ELEG/CISC 867: Advanced Machine Learning **Spring 2019**

Lecture 11: Regularization and Stability

Lecturer: Xiugang Wu *04/30/2019*

Last time, we have seen that convex-Lipschitz/smooth-bounded problems are learnable via stochastic gradient descent. In this lecture, we will introduce a new learning rule, called Regularized Loss Minimization (RLM), and will demonstrate that convex-Lipschitz/smooth-bounded problems can also be learned using RLM; in particular, we show this by interpreting regularization as a stabilizer of the learning algorithm and proving that stable algorithms do not overfit (i.e. have small generalization error).

1 Regularized Loss Minimization

RLM is a learning paradigm where we jointly minimize the empirical risk and a regularization function, i.e.,

$$\min_{\mathbf{w}\in\mathcal{H}} L(\mathbf{w}, Z^n) + R(\mathbf{w}).$$

One can intuitively think of the regularizer $R(\mathbf{w})$ as measuring the complexity of the hypothesis $\mathbf{w}$, and hence the algorithm balance between low empirical risk and less complex hypothesis; recall bias-complexity tradeoff. Another view of regularization is as a stabilizer of the learning algorithm; we will adopt the stability view in this lecture and prove the close relation between stability and learnability.

There are many possible regularizers one can use, reflecting the prior belief about the problem. This lecture will focus on the so-called ℓ_2 regularization or Tikhonov regularization: $\lambda\|\mathbf{w}\|^2$ where $\lambda > 0$ is a scalar and the norm is the ℓ_2 norm. This yields the learning rule:

$$A(Z^n) = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{H}} L(\mathbf{w}, Z^n) + \lambda\|\mathbf{w}\|^2.$$

1.1 Ridge Regression

Applying RLM with Tikhonov regularization to linear regression with square loss, we obtain the so-called ridge regression problem, given by

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 + \lambda\|\mathbf{w}\|^2.$$

To solve this problem we take the gradient of the above and compare it to zero, yielding:

$$(X^TX + n\lambda I)\mathbf{w} = X^Ty^n \tag{1}$$

where

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix},$$

and I is the identity matrix. Since $(X^TX + n\lambda I)$ is always invertible, the solution to the ridge regression problem is given by

$$\mathbf{w} = (X^TX + n\lambda I)^{-1}X^Ty^n. \tag{2}$$

Note that in the above if $\lambda = 0$ then the problem (1) reduces to the ordinary linear regression (OLS) formulation $X^TX\mathbf{w} = X^Ty^n$ and the solution (2) holds with $\mathbf{w} = (X^TX)^{-1}X^Ty^n$ provided that X^TX is invertible.

2 Stable Algorithms Do Not Overfit

We now show how regularization stabilizes the algorithm and prevents overfitting. Roughly speaking, a learning algorithm is stable if a small change of the input to the algorithm does not change the output of the algorithm much. This idea can be formalized as follows. Given the training set Z^n and an additional example Z_{n+1}, let $Z^n_{i/n+1}$ be the training set obtained by replacing the i-th example of Z^n with Z_{n+1}; namely, $Z^n_{i/n+1} = (Z_1, Z_2, \ldots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \ldots, Z_n)$. In our definition of stability, "a small change of the input" means that we feed algorithm A with $Z^n_{i/n+1}$ instead of with Z^n, i.e. we only change one training example. We measure the effect of this small change of the input on the output of A, by comparing the loss of the hypothesis $A(Z^n)$ on Z_i to the loss of the hypothesis $A(Z^n_{i/n+1})$ on Z_i. Intuitively, a good learning algorithm will have $L(A(Z^n_{i/n+1}), Z_i) - L(A(Z^n), Z_i) > 0$ on Z_i, since in the first term the learning algorithm does not observe the example Z_i while in the second term Z_i is indeed observed. If the preceding difference is very large, i.e. the learning algorithm is not stable, we suspect that the learning algorithm might lead to overfitting because the algorithm drastically changes its prediction on Z_i if it observes it in the training set. This relation between stability and overfitting (i.e. generalization error) is captured by the following theorem.

Theorem 2.1

$$\mathbb{E}_{Z^n\sim P^n}[L(A(Z^n), P) - L(A(Z^n), Z^n)] = \mathbb{E}_{Z^{n+1}\sim P^{n+1}, I\in U[1:n]}[\ell(A(Z^n_{I/n+1}), Z_I) - \ell(A(Z^n), Z_I)].$$

Proof: Note that the expected true error satisfies

$$\mathbb{E}[L(A(Z^n), P)] = \mathbb{E}[\ell(A(Z^n), Z_{n+1})] = \mathbb{E}[\ell(A(Z^n_{i/n+1}), Z_i)]$$

for any $i \in [1:n]$, while the training error over any z^n satisfies

$$L(A(z^n), z^n) = L(A(z^n), P_n) = \mathbb{E}[\ell(A(z^n), z_I)].$$

□

Let $\epsilon(n)$ be a monotonically decreasing function. We say that a learning algorithm A is on-average-replace-one-stable with rate $\epsilon(n)$ if for any distribution P,

$$\mathbb{E}_{Z^{n+1}\sim P^{n+1}, I\in U[1:n]}[\ell(A(Z^n_{I/n+1}), Z_I) - \ell(A(Z^n), Z_I)] \le \epsilon(n).$$

The theorem says that a learning algorithm does not overfit if and only if it is on-average-replace-one-stable. Of course, a learning algorithm that does not overfit is not necessarily a good learning algorithm; imagine a bad stable algorithm A that always outputs the same hypothesis so that its expected training error equals the true error. A useful algorithm should find a hypothesis that on one hand fits the training set (i.e., has a low empirical risk) and on the other hand does not overfit. Or, in light of the theorem, the algorithm should both fit the training set and at the same time be stable. As we shall see, the parameter λ of the RLM rule balances between fitting the training set and being stable.

3 Tikhonov Regularization as Stabilizer

We now show that RLM with Tikhonov regularization is a stable learning algorithm given that the loss function is convex Lipschitz/smooth. This is mainly because Tikhonov regularization makes the objective of RLM strongly convex.

Definition 3.1 (Strongly Convex Functions) *A function f is λ-strongly convex if for all $\mathbf{w}, \mathbf{u}$ and $\alpha \in [0,1]$ we have*

$$f(\alpha \mathbf{w} + (1-\alpha)\mathbf{u}) \leq \alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2.$$

The following lemma implies that the objective of RLM is 2λ-strongly convex; it also underscores an important property of strong convexity.

Lemma 3.1 *The function $f(\mathbf{w}) = \lambda\|\mathbf{w}\|^2$ is 2λ-strongly convex. If f is λ-strongly convex and g is convex, then $f+g$ is λ-strongly convex. If f is λ-strongly convex and $\mathbf{u}$ is a minimizer of f, then for any $\mathbf{w}$, $f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\|^2$.*

We are in a position to prove that RLM is stable. Denoting $f(\mathbf{w}) = L(\mathbf{w}, z^n) + \lambda\|\mathbf{w}\|^2$ and using the above lemma, we have

$$f(A(z^n_{i/n+1})) - f(A(z^n)) \geq \lambda \|A(z^n_{i/n+1}) - A(z^n)\|^2.$$

On the other hand, we have

$$\begin{aligned}
f(A(z^n_{i/n+1})) - f(A(z^n)) &= L(A(z^n_{i/n+1}), z^n) + \lambda\|A(z^n_{i/n+1})\|^2 - [L(A(z^n), z^n) + \lambda\|A(z^n)\|^2] \\
&= L(A(z^n_{i/n+1}), z^n_{i/n+1}) + \lambda\|A(z^n_{i/n+1})\|^2 - [L(A(z^n), z^n_{i/n+1}) + \lambda\|A(z^n)\|^2] \\
&\quad + \frac{1}{n}[\ell(A(z^n_{i/n+1}), z_i) - \ell(A(z^n_{i/n+1}), z_{n+1})] - \frac{1}{n}[\ell(A(z^n), z_i) - \ell(A(z^n), z_{n+1})] \\
&\leq \frac{1}{n}[\ell(A(z^n_{i/n+1}), z_i) - \ell(A(z^n), z_i)] + \frac{1}{n}[\ell(A(z^n), z_{n+1}) - \ell(A(z^n_{i/n+1}), z_{n+1})].
\end{aligned}$$

Combining the above, we obtain that

$$\lambda\|A(z^n_{i/n+1}) - A(z^n)\|^2 \leq \frac{1}{n}[\ell(A(z^n_{i/n+1}), z_i) - \ell(A(z^n), z_i)] + \frac{1}{n}[\ell(A(z^n), z_{n+1}) - \ell(A(z^n_{i/n+1}), z_{n+1})]. \quad (3)$$

To proceed, we now need the Lipschitzness or smoothness of the loss function.

3.1 Lipschitz Loss

If ℓ is ρ-Lipschitz, we have

$$\ell(A(z^n_{i/n+1}), z_i) - \ell(A(z^n), z_i) \leq \rho\|A(z^n_{i/n+1}) - A(z^n)\|$$

and

$$\ell(A(z^n), z_{n+1}) - \ell(A(z^n_{i/n+1}), z_{n+1}) \leq \rho\|A(z^n_{i/n+1}) - A(z^n)\|,$$

which lead to

$$\|A(z^n_{i/n+1}) - A(z^n)\| \leq \frac{2\rho}{\lambda n}$$

in light of (3). This in turn implies that

$$\ell(A(z^n_{i/n+1}), z_i) - \ell(A(z^n), z_i) \leq \frac{2\rho^2}{\lambda n}$$

and hence the generalization error can be bounded as

$$\mathbb{E}[L(A(Z^n), P) - L(A(Z^n), Z^n)] = \mathbb{E}[\ell(A(Z^n_{I/n+1}), Z_I) - \ell(A(Z^n), Z_I)] \leq \frac{2\rho^2}{\lambda n}.$$

3.2 Smooth Loss

If ℓ is β-smooth and nonnegative, we have

$$\begin{aligned}
\ell(A(z^n_{i/n+1}), z_i) - \ell(A(z^n), z_i) &\leq \nabla \ell(A(z^n), z_i)^T [A(z^n_{i/n+1}) - A(z^n)] + \frac{\beta}{2}\|A(z^n_{i/n+1}) - A(z^n)\|^2 \\
&\leq \|\nabla \ell(A(z^n), z_i)\| \cdot \|A(z^n_{i/n+1}) - A(z^n)\| + \frac{\beta}{2}\|A(z^n_{i/n+1}) - A(z^n)\|^2 \\
&\leq \sqrt{2\beta \ell(A(z^n), z_i)} \cdot \|A(z^n_{i/n+1}) - A(z^n)\| + \frac{\beta}{2}\|A(z^n_{i/n+1}) - A(z^n)\|^2
\end{aligned}$$

where the first inequality holds because one can upper bound a convex smooth function in terms of its first order Taylor approximation, the second inequality is due to Cauchy-Schwartz inequality, and the last inequality follows from the self-boundedness of smooth functions. Similarly, we have

$$\ell(A(z^n), z_{n+1}) - \ell(A(z^n_{i/n+1}), z_{n+1}) \leq \sqrt{2\beta \ell(A(z^n_{i/n+1}), z_{n+1})} \cdot \|A(z^n_{i/n+1}) - A(z^n)\| + \frac{\beta}{2}\|A(z^n_{i/n+1}) - A(z^n)\|^2.$$

Combining these with (3) we obtain

$$\|A(z^n_{i/n+1}) - A(z^n)\| \leq \frac{\sqrt{2\beta}}{\lambda n - \beta}\left[\sqrt{\ell(A(z^n), z_i)} + \sqrt{\ell(A(z^n_{i/n+1}), z_{n+1})}\right]$$

and therefore, if λ is such that $\beta \leq \lambda n/2$,

$$\begin{aligned}
\ell(A(z^n_{i/n+1}), z_i) - \ell(A(z^n), z_i) &\leq \sqrt{2\beta \ell(A(z^n), z_i)} \cdot \|A(z^n_{i/n+1}) - A(z^n)\| + \frac{\beta}{2}\|A(z^n_{i/n+1}) - A(z^n)\|^2 \\
&\leq \left(\frac{4\beta}{\lambda n} + \frac{8\beta^2}{(\lambda n)^2}\right)\left[\sqrt{\ell(A(z^n), z_i)} + \sqrt{\ell(A(z^n_{i/n+1}), z_{n+1})}\right]^2 \\
&\leq \frac{8\beta}{\lambda n}\left[\sqrt{\ell(A(z^n), z_i)} + \sqrt{\ell(A(z^n_{i/n+1}), z_{n+1})}\right]^2 \\
&\leq \frac{24\beta}{\lambda n}\left[\ell(A(z^n), z_i) + \ell(A(z^n_{i/n+1}), z_{n+1})\right],
\end{aligned}$$

where in the last inequality we have used the fact that $(a+b)^2 \leq 3(a^2+b^2)$. This allows us to bound the generalization error for $\lambda \geq 2\beta/n$:

$$\begin{aligned}
\mathbb{E}[L(A(Z^n), P) - L(A(Z^n), Z^n)] &= \mathbb{E}[\ell(A(Z^n_{I/n+1}), Z_I) - \ell(A(Z^n), Z_I)] \\
&\leq \frac{24\beta}{\lambda n}\left[\ell(A(Z^n), Z_I) + \ell(A(Z^n_{I/n+1}), Z_{n+1})\right] \\
&= \frac{48\beta}{\lambda n}\mathbb{E}[L(A(Z^n), Z^n)].
\end{aligned}$$