Train faster, generalize better: Stability of stochastic gradient descent

Presenter: Xuezhou Zhang

December 27, 2016

Outline

General Setting of Supervised Learning

- We receive a sample $S = (z_1, \ldots, z_n)$ of n examples drawn i.i.d. from a distribution $\mathcal{D}$. Our goal is to find a model w with small **population risk**, defined as

$$R[w] = \mathbb{E}_{z \sim \mathcal{D}} f(w; z) \tag{1}$$

  where $f(w; z)$ is the **loss** of the model described by w encountered on example z.
- Since we cannot measure the objective $R[w]$ directly, we often instead try to minimize the **empirical risk**, defined as:

$$R_S[w] = \frac{1}{n} \sum_{i=1}^{n} f(w; z_i) \tag{2}$$

General Setting of Supervised Learning

- The **generalization error** of a model w is the difference

$$R_S[w] - R[w]. \tag{3}$$

- When $w = A(S)$ is chosen by a potentially randomized algorithm A, it makes sense to consider the **expected generalization error**:

$$\epsilon_{gen} = \mathbb{E}_{S,A}[R_S[A(S)] - R[A(S)]], \tag{4}$$

where the expectation is over the randomness of A and the sample S.

Recall from the last talk..

Definition (2.1)

*A randomized algorithm A is ϵ-**uniformly stable** if for all data sets $S, S' \in Z^n$ such that S and S' differ in at most one example, we have*

$$\sup_z \mathbb{E}_A[f(A(S); z) - f(A(S'); z)] \leq \epsilon. \tag{5}$$

We will denote by $\epsilon_{stab}(A, n)$ the infimum over all ϵ for which (5) holds.
***Note**: This stability constant implicitly depends on n.*

Theorem (2.2)

[Generalization in expectation] Let A be ϵ-uniformly stable. Then,

$$|\mathbb{E}_{S,A}[R_S[A(S)] - R[A(S)]]| \leq \epsilon. \tag{6}$$

Stochastic Gradient Descent

- Given n labeled examples $S = (z_1, \ldots, z_n)$ where $z_i \in Z$, consider a decomposable objective function

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f(w; z_i) \tag{7}$$

where $f(w; z_i)$ denotes the loss of w on the example z_i. The **stochastic gradient update** for this problem with learning rate $\alpha_t > 0$ is given by

$$w_{t+1} = G_{f,\alpha_t}(w_t) = w_t - \alpha_t \nabla_w f(w_t; z_{i_t}). \tag{8}$$

- **Stochastic Gradient Descent** (SGD) is the algorithm resulting from performing stochastic gradient updates T times where the indices i_t are randomly chosen.

Outline

Stability of Sochastic Gradient Descent

- Our goal is to show that SGD is uniformly-stable in three different cases, namely with convex, strongly convex and non-convex objective functions.
- Then, theorem 2.2 would imply that SGD generalizes well.
- Recall that to show that a learning algorithm is stable, we want to establish some bound on the quantity

$$\mathbb{E}_A[f(A(S); z) - f(A(S'); z)]. \tag{9}$$

Stability of Sochastic Gradient Descent

- We say that f is L-**Lipschitz** if for all points x in the domain of f we have $\|\nabla f(x)\| \leq L$. This implies that for all $v, w \in \Omega$ we have

$$|f(v) - f(w)| \leq L\|v - w\| \tag{10}$$

- Note that if the loss function f is L-Lipschitz for every example z, then we have

$$\mathbb{E}_A|f(w; z) - f(w'; z)| \leq L\,\mathbb{E}_A\|w - w'\| \tag{11}$$

- Hence, it suffices to analyze how w_t and w'_t diverge in the domain as a function of iteration t, and we can do that with the help of the next two lemmas.

Proof Tools

Definition (2.3)

*We consider general update rules of the form $G : \Omega \to \Omega$ which map a point $w \in \Omega$ in the parameter space to another point $G(w)$. An update rule is η-**expansive** if*

$$\sup_{v,w \in \Omega} \frac{\|G(v) - G(w)\|}{\|v - w\|} \leq \eta \tag{12}$$

Definition (2.4)

*An update rule is σ-**bounded** if*

$$\sup_{w \in \Omega} \|w - G(w)\| \leq \sigma. \tag{13}$$

Proof Tools

Lemma (Growth recursion)

Fix an arbitrary sequence of updates $G_1, \ldots, G_T$ and another sequence $G'_1, \ldots, G'_T$. Let $w_0 = w'_0$ be a starting point in Ω and define $\delta_t = \|w_t - w'_t\|$ where w_t, w'_t are defined recursively through

$$w_{t+1} = G_t(w_t) \qquad w'_{t+1} = G'_t(w'_t). \tag{14}$$

Then we have the recurrence relation

$$\delta_0 = 0 \tag{15}$$

$$\delta_{t+1} \leq \begin{cases} \eta\delta_t & G_t = G'_t \text{ is } \eta\text{-expansive} \\ \min(\eta, 1)\delta_t + 2\sigma_t & G_t \text{ and } G'_t \text{ are } \sigma\text{-bounded,} \\ & \text{and } G_t \text{ is } \eta\text{-expansive} \end{cases} \tag{16}$$

Note: This is saying if the update rules are expansive and bounded, then we can bound δ_t recursively.

Proof Tools

Definition (3.6)

*A function $f : \Omega \to \mathbb{R}$ is β-**smooth** if for all $v, w \in \Omega$ we have*

$$\|\nabla f(v) - \nabla f(w)\| \leq \beta \|v - w\| \tag{17}$$

Definition (3.4)

*A function $f : \Omega \to \mathbb{R}$ is **convex** if for all $v, w \in \Omega$ we have*

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle. \tag{18}$$

Definition (3.5)

*A function $f : \Omega \to \mathbb{R}$ is γ-**strongly convex** if for all $v, w \in \Omega$ we have*

$$f(v) \geq f(w) + \langle \nabla f(w), w - v \rangle + \frac{\gamma}{2} \|v - w\|^2. \tag{19}$$

Proof Tools

Lemma (3.7)

Assume that f is β-smooth. Then the following properties hold.
1. the gradient update $G_{f,\alpha}$ is $(1+\alpha\beta)$-expansive.
2. Assume in addition that f is convex. Then, for any $\alpha \leq \frac{2}{\beta}$, $G_{f,\alpha}$ is 1-expansive.
3. Assume in addition that f is γ-strongly convex. Then, for $\alpha \leq \frac{2}{\beta+\gamma}$, $G_{f,\alpha}$ is $\left(1-\frac{\alpha\beta\gamma}{\beta+\gamma}\right)$-expansive.

Note: That is saying, the gradient update $G_{f,\alpha}$ is expansive with some constant if f is smooth and α is small enough.

Stability of SGD with convex objectives

Theorem (3.8)

Assume that the loss function $f(\cdot; z)$ is β-smooth, convex and L-Lipschitz for every z. Suppose that we run SGD with step sizes $\alpha_t \leq \frac{2}{\beta}$ for T steps. Then, SGD satisfies uniform stability with

$$\epsilon_{stab} \leq \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t. \tag{20}$$

Note: Intuitively, the theorem says if the step size α is small, number of steps T is small, sample size n is large, then SGD is stable.

Stability of SGD with convex objectives

Proof sketch.

- Let S and S' be two samples of size n differing in only a single example. Consider the gradient updates $G_1, \ldots, G_T$ and $G'_1, \ldots, G'_T$ induced by running SGD on sample S and S', respectively.
- Observe that at step t, with probability $1 - 1/n$, the example selected by SGD is the same in both S and S'. In this case, we have $G_t = G'_t$ and we can use the 1-expansivity of G_t by the lemmas. With probability 1/n the selected example is different in which case we use that both G_t and G'_t are $\alpha_t L$-bounded. Hence, this gives

$$\mathbb{E}[\delta_{t+1}] \leq \left(1 - \frac{1}{n}\right) \mathbb{E}[\delta_t] + \frac{1}{n} \left(\mathbb{E}[\delta_t] + 2\alpha_t L\right) = \mathbb{E}[\delta_t] + \frac{2\alpha_t L}{n} \tag{21}$$

- Unraveling the recursion gives us the result.

Stability of SGD with strongly convex objectives

Theorem (3.9)

Assume that the loss function $f(\cdot; z)$ is β-smooth, L-Lipschitz and γ-strongly convex for every z. Suppose that we run the (projected) SGD with constant step size $\alpha \leq \frac{1}{\beta}$ for T steps. Then, SGD satisfies uniform stability with

$$\epsilon_{stab} \leq \frac{2L^2}{\gamma n}. \tag{22}$$

- In the strongly convex case, we restrict our attention to a compact, convex set Ω over which we wish to optimize, as a strongly convex function has unbounded gradient in $\mathbb{R}^n$.
- Intuitively, the theorem says the stability of SGD is only affected by the sample size, with no dependence on the size or number of steps at all.

Stability of SGD with non-convex objectives

Theorem (3.12)

Assume that $f(\cdot; z) \in [0, 1]$ is a β-smooth, L-Lipschitz loss function for every z. Suppose that we run SGD for T steps with monotonically non-increasing step sizes $\alpha_t \leq c/t$. Then, SGD satisfies uniform stability with

$$\epsilon_{stab} \leq \frac{1 + 1/\beta c}{n - 1} (2cL)^{\frac{1}{\beta c + 1}} T^{\frac{\beta c}{\beta c + 1}}. \tag{23}$$

In particular, omitting constant factors, we get

$$\epsilon_{stab} \lesssim \frac{T^{1 - 1/(\beta c + 1)}}{n}. \tag{24}$$

Note: This theorem shows that the number of steps of SGD can grow as n^c for a small $c > 1$, without sacrificing stability. This provides some explanation to why neural networks can be trained for multiple epochs of stochastic gradient and still exhibit excellent generalization.

Outline

Weight Decay and Regularization

Definition

Let $f : \Omega \to \Omega$ be a differentiable function. We define the gradient update with ***weight decay*** *at rate μ as*

$$G_{f,\mu,\alpha}(w) = (1 - \alpha\mu)w - \alpha\nabla f(w). \tag{25}$$

It is easy to verify that the above update rule is equivalent to performing a gradient update on the l_2-regularized objective

$$g(w) = f(w) + \frac{\mu}{2}\|w\|^2. \tag{26}$$

Lemma

Assume that f is β-smooth. Then, $G_{f,\mu,\alpha}$ is $(1 + \alpha(\beta - \mu))$-expansive.

Dropout

Definition

*We say that a randomized map $D : \Omega \to \Omega$ is a **dropout operator** with dropout rate s if for every $v \in \Omega$ we have $\mathbb{E}\|Dv\| = s\|v\|$. For a differentiable function $f : \Omega \to \Omega$, we let $DG_{f,\alpha}$ denote the **dropout gradient update** defined as $DG_{f,\alpha}(v) = v - \alpha D(\nabla f(v))$.*

Lemma

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