10-716: Advanced Machine Learning **Spring 2019**

Lecture 17: March 28

Lecturer: Pradeep Ravikumar *Scribes: Ksenia Korovina, Renato Negrinho*

17.1 Stability of Gradient descent

In the previous lecture, we introduced the notion of stability of an algorithm, and showed that an ϵ-stable algorithm is has ϵ-bounded statistical complexity. In this part of the lecture, we will consider stability of Gradient descent (GD) algorithm.

To recap, an algorithm $\mathcal{A}$ is ϵ-stable if, for any samples S and S' that differ at one point, the difference of losses is uniformly bounded by ϵ:

$$\sup_z |\ell(\theta(S), z) - \ell(\theta(S'), z)| \leq \epsilon$$

Gradient descent with step size η performs the following steps for $t = 1 \dots T$:

$$\theta(t) \leftarrow \theta(t-1) - \eta \nabla R_n(\theta(t-1))$$

We will show the following stability bound:

Theorem 17.1 *Assume the following regularity conditions on the loss ℓ:*

1. *Loss ℓ is convex wrt θ*
2. *Gradients of ℓ are β-Lipshitz over parameters: $\|\nabla \ell(\theta, z) - \nabla \ell(\theta', z)\| \leq \beta \|\theta - \theta'\|$*
3. *ℓ is L-Lipshitz over parameters: $|\ell(\theta, z) - \ell(\theta', z)| \leq L\|\theta - \theta'\|$*

*Then, under Assumptions **(1)-(3)** on ℓ, if step size $\eta \leq \frac{1}{\beta}$,*

$$\epsilon_{STAB} \leq 2\eta \frac{L^2 T}{n}$$

Proof: Take $S = \{z_1, \dots, z_n\}$, $S' = \{z_1, \dots, z_i', \dots, z_n\}$. Denote by $\theta(t)$ the GD outcome at step t when run on sample S, and $\theta'(t)$ on sample S'. By Assumption 3, we can bound the end difference of losses at an arbitrary test point z

$$|\ell(\theta(T), z) - \ell(\theta'(T), z)| \leq L\|\theta(T) - \theta'(T)\|_2$$

Now we need to bound the norm of difference between parameters. Note that for any $t \in \{1, \ldots, T\}$,

$$\theta(t) = \theta(t-1) - \eta \frac{1}{n} \sum_{j=1}^{n} \nabla \ell(\theta(t-1), z_j),$$

$$\theta'(t) = \theta'(t-1) - \eta \left(\frac{1}{n} \sum_{j=1}^{n} \nabla \ell(\theta'(t-1), z_j) - \nabla \ell(\theta'(t-1), z_i) + \nabla \ell(\theta'(t-1), z_i') \right)$$

Let us denote the gradient update operator by $\mathcal{G}$:

$$\mathcal{G}(\theta) = \theta - \eta \nabla R_n(\theta)$$

Then by grouping the terms and applying the triangle inequality, we get

$$\|\theta(t) - \theta'(t)\| \leq \|\mathcal{G}(\theta(t-1)) - \mathcal{G}(\theta'(t-1))\| + \|\frac{\eta}{n}[\nabla \ell(\theta'(t-1), z_i) - \nabla \ell(\theta'(t-1), z_i')]\|$$

We use the fact that $\mathcal{G}$ is contractive (see Lemma 17 and Corollary 18 in [CJY] for a proof) to bound the first term:

$$\|\mathcal{G}(\theta(t-1)) - \mathcal{G}(\theta'(t-1))\| \leq \|\theta(t-1) - \theta'(t-1)\|$$

For the second term, we use the L-Lipshitz assumption (3) and the fact that gradient norms of an L-Lipshitz function are bounded by L to get

$$\|\frac{\eta}{n}[\nabla \ell(\theta'(t-1), z_i) - \nabla \ell(\theta'(t-1), z_i')]\| \leq \frac{\eta}{n}\|[\nabla \ell(\theta'(t-1), z_i)\| + \frac{\eta}{n}\|\nabla \ell(\theta'(t-1), z_i')]\| \leq 2\frac{\eta}{n}L.$$

Therefore

$$\|\theta(T) - \theta'(T)\| \leq \frac{2\eta L T}{n}$$

and

$$|\ell(\theta(T), z) - \ell(\theta'(T), z)| \leq \frac{2\eta L^2 T}{n}.$$

■

Moreover, we can show that this bound is tight:

Proposition 17.2 *The bound of Theorem 17.1 is tight: there exists a loss problem that satisfies the conditions **(1)-(3)** and requires stability ϵ at least $2\eta\frac{L^2 T}{n}$.*

Proof: Consider the following loss function:

$$\ell(\theta, z) = \begin{cases} L\theta, z \in C \\ -L\theta, z \in C^c \end{cases}$$

Consider samples $S = \{z_1, \ldots, z_n\}$, and $S' = \{z_1, \ldots, z_i', \ldots, z_n\}, z_j \in C, z_i' \in C^c$. Then

$$R_n(S) = L\theta, \quad R_n(S') = \frac{n-1}{n} L\theta - \frac{L\theta}{n} = \frac{n-2}{n} L\theta$$

Computing derivatives, we see that

$$\theta(T) = -\eta LT, \quad \theta'(T) = -\eta \frac{n-2}{n} LT$$

therefore

$$|\theta(T) - \theta'(T)| = \frac{2LT\eta}{n}.$$

$$|\ell(\theta(T), z) - \ell(\theta'(T), z)| = \frac{2L^2T\eta}{n}$$

■

The bound of Theorem 17.1 can be roughly interpreted as follows: with larger T, larger subsets of the parameter space can be explored, while reducing stability-related generalization guarantees. Based on that, we can examine the consequences of the proved result on choosing the stopping time of gradient descent. Recalling from the convex optimization theory that conditions in Theorem 17.1 also imply optimization error at most $O(\frac{1}{T})$ after T steps of GD, we get a bound on generalization performance of Gradient descent:

$$R(\tilde{\theta}_n) - R(\theta_0^\star) = \epsilon_{GD} \leq \epsilon_{STAB}^{GD} + \epsilon_{OPT}^{GD} \leq 2\eta \frac{L^2 T}{n} + \frac{c}{T}$$

We can minimize this upper bound to get the order of the optimal stopping time as a function of the sample size:

$$T^* \asymp \sqrt{n}$$

and then

$$\epsilon_{GD} \leq \frac{c}{\sqrt{n}},$$

which matches the information theoretic lower bound.

This analysis justifies early stopping in terms of statistical guarantees. In the next section, we provide another insight into this procedure: it turns out that early stopping acts as a regularizer.

17.2 Regularized estimation and gradient descent

In this section, we relate gradient descent to ridge regression, allowing us to transfer well studied statistical properties of regularized estimation to the iterates of gradient descent. We will study the continuous time version of gradient descent described by the differential equation

$$\dot{\theta}(t) = \frac{d}{dt}\theta(t) = -\nabla f(\theta(t))$$

which can be thought as the limit of

$$\theta(t + \Delta t) = \theta(t) - \Delta t \nabla f(\theta(t)),$$

when Δt approaches 0. The initial conditions for the differential equation are $\theta(0) = \theta_0$. We will specifically be interested in the case where $f = R_n$, i.e., the empirical risk R_n.

The regularized estimation path is

$$\underline{\theta}(v) = \arg\inf_{\theta} f(\theta) + \frac{1}{2v}||\theta - \theta_0||_2^2.$$

for $v \in [0, \infty)$. We have that $\theta(0) = \underline{\theta}(0) = \theta_0$ and $\lim_{t\to\infty} \theta(t) = \lim_{v\to\infty} \underline{\theta}(v) = \hat{\theta}$. For $f = R_n$ and $\theta_0 = 0$, we recover the typical ℓ_2 squared regularized estimation problem.

17.2.1 Distance between gradient descent and regularized estimation paths

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a twice differentiable strongly convex function and smooth function with parameters $m, M > 0$, i.e.,$mI \preceq \nabla^2 f(\theta) \preceq MI$ for all $\theta \in \mathbb{R}^p$.

Theorem 17.3 *Let $\hat{\theta}$ be the minimizer of $f(\theta)$, $\kappa = \frac{m}{M}$, and $c = \frac{2\kappa}{\kappa+1}$. Furthermore, let the regularization penalty v be described as a function of t such that $v(t) = \frac{1}{cm}\left(e^{cMt} - 1\right)$. If gradient descent is started at θ_0, we have*

$$||\theta(t) - \underline{\theta}(v(t))||_2 \leq \frac{||\nabla f(\theta_0)||_2}{m}\left(e^{-mt} + \frac{c}{1 - c - e^{-cMt}}\right).$$

When $\kappa = 1$, we have $\underline{\theta}(v(t)) = \theta(t)$ for all $t \in [0, \infty)$. The above theorem gives us an upper bound of $O(e^{-mt} - e^{-Mt})$ for $||\theta(t) - \underline{\theta}(v(t))||_2$.

17.2.2 Excess risk of the iterates of continuous gradient descent

Given n i.i.d. samples $D_n = \{x_i\}_{1=1}^n$, where $x_i \in \mathcal{X}$ drawn from a distribution P. Let $\ell : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ be a loss function assigning cost $\ell(\theta, x)$ for an observation x. Let the population risk be defined as $R(\theta) = \mathbb{E}_{X\sim P}\left[\ell(\theta, X)\right]$ with minimizer θ^*. Our goal is to obtain an estimate $\hat{\theta}$ with low excess risk using D_n, where the excess risk of $\hat{\theta}$ is defined as $R(\hat{\theta}) - \min_\theta R(\theta)$.

For the empirical risk $R_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta, x_i)$, we consider the following regularized estimation problem:

$$\min_{\theta} R_n(\theta) + \frac{1}{2v}||\theta||_2^2.$$

Theorem 17.4 *Suppose the empirical risk $R_n(\theta)$ is m strongly convex and M smooth. Consider the regularized problem with a regularization penalty $\frac{1}{v} \geq 2\frac{||\nabla R_n(\theta^*)||_2}{||\theta^*||}$. Then the optimal solution $\underline{\theta}(v)$ satisfies*

$$||\underline{\theta}(v) - \theta^*||_2 \leq \frac{3}{mv}||\theta^*||_2.$$

Combining Theorem 17.3 and Theorem 17.4 allows us to bound the parameter error of continuous time gradient descent. This is captured in the following theorem:

Theorem 17.5 *Let the conditions of Theorem 17.3 be satisfied. Let $t \leq \frac{1}{cM}\log\left(1 + \frac{cm||\theta^*||}{2||\nabla R_n(\theta^*)||_2}\right)$, where $c = \frac{2m}{m+M}$. Then $\theta(t)$ satisfies the following error bound*

$$||\theta(t) - \theta^*||_2 \leq \frac{||\nabla R_n(\theta_0)||_2}{m}\left(e^{-mt} + \frac{c}{1-c+e^{cMt}}\right) + \frac{3}{c}\frac{e^{-cMt}}{1-e^{-cMt}}||\theta^*||_2.$$

The above theorem proves a family of deterministic error bounds for each value of v and t. The random quantities $m, M, ||\nabla R_n(\theta^*)||_2$ need to be bounded for the specific learning problem under consideration (recall that R_n is a random quantity depending on the data).

Assuming that the population risk $R : \mathbb{R}^p \to \mathbb{R}$ is strongly convex and strongly smooth with parameters $\bar{m}$ and $\bar{M}$, we can prove under certain regularity conditions on distribution P and $R(\theta)$ that $||\nabla R_n(\theta^*)||_2, ||\nabla R_n(0)||_2$ are in $O(\sqrt{\frac{p}{n}})$ and $O(\sqrt{\frac{p}{n}} + ||\theta^*||_2)$, and m, M are with high probability close to $\bar{m}, \bar{M}$. Substituting these results in Theorem 17.5 for

$$t = \frac{1}{\bar{c}\bar{M}}\log\left(1 + \frac{1}{\bar{c}\bar{m}||\theta^*||_2}\sqrt{\frac{n}{p}}\right),$$

gives us

$$||\theta(t) - \theta^*||_2 = O\left(\left(e^{-\bar{m}t} - \bar{c}e^{-\bar{M}t}\right) + \sqrt{\frac{p}{n}}\right),$$

with $\bar{c} = \frac{2\bar{m}}{\bar{m}+\bar{M}}$.

The standard parametric rate of $O(\sqrt{\frac{p}{n}})$ is achieved when $\bar{m} = \bar{M}$ at

$$t = \frac{1}{M}\log\left(1 + \frac{\bar{M}||\theta^*||}{2}\sqrt{\frac{n}{p}}\right).$$

Note that this analysis that gives us insight about early stopping, i.e., rather than run gradient descent to convergence, from a statistical estimation perspective, it makes sense to stop at an earlier time t.

17.3 Statistical query model

Here we briefly introduce the Statistical query model, which includes Empirical risk minimization with gradient descent as a particular example. Assume we have a set of queries $Q = \{q : \mathcal{Z} \to \mathbb{R}^d\}$ and, instead of directly accessing the distribution $\mathbb{P}$, we have access to an oracle that returns a noisy evaluation $z_q \in \mathbb{R}^d$ of the mean given the query, with the guarantee that

$$\mathbb{P}\left(\sup_{q\in Q}|z_q - \mathbb{E}_{z\sim\mathbb{P}}[q(z)]| > \epsilon\right) \leq \delta$$

Let's consider a particular instantiation of this setting: Empirical risk minimization. In it, gradient values at point z are queried to the oracle:

$$q_\theta(z) = \nabla \ell(\theta, z),$$

and an oracle returns sample average (over an "internally stored" sample):

$$\frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta, z_i)$$

which is assumed to be concentrated around $\mathbb{E}\nabla \ell(\theta, z)$. Given this information, we need to minimize empirical risk, while being restricted to calls to the oracle. Using this model allows to establish general lower bounds on sample complexity under a computational budget constraint, as we will see in the next lecture.

References

[W] M. Wainwright, "High Dimensional Statistics," *Prerelease*, 2019

[CJY18] Yuansi Chen, Chi Jin, Bin Yu, "Stability and Convergence Trade-off of Iterative Optimization Algorithms," 2018

[SPR] Arun Sai Suggala, Adarsh Prasad, Pradeep Ravikumar, "Connecting Optimization and Regularization Paths," NeurIPS 2018