Notes on Generalization Error Bounds

1 Preliminaries

We view a dataset of size n as a collection of n loss functions $\{f_i : i \in [n]\}$, where f_i denotes the loss of a certain parameter configuration on the i-th sample. We make the following assumption on the loss functions.

Assumption 1.1. *Each loss function f_i is differentiable, C-bounded and L-lipschitz.*

The following lemma allows us to reduce the proof of algorithmic stability to the analysis of a single update. Let $\mathrm{KL}(P, Q)$ denote the KL-divergence from Q to P.

Lemma 1.2. *Let $(X_0, X_1, \ldots, X_T)$ and $(X_0', X_1', \ldots, X_T')$ be two Markov chains such that for each $t \in \{0, 1, \ldots, T\}$, X_t and X_t' have the same support. Suppose that the following two conditions hold:*

1. *X_0 and X_0' follow the same distribution.*
2. *For any $t \in [T]$ and any x_0 in the support of X_{t-1}, $\mathrm{KL}\left(X_t | X_{t-1} = x_0, X_t' | X_{t-1}' = x_0\right) \le \alpha_t$.*

Then it holds that

$$\mathrm{KL}\left(X_T, X_T'\right) \le \sum_{t=1}^{T} \alpha_t.$$

Proof. The chain rule of KL-divergence implies that

$$\begin{aligned}
\mathrm{KL}\left(X_t, X_t'\right) &\le \mathrm{KL}\left((X_{t-1}, X_t), (X_{t-1}', X_t')\right) \\
&= \mathrm{KL}\left(X_{t-1}, X_{t-1}'\right) + \mathop{\mathrm{E}}_{x \sim X_{t-1}} \left[\mathrm{KL}\left(X_t | X_{t-1} = x, X_t' | X_{t-1}' = x\right)\right] \\
&\le \mathrm{KL}\left(X_{t-1}, X_{t-1}'\right) + \alpha_t.
\end{aligned}$$

A summation over $t = 1, 2, \ldots, T$ proves the lemma. □

2 Stability Bound for Langevin Monte Carlo

We define Langevin Monte Carlo (LMC) on dataset $S = \{f_i : i \in [n]\}$ as the following procedure:

$$X_{t+1} \leftarrow X_t - \gamma \nabla \overline{f}(X_t) + \zeta_t.$$

Here γ is a step size and $\overline{f} = \frac{1}{n} \sum_{i=1}^{n} f_i$ denotes the average loss on the samples in S. Noise ζ_t is drawn from the standard Gaussian distribution $\mathcal{N}(0, I)$.

We consider two datasets S and S' of size n that differ by at most one loss function. Let $\overline{f}$ and $\overline{f}'$ denote the average loss on samples in S and S', respectively. Let random variables X_t and X_t' denote the parameter after t steps of LMC on datasets S and S', respectively.

The following lemma bounds the contribution of each iteration in LMC to the KL-divergence.

Lemma 2.1. *Under Assumption 1.1, for any time step t and x_0 in the parameter space,*

$$\mathrm{KL}\left(X_t|X_{t-1}=x_0, X'_t|X'_{t-1}=x_0\right) \le \frac{4\gamma^2L^2}{n^2}.$$

Proof. Let $\mu = x_0 - \gamma\nabla\overline{f}(x_0)$ and $\mu' = x_0 - \gamma\nabla\overline{f}'(x_0)$. Since $\overline{f}$ and $\overline{f}'$ differ by a single L-lipschitz loss function,

$$\left\|\nabla\overline{f}(x) - \nabla\overline{f}(x)\right\| \le \frac{2L}{n}.$$

It then follows that $\|\mu - \mu'\| \le \frac{2\gamma L}{n}$. Since the conditional distributions of X_t and X'_t are given by $\mathcal{N}(\mu, I)$ and $\mathcal{N}(\mu', I)$,

$$\mathrm{KL}\left(X_t|X_{t-1}=x_0, X'_t|X'_{t-1}=x_0\right) \le \left\|\mu-\mu'\right\|^2 \le \frac{4\gamma^2L^2}{n^2}.$$

$\square$

By Lemmas 1.2 and 2.1,

$$\mathrm{KL}\left(X_T, X'_T\right) \le \frac{4\gamma^2L^2T}{n^2}.$$

Then a standard argument shows that, for any C-bounded loss function f,

$$\begin{aligned}
|f(X_T) - f(X'_T)| \le& 2C \cdot \mathrm{TV}\left(X_T, X'_T\right) && (C\text{-boundedness})\\
\le& 2C \cdot \sqrt{\frac{1}{2}\mathrm{KL}\left(X_T, X'_T\right)} && (\text{Pinsker's inequality})\\
\le& \frac{\gamma LC\sqrt{8T}}{n}.
\end{aligned}$$

Here $\mathrm{TV}\left(P, Q\right)$ denote the total variation distance between distributions P and Q.

3 Stability Bound for Stochastic Gradient Langevin Dynamics

Stochastic Gradient Langevin Dynamics (SGLD) on dataset $S = \{f_i : i \in [n]\}$ is defined as follows:

$$X_{t+1} \leftarrow X_t - \gamma\nabla f_{i_t}(X_t) + \zeta_t.$$

Here γ is the step size, index i_t is drawn uniformly from $[n]$, and noise ζ_t is drawn from $\mathcal{N}(0, I)$.

Let $S = \{f_1, f_2, \ldots, f_n\}$ and $S' = \{f'_1, f_2, \ldots, f_n\}$ be two datasets of size n that differ by at most one sample. Suppose we run SGLD on both datasets and obtain two sequences of parameters $(X_0, X_1, \ldots)$ and $(X'_0, X'_1, \ldots)$. The following lemma proves a bound for SGLD, similar to Lemma 2.1.

Lemma 3.1. *If $n \ge 2$, $\gamma L \le \frac{1}{10}$, and Assumption 1.1 holds, for any time step t and any point x_0 in the parameter space, it holds that*

$$\mathrm{KL}\left(X_t|X_{t-1}=x_0, X'_t|X'_{t-1}=x_0\right) \le 44\ln 2 \cdot \frac{\gamma^2L^2}{n^2}.$$

Proof. Let $\mu_i = x_0 - \gamma \nabla f_i(x_0)$ for each $i \in [n]$ and $\mu_1' = x_0 - \gamma \nabla f_1'(x_0)$. Since the loss functions are L-lipschitz, μ_1' and each μ_i is in the Euclidean ball of radius γL centered at x_0.

Define probability distributions $A = \frac{1}{n-1}\sum_{i=2}^{n} \mathcal{N}(\mu_i, I)$, $B = \mathcal{N}(\mu_1, I)$ and $C = \mathcal{N}(\mu_1', I)$. Then according to the update rule of SGLD, the conditional distribution of X_t and X_t', denoted by P and P', can be written as

$$P = \frac{1}{n}\sum_{i=1}^{n} \mathcal{N}(\mu_i, I) = \left(1 - \frac{1}{n}\right) A + \frac{1}{n} B \tag{1}$$

and

$$P' = \frac{1}{n}\left(\mathcal{N}(\mu_1', I) + \sum_{i=2}^{n} \mathcal{N}(\mu_i, I)\right) = \left(1 - \frac{1}{n}\right) A + \frac{1}{n} C. \tag{2}$$

By [1, Theorem 3], the KL divergence $\mathrm{KL}\left(P, P'\right)$ is bounded (up to a constant factor) by the *directional triangular discrimination* from P to P', defined as

$$\Delta^*\left(P, P'\right) = \sum_{k=0}^{+\infty} \Delta\left(2^{-k}P + (1-2^{-k})P', P'\right),$$

where each term $\Delta\left(2^{-k}P + (1-2^{-k})P', P'\right)$ is the integral of

$$\frac{\left[2^{-k}P(x) + (1-2^{-k})P'(x) - P'(x)\right]^2}{2^{-k}P(x) + (1-2^{-k})P'(x) + P'(x)} = \frac{4^{-k}(P(x) - P'(x))^2}{2^{-k}P(x) + (2-2^{-k})P'(x)}$$

over the whole parameter space. Plugging (1) and (2) into the integrand gives

$$\frac{4^{-k}\cdot \frac{1}{n^2}(B(x) - C(x))^2}{2(1-\frac{1}{n})A(x) + 2^{-k}\cdot\frac{1}{n}B(x) + (2-2^{-k})\cdot\frac{1}{n}C(x)} \le \frac{4^{-k}}{n^2}\cdot\frac{(B(x)-C(x))^2}{A(x)}.$$

Thus, the directional triangular discrimination from P to P' is bounded by

$$\Delta^*\left(P, P'\right) \le \sum_{k=0}^{+\infty}\int \frac{4^{-k}}{n^2}\cdot\frac{(B(x)-C(x))^2}{A(x)}\,\mathrm{d}x = \frac{4}{3n^2}\int\frac{(B(x)-C(x))^2}{A(x)}\,\mathrm{d}x.$$

It remains to prove that the integral of $\frac{(B(x)-C(x))^2}{A(x)}$ over the parameter space is upper bounded by $44\gamma^2L^2$ under the following conditions:

1. A is a mixture of Gaussian distributions, each with covariance matrix I.
2. B and C are Gaussian distributions with covariance matrix I.
3. There exists a ball of radius γL that contains the means of all Gaussian distribution mentioned above.

Note that the term $\frac{(B(x)-C(x))^2}{A(x)}$ is convex in $A(x)$, so it suffices to consider the case where $A(x)$ is a single Gaussian distribution. The proof for this part is technical and relegated to Lemma A.1 in Appendix A.

Therefore, we conclude that

$$\mathrm{KL}\left(P, P'\right) \le \ln 2\cdot\Delta^*\left(P, P'\right) \le \frac{4\ln 2}{3n^2}\cdot 33\gamma^2L^2 = 44\ln 2\cdot\frac{\gamma^2L^2}{n^2}.$$

□

A Missing Proofs in Section 3

Lemma A.1. *Let $A = \mathcal{N}(\mu_A, I)$, $B = \mathcal{N}(\mu_B, I)$ and $C = \mathcal{N}(\mu_C, I)$ be three Gaussian distributions on $\mathbb{R}^d$ such that μ_A, μ_B, μ_C are in a Euclidean ball of radius $R \in \left[0, \frac{1}{10}\right]$. Then it holds that*

$$\int_{\mathbb{R}^d} \frac{(B(x) - C(x))^2}{A(x)}\ dx \leq 33R^2.$$

Proof of Lemma A.1. By applying a translation and a rotation, we could assume without loss of generality that $\mu_A = 0$, and the last $d-2$ coordinates of μ_B and μ_C are all zero. Observe that the integral is unchanged when we project the space to the two-dimensional subspace corresponding to the first two coordinates. Thus, it suffices to prove the lemma for $d = 2$.

Let x be a point in $\mathbb{R}^d$ with $\|x\| = r$. Observe that $\|x - \mu_A\| = r$ and

$$\|x - \mu_B\|, \|x - \mu_C\| \in [\max(r - 2R, 0), r + 2R].$$

Thus, the term $\frac{(B(x)-C(x))^2}{A(x)}$ is upper bounded by:

$$\frac{1}{2\pi} \cdot \frac{\left[e^{-\frac{\max(r-2R,0)^2}{2}} - e^{-\frac{(r+2R)^2}{2}}\right]^2}{e^{-\frac{r^2}{2}}}.$$

Therefore, we can bound the integral by

$$\begin{aligned}\int_{\mathbb{R}^2} \frac{(B(x) - C(x))^2}{A(x)}\ \mathrm{d}x \leq& \frac{1}{2\pi} \int_0^{+\infty} \frac{\left[e^{-\frac{\max(r-2R,0)^2}{2}} - e^{-\frac{(r+2R)^2}{2}}\right]^2}{e^{-\frac{r^2}{2}}} \cdot 2\pi r\ \mathrm{d}r \\ =& 2\sqrt{2\pi}Re^{4R^2}\left[\mathrm{erf}\left(\sqrt{2} \cdot R\right) + \mathrm{erf}\left(3\sqrt{2} \cdot R\right)\right] + e^{-14R^2} - 2e^{-6R^2} + e^{2R^2}.\end{aligned} \tag{3}$$

Here $\mathrm{erf}(\cdot)$ is the error function defined as $\mathrm{erf}(x) := \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}\ \mathrm{d}t$. Finally, it can be verified that for any $R \in \left[0, \frac{1}{10}\right]$, the right-hand side of (3) is upper bounded by $33R^2$. □

References

[1] Flemming Topsoe. Some inequalities for information divergence and related measures of discrimination. *Transactions on Information Theory (TIT)*, 46(4):1602–1609, 2000.