Algorithmic Stability Analysis

Mathematical Analysis of Machine Learning Algorithms
(Chapter 7)

Algorithmic Stability

We consider a more general setting where the training algorithm may not necessarily correspond to an ERM method.

We are still interested in bounding the difference of training error and generalization of such an algorithm. We introduce the notation of algorithmic stability as follows.

Definition 1

An algorithm $\mathcal{A}$ is ϵ-uniformly stable if for all $\mathcal{S}_n$ and $\mathcal{S}'_n$ that differ by only one element:

$$\sup_{z \in \mathcal{Z}} [\mathbb{E}_{\mathcal{A}} \phi(\mathcal{A}(\mathcal{S}'_n), z) - \mathbb{E}_{\mathcal{A}} \phi(\mathcal{A}(\mathcal{S}_n), z)] \leq \epsilon,$$

where $\mathbb{E}_{\mathcal{A}}$ denotes the expectation over the internal randomization of the algorithm.

Expected Generalization Bound

Theorem 2 (Thm 7.2)

If an algorithm $\mathcal{A}$ is ϵ-uniformly stable, then for $\mathcal{S}_n \sim \mathcal{D}^n$:

$$\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), \mathcal{D}) \leq \mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) + \epsilon.$$

Proof of Theorem 2

Consider two independent samples of size n: $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$ and $\mathcal{S}'_n = \{Z'_1, \ldots, Z'_n\}$. Let $\mathcal{S}_n^{(i)} = \{Z_1, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_n\}$. Let $p_t^{(i)}$ be the distribution obtained by $\mathcal{A}$ with $\mathcal{S}_n^{(i)}$. We have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), \mathcal{D}) - \mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) \\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{S}'_n}\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n^{(i)}), Z_i) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), Z_i) \\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{S}'_n}\mathbb{E}_{\mathcal{S}_n}[\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n^{(i)}), Z_i) - \mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), Z_i)] \leq \epsilon.
\end{aligned}
$$

The first equation used the fact that Z_i is independent of $\mathcal{S}_n^{(i)}$, and thus the distribution of $\phi(\mathcal{A}(\mathcal{S}_n^{(i)}), Z_i)$ is the same as that of $\phi(\mathcal{A}(\mathcal{S}_n), Z)$ with $Z \sim \mathcal{D}$. The inequality used the definition of uniform stability.

Large Probability Bound

Theorem 3 (Thm 7.3)

Assume that $\mathcal{A}$ is ϵ uniformly stable. Let $S_n = \{Z_1, \ldots, Z_n\} \sim \mathcal{D}^n$ and $S'_n = \{Z'_1, \ldots, Z'_n\} \sim \mathcal{D}^n$ be independent training and validation sets of iid data from $\mathcal{D}$. Assume that for some $\delta \in (0, 1)$, we have the following inequality between the expected validation loss and the expected test loss. With probability at least $1 - \delta$,

$$\mathbb{E}_{S_n}\, \mathbb{E}_{\mathcal{A}}\, \phi(\mathcal{A}(S_n), \mathcal{D}) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{S_n}\, \mathbb{E}_{\mathcal{A}}\, \phi(\mathcal{A}(S_n), Z'_i) + \epsilon_n(\delta). \qquad (1)$$

Then with probability at least $1 - \delta$:

$$\begin{aligned} \mathbb{E}_{\mathcal{A}} \phi(\mathcal{A}(S_n), \mathcal{D}) \leq & \mathbb{E}_{\mathcal{A}} \phi(\mathcal{A}(S_n), S_n) \\ & + \epsilon_n(\delta/2) + (2 + 5\lceil \log_2 n \rceil)\epsilon \ln(2/\delta) + 2\epsilon. \end{aligned}$$

Example

Example 4 (Expl 7.4)

For bounded loss $\phi(\cdot, \cdot) \in [0, 1]$, we can apply the additive Chernoff bound and let

$$\epsilon_n(\delta) = \sqrt{\frac{\ln(1/\delta)}{2n}}$$

in (1). This leads to the following inequality. With probability at least $1 - \delta$:

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), \mathcal{D}) \leq & \mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) + (2 + 5\lceil \log_2 n \rceil)\epsilon \ln(2/\delta) \\ & + 2\epsilon + \sqrt{\frac{\ln(2/\delta)}{2n}}.\end{aligned}$$

Leave-One-Out Stability

Definition 5

Given datasets $\mathcal{S}_n = \{Z_1, \ldots, Z_n\} \subset \mathcal{S}_{n+1} = \{Z_1, \ldots, Z_n, Z_{n+1}\}$. Let $\epsilon(\cdot, \cdot)$ be a function $\mathcal{Z} \times \mathcal{Z}^{n+1} \to \mathbb{R}$. The algorithm $\mathcal{A}(\mathcal{S}_n)$ is $\epsilon(\cdot, \cdot)$ leave-one-out stable if there exists $\bar{\mathcal{A}}(\mathcal{S}_{n+1})$ such that for all $(Z_{n+1}, \mathcal{S}_{n+1})$:

$$\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), Z_{n+1}) - \mathbb{E}_{\bar{\mathcal{A}}}\phi(\bar{\mathcal{A}}(\mathcal{S}_{n+1}), Z_{n+1}) \leq \epsilon(Z_{n+1}, \mathcal{S}_{n+1}),$$

where $\mathbb{E}_{\mathcal{A}}$ denotes the expectation over the internal randomization of the algorithm.

Generalization Bound

Theorem 6 (Thm 7.7)

If an algorithm $\mathcal{A}$ is $\epsilon(\cdot,\cdot)$-leave-one-out stable, then

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n),\mathcal{D}) \leq & \mathbb{E}_{\mathcal{S}_{n+1}}\mathbb{E}_{\bar{\mathcal{A}}}\phi(\bar{\mathcal{A}}(\mathcal{S}_{n+1}),\mathcal{S}_{n+1}) \\ & + \mathbb{E}_{\mathcal{S}_{n+1}}\frac{1}{n+1}\sum_{Z\in\mathcal{S}_{n+1}}\epsilon(Z,\mathcal{S}_{n+1}).\end{aligned}$$

Convexity

Definition 7

Given $\lambda \geq 0$. A function $\phi(w)$ is λ-strongly convex in w if for all $w, w' \in \Omega$:

$$\phi(w') \geq \phi(w) + \nabla\phi(w)^\top (w' - w) + \frac{\lambda}{2}\|w - w'\|_2^2.$$

A function $\phi(w)$ is convex if it is 0-strongly convex.

Properties of convexity can be found in Appendix A.

Stability of ERM under Strong Convexity

Theorem 8 (Simplification with $h(\cdot) = 0$, Thm 7.8)

Assume that $\phi(w, z)$ is $G(z)$-Lipschitz in w on a closed convex set Ω. The training loss $\phi(w, \mathcal{S}_n)$ is λ strongly convex. Then the regularized empirical risk minimization method

$$\mathcal{A}(\mathcal{S}_n) = \arg\min_{w \in \Omega} \phi(w, \mathcal{S}_n)$$

is $\epsilon(Z_{n+1}, \mathcal{S}_{n+1}) = G(Z_{n+1})^2/(\lambda(n+1))$ leave-one-out stable. If moreover we have $\sup_z G(z) \leq G$, then it is $\epsilon = 2G^2/(\lambda n)$ uniformly stable.
Then the following expected oracle inequality holds:

$$\mathbb{E}_{\mathcal{S}_n} \phi(\mathcal{A}(\mathcal{S}_n), \mathcal{D}) \leq \inf_{w \in \Omega} [\phi(w, \mathcal{D})] + \frac{\mathbb{E}_Z G(Z)^2}{\lambda(n+1)}.$$

Example: Stability of SVM

We consider the binary linear support vector machine (SVM) formulation with $y \in \{\pm 1\}$, which employs the hinge loss

$$L(f(w,x),y) = \max(1 - f(w,x)y, 0), \quad g(w) = \frac{\lambda}{2}\|w\|_2^2,$$

with linear function class $\{f(w,x) = w^\top \psi(x) : w \in \mathbb{R}^d\}$, where $\psi(x) \in \mathbb{R}^d$ is a known feature vector. Let

$$\phi(w,z) = L(f(w,x),y) + g(w).$$

We will prove $G(Z) = \|\psi(X)\|_2 + \sqrt{2\lambda}$ Lipschitz result for the ERM solutions, which implies the following result from Theorem 8.

SVM Generalization Bound

$$\mathbb{E}_{\mathcal{S}_n}\phi(\mathcal{A}(\mathcal{S}_n),\mathcal{D}) \leq \inf_{w\in\mathbb{R}^d} \phi(w,\mathcal{D}) + \frac{\mathbb{E}_X(\|\psi(X)\|_2 + \sqrt{2\lambda})^2}{\lambda(n+1)}.$$

Stability of SVM: Lipschitz Constant for ERM Solutions

The loss $\phi(w, z) = L(f(w, x), y) + g(w)$ is λ strongly convex. Moreover, the empirical minimizer $\mathcal{A}(\mathcal{S}_n)$ satisfies

$$\phi(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) \leq \phi(0, \mathcal{S}_n) = 1.$$

Therefore $\|\mathcal{A}(\mathcal{S}_n)\|_2 \leq \sqrt{2/\lambda}$. This implies that we may consider the restriction of SVM to

$$\Omega = \left\{ w : \|w\|_2 \leq \sqrt{2/\lambda} \right\}$$

without changing the solution. It is clear that on Ω, $\phi(w, Z)$ with $Z = (X, Y)$ is

$$G(Z) = \|\psi(X)\|_2 + \sqrt{2\lambda}$$

Lipschitz.

Stochastic Gradient Descent (SGD)

One advantage of stability analysis is that it can be applied to computational procedures such as SGD, which cannot be handled directly by empirical process analysis.

Let Ω be a convex set, the projected SGD method is described below.

Algorithm 1: Stochastic Gradient Descent Algorithm

Input: $\mathcal{S}_n$, $\bar{\phi}(w, z)$, w_0, learning rates $\{\eta_t\}$

Output: w_T

1 **for** $t = 1, 2, \ldots, T$ **do**

2 Randomly pick $Z \sim \mathcal{S}_n$

3 Let $w_t = \mathrm{proj}_\Omega(w_{t-1} - \eta_t \nabla \bar{\phi}(w_{t-1}, Z))$

4 where $\mathrm{proj}_\Omega(v) = \arg\min_{u \in \Omega} \|u - v\|_2^2$

Return: w_T

Contraction of SGD

Definition 9

A function $\bar{\phi}(w) : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and λ-strongly convex if $\forall w, w' \in \mathbb{R}^d$:

$$\frac{\lambda}{2}\|w - w'\|_2^2 \leq \bar{\phi}(w') - \bar{\phi}(w) - \nabla\bar{\phi}(w)^\top(w' - w) \leq \frac{L}{2}\|w - w'\|_2^2.$$

The following contraction property of SGD used to derive its stability.

Lemma 10 (SGD contraction, Lem 7.12)

Assume $\bar{\phi}(w)$ is an L-smooth and λ-strongly convex function of w on $\mathbb{R}^d$. Then for all $w, w' \in \mathbb{R}^d$ and $\eta \in [0, 1/L]$:

$$\|\mathrm{proj}_\Omega(w - \eta\nabla\bar{\phi}(w)) - \mathrm{proj}_\Omega(w' - \eta\nabla\bar{\phi}(w'))\|_2 \leq (1 - \lambda\eta)\|w - w'\|_2.$$

Uniform Stability of SGD

We have the following uniform stability result for the SGD procedure under convex loss functions.

Theorem 11 (Thm 7.13)

Assume that $\bar{\phi}(w, z) = \phi(w, z) + h(w)$ is λ-strongly convex and L-smooth in w on $\mathbb{R}^d$. Moreover, assume $\phi(w, z)$ is G Lipschitz on Ω. Define $b_0 = 0$, and for $t \geq 1$:

$$b_t = (1 - \eta_t \lambda) b_{t-1} + \frac{2\eta_t}{n} G^2,$$

where $\eta_t \in [0, 1/L]$. Then after T steps, Algorithm 1 is $\epsilon = b_T$ uniformly stable with respect to $\phi(w, z)$.

The result also holds for an arbitrary convex combination of the form $\sum_{t=0}^{T} \alpha_t w_t$ as the output of Algorithm 1.

Proof of Theorem 11 (I/II)

Let w_t be the intermediate steps of SGD on $\mathcal{S}_n$, and w'_t be the intermediate steps of SGD on $\mathcal{S}'_n = (\mathcal{S}_n \setminus \{Z_n\}) \cup \{Z'_n\}$. We consider a coupling of w_t and w'_t, with the same randomization for w_t and w'_t, except when we choose $Z = Z'_n$ for update of w'_t, we choose $Z = Z_n$ for updating of w_t with i drawn uniformly from $[n]$.
It follows from Lemma 10 that with this coupling, at each t, with probability $\frac{n-1}{n}$, we choose the same Z_i to update both w_t and w'_t:

$$\|w_t - w'_t\|_2 \leq (1 - \lambda\eta_t)\|w_{t-1} - w'_{t-1}\|_2.$$

With probability $1/n$, we have

$$\begin{aligned}\|w_t - w'_t\|_2 \leq & \|[w_{t-1} - \eta_t \nabla\bar{\phi}(w_{t-1}, Z_n)] - [w'_{t-1} - \eta_t \nabla\bar{\phi}(w'_{t-1}, Z_n)]\|_2 \\ & + \eta_t \|\nabla\bar{\phi}(w'_{t-1}, Z_n) - \nabla\bar{\phi}(w'_{t-1}, Z'_n)\|_2 \\ \leq & (1 - \lambda\eta_t)\|w_{t-1} - w'_{t-1}\|_2 + 2G\eta_t,\end{aligned}$$

where i is uniformly from $[n]$. The second inequality used Lemma 10 again.

Proof of Theorem 11 (II/II)

Therefore

$$\mathbb{E}_{\mathcal{A}}\|w_t - w_t'\|_2 \leq (1-\eta_t\lambda)\mathbb{E}_{\mathcal{A}}\|w_{t-1} - w_{t-1}'\|_2 + \frac{2\eta_t G}{n}.$$

We now define

$$s_t = \mathbb{E}_{\mathcal{A}}\,\|w_t - w_t'\|_2 G,$$

then we have

$$s_t \leq (1-\eta_t\lambda)s_{t-1} + \frac{2\eta_t}{n}G.$$

It follows from the definition of b_t that $s_t \leq b_t$. Therefore

$$\mathbb{E}_{\mathcal{A}}\,\|w_T - w_T'\|_2 G \leq b_T. \tag{2}$$

Let $\epsilon(Z) = \mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), Z) - \mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n'), Z)$, then from the Lipschitz condition of $\phi(w, Z)$ and (2), we obtain

$$\epsilon(Z) \leq \mathbb{E}_{\mathcal{A}}\|\mathcal{A}(\mathcal{S}_n) - \mathcal{A}(\mathcal{S}_n')\|_2 G \leq b_T.$$

This proves the desired result.

Example (Example 7.15 in the Book)

Convergence of SGD (Thm 14.5 in the Book)

Consider a constant learning rate η for T steps, and a final estimator w_t from Algorithm 1, with t drawn uniformly from 0 to $T-1$. Algorithm 1 approximately solves the ERM problem with $\lambda = 0$:

$$\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) \leq \inf_{w\in\Omega}\left[\phi(w, \mathcal{S}_n) + \frac{\|w_0 - w\|_2^2}{2T\eta}\right] + \frac{\eta}{2}G^2,$$

where we assume that $\|\nabla\phi(w, z)\|_2 \leq G$.

We can take $b_t = 2\eta t G^2/n$ in Theorem 11 with $\lambda = 0$. This implies a generalization bound

$$\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), \mathcal{D}) \leq \inf_{w\in\Omega}\left[\phi(w, \mathcal{D}) + \frac{\|w_0 - w\|_2^2}{2T\eta}\right] + \frac{\eta}{2}G^2 + \frac{2\eta T G^2}{n}.$$

Note that Example 7.15 uses leave-one-out stability bound (Thm 7.14 in the book), leading to a slightly better result.

Gibbs Algorithm for Nonconvex Loss

For nonconvex problems, ERM is not necessarily stable.

Gibbs Algorithm

Gibbs algorithm is a learning algorithm that randomly draws w from the following "posterior distribution", also referred to as the *Gibbs distribution*:

$$p(w|\mathcal{S}_n) \propto p_0(w) \exp\left(-\beta \sum_{Z \in \mathcal{S}_n} \phi(w, Z)\right), \tag{3}$$

where $\beta > 0$ is a tuning parameter, $p_0(w)$ is a prior on Ω.

The test performance of Gibbs algorithm is measured by the expectation:

$$\mathbb{E}_{\mathcal{A}} \phi(\mathcal{A}(\mathcal{S}_n), \mathcal{D}) = \mathbb{E}_{w \sim p(w|\mathcal{S}_n)} \phi(w, \mathcal{D}).$$

Uniform Stability of Gibbs Algorithm

Theorem 12 (Thm 7.17)

Consider the Gibbs algorithm $\mathcal{A}$ described in (3)*. If for all z:*

$$\sup_{w \in \Omega} \phi(w, z) - \inf_{w \in \Omega} \phi(w, z) \leq M,$$

then $\mathcal{A}$ is $\epsilon = 0.5(e^{2\beta M} - 1)M$ uniformly stable.

Proof of Theorem 12 (I/II)

Consider $\mathcal{S}_n$ and $\mathcal{S}'_n$ that differ by one element. It follows that for any w:

$$\exp(-\beta M) \leq \frac{\exp(-\beta\phi(w, \mathcal{S}'_n))}{\exp(-\beta\phi(w, \mathcal{S}_n))} \leq \exp(\beta M).$$

This implies that

$$\exp(-\beta M) \leq \frac{\mathbb{E}_{w\sim p_0} \exp(-\beta\phi(w, \mathcal{S}'_n))}{\mathbb{E}_{w\sim p_0} \exp(-\beta\phi(w, \mathcal{S}_n))} \leq \exp(\beta M).$$

Therefore

$$\frac{p(w|\mathcal{S}'_n)}{p(w|\mathcal{S}_n)} = \frac{\exp(-\beta\phi(w, \mathcal{S}'_n))}{\exp(-\beta\phi(w, \mathcal{S}_n))} \frac{\mathbb{E}_{w\sim p_0} \exp(-\beta\phi(w, \mathcal{S}_n))}{\mathbb{E}_{w\sim p_0} \exp(-\beta\phi(w, \mathcal{S}'_n))} \leq e^{2\beta M}.$$

Proof of Theorem 12 (II/II)

This implies that

$$\left|\frac{p(w|\mathcal{S}'_n)}{p(w|\mathcal{S}_n)} - 1\right| \leq \max\left(1 - e^{-2\beta M}, e^{2\beta M} - 1\right) \leq e^{2\beta M} - 1.$$

Now let $\bar{\phi}(z) = \inf_w \phi(w, z) + 0.5M$. We know that $|\phi(w, z) - \bar{\phi}(z)| \leq 0.5M$. Therefore

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}'_n), z) - \mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n), z) \\
=&\mathbb{E}_{w\sim p(\cdot|\mathcal{S}_n)}\left(\frac{p(w|\mathcal{S}'_n)}{p(w|\mathcal{S}_n)} - 1\right)[\phi(w, z) - \bar{\phi}(z)] \\
\leq&\mathbb{E}_{w\sim p(\cdot|\mathcal{S}_n)}\left|\frac{p(w|\mathcal{S}'_n)}{p(w|\mathcal{S}_n)} - 1\right| |\phi(w, z) - \bar{\phi}(z)| \\
\leq&(e^{2\beta M} - 1) \cdot 0.5M.
\end{aligned}$$

This proves the desired result.

Example

Example 13 (Expl 7.18)

Consider the Gibbs algorithm $\mathcal{A}$ described in (3) with bounded loss as in Theorem 12. We have the following expected oracle inequality.

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n),\mathcal{D}) \\
\leq&\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(\mathcal{A}(\mathcal{S}_n),\mathcal{S}_n) + 0.5(e^{2\beta M}-1)M \\
\leq&\mathbb{E}_{\mathcal{S}_n}\left[\mathbb{E}_{w\sim p(\cdot|\mathcal{S}_n)}\phi(w,\mathcal{S}_n) + \frac{1}{\beta n}\mathrm{KL}(p(\cdot|\mathcal{S}_n)||p_0)\right] + 0.5(e^{2\beta M}-1)M \\
\leq&\inf_p\left[\mathbb{E}_{w\sim p}\phi(w,\mathcal{D}) + \frac{1}{\beta n}\mathrm{KL}(p||p_0)\right] + 0.5(e^{2\beta M}-1)M.
\end{aligned}
$$

The last inequality used (Eq 7.10) in the book, which states

$$
p(w|\mathcal{S}_n) = \arg\min_{p\in\Delta(\Omega)}\left[\mathbb{E}_{w\sim p}\phi(w,\mathcal{S}_n) + \frac{1}{\beta n}\mathrm{KL}(p||p_0)\right].
$$

Stochastic Gradient Langevin Dynamics

Algorithm 2: Stochastic Gradient Langevin Dynamics Algorithm

Input: $\mathcal{S}_n$, $\bar{\phi}(w, z)$, p_0, learning rates $\{\eta_t\}$

Output: w_T

Draw $w_0 \sim p_0$

for $t = 1, 2, \ldots, T$ **do**

  Randomly pick $Z \sim \mathcal{S}_n$ uniformly at random

  Randomly generate $\epsilon_t \sim N(0, I)$

  Let $\tilde{w}_t = w_{t-1} - \eta_t \nabla \bar{\phi}(w_{t-1}, Z) + \sqrt{2\eta_t/\beta}\epsilon_t$

  Let $w_t = \mathrm{proj}_\Omega(\tilde{w}_t)$, where $\mathrm{proj}_\Omega(v) = \arg\min_{u \in \Omega} \|u - v\|_2^2$

Return: w_T

Similar to SGD, which solves ERM, the stochastic gradient Langevin dynamics (SGLD) algorithm can be used to sample from the Gibbs distribution.

Stability of SGLD: Convex Functions

Theorem 14 (Thm 7.22)

Assume that $\bar{\phi}(w, z) = \phi(w, z) + h(w)$ is λ-strongly convex and L-smooth in w on $\mathbb{R}^d$. Moreover, assume $\phi(w, z)$ is G Lipschitz on Ω. Define $b_0 = 0$, and for $t \geq 1$:

$$b_t = (1 - \eta_t \lambda) b_{t-1} + \frac{2\eta_t}{n} G^2,$$

where $\eta_t \in [0, 1/L]$. Then after T steps, Algorithm 2 is $\epsilon = b_T$ uniformly stable. The result also holds for any random convex combinations of $\{w_t : t \leq T\}$ with combination coefficients from a known distribution.

Proof.

Since the addition of Gaussian noise is independent of the data, the same stability analysis of SGD still holds. □

Stability of SGLD: Non-Convex Functions

It is simpler to analyze the non-stochastic version (often referred to as unadjusted Langevin algorithm, or ULA), where line 5 of Algorithm 2 is replaced by the full gradient

$$\tilde{w}_t = w_{t-1} - \eta_t \nabla \bar{\phi}(w_{t-1}, \mathcal{S}_n) + \sqrt{2\eta_t/\beta}\epsilon_t. \tag{4}$$

Theorem 15 (Thm 7.23)

Assume that for all z, z', $\bar{\phi}(w, z) - \bar{\phi}(w, z')$ is a G-Lipschitz function of w on $\Omega \subset \mathbb{R}^d$ (but $\bar{\phi}$ is not necessarily convex):

$$\|\nabla \bar{\phi}(w, z) - \nabla \bar{\phi}(w, z')\|_2 \leq G.$$

Assume also that $\sup_{w,w' \in \Omega}[\phi(w, z) - \phi(w', z)] \leq M$ for all z. Then after T steps, ULA (with line 5 of Algorithm 2 replaced by (4)) is ϵ_T uniformly stable with $\epsilon_T = \frac{MG}{4n}\sqrt{2\beta \sum_{t=1}^{T} \eta_t}$.

Summary (Chapter 7)

- Stability can be used to derive generalization bound for any algorithm.
- ERM with Strongly Convex Loss is stable.
- SGD with Strongly Convex Loss is stable.
- Gibbs Algorithm with Non-convex Loss is stable.
- Stochastic Gradient Langevin Dynamics