Convex Optimization with Unbounded Nonconvex Oracles using Simulated Annealing

Oren Mangoubi
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Nisheeth K. Vishnoi
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We consider the problem of minimizing a convex objective function F when one can only evaluate its noisy approximation $\hat{F}$. Unless one assumes some structure on the noise, $\hat{F}$ may be an arbitrary nonconvex function, making the task of minimizing F intractable. To overcome this, prior work has often focused on the case when $F(x)-\hat{F}(x)$ is uniformly-bounded. In this paper we study the more general case when the noise has magnitude $\alpha F(x)+\beta$ for some $\alpha, \beta > 0$, and present a polynomial time algorithm that finds an approximate minimizer of F for this noise model. Previously, Markov chains, such as the stochastic gradient Langevin dynamics, have been used to arrive at approximate solutions to these optimization problems. However, for the noise model considered in this paper, no single temperature allows such a Markov chain to both mix quickly and concentrate near the global minimizer. We bypass this by combining "simulated annealing" with the stochastic gradient Langevin dynamics, and gradually decreasing the temperature of the chain in order to approach the global minimizer. As a corollary one can approximately minimize a nonconvex function that is close to a convex function; however, the closeness can deteriorate as one moves away from the optimum.

Keywords: Nonconvex optimization, Stochastic gradient Langevin dynamics, Simnulated annealing

1. Introduction

A general problem that arises in machine learning, computational mathematics and optimization is that of minimizing a convex objective function $F : \mathcal{K} \to \mathbb{R}$, where $\mathcal{K} \subseteq \mathbb{R}^d$ is convex, and one can only evaluate F *approximately*. Let $\hat{F}$ denote this "noisy" approximation to F. In this setting, even though the function F is convex, we can no longer assume that $\hat{F}$ is convex. However, if one does not make any assumption on the noise function, the problem of minimizing F can be shown to be arbitrarily hard. Thus, having some restrictions on the noise function is necessary.

A well studied setting is that of "additively" bounded noise (Applegate and Kannan, 1991; Singer and Vondrák, 2015; Risteski and Li, 2016; Zhang et al., 2017). Here, the noise $N(x) := F(x) - \hat{F}(x)$ is assumed to have a uniform bound on $\mathcal{K}$: $\sup_{x\in\mathcal{K}} |N(x)| \leq \beta$ for some $\beta \geq 0$. In practice, however, the strongest bound we might have for the noise may not be uniform on $\mathcal{K}$. One such noise model is that of "multiplicative" noise where one assumes that $|\hat{F}(x) - F(x)| \leq \alpha F(x)$, for all $x \in \mathcal{K}$ and some $\alpha \geq 0$. In other words, $|N(x)| = |\hat{F}(x) - F(x)| = |F(x)| \times \alpha$, which motivates the name. One situation where multiplicative noise arises is when F decomposes into a sum of functions that are easier to compute, but these component functions are computed via Monte Carlo integration and the stopping criteria of these integration methods depend on the computed

value of the component function (Chen, 2015). For other natural settings where multiplicative noise arises see (Chen et al., 2015; Jebalia and Auger, 2008; Jebalia et al., 2011).

More generally, one can model the noisy function $\hat{F}$ by decomposing it into additive and multiplicative components, in the following sense:

Definition 1 *We say that $\hat{F}$ has additive and multiplicative noise levels (β, α) if*

$$|\hat{F}(x) - F(x)| \leq \alpha(F(x) - F(x^*)) + \beta \tag{1}$$

Note that this noise model has the natural property that the noise level does not change if we replace $F(x)$ and $\hat{F}(x)$ with a new objective function $F(x) + C$ and a new oracle $\hat{F}(x) + C$ for the same number $C > 0$. To motivate this definition, we consider a situation where the error in computing the objective function depends on the amount of time spent on the computation. For instance, to compute the objective function one may be required to solve a complicated system of partial differential equations, where a finer discretization leads to a greater accuracy but also to a longer computation time (Conrad et al., 2018; Cliffe et al., 2011). In this case, one can start by using a short computation time for each evaluation and gradually increase the computation time as one approaches the minimum value. Whereas a purely additive noise model would require one to have a uniform computational cost at each step, the multiplicative noise model allows one to analyze methods where one has the flexibility to use a different cost at each step.

As another application, we consider the problem of solving a system of noisy linear or nonlinear black-box equations where one wishes to find a value of x such that $h_i(x) = 0$ for each component function h_i (Chen et al., 2015). Since each equation $h_i(x) = 0$ must be satisfied simultaneously for a single value of x, it is not enough to solve each equation individually. One way in which we may solve this system of equations is by minimizing an objective function of the form $F(x) = \frac{1}{n}\sum_{i=1}^{n}(h_i(x))^2$ since any value of x that minimizes $F(x)$ also solves the system of equations $h_i(x) = 0$ for every i, provided that such a solution exists. While it is true that one may instead minimize the objective function $\frac{1}{n}\sum_{i=1}^{n}|h_i(x)|$ to solve the same system of equations, it is oftentimes preferable to use the quadratic objective function $F(x) = \frac{1}{n}\sum_{i=1}^{n}(h_i(x))^2$ since it is much smoother and can lead to faster convergence in practice (Chen et al., 2015). Rather than having access to an exact computation oracle for $h_i(x)$ one may instead only have access to a perturbed function $\hat{h}_i(x) = h_i(x) + N_i(x)$. Here $N_i(x)$ is a noise term that may have additive or multiplicative noise (or both), that is, $|N_i(x)| \leq b + ah_i(x)$ for some $a, b \geq 0$. Hence, instead of minimizing the objective function F, one must try to minimize a noisy function of the form $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}(\hat{h}_i(x))^2$. A straightforward calculation shows that the fact that $|N_i(x)| \leq b + ah_i(x)$ for all i implies that

$$|\hat{F}(x) - F(x)| \leq (2a + a^2 + 2b + 2ab)(F(x) - F(x^*)) + \frac{1}{2}(b + ab) + b^2,$$

where $F(x^*) = 0$. Thus, $\hat{F}$ can be modeled as having additive noise level $\beta = \frac{1}{2}(b + ab) + b^2$ together with multiplicative noise level $\alpha = 2a + a^2 + 2b + 2ab$. In particular, even if each component function only has additive noise (that is, if $a = 0$), $\hat{F}$ will still have nonzero multiplicative noise $\alpha = 2b$. Thus we arrive at the following general problem.

Problem 1 *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex body and $F : \mathcal{K} \to \mathbb{R}$ be a convex function, where $F(x^*) = 0$ and $x^\star$ is a minimizer of F in $\mathcal{K}$. Given access to a noisy oracle $\hat{F}$ for F that has additive and multiplicative noise levels (β, α). The problem is to find an approximate minimizer $\hat{x}$ for F such that $F(\hat{x}) \leq \hat{\varepsilon}$ for a given $\hat{\varepsilon} > 0$.*

One of the first papers to study this problem was by Applegate and Kannan (1991) in the special case of additive noise (where $\alpha = 0$). Specifically, they studied the related problem of sampling from the canonical distribution $\frac{1}{\int_{\mathcal{K}} e^{-\xi \hat{F}(y)} \mathrm{d}y} e^{-\xi \hat{F}(x)}$ when $\hat{F}$ is an additively noisy version of a convex function. Roughly, their algorithm discretized $\mathcal{K}$ with a grid and ran a simple random walk on this grid. Using their Markov chain one can solve Problem 1 in the special case of $\alpha = 0$ for some error $\hat{\varepsilon} = \tilde{O}(d\beta)$ with running time that is polynomial in d and various other parameters as well.

In Belloni et al. (2015), Problem 1 was studied in a special case where the noise decreases to zero near the global minimum[1] and F is $\mathfrak{m}$-strongly convex. Specifically, they study the situation where the noise is bounded by $|N(x)| \leq c\|x - x^\star\|^p$, for some $0 < p < 2$ and some $c > 0$. Roughly speaking, in this regime they show that one can obtain an approximate minimizer $\hat{x}$ such that $F(\hat{x}) - F(x^\star) \leq O((\frac{d}{\mathfrak{m}})^{\frac{1}{2-p}})$ in polynomial time. To find an approximate minimizer $\hat{x}$, they repeatedly run a simulated annealing Markov chain based on the "hit-and-run" algorithm. They state that they are "not aware of optimization methods for such a problem" outside of their work, and that "it is rather surprising that one may obtain provable guarantees through simulated annealing" under noise with non-uniform bounds even in the special case of strong convexity.

Problem 1 was also studied by Zhang et al. (2017) in the special case of additive noise (where $\alpha = 0$ but $\beta \geq 0$). The main component of their algorithm is the stochastic gradient Langevin dynamics (SGLD) Markov chain that runs at a fixed "temperature" parameter to find an approximate minimizer $\hat{x}$. In particular, they show that one can solve Problem 1 in the special case of $\alpha = 0$ for some error $\hat{\varepsilon} = \tilde{O}(d\beta)$ with running time that is polynomial in d and β and various smoothness parameters. Other related works that have studied various aspects of optimization under additive noise include (Singer and Vondrák, 2015; Hazan et al., 2016; Risteski and Li, 2016).

The difficulty of extending these results to the general case when both $\alpha, \beta > 0$, and F is not necessarily strongly convex arises from the fact that, in this setting, the noise can become unbounded and the prior Markov chain approaches do not seem to work. Roughly, the Markov chains of (Applegate and Kannan, 1991; Zhang et al., 2017) run at a fixed temperature and, due to the fact that the noise can be very different at different levels of F, would either get stuck in a local minimum or never come close to the minimzer; see Figure 2 for an illustration. The Markov chain of Belloni et al. (2015) on the other hand varies the temperature but the strong convexity of F makes the task of estimating progress significantly simpler.

1.1. Our contributions

The main result of this paper is the first polynomial time algorithm that solves Problem 1 when $\alpha, \beta > 0$ without assuming that F is strongly convex. Our algorithm combines simulated annealing (as in Belloni et al. (2015)) with the stochastic gradient Langevin dynamics (as in Zhang et al. (2017)). We assume that $\|\nabla F\| \leq \lambda$ and that $\mathcal{K}$ is contained in a bounding ball of radius $R > 0$, and that $\mathcal{K} = \mathcal{K}' + B(0, r')$ for some $r' > 0$, where "+" denotes the Minkowski sum. Note that, given bounds λ and R, one can deduce an upper bound of λR on the value of F in $\mathcal{K}$. Also note that while the Lipschitz gradient assumption helps us prove running time bounds for our algorithm, it is likely not needed to solve the problem.

1. Belloni et al. (2015) also study separately the special case of purely additive noise, but not simultaneously in the presence of a non-uniformly bounded noise component.

Theorem 2 [Informal; see Section B.2 for a formal description] *For any desired accuracy level $\hat{\varepsilon}$, additive noise level $\beta = O(\hat{\varepsilon})$, and a multiplicative noise level α that is a sufficiently small constant, there exists an algorithm that solves Problem 1 and outputs $\hat{x}$ with high probability such that $F(\hat{x}) - F(x^\star) \leq \hat{\varepsilon}$. The running time of the algorithm is polynomial in d, R, $1/r'$, and λ, whenever $\alpha \leq \tilde{O}(\frac{1}{d})$ and $\beta \leq \tilde{O}(\frac{\hat{\varepsilon}}{d})$.*

When the multiplicative noise coefficient satisfies $\alpha \leq \tilde{O}(\frac{1}{d})$, Theorem 2 guarantees that one can obtain an approximate minimizer $\hat{x}$ such that $F(\hat{x}) - F(x^\star) \leq \hat{\varepsilon}$ for arbitrarily small $\hat{\varepsilon}$ in polynomial time. Also note that related work (Applegate and Kannan, 1991) for additive noise does not require a Lipschitz gradient or a bound on the diameter of $\mathcal{K}$, although they still require a bound on the range of the objective function.

The requirement that $\beta \leq \tilde{O}(\frac{\hat{\varepsilon}}{d})$ in order to get a polynomial running time can be shown to be necessary using results from the work of Blum and Rivest (1989) (as done by Zhang et al. (2017)). If the additive noise β was required to be any lower than $\Omega(\frac{\hat{\varepsilon}}{d})$, the algorithm would take an exponentially long time to escape the local minima (Figure 1). We believe that the requirement that $\alpha \leq \tilde{O}(\frac{1}{d})$ in order to get a polynomial running time is also tight for a similar reason. This is because a sub-level set U of F of height $\hat{\varepsilon}$, i.e., $U = \{x \in \mathcal{K} : F(x) \leq \hat{\varepsilon}\}$, will have a uniform bound on the noise of size $\sup_{x \in U} \alpha F(x) \leq \alpha\hat{\varepsilon}$ in the presence of multiplicative noise level α. This is equivalent to having additive noise level $\tilde{O}(\frac{\hat{\varepsilon}}{d})$, which is required for the Markov chain to quickly escape the local minima of that sub-level set. Establishing this formally is an interesting open problem. While our algorithm's running time is polynomial in various parameters, we believe that it is not tight and can be improved with a more thorough analysis of the underlying Markov chain. The results of Zhang et al. (2017) for the additive noise is more general; their algorithm works for a class of nonconvex functions F with a certain saddle-point property. It would therefore be interesting to see if we can solve Problem 1 for this class of nonconvex functions F but under the more general noise model where we have both additive and multiplicative noise. We note the following obvious but important corollary of our main result for nonconvex functions: Suppose we are given oracle access to a nonconvex function $\hat{F}$ with a guarantee that there is a convex function F such that $|\hat{F}(x) - F(x)| \leq \alpha(F(x) - F(x^\star)) + \beta$ (as in Definition 1), then there is an algorithm to minimize $\hat{F}$.

1.2. On the assumption that $F(x^\star) = 0$.

Suppose that we are given a function $\mathcal{F}$ with noisy oracle $\hat{\mathcal{F}}$ with additive and multiplicative noise level α, β, but $\mathcal{F}(x^*) \neq 0$. Then, if we know the value of $\mathfrak{m} = \mathcal{F}(x^*)$, we can put this function in the form of Problem 1 by defining a "shifted" objective function $F(x) := \mathcal{F}(x) - \mathcal{F}(x^*)$ and "shifted" oracle $\hat{F}(x) := \hat{\mathcal{F}}(x) - \mathfrak{m}$. In practice, we do not know the minimizing value $\mathfrak{m}$, but we can still obtain a noisy oracle $\hat{F}$ for F by guessing a value for $\mathfrak{m}'$ and setting $\hat{F}'(x) = \hat{\mathcal{F}}(x) - \mathfrak{m}'$, although $\hat{F}'$ will have a larger additive noise level $|\mathfrak{m}' - \mathcal{F}(x^*)| + \beta$ depending on the accuracy $|\mathfrak{m}' - \mathcal{F}(x^*)|$ of our guess. In practice, if we know β, then we can get around this problem by performing a binary search, by repeatedly running our algorithm using a sequence of noisy oracles $\hat{F}_1, \hat{F}_2, \ldots$ obtained with different guesses $\mathfrak{m}'_1, \mathfrak{m}'_2, \ldots$. If we make a guess $\mathfrak{m}'_j$ and our algorithm returns a value $\hat{\mathcal{F}}(\hat{x}) \leq \mathfrak{m}'_j + \beta$, then our next guess $\mathfrak{m}'_{j+1}$ should be lower; otherwise it should be higher. The number of times we must run our algorithm is therefore only logarithmic in the desired accuracy $\hat{\varepsilon}^{-1}$. If we do not know β, then the number of times we must run our algorithm will instead be polynomial in $\hat{\varepsilon}^{-1}$, λ and R.

Since our framework (Problem 1) assumes $F(x^\star) = 0$, the value of $\hat{F}(x)$ gives us a good estimate for the amount of (multiplicative) noise near a point x. Therefore, $\hat{F}$ can help us choose the temperature parameter at each step in our algorithm: a larger value of $\hat{F}$ means that there may be more noise present and we require a higher temperature, while a lower value of $\hat{F}$ means that we can lower temperature. (See Section 1.2 for how our framework can be generalized to $F(x^*) \neq 0$).

1.3. Short summary of techniques

To find an approximate global minimum of the objective function F, we must try to find an approximate global minimum of the noisy approximation $\hat{F}$. One method of optimizing a nonconvex or approximately convex function $\hat{F}$ is to generate a Markov chain with stationary distribution approximating the canonical distribution $\hat{\pi}^{(\xi)}(x) := \frac{1}{\int_{\mathcal{K}} e^{-\xi \hat{F}(y)} \mathrm{d}y} e^{-\xi \hat{F}(x)}$, where ξ is thought of as an "inverse temperature" parameter. If the "temperature" ξ^{-1} is small, then $\hat{\pi}^{(\xi)}$ concentrates near the global minima of $\hat{F}$. On the other hand, to escape local minima of "depth" $\beta > 0$ in polynomial time, one requires the temperature ξ^{-1} to be at least $\Omega(\beta)$ (see Figure 1). Now consider the random variable $Z \sim N(0, \xi^{-1} I_d)$ with $\pi^{(\xi)}(x) := \frac{1}{\int_{\mathbb{R}^d} e^{-\xi \frac{1}{2}\|y\|^2} \mathrm{d}y} e^{-\xi \frac{1}{2}\|x\|^2}$. Then $F(Z)$ concentrates near $d\xi^{-1}$ with high probability. This suggests that for a noisy function $\hat{F}$ where we are given a bound on the additive noise level $\beta > 0$, the best we can hope to achieve in polynomial time is to find a point $\hat{x}$ such that $|F(\hat{x}) - F(x^\star)| \leq \tilde{O}(d\beta)$, since there may be sub-optimal local minima in the vicinity of $x^\star$ that have depth $O(\beta)$, requiring the temperature ξ^{-1} to be at least $\Omega(\beta)$ (Figure 1).

As mentioned earlier, optimization of a noisy function under additive noise is studied by Zhang et al. (2017), who analyze the stochastic gradient Langevin dynamics (SGLD) Markov chain. The SGLD chain approximates the Langevin diffusion, which has stationary distribution $\hat{\pi}^{(\xi)}$. They show that by running SGLD at a single fixed temperature ξ one can obtain an approximate global minimizer $\hat{x}$ of F such that $|F(\hat{x}) - F(x^\star)| < \tilde{O}(\hat{\varepsilon})$ with high probability with running time that is polynomial in d, $e^{d\beta/\hat{\varepsilon}}$, and various smoothness bounds on F. In particular, for the algorithm to get a polynomial running time in d and β one must choose $\hat{\varepsilon} = \Omega(d\beta)$. Thus, the SGLD algorithm returns an approximate minimizer such that $|F(\hat{x}) - F(x^\star)| \leq \tilde{O}(d\beta)$ in polynomial time in the additive case.

More generally, if multiplicative noise is present one may have many local minima of very different sizes, so our bound on the "depth" of the local minima is not uniform over $\mathcal{K}$. In this case the approach by Zhang et al. (2017) of using a single fixed temperature will lead to either a very long running time or a very large error $\hat{\varepsilon}$: If the temperature is hot enough to escape even the deepest the local minima, then the Markov chain will not concentrate near the global minimum and the error $\hat{\varepsilon}$ will be large (Figure 2(b)). If the temperature is chosen to be too cold, then the algorithm will take a very long time to escape the deeper local minima (Figure 2(c)). Instead of using a fixed temperature, we search for the global minimum by starting the Markov chain at a high temperature and then slowly lowering the temperature at each successive step of the chain (Figure 2(d)). This approach is referred to as "simulated annealing" in the literature (Kirkpatrick et al., 1983).

The only non-asymptotic analysis we are aware of where the bound on the noise is not uniform involves a simulated annealing technique based on the hit-and-run algorithm (Belloni et al., 2015). Specifically, Belloni et al. (2015) show that if F is $\mathfrak{m}$-strongly convex, then one can compute an approximate global minimizer $\hat{x}$ such that $|F(\hat{x}) - F(x^\star)| < (\frac{d}{\mathfrak{m}})^{\frac{1}{2-p}}$ with running time $\tilde{O}(d^{4.5})$, as long as $N(x) \leq c\|x\|^p$ for some $0 < p < 2$ and some $c > 0$. The algorithm used by Belloni

et al. (2015) runs a sequence of subroutine Markov chains. Each of these subroutine Markov chains is restricted to a ball $B(y_k, r_k)$ centered at the point y_k returned by the subroutine chain from the last epoch. Crucially, for this algorithm to work, r_k must be chosen such that $B(y_k, r_k)$ contains the minimizer $x^\star$ at each epoch k. Towards this end, Belloni et al. (2015) show that since the temperature is decreased at each epoch, $F(y_k)$ is much smaller than $F(y_{k-1})$ at each epoch k. Since F is assumed to be strongly convex, Belloni et al. (2015) show that this decrease in F implies a contraction in the distance $\|y_k - x^\star\|$ at each epoch k, allowing one to choose a sequence of radii r_k that contract as well at each step but still have the property that $x^\star \in B(y_k, r_k)$.

One obstacle in generalizing the results by Belloni et al. (2015) to the non-strongly convex case is that we do not have an oracle for the sub-level sets of F, but only for $\hat{F}$, whose sub-level sets may not even be connected. Instead, we show that the SGLD Markov chain concentrates inside increasingly smaller sub-level sets of F as the temperature parameter is decreased. To analyze the behavior of the SGLD Markov chain at each temperature, we build several new tools and use some from the past work. Our results make important contributions to the growing body of work on non-asymptotic analysis of simulated annealing, Langevin dynamics and their various combinations (Raginsky et al., 2017; Bubeck et al., 2015; Welling and Teh, 2011; Lee and Vempala, 2017).

1.4. Organization of the rest of the paper

In the main body of the paper we present a detailed but informal primer of the algorithm followed by the key steps and ideas involved in the proof of Theorem 2 in Section 2. The precise description of the algorithm and the full proofs are quite technical and have been moved to the appendix due to space constraints. In the Appendix, we present the notation and other preliminaries in Section A. This is followed by a formal presentation of the algorithm and the statement of the main results in Sections B.1 and B.2. Section C contains the detailed mathematical proof of our main theorem.

2. Overview of Our Contributions

The model and the problem. Let $\mathcal{K} \subseteq \mathbb{R}^d$ be the given convex body contained in a ball of radius $R > 0$ and $F : \mathcal{K} \to \mathbb{R}$ the given convex function. We assume that F has gradient bounded by some number $\lambda > 0$, and that $\mathcal{K} = \mathcal{K}' + B(0, r')$ for some $r' > 0$, where "+" denotes the Minkowski sum and $\mathcal{K}'$ is a convex body. Let $x^\star$ be a minimizer of F over $\mathcal{K}$. Recall that our goal is to find an approximate minimizer $\hat{x}$ of F on $\mathcal{K}$, such that $F(\hat{x}) - F(x^\star) \leq \hat{\varepsilon}$ for a given $\hat{\varepsilon} > 0$.

We assume that we have access to a membership oracle for $\mathcal{K}$ and a noisy oracle $\hat{F}$ for F. Recall that in our model of noise, since $F(x^*) = 0$, we may assume that there exist functions $\varphi : \mathcal{K} \to \mathbb{R}$ and $\psi : \mathcal{K} \to \mathbb{R}$ and numbers $\alpha, \beta \geq 0$, with $|\varphi(x)| \leq \beta$ and $|\psi(x)| \leq \alpha$ for all $x \in \mathcal{K}$, such that

$$\hat{F}(x) = F(x)(1 + \psi(x)) + \varphi(x) \qquad \forall x \in \mathcal{K}. \tag{2}$$

We say that $\hat{F}$ has "additive noise" φ of level β and "multiplicative noise" ψ of level α. To simplify our analysis, we assume that $F \geq 0$ and that F has minimizer $x^\star \in \mathcal{K}$ such that $F(x^\star) = 0$ (if not, we can always shift F and $\hat{F}$ down by the constant $F(x^\star)$ to satisfy this assumption). That way, the multiplicative noise ψ has the convenient property that it goes to zero as we approach $x^\star$.

We first describe our algorithm and the proof assuming that $\hat{F}$ is smooth and we have access to its gradient $\nabla \hat{F}$. Specifically, we assume that $\|\nabla \hat{F}\|$ is bounded above by some number $\tilde{\lambda} > 0$ and that the Hessian of $\hat{F}$ has singular values bounded above by $L > 0$. This simplifies the presentation considerably and we explain how to deal with the non-smooth case at the end of this section.

Our algorithm. To find an approximate minimizer of F, we would like to design a Markov chain whose stationary distribution concentrates in a subset of $\mathcal{K}$ where the values of F are small. The optimal choice of parameters for this Markov chain will depend on the amount of noise present. Since the bounds on the noise are not uniform, the choice of these parameters will depend on the current state of the chain. To deal with this fact, we will run a sequence of Markov chains in different epochs, where the parameters of the chain are fixed throughout each epoch. Our algorithm runs for $k_{\max}$ epochs, with each epoch indexed by k.

In epoch k, we run a separate Markov chain $\{X_i^{(k)}\}_{i=1}^{i_{\max}}$ over $\mathcal{K}$ for the same number of iterations $i_{\max}$. Each such Markov chain has parameters ξ_k and η_k that depend on k. We think of ξ_k^{-1} as the "temperature" of the Markov chain and η_k as the step size. At the beginning of each epoch, we decrease the temperature and step size, and keep them fixed throughout the epoch. We explain quantitatively how we set the temperature a bit later. Each Markov chain also has an initial point $X_0^{(k)} \in \mathcal{K}$. This initial point is chosen from the uniform distribution on a small ball centered at the point in the Markov chain of the previous epoch $(k-1)$ with the smallest value of $\hat{F}$. In the final epoch, the algorithm outputs a solution $\hat{x}$, where $\hat{x}$ is chosen to be the point in the Markov chain of the final epoch with the smallest value of $\hat{F}$.

Description of the Markov chain in a single epoch. We now describe how the Markov chain, at the point $X_i^{(k)}$ in the k-th epoch, chooses the next point $X_{i+1}^{(k)}$. First, we compute the gradient $\nabla\hat{F}(X_i^{(k)})$. Then we compute a "proposal" X'_{i+1} for the next point as follows

$$X'_{i+1} = X_i^{(k)} - \eta_k \nabla\hat{F}(X_i^{(k)}) + \sqrt{\frac{2\eta_k}{\xi_k}} P_i, \tag{3}$$

where P_i is sampled from $N(0, I_d)$. If X'_{i+1} is inside the domain $\mathcal{K}$, then we accept the proposal and set $X_{i+1}^{(k)} = X'_{i+1}$; otherwise we reject the proposal and we set $X_{i+1}^{(k)} = X_i^{(k)}$ – which is the old point. The update rule in Equation (3) is called the Langevin dynamics. This is a version of gradient descent injected with a random term. The amount of randomness is controlled by the temperature ξ_k^{-1} and the step size η_k. This randomness allows the Markov chain to escape local minima when $\hat{F}$ is not convex. Although the stationary distribution of this Markov chain is not known exactly, roughly speaking it is approximately proportional[2] to $e^{-\xi_k \hat{F}}$. This completes the description of our algorithm in the smooth case and we now turn to explaining the steps involved in bounding its running time for a given bound on the error $\hat{\varepsilon}$.

Steps in bounding the running time. In every epoch, the algorithm makes multiplicative progress so that the smallest value of F achieved by the Markov chain decreases by a factor of $1/10$. To achieve an error $\hat{\varepsilon}$, our algorithm therefore requires $k_{\max} = O(\log \frac{M}{\hat{\varepsilon}})$ epochs, where M is the maximum value of $\hat{F}$ on $\mathcal{K}$ ($M \leq \lambda R$). The running time of our algorithm is given by the number of epochs $k_{\max}$ multiplied by the number of steps $i_{\max}$ taken by the Markov chain within each epoch. For simplicity, we will run the Markov chain at each epoch for the same number of steps $i_{\max}$. For the value of F to decrease by a factor of $1/10$ in each epoch, we must set the number

2. By this we actually mean that the Markov chain is ε'-close to another Markov chain Z with stationary distribution $\propto e^{-\xi_k \hat{F}}$ (see Definition 5 and Theorem 12). Specifically, a Markov chain X on a space S with transition kernel Q_X is said to be ε'-close to a Markov chain Z w.r.t. a set U if $Q_Z(x, A) \leq Q_X(x, A) \leq (1+\varepsilon)Q_Z(x, A)$ for all $x \in \mathcal{K}\backslash U$ and $A \subseteq \mathcal{K}\backslash\{x\}$ (Zhang et al., 2017)

of steps $i_{\max}$ taken by the Markov chain during each epoch to be no less than the hitting time of the Markov chain for epoch k to a sub-level set $U_k \subseteq \mathcal{K}$ of F, where the "height" of U_k is one-tenth the value of F at the initial point in this Markov chain. By the height of a sub-level set, we mean the largest value of F achieved at any point on that sub-level set, that is the sub-level set $\{y \in \mathcal{K} : F(y) \leq h\}$ has height h. Thus, bounding the hitting time will allow us to bound the number of steps $i_{\max}$ for which we must run each Markov chain. Specifically, we should choose $i_{\max}$ to be no less than the greatest hitting time in any of the epochs with high probability.

This approach was used in the simpler setting of additive noise and a non-iterative way by Zhang et al. (2017). Thus, the running time is roughly the product of the number of epochs and the hitting time to the sub-level set U_k, and having determined the number of epochs required for a given accuracy, we proceed to bounding the hitting time.

Bounding the hitting time and the Cheeger constant. To bound the hitting time of the Markov chain in a single epoch, we use the strategy of Zhang et al. (2017), who bound the hitting time of the Langevin dynamics Markov chain in terms of the Cheeger constant. Since the Markov chain has approximate stationary measure induced by $e^{-\xi_k \hat{F}}$, we consider the Cheeger constant with respect to this measure, defined as follows:

Given a probability measure μ on some domain, we consider the ratio of the measure of the boundary of an arbitrary subset A of the domain to the measure of A itself. The Cheeger constant of a set V is the infimum of these ratios over all subsets $A \subseteq V$ (see Definition 4 in Section C.2 for a formal definition). We use some of the results of Zhang et al. (2017) to show a bound on the hitting time to the sub-level set U_k contained in a larger sub-level set U_k' in terms of the Cheeger constant $\hat{\mathcal{C}}_k$, with respect to the measure induced by $e^{-\xi_k \hat{F}(x)}$ on U_k'. Specifically, we set U_k' to be the sub-level set of height $\hat{F}(X_0^{(k)}) + \xi_k^{-1} d$ and U_k to be the sub-level set of height $\frac{1}{10} F(X_0^{(k)})$ and show that for a step size

$$\eta_k = \frac{(\hat{\mathcal{C}}_k)^2}{d^3((\xi_k \tilde{\lambda})^2 + \xi_k L)^2},$$

the hitting time to U_k is bounded by $\frac{R\tilde{\lambda}\xi_k + d}{\sqrt{\frac{\eta_k}{d}}\hat{\mathcal{C}}_k}$; see Section C.5. Thus, to complete our bound on the hitting time we need to bound the corresponding Cheeger constants.

Bounding the Cheeger constant. We would like to bound the Cheeger constant of the measure induced by $e^{-\xi_k \hat{F}(x)}$. However, $\hat{F}$ is not convex, so we cannot directly apply the usual approach of Lovász and Simonovits (1993) for convex functions. Instead, we first apply their result to bound the Cheeger constant of the convex function F. We then bound the Cheeger constant of the nonconvex function $\hat{F}$ in terms of the Cheeger constant of the convex function F, using a very useful stability property satisfied by the Cheeger constant.

Roughly speaking, we show that the Cheeger constant of $U_k' \backslash U_k$ is bounded below by $1/R$ (where R is the radius of the bounding ball for $\mathcal{K}$) as long as the inverse temperature satisfies

$$\xi_k \geq \frac{d}{1/10 F(X_0^{(k)})}$$

(see Lemma 6). However, the difficulty is that since U_k' may have sharp corners, the volume of U_k might be so small that U_k would have much smaller measure than $U_k' \backslash U_k$, leading to a very small Cheeger constant. To get around this problem, we instead consider a slightly "rounded" version of

$\mathcal{K}$, where we take $\mathcal{K}$ to be the Minkowski sum of another convex body with a ball of very small radius r'. The roundness allows us to show that U_k contains a ball of even smaller radius $\hat{r}$ such that the measure is much larger on this ball than at any point in $U_k' \backslash U_k$. This in turn allows us to apply the results of Lovász and Simonovits (1993) to show that the Cheeger constant is bounded below by $1/R$ (see Lemma 6). Note our Cheeger bound is more general (for convex functions) than that obtained by Zhang et al. (2017), where the constraint set is assumed to be a ball.

The Cheeger constant has the following useful stability property that allows us to bound the Cheeger constant of the nonconvex $\hat{F}$ with respect to the convex F: if $|\hat{F}(x) - F(x)| \leq N_k$ for all $x \in U_k'$, then the Cheeger constant for the measures proportional to $e^{-\xi_k F}$ and $e^{-\xi_k \hat{F}}$ differ by a factor of at most $e^{-2\xi_k N_k}$. For our choice of U_k', we have

$$N_k \approx \alpha[F(X_0^{(k)}) + \xi_k^{-1} d] + \beta.$$

We can then use the stability property to show that the Cheeger constants of $\hat{F}$ and F differ by a factor of at most $e^{-2\xi_k N_k}$, allowing us to get a large bound for the Cheeger constant of $\hat{F}$ in terms of our bound for the Cheeger constant of F as long as the bound on the noise N_k on U_k' is not too large, namely we get that the Cheeger constant is bounded below by

$$\frac{1}{R} e^{-2\xi_k N_k} \approx \frac{1}{R} \exp\left(-\alpha d - \frac{d}{F(X_0^{(k)})}\beta\right)$$

if we choose $\xi_k = \frac{d}{\frac{1}{10}F(X_0^{(k)})}$.

At this point we mention the key difference between the approach of Zhang et al. (2017) and ours in bounding the hitting time. As Zhang et al. (2017) assume a uniform bound on the noise they only consider the Cheeger constant of $\mathcal{K} \backslash U_k$, where $\mathcal{K}$ is the entire constraint set and is assumed to be a ball. Since the noise in our model depends on the "height" of the level sets, we instead need to bound the Cheeger constant of $U_k' \backslash U_k$, where U_k' is the level set of height $\hat{F}(X_0^{(k)}) + \xi_k^{-1} d$ and U_k is the level set of height $\frac{1}{10}F(X_0^{(k)})$.

In order to complete our bound for the Cheeger constant of $\hat{F}$, we still need to verify that we can choose a temperature such that the Cheeger constant of F is large and the Cheeger constants of F and $\hat{F}$ are close at this same temperature.

Requirements on the temperature to bound the Cheeger constant. To get a large bound for the Cheeger constant of $\hat{F}$, we need to use a temperature ξ_k^{-1} such that the following competing requirements are satisfied:

1. We want the Cheeger constant of the convex objective function F on $U_k' \backslash U_k$ to be bounded below by $1/R$. We can show such a bound on the Cheeger constant if the temperature is *low* enough, in particular a temperature of $\xi_k^{-1} \approx \frac{\frac{1}{10}F(X_0^{(k)})}{d}$ suffices.
2. We need the Markov chain to stay inside a level set on which the upper bound N_k on the noise is at most $\alpha[F(X_0^{(k)}) + \xi_k^{-1} d] + \beta$, to show that the Cheeger constants of F and $\hat{F}$ are close. That is, we need to show that the ratio $e^{-2\xi_k N_k}$ of the Cheeger constants of F and $\hat{F}$ is not too small, roughly $e^{-2\xi_k N_k} \geq \exp\left(-\alpha d - \frac{d}{F(X_0^{(k)})}\beta\right)$. This again requires the temperature to be *low* enough, with $\xi_k^{-1} \approx \frac{\frac{1}{10}F(X_0^{(k)})}{d}$ sufficing.

3. To show that the ratio of the Cheeger constants roughly satisfies $e^{-2\xi_k N_k} \geq \exp\left(-\alpha d - \frac{d}{F(X_0^{(k)})}\beta\right)$, we also need the temperature to be *high* enough. Specifically, a temperature of $\xi_k^{-1} \approx \frac{\frac{1}{10}F(X_0^{(k)})}{d}$ suffices for this requirement as well.

At some epoch k, the value of F becomes too low for all three of these requirements on the temperature to be satisfied simultaneously. At this point the Cheeger constant and hitting time to U_k become very large no matter what temperature we use, so that the minimum value of F obtained by the Markov chain no longer decreases by a large factor in $i_{\max}$ steps.

Quantitative error and running time bounds. We now give a more quantitative analysis to determine at what point F stops decreasing. The value of F at this point determines the error $\hat{\varepsilon}$ of the solution returned by our algorithm. Towards this end, we set the inverse temperature to be $\xi_k = \frac{d}{\frac{1}{10}F(X_0^{(k)})}$ and check to what extent all 3 requirements above are satisfied.

1. We start by showing that if the temperature roughly satisfies $\xi_k^{-1} \leq \frac{\frac{1}{10}F(X_0^{(k)})}{d}$ then the Cheeger constant for F on $U_k' \backslash U_k$ is bounded below by $1/R$ (see Lemma 6).

2. We then show that at each epoch the Markov chain remains with high probability in the level set U_k' of height $\hat{F}(X_0^{(k)}) + \xi_k^{-1}d$ (Lemma 10). The fact that the noise satisfies $|\hat{F}(x) - F(x)| \leq \alpha F(x) + \beta$ (note that we assume $F \geq 0$), implies that the noise is roughly bounded above by $N_k = \alpha[F(X_0^{(k)}) + \xi_k^{-1}d] + \beta$ on this level set.

3. Since we chose the temperature to be $\xi_k^{-1} = \frac{\frac{1}{10}F(X_0^{(k)})}{d}$, we have that

$$\xi_k N_k \approx \alpha d + \frac{d}{F(X_0^{(k)})}\beta.$$

Combing these three facts we get that the Cheeger constant is bounded below by $\frac{1}{R}e^{-2\xi_k N_k} \approx \frac{1}{R}\exp\left(-\alpha d - \frac{d}{F(X_0^{(k)})}\beta\right)$. If we run the algorithm for enough epochs to reach $F(X_0^{(k)}) \leq \hat{\varepsilon}$ for any desired error $\hat{\varepsilon} > 0$, the Cheeger constant will be roughly bounded below by $\frac{1}{R}\exp(-\alpha d - \frac{d}{\hat{\varepsilon}}\beta)$.

Recall that the hitting time is bounded by $\frac{R\tilde{\lambda}\xi_k + d}{\sqrt{\frac{\eta_k}{d}}\hat{\mathcal{C}}_k}$, for stepsize $\eta_k \approx \frac{(\hat{\mathcal{C}}_k)^2}{Rd^3((\xi_k\tilde{\lambda})^2 + \xi_k L)^2}$. Choosing $i_{\max}$ to be equal to our bound on the hitting time, and recalling that $k_{\max} = \tilde{O}(1)$, we get a running time of roughly

$$\tilde{O}\left(R^{\frac{3}{2}}\left[d^5\frac{\tilde{\lambda}^3}{\hat{\varepsilon}^3} + d^{\frac{5}{2}}\frac{L}{\hat{\varepsilon}}\right]\exp(c[\alpha d + \frac{d}{\hat{\varepsilon}}\beta])\right),$$

for some $c = \tilde{O}(1)$.

Therefore, for our choice of inverse temperature $\xi_k = \frac{d}{\frac{1}{10}F(X_0^{(k)})}$, the running time is polynomial in d, R, λ and $\tilde{\lambda}$ whenever the multiplicative noise level satisfies $\alpha \leq \tilde{O}(\frac{1}{d})$ and the additive noise level satisfies $\beta \leq \tilde{O}(\frac{\hat{\varepsilon}}{d})$. As discussed in the introduction, the requirements that $\alpha \leq \tilde{O}(\frac{1}{d})$ and $\beta \leq \tilde{O}(\frac{\hat{\varepsilon}}{d})$ are not an artefact of the analysis or algorithm and are in fact tight.

Drift bounds and initialization. So far we have been implicitly assuming that the Markov chain does not leave U'_k, so that we could analyze the Markov chain using the Cheeger constant on U'_k. We now show that this assumption is indeed true with high probability. This is important to verify, since there are examples of Markov chains where the Markov chain may have a high probability of escaping a level set U'_k, even if this level set contains most of the stationary measure, provided that the Markov chain is started far from the stationary distribution.

To get around this problem, at each epoch we choose the initial point $X_0^{(k)}$ from the uniform distribution on a ball of radius r centered at the point in the Markov chain of the previous epoch $k-1$ with the smallest value of $\hat{F}$. We then show that if the Markov chain is initialized in this small ball, it has a low probability of leaving the level set U'_k (see Propositions 8, 10 and Lemma 10).

Our method of initialization is another crucial difference between our algorithm and the algorithm by Zhang et al. (2017) and Belloni et al. (2015), since it allows us to effectively restrict the Markov chain to a sub-level set of the objective function F, which we do not have direct oracle access to, rather than restricting the Markov chain to a large ball as by Belloni et al. (2015) or the entire constraint set $\mathcal{K}$ as by Zhang et al. (2017) for which we have a membership oracle. This in turn allows us to get a tighter bound on the multiplicative noise than would otherwise be possible, since the amount of multiplicative noise depends, by definition, on the sub-level set.

We still need to show that the chain $X^{(k)}$ does not leave the set U'_k with high probability. To bound the probability that $X^{(k)}$ leaves U'_k, we would like to use the fact that most of its stationary distribution is concentrated in U'_k. However, the problem is that we do not know the stationary distribution of $X^{(k)}$. To get around this, we consider a related Markov chain $Y^{(k)}$ with known stationary distribution. The chain $Y^{(k)}$ evolves according to the same update rules as $X^{(k)}$, using the same sequence of Gaussian random vectors $P_1, P_2, \ldots$ and the same starting point, except that it performs a Metropolis "accept-reject" step that causes its stationary distribution to be proportional to $e^{-\xi_k \hat{F}}$. The fact that we know the stationary distribution of $Y^{(k)}$ is key to showing that $Y^{(k)}$ stays in the subset U'_k with high probability (see Proposition 9). We then argue that $Y^{(k)} = X^{(k)}$ with high probability, implying that $X^{(k)}$ also stays inside U'_k with high probability (see Lemma 10).

Another coupled toy chain. So far we have shown that the Markov chain $X^{(k)}$ stays inside the set U'_k *with high probability*. However, to use the stability property to bound the hitting time of the Markov chain $X^{(k)}$ to the set U_k, we actually want $X^{(k)}$ to be *restricted* to the set U'_k where the noise is not too large. In reality, however, the domain of $X^{(k)}$ is all of $\mathcal{K}$, so we cannot directly bound the hitting time of $X^{(k)}$ with the Cheeger constant of $U'_k \backslash U_k$. Instead, we consider a Markov chain $\hat{X}^{(k)}$ that evolves according to the same rules as $X^{(k)}$, except that it rejects any proposal outside of U'_k. Since $\hat{X}^{(k)}$ has domain U'_k, we can use our bound on the Cheeger constant of $U'_k \backslash U_k$ to obtain a bound on the hitting time of $\hat{X}^{(k)}$. Then, we argue that since $X^{(k)}$ stays in U'_k with high probability, and $\hat{X}^{(k)}$ and $X^{(k)}$ evolve according to the same update rules as long as $X^{(k)}$ stays inside U'_k, $\hat{X}^{(k)} = X^{(k)}$ with high probability as well, implying a hitting time bound for $X^{(k)}$.

Rounding the sub-level sets. We must also show a bound on the roundness of the sets U'_k, to avoid the possibillty of the Markov chain getting stuck in "corners". Zhang et al. (2017) take this as an assumption about the constraint set. However, since we must consider the Cheeger constant on sub-level sets U'_k rather than just on the entire constraint set, we must make sure that these sub-level sets are "round enough". Towards this end we consider "rounded" sub-level sets where we take the Minkowski sum of U'_k with a ball of a small radius r'. We then apply the Hanson-Wright inequality

to show that any Gaussian random variable with center inside this rounded sub-level set and small enough covariance remains inside the rounded sub-level set with high probability (see Lemma 17).

Smoothing a non-differentiable noisy oracle. Finally, so far we have considered the special case where $\hat{F}$ is smooth. However, $\hat{F}$ may not be smooth or may not even be differentiable, so we may not have access to a well-behaved gradient which we need to compute the Langevin dynamics Markov chain (Equation 3). To get around this problem, we follow the approach of Duchi et al. (2015) and Zhang et al. (2017). We define a smoothed function

$$\tilde{f}_\sigma(x) := \mathbb{E}_Z[\hat{F}(x+Z)]$$

where $Z \sim N(0, \sigma I_d)$ and $\sigma > 0$ is a parameter we must fix. The smoothness of $\tilde{f}_\sigma$ comes from the fact that $\tilde{f}_\sigma$ is a convolution of $\hat{F}$ with a Gaussian distribution.

When choosing σ, we want σ to be small enough so that we get a good bound on the noise $|\tilde{f}_\sigma(x) - F(x)|$. Specifically, we need

$$\sigma = \tilde{O}\left(\min\left(\frac{r}{\sqrt{d}}, \frac{\beta + \hat{\varepsilon}\alpha + \hat{\varepsilon}/d}{\lambda\sqrt{d}}\right)\right),$$

where λ is a bound on $\|\nabla F\|$. On the other hand, we also want σ not to be too small so that we get a good bound on the smoothness of $\tilde{f}_\sigma$.

Further, so far we have also assumed that we have access to the full gradient of $\hat{F}$, but in general $\hat{F}$ may not even have a gradient. Instead, we would like to use the gradient of $\tilde{f}_\sigma$ to compute the proposal for the Langevin dynamics Markov chain (Equation (3)). However, computing the full gradient of $\tilde{f}_\sigma$ can be expensive, since we do not even have direct oracle access to $\tilde{f}_\sigma$. Instead, we compute a projection $g(x)$ of $\nabla\tilde{f}_\sigma$, where

$$g(x) = \frac{Z}{\sigma^2}(\hat{F}(x+Z) - \hat{F}(x))$$

Since g has the property that $\mathbb{E}[g(x)] = \nabla\tilde{f}_\sigma$, g is called a "stochastic gradient" of $\tilde{f}_\sigma$. We use this stochastic gradient g in place of the full gradient of $\hat{F}$ when computing the proposal for the Langevin dynamics Markov chain (Equation 3). This gives rise to the following Markov chain proposal, also known as stochastic gradient Langevin dynamics (SGLD):

$$X'_{i+1} = X_i^{(k)} - \eta_k g(X_i^{(k)}) + \sqrt{\frac{2\eta_k}{\xi_k}} P_i.$$

To bound the running time of SGLD, we will need a bound on the magnitude of the gradient of $\tilde{f}_\sigma$ (see Lemma 13), bounds on the Hessian and tails of $\tilde{f}_\sigma$, which we obtain from Zhang et al. (2017) (see Lemma 15), and bounds on the noise of the smoothed function, $|\tilde{f}_\sigma - F(x)|$ (see Lemma 16).

Although in this technical overview we largely showed running time and error bounds assuming access to a full gradient, in reality we prove Theorem 2 for the more general stochastic gradient Langevin dynamics algorithm, where we only assume access to a stochastic gradient of a smooth function. Therefore, the bounds on the noise and smoothness of $\tilde{f}_\sigma$ allow us to extend the error and polynomial running time bounds shown in this overview to the more general case where $\hat{F}$ may not be differentiable.

References

David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 156–163. ACM, 1991.

Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Conference on Learning Theory*, pages 240–265, 2015.

Avrim Blum and Ronald L Rivest. Training a 3-node neural network is NP-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

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.

Ruobing Chen. *Stochastic derivative-free optimization of noisy functions*. Ph.D. Thesis, Lehigh University, 2015.

Ruobing Chen, Matt Menickelly, and Katya Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, pages 1–41, 2015.

K Andrew Cliffe, Mike B Giles, Robert Scheichl, and Aretha L Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Computing and Visualization in Science*, 14(1):3, 2011.

Patrick R Conrad, Andrew D Davis, Youssef M Marzouk, Natesh S Pillai, and Aaron Smith. Parallel local approximation MCMC for expensive models. *SIAM/ASA Journal on Uncertainty Quantification*, 6(1):339–373, 2018.

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.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Elad Hazan, Kfir Yehuda Levy, and Shai Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *International Conference on Machine Learning*, pages 1833–1841, 2016.

Mohamed Jebalia and Anne Auger. On multiplicative noise models for stochastic search. In *International Conference on Parallel Problem Solving from Nature*, pages 52–61. Springer, 2008.

Mohamed Jebalia, Anne Auger, and Nikolaus Hansen. Log-linear convergence and divergence of the scale-invariant $(1 + 1)$-ES in noisy environments. *Algorithmica*, 59(3):425–460, 2011.

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

Yin Tat Lee and Santosh S Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. *To appear in Proceedings of STOC 2018, arXiv preprint arXiv:1710.06261*, 2017.

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

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

Andrej Risteski and Yuanzhi Li. Algorithms and matching lower bounds for approximately-convex optimization. In *Advances in Neural Information Processing Systems*, pages 4745–4753, 2016.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab*, 18(82):1–9, 2013.

Yaron Singer and Jan Vondrák. Information-theoretic lower bounds for convex optimization with erroneous oracles. In *Advances in Neural Information Processing Systems*, pages 3204–3212, 2015.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022, 2017.

Appendix A. Preliminaries

In this section we go over notation and assumptions that we use to state our algorithm and prove our main result. We start by giving assumptions we make about the convex objective function F. We then explain how to obtain an oracle for the gradient of the smoothed function $\tilde{f}_\sigma$ if we only have access to the non-smooth oracle $\hat{F}$.

A.1. Notation

In this section we define the notation we use to prove our main result. For any set $S \subseteq \mathbb{R}^d$ and $t \geq 0$ define $S_t := S + B(0,t)$ where "+" denotes the Minkowski sum. We denote the ℓ_2-norm by $\|\cdot\|$, and the $d \times d$ identity matrix by I_d. We denote by $\|\cdot\|_{\mathrm{op}}$ the operator norm of a matrix, that is, its largest singular value. We define $B(a,t)$ to be the closed Euclidean ball with center a and radius t. Denote the multivariate normal distribution with mean m and covariance matrix Σ by $N(m,\Sigma)$. Let $x^\star$ denote a minimizer of F on $\mathcal{K}$.

A.2. Assumptions on the convex objective function and the constraint set

We make the following assumptions about the convex objective function F and $\mathcal{K}$:

- $\mathcal{K}$ is contained in a ball, with $\mathcal{K} \subseteq B(c,R)$ for some $c \in \mathbb{R}^d$.
- $F(x^\star) = 0$.[3]
- There exists an $r' > 0$ and a convex body $\mathcal{K}'$ such that $\mathcal{K} = \mathcal{K}' + B(0,r')$ for some convex body. (This assumption is necessary to ensure that our convex body does not have "pointy" edges, so that the Markov chain does not get stuck for a long time in a corner.)
- F is convex over $\mathcal{K}_{\mathsf{r}}$ for some $\mathsf{r} > 0$.
- $\|\nabla F(x)\| \leq \lambda$ for all $x \in \mathcal{K}_{\mathsf{r}}$, where $\lambda > 0$.

A.3. A smoothed oracle from a non-smooth one

In this section we show how to obtain a smooth noisy oracle for F if one only has access to a non-smooth and possibly non-continuous noisy oracle $\hat{F}$. Our goal is to find an approximate minimum for F on the constraint set $\mathcal{K}$. (We consider the thickened set $\mathcal{K}_{\mathsf{r}}$ only to help us compute a smooth oracle for F on $\mathcal{K}$). We assume that we have access to a noisy function $\hat{F}$ of the form

$$\hat{F}(x) = F(x)(1+\psi(x)) + \varphi(x), \tag{4}$$

where $|\psi(x)| < \alpha$, and $|\varphi(x)| < \beta$ for every $x \in \mathcal{K}_{\mathsf{r}}$, for some $\alpha, \beta \geq 0$. We extend $\hat{F}$ to values outside $\mathcal{K}_{\mathsf{r}}$ by setting $\hat{F}(x) = 0$ for all $x \notin \mathcal{K}_{\mathsf{r}}$. Since $\hat{F}$ need not be smooth, as in Duchi et al. (2015) and Zhang et al. (2017) we will instead optimize the following smoothed function

$$\tilde{f}_\sigma(x) := \mathbb{E}_Z[\hat{F}(x+Z)] \tag{5}$$

3. If $F(x^\star)$ is nonzero, we can define a new objective function $F'(x) = F(x) - F(x^\star)$ and a new noisy function $\tilde{f}'(x) = \tilde{f}(x) - F(x^\star)$. The noise $N'(x) = \tilde{f}'(x) - F'(x)$ can then be modeled as having additive noise of level $\beta' = \beta + \alpha F(x^\star)$ and multiplicative noise of level $\alpha' = \alpha$, if $N(x) = \tilde{f}(x) - F(x)$ has additive noise of level β and multiplicative noise of level α.

where $Z \sim \mathcal{N}(0, \sigma I_d)$, for some $\sigma > 0$. The parameter σ determines the smoothness of $\tilde{f}_\sigma$; a larger value of σ will mean that $\tilde{f}_\sigma$ will be smoother. The gradient of $\tilde{f}_\sigma(x)$ can be computed using a stochastic gradient $g(x)$, where

$$g(x) \equiv g_Z(x) := \frac{1}{\sigma^2} Z \left(\hat{F}(x+Z) - \hat{F}(x) \right), \qquad \nabla \tilde{f}_\sigma(x) = \mathbb{E}_Z[g(x)].$$

Appendix B. Our Contribution

B.1. Our Algorithm

In this section we state our simulated annealing algorithm (Algorithm 2) that we use to obtain a solution to Problem 1. At each epoch, our algorithm uses the SGLD Markov chain as a subroutine, which we describe first in Algorithm 1. The SGLD Markov chain we describe here is the same algorithm used in Zhang et al. (2017), except that we allow the user to specify the initial point.

Algorithm 1 Stochastic gradient Langevin dynamics (SGLD)

input: Convex constraint set $\hat{\mathcal{K}} \subseteq \mathbb{R}^d$, inverse temperature $\xi > 0$, step size $\eta > 0$, parameters $i_{\max} \in \mathbb{N}$ and $D > 0$, and a stochastic gradient oracle g for some $\tilde{f} : \mathcal{K} \to \mathbb{R}$.
input: Initial point $X_0 \in \hat{\mathcal{K}}$.
for $i = 0$ *to* $i_{\max}$ **do**
1. Sample $P_i \sim N(0, I_d)$.
2. Set $X'_{i+1} = X_i - \eta g(X_i) + \sqrt{\frac{2\eta}{\xi}} P_i$.
3. Set $X_{i+1} = X'_{i+1}$ if $X'_{i+1} \in \hat{\mathcal{K}} \cap B(X_i, D)$. Otherwise, set $X_{i+1} = X_i$.

end
output: $X_{i^\star}$, where $i^\star := \operatorname{argmin}_i \{\hat{F}(X_i)\}$

Using Algorithm 1 as a subroutine, we define the following simulated annealing algorithm:

Algorithm 2 Simulated annealing SGLD

input: Convex constraint set $\hat{\mathcal{K}} \subseteq \mathbb{R}^d$, initial point $x_0 \in \hat{\mathcal{K}}$, inverse temperatures $\xi_0, \xi_1, \ldots, \xi_{k_{\max}}$, step sizes $\eta_0, \eta_1, \ldots, \eta_{k_{\max}}$, parameters $k_{\max}, i_{\max} \in \mathbb{N}$, $D > 0$ and $r > 0$, and a stochastic gradient oracle g for some $\tilde{f} : \hat{\mathcal{K}} \to \mathbb{R}$.
1. Sample y_0 from the uniform distribution on $B(x_0, r) \cap \hat{\mathcal{K}}$.
for $k = 0$ *to* $k_{\max}$ **do**
2. Run Algorithm 1 on $\hat{\mathcal{K}}$, inverse temperature $\xi = \xi_k$, and step size η_k, $i_{\max}$, the oracle g, and the initial point $X_0 = y_k$. Let x_{k+1} be the output of Algorithm 1.
3. Sample y_{k+1} from the uniform distribution on $B(x_{k+1}, r) \cap \hat{\mathcal{K}}$.

end
output: $x_{k_{\max}}$

B.2. Statement of Our Main Theorem

We now formally state our main result, where we bound the error and running time when Algorithm 2 is used to solve Problem 1, assuming access to an oracle $\hat{F}$ that may be non-smooth or even non-continuous.

Theorem 3 ***(Main Theorem: Error bounds and running time for Algorithm 2)*** *Let $F : \mathcal{K} \to \mathbb{R}$ be a convex function, and $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set that satisfy the assumptions stated in Section A.1. Let $\hat{F}$ be a noisy oracle for F with multiplicative noise of level $\alpha \le O(1)$ and additive noise of level β, as in Equation (4). Let $\hat{\varepsilon} \ge 75\beta$ and $\delta' > 0$. Then there exist parameters i_{max}, k_{max}, $(\xi_k, \eta_k)_{k=0}^{k_{\max}}$, and σ, such that if we run Algorithm 2 with a smoothed version $\tilde{f}_\sigma$ of the oracle $\hat{F}$ (as defined in Section A.3), with probability at least $1 - \delta'$, the algorithm outputs a point $\hat{x}$ such that*

$$F(\hat{x}) - F(x^\star) \le \hat{\varepsilon},$$

with running time that is polynomial in d, $e^{\left(d\alpha + d\frac{\beta}{\hat{\varepsilon}}\right)\mathsf{c}}$, R, λ, $\frac{1}{r'}$, β, $\frac{1}{\hat{\varepsilon}}$, and $\log\frac{1}{\delta'}$, where $\mathsf{c} = O(\log(R + \lambda))$.

We give a proof of Theorem 3 in Section C.8.
The precise values of the parameters in this theorem are quite involved and appear in the proofs at the following places: ξ_k appears in (39), η_k in (41), $i_{\max}$ in (40), and the expression for $k_{\max}$ can be found in (38). Below we present their approximate magnitudes. The inverse temperature parameter ξ_k, the smoothing parameter σ, and the number of epochs $k_{\max}$ satisfy:

$$\tilde{\Omega}\left(\frac{d}{\lambda R}\right) \le \xi_k \sim \tilde{O}\left(d \cdot \max\left\{\frac{1}{\hat{\varepsilon}}, \hat{F}(X_0^{(0)}) \cdot 10^k\right\}\right) \le \tilde{O}\left(\frac{d}{\hat{\varepsilon}}\right),$$

$$\sigma = \frac{1}{2}\min\left(\frac{\beta}{\lambda(1+\alpha)\sqrt{d}}, \frac{\mathsf{r}}{\sqrt{\log(\frac{1}{\alpha}) + d}}\right),$$

$$k_{\max} \sim \log\frac{R}{\hat{\varepsilon}}.$$

To make the expressions for η_k and $i_{\max}$ understandable, assume that $\lambda R > 1, \beta$, that $\mathsf{r} > \frac{\beta}{\lambda}$, $d > \hat{\varepsilon}$, and that $R > \lambda$. Then

$$\eta_k \sim \frac{\hat{\varepsilon}^4}{d^9 R^5 \lambda^8 \beta^4} e^{-d\left(\alpha + \frac{\beta}{\hat{\varepsilon}}\right)\mathsf{c}'} 10^{-k}$$

$$1 \le i_{\max} \le \left[\frac{d^{6.5}}{\hat{\varepsilon}^3} R^{\frac{11}{2}} \lambda^6 \beta e^{d\left(\alpha + \frac{\beta}{\hat{\varepsilon}}\right)\mathsf{c}'}\right]^{1+\mathsf{c}''\alpha},$$

where $\mathsf{c}' \sim \log\frac{R^2}{rd\delta\min\{\hat{\varepsilon}/\lambda, r'\}}$ and c'' is a constant factor. In particular, the running time is given by $i_{\max} \times k_{\max}$.

Appendix C. Proofs

C.1. Assumptions about the smooth oracle

We first show how to optimize F if one has access to a smooth noisy objective function $\tilde{f} : \mathcal{K} \to \mathbb{R}$ (Sections C.2, C.3, C.5). Then, in Section C.6, we show how one can obtain a smooth noisy objective function from a non-smooth and possibly non-continuos noisy objective function $\hat{F}$. We will make the following assumptions (we prove in Section C.6 that these assumptions hold for a smoothed version $\tilde{f}_\sigma$ of a non-smooth noisy objective function $\hat{F}$). We assume the following noise model for $\tilde{f}$:

$$\tilde{f}(x) = F(x)(1 + \psi(x)) + \varphi(x),$$

for all $x \in \mathcal{K}$ where $|\psi(x)| \leq \alpha$ and $|\varphi(x)| \leq \beta$. Note that, with a slight abuse of notation, in Section A.3 we also used the letters α and β to denote the noise levels of the non-smooth oracle $\hat{F}$, even though typically $\hat{F}$ will have lower noise levels than $\tilde{f}$. In this section, as well as in Sections C.2-C.5 where we assume direct access to a stochastic gradient for the smooth oracle $\tilde{f}$, we will instead refer to the noise levels of $\hat{F}$ by "$\hat{\alpha}$" and "$\hat{\beta}$". In Section B.2, on the other hand, "α" and "β" will be used exclusively to denote the noise levels of $\hat{F}$. We also assume that

$$\alpha \geq \hat{\alpha} \quad \text{and} \quad \beta \geq \hat{\beta}. \tag{6}$$

We make the following assumptions about $\tilde{f}$:

- $\psi(x) > -\alpha^\dagger$ for some $0 \leq \alpha^\dagger < 1$. This assumption is needed because if not we might have $\psi(x) = -1$ for all $x \in \mathcal{K}$, in which case $\tilde{f}(x)$ would give no information about F.
- $\|\nabla \tilde{f}(x)\| \leq \tilde{\lambda}$ for all $x \in \mathcal{K}$.
- We assume that we have access to a stochastic gradient g such that $\nabla \tilde{f}(x) = \mathbb{E}[g(x)]$ for every $x \in \mathcal{K}$. However, we do *not* assume that we have oracle access to $\tilde{f}$ itself.

Assumption 1 *(**Based on assumption A in Zhang et al. (2017)**) Let $\tilde{f} : \mathcal{K} \to \mathbb{R}^d$ be differentiable, and let $g \equiv g_w : \mathcal{K} \to \mathbb{R}^d$ be such that $\nabla \tilde{f}(x) = \mathbb{E}[g_W(x)]$ where W is a random variable. We will assume that*

1. *There exists $\zeta_{\max} > 0$ such that for every compact convex $\hat{\mathcal{K}} \subseteq \mathbb{R}^d$, every $x \in \hat{\mathcal{K}}_{r'}$, and every $0 \leq \zeta \leq \zeta_{\max}$, the random variable $Z \sim N(x, 2\zeta I_d)$ satisfies $\mathbb{P}(Z \in \mathcal{K}) \geq \frac{1}{3}$. We prove this assumption in Lemma 17.*
2. *There exists $L > 0$ such that $|\tilde{f}(y) - \tilde{f}(x) - \langle y - x, \nabla \tilde{f}(x)\rangle| \leq \frac{L}{2}\|y - x\|^2 \qquad \forall x, y \in \mathcal{K}$.*
3. *There exists $b_{\max} > 0$ and $G > 0$ such that for any $u \in \mathbb{R}^d$ with $\|u\| \leq b_{\max}$ the stochastic gradient $g(x)$ satisfies $\mathbb{E}[e^{\langle u, g(x)\rangle^2} | x] \leq e^{G^2\|u\|^2}$.*

C.2. Conductance and bounding the Cheeger constant

To help us bound the convergence rate, we define the Cheeger constant of a distribution, as well as the conductance of a Markov chain. For any set $\hat{\mathcal{K}}$ and any function $f : \hat{\mathcal{K}} \to \mathbb{R}$, define

$$\mu_f^{\hat{\mathcal{K}}}(x) := \frac{e^{-f(x)}}{\int_{\hat{\mathcal{K}}} e^{-f(x)}} \qquad \forall x \in \hat{\mathcal{K}}.$$

Definition 4 ***(Cheeger constant)*** *For all $V \subseteq \hat{\mathcal{K}}$, define the Cheeger constant to be*

$$\mathcal{C}_f^{\hat{\mathcal{K}}}(V) := \liminf_{\varepsilon \downarrow 0} \inf_{A \subseteq V} \frac{\mu_f^{\hat{\mathcal{K}}}(A_\varepsilon) - \mu_f^{\hat{\mathcal{K}}}(A)}{\varepsilon \mu_f^{\hat{\mathcal{K}}}(A)}.$$

recalling that $A_\varepsilon = A + B(0, \varepsilon)$.

For a Markov chain $Z_0, Z_1, \ldots$ on $\hat{\mathcal{K}}$ with stationary distribution μ_Z and transition kernel Q_Z, we define the conductance on a subset V to be

$$\Phi_Z^{\hat{\mathcal{K}}}(V) := \inf_{A \subseteq V} \frac{\int_A Q_Z(x, \hat{\mathcal{K}} \backslash A) \mu_Z(x) \mathrm{d}x}{\mu_Z(A)} \qquad \forall V \subseteq \hat{\mathcal{K}}$$

and the hitting time

$$\tau_Z(A) := \inf\{i : Z_i \in A\} \qquad \forall A \subseteq \hat{\mathcal{K}}.$$

Finally, we define the notion of two Markov chains being ε'-close:

Definition 5 *If $W_0, W_1, \ldots$ and $Z_0, Z_1, \ldots$ are Markov chains on a set $\hat{\mathcal{K}}$ with transition kernels Q_W and Q_Z, respectively, we say that W is ε'-close to Z with respect to a set $U \subseteq \hat{\mathcal{K}}$ if*

$$Q_Z(x, A) \leq Q_W(x, A) \leq (1 + \varepsilon') Q_Z(x, A)$$

for every $x \in \hat{\mathcal{K}} \backslash U$ and $A \subseteq \hat{\mathcal{K}} \backslash \{x\}$.

We now give a generalization of Proposition 2 in Zhang et al. (2017):

Lemma 6 ***(Bounding the Cheeger constant)*** *Assume that $\hat{\mathcal{K}} \subseteq \mathcal{K}'$ is convex, and that F is convex and λ-Lipschitz on $\hat{\mathcal{K}}_{r'}$. Then for every $\varepsilon > 0$ and all $\xi \geq \frac{4d \log(R/\min(\frac{\varepsilon}{2\lambda}, r'))}{\varepsilon}$ we have*

$$\mathcal{C}_{\xi F}^{\hat{\mathcal{K}}_{r'}}(\hat{\mathcal{K}}_{r'} \backslash U^\varepsilon) \geq \frac{1}{R}.$$

Proof Let $\hat{x}^\star$ be a minimizer of F on $\hat{\mathcal{K}}_{r'}$. Let $\hat{r} = \min(\frac{\varepsilon}{2\lambda}, r')$. Then since $\hat{\mathcal{K}}_{r'} = \hat{\mathcal{K}} + B(0, r')$, for some $a \in \mathcal{K}'$ there is a closed ball $B(a, \hat{r}) \subseteq \hat{\mathcal{K}}_{r'}$, with $x^\star \in B(a, \hat{r})$. By the Lipschitz property, we have

$$\sup\{F(x) : x \in B(a, \hat{r})\} \leq F(x^\star) + 2\hat{r}\lambda \leq F(x^\star) + \frac{\varepsilon}{2}.$$

Therefore,

$$\inf\left\{ \frac{e^{-\xi F(x)}}{e^{-\xi F(y)}} \ : \ x \in B(a, \hat{r}), y \in \mathcal{K}' \backslash U^\varepsilon \right\} \geq e^{\xi\varepsilon/2}. \tag{7}$$

Then Equation (7) implies that

$$\frac{\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(\hat{\mathcal{K}}_{r'} \backslash U_\varepsilon)}{\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(U_\varepsilon)} \leq e^{-\xi\varepsilon/2} \frac{\mathrm{Vol}(B(c, R))}{\mathrm{Vol}(B(a, \hat{r}))}$$

$$
\begin{aligned}
&= e^{-\xi\varepsilon/2}(\frac{R}{\hat{r}})^d \\
&= e^{-\xi\varepsilon/2+d\log(R/\hat{r})} \\
&\leq \frac{1}{2},
\end{aligned}
$$

which implies that

$$\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(\hat{\mathcal{K}}_{r'}\backslash U_\varepsilon) \leq \frac{1}{2}. \tag{8}$$

Then by Theorem 2.6 of Lovász and Simonovits (1993) for all $A \subseteq \hat{\mathcal{K}}_{r'}\backslash U^\varepsilon$ for any $0 < \delta < 2R$ we have

$$
\begin{aligned}
\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A_\delta\backslash A) &\geq \frac{2\frac{\delta}{2R}}{1-\frac{\delta}{2R}} \min(\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A), \mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(\hat{\mathcal{K}}_{r'}\backslash A_\delta)) \\
&= \frac{2\frac{\delta}{2R}}{1-\frac{\delta}{2R}} \min(\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A), 1-\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A_\delta)) \\
&= \frac{2\frac{\delta}{2R}}{1-\frac{\delta}{2R}} \min(\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A), 1-\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A_\delta\backslash A)-\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A)) \\
&\overset{\text{Eq. (8)}}{\geq} \frac{2\frac{\delta}{2R}}{1-\frac{1}{2R}} \min(\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A), 1-\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A_\delta\backslash A)-\frac{1}{2}) \\
&= \frac{2\frac{\delta}{2R}}{1-\frac{\delta}{2R}} \min(\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A), \frac{1}{2}-\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A_\delta\backslash A)) \\
&= \frac{2\frac{\delta}{2R}}{1-\frac{\delta}{2R}} \mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A),
\end{aligned}
$$

provided that $0 < \delta < \Delta_A$ for some small enough value $\Delta_A > 0$ that depends on A. Therefore for every $A \subseteq \hat{\mathcal{K}}_{r'}\backslash U_\varepsilon$ there exists $\Delta_A > 0$ such that

$$\frac{\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A_\delta\backslash A)}{\delta\mu_{\xi F}^{\hat{\mathcal{K}}_{r'}}(A)} \geq \frac{2\frac{1}{2R}}{1-\frac{\delta}{2R}} \qquad \forall 0 < \delta < \Delta_A.$$

Taking $\delta \to 0$, we get

$$\mathcal{C}_{\xi F}^{\hat{\mathcal{K}}_{r'}}(\hat{\mathcal{K}}_{r'}\backslash U^\varepsilon) \geq \frac{1}{R}.$$

■

C.3. Bounding the escape probability

We will use the Lemma proved in this section (Lemma 10) to show that the SGLD chain X defined in Algorithm 1 does not drift too far from its initial objective function value with high probability.

This will allow us to bound the noise, since the noise is proportional to the objective function F. The organization of this section is as follows: we first define a "toy" algorithm and an associated Markov chain Y that will allow us to prove Lemma 10 (and which we will later use to prove Theorem 2). We then prove Propositions 8 and 9, and Lemma 10. Proposition 8 is used to prove Proposition 9, which in turn is used to prove Lemma 10.

We begin by recalling the Metropolis-adjusted version of Algorithm 1 defined in Zhang et al. (2017), which defines a Markov chain $Y_0, Y_1, \ldots$ with stationary distribution $\mu_{\xi\tilde{f}}^{\mathcal{K}}$. Note that this is a"toy" algorithm which is not meant to be implemented; rather we state this algorithm only to define the Markov chain $Y_0, Y_1, \ldots$, which we will use as a tool to prove Lemma 10 and Theorem 2.

Algorithm 3 Lazy Metropolis-adjusted SGLD

input: Convex constraint set $\hat{\mathcal{K}} \subseteq \mathbb{R}^d$, inverse temperature $\xi > 0$, step size $\eta > 0$, parameters $i_{\max} \in \mathbb{N}$ and $D > 0$, stochastic gradient oracle g for some $\tilde{f} : \mathcal{K} \to \mathbb{R}$,
input: Initial point $Y_0 \in \mathbb{R}^d$.
for $i = 0$ *to* $i_{\max}$ **do**
1. Sample $P_i \sim N(0, I_d)$.
2. Set $Y'_{i+1} = Y_i - \eta g(Y_i) + \sqrt{\frac{2\eta}{\xi}} P_i$.
3. Set $Y''_{i+1} = Y'_{i+1}$ if $Y'_{i+1} \in \hat{\mathcal{K}} \cap B(Y_i, D)$. Otherwise, set $Y''_{i+1} = Y_i$.
4. Set $Y'''_{i+1} = Y''_{i+1}$ with probability $\min\left(1, \frac{\mathbb{E}[e^{-\frac{1}{4\eta}\|Y_i - Y''_{i+1} + \eta g(Y''_{i+1})\|}]}{\mathbb{E}[e^{-\frac{1}{4\eta}\|Y''_{i+1} - Y_i + \eta g(Y_i)\|}]} e^{\tilde{f}(Y_i) - \tilde{f}(Y''_{i+1})}\right)$. Otherwise, set $Y'''_{i+1} = Y_i$.
5. Set $V_i = 1$ with probability $\frac{1}{2}$ and set $V_i = 0$ otherwise. Let $Y_{i+1} = Y'''_{i+1}$ if $V_i = 1$; otherwise, let $Y_{i+1} = Y_i$.

end
output: $x_{i^\star}$, where $i^\star := \operatorname{argmin}_i \{\tilde{f}(x_i)\}$.

We now define a coupling of three Markov chains. We will use this coupling to prove Lemma 10 and Theorem 2.

Definition 7 ***(Coupled Markov chains)*** *Let X and $\hat{X}$ be Markov chains generated by Algorithm 1 with constraint set $\mathcal{K}$ and $\hat{\mathcal{K}}_{r'}$, respectively, where $\hat{\mathcal{K}} \subseteq \mathcal{K}$ and $\hat{\mathcal{K}}_{r'} = \hat{\mathcal{K}} + B(0, r')$. Let Y be the Markov chain generated by Algorithm 3. We define a coupling of the Markov chains X, $\hat{X}$ and Y in the following way: Define recursively, $t(0) = 0$,*

$$t(i+1) = \min\{j \in \mathbb{N} : j > i, V_j = 1\}.$$

Let $Q_0, Q_1, \ldots \sim N(0, I_d)$ be i.i.d. Let $X_0 = Y_0 = \hat{X}_0$. Let Y_i be the chain in Algorithm 3 generated by setting $P_i = Q_i$ for all $i \geq 0$ with constraint set $\mathcal{K}$. Let X be the chain in Algorithm 3 generated by setting $P_i = Q_{t(i)}$ for all $i \geq 0$ with constraint set $\mathcal{K}$. Let $\hat{X}$ be the chain in Algorithm 3 generated by setting $P_i = Q_{t(i)}$ for all $i \geq 0$ with constraint set $\hat{\mathcal{K}}_{r'}$.

We now bound the escape probability of the Markov chain Y from a sub-level set of a given height, assuming that it is initialized from its stationary distribution conditioned on a small ball.

Proposition 8 ***(Escape probability from stationary distribution on a small ball)*** *Let $r > 0$ be such that $r' \geq r > 0$ and let $\xi > 0$. Let $Y_0, Y_1, \ldots$ be the Markov chain defined in Algorithm 3 with stationary distribution $\pi = \mu_{\xi\tilde{f}}^{\mathcal{K}}$, and let Y_0 be sampled from $\pi_0 := \mu_{\xi\tilde{f}}^{B(y,r)\cap\mathcal{K}}$, where π_0 is the distribution of π conditioned on $B(y,r)\cap\mathcal{K}$ for some $y \in \mathcal{K}$. Then for every $i \geq 0$ we have*

$$\mathbb{P}(\tilde{f}(Y_i) \geq h) \leq e^{\xi[\tilde{f}(y)+\tilde{\lambda}r]-\xi h+d\log(\frac{2R}{r})} \qquad \forall h \geq 0.$$

Proof Fix $h \geq 0$. Define $S_1 := B(y,r)\cap\mathcal{K}$ and $S_2 := \{x \in \mathcal{K} : \tilde{f}(x) \geq h\}$. Let $c_\pi = \int_{\mathcal{K}} e^{-\xi\tilde{f}(x)}\mathrm{d}x$ be the normalizing constant of π. Since π is the stationary distribution of Y,

$$\mathbb{P}(Y_i \in S_2) \leq \frac{\pi(S_2)}{\pi(S_1)} \qquad \forall\, i \in \{0, \ldots, i_{\max}\} \tag{9}$$

We can see why Inequality (9) is true by the following argument: Let Z be a copy of the Y chain started at stationarity. Let $\mathcal{E}$ be the event that $Z_0 \in S_1$. Then $Z_0|\mathcal{E}$ (Z_0 conditioned on the event $\mathcal{E}$) has the same distribution as $Y_0 \sim \pi_0$. Therefore, $Z_i|\mathcal{E}$ has the same distribution as Y_i (since the Z and Y chains have the same transition kernel). Therefore, $\mathbb{P}(Z_i \in S_2|\mathcal{E}) = \mathbb{P}(Y_i \in S_2)$. Hence,

$$\pi(S_2) = \mathbb{P}(Z_i \in S_2) \geq \mathbb{P}(\{Z_i \in S_2\}\cap\mathcal{E}) = \mathbb{P}(Z_i \in S_2|\mathcal{E})\mathbb{P}(\mathcal{E}) = \mathbb{P}(Y_i \in S_2)\mathbb{P}(Z_0 \in S_1) = \mathbb{P}(Y_i \in S_2)\pi(S_1),$$

which implies Inequality (9).

But $\|\nabla\xi\tilde{f}\| = \|\xi\nabla\tilde{f}\| \leq \xi\tilde{\lambda}$, implying that

$$\begin{aligned}\pi(S_1) = \pi(B(y,r)) &\geq c_\pi e^{-[\xi\tilde{f}(y)+\xi\tilde{\lambda}r]} \times \mathrm{Vol}(B(y,r)\cap\mathcal{K}) \\ &\geq c_\pi e^{-\xi[\tilde{f}(y)+\tilde{\lambda}r]} \times \mathrm{Vol}(B(0,\frac{1}{2}r)),\end{aligned} \tag{10}$$

since $B(y,r)\cap\mathcal{K}$ contains a ball of radius $\frac{1}{2}r$ because $r \leq r'$. Also,

$$\pi(S_2) = \pi(\{x : \tilde{f}(x) \geq h\}) \leq c_\pi e^{-\xi h}\mathrm{Vol}(\mathcal{K}) \leq c_\pi e^{-\xi h}\mathrm{Vol}(B(0,R)). \tag{11}$$

Therefore,

$$\begin{aligned}\mathbb{P}(Y_i \in S_2) &\overset{\text{Eq. (9)}}{\leq} \frac{\pi(S_2)}{\pi(S_1)} \\ &\overset{\text{Eq. (10), (11)}}{\leq} e^{\xi[\tilde{f}(y)+\tilde{\lambda}r]-\xi h} \times \left(\frac{R}{\frac{1}{2}r}\right)^d \\ &= e^{\xi[\tilde{f}(y)+\tilde{\lambda}r]-\xi h+d\log(\frac{2R}{r})}.\end{aligned}$$

■

We now extend our bound for the escape probability of the Markov chain Y (Proposition 8) to the case where Y is instead initialized from the *uniform* distribution on a small ball:

Proposition 9 ***(Escape probability from uniform distribution on a small ball)*** *Let $r > 0$ be such that $r' \geq r > 0$ and let $\xi > 0$. Let ν_0 be the uniform distribution on $B(y,r)\cap\mathcal{K}$ for some $y \in \mathcal{K}$.*

Let $Y_0, Y_1, \ldots$ be the Markov chain defined in Algorithm 3 with stationary distribution $\pi = \mu^{\mathcal{K}}_{\xi\tilde{f}}$, and let Y_0 be sampled from ν_0. Then for every $i \geq 0$ we have

$$\mathbb{P}(\tilde{f}(Y_i) \geq h) \leq e^{\xi[\tilde{f}(y)+\tilde{\lambda}r]-\xi h+d\log(\frac{2R}{r})} + 2r\tilde{\lambda}\xi \qquad \forall h \geq 0. \tag{12}$$

Moreover, for every $A \subseteq \mathcal{K}$, we have

$$\nu_0(A) \leq e^{2R\tilde{\lambda}\xi+d\log(\frac{2R}{r})}\pi(A). \tag{13}$$

Proof Since $\|\nabla\xi\tilde{f}(x)\| \leq \xi\tilde{\lambda}$,

$$\sup_{x\in B(y,r)\cap\mathcal{K}} \xi\tilde{f}(x) - \inf_{x\in B(y,r)\cap\mathcal{K}} \xi\tilde{f}(x) \leq 2r\tilde{\lambda}\xi,$$

and hence

$$\frac{\inf_{x\in B(y,r)\cap\mathcal{K}} \pi(x)}{\sup_{x\in B(y,r)\cap\mathcal{K}} \pi(x)} \geq e^{-2r\tilde{\lambda}\xi}. \tag{14}$$

Define $\pi_0 := \mu^{B(y,r)\cap\mathcal{K}}_{\xi\tilde{f}}$ to be the distribution of π conditioned on $B(y,r)\cap\hat{\mathcal{K}}'$. Let Z be sampled from the distribution π_0. Let $Z' = Y_0$ with probability $\min(\frac{\pi_0(Y_0)}{\nu_0(Y_0)}, 1)$; otherwise let $Z' = Z$. Then Z' has distribution π_0. Moreover, by Equation (14), $Z' = Y_0$ with probability at least $e^{-2r\tilde{\lambda}\xi}$. Therefore, by Proposition 8

$$\begin{aligned}\mathbb{P}(\tilde{f}(Y_i) \geq h) &\leq e^{\xi[\tilde{f}(y)+\tilde{\lambda}r]-\xi h+d\log(\frac{2R}{r})} + 1 - e^{-2r\tilde{\lambda}\xi} \\ &\leq e^{\xi[\tilde{f}(y)+\tilde{\lambda}r]-\xi h+d\log(\frac{2R}{r})} + 2r\tilde{\lambda}\xi \qquad \forall h \geq 0.\end{aligned}$$

This proves Equation (12). Now, since $\|\xi\nabla\tilde{f}(x)\| \leq \xi\tilde{\lambda}$ and $\mathcal{K} \subseteq B(c,R)$,

$$\sup_{x\in\mathcal{K}} \xi\tilde{f}(x) - \inf_{x\in\mathcal{K}} \xi\tilde{f}(x) \leq 2R\tilde{\lambda}\xi,$$

implying that

$$\frac{\inf_{x\in\mathcal{K}} \pi(x)}{\sup_{x\in\mathcal{K}} \pi(x)} \geq e^{-2R\tilde{\lambda}\xi}. \tag{15}$$

Therefore, for every $z \in \mathcal{K}$ we have

$$\begin{aligned}\frac{\pi(z)}{\nu_0(z)} &= \mathrm{Vol}(B(y,r)\cap\mathcal{K}) \times \pi(z) \\ &\geq \mathrm{Vol}(B(0,\frac{1}{2}r)) \times \pi(z) \\ &\geq \mathrm{Vol}(B(0,\frac{1}{2}r)) \times \frac{1}{\mathrm{Vol}(B(0,2R))} \frac{\inf_{x\in\mathcal{K}} \pi(x)}{\sup_{x\in\mathcal{K}} \pi(x)} \\ &\overset{\text{Eq. (15)}}{\geq} \left(\frac{2R}{r}\right)^{-d} e^{-2R\tilde{\lambda}\xi}\end{aligned} \tag{16}$$

$$= e^{-2R\tilde{\lambda}\xi - d\log(\frac{2R}{r})}.$$

Where the second inequality holds since $r \leq r'$. This proves Equation (13). ■

We are now ready to bound the escape probability of the SGLD Markov chain X defined in Algorithm 1 when it is initialized from the uniform distribution on a small ball:

Lemma 10 ***(Escape probability for unadjusted SGLD chain)*** *Let $r > 0$ be such that $r' \geq r > 0$ and let $\xi > 0$. Let ν_0 be the uniform distribution on $B(y, r) \cap \mathcal{K}$ for some $y \in \mathcal{K}$, and let X_0 be sampled from ν_0. Let $X_0, X_1, \ldots$ be the Markov chain generated by Algorithm 1 with constraint set $\mathcal{K}$. Let $\delta \leq \frac{1}{4}$ and let $0 < \eta \leq \frac{\delta}{i_{\max} \times 16d(G^2+L)}$ then*

$$\mathbb{P}(\tilde{f}(X_i) \geq h) \leq e^{\xi[\tilde{f}(y)+\tilde{\lambda}r] - \xi h + d\log(\frac{2R}{r})} + 2r\tilde{\lambda}\xi + \delta \qquad \forall h \geq 0.$$

Proof

Let $Y_0, Y_1, \ldots$ be the Markov chain generated by Algorithm 3, and let $X_0, X_1, \ldots$ be the Markov chain defined in Algorithm 1, where both chains have constraint set $\mathcal{K}$. Couple the Markov chains X and Y as in Definition 7. By Claim 2 in the proof of Lemma 13 of Zhang et al. (2017), for each $i \geq 0$ the rejection probability $\mathbb{P}(Y_{i+1} = Y_i)$ is bounded above by $1 - e^{-16\eta d(G^2+L)} \leq 1 - e^{-\frac{\delta}{i_{\max}}} \leq \frac{\delta}{i_{\max}}$. Hence, for all $0 \leq i \leq i_{\max}$ we have

$$\mathbb{P}(X_j = Y_j \ \forall 0 \leq j \leq i) \geq (1 - \frac{\delta}{i_{\max}})^i \geq 1 - \delta. \tag{17}$$

Thus,

$$\begin{aligned}
\mathbb{P}(\tilde{f}(X_i) \geq h) &\leq \mathbb{P}(\tilde{f}(Y_i) \geq h) + \mathbb{P}(X_j \neq Y_j \text{ for some } 0 \leq j \leq i) \\
&\overset{\text{Eq. (17)}}{\leq} \mathbb{P}(\tilde{f}(Y_i) \geq h) + \delta \qquad \forall h \geq 0 \\
&\overset{\text{Proposition 9}}{\leq} e^{\xi[\tilde{f}(y)+\tilde{\lambda}r] - \xi h + d\log(\frac{2R}{r})} + 2r\tilde{\lambda}\xi + \delta \qquad \forall h \geq 0.
\end{aligned}$$

■

C.4. Comparing noisy functions

In this section we bound the ratio of $\hat{F}$ to $\tilde{f}$. We use this bound to prove Theorem 2 in Section C.5.

Lemma 11 ***(Bounding the ratio of two noisy objective functions)*** *Fix $x \in \mathcal{K}$ and let $t \geq 5\beta$. Define $\hat{H} = \max\{\tilde{f}(x), t\}$ and let $\hat{J} = \max\{\hat{F}(x), t\}$. Then,*

$$\frac{1}{5}\hat{H} \leq \hat{J} \leq 5\hat{H}.$$

Proof By our assumption in Equation (6), we have that

$$|F(x) - \hat{F}(x)| \leq \hat{\alpha}F(x) + \hat{\beta} \leq \alpha F(x) + \beta.$$

Since $\alpha < \frac{1}{2}$, we have,

$$F(x) \leq 2\hat{F}(x) + 2\beta. \tag{18}$$

We also have that,

$$|\tilde{f}(x) - F(x)| \leq \alpha F(x) + \beta$$

implying that

$$\tilde{f}(x) \leq 4F(x) + \beta. \tag{19}$$

Therefore, combining Equations (18) and (19), we have

$$\tilde{f}(x) \leq 4\hat{F}(x) + 5\beta, \tag{20}$$

implying that

$$\max(\tilde{f}(x), 5\beta) \leq \max(4\hat{F}(x) + 5\beta, 20\beta).$$

Thus,

$$\max(\tilde{f}(x), 5\beta) \leq 5\max(\hat{F}(x), 5\beta).$$

Thus, we have $\hat{H} \leq 5\hat{J}$. By a similar argument as above, we can also show that $\hat{J} \leq 5\hat{H}$. ■

C.5. Bounding the error and running time: The smooth case

In this section we will show how to bound the error and running time of Algorithm 1, if we assume that we have access to a stochastic gradient oracle g for a smooth noisy function $\tilde{f}$, which approximates the convex function F. In particular, we do *not* assume access to the smooth function $\tilde{f}$ itself, only to g. We also assume access to a non-smooth oracle $\hat{F}$, which we use to determine the temperature parameter for our Markov chain based on the value of $\hat{F}(X_k^0)$ at the beginning of each epoch. To prove the running time and error bounds, we will use the results of Sections C.2 and C.3.

Recall that in this section α and β refer exclusively to the multiplicative and additive noise levels of $\tilde{f}$. We must first define parameters that will be needed to formally state and prove our error and running time bounds:

- Fix $0 \leq \varepsilon < \frac{1}{25}$ and $\delta > 0$.
- Set parameters of Algorithms 1 and 2 as follows:
  - Let $y_0 \in \mathcal{K}$ and let $H_0 := \tilde{f}(y_0)$.
  - Fix $\mathfrak{D} \geq \frac{1}{\varepsilon}\beta$. For every $0 \leq k \leq k_{\max}$, let $H_k := \tilde{f}(x_k)$ and define $\hat{H}_k := \max(H_k, \mathfrak{D})$.

- Assume, without loss of generality, that $r' \leq \frac{\mathfrak{D}}{\lambda}$. [4]
- For every $0 \leq k \leq k_{\max}$, let $J_k := \hat{F}(x_k)$. Define $\hat{J}_k := \max(J_k, \mathfrak{D})$.
- Set the number of epochs to be $k_{\max} = \lceil \frac{\log(5J_0/\mathfrak{D})}{\log(\frac{1}{25\varepsilon})} \rceil + 1$.
- At every $k \geq 0$, set the temperature to be $\xi_k = \frac{4d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}{\frac{1}{5}\varepsilon\hat{J}_k}$. Define $\bar{\xi} := \frac{4d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}{\frac{1}{25}\varepsilon\mathfrak{D}}$.
- Set $r = \frac{\delta}{\bar{\xi}\tilde{\lambda}}$.
- Define

$$\bar{\eta}^\dagger := c\min\left\{\zeta_{\max}, d\frac{\omega^2}{\lambda^2}, \frac{b_{\max}^2}{d}, \frac{1}{Rd^3((\bar{\xi}G)^2 + \bar{\xi}L)^2}\right\}$$

  and

$$\mathfrak{B}' := \frac{\left(d\log(2\frac{R}{r}) + \delta + 1 + \log(\frac{1}{\delta})\right))}{2d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}.$$

- Set the number of steps $i_{\max}$ for which we run the the Markov chain X in each epoch to be

$$i_{\max} = \left\lceil \left(\frac{8R\tilde{\lambda}\xi_k + 4d(1 + \log(1+\bar{\xi}) + \log(\frac{2R\tilde{\lambda}}{\delta})) + 4\log(\frac{1}{\delta})}{\left(\frac{1}{1536R}\sqrt{\bar{\eta}^\dagger/d}e^{-\frac{150d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}'+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}\right)^2} \right)^{\frac{1}{1-\frac{150}{\varepsilon}\alpha}} \right\rceil + 1.$$

- Define $\mathfrak{B} := \frac{\left(d\log(2\frac{R}{r}) + \delta + \log(i_{\max}+1) + \log(\frac{1}{\delta})\right))}{2d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}$.
- For every $\xi > 0$ define

$$\eta(\xi) := c\min\left\{\zeta_{\max}, d\frac{\omega^2}{\lambda^2}, \frac{b_{\max}^2}{d}, \frac{(e^{-\frac{100d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))})^2}{Rd^3((\xi G)^2 + \xi L)^2}\right\},$$

  where $\omega = \varepsilon\mathfrak{D}$, and c is the universal constant in Lemma 15 of Zhang et al. (2017). Set the step size at each epoch to be $\eta_k = \eta(\xi_k)$. Also define $\bar{\eta} = \eta(\bar{\xi})$.
- Set $D = \sqrt{2\bar{\eta}d}$.

We now state the error and running time bounds:

Theorem 12 ***(Error and running time bounds when using a smooth noisy objective function)*** *Assume that $\alpha \leq \frac{\varepsilon}{32}$. Then with probability at least $1 - 6\delta(k_{\max}+1)$ Algorithm 2 returns a point $\hat{x} = x_{k_{\max}}$ such that*

$$F(\hat{x}) - F(x^\star) \leq \frac{1}{1-\alpha}(\mathfrak{D} + \beta),$$

4. This is without loss of generality since if there exists a convex body $\mathcal{K}'$ such that $\mathcal{K}' + B(0, r') = \mathcal{K}$, then for every $0 < \rho \leq r'$ there must also exist a convex body $\mathcal{K}''$ such that $\mathcal{K}'' + B(0, \rho) = \mathcal{K}$, namely $\mathcal{K}'' = \mathcal{K}' + B(0, r' - \rho)$.

with running time that is polynomial in d, $e^{\frac{d}{\varepsilon/150}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}'+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D},r'))}$, R, λ, $\tilde{\lambda}$, L, G, $\zeta_{\max}$, $b_{\max}$, *and* $\log(\frac{1}{\delta})$.

Proof Set notation as in Algorithms 1 and 2. Denote by $X^{(k)}$ the Markov chain generated by Algorithm 1 as a subroutine in the k'th epoch of Algorithm 2 with constraint set $\mathcal{K}$.

Set $h_k = \hat{H}_k + \xi_k^{-1}\left(d\log(\frac{2R}{r}) + \delta + \log(i_{\max}+1) + \log(\frac{1}{\delta})\right)$. Then by Lemma 10

$$\begin{aligned}\mathbb{P}(\sup_{0\le i\le i_{\max}} \tilde{f}(X_i^{(k)}) \ge h_k) &\le (i_{\max}+1)\times[e^{\xi_k[\hat{H}_k+\tilde{\lambda}r]-\xi_k h_k + d\log(\frac{2R}{r})} + 2r\tilde{\lambda}\xi_k + \delta] \qquad (21)\\ &\le e^{\xi_k\hat{H}_k+\delta-\xi_k h_k+d\log(\frac{2R}{r})+\log(i_{\max}+1)} + 4\delta]\\ &= 5\delta,\end{aligned}$$

where the second inequality holds since $r = \frac{\delta}{\bar{\xi}\tilde{\lambda}}$ and $\xi_k \le \bar{\xi}$ for all k.

But $\tilde{f}(X_i^{(k)}) \ge h_k$ if and only if

$$F(X_i^{(k)})(1+\psi(X_i^{(k)})) + \varphi(X_i^{(k)}) \ge h_k$$

if and only if

$$F(X_i^{(k)}) \ge \frac{1}{1+\psi(X_i^{(k)})}(h_k - \varphi(X_i^{(k)})),$$

since $1+\psi(X_i^{(k)}) \ge 0$. Also,

$$\frac{1}{1+\psi(X_i^{(k)})}(h_k-\varphi(X_i^{(k)})) \le \frac{1}{1-\alpha^\dagger}(h_k+\beta),$$

since $\psi(X_i^{(k)}) \ge -\alpha^\dagger > -1$ and $|\varphi(X_i^{(k)})| < \beta$.
Hence,

$$\begin{aligned}5\delta &\overset{\text{Eq. (21)}}{\ge} \mathbb{P}\left(\sup_{0\le i\le i_{\max}} \tilde{f}(X_i^{(k)}) \ge h_k\right) \qquad (22)\\ &\ge \mathbb{P}\left(\sup_{0\le i\le i_{\max}} F(X_i^{(k)}) \ge \frac{1}{1-\alpha^\dagger}(h_k+\beta)\right).\end{aligned}$$

Define $\hat{\mathcal{K}}^{(k)} := (\mathcal{K}' \cap \{x\in\mathbb{R}^d : F(x) \le \frac{1}{1-\alpha^\dagger}(h_k+\beta)+\lambda r'\}) + B(0,r')$. Then

$$\left\{x\in\mathcal{K} : F(x) \le \frac{1}{1-\alpha^\dagger}(h_k+\beta)\right\} \subseteq \hat{\mathcal{K}}^{(k)}, \qquad (23)$$

since $\|\nabla F\| \le \lambda$. Thus, by Equations (22) and (23),

$$\mathbb{P}\left(X_i^{(k)} \in \hat{\mathcal{K}}^{(k)}\ \forall 0\le i\le i_{\max}\right) \ge 1-5\delta. \qquad (24)$$

Also, for every $x \in \hat{\mathcal{K}}^{(k)}$, since $r' \le \frac{\mathfrak{D}}{\lambda}$, we have

$$\begin{aligned} F(x) &\le \frac{1}{1-\alpha^\dagger}(h_k+\beta) + 2\lambda r' \\ &\le \frac{1}{1-\alpha^\dagger}(h_k+\beta) + 2\mathfrak{D} \\ &= \frac{1}{1-\alpha^\dagger}\left(\hat{H}_k + \xi_k^{-1}\left(d\log(4\frac{R}{r}) + \delta + \log(i_{\max}+1) + \log(\frac{1}{\delta})\right) + \beta\right) + 2\mathfrak{D} \\ &= \frac{1}{1-\alpha^\dagger}\left(\hat{H}_k + \frac{1}{5}\varepsilon\hat{J}_k\frac{\left(d\log(4\frac{R}{r}) + \delta + \log(i_{\max}+1) + \log(\frac{1}{\delta})\right))}{2d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))} + \beta\right) + 2\mathfrak{D} \\ &\overset{\text{Lemma 11}}{\le} \frac{1}{1-\alpha^\dagger}\left(\hat{H}_k + \varepsilon\hat{H}_k\mathfrak{B} + \beta\right) + 2\mathfrak{D}. \end{aligned} \tag{25}$$

Thus, for every $x \in \hat{\mathcal{K}}^{(k)}$,

$$\begin{aligned} |N(x)| &\le \alpha F(x) + \beta \\ &\overset{\text{Eq. (25)}}{\le} \frac{\alpha}{1-\alpha^\dagger}\left(\hat{H}_k(1+\varepsilon\mathfrak{B}) + \beta + 2\mathfrak{D}\right) + \beta := N_k. \end{aligned} \tag{26}$$

Define $U_k^{\varepsilon''} := \{x \in \hat{\mathcal{K}}^{(k)} : F(x) \le \varepsilon''\}$ for every $\varepsilon'' > 0$. Then by Lemma 6, $\mathcal{C}^{\hat{\mathcal{K}}^{(k)}}_{(\xi F)}(\hat{\mathcal{K}}^{(k)} \backslash U_k^{\varepsilon\hat{H}_k}) \ge \frac{1}{R}$ for any $\xi \ge \frac{4d\log(R/\min(\frac{\varepsilon}{2\lambda}\hat{H}_k, r'))}{\varepsilon\hat{H}_k}$.
But by Lemma 11, $\xi_k = \frac{4d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}{\frac{1}{5}\varepsilon\hat{J}_k} \ge \frac{4d\log(R/\min(\frac{\varepsilon}{2\lambda}\hat{H}_k, r'))}{\varepsilon\hat{H}_k}$, implying that

$$\begin{aligned} \mathcal{C}^{\hat{\mathcal{K}}^{(k)}}_{(\xi_k\tilde{f})}(\hat{\mathcal{K}}^{(k)} \backslash U_k^{\varepsilon\hat{H}_k}) &\overset{\text{Eq. (26)}}{\ge} e^{-2\xi_k N_k}\mathcal{C}^{\hat{\mathcal{K}}^{(k)}}_{(\xi_k F)}(\hat{\mathcal{K}}^{(k)} \backslash U_k^{\varepsilon\hat{H}_k}) \\ &\overset{\text{Lemma 6}}{\ge} \frac{1}{R}e^{-2\xi_k N_k} \\ &= \frac{1}{R}e^{-\frac{4d}{\varepsilon}\frac{N_k}{\frac{1}{5}\hat{J}_k}\log(R/\min(\frac{\varepsilon}{2\lambda}\frac{1}{5}\mathfrak{D}, r'))} \\ &\overset{\text{Lemma 11}}{\ge} \frac{1}{R}e^{-\frac{4d}{\varepsilon}\frac{N_k}{\frac{1}{25}\hat{H}_k}\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))} \\ &= \frac{1}{R}e^{-\frac{4d}{\varepsilon}\frac{\frac{\alpha}{1-\alpha^\dagger}\left(\hat{H}_k(1+\varepsilon\mathfrak{B})+2\mathfrak{D}+\beta\right)+\beta}{\frac{1}{25}\hat{H}_k}\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))} \\ &\ge \frac{1}{R}e^{-\frac{100d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}, \end{aligned} \tag{27}$$

where the first inequality holds by the stability property of the Cheeger constant, and the last inequality is true since $\hat{H}_k \ge \mathfrak{D}$ by definition.
Recall that

$$\eta_k = c\min\left\{\zeta_{\max}, d\frac{\omega^2}{\lambda^2}, \frac{b_{\max}^2}{d}, \frac{(e^{-\frac{100d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))})^2}{Rd^3((\xi_k G)^2 + \xi_k L)^2}\right\} \tag{28}$$

$$\overset{\text{Eq. (27)}}{\leq} c \min\left\{\zeta_{\max}, d\frac{\omega^2}{\lambda^2}, \frac{b_{\max}^2}{d}, \frac{(\mathcal{C}^{\hat{\mathcal{K}}^{(k)}}_{(\xi_k\tilde{f})}(\hat{\mathcal{K}}^{(k)} \backslash U_k^{\varepsilon \hat{H}_k}))^2}{d^3((\xi_k G)^2 + \xi_k L)^2}\right\},$$

where $\omega = \varepsilon\mathfrak{D}$.

Recall that $X^{(k)}$ is the subroutine Markov chain described in Algorithm 1 with inputs specified by Algorithm 2 and constraint set $\mathcal{K}$. Let $\hat{X}^{(k)}$ be the Markov chain generated by Algorithm 1 with constraint set $\hat{\mathcal{K}}^{(k)}_{r'}$ and initial point $X_0^{(k)} = \hat{X}_0^{(k)}$. Let $Y^{(k)}$ be the Markov chain generated by Algorithm 3 with constraint set $\hat{\mathcal{K}}^{(k)}_{r'}$. Couple the Markov chains as in definition 7.
Write

$$(U_k^{\varepsilon \hat{H}_k})_{\omega/\lambda} := (U_k^{\varepsilon \hat{H}_k} + B(0, \omega/\lambda)) \cap \hat{\mathcal{K}}^{(k)}$$

as shorthand. Then by Lemma 15 of Zhang et al. (2017) and by Equation (28), the Markov chain $\hat{X}^{(k)}$ is ε'-close to $Y^{(k)}$ with $\varepsilon' \leq \frac{1}{4}\Phi_Y(\hat{\mathcal{K}}^{(k)} \backslash (U_k^{\varepsilon \hat{H}_k})_{\omega/\lambda})$ and

$$\begin{aligned}\Phi_Y(\hat{\mathcal{K}}^{(k)} \backslash (U_k^{\varepsilon \hat{H}_k})_{\omega/\lambda})) &\geq \frac{1}{1536}\sqrt{\eta_k/d}\mathcal{C}^{\hat{\mathcal{K}}^{(k)}}_{(\xi_k\tilde{f})}(\hat{\mathcal{K}}^{(k)} \backslash (U_k^{\varepsilon \hat{H}_k})_{\omega/\lambda}) \\ &\overset{\text{Eq. (27)}}{\geq} \frac{1}{1536R}\sqrt{\eta_k/d}e^{-\frac{100d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}.\end{aligned} \tag{29}$$

Recall that by Equation (13) of Proposition 9, for every $A \subseteq \mathcal{K}'$, we have

$$\nu_0(A) \leq e^{4R\tilde{\lambda}\xi_k + d\log(\frac{2R}{r})}\mu^{\hat{\mathcal{K}}^{(k)}}_{\xi_k\tilde{f}}(A).$$

Therefore, since $\hat{X}^{(k)}$ is ε'-close to $Y^{(k)}$, by Lemma 11 of Zhang et al. (2017), with probability at least $1-\delta$ we have

$$\begin{aligned}\tau_{\hat{X}^{(k)}}((U_k^{\varepsilon \hat{H}_k})_{\omega/\lambda}) &\leq \frac{4\log(e^{2R\tilde{\lambda}\xi_k + d\log(\frac{2R}{r})}/\delta)}{\Phi_Y^2(\hat{\mathcal{K}}^{(k)} \backslash (U_k^{\varepsilon \hat{H}_k})_{\omega/\lambda})} \\ &\overset{\text{Eq. (29)}}{\leq} \frac{8R\tilde{\lambda}\xi_k + 4d\log(\frac{2R}{r}) + 4\log(\frac{1}{\delta})}{\left(\frac{1}{1536R}\sqrt{\eta_k/d}e^{-\frac{100d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}\right)^2} \\ &= \frac{8R\tilde{\lambda}\xi_k + 4d(\log(\bar{\xi}) + \log(\frac{2R\tilde{\lambda}}{\delta})) + 4\log(\frac{1}{\delta})}{\left(\frac{1}{1536R}\sqrt{\eta_k/d}e^{-\frac{100d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}\right)^2} \\ &\leq \frac{8R\tilde{\lambda}\xi_k + 4d(\log(1+\bar{\xi}) + \log(\frac{2R\tilde{\lambda}}{\delta})) + 4\log(\frac{1}{\delta})}{\left(\frac{1}{1536R}\sqrt{\eta_k/d}e^{-\frac{100d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}\right)^2} \\ &\leq \frac{8R\tilde{\lambda}\xi_k + 4d(1+\log(1+\bar{\xi}) + \log(\frac{2R\tilde{\lambda}}{\delta})) + 4\log(\frac{1}{\delta})}{\left(\frac{1}{1536R}\sqrt{\bar{\eta}/d}e^{-\frac{100d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}\right)^2}\end{aligned} \tag{30}$$

$$\leq \frac{8R\tilde{\lambda}\xi_k + 4d(1+\log(1+\bar{\xi})+\log(\frac{2R\tilde{\lambda}}{\delta})) + 4\log(\frac{1}{\delta})}{\left(\frac{1}{1536R}\sqrt{\bar{\eta}^\dagger/de}^{-\frac{150d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}'+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\tilde{\lambda}}\mathfrak{D},r'))-\frac{75}{\varepsilon}\alpha\log(i_{\max}+1)}\right)^2}$$

$$\leq \frac{8R\tilde{\lambda}\xi_k + 4d(1+\log(1+\bar{\xi})+\log(\frac{2R\tilde{\lambda}}{\delta})) + 4\log(\frac{1}{\delta})}{\left(\frac{1}{1536R}\sqrt{\bar{\eta}^\dagger/de}^{-\frac{150d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}'+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\tilde{\lambda}}\mathfrak{D},r'))}\right)^2} \times (i_{\max}+1)^{\frac{150}{\varepsilon}\alpha}$$

$$\leq i_{\max},$$

where the first equality is true since $r = \frac{\delta}{\xi\tilde{\lambda}}$, the fourth inequality is true by the definition of $\bar{\eta}$, the fifth inequality is true by the definition of $\bar{\eta}^\dagger$, and the last inequality is true by our choice of $i_{\max}$.

But by Equation (24), $X_i^{(k)} = \hat{X}_i^{(k)}$ with probability at least $1-5\delta$. Therefore, since Equation (30) holds with probability at least $1-\delta$, we have that

$$\tau_{X^{(k)}}((U_k^{\varepsilon\hat{H}_k})_{\omega/\lambda}) \leq i_{\max}. \tag{31}$$

with probability at least $1-6\delta$.

Therefore, by Equation (31), with probability at least $1-6\delta$ for some $0 \leq i_k^\circ \leq i_{\max}$ we have $X_{i_k^\circ}^{(k)} \in (U_k^{\varepsilon\hat{H}_k})_{\omega/\tilde{\lambda}}$ and hence that

$$F(X_{i_k^\circ}^{(k)}) \leq \varepsilon\hat{H}_k + \tilde{\lambda}\times\frac{\omega}{\tilde{\lambda}} = \varepsilon\hat{H}_k + \varepsilon\mathfrak{D} \leq 2\varepsilon\hat{H}_k$$

and therefore, since $0 \leq \alpha < 1$,

$$\frac{1}{5}\tilde{f}(x_{k+1}) \overset{\text{Lemma 11}}{\leq} \hat{F}(x_{k+1}) = \min_{0\leq i\leq i_{\max}} \hat{F}(X_i^{(k)}) \leq \hat{F}(X_{i_k^\circ}^{(k)}) \leq 2F(X_{i_k^\circ}^{(k)}) + \beta \leq 4\varepsilon\hat{H}_k + \beta \leq 5\varepsilon\hat{H}_k.$$

Hence, for every $0 \leq k \leq k_{\max}$ we have

$$\tilde{f}(x_{k+1}) = H_{k+1} \leq 25\varepsilon\hat{H}_k = 25\varepsilon\max(H_k, \mathfrak{D}) \tag{32}$$

with probability at least $1-6\delta$.

Therefore, by induction on Equation (32), for every $0 \leq k \leq k_{\max}$, we have

$$H_{k+1} \leq 25\varepsilon \times \max\left((25\varepsilon)^k H_0, \mathfrak{D}\right) \tag{33}$$

with probability at least $1-6\delta(k+1)$.

By Lemma 11, we have $k_{\max} = \lceil\frac{\log(5J_0/\mathfrak{D})}{\log(\frac{1}{25\varepsilon})}\rceil + 1 \geq \lceil\frac{\log(H_0/\mathfrak{D})}{\log(\frac{1}{25\varepsilon})}\rceil + 1$. Then, with probability at least $1-6\delta(k_{\max}+1)$,

$$\begin{aligned}\tilde{f}(x_{k_{\max}}) - F(x^\star) &= \tilde{f}(x_{k_{\max}}) \\ &= H_{k_{\max}} \\ &\overset{\text{Eq. (33)}}{\leq} 25\varepsilon \times \max\left((25\varepsilon)^{k_{\max}-1}H_0, \mathfrak{D}\right)\end{aligned} \tag{34}$$

$$\leq 25\varepsilon \times \mathfrak{D}$$
$$\leq \mathfrak{D},$$

since $0 \leq \varepsilon < \frac{1}{25}$ implies that $0 \leq 25\varepsilon < 1$.
Hence,

$$\begin{aligned} F(x_{k_{\max}}) - F(x^\star) &= F(x_{k_{\max}}) \\ &\leq \frac{1}{1-\alpha}(\tilde{f}(x_{k_{\max}}) + \beta) \\ &\leq \frac{1}{1-\alpha}(\mathfrak{D} + \beta), \end{aligned}$$

where the first equality holds since $F(x^\star) = 0$. ■

C.6. The non-smooth case

In this section we bound the gradient, supremum, and smoothness of the smoothed function f_σ obtained from F (Propositions 13 and 14 and Lemma 15), where f_σ is defined in Equation (5). We also bound the noise $|F(x) - f_\sigma(x)|$ of f_σ (Lemma 16). We use these bounds in Section C.8 to Prove our main result (Theorem 3).

Proposition 13 ***(Gradient bound for smoothed oracle)***
For every $x \in \mathcal{K}$ we have

$$\|\nabla \tilde{f}_\sigma(x)\| \leq \frac{\sqrt{2d}}{\sigma}(2\lambda R(1+2\alpha) + 2\beta).$$

Proof

$$\begin{aligned} \|\nabla \tilde{f}_\sigma(x)\| &\leq \mathbb{E}_Z \left[\frac{1}{\sigma^2}\|Z\| \left|\hat{F}(x+Z) - \hat{F}(x)\right|\right] \\ &\leq \mathbb{E}_Z \left[\frac{1}{\sigma^2}\|Z\| \max_{y_1,y_2 \in \mathcal{K}} |\hat{F}(y_2) - \hat{F}(y_1)|\right] \\ &\leq \frac{1}{\sigma} \max_{y_1,y_2 \in \mathcal{K}} |\hat{F}(y_2) - \hat{F}(y_1)| \mathbb{E}_Z \left[\frac{1}{\sigma}\|Z\|\right] \\ &\leq \frac{1}{\sigma} \max_{y_1,y_2 \in \mathcal{K}} |\hat{F}(y_2) - \hat{F}(y_1)| \mathbb{E}_Z \left[\frac{1}{\sigma}\|Z\|\right] \\ &= \frac{1}{\sigma} \max_{y_1,y_2 \in \mathcal{K}} |\hat{F}(y_2) - \hat{F}(y_1)| \sqrt{2}\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} \\ &\leq \frac{\sqrt{2d}}{\sigma} \max_{y_1,y_2 \in \mathcal{K}} |\hat{F}(y_2) - \hat{F}(y_1)| \\ &\leq \frac{\sqrt{2d}}{\sigma}(2\lambda R(1+2\alpha) + 2\beta), \end{aligned}$$

where the equality is true since $\frac{1}{\sigma}\|Z\|$ has χ distribution with d degrees of freedom, and the second-to-last inequality is true since $\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} \leq \sqrt{d}$. The last ineqaulty is true because F is λ-Lipschitz, and because of our assumption on the noise (Equation (4)). ■

Proposition 14 *(maximum value of non-smooth noisy oracle) For every $x \in \mathcal{K}_\mathsf{r}$, we have*

$$\hat{F}(x) \leq (1+\alpha)2\lambda(R+\mathsf{r}) + \beta.$$

Proof Since F is λ-Lipschitz,

$$F(x) \leq 2\lambda(R+\mathsf{r}) \qquad \forall x \in \mathcal{K}_\mathsf{r}. \tag{35}$$

Thus,

$$\hat{F}(x) \leq (1+\alpha)|F(x)| + \beta \leq (1+\alpha)2\lambda(R+\mathsf{r}) + \beta.$$

■

We recall the following Lemma from Zhang et al. (2017):

Lemma 15 ***(Lemma 17 in Zhang et al. (2017))*** *Suppose that $\hat{M} > 0$ is a number such that $0 \leq \hat{F}(x) \leq \hat{M}$ for all $x \in \mathcal{K}_\mathsf{r}$ then*

1. $$\mathbb{E}_Z[g_Z(x)] = \nabla \tilde{f}_\sigma(x) \qquad \forall x \in \mathcal{K}.$$

2. *For every $u \in \mathbb{R}^d$,*
$$\mathbb{E}_Z[e^{\langle u, g_Z(x)\rangle (2\hat{M}/\sigma)^2}] \leq e^{\frac{4\hat{M}^2}{\sigma^2}\|u\|^2}.$$

3. $$\|\nabla^2 \tilde{f}_\sigma(x)\|_{\mathrm{op}} \leq \frac{2\hat{M}}{\sigma^2}.$$

We show that the smoothed gradient is a good approximation of F for sufficiently small σ:

Lemma 16 ***(Noise of smoothed oracle)*** *Let $A \subseteq A_t \subseteq \mathcal{K}_\mathsf{r}$ for some $A \subseteq \mathcal{K}_\mathsf{r}$ and some $t > 0$. Let $H' = \sup_{y \in A} F(y)$. Then*

$$|\tilde{f}_\sigma(x) - F(x)| \leq \lambda\sigma(1+\alpha)\sqrt{d} + H' \times e^{-\frac{t^2/\sigma^2 - d}{8}} + \alpha H' + \beta$$

for every $x \in A$.

Proof Define $\mathsf{N}(x) := \hat{F}(x) - F(x)$. For any function $h : \mathbb{R}^d \to \mathbb{R}$, define

$$\tilde{h}_\sigma(x) := \mathbb{E}_Z[h(x+Z)],$$

where $Z \sim \mathcal{N}(0, \sigma^2 I_d)$. Then for every $x \in A$ we have,

$$\begin{aligned} |\tilde{f}_\sigma(x) - F(x)| &= |\tilde{F}_\sigma(x) + \tilde{\mathsf{N}}_\sigma(x) - F(x)| \\ &\leq |\tilde{F}_\sigma(x) - F(x)| + |\tilde{\mathsf{N}}_\sigma(x)| \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_Z[|F(x+Z) - F(x)|] + |\mathbb{E}_Z[\mathsf{N}(x+Z)]| \\
&\leq \mathbb{E}_Z[|F(x+Z) - F(x)|] + \mathbb{E}_Z[\alpha(H' + \lambda\|Z\|) + \beta] \\
&\leq \mathbb{E}_Z[\lambda\|Z\|] + H' \times \mathbb{P}(\|Z\| \geq t) + \mathbb{E}_Z[\alpha(H' + \lambda\|Z\|) + \beta] \\
&= \lambda(1+\alpha)\mathbb{E}_Z[\|Z\|] + H' \times \mathbb{P}(\|Z\| \geq t) + \alpha H' + \beta \\
&= \lambda\sigma(1+\alpha)\mathbb{E}_Z[\frac{1}{\sigma}\|Z\|] + H' \times \mathbb{P}(\frac{1}{\sigma}\|Z\| \geq \frac{t}{\sigma}) + \alpha H' + \beta \\
&\leq \lambda\sigma(1+\alpha)\sqrt{d} + H' \times \mathbb{P}(\frac{1}{\sigma}\|Z\| \geq \frac{t}{\sigma}) + \alpha H' + \beta \\
&\leq \lambda\sigma(1+\alpha)\sqrt{d} + H' \times e^{-\frac{t^2/\sigma^2 - d}{8}} + \alpha H' + \beta,
\end{aligned}
$$

where the second inequality holds because F is λ-Lipschitz on $\mathcal{K}_\mathsf{r}$ and also since F is defined to be zero outside $\mathcal{K}_\mathsf{r}$ with $x \in \mathcal{K} \subseteq \mathcal{K}_\mathsf{r}$. The third inequality holds by our assumption on the noise (Equation (4)), and since F is defined to be zero outside $\mathcal{K}_\mathsf{r}$. The fourth inequality holds because $\frac{1}{\sigma}\|z\|$ is χ-distributed with d degrees of freedom. The last inequality holds by the Hanson-Wright inequality (see for instance Hanson and Wright (1971), Rudelson and Vershynin (2013)). ■

C.7. Rounding the domain of the Markov Chain

We now show that our constraint set $\hat{\mathcal{K}}$ is sufficiently "rounded". This roundness property is used to show that the Markov chain does not get stuck for a long time in corners of the constraint set.

Lemma 17 ***(Roundness of constraint set)*** *Let* $\zeta_{\max} = (\frac{r'}{10\sqrt{2}(d+20)})^2$. *Let* $\hat{\mathcal{K}} \subseteq \mathcal{K}'$ *be a convex set. Then for any* $\zeta \leq \zeta_{\max}$ *and any* $x \in \hat{\mathcal{K}}_{r'}$ *the random variable* $W \sim \mathcal{N}(0, I_d)$ *satisfies*

$$\mathbb{P}(\sqrt{2\zeta}W + x \in \hat{\mathcal{K}}_{r'}) \geq \frac{1}{3}.$$

Proof Without loss of generality, we may assume that x is the origin and that $\hat{\mathcal{K}}_{r'}$ contains the ball $B(a, r')$ where $a = (r', 0, \ldots, 0)^\top$ (since $\hat{\mathcal{K}}_{r'} = \hat{\mathcal{K}} + B(0, r')$ implies that there is a ball contained in $\hat{\mathcal{K}}_{r'}$ that also contains x on its boundary. We can then translate and rotate $\hat{\mathcal{K}}_{r'}$ to put x and a in the desired position).

Since $\mathbb{P}(\frac{1}{10} \leq W_1 \leq 100) \geq 0.45$, with probability at least 0.45 we have that

$$\frac{1}{10} \leq W_1 \leq 100$$

but $\zeta_{\max} = (\frac{r'}{10\sqrt{2}(d+20)})^2$, and hence, with probability at least 0.45,

$$\frac{\sqrt{2\zeta}}{r'}(d+20) \leq W_1 \leq \frac{r'}{\sqrt{2\zeta}}.$$

But our choice of $\zeta_{\max}$ implies that $\sqrt{\frac{(r')^2}{2\zeta}} - \frac{1}{\sqrt{\frac{(r')^2}{2\zeta}}}(d+20) > 0$, implying that

$$\frac{r'}{\sqrt{2\zeta}} - \left(\sqrt{\frac{(r')^2}{2\zeta}} - \frac{1}{\sqrt{\frac{(r')^2}{2\zeta}}}(d+20)\right) \leq W_1 \leq \frac{r'}{\sqrt{2\zeta}} + \left(\sqrt{\frac{(r')^2}{2\zeta}} - \frac{1}{\sqrt{\frac{(r')^2}{2\zeta}}}(d+20)\right).$$

But for any $\mathsf{a} > 0$, we have $\sqrt{\mathsf{a}} - \frac{t}{\sqrt{\mathsf{a}}} \leq \sqrt{\mathsf{a} - t}$ for every $t \in [0, \mathsf{a})$, which implies

$$\frac{r'}{\sqrt{2\zeta}} - \sqrt{\frac{(r')^2}{2\zeta} - (d+20)} \leq W_1 \leq \frac{r'}{\sqrt{2\zeta}} + \sqrt{\frac{(r')^2}{2\zeta} - (d+20)}.$$

Therefore

$$r' - \sqrt{(r')^2 - 2\zeta(d+20)} \leq \sqrt{2\zeta} W_1 \leq r' + \sqrt{(r')^2 - 2\zeta(d+20)}.$$

Hence,

$$(\sqrt{2\zeta} W_1 - r')^2 \leq (r')^2 - 2\zeta(d+20),$$

which implies that

$$(\sqrt{2\zeta} W_1 - r')^2 + 2\zeta(d+20) \leq (r')^2. \tag{36}$$

But by the Hanson-Wright inequality

$$\mathbb{P}\left(\sum_{j=2}^{d} W_j^2 \geq d + 20\right) \leq e^{-\frac{21}{8}} < \frac{1}{10}. \tag{37}$$

Thus, Equations (37) and (36) imply that with probability at least $0.45 - \frac{1}{10} \geq \frac{1}{3}$ we have

$$\begin{aligned} \|\sqrt{2\zeta} W - a\| &= \|\sqrt{2\zeta} W - (r', 0, \ldots, 0)^\top\|^2 \\ &= (\sqrt{2\zeta} W_1 - r')^2 + 2\zeta \sum_{j=2}^{d} W_j^2 \\ &\leq (\sqrt{2\zeta} W_1 - r')^2 + 2\zeta(d+20) \\ &\overset{\text{Eq. (36)}}{\leq} (r')^2, \end{aligned}$$

implying that $W \in B(a, r') \subseteq \hat{\mathcal{K}}_{r'}$ with probability at least $\frac{1}{3}$. ■

C.8. Proof of Main Result (Theorem 3)

In this section, we prove Theorem 3. We do so by applying the bounds on the smoothness of f_σ of Section C.6 to Theorem 12.

We note that in this section we will use "α" and "β" exclusively to denote the multiplicative and additive noise levels of F. We will then set the smooth oracle $\tilde{f}$ to be $\tilde{f} = \tilde{f}_\sigma$, where $\tilde{f}_\sigma$ is the smooth function obtained from F, defined in Equation (5). As an intermediate step in proving the main result, we show that $\tilde{f}_\sigma$ has multiplicative noise level 2α and additive noise level 2β.

Proof We will assume that $\alpha < \frac{1}{800}$. This assumption is consistent with the statment of Theorem 3, which assumes that $\alpha = O(1)$.

Define the following constants: $M = 2\lambda R + 2\beta$, $\hat{M} = 6\lambda R + \beta$, $L = \frac{4\hat{M}}{\sigma^2}$, $G = \frac{2\hat{M}}{\sigma}$, $b_{\max} = 1$, $\zeta_{\max} = \left(\frac{r'}{10\sqrt{2}(d+20)}\right)^2$, and $\tilde{\lambda} = \frac{\sqrt{2d}}{\sigma}(2\lambda R(1+2\alpha) + 2\beta)$.

We set $\sigma = \frac{1}{2}\min\left(\frac{\beta}{\lambda(1+\alpha)\sqrt{d}}, \frac{\mathsf{r}}{\sqrt{\log(\frac{1}{\alpha})+d}}\right)$. Recall from Section A.3 that σ determines the amount of smoothness in $\tilde{f}_\sigma$. A larger value of σ means that $\tilde{f}_\sigma$ will be smoother, decreasing the running time of the algorithm. On the other hand, a smaller value of σ means that $\tilde{f}_\sigma$ will be a closer approximation to F, and consequently lead to a lower error. We choose σ in such a way so that the error $\hat{\varepsilon}$ will be bounded by the desired value $\hat{\varepsilon}$.

Set parameters of Algorithms 1 and 2 as follows:

- Fix $\varepsilon = \frac{1}{50}$.
- Let $\mathfrak{D} = \frac{2}{3}\hat{\varepsilon}$.
- Define $J_0 := \hat{F}(x_0)$ and set the number of epochs to be

$$k_{\max} = \left\lceil \frac{\log(5J_0/\mathfrak{D})}{\log(2)} \right\rceil + 1. \tag{38}$$

- For every $0 \le k \le k_{\max}$, let $J_k := \hat{F}(x_k)$, and define $\hat{J}_k := \max(J_k, \mathfrak{D})$.
- Fix $\delta = \frac{\delta'}{6(k_{\max}+1)}$.
- At every $k \ge 0$, set the temperature to be

$$\xi_k = \frac{4d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}{\frac{1}{5}\varepsilon\hat{J}_k}. \tag{39}$$

  Define $\bar{\xi} := \frac{4d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}{\frac{1}{25}\varepsilon\mathfrak{D}}$.
- Set $r = \frac{\delta}{\bar{\xi}\tilde{\lambda}}$.
- Define

$$\bar{\eta}^\dagger := c\min\left\{\zeta_{\max}, d\frac{\omega^2}{\lambda^2}, \frac{b_{\max}^2}{d}, \frac{1}{Rd^3((\bar{\xi}G)^2 + \bar{\xi}L)^2}\right\}$$

  and

$$\mathfrak{B}' := \frac{\left(d\log(2\frac{R}{r}) + \delta + 1 + \log(\frac{1}{\delta})\right))}{2d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}.$$

- Set the number of steps $i_{\max}$ for which we run the the Markov chain X in each epoch to be

$$i_{\max} = \left\lceil \left(\frac{8R\tilde{\lambda}\xi_k + 4d(1+\log(1+\bar{\xi}) + \log(\frac{2R\tilde{\lambda}}{\delta})) + 4\log(\frac{1}{\delta})}{\left(\frac{1}{1536R}\sqrt{\bar{\eta}^\dagger/d}e^{-\frac{150d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}'+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D}, r'))}\right)^2} \right)^{\frac{1}{1-\frac{150}{\varepsilon}\alpha}} \right\rceil + 1. \tag{40}$$

- Define $\mathfrak{B} := \frac{\left(d\log(2\frac{R}{r})+\delta+\log(i_{\max}+1)+\log(\frac{1}{\delta})\right))}{2d\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D},r'))}$.
- For every $\xi > 0$ define

$$\eta(\xi) := c\min\left\{\zeta_{\max}, d\frac{\omega^2}{\lambda^2}, \frac{b_{\max}^2}{d}, \frac{(e^{-\frac{100d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D},r'))})^2}{Rd^3((\xi G)^2+\xi L)^2}\right\}, \quad (41)$$

  where $\omega = \varepsilon\mathfrak{D}$, and c is the universal constant in Lemma 15 of Zhang et al. (2017). We set the step size at each epoch k to be $\eta_k = \eta(\xi_k)$. We also define $\bar{\eta} := \eta(\bar{\xi})$.
- Set $D = \sqrt{2\bar{\eta}d}$.

We determine the constants for which $\tilde{f} = \tilde{f}_\sigma$ satisfies the various assumptions of Theorem 12.

Since $\sigma = \frac{1}{2}\min\left(\frac{\beta}{\lambda(1+\alpha)\sqrt{d}}, \frac{\mathsf{r}}{8\sqrt{\log(\frac{1}{\alpha})+d}}\right)$, by Lemma 16, we have that

$$|\tilde{f}_\sigma(x) - F(x)| \le 2\alpha F(x) + 2\beta \qquad \forall x \in \mathcal{K}.$$

So, with a slight abuse of notation, we may state that $\tilde{f} = \tilde{f}_\sigma$ has multiplicative noise of level 2α and additive noise of level 2β, if we use "α" and "β" to denote the noise levels of $\hat{F}$.

Hence, $M = 2\lambda R + 2\beta \ge \sup_{x\in\mathcal{K}} \tilde{f}_\sigma(x)$. By Lemma 17, part 1 of Assumption 1 is satisfied with constant $\zeta_{\max}$. By Lemma 15 and Proposition 14, $\tilde{f}_\sigma$ satisfies parts 2 and 3 of Assumption 1 with constants L, G and $b_{\max}$, (recall that we defined these constants at the beginning of this proof). By Proposition 13, $\|\nabla\tilde{f}_\sigma(x)\| \le \tilde{\lambda}$ for all $x \in \mathcal{K}$. Therefore, applying Theorem 2 with the above constants and the smoothed function f_σ, we have,

$$F(\hat{x}) - F(x^\star) \le \frac{1}{1-2\alpha}(\mathfrak{D}+2\beta) \le \hat{\varepsilon},$$

with running time that is polynomial in d, $e^{\frac{8d}{\varepsilon}\left[\frac{\alpha}{1-\alpha^\dagger}\left(3+\varepsilon\mathfrak{B}'+\frac{\beta}{\mathfrak{D}}\right)+\frac{\beta}{\mathfrak{D}}\right]\log(R/\min(\frac{\varepsilon}{2\lambda}\mathfrak{D},r'))}$, R, λ, $\tilde{\lambda}$, L, G, $\zeta_{\max}$, $b_{\max}$, and $\log(\frac{1}{\delta})$. This completes the proof of the Theorem. ■

Appendix D. Figures

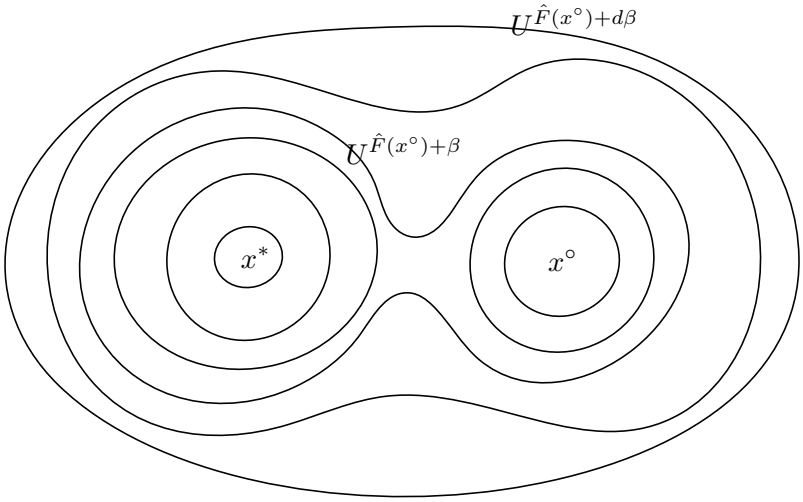

Figure 1: To quickly escape a local minimizer x° of "depth" β, a Markov chain must run at a temperature β. At this temperature, the Markov chain will concentrate in a sub-level set of height $d\beta$. This sub-level set does not have a narrow bottleneck, so a Markov chain running at temperature β will quickly escape the local minimum at x°.

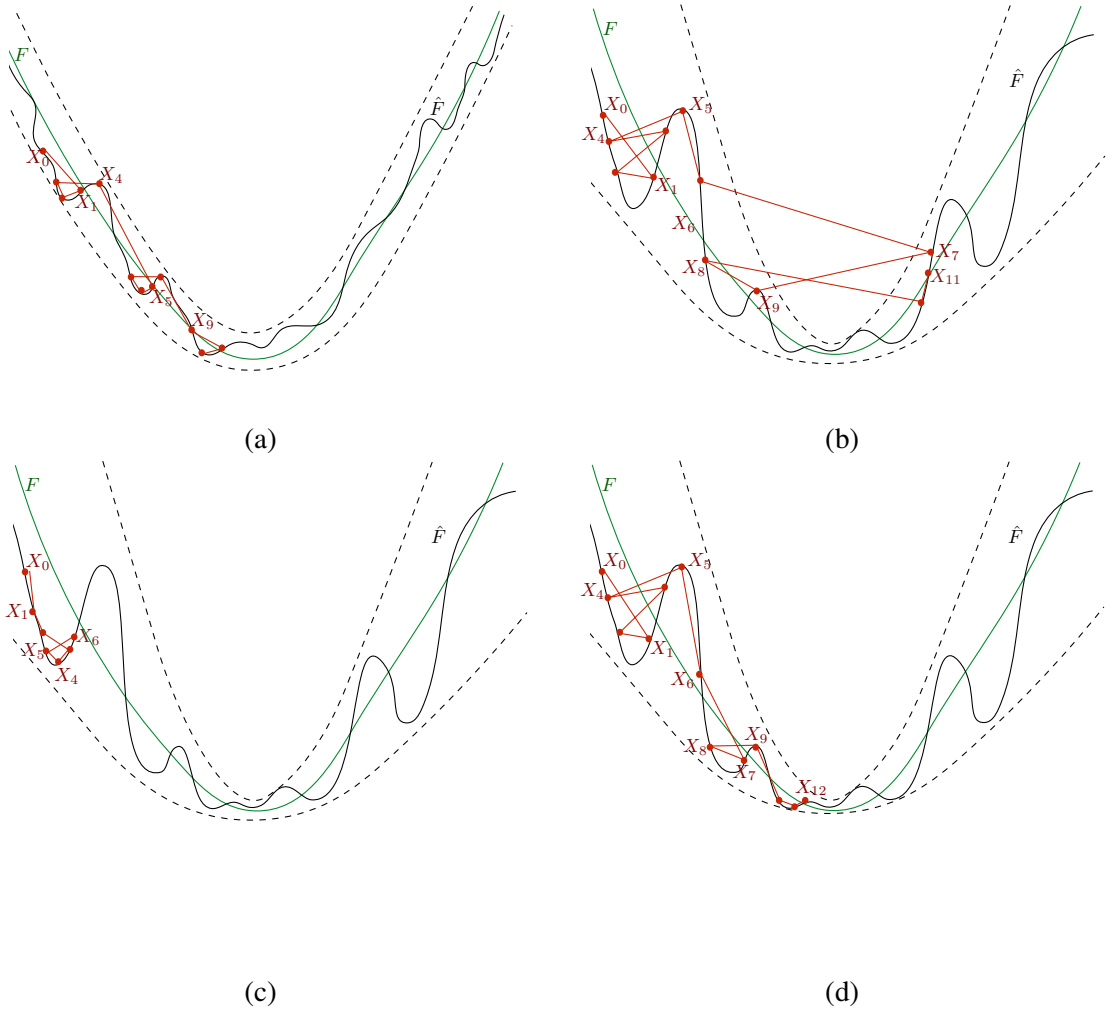

Figure 2: (a) Optimization of a convex function F (green) with noisy oracle $\hat{F}$ (black) under bounded additive noise. Since the gap between the noise bounds (dashed lines) is constant, the Markov chain (red) can be run at a single temperature that is both hot enough to quickly escape any local minimum but also cold enough so that the Markov chain eventually concentrates near the global minimum. (b) and (c) Optimization of a convex function F (green) with noisy oracle $\hat{F}$ (black) when both additive and multiplicative noise are present, if we run the Markov chain at single a fixed temperature. If the temperature is hot enough to escape even the deepest local minima (b), then the Markov chain will not concentrate near the global minimum, leading to a large error. If instead the Markov chain is run at a colder temperature (c), it will take a very long time to escape the deeper local minima. (d) Optimization of a convex function F (green) with noisy oracle $\hat{F}$ (black) under both additive and multiplicative noise, when using a gradually decreasing temperature. If multiplicative noise is present the local minima of $\hat{F}$ are very deep for large values of F. To quickly escape the deeper local minima, the Markov chain is started at a high temperature. As the Markov chain concentrates in regions where F is smaller, the local minima become shallower, so the temperature may be gradually decreased while still allowing the Markov chain to escape nearby local minima. As the temperature is gradually decreased, the Markov chain concentrates in regions with successively smaller values of $\hat{F}$.