ANALYSIS OF STOCHASTIC GRADIENT DESCENT

A PREPRINT

Haoliang Yue
hyue3@illinois.edu

May 10, 2019

ABSTRACT

This paper analyzes stochastic gradient descent (SGD) in two aspects. For the first topic, it mainly discusses the paper by Kuzborskij et al. [1], accepted by international conference on machine learning (ICML) in 2018, about data-dependent stability of SGD. Kuzborskij et al. [1] heavily based the analysis on a previous work by Hardt et al. [2] in 2016, which is overlapped with chapter 10 of the lecture note. Therefore, this paper will combine the knowledge from lecture with literature review to provide and compare several bounds on stability. The second half will talk about why SGD performs well in neural network (NN), to be specific, about when does SGD escape a bad local minima and get close to a desired solution. This part refers to Kleinberg et al. [3] accepted by ICML in 2018 as well.

Keywords Statistical learning theory, Stochastic gradient descent, Neural network

1 Introduction

Stochastic gradient descent (SGD) is widely used in machine learning. Although in most cases, it serves as a faster but less accurate alternative to gradient descent (GD), SGD always finds better solutiosn than GD in training highly complex and non-convex models, such as neural network (NN).

2 Problem Setup

The problem setup is basically the same as in the lecture course, so the notation is rewritten correspondingly. A example space Z represents a set of possible values of data samples, and its member is indicated by $Z \in \mathsf{Z}$. For instance, in a supervised setting $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$, such that X is the input and Y is the output space of a learning problem. P is a probability distribution for Z-valued random variables, where training and testing examples are assumed to be drawn independent and identically distributed (iid) from. In particular, the training set is denoted as $Z^n = \{Z_i\}_{i=1}^n \sim P^n$.

For a parameter space $\mathcal{F} \subseteq \mathbb{R}^d$, which is nonempty and closed, we define a learning algorithm as a map $A : Z^* \to \mathcal{F}$. To measure the accuracy of a learning algorithm A, a loss function $\ell(f, Z)$ measures the cost incurred by predicting with parameters $f \in \mathcal{F}$ on an example Z. The risk of $\mathcal{F}$ with respect to the distribution P and the empirical risk given a training dataset Z^n are defined as $L(f) := E_{Z\sim P}[\ell(f, Z)]$ and $L_n(f) := \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i)$. Finally, define $L^*(f) := \inf_{f\in\mathcal{F}} L(f)$.

For SGD, step sizes $\{\alpha_t\}_{t=1}^T$, random indices $I = \{j_t\}_{t=1}^T$, an initialization point f_1, and a update formula

$$f_{t+1} = f_t - \alpha_t \nabla \ell(f_t, Z_{i_t})$$

for $T \leq n$ steps are needed. Here, the varience of stochastic gradients is assumed to obey

$$\mathbf{E}_{Z^n,Z}[||\nabla \ell(f_{Z^n,t}, Z) - \nabla L(f_{Z^n,t})||^2] \leq \sigma^2$$

, $\forall t \in [T]$. The notation $f_{Z^n,t}$ means the output of SGD ran on a training dataset Z^n at step t, and $[T]$ denotes the enumeration.

The SGD gradient update rule can be regarded as an operate $G_t : \mathcal{F} \to \mathcal{F}$, such that $G_t(f) := f - \alpha_t \nabla \ell(f, Z_{i_t})$. In addition, define $\delta_t(Z^n, Z) := ||f_{Z^n,t} - f_{Z^{n-(i)},t}||$.

3 Stability

In this section, Kuzborskij et al. [1] and Hardt et al. [2] is focused to evaluate the stability of SGD and obtain some bounds that are useful to bound the error.

3.1 Uniform Bound

Before continuing to derive the main result, some preliminaries about stability are necessary to be introduced. Uniform stability is defined according to both Kuzborskij et al. [1] and the lecture note.

Definition 1 (Uniform stability) **.** *A randomized algorithm A is ε-uniformly stable if for all datasets $Z^n, Z^n_{(i)} \in Z^*$ such that Z^n and $Z^n_{(i)}$ differ in the i-th example, we have*

$$\sup_{Z \in \mathsf{Z}, i \in [n]} \mathbf{E}_A[\ell(A(Z^n), Z) - \ell(A(Z^n_{(i)}), Z)] \leq \varepsilon$$

.

As Z^n and $Z^n_{(i)}$ only differ in the i-th example, when outputs of a learning algorithm A on the original Z^n and the modified $Z^n_{(i)}$ are compared, A is said to be stable if this small perturbation in training dataset does not affects its output to much.

Theorem 1. *Let A be ε-uniformly stable. Then,*

$$|\mathbf{E}_{Z^n, A}[L_n(A(Z^n)) - L(A(Z^n))]| \leq \varepsilon$$

.

The above theorem indicates that if an algorithm is stable, it also generalizes from the training dataset to the unseen dataset sampled from the same distribution. In particular, an algorithm is stable on average if and only if it generalizes on averages.

3.2 Data-Dependent Bound

Data-dependent stability states not only the property of the learning algorithm but also the additional parameters, denoted by θ, of the algorithm. In the following section, θ will describe the data-generation distribution and the initialization point of SGD.

Definition 2 (On-Average stability) **.** *A randomized algorithm A is $\varepsilon(\theta)$-on-average stable if it is true that*

$$\sup_{i \in [n]} \mathbf{E}_A \mathbf{E}_{Z^n, Z}[\ell(A(Z^n), Z) - \ell(A(Z^n_{(i)}), Z)] \leq \varepsilon(\theta)$$

, where $Z^n \sim P^n$ and $Z^n_{(i)}$ is its copy with i-th example replaced by $Z \sim P$.

Theorem 2. *Let an algorithm A be ε-on-average stable. Then,*

$$\mathbf{E}_{Z^n} \mathbf{E}_A[L(A(Z^n)) - L_n(A(Z^n))] \leq \varepsilon$$

.

Same as the uniform case, on-average stable algorithm also guarantees to generalize in expectation.

Also, a few conditions are required to derive the main result.

Definition 3 (Lipschitz) **.** *A loss function ℓ is L-Lipschitz if $||\nabla \ell(f, Z)|| \leq L$, $\forall f \in \mathcal{F}$ and $\forall Z \in \mathsf{Z}$. Note that this also implies that $|\ell(f, Z) - \ell(g, Z)| \leq L||f - g||$.*

Definition 4 (Smooth) **.** *A loss function is M-smooth if $\forall f, g \in \mathcal{F}$ and $\forall Z \in \mathsf{Z}$, $||\nabla \ell(f, Z) - \nabla \ell(g, Z)|| \leq M||f - g||$, which also implies $\ell(f, Z) - \ell(g, Z) \leq \nabla \ell(g, Z)^\top (f - g) + \frac{M}{2}||f - g||^2$.*

Definition 4 (Lipschitz Hessian) **.** *A loss function ℓ has ρ-Lipschitz Hessian if $\forall f, g \in \mathcal{F}$ and $\forall Z \in \mathsf{Z}$, $||\nabla^2 \ell(f, Z) - \nabla^2 \ell(g, Z)||_2 \leq \rho||f - g||$.*

Lipschitz Hessian holds whenever ℓ has a bounded third derivative, and it is occasionally used in analysis of SGD.

3.2.1 Convex Loss

Kuzborskij et al. [1] states a data-dependent stability result for convex loss in Theorem 3.

Theorem 3. *Assume that ℓ is convex, and that SGD's step sizes satisfy $\alpha_t = \frac{c}{\sqrt{t}} \leq \frac{1}{M}$, $\forall t \in [T]$. Then SGD is $\epsilon(P, f_1)$-on-average stable with*

$$\epsilon(P, f_1) = \mathcal{O}(\sqrt{c(L(f-1) - L^*)} \cdot \frac{\sqrt[4]{T}}{n} + c\sigma\frac{\sqrt{T}}{n})$$

.

Theorem 4. *Assume that the loss function $\ell(\cdot, Z)$ is M-smooth, convex and L-Lipschitz for every Z. Suppose that we run SGD with step sizes $\alpha_t \leq \frac{2}{M}$ for T steps. Then, SGD satisfies uniform stability with*

$$\varepsilon \leq \frac{2L^2}{n}\sum_{t=1}^{T}\alpha_t$$

.

However, in theorem 4 from Hardt et al. [2] and the lecture note, under the same assumption, it implies a uniform stability bound that $\varepsilon = \mathcal{O}(\frac{\sqrt{T}}{n})$. The first bound differs in that it has a multiplicative risk at the initialization point. Therefore, intuitively, if starting at a good location of the objective function, the algorithm will be more stable and generalizes better. For example, when $L(f_1) = L*$, the theorem confirms that SGD is perfectly stable and does not need any further updates. However, it is only the case when the variance of stochastic gradients is not too large. On the other hand, when the variance is large enough to make the second term dominant, the bound dose not offer any improvement compared to the uniform one. At least, taking the minimum of two bounds can always tighten the result.

3.2.2 Non-Convex Loss

For loss function that is not convex, it is assumed to have a ρ-Lipschitz Hessian.

Theorem 5. *Assume that $\ell(\cdot, Z) \in [0, 1]$ and has a ρ-Lipschitz Hessian, and that step sizes a form $\alpha_t = \frac{c}{t}$ satisfy $c \leq \min(\frac{1}{M}, \frac{1}{4(2M\ln(T))^2})$. Then, SGD is $\varepsilon(P, f_1)$-on-average stable with*

$$\varepsilon(P, f_1) \leq \frac{1 + \frac{1}{c\gamma}}{n}(2cL^2)^{\frac{1}{1+c\gamma}}(\mathbf{E}_{Z^n,A}[L(A(Z^n))] \cdot T)^{\frac{c\gamma}{1+c\gamma}}$$

, where

$$\gamma := \tilde{\mathcal{O}}(\min(M, \mathbf{E}_Z[||\nabla^2\ell(f_1, Z)||_2] + \Delta^*_{1,\sigma^2}))$$

,

$$\Delta^*_{1,\sigma^2} := \rho(c\sigma + \sqrt{c(L(f_1) - L*)})$$

.

This theorem is the first to establish a theoretical link between the curvature of the loss function and the generalization ability of SGD in a data-dependent sense. Stability is controlled by the curvature, the risk of the initialization point, and the variance of the stochastic gradient. In particular, γ characterizes how the curvature at the initialization point affects stability and the generalization error of SGD. Since γ heavily affects the rate of convergence in the inequality, in most situations, smaller γ yields higher stability. Therefore, it suggests that starting from a point in a less curved region with low risk is supposed to yield higher stability and allow faster generalization.

Lemma 1. *Assume that the loss function $\ell(\cdot, Z)$ is M-smooth and that its Hessian is ρ-Lipschitz. Then,*

$$||G_t(f_{Z^n,t} - f_{Z^n_{(i)},t})|| \leq (1 + \alpha_t\xi_t(Z^n, Z))\delta_t(Z^n, Z)$$

where

$$\xi_t(Z^n, Z) := ||\nabla^2\ell(f_1, Z_t)||_2 + \frac{\rho}{2}||\sum_{k=1}^{t-1}\alpha_k\nabla\ell(f_{Z^n,k}, Z_k)|| + \frac{\rho}{2}||\sum_{k=1}^{t-1}\alpha_k\nabla\ell(f_{Z^n_{(i)},k}, Z'_k)||$$

. Furthermore, for any $t \in [T]$,

$$\mathbf{E}_{Z^n,Z}[\xi_t(Z^n, Z)] \leq E_{Z^n,Z}[||\nabla^2\ell(f_1, z_t)||_2 + 2\rho\sqrt{(L(f_1) - L*)c(1 + \ln(T))} + \rho\sigma(\sqrt{2cM} + c(1 + ln(T)))]$$

.

Lemma 2. *Assume that the loss function $\ell(\cdot, Z) \in [0,1]$ is L-Lipschitz for all Z. Then, for every $t_0 \in 0, 1, 2, ..., n$ we have that,*

$$\mathbf{E}_{Z^n,Z}\mathbf{E}_A[\ell(f_{Z^n,T}, Z) - \ell(f_{Z^n_{(i)},T}, Z)] \leq L\mathbf{E}_{Z^n,Z}[\mathbf{E}_A[\delta_T(Z^n, Z)|\delta_{t_0}(Z^n, Z) = 0]] + \mathbf{E}_{Z^n,A}[L(A(Z^n))]\frac{t_0}{n}$$

.

Here comes the outline about how to prove theorem 5. Denoting $\mathbf{E}_A[\delta_t(Z^n, Z)|\delta_{t_0}(Z^n, Z) = 0]$ as $\Delta_t(Z^n, Z)$, the proof starts by bounding $\Delta_T(Z^n, Z)$ in lemma 2 using recursion. By applying lemma 1, it ends that

$$\Delta_T(Z^n, Z) \leq \sum_{t=t_0+1}^{T} \exp(c \sum_{k=t+1}^{T} \frac{\psi_k(Z^n, Z)}{k}) \frac{2cL}{nt}$$

, where $\psi_k(Z^n, Z) := \min(\xi_t(Z^n, Z), M)$. Then, $\mathbf{E}_{Z^n,Z}[\exp(c\sum_{k=t+1}^{T} \frac{\psi_k(Z^n,Z)}{k})]$ can be bounded as

$$\mathbf{E}_{Z^n,Z}[\exp(c \sum_{k=t+1}^{T} \frac{\psi_k(Z^n, Z)}{k})] \leq \exp(c \sum_{k=t+1}^{T} \frac{2\mu_k}{k})$$

, where $\mu_k := \mathbf{E}_{Z^n,Z}[\psi_k(Z^n, Z)] \leq \gamma$. Consequently,

$$\mathbf{E}_{Z^n,Z}[\Delta_T(Z^n, Z)] \leq \frac{1}{2c\gamma} \frac{2cL}{n} (\frac{T}{t_0})^{2c\gamma}$$

. By plugging this inequality back to lemma 2,

$$\mathbf{E}_{Z^n,Z}\mathbf{E}_A[\ell(f_{Z^n,T}, Z) - \ell(f_{Z^n_{(i)},T}, Z)] \leq \frac{L^2}{\gamma n} (\frac{T}{t_0})^{2c\gamma} + r\frac{t_0}{n}$$

, where $r := \mathbf{E}_{Z^n,Z}[L(A(Z^n))]$ for brevity. Since t_0 is a free parameter, and eventually by tuning of t_0, the right hand side of the previous inequality is minimized to be

$$\frac{1 + \frac{1}{q}}{n} (2cL^2)^{\frac{1}{1+q}} (rT)^{\frac{q}{1+q}}$$

, where $q := 2c\gamma$, when

$$t_0 = (\frac{2cL^2}{r})^{\frac{1}{1+q}} T^{\frac{q}{1+q}}$$

.

Corollary 1. *Under conditions of Theorem 5 we have that SGD is $\varepsilon(P, f_1)$-on-average stable with*

$$\varepsilon(P, f_1) = \mathcal{O}(\frac{1 + \frac{1}{c\gamma}}{n} (L(f_1) \cdot T)^{\frac{c\gamma}{1+c\gamma}})$$

.

In corollary 1, for the risk term in $(L(f_1) \cdot T)^{\frac{c\gamma}{1+c\gamma}}$, if $L(f_1)$ goes to 0, ε also goes to 0. In other words, the generalization error approaches zero as the risk of the initialization point vanishes. However, uniform stability can not capture this fact in that it is distribution-free.

Corollary 2. *Under conditions of Theorem 5 we have that the output of SGD obeys*

$$\mathbf{E}_{Z^n,A}[L(A(Z^n)) - L_n(A(Z^n))] = \mathcal{O}(\frac{1 + \frac{1}{c\gamma}}{n} \cdot \max((\mathbf{E}_{Z^n,A}[L_n(A(Z^n))] \cdot T)^{\frac{c\gamma}{1+c\gamma}}, (\frac{T}{n})^{c\gamma}))$$

.

Kuzborskij et al. [1] empirically evaluate the tightness of this non-convex generalization bounds on real data using the MINST dataset. The network involves three convolutional layers interlaced with max-pooling layers and then followed by fully connected layers.

In figure 1, the blue dash line represents the bound given in Theorem 4 by Hardt et al. [2] and the lecture note, and colored lines are data-dependent bounds for multiple start points. It can be clearly seen that data-dependent bound gives tighter estimations, and moreover, using pre-trained positions to start suggests even tighter estimations.

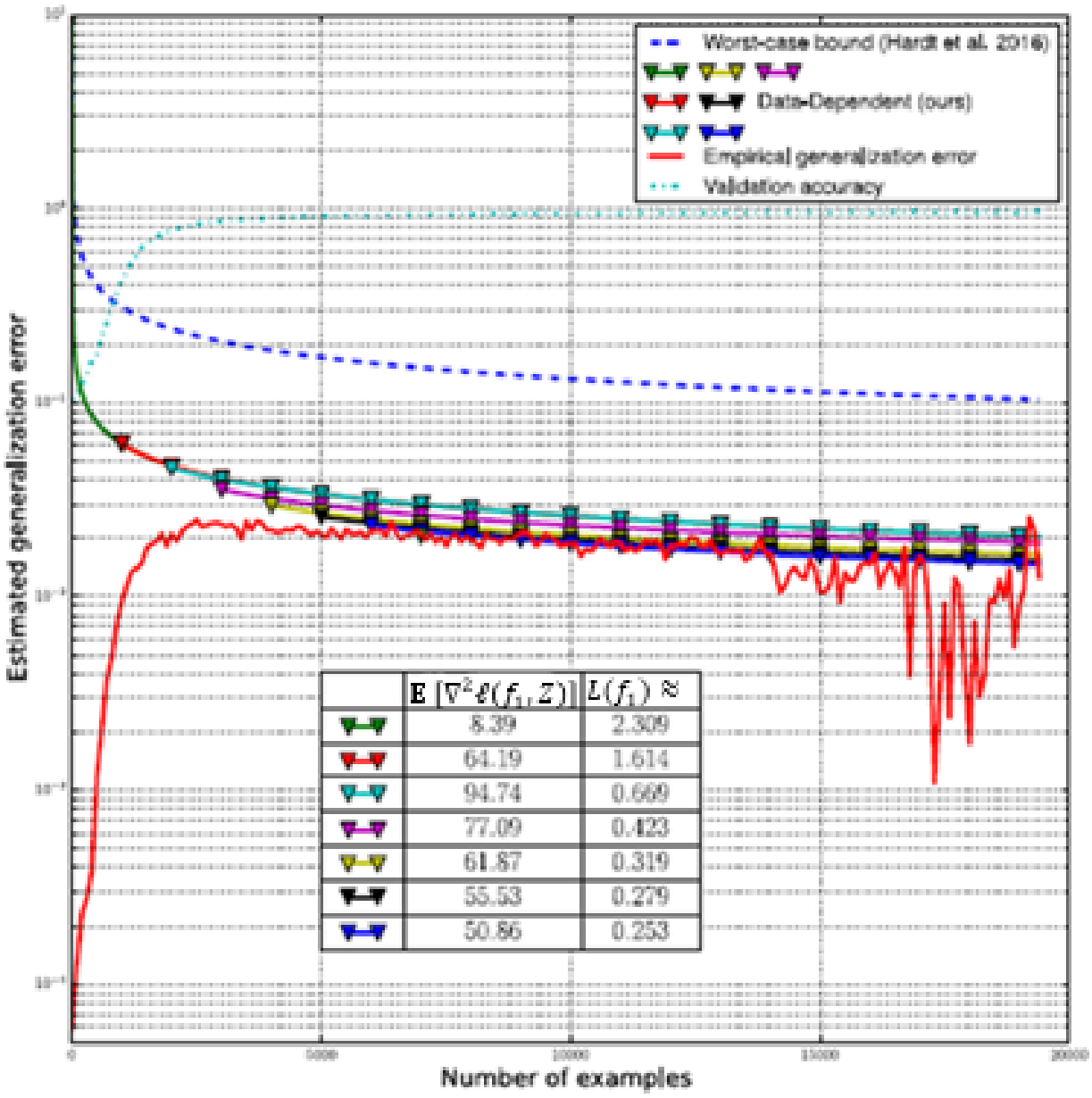

Figure 1: Comparison.

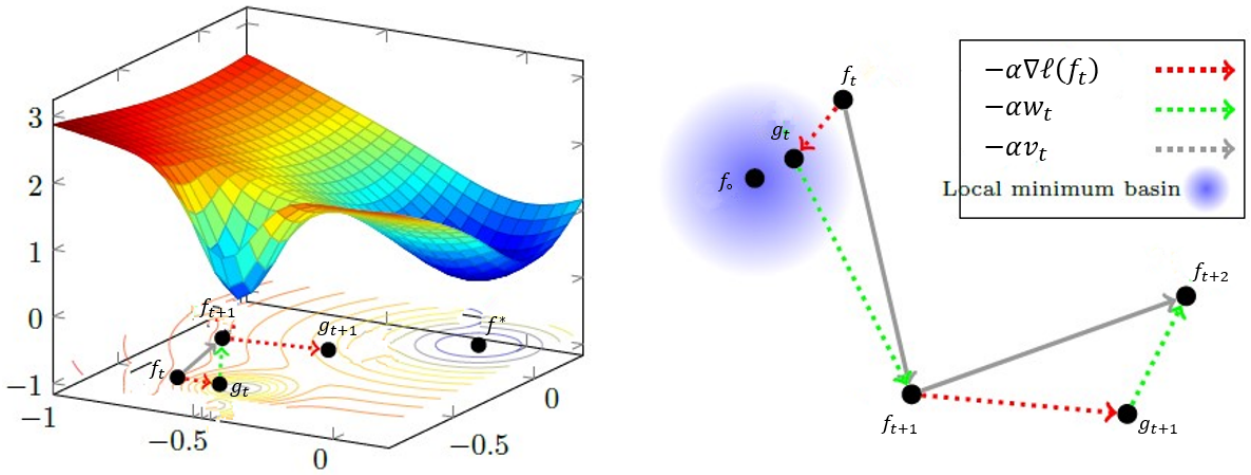

Figure 2: Trajectory.

4 Optimization

This section will discuss the phenomenon that the performance of SGD overwhelms GD in both efficiency and accuracy for non-convex optimization problem, such as NN, and develop theorems to support this proposition.

Assume step sizes α is a fixed value here, and let v_t denotes the stochastic gradients that statisfies $\mathbf{E}[v_t] = \nabla\ell(f_t)$ and w_t denotes the noise with $\mathbf{E}[w_t] = 0$. Figure 2 gives an intuition that for some f_t, instead of pointing to the solution $f*$, its negative gradient points to a bad local minimum. If using GD, it will arrive at $g_t := f_t - \alpha\nabla\ell(f_t)$, but if using SGD, it will arrive at $f_{t+1} = f_t - \alpha v_t$. So, if a large α is taken, the negative gradient at x_{t+1} may not point to $x_\circ$), and it is possible to get out of the basin with the help of the noise.

Since, $-\alpha v_t = -\alpha(\nabla\ell(f_t) + w_t)$ and then $f_{t+1} = g_t - \alpha w_t$, it follows that $g_{t+1} = g_t - \alpha w_t - \alpha\nabla\ell(g_t - \alpha w_t)$. Since the noise has expectation zero, $\mathbf{E}_{w_t}[g_{t+1} = g_t - \alpha\nabla\mathbf{E}_{w_t}[\ell(g_t) - \alpha w_t]]$. Therefore, the function $\ell_t'(f)\mathbf{E}_{w_t}[\ell(g - \alpha w_t)]$ is simply the original ℓ convolved with the α-scaled gradient noise, and the sequence g_t is approximately doing GD on the sequence of functions $\ell_t'(f)$.

Assumption 1. For a fixed point f^*, noise distribution $W(f)$, step size α, the function ℓ is m-one point strongly convex with respect to f^* after convolved with nose. That is, for any f, g in domain $\mathcal{F}$ s.t. $g = f - \alpha\nabla\ell(f)$,

$$\langle -\nabla\mathbf{E}_{w\in W(f)}\ell(g - \alpha w), f^* - g\rangle \geq c||f^* - g||_2^2$$

.

According to Kleinberg et al. [3], for point g, since the direction $f^* - g$ points to f^*, by having positive inner product with $f^* - g$, the direction $-\alpha\nabla\ell(g_t - \alpha w_t)$ approximately points to f^* in expectation. Therefore, g_t will converge to f^* with decent probability.

Theorem 6. *Assume ℓ is M-smooth, for every $f \in \mathcal{F}$, $W(f)$ s.t., $\max_{w\sim W(f)}||w||_2 \leq r$. For a fixed target solution f^*, if there exists constant $c, \alpha > 0$, such that Assumption 1 holds with f^*, α, c, and $\alpha < \min(\frac{1}{2M}, \frac{c}{M^2}, \frac{1}{2c})$, $\lambda := 2\alpha c - \alpha^2M^2$, $b := \alpha^2r^2(1 + \alpha M)^2$. Then For any fixed $T_1 \geq \frac{\log(\frac{\lambda||g_0 - f^*||_2^2}{b})}{\lambda}$ and $T_2 > 0$, with probability at least $\frac{1}{2}$, we have $||g_T - f^*||_2^2 \leq \frac{20b}{\lambda}$ and $||g_t - f^*||_2^2 \leq \mathcal{O}(\frac{\log(T_2)b}{\lambda})$ for all t s.t., $T_1 + T_2 \geq t \geq T_1$.*

This theorem helps to explain why SGD can escape sharp local minima and converge to flat local minima. The sharp local minima have small loss value and small diameter, so after convolved with the noise kernel, they easily disappear, but flat local minima survive due to their large diameter. See figure 3 as an example. The function ℓ is an approximately convex function but very spiky. Therefore, GD gets stuck at various local minima as shown in the bottom left subplot. For using SGD, on one hand, if the noise is small, the convolved ℓ is still somewhat non-convex, then SGD may find a few bad local minima; on the other hand, if the noise is too large, the noise dominates the gradient, and SGD will act like random walk. In this example, using noise in $[-0.3, 0.3]$ seems to be nice trade-off.

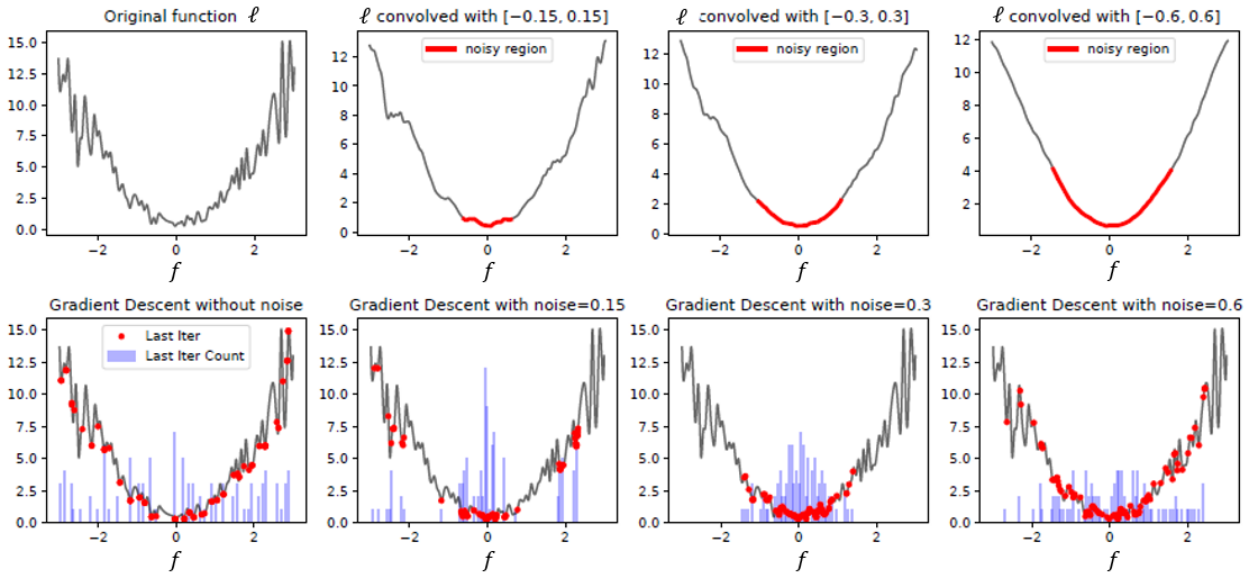

Figure 3: Example.

This theorem not only says SGD will get close to f^* but also says with constant probability, SGD will stay close to f^* for the future T_2 steps. Within $T_1 + T_2$ steps, SGD will stay in a local region centered at f^* with diameter $\mathcal{O}(\frac{\log(T_2)b}{\lambda})$. However, for fixed c, there exists a lower bound on α to satisfy assumption 1, so α cannot be arbitrarily small.

References

[1] Ilja Kuzborskij and Christoph H Lampert. Data-dependent stability of stochastic gradient descent. *arXiv preprint arXiv:1703.01678*, 2017.

[2] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

[3] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.