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 $X \times Y$, where $X \subseteq \mathbb{R}^d$ and Y is some discrete set of class labels. For instance, in a binary classification tasks with two labels Y might be $Y = \{-1, 1\}$.

- A "model" is specified by a set of parameters $w \in \Omega \subseteq \mathbb{R}^n$
- The "loss function" is denoted by $\ell \colon \Omega \times (X \times Y) \to \mathbb{R}$, note that $\ell(w, z)$ gives the loss of model w on instance z.
- The risk of a model is defined as $R(w) = \mathbb{E}_{z \sim D}[\ell(w, z)]$.

Our goal is to find a model w that minimizes $R(w)$.

One way to accomplish this is to use stochastic optimization directly on the population objective:

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

When given a finite data set, however, it is usually effective to make multiple passes over the data. In this case, the stochastic gradient method may no longer optimize risk directly.

11.1 Empirical risk and generalization error

Consider a finite sample Suppose $S = ((x_1, y_1), ..., (x_m, y_m)) \in (X \times Y)^m$, where $z_i = (x_i, y_i)$ represents the i-th labeled example. The empirical risk is defined as

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

Empirical risk minimization is commonly used as a proxy for minimizing the unknown population risk. But how good is this proxy? Ideally, we would like that the point w that we find via empirical risk minimization has $R_S(w) \approx R(w)$. However, this may not be the case, since the risk $R(w)$ captures loss on unseen example, while the empirical risk $R_S(w)$ captures loss on seen examples. Generally, we expect to do much better on seen examples than unseen examples. This performance gap between seen and unseen examples is what we call *generalization error*.

Definition 11.1 (Generalization error)**.** We define the *generalization error* of a model w as

$$\epsilon_{\text{gen}}(w) = R(w) - R_s(w) \, .$$

Note the following tautological, yet important identity:

$$R(w) = R_S(w) + \epsilon_{\text{gen}}(w) \tag{1}$$

What this shows in particular is that if we manage to make the empirical risk $R_S(w)$ small through optimization, then all that remains to worry about is generalization error.

So, how can we bound generalization error? The fundamental relationship we'll establish in this lecture is that generalization error equals an algorithmic robustness property that we call *algorithmic stability*. Intuitively, algorithmic stability measures how sensitive an algorithm is to changes in a single training example.

11.2 Algorithmic stability

To introduce the idea of stability, we choose two independent samples $S = (z_1, ..., z_m)$ and $S' = (z'_1, ..., z'_m)$, each drawn independently and identically from D. Here, the second sample S' is called a *ghost sample* and mainly serves an analytical purpose.

Correlating the two samples in a single point, we introduce the hybrid sample $S^{(i)}$ as:

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

Note that here the i-th example comes from S', while all others come from S.

With this notation at hand, we can introduce a notion of average stability.

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

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

This definition can be interpreted as comparing the performance of the algorithm on an unseen versus a seen example. This is the intuition why average stability, in fact, equals generalization error.

Theorem 11.3.

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

Proof. Note that

$$\begin{aligned}
\mathbb{E}[\epsilon_{\text{gen}}(A)] &= R(A(S)) - R_S(A(S))\,, \\
\mathbb{E}\left[R_S(A(S))\right]] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(A(S), z_i)\right]\,, \\
\mathbb{E}[R(A(S))] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(A(S), z_i')\right]\,.
\end{aligned}$$

At the same time, since z_i and z_i' are identically distributed and independent of the other examples, we have

$$\mathbb{E}\,\ell(A(S), z_i) = \mathbb{E}\,\ell(A(S^{(i)}), z_i')\,.$$

Applying this identity to each term in the empirical risk above, and comparing with the definition of $\Delta(A)$, we conclude $\mathbb{E}[R(A(S)) - R_S(A(S))] = \Delta(A)$ ■

11.2.1 Uniform stability

While average stability gave us an exact characterization of generalization error, it can be hard to work with the expectation over S and S'. Uniform stability replaces the averages by suprema, leading to a stronger but useful notion [BE02].

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

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

Since uniform stability upper bounds average stability, we know that uniform stability upper bounds generalization error (in expectatation).

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

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 as we will show next.

11.3 Stability of empirical risk minimization

The next theorem due to [SSSSS10] shows that empirical risk minimization of a strongly convex loss function is uniformly stable.

Theorem 11.6. *Assume $\ell(w,z)$ is α-strongly convex over the domain Ω and L-Lipschitz. Let $\widehat{w}_S = \arg\min_{w\in\Omega} \frac{1}{m}\sum_{i=1}^m \ell(w,z_i)$ denote the empirical risk minimizer (ERM). Then, ERM satisfies*

$$\Delta_{\sup}(ERM) \leqslant \frac{4L^2}{\alpha m}\,.$$

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

Proof. Denote by $\widehat{w}_S$ the empirical risk minimizer on a sample S. Fix arbitrary samples S, S' of size m and an index $i \in [m]$. We need to show that

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

On one hand, by strong convexity we know that

$$R_S(\widehat{w}_{S^{(i)}}) - R_S(\widehat{w}_S) \geqslant \frac{\alpha}{2}\|\widehat{w}_S - \widehat{w}_{S^{(i)}}\|^2\,.$$

On the other hand,

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

Here, we used that

$$R_{S^{(i)}}(\widehat{w}_{S^{(i)}}) - R_{S^{(i)}}(\widehat{w}_S)) \leqslant 0$$

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

$$\frac{1}{m}|(\ell(\widehat{w}_{S^{(i)}}, z_i') - \ell(\widehat{w}_S, z_i'))| \leqslant L\|\widehat{w}_{S^{(i)}} - \widehat{w}_S\| \leqslant \frac{4L^2}{\alpha m}\,.$$

Hence, $\Delta_{\sup}(\mathrm{ERM}) \leqslant \frac{4L^2}{\alpha m}$.

■

11.4 Regularization

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

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

The regularized loss $r(w,z)$ is $\alpha-$strongly convex. The last term is named ℓ_2-regularization, weight decay or Tikhonov regularization depending on the field you work on. Therefore, we now have the following chain of implications:

$$\text{regularization} \Rightarrow \text{strong convexity} \Rightarrow \text{uniform stability} \Rightarrow \text{generalization}$$

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] for details.

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

References

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

[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.