IIIS 2018 Spring: ATCS - Learning and Optimization - Theory and Practice
Lecture date: May 14, 2018
Instructor: Jian Li **Scribe: Jiezhong Qiu**

1 Stability Bound for Stochastic Gradient Method

1.1 Preliminaries

Consider the following general setting of supervised learning.

- An unknown distribution $\mathcal{D} \in \Delta(Z)$. We receive a sample $S = (z_1, ..., z_n)$ of n examples drawn i.i.d. from $\mathcal{D}$.
- The goal is to find a model w with small population risk, defined as: $R[w] := \mathbb{E}_{z\sim\mathcal{D}} f(w; z)$, where $f(w; z)$ is the loss of the model parameterized by w encountered on example z.
- However, we cannot measure $R[w]$ directly. The idea is to use a sample-averaged proxy, the empirical risk, defined as $R_S[w] := \frac{1}{n}\sum_{i=1}^{n} f(w; z_i)$.

Definition 1 *A randomized algorithm A is ϵ-uniformly stable if for all data sets $S, S' \in \mathbb{Z}^n$ such that S and S' differ in at most one example, we have*

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

Recall the important theorem that uniform stability implies generalization in expectation — if an algorithm is uniformly stable, then its generalization error is small.

Theorem 2 *[2] Let algorithm A be ϵ-uniformly stable. Then,*

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

Consider a general updating rule $G : \Omega \to \Omega$. For example, gradient descent update rule or stochastic gradient descent.

Definition 3 *An update rule is η-expansive if for all $v, w \in \Omega$, $\|G(v) - G(w)\| \leq \eta \|v - w\|$[1]. It is σ-bounded if $\|w - G(w)\| \leq \sigma$.*

Definition 4 *A function $f : \Omega \to \mathbb{R}$ is β-smooth if for all $u, v \in \Omega$, we have $\|\nabla f(u) - \nabla f(v)\| \leq \beta \|u - v\|$.*

Theorem 5 *Assume that f is L-Lipschitz. Then the gradient update G is (αL)-bounded.*

Proof: $G(w) = w - \alpha\nabla f(w)$. By Lipschitz condition, $\|w - G(w)\| = \|\alpha\nabla f(w)\| \leq \alpha L$. □

[1] If not specified, we consider 2-norm in this note.

Theorem 6 *If f is β-smooth. The following properties hold.*

- *if f is convex and $\alpha < 2/\beta$, then G is 1-expansive.*
- *if f is γ-strongly convex and $\alpha \le \frac{2}{\beta+\gamma}$, then G is $\left(1 - \frac{\alpha\beta\gamma}{\beta+\gamma}\right)$-expansive.*

Proof:

- Convexity and β-smooth implies the gradients are co-coercive, namely

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

  To see why it is true, on can refer to this link[2]. Then

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

- Refer to [2] for a detailed proof.

□

1.2 Convex Optimization

Theorem 7 *$f(\cdot; z)$ is β-smooth, convex, and L-Lipchitz. If step size $\alpha_t \le 2/\beta$, then*

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

Proof: Let S and S' be two samples of size n differing in only a single example. Consider the stochastic gradient updates $G_1, \cdots, G_T$ and $G'_1, \cdots, G'_T$ induced by running SGM on sample S and S' , respectively. Let w_T and w'_T denote the corresponding outputs. Let $\delta_t = \|w_t - w'_t\|$. For each step t, there are two cases:

- The examples sampled by SGM are the same one w.p. $1 - \frac{1}{n}$. In this case, the function form of G_t and G'_t are the same. We can use the 1-expansivity of G_t and $\alpha_t \le 2/\beta$ s.t. $\delta_{t+1} \le \delta_t$.
- The examples sampled are different w.p. $\frac{1}{n}$. In this case, by using the $(\alpha_t L)$-boundness and 1-expansivity of G_t and G'_t, we have

$$\delta_{t+1} = \|G_t(w_t) - G'_t(w'_t)\| \le \|G_t(w_t) - G_t(w'_t)\| + \|G_t(w'_t) - w'_t\| + \|w'_t - G'_t(w'_t)\| \le 2\alpha_t L + \delta_t.$$

In summary:

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

Thus $\mathbb{E}[\delta_T] \le \frac{2L}{n} \sum_{t=1}^{T} \alpha_t$, since $f(\cdot; z)$ is Lipchitz, $\mathbb{E}\left|f(w_T; z) - f(w'_T; z)\right| \le L\mathbb{E}[\delta_T] \le \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t$.
□

[2] http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf

1.3 Strongly Convex Optimization

Consider the projected stochastic gradient method $w_{t+1} = \Pi_\Omega(w_t - \alpha_t \nabla f(w_t; z_t))$, where Π_Ω is the Euclidean projection onto the set Ω, namely $\Pi_\Omega(v) = \arg\min_{w\in\Omega} \|w - v\|$.

Theorem 8 *Assume that the loss function $f(\cdot; z)$ is γ-strongly convex and β-smooth for all z. Suppose we run the projected SGM iteration with constant step size $\alpha \le 1/\beta$ for T steps. Then, SGM satisfies uniform stability with*

$$\epsilon_{stab} \le \frac{2L^2}{\gamma n}$$

where L is at most $\beta diam(\Omega)$.

Proof: First show $f(\cdot; z)$ is L-Lipchitz and L is finite and bounded. Let $L = \sup_{w\in\Omega} \sup_z \|\nabla f(w; z)\|$. Since $f(\cdot; z)$ is β-smooth and convex, for $\forall w \in \Omega$, $\|\nabla f(w; z) - \nabla f(w^*; z)\| \le \beta \|w - w^*\|$ where w^* is the minimizer. Thus $\|\nabla f(w; z)\| \le \beta \text{diam}(\Omega)$. Thus L is at most $\beta\text{diam}(\Omega)$.

Let $\delta_t = \|w_t - w_t'\|$. For each step t, there are two cases:

- The examples sampled by projected SGM are the same one w.p. $1 - \frac{1}{n}$. In this case, the function form of G_t and G_t' are the same. Note that by Theorem 6, if $\alpha \le 1/\beta$, $\frac{2\alpha\beta\gamma}{\beta+\gamma} \ge \alpha\gamma$ and $\alpha\gamma < 1$, thus G_t is $(1 - \alpha\gamma)$-expansive and

$$\delta_{t+1} = \|\Pi_\Omega(G_t(w_t) - \Pi_\Omega(G_t w_t')\| \le \|G_t(w_t) - G_t(w_t')\| \le (1 - \alpha\gamma)\delta_t.$$

- The examples sampled are different w.p. $\frac{1}{n}$. In this case, by using the (αL)-boundness and $(1 - \alpha\gamma)$-expansivity of G_t and G_t', we have

$$\delta_{t+1} \le \|G_t(w_t) - G_t'(w_t')\| \le \|G_t(w_t) - G_t(w_t')\| + \|G_t(w_t') - w_t'\| + \|w_t' - G_t'(w_t')\| \le 2\alpha_t L + (1-\alpha\gamma)\delta_t.$$

In summary, $\mathbb{E}[\delta_{t+1}] \le (1 - \alpha\gamma)\mathbb{E}[\delta_t] + \frac{2\alpha L}{n}$ and $\mathbb{E}[\delta_T] \le \frac{2\alpha L}{n} \sum_{t=1}^{T} (1 - \alpha\gamma)^t \le \frac{2L}{\gamma n}$. Since $f(\cdot; z)$ is Lipchitz,

$$\mathbb{E}\,|f(w_T; z) - f(w_T'; z)| \le L\mathbb{E}[\delta_T] \le \frac{2L^2}{\gamma n}.$$

□

2 Stability Bound for Stochastic Gradient Langevin Dynamics

In this section, we introduce the stability bound for stochastic gradient Langevin Dynamics (SGLD). We first introduce what is SGLD[3]. Consider SGM where at each step t we sample i_t i.i.d. uniformly from $[n]$ and perform the following updating rule:

$$w_{t+1} = w_t - \alpha \nabla f(w_t; z_{i_t}) = w_t - \alpha \nabla f(w_t) + S_t,$$

where $S_t = \alpha\nabla f(w_t) - \alpha \nabla f(w_t; z_{i_t})$ can been viewed as some noise with 0 mean. We then make an assumption that S_t is a Gaussian with 0 mean and unit variance, i.e. $w_{t+1} = w_t - \alpha\nabla f(w_t) + \mathcal{N}(0, 1)$. We name it as Stochastic Gradient Langevin Dynamics (SGLD). It has a strong connection with Langevin dynamic: $dw = -\alpha\nabla f(w)dt + dB_t$.

[3]This section is a subset of `http://iiis.tsinghua.edu.cn/~jianli/courses/ATCS2018spring/gen-error-bounds.pdf`.

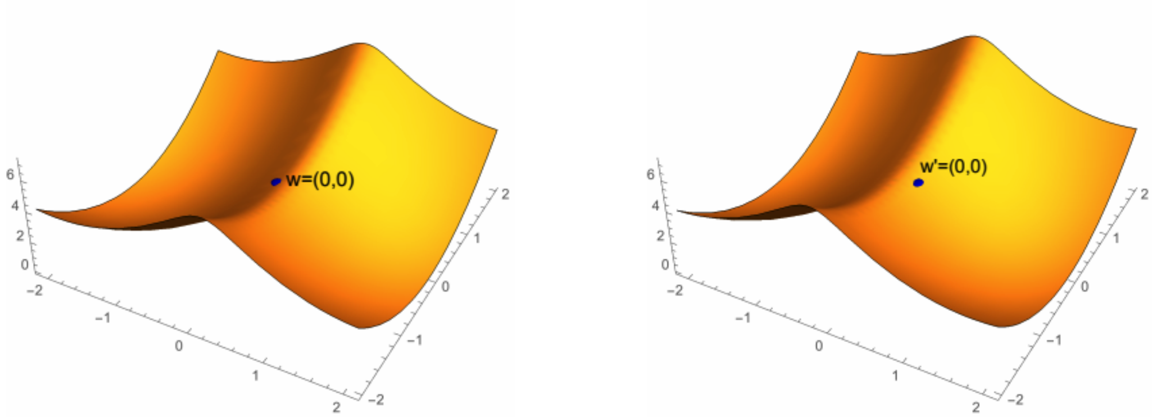

Figure 1: An example from [3]. if the initial point is close to the saddle point, small shift on the loss surface leads to completely different local minimum.

Assumption 9 *Each loss function $f(\cdot; z)$ is differentiable, C-bounded and L-lipschitz.*

Note that we do not assume the loss function to be convex anymore. Under this assumption, traditional SGM is not stable anymore. See Figure 1 for an example.

Theorem 10 *Consider two Markov Chain $(w_0, w_1, \cdots, w_T)$ and $(w'_0, w'_1, \cdots, w'_T)$ with $w_0 = w'_0$. If for $\forall w$, $\mathrm{KL}\left(w_t | w_{t-1} = w || w'_t | w'_{t-1} = w \right) \leq \alpha_t$, then $\mathrm{KL}\left(w_T || w'_T \right) \leq \sum_{t=1}^{T} \alpha_t$.*

Proof: Suppose we have two joint distribution $p(x, y)$ and $q(x, y)$ for r.v. (x, y) and (x', y'), respectively. We have the following observation:

$$\begin{aligned}
\mathrm{KL}\left((x, y) || (x', y') \right) &= \int p(x, y) \log \frac{p(x, y)}{q(x, y)} dx dy \\
&= \int p(x) \log \frac{p(x)}{q(x)} dx + \int \left(\int p(y|x) \log \frac{p(y|x)}{q(y|x)} dy \right) dx \\
&= \mathrm{KL}\left(x || y \right) + \mathbb{E}_{x_0} [\mathrm{KL}\left(y | x = x_0 || y' | x' = x_0 \right)].
\end{aligned}$$

Due to the non-negativity of KL-Divergence,

$$\begin{aligned}
\mathrm{KL}\left(w_t || w'_t \right) &\leq \mathrm{KL}\left((w_{t-1}, w_t) || (w'_{t-1}, w'_t) \right) \\
&= \mathrm{KL}\left(w_{t-1} || w'_{t-1} \right) + \mathbb{E}_w [\mathrm{KL}\left(w_t | w_{t-1} = w || w'_t | w'_{t-1} = w \right)] \\
&\leq \mathrm{KL}\left(w_{t-1} || w'_{t-1} \right) + \alpha_t.
\end{aligned}$$

Thus $\mathrm{KL}\left(w_T || w'_T \right) \leq \sum_{t=1}^{T} \alpha_t$. □

Theorem 11 *Under Assumption 9, for any t and w,*

$$\mathrm{KL}\left(w_t | w_{t-1} = w || w'_t | w'_{t-1} = w \right) \leq \frac{4\alpha^2 L^2}{n^2}.$$

Proof: Let $\mu = w - \alpha\nabla f(w)$ and $\mu' = w - \alpha\nabla f'(w)$. Since f and f' differs by a single L-Lipchitz function,

$$\|\mu - \mu'\| = \alpha\,\|\nabla f(w) - \nabla f'(w)\| \leq \frac{2\alpha L}{n}. \tag{1}$$

Since the conditional distributions of w_t and w_t' are given by $\mathcal{N}(\mu, I)$ and $\mathcal{N}(\mu', I)$, respectively. Following the property of Gaussian, we have

$$\mathrm{KL}\left(\, w_t|w_{t-1} = w||\, w_t'|w_{t-1}' = w\right) \leq \|\mu - \mu'\|^2 = \frac{4\alpha^2 L^2}{n^2}.$$

□

Then for any C-bounded loss function f:

$$\begin{aligned}
|\mathbb{E}[f(w_T)] - \mathbb{E}[f(w_T')]| &= \left|\int (p(w)f(w) - q(w)f(w))\,dw\right| \\
&\leq C \cdot \int |p(w) - q(w)|\,dw = C \cdot \mathrm{TV}(w_T, w_T') \\
&\leq C \cdot \sqrt{\frac{1}{2}\mathrm{KL}\left(\, w_T||\, w_T'\right)} \\
&\leq \frac{\alpha L C\sqrt{2T}}{n},
\end{aligned}$$

where the first inequality is due to C-boundness, the second inequality is due to Pinsker Inequality [1].

References

[1] Pinsker's inequality — Wikipedia, the free encyclopedia, 2018.

[2] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[3] Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. *arXiv preprint arXiv:1707.05947*, 2017.