/\nu]$ hold, satisfying Theorem 7's condition on (ν, ξ). As a result, Theorem 7 implies that the iteration complexity of SGLD is bounded by the restricted Cheeger constant $\mathcal{C}_{(\xi F)}(K \backslash U)$. By Lemma 19, the restricted Cheeger constant is lower bounded by $\Omega(d)$, so that the iteration complexity is polynomial in $(d, 1/q_0, 1/\epsilon, \log(1/\delta))$.

G.1. Proof of Lemma 18

(1) Let $x \in K$ be an arbitrary point and let $z \sim N(0, 2hI)$. An equivalent way to express the relation $x + z \in K$ is the following sandwich inequality:

$$1/4 - \|x\|_2^2 - \|z\|_2^2 \leq 2\langle x, \, z \rangle \leq 1 - \|x\|_2^2 - \|z\|_2^2. \tag{95}$$

For any $t > 0$, we consider a sufficient condition for inequality (95):

$$\|z\|_2^2 \leq 2thd \quad \text{and} \quad 2\langle x, \, z \rangle \in I_x := [1/4 - \|x\|_2^2, 1 - \|x\|_2^2 - 2thd].$$

The random variable $\frac{\|z\|_2^2}{2h}$ satisfies a chi-square distribution with d degrees of freedom. By Lemma 16, for any $t \geq 5$, the condition $\|z\|_2^2 \leq 2thd$ holds with probability at least $1 - e^{-\Omega(td)}$.

Suppose that t is a fixed constant, and h is chosen to be $h := \frac{c^2}{2td^2}$ for a constant $c > 0$. Then the random variable $w_x := 2\langle x, \, z \rangle$ satisfies a normal distribution $N(0; \frac{4\|x\|_2^2 (c/d)^2}{t})$. The interval I_x, no matter how $x \in K$ is chosen, covers either $[-1/4, -(c/d)^2]$ or $[(c/d)^2, 1/4]$. For $c \to 0$, we have $(c/d)^2 \ll \frac{2\|x\|_2}{t}(c/d) \ll 1/4$, so that the probability of $w_x \in I_x$ is asymptotically lower bounded by 0.5. It implies that there is a strictly positive constant c (depending on the value of t) such that $P(w_x \in I_x) \geq 0.4$ for all $x \in K$. With this choice of c, we apply the union bound:

$$P(x + z \in K) \geq P(w_x \in I_x) - P(\|z\|_2^2 > 2thd) \geq 0.4 - e^{-\Omega(td)}.$$

By choosing t to be a large enough constant, the above probability is lower bounded by $1/3$.

(2) For two vectors $x, y \in \overline{K}$, the loss values $\ell(x; a)$ and $\ell(y; a)$ are non-equal only when $\text{sign}(\langle x, a \rangle) \neq \text{sign}(\langle y, a \rangle)$. Thus, we have the upper bound

$$|F(x) - F(y)| \leq P\Big(\text{sign}(\langle x, a \rangle) \neq \text{sign}(\langle y, a \rangle)\Big). \tag{96}$$

If we change the distribution of a from uniform distribution to a normal distribution $N(0, I_{d \times d})$, the right-hand side of inequality (96) won't change. Both $\langle x, a \rangle$ and $\langle x, b \rangle$ become normal random variables with correlation coefficient $\frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2}$. Under this setting, Tong (2012) proved that the right-side is equal to

$$P\Big(\text{sign}(\langle x, a \rangle) \neq \text{sign}(\langle y, a \rangle)\Big) = \frac{1}{\pi} \arccos\Big(\frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2}\Big)$$

By simple algebra and using the fact that $\|x\|_2, \|y\|_2 \geq 1/4$, we have

$$\frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2} = \frac{1}{2}\Big(\frac{\|x\|_2}{\|y\|_2} + \frac{\|y\|_2}{\|x\|_2} - \frac{\|x - y\|_2^2}{\|x\|_2 \|y\|_2}\Big) \geq 1 - 8\|x - y\|_2^2,$$

Combining the above relations, and using the fact that $\arccos(t) \leq 3\sqrt{1 - t}$ for any $t \in [-1, 1]$, we obtain

$$|F(x) - F(y)| = \frac{1}{\pi} \arccos\Big(\frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2}\Big) \leq \frac{3\sqrt{8}}{\pi} \|x - y\|_2 \leq 3\|x - y\|_2,$$

which shows that the function F is 3-Lipschitz continuous in $\overline{K}$.

(3) Since function f is the empirical risk of a linear classifier, its uniform convergence rate can be characterized by the VC-dimension. The VC-dimension of linear classifiers in a d-dimensional space is equal to $d + 1$. Thus, the concentration inequality of Vapnik (1999) implies that with probability at least $1 - \delta$, we have

$$\sup_{x \in \mathbb{R}^d} |f(x) - F(x)| \leq U(n) := c\sqrt{\frac{d(\log(n/d) + 1) + \log(1/\delta)}{n}},$$

where $c > 0$ is a universal constant. When $n \geq d$, the upper bound $U(n)$ is a monotonically decreasing function of n. In order to guarantee $U(n) \leq \nu$, it suffices to choose $n \geq n_0$ where the number n_0 satisfies:

$$n_0 = \max\{n \in \mathbb{R} : U(n) = \nu\}. \tag{97}$$

It is easy to see that n_0 is polynomial in $(d, 1/\nu, 1/\delta)$. Thus, by the definition of $U(n)$, we have

$$\nu = c\sqrt{\frac{d(\log(n_0/d) + 1) + \log(1/\delta)}{n_0}} \leq c\sqrt{\frac{d \cdot \text{polylog}(d, 1/\nu, 1/\delta)}{n_0}}.$$

It implies $n_0 \lesssim \frac{d}{\nu^2}$, thus completes the proof.

G.2. Proof of Lemma 19

Note that the population risk F can be non-differentiable. In order to apply Lemma 3 to lower bound the restricted Cheeger constant, we define a smoothed approximation of F, and apply Lemma 3 on the smoothed approximation. For an arbitrary $\sigma > 0$, we define $\tilde{F}_\sigma$ to be a smoothed approximation

of the population risk:

$$\tilde{F}_\sigma(x) := \mathbb{E}[F(x+z)] \quad \text{where} \quad z \sim N(0, \sigma^2 I_{d \times d}).$$

By Lemma 18, the function F is 3-Lipschitz continuous, so that $\tilde{F}_\sigma$ uniformly converges to F as $\sigma \to 0$. It means that

$$\lim_{\sigma \to 0} \mathcal{C}_{(\xi \tilde{F}_\sigma)}(K \backslash U) = \mathcal{C}_{(\xi F)}(K \backslash U).$$

It suffices to lower bound $\mathcal{C}_{(\xi \tilde{F}_\sigma)}(K \backslash U)$ and then take the limit $\sigma \to 0$. The function $\tilde{F}_\sigma$ is continuously differentiable, so that we can use Lemma 3 to lower bound $\mathcal{C}_{(\xi \tilde{F}_\sigma)}(K \backslash U)$.

Consider an arbitrary constant $0 < t \le 1/6$. We choose a small enough $\sigma > 0$ such that for $z \sim (0, \sigma^2 I)$, the inequality $\mathbb{E}[\|z\|_2] \le t$ holds, and the event $\mathcal{E}_t := \{\|z\|_2 \le t\}$ holds with probability at least $1/2$. It is clear that the choice of σ depends on that of t, and as $t \to 0$, we must have $\sigma \to 0$.

The first step is to define a vector field that satisfies the conditions of Lemma 3. For arbitrary $\delta \in [0, 1]$, we define a vector field ϕ_δ such that:

$$\phi_\delta(x) := \frac{1}{3}(\langle x, x^* \rangle x - \|x\|_2^2 x^*) + \frac{\delta}{3}(\|x\|_2^2 - 5/8) x, \tag{98}$$

and make the following claim.

Claim 5 *For any $\delta \in (0, 1]$, we can find a constant $\epsilon_0 > 0$ such that $\|\phi_\delta(x)\|_2 \le 1$ and $x - \epsilon\phi_\delta(x) \in K$ holds for arbitrary $x \in K$ and $\epsilon \in (0, \epsilon_0]$.*

The claim shows that ϕ_δ satisfies the conditions of Lemma 3 for any $\delta \in (0, 1]$, so that given a scalar $\xi > 0$, the lemma implies

$$\mathcal{C}_{\xi \tilde{F}_\sigma(x)}(K \backslash U) \ge \inf_{x \in K \backslash U} \left\{ \xi \langle \phi_\delta(x), \nabla \tilde{F}_\sigma(x) \rangle - \operatorname{div} \phi_\delta(x) \right\}.$$

The right-hand side is uniformly continuous in δ, so that if we take the limit $\delta \to 0$, we obtain the lower bound:

$$\mathcal{C}_{\xi \tilde{F}_\sigma(x)}(K \backslash U) \ge \inf_{x \in K \backslash U} \left\{ \xi \langle \phi_0(x), \nabla \tilde{F}_\sigma(x) \rangle - \operatorname{div} \phi_0(x) \right\}. \tag{99}$$

It remains to lower bound the right-hand side of inequality (99). Recall that $\tilde{F}_\sigma(x) = \mathbb{E}[F(x+z)]$. The definition of the gradient of $\tilde{F}_\sigma$ implies

$$\langle \phi_0(x), \nabla \tilde{F}_\sigma(x) \rangle = \lim_{\epsilon \to 0} \frac{\tilde{F}_\sigma(x + \epsilon\phi_0(x)) - \tilde{F}_\sigma(x)}{\epsilon} = \lim_{\epsilon \to 0} \mathbb{E}\left[\frac{F(x + z + \epsilon\phi_0(x)) - F(x+z)}{\epsilon} \right]$$

For the right-hand side, we prove lower bound for it using the Massart noise model. We start by simplifying the fraction term inside the expectation. Without loss of generality, assume that $\epsilon \in (0, 0.2]$, then the definition of ϕ_0 implies:

$$
\begin{aligned}
F(x + z + \epsilon\phi_0(x)) &= F\left((1 + \frac{\epsilon}{3}\langle x, x^* \rangle)x + z - \frac{\epsilon}{3}\|x\|_2^2 x^* \right) \\
&\overset{(i)}{=} F\left(x + \frac{z}{1 + \frac{\epsilon}{3}\langle x, x^* \rangle} - \frac{\frac{\epsilon}{3}\|x\|_2^2 x^*}{1 + \frac{\epsilon}{3}\langle x, x^* \rangle} \right) \\
&\overset{(ii)}{\ge} F\left(x + z - \frac{\epsilon\|x\|_2^2}{3} x^* \right) - 2\epsilon\|z\|_2 - \frac{2\epsilon^2}{3},
\end{aligned}
$$

where equation (i) uses $F(x) = F(\alpha x)$ for any $\alpha > 0$. To derive inequality (ii), we used the fact that $1/(1 + \frac{\epsilon}{3}\langle x, x^* \rangle) \in [1 - \frac{2\epsilon}{3}, 1 + \frac{2\epsilon}{3}]$ for any $x \in K$, $\epsilon \in (0, 0.2]$, and the property that F is 3-Lipschitz continuous. Combining the two equations above, and using the fact $\mathbb{E}[\|z\|_2] \leq t$, we obtain:

$$\langle \phi_0(x), \nabla \tilde{F}_\sigma(x) \rangle \geq \lim_{\epsilon \to 0} \mathbb{E}\Big[\frac{F(x + z - \frac{\epsilon\|x\|_2^2}{3}x^*) - F(x + z)}{\epsilon}\Big] - 2t$$

$$= \lim_{\epsilon \to 0} \mathbb{E}\Big[\frac{F(\frac{3(x+z)}{\|x\|_2^2} - \epsilon x^*) - F(\frac{3(x+z)}{\|x\|_2^2})}{\epsilon}\Big] - 2t. \tag{100}$$

We further simplify the lower bound (100) by the following claim, which is proved using properties of the Massart noise.

Claim 6 *For any $x \in \mathbb{R}^d$ and any $\epsilon > 0$, we have $F(x - \epsilon x^*) - F(x) \geq 0$. Moreover, for any $x \in \mathbb{R}^d : \|x\|_2 \geq 1$, let α be the angle between x and x^*, then we have:*

$$\frac{F(x - \epsilon x^*) - F(x)}{\epsilon} \geq \frac{3q_0|\sin(\alpha)|}{5\pi\|x\|_2}\Big(\frac{|\sin(\alpha)|\sqrt{1 - \epsilon^2}}{2\sqrt{d}} - \epsilon\Big).$$

When the event $\mathcal{E}_t$ holds, we have $\|z\|_2 \leq t \leq 1/6 \leq \|x\|_2/3$, so that $\|\frac{3(x+z)}{\|x\|_2^2}\|_2 \in [2, 4]$. Combining with inequality (100) and Claim 6, we have

$$\langle \phi_0(x), \nabla \tilde{F}_\sigma(x) \rangle \geq P(\mathcal{E}_t) \lim_{\epsilon \to 0} \mathbb{E}\Big[\frac{F(\frac{3(x+z)}{\|x\|_2^2} - \epsilon x^*) - F(\frac{3(x+z)}{\|x\|_2^2})}{\epsilon} \mid \mathcal{E}_t\Big] - 2t$$

$$\geq \frac{1}{2} \times \lim_{\epsilon \to 0} \mathbb{E}\Big[\frac{3q_0|\sin(\alpha_{x+z})|}{5\pi \times 4}\Big(\frac{|\sin(\alpha_{x+z})|\sqrt{1 - \epsilon^2}}{2\sqrt{d}} - \epsilon\Big) \mid \mathcal{E}_t\Big] - 2t$$

$$= \frac{3q_0\mathbb{E}[\sin^2(\alpha_{x+z}) \mid \mathcal{E}_t]}{80\pi\sqrt{d}}, \tag{101}$$

where α_{x+z} represents the angle between $x + z$ and x^*.

In order to lower bound the divergence term $\operatorname{div} \phi_0(x)$, let $H(x) \in \mathbb{R}^{d \times d}$ be the Jacobian matrix of ϕ_0 at point x (i.e. $H_{ij} := \frac{\partial(\phi_0(x))_i}{\partial x_j}$), then we have

$$H(x) = \frac{1}{3}\Big(\langle x, x^* \rangle I + x(x^*)^\top - 2x^* x^\top\Big).$$

It means that $\operatorname{div} \phi_0(x) = \operatorname{tr}(H(x)) = \frac{d-1}{3}\langle x, x^* \rangle = \frac{(d-1)\|x\|_2}{3}\cos(\alpha_x)$. Combining this equation with inequalities (99), (101), and taking the limits $t \to 0$, $\sigma \to 0$, we obtain:

$$\mathcal{C}_{\xi F(x)}(K \backslash U) \geq \inf_{x \in K \backslash U} \Big\{ \underbrace{\frac{3\xi q_0 \sin^2(\alpha_x)}{80\pi\sqrt{d}} - \frac{(d-1)\|x\|_2}{3}\cos(\alpha_x)}_{:=L(x)} \Big\}. \tag{102}$$

where α_x represents the angle between x and x^*.

According to inequality (102), for any $\alpha_x \in (\pi - \alpha_0, \pi]$, we have

$$L(x) \geq -\frac{(d-1)\|x\|_2}{3}\cos(\alpha_x) \geq \frac{(d-1)\|x\|_2}{3}\cos(\alpha_0) \geq \frac{(d-1)}{6\sqrt{2}},$$

where the last inequality follows since $\|x\|_2 \geq 1/2$ and $\alpha_0 \in (0, \pi/4]$. Otherwise, if $\alpha_x \in [\alpha_0, \pi - \alpha_0]$, then we have

$$L(x) \geq \frac{3\xi q_0 \sin^2(\alpha_0)}{80\pi\sqrt{d}} - \frac{d-1}{3}.$$

Once we choose $\xi \geq \frac{160\pi}{3} \frac{d^{3/2}}{q_0 \sin^2(\alpha_0)}$, the above expression will be lower bounded by $d/3$, which completes the proof.

Proof of Claim 5 Since $\|x\|_2 \leq 1$ for any $x \in K$, it is easy to verify that $\|\phi_\delta(x)\|_2 \leq 1$. In order to verify $x - \epsilon\phi_\delta(x) \in K$, we notice that

$$\|x - \epsilon\phi_\delta(x)\|_2^2 = \|x\|_2^2 + \epsilon^2\|\phi_\delta(x)\|_2^2 - \frac{2\epsilon\delta}{3}(\|x\|_2^2 - 5/8)\|x\|_2^2.$$

As a consequence, we have

$$\left| \|x - \epsilon\phi_\delta(x)\|_2^2 - 5/8 \right| \leq \left| \|x\|_2^2 - 5/8 \right| \left(1 - \frac{2\epsilon\delta}{3}\|x\|_2^2 \right) + \epsilon^2$$

The right-hand side will be maximized if $\|x\|_2^2 = 1/4$. Thus, if we assume $\delta > 0$, then for any $\epsilon < \delta/16$, it is easy to verify that the right-hand side is bounded by $3/8$. As a consequence, we have $\|x - \epsilon\phi_\delta(x)\|_2^2 \in [1/4, 1]$, which verifies that $x - \epsilon\phi_\delta(x) \in K$.

Proof of Claim 6 When $x = 0$ or $x - \epsilon x^* = 0$, it is easy to verify that $F(x - \epsilon x^*) - F(x) \geq 0$ by the definition of the loss function. Otherwise, we assume that $x \neq 0$ and $x - \epsilon x^* \neq 0$. In these cases, the events $\langle x, a \rangle \neq 0$ and $\langle x - \epsilon x^*, a \rangle \neq 0$ hold almost surely, so that we can assume that the loss function always takes zero-one values.

When the parameter changes from x to $x - \epsilon x^*$, the value of $\ell(x; a)$ and $\ell(x - \epsilon x^*; a)$ are non-equal if and only if $\text{sign}(\langle x, a \rangle) \neq \text{sign}(\langle x - \epsilon x^*, a \rangle)$. This condition is equivalent of

$$\text{sign}(\langle x, a \rangle) = \text{sign}(\langle x^*, a \rangle) \quad \text{and} \quad |\langle x, a \rangle| < \epsilon|\langle x^*, a \rangle|. \tag{103}$$

Let $\mathcal{E}$ be the event that condition (103) holds. Under this event, when the parameter changes from x to $x - \epsilon x^*$, the loss changes from $\frac{1-q(a)}{2}$ to $\frac{1+q(a)}{2}$. It means that the loss is non-decreasing with respect to the change $x \to x - \epsilon x^*$, and as a consequence, we have $F(x - \epsilon x^*) - F(x) \geq 0$.

In order to establish the lower bound in Claim 6, we first lower bound the probability of event $\mathcal{E}$. In the proof of Lemma 18. we have shown that this probability is equal to:

$$P(\mathcal{E}) = \frac{1}{\pi} \arccos \left(\frac{\langle x, x - \epsilon x^* \rangle}{\|x\|_2 \|x + \epsilon x^*\|_2} \right)$$

Let β be the angle between x and $x - \epsilon x^*$, then the right-hand side is equal to β/π. Using the geometric property that $\frac{\|x\|_2}{|\sin(\alpha)|} = \frac{\epsilon\|x^*\|_2}{|\sin(\beta)|}$, we have

$$P(\mathcal{E}) = \frac{\beta}{\pi} \geq \frac{|\sin\beta|}{\pi} = \frac{\epsilon|\sin(\alpha)|}{\|x\|_2\pi} \geq \frac{\epsilon|\sin(\alpha)|}{\|x\|_2\pi}. \tag{104}$$

Conditioning on the event $\mathcal{E}$, when the parameter moves $x \to x - \epsilon x^*$, the loss $\ell(x; a)$ changes by amount $q(a)$. Since the Massart noise forces $q(a) \geq q_0|\langle x^*, a \rangle|$, we can lower bound the gap $F(x - \epsilon x^*) - F(x)$ by lower bounding the expectation $\mathbb{E}[|\langle x^*, a \rangle| \,\big|\, \mathcal{E}]$. We decompose the vector a into two components: the component a_1 that is parallel to x and the component a_2 that is orthogonal to x. Similarly, we can decompose the vector x^* into two components x_1^* and x_2^*, parallel to and

orthogonal to the vector x respectively. The decomposition implies

$$\langle x^*, a \rangle = \langle x_1^*, a_1 \rangle + \langle x_2^*, a_2 \rangle.$$

For the first term on the right-hand side, we have $|\langle x_1^*, a_1 \rangle| \leq \|a_1\|_2 = \frac{|\langle x, a \rangle|}{\|x\|_2} \leq \frac{\epsilon}{\|x\|_2} \leq \epsilon$ by condition (103) and the assumption that $\|x\|_2 \geq 1$. For the second term, if we condition on a_1, then the vector a_2 is uniformly sampled from a $(d-1)$-dimensional sphere of radius $\sqrt{1 - \|a_1\|_2^2}$ that centers at the origin. The vector x_2^*, constructed to be orthogonal to x, also belongs to the same $(d-1)$-dimensional subspace. Under this setting, Awasthi et al. (2015, Lemma 4) proved that

$$P\left(|\langle x_2^*, a_2 \rangle| > \frac{\|x_2^*\|_2 \sqrt{1 - \|a_1\|_2^2}}{2\sqrt{d}} \,\Big|\, \mathcal{E}, a_1 \right) \geq 1 - \sqrt{\frac{1}{2\pi}} \geq 3/5.$$

Using the bound $\|a_1\|_2 = \frac{|\langle x, a \rangle|}{\|x\|_2} \leq \frac{\epsilon}{\|x\|_2} \leq \epsilon$, we marginalize a_1 to obtain

$$P\left(|\langle x_2^*, a_2 \rangle| > \frac{\|x_2^*\|_2 \sqrt{1 - \epsilon^2}}{2\sqrt{d}} \,\Big|\, \mathcal{E} \right) \geq 3/5.$$

Recall that $\langle x^*, a \rangle = \langle x_1^*, a_1 \rangle + \langle x_2^*, a_2 \rangle$ and $|\langle x_1^*, a_1 \rangle| \leq \epsilon$. These two relations imply $|\langle x^*, a \rangle| \geq |\langle x_2^*, a_2 \rangle| - |\langle x_1^*, a_1 \rangle| \geq |\langle x_2^*, a_2 \rangle| - \epsilon$. Combining it with the relation $\|x_2^*\|_2 = |\sin(\alpha)|$, we obtain

$$P\left(|\langle x^*, a \rangle| \geq \frac{|\sin(\alpha)| \sqrt{1 - \epsilon^2}}{2\sqrt{d}} - \epsilon \,\Big|\, \mathcal{E} \right) \geq 3/5.$$

As a consequence, we have

$$\mathbb{E}[|\langle x^*, a \rangle| \,|\, \mathcal{E}] \geq \frac{3}{5}\left(\frac{|\sin(\alpha)| \sqrt{1 - \epsilon^2}}{2\sqrt{d}} - \epsilon \right). \tag{105}$$

Combining inequalities (104), (105) with the relation that $q(a) \geq q_0 |\langle x^*, a \rangle|$, we obtain

$$F(x - \epsilon x^*) - F(x) \geq P(\mathcal{E}) \times q_0 \, \mathbb{E}[|\langle x^*, a \rangle| \,|\, \mathcal{E}] \geq \frac{\epsilon |\sin(\alpha)|}{\|x\|_2 \pi} \times q_0 \frac{3}{5}\left(\frac{|\sin(\alpha)| \sqrt{1 - \epsilon^2}}{2\sqrt{d}} - \epsilon \right).$$

which completes the proof.