ESE 680-004: Learning and Control **Fall 2019**

Lecture 16: Algorithmic Stability and Stochastic Gradient Descent

Lecturer: Nikolai Matni *Scribes: Jialin Mao*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

1 Introduction

In this lecture we continue our study on generalization error bounds. Let us first recall the problem setup. Let $\mathcal{X}, \mathcal{Y}$ be an input and an output space, respectively. Denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and let $\mathcal{D}$ be an unknown distribution on $\mathcal{Z}$. Given a function class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ and a loss function $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$, we would like to find $f \in \mathcal{F}$ that minimizes the risk

$$R[f] = \mathbb{E}_{z \sim \mathcal{D}}[\ell(f, z)].$$

The difficulty here is that the distribution is unknown and we only have access to a data set $S = \{z_1, \ldots, z_N\}$ consisting of i.i.d. samples from the distribution. A natural workaround is to find the minimizer of the empirical risk

$$R_S[f] = \frac{1}{N} \sum_{i=1}^{N} \ell(f, z_i)$$

and the crucial problem here is to bound the generalization error

$$R[f] - R_S[f].$$

In the previous lecture we proved uniform generalization bounds based on concentration inequalities and some measure of the function class complexity. In this lecture, we will adopt an alternative perspective and view the problem through the lens of stochastic optimization. Instead of considering all possible functions within a function class, we will focus on functions found by stochastic optimization and prove generalization bounds based on algorithmic stability.

The lecture note will be organized as follows: we will start by introducing notions of stochastic gradient algorithms in section 2 and show that they find solutions with vanishing excess risk for convex loss function. After introducing notions of algorithmic stability in section 3, we will show how stability guarantees generalization error and proceed to prove the stability of stochastic gradient methods in section 4. We will conclude with stability inducing property of many commonly used optimization techniques in section 5.

2 Stochastic Gradient Method

Consider a function class parameterized by $\theta \in \Theta$, where Θ is convex and compact. Then each function $f \in \mathcal{F}$ can be characterized by its parameterization θ and we may write

$$R[\theta] \coloneqq R[f_\theta], \qquad R_S[\theta] \coloneqq R[f_\theta], \qquad \ell(\theta, z) \coloneqq \ell(f_\theta, z).$$

Stochastic gradient descent (SGD) is an iterative algorithm that updates θ based on the gradient of the loss function at a randomly sampled data point $z_k = (x_k, y_k) \sim \mathcal{D}$:

$$\theta_{k+1} = G_{\ell,\alpha}(\theta_k) \coloneqq \Pi_\Theta \left(\theta_k - \alpha_k \nabla \ell(\theta_k, z_k)\right), \tag{1}$$

where α_k is the learning rate and Π_Θ is the Euclidean projection onto Θ. $G_{\ell,\alpha}(\theta)$ is called the update rule.

2.1 Stochastic Optimization Bounds

Let $\{\theta_i\}_{i=1}^n$ be the trajectory of SGD according to the update rule (1), and let

$$\bar{\theta}_n := \sum_{i=1}^n w_i \theta_i, \qquad \text{where } w_i = \frac{\alpha_i}{\sum_{i=1}^n \alpha_i} \tag{2}$$

be the weighted running average of the trajectory. Our first result below shows that for loss functions that are convex in θ, $\bar{\theta}_n$ asymptotically minimizes the population risk (over Θ):

Theorem 1. *Suppose $\ell(\cdot, z)$ is a convex function for all $z \in \mathcal{Z}$, and $||\nabla \ell(\theta, z)|| \leq G$ for all $\theta \in \Theta, z \in \mathcal{Z}$, and suppose diam$(\Theta) \leq D$. Let $R_\star = \min_{\theta \in \Theta} R[\theta]$ be the best possible true risk achievable by a $\theta \in \Theta$. Set the step size $\alpha_i = \frac{D}{G\sqrt{n}}$. Then, with probability at least $1 - \delta$, we have*

$$R[\bar{\theta}_n] \leq R_\star + \frac{DG(1 + \sqrt{2\log(1/\delta)})}{\sqrt{n}} \tag{3}$$

Proof. Denote

$$\theta_\star = \arg\min_{\theta \in \Theta} R[\theta], \qquad g_i = \nabla \ell(\theta_i, (x_i, y_i)), \qquad D_i = ||\theta_i - \theta_\star||, \qquad \Delta_i = g_i - \nabla R[\theta_i].$$

Then

$$\begin{aligned}
R[\bar{\theta}_n] - R[\theta_\star] &\leq \mathbb{E}_z \left[\sum_{i=1}^n w_i \ell(\theta_i, z) \right] - R[\theta_\star] \qquad \text{(Jensen's Inequality)} \\
&= \sum_{i=1}^n w_i (R[\theta_i] - R_\star) \leq \sum_{i=1}^n w_i \langle \nabla R(\theta_i), \theta_i - \theta_\star \rangle \qquad \text{(convexity)} \\
&= \frac{1}{\sum_{i=1}^n \alpha_i} \sum_{i=1}^n [\alpha_i \langle g_i, \theta_i - \theta_\star \rangle - \alpha_i \langle \Delta_i, \theta_i - \theta_\star \rangle] \\
&\leq \frac{1}{2\sum_{i=1}^n \alpha_i} \sum_{i=1}^n \left(D_i^2 - D_{i+1}^2 + \alpha_i^2 ||g_i||^2 - 2\alpha_i \langle \Delta_i, \theta_i - \theta_\star \rangle \right) \qquad (\star) \\
&= \frac{1}{2\sum_{i=1}^n \alpha_i} \left(D_1^2 - D_{n+1}^2 + \sum_{i=1}^n \left(\alpha_i^2 ||g_i||^2 - 2\alpha_i \langle \Delta_i, \theta_i - \theta_\star \rangle \right) \right) \\
&\leq \frac{D^2 + G^2 \sum_{i=1}^n \alpha_i^2}{2\sum_{i=1}^n \alpha_i} - \sum_{i=1}^n w_i \langle \Delta_i, \theta_i - \theta_\star \rangle .
\end{aligned}$$

In $(\star)$ we used a key inequality:

$$||\theta_{i+1} - \theta_\star||^2 \leq ||\theta_i - \theta_\star||^2 - 2\alpha_i \langle g_i, \theta_i - \theta_\star \rangle + \alpha_i^2 ||g_i||^2 .$$

To see this, notice

$$\begin{aligned}
||\theta_{i+1} - \theta_\star||^2 &= ||\theta_{i+1} - \theta_i + \theta_i - \theta_\star||^2 \\
&= ||\theta_{i+1} - \theta_i||^2 + 2\langle \theta_{i+1} - \theta_i, \theta_i - \theta_\star \rangle + ||\theta_i - \theta_\star||^2 \\
&= ||\theta_i - \theta_\star||^2 + ||\Pi_\Theta(\theta_i - \alpha_i g_i) - \Pi_\Theta(\theta_i)||^2 + 2\langle \Pi_\Theta(\theta_i - \alpha_i g_i), \theta_i - \theta_\star \rangle \\
&\leq ||\theta_i - \theta_\star||^2 + \alpha_i^2 ||g_i||^2 + 2\langle \Pi_\Theta(\theta_i - \alpha_i g_i), \theta_i - \theta_\star \rangle \\
&= ||\theta_i - \theta_\star||^2 + \alpha_i^2 ||g_i||^2 - 2\alpha_i \langle g_i, \theta_i - \theta_\star \rangle ,
\end{aligned}$$

where for the inequality we used the fact that $||\Pi_\Theta(u) - \Pi_\Theta(v)|| \leq ||u - v||$ for any convex set Θ (see Lemma 5 and comments for a proof), and in the last step we use the fact that $\theta_i - \theta_\star \in \Theta$, so $\langle v, \theta_i - \theta_\star \rangle = \langle \Pi_\Theta(v), \theta_i - \theta_\star \rangle$ for any v in the ambient space.

Setting $\alpha_i = \frac{D}{G\sqrt{n}}$, we have

$$R[\theta_n] - R[\theta_\star] \leq \frac{DG}{\sqrt{n}} - \frac{1}{n}\sum_{i=1}^{n} \langle \Delta_i, \theta_i - \theta_\star \rangle \,. \tag{4}$$

Now notice that

$$X_j := \frac{1}{n}\sum_{i=1}^{j} \langle \Delta_i, \theta_i - \theta_\star \rangle$$

is a martingale (since $\mathbb{E}[\Delta_{j+1}|\theta_1, \ldots, \theta_j] = 0$) and by Cauchy-Schwartz we have

$$|X_{j+1} - X_j| = \frac{1}{n}\left|\langle \Delta_i, \theta_i - \theta_\star \rangle\right| \leq \frac{2GD}{n},$$

so by Azuma's inequality we have

$$\mathbb{P}[-X_n \geq t] \leq \exp\left(-\frac{nt^2}{2G^2D^2}\right).$$

Inverting the probability, we have with probability at least $1 - \delta$,

$$-\frac{1}{n}\sum_{i=1}^{n} \langle \Delta_i, \theta_i - \theta_\star \rangle \leq DG\sqrt{\frac{2\log(1/\delta)}{n}}$$

Combining this with (4), we have the desired high probability bound. □

Notice that this is not a generalization bound, but can be combined with a generalization bound to give a bound on the population risk of SGD.

3 Algorithmic Stability

In this section we will introduce the notion of algorithmic stability. Consider a learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{F}$ that maps a training set S into a function $\mathcal{A}(S)$. The notion of stability we will be using in this note is uniform stability (see other forms of stability in [1]). Informally, an algorithm is uniformly stable if a single change in the input results in almost no change in prediction, the formal definition is given below:

Definition 1 (Uniform Stability [1])**.** *A randomized algorithm $\mathcal{A}$ is ϵ-uniformly stable if for all datasets $S, S' \in \mathcal{Z}^n$ such that S, S' differ in at most one example, we have*

$$\sup_z \mathbb{E}_\mathcal{A}[\ell(\mathcal{A}(S); z) - \ell(\mathcal{A}(S'), z)] \leq \epsilon.$$

3.1 Generalization Bounds Through Algorithmic Stability

Now we show how algorithmic stability implies small generalization error. The following theorem proves generalization in expectation:

Theorem 2 (Theorem 2.2 [2])**.** *Let $\mathcal{A}$ be ϵ-uniformly stable, then*

$$\left|\mathbb{E}_{S,\mathcal{A}}\left[R_S(\mathcal{A}(S)) - R(\mathcal{A}(S))\right]\right| \leq \epsilon.$$

Proof. Let $S = (z_1, \dots, z_n)$ and $S' = (z'_1, \dots, z'_n)$ be two independent random samples from D, and let

$$S^i = \{z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_n\},$$

be a copy of S with the i-th element replaced by z'_i. Then

$$\mathbb{E}_S \mathbb{E}_{\mathcal{A}}[R_S[\mathcal{A}(S)]] = \mathbb{E}_S \mathbb{E}_{\mathcal{A}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\mathcal{A}(S), z_i) \right] = \mathbb{E}_{S,S'} \mathbb{E}_{\mathcal{A}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\mathcal{A}(S^i), z'_i) \right],$$

and

$$\mathbb{E}_S \mathbb{E}_{\mathcal{A}}[R[\mathcal{A}(S)]] = \mathbb{E}_S \mathbb{E}_{\mathcal{A}} \mathbb{E}_z[\ell(\mathcal{A}(S), z)] = \mathbb{E}_{S,S'} \mathbb{E}_{\mathcal{A}}[\ell(\mathcal{A}(S), z'_i)],$$

so their difference can be expressed as

$$\mathbb{E}_{S,\mathcal{A}}[R_S[\mathcal{A}(S)] - R(\mathcal{A}(S))] = \mathbb{E}_{S,S',\mathcal{A}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\ell(\mathcal{A}(S^i), z'_i) - \ell(\mathcal{A}(S), z'_i) \right) \right] \leq \epsilon,$$

since each term in the sum is upper bounded by ϵ by uniform stability. □

We can get high probability bound from theorem 2. Notice that the algorithmic stability also implies that

$$\left| R[\mathcal{A}(S)] - R[\mathcal{A}(S^i)] \right| \leq \epsilon,$$

and

$$\left| R_S[\mathcal{A}(S)] - R_{S^i}[\mathcal{A}(S^i)] \right| \leq \epsilon,$$

so the function $F(S, z) := R[\mathcal{A}(S)] - R_S[\mathcal{A}(S)]$ satisfies the bounded difference property and by McDiarmid's inequality and we get

$$\mathbb{P}\left[R[\mathcal{A}(S)] - R_S[\mathcal{A}(S)] \geq \epsilon_{\text{stab}} + t \right] \leq \exp\left(-\frac{t^2}{2\epsilon_{\text{stab}}^2 n} \right)$$

Inverting probability, with probability at least $1 - \delta$, we have

$$R[\mathcal{A}(S)] \leq R_S[\mathcal{A}(S)] \leq \epsilon_{\text{stab}} \left(1 + \sqrt{2n \log(1/\delta)} \right) \tag{5}$$

Notice that this bound is nonvacuous only if $\epsilon_{\text{stab}} = o(1/\sqrt{n})$, with $\epsilon_{\text{stab}} = c/\sqrt{n}$, we can recover the classical bound $O(1/\sqrt{n})$. There are more recent works like [3] that give better high probability bounds.

4 Algorithmic Stability of SGD

This section will be devoted to proving the algorithmic stability of SGD. We will mainly focus on the case of convex loss function in this note. Some results do carry over to the case where the loss-function is non-convex, provided that the steps are sufficiently small and the number of iterations is not too large. See section 3.5 of [2] for detailed analysis.

We will define here several properties of the update rules that characterize how they drive update sequences:

Definition 2 (Expansiviy)**.** *An update rule is η-expansive if*

$$\sup_{v,w \in \Theta} \frac{||G(v) - G(w)||}{||v - w||} \leq \eta$$

Definition 3 (Boundedness)**.** *An update rule is σ-bounded if*

$$\sup_{w \in \Theta} ||w - G(w)|| \leq \sigma.$$

The following lemma explicitly shows how expansiveness and boundedness control the divergence between update sequences:

Lemma 1 (Lemma 2.5 [2])**.** *For two fixed sequences of update rules $G_1, \dots, G_T$ and $G'_1, \dots, G'_T$, let $w_0 = w'_0$ be a starting point and let $w_{t+1} = G_t(w_t)$, $w'_{t+1} = G'_t(w'_t)$ be the sequence of updates evolving according to the update rules. Denote $\delta_t := ||w_t - w'_t||$, then $\delta_0 = 0$ and we have the recurrence relation*

$$\delta_{t+1} \leq \begin{cases} \eta\delta_t & G_t = G'_t \text{ is } \eta\text{-expansive} \\ \min(\eta, 1)\delta_t + 2\sigma_t & G_t \text{ and } G'_t \text{ are } \sigma\text{-bounded,} \\ & G_t \text{ is } \eta\text{-expansive} \end{cases}$$

Proof. The first inequality follows directly from the definition of η-expansiveness. The second inequality is essentially triangular inequality:

$$\begin{aligned} \delta_{t+1} &= \|G_t(w_t) - G'_t(w'_t)\| \\ &\leq \|G_t(w_t) - w_t + w'_t - G'_t(w'_t)\| + \|w_t - w'_t\| \\ &\leq \delta_t + \|G_t(w_t) - w_t\| + \|G_t(w'_t) - w'_t\| \\ &\leq \delta_t + 2\sigma \end{aligned}$$

and alternatively:

$$\begin{aligned} \delta_{t+1} &= \|G_t(w_t) - G'_t(w'_t)\| \\ &= \|G_t(w_t) - G_t(w'_t) + G_t(w'_t) - G'_t(w'_t)\| \\ &\leq \|G_t(w_t) - G_t(w'_t)\| + \|G_t(w'_t) - G'_t(w'_t)\| \\ &\leq \|G_t(w_t) - G_t(w'_t)\| + \|w'_t - G_t(w'_t)\| + \|w'_t - G'_t(w'_t)\| \\ &\leq \eta\delta_t + 2\sigma \end{aligned}$$

□

The following lemma shows that smoothness of the loss function and its derivative implies expansiveness of the gradient update rule (1):

Lemma 2 (Lemma 3.3, 3.7 [2])**.** *Assume f is L-Lipschitz and the gradient of f is β-Lipschitz, then the follow properties hold:*

1. *$G_{f,\alpha}$ is (αL)-bounded.*
2. *Assume in addition that f is convex. Then for any $\alpha \leq 2/\beta$, the gradient update $G_{f,\alpha}$ is 1-expansive.*

Proof.

1. The first property follows directly from definition and the 1-expansivity of Euclidean projection:

$$||w - G_{f,\alpha}(w)|| \leq ||\alpha \nabla f(w)|| \leq \alpha L.$$

2. Lipschitz assumption on the gradient and convexity together imply the co-coercivity of the gradients (a proof can be found in Theorem 2.1.5 of [4]):

$$\langle \nabla f(v) - \nabla f(w), v - w \rangle \geq \frac{1}{\beta}\|\nabla f(v) - \nabla f(w)\|^2$$

Therefore

$$\begin{aligned} \|G_{f,\alpha}(v) - G_{f,\alpha}(w)\|^2 &= \|v - w\|^2 - 2\alpha\langle \nabla f(v) - \nabla f(w), v - w \rangle + \alpha^2\|\nabla f(v) - \nabla f(w)\|^2 \\ &\leq \|v - w\|^2 - \left(\frac{2\alpha}{\beta} - \alpha^2\right)\|\nabla f(v) - \nabla f(w)\|^2 \\ &\leq \|v - w\|^2 \end{aligned}$$

□

The lemmas above provide all the essential tools we need to prove the stability of SGD. Intuitively, on two datasets that differ only in one example, SGD performs the same update with high probability, so the final output are close to each other by the expansiveness of the gradient update. In the rare case where SGD selects different example, we can make use of the smoothness properties of the loss function to bound the growth of difference. A formal proof is given below:

Theorem 3 (Theorem 3.8 [2])**.** *Assume the loss function is convex and L-Lipschitz, and its gradient is β-Lipschitz. Suppose we run SGD with step size $\alpha_t \leq 2/\beta$ for T steps, then the procedure satisfies uniform stability with*

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

Proof. Let S, S' denote two datasets differing only in one point, we want to show

$$\sup_{S,S',z} \mathbb{E}\left|\ell(\mathcal{A}(S); z) - \ell(\mathcal{A}(S'), z)\right| \leq \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t. \tag{6}$$

Consider the gradient updates $G_1, \dots, G_T$ and $G'_1, \dots, G'_T$ induced by running SGD on S, S', respectively. Let w_T, w'_T denote the corresponding outputs.

For a fixed example $z \in \mathcal{Z}$, the Lipschitz condition gives

$$\mathbb{E}\left|f\left(w_T; z\right) - f\left(w'_T; z\right)\right| \leq L\mathbb{E}\left[\delta_T\right], \tag{7}$$

where $\delta_t = ||w_t - w'_t||$ as before. Notice at time step t, with probabiliy $1 - 1/n$, the element selected by SGD is the same in both S, S'. In this case we have $G_t = G'_t$ so by Lemma 2.2, the update is 1-expansive and we have $\delta_{t+1} \leq \delta_t$.

With probability $1/n$, the selected sample is different, in which case we will use that both G_t and G'_t are $\alpha_t L$-bounded by Lemma 2.1. Then by the second inequality in Lemma 1 we have $\delta_{t+1} \leq \delta_t + 2\alpha_t L$. Then by linearity of expectation we have

$$\mathbb{E}\left[\delta_{t+1}\right] \leq \left(1 - \frac{1}{n}\right)\mathbb{E}\left[\delta_t\right] + \frac{1}{n}\mathbb{E}\left[\delta_t\right] + \frac{2\alpha_t L}{n} = \mathbb{E}\left[\delta_t\right] + \frac{2L\alpha_t}{n}$$

for every $0 \leq t \leq T$, which gives

$$\mathbb{E}\left[\delta_T\right] \leq \frac{2L}{n} \sum_{t=1}^{T} \alpha_t.$$

Combining this with (7), we have

$$\mathbb{E}\left|f\left(w_T; z\right) - f\left(w'_T; z\right)\right| \leq \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t,$$

since this holds for all S, S', z, we know the desired inequality (6) holds. □

If we use a constant learning rate α and output the averaged model $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$, we could improve the bound by a constant factor (see proof in 4) and get

$$\epsilon_{\text{stab}} \leq \frac{\alpha T L^2}{n}.$$

Combining this with the high probability bound (5), we have with probability at least $1 - \delta$,

$$R[\theta_T] \leq R_S[\theta_T] + \frac{\alpha T L^2}{n}\left(1 + \sqrt{2n\log(1/\delta)}\right).$$

5 Stability-Inducing Operations

In this section we show that several popular heuristic operations performed in practice do increase the stability of stochastic gradient method.

5.1 Weight Decay

Given an objective function f, a learning rate α and a weight decay rate μ, the gradient update with weight decay is defined to be

$$G_{f,\mu,\alpha}(w) = (1-\alpha\mu)w - \alpha\nabla f(w).$$

Notice this is equivalent to a gradient step with step size α on the regularized objective

$$g(w) = f(w) + \frac{\mu}{2}\left\|w\right\|^2.$$

The effect of weight decay on gradient descent is shown in the following lemma:

Lemma 3 (Lemma 4.2 [2])**.** *Assume f has β-Lipschitz gradients, then $G_{f,\mu,\alpha}$ is $(1+\alpha(\beta-\mu))$-expansive.*

Proof. Denote $G = G_{f,\mu,\alpha}$, then by triangular inequality and Lipschitz condition, we have

$$\begin{aligned}\|G(v)-G(w)\| &\le (1-\alpha\mu)\|v-w\| + \alpha\|\nabla f(w)-\nabla f(v)\| \\ &\le (1-\alpha\mu)\|v-w\| + \alpha\beta\|w-v\| \\ &= (1-\alpha\mu+\alpha\beta)\|v-w\|\end{aligned}$$

□

In other words, using weight decay improves the smoothness of the function and replaces any dependence on β with $\beta-\mu$. In particular, once $\mu > \beta$, the update becomes contractive.

5.2 Gradient Clipping

When training deep neural networks it is a common practice to restrict the magnitude of the gradient, typically through truncation, scaling, or dropping of examples that cause an exceptionally large value of the gradient norm. This leads to a bound on the Lipschitz parameter L of the loss.

5.3 Dropout

Dropout is a popular heuristic often used for preventing overfitting. Practically, applying dropout is equivalent to placing a mask on the gradient that sends a fraction of the gradient to zero, i.e., replace $\nabla\ell(\theta,z)$ with $D\nabla\ell(\theta,z)$.

Definition 4. *We say a randomized map $D:\Theta\to\Theta$ is a dropout operator with dropout rate s if for all $v\in\Theta$, we have $\mathbb{E}\left||Dv\right|| = s\left||v\right||$. For a differentiable function $f:\Theta\to\Theta$, we let*

$$DG_{f,\alpha} \coloneqq v - \alpha D(\nabla f(v))$$

denote the dropout gradient update.

Dropout operators improve the effective Lipschitz constant of the objective function.

Lemma 4. *Assume f is L-Lipschitz. Then the dropout update $DG_{f,\alpha}$ with rate s is $s\alpha L$-bounded.*

Proof. Essentially by definition:

$$\mathbb{E}\left||DG_{f,\alpha}(v)-v\right|| = \alpha\mathbb{E}\left||D\nabla f(v)\right|| = \alpha s\mathbb{E}\left||\nabla f(v)\right|| \le \alpha sL.$$

□

5.4 Projections and Proximal Steps

The proximal update rule is defined by

Definition 5 ((Proximal Update) [2])**.** *For a nonnegative step size $\alpha \geq 0$ and a function $f : \Theta \to \mathbb{R}$, we define the proximal update rule $P_{f,\alpha}$ as*

$$P_{f,\alpha}(\theta) = \arg\min_{v} \frac{1}{2} ||\theta - v||^2 + \alpha f(v)$$

The following lemma shows why proximal steps could improve stability:

Lemma 5. *If f is convex, the proximal update is* 1*-expansive.*

Proof. Define

$$P_\alpha(w) := \arg\min_{v} \frac{1}{2\alpha} ||w - v|| + f(v),$$

and define the map $Q_\alpha(w) := w - P_\alpha(w)$. Then by the optimality conditions we know

$$\alpha^{-1} Q_\alpha(w) \in \partial f(P_\alpha(w)).$$

Then by convexity of f we have

$$\langle P_\alpha(v) - P_\alpha(w), Q_\alpha(v) - Q_\alpha(w) \rangle,$$

so we have

$$\begin{aligned} ||v - w||^2 &= ||P_\alpha(v) - P_\alpha(w) + Q_\alpha(v) - Q_\alpha(w)||^2 \\ &= ||P_\alpha(v) - P_\alpha(w)|| + 2 \langle P_\alpha(v) - P_\alpha(w), Q_\alpha(v) - Q_\alpha(w) \rangle + ||Q_\alpha(v) - Q_\alpha(w)||^2 \\ &\geq ||P_\alpha(v) - P_\alpha(w)|| . \end{aligned}$$

□

Notice that for a convex set Θ, the indicator function $\mathbb{I}_\Theta$ is convex, and hence the Euclidean projection onto Θ is 1-expansive. In many cases, proximal operators are actually contractive. A notable case is when $f(\cdot)$ is the Euclidean norm, in which case the update rule is η-expansive with $\eta = (1 + \alpha)^{-1}$, so choosing an appropriate proximal update could induce better stability.

5.5 Model Averaging

The idea of model averaging is to output the weighted average of iterates w_t obtained at each time step, as introduced in 2. The following theorem shows that model averaging improves the bound in Theorem 3 by a constant factor

Theorem 4. *Assume $f : \Theta \to [0, 1]$ is convex, L-Lipschitz and has β-Lipschitz gradient, then if we run SGD with step size $\alpha_T \leq \alpha \; leq 2/\beta$ for T steps, the average of the first T iterates of SGD has uniform stability of $\epsilon_{stab} \leq \frac{\alpha(T+1)L^2}{n}$.*

Proof. Let $\bar{w}_T = \frac{1}{T} \sum_{t=1}^{T} w_t$, then since

$$w_t = \sum_{k=1}^{t} \alpha \nabla f(w_k, z_k),$$

we have

$$\bar{w}_T = \alpha \sum_{t=1}^{T} \sum_{k=1}^{t} \nabla f(w_k, z_k) = \alpha \sum_{k=1}^{T} \sum_{t=k}^{T} \nabla f(w_k, z_k) = \alpha \sum_{k=1}^{T} \frac{T - k + 1}{T} \nabla f(w_k, z_k).$$

Follow a similar argument as in theorem 3, we have

$$\mathbb{E}[\delta_t] \leq (1 - 1/n)\mathbb{E}[\delta_{t-1}] + \frac{1}{n} \left(\mathbb{E}[\delta_{t-1}] + 2\alpha L \frac{T - t + 1}{T} \right),$$

summing up both sides and using the Lipschitz continuity of f as in theorem 3, we have the desired bound. □

References

[1] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 06 2002.

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

[3] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *arXiv e-prints*, page arXiv:1902.10710, Feb 2019.

[4] Yurii Nesterov. *Smooth Convex Optimization*, pages 59–137. Springer International Publishing, Cham, 2018.