Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: hardt+ee227c@berkeley.edu

Graduate Instructor: Max Simchowitz

Email: msimchow+ee227c@berkeley.edu

March 5, 2018

11 Lecture 11: Learning, Stability, Regularization

In this lecture we take a look at machine learning, and empirical risk minimization in particular. We define the distribution of our data as D over $X \times Y$, $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^{m'}$. For instance, in a classification tasks with two labels Y is usually specified as $Y = \{-1, 1\}$.

- A "model" is specified by a set of parameters $w \in \Omega \in \mathbb{R}^n$
- The "loss function" is denoted by $\ell : \Omega \times (X \times Y) \to \mathbb{R}$, note that $\ell(w, z)$ gives the loss of model w on instance z
- Population objective (*Risk*): $R(w) = \mathop{\mathbb{E}}_{z \sim D}[\ell(w, z)]$
- Goal: Find w that minimizes $R(w)$

One way to accomplish this is to use stochastic optimization:

$$w_{t+1} = w_t - \eta \nabla \ell(w_t, z_t) \quad z \in D$$

11.1 Empirical Risk

Suppose $S \in (X \times Y)^m$, $S = ((x_1, y_1), ..., (x_m, y_m))$, and z_i represents the instance $(x_i, y_i), i \in \{1, ..., m\}$.The empirical risk is define as:

$$R_S(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, z_i)$$

Our goal is to minimize this empirical risk. However, what we really want is : $R_S(w) = R(w)$.

- $R(w)$ captures loss on unseen example
- $R_S(w)$ captures loss on seen example

Definition 11.1 (Generalization error)**.**

$$\mathcal{E}_{\text{gen}}(w) = R(w) - R_s(w)$$

Then: $R(w) = R_s(w) + \mathcal{E}_{\text{gen}}(w)$, where optimization can handle the part $R_s(w)$ pretty well.

How do we bound $\mathcal{E}_{\textbf{gen}}(w)$?

- Principle: generalization error = stability
- Stability: How much does your model change if you change one training point

Choose two independent samples $S = (z_1, ..., z_m)$ $\quad S' = (z'_1, ..., z'_m)$. Where S and S' can be completely different, we denote $S^{(i)}$ as:

$$S^{(i)} = (z_1,\ ...,\ z_{i-1},\ \underline{\mathbf{z'_i}},\ z_{i+1},\ \dots, z_m)$$

Definition 11.2 (Average stability)**.** The average stability of an algorithm $A : (X \times Y)^m \rightarrow \Omega$:

$$\Delta(A) = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m}[\ell(A(s), z'_i) - \ell(A(s^{(i)}), z'_i)]]$$

This can be interpreted as performance on something unseen versus something seen.

Theorem 11.3.

$$\mathbb{E}[\mathcal{E}_{(gen)}(A)] = \Delta(A)$$

Proof.

$$\mathbb{E}[\mathcal{E}_{(gen)}(A)] = R(A(s)) - R_s(A(s))$$
$$\mathbb{E}[R_s(A(s))] = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m}\ell(A(s), z_i)]$$
$$\mathbb{E}[R(A(s))] = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m}\ell(A(s), z'_i)]$$

Hence, since $\mathbb{E}[\ell(A(s), z_i)] = \mathbb{E}\ell(A(s^{(i)}), z'_i)$ $\mathbb{E}[R - R_s] = \Delta(A)$ ■

11.2 Uniform Stability

We would now show how stability is related to generalization. For doing so we need to define the concept of *uniform stability*.

Definition 11.4 (Uniform stability)**.** The uniform stability of an algorithm A is defined as

$$\Delta_{\sup}(A) = \sup_{\mathcal{S},\mathcal{S}' \in (\mathcal{X},\mathcal{Y})^m} \sup_{i \in [m]} |\ell(A(S), z_i') - \ell(A(S^{(i)}, z_i')|$$

Corollary 11.5. $\mathbb{E}[\mathcal{E}_{gen}(A)] \leqslant \mathbb{E}[\Delta_{\sup}(A)]$

This corollary turns out to be surprisingly useful since many algorithms are uniformly stable. For instance, strongly convex loss function is sufficient for stability, and hence generalization.

11.3 Empirical Risk Minimization (ERM)

Theorem 11.6. *Assume $\ell(w, z)$ is α-strongly convex with respect $w \in \Omega$ abd L-Lipschitz. Let $\widehat{w}_S = \arg\min_{w \in \Omega} \frac{1}{m} \sum_{i=1}^m l(w, z_i)$. Then, ERM satisfies:*

$$\Delta_{\sup}(ERM) \leqslant \frac{4L^2}{\alpha m} = \mathcal{O}(\frac{1}{m})$$

An interesting point is that there is no explicit reference to the complexity of the class. In the following we present the proof.

Proof. We need to show that $|(\ell(\widehat{w}_{S^{(i)}}, z_i') - \ell(\widehat{w}_S, z_i'))| \leqslant \frac{4L^2}{\alpha m}$.

1. On one hand, by strong convexity we know that $R_S(\widehat{w}_{S^{(i)}}) - R_S(\widehat{w}_S) \geqslant \frac{\alpha}{2} \|\widehat{w}_S - \widehat{w}_{S^{(i)}}\|^2$.

2. On the other hand,

$$\begin{aligned}
&R_S(\widehat{w}_{S^{(i)}}) - R_S(\widehat{w}_S) \\
=&\frac{1}{m}(\ell(\widehat{w}_{S^{(i)}}, z_i) - \ell(\widehat{w}_S, z_i)) + \frac{1}{m}\sum_{i \neq j}(\ell(\widehat{w}_{S^{(i)}}, z_j) - \ell(\widehat{w}_S, z_j)) \\
\leqslant&\frac{1}{m}|\ell(\widehat{w}_{S^{(i)}}, z_i) - \ell(\widehat{w}_S, z_i)| + \frac{1}{m}|(\ell(\widehat{w}_{S^{(i)}}, z_i') - \ell(\widehat{w}_S, z_i'))| + (R_{S^{(i)}}(\widehat{w}_{S^{(i)}}) - R_{S^{(i)}}(\widehat{w}_{S^{(i)}})) \\
\leqslant&\frac{2L}{m}\|\widehat{w}_{S^{(i)}} - \widehat{w}_S\|
\end{aligned}$$

   In the last inequality we have used that $(R_{S^{(i)}}(\widehat{w}_{S^{(i)}}) - R_{S^{(i)}}(\widehat{w}_{S^{(i)}})) \leqslant 0$, and the fact that ℓ is $L-$lipschitz.

Putting it all together $\|\widehat{w}_{S^{(i)}} - \widehat{w}_S\| \leqslant \frac{4L}{\alpha m}$. Then by the Lipschitz condition we have that $\frac{1}{m}|(\ell(\widehat{w}_{S^{(i)}}, z_i') - \ell(\widehat{w}_S, z_i'))| \leqslant L\|\widehat{w}_{S^{(i)}} - \widehat{w}_S\| \leqslant \frac{4L^2}{\alpha m} \Rightarrow \Delta_{\sup}(\mathrm{ERM}) \leqslant \frac{4L^2}{\alpha m}$.

■

11.4 Regularization

Not all the ERM problems are strongly convex. However, if the problem is convex we can consider the regularized objective

$$r(w,z) = \ell(w,z) + \frac{\alpha}{2}\|w\|^2$$

$r(w,z)$ is $\alpha-$strongly convex. The last term is named l2-regularization, weight decay or Tikhonov regularization depending on the field you work on. Therefore, we now have the following chain of implications:

regularization ⇒ strong convexity ⇒ uniform stability ⇒ generalization

We can also show that solving the regularized objective also solves the unregularized objective. Assume that $\Omega \subseteq \mathcal{B}_2(R)$, by setting $\alpha \approx \frac{L^2}{R^2 m}$ we can show that the minimizer of the regularized risk also minimizes the unregularized risk up to error $\mathcal{O}(\frac{LR}{\sqrt{m}})$. Moreover, by the previous theorem the generalized error will also be $\mathcal{O}(\frac{LR}{\sqrt{m}})$. See Theorem 3 in [SSSSS10].

11.5 Implicit Regularization

In implicit regularization the algorithm itself regularizes the objective, instead of explicitly adding a regularization term. The following theorem describes the regularization effect of the Stochastic Gradient Method (SGM).

Theorem 11.7. *Assume $\ell(\cdot,z)$ is convex, β-smooth and L-Lipschitz. If we run SGM for T steps, then the algorithm has uniform stability*

$$\Delta_{\sup}(SGM_T) \leqslant \frac{2L^2}{m}\sum_{t=1}^{T}\eta_t$$

Note for $\eta_t \approx \frac{1}{m}$ then $\Delta_{\sup}(\mathrm{SGM}_T) = \mathcal{O}(\frac{\log(T)}{m})$, and for $\eta_t \approx \frac{1}{\sqrt{m}}$ and $T = \mathcal{O}(m)$ then $\Delta_{\sup}(\mathrm{SGM}_T) = \mathcal{O}(\frac{1}{\sqrt{m}})$. See [HRS15] for proof.

References

[HRS15] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *CoRR*, abs/1509.01240, 2015.

[SSSSS10] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.