Regularization and Stability

Peng Lingwei

August 1, 2019

Contents

13 Regularization and Stability

Regularized Loss Minimization will learn all convex-Lipschitz-bounded and convex-smooth-bounded learning problems.

An algorithm is considered stable if a slight change of its input does not change its output much.It's closed to learnability.

13.1 REGULARIZED LOSS MINIMIZATION

Regularized Loss Minimization (RLM):

$$\operatorname*{arg\,min}_{\mathbf{w}} \left(L_S(\mathbf{w}) + R(\mathbf{w})\right).$$

Tikhonov regularization: $\lambda\|\mathbf{w}\|^2$

A learning rule: $A(S) = \operatorname*{arg\,min}_{\mathbf{w}} \left(L_S(\mathbf{w}) + \lambda\|\mathbf{w}\|^2\right)$ has two interpretation:

- Structural risk minimization. We define $\mathcal{H} = \cup\mathcal{H}_n$, which satisfies: $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{H}_3 \ldots$, where $\mathcal{H}_i = \{\mathbf{w} : \|\mathbf{w}\| \leq i\}$.
- Stabilizer.

13.1.1 Ridge Regression

Definition 13.1. *(ridge regression). Performing linear regression using following equation:*

$$\operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^d} \left(\lambda\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\langle\mathbf{w},\mathbf{x}_i\rangle - y_i)^2\right) \tag{13.1}$$

The solution to ridge regression becomes:

$$\mathbf{w} = (2\lambda m I + A)^{-1}\mathbf{b} \tag{13.2}$$

in which, A is a positive semidefinite matrix.

Theorem 13.1. *Let $\mathcal{X} \times [-1,1] \sim \mathcal{D}$, where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\}$, and $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$. $\forall \epsilon \in (0,1)$, let $m \geq 150B^2/\epsilon^2$. Then, applying the ridge regression algorithm with parameter $\lambda = \epsilon/(3B^2)$ satisfies*

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon.$$

Proof. The proof is in the next section. □

Exercise **??** tells us how an algorithm with a bounded expected risk can be used to construct an agnostic PAC learner.

Example 13.1. ***From Bounded Expected Risk to Agnostic PAC Learning***: *Let A be an algorithm that guarantees the following: If $m \geq m_{\mathcal{H}}(\epsilon)$ then for every distribution $\mathcal{D}$ it holds that*

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

We can get $m_{\mathcal{H}}(\epsilon,\delta)$ from Bounded Expected Risk.

Proof. Step 1: If $m \geq m_{\mathcal{H}}(\epsilon\delta)$, then

$$\mathbb{P}\{L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) > \epsilon\} \leq \frac{1}{\epsilon}\mathbb{E}\{L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)\} \leq \delta$$

Step 2: We devided data into k+1 chunks, which $k = \lceil \log_2(2/\delta) \rceil$. For the first k chunks, each chunk is larger than $m_{\mathcal{H}}(\epsilon/4)$, then we have,

$$\mathbb{P}\{\min_{i\in[k]} L_{\mathcal{D}}(A(S_i)) > \min_{h\in\mathcal{H}} L_D(h) + \epsilon/2\} < \frac{1}{2^k} < \frac{\delta}{2}$$

Step 3: Then we apply ERM over finite class $\{h_1, \ldots, h_k\}$ on the last chunk. If we want get

$$\mathbb{P}\{L_{\mathcal{D}}(A_2(S_{k+1})) > \min_{i\in[k]} L_{\mathcal{D}}(h_k) + \epsilon/2\} < \frac{\delta}{2}$$

we need

$$m \geq m_{\mathcal{H}}(\epsilon/2, \delta/2) \geq m_{\mathcal{H}}^{UC}(\epsilon/4, \delta/2) \geq 8\left\lceil \frac{\log(4/\delta) + \log(\lceil \log_2(2/\delta) \rceil)}{\epsilon^2} \right\rceil$$

Overall, we have

$$m_{\mathcal{H}}(\epsilon,\delta) = m_{\mathcal{H}}(\epsilon/4)\lceil \log_2(2/\delta) \rceil + 8\left\lceil \frac{\log(4/\delta) + \log(\lceil \log_2(2/\delta) \rceil)}{\epsilon^2} \right\rceil$$

□

13.2 STABLE RULES DO NOT OVERFIT

Symbols in following sections:

- Training set: $S = (z_1, \ldots, z_m)$.
- An additional example z'.
- Replacing training set: $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$.
- Uniform distribution over $[m]$: $U(m)$.

Theorem 13.2.

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] = \mathop{\mathbb{E}}_{(S,z')\sim\mathcal{D}^{m+1}, i\sim U(m)}[l(A(S^{(i)}, z_i)) - l(A(S), z_i)] \tag{13.3}$$

Proof. The proof is trivial. □

When the right-hand side of Equation 13.3 is small, we say that A is a stable algorithm. In light of Theorem **??**, the algorithm should both fit the training set and at the same time be stable.

Definition 13.2. *(On-Average-Replace-One-Stable). Let $\epsilon(m) : \mathbb{N} \to \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm A is on-average-replace-one-stable with rate $\epsilon(m)$ if for every distribution $\mathcal{D}$*

$$\mathop{\mathbb{E}}_{(S,z')\sim\mathcal{D}^{m+1}, i\sim U(m)}[l(A(S^{(i)}), z_i) - l(A(S), z_i)] \leq \epsilon(m) \tag{13.4}$$

13.3 TIKHONOV REGULARIZATION AS A STABILIZER

Tikhonov regularization leads to a stable algorithm.

Definition 13.3. *(Strongly Convex Functions).For $\alpha \in (0, 1)$*

$$f(\alpha\mathbf{w} + (1-\alpha)\mathbf{u}) \leq \alpha f(\mathbf{w}) + (1-\alpha)f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w} - \mathbf{u}\|^2 \tag{13.5}$$

We have

$$f(\mathbf{w}) - f(\mathbf{w}^*) \geq \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}^*\|^2.$$

($\mathbf{w}^*$ is minimum point).

Let $A(S) = \arg\min_{\mathbf{w}} \left(L_S(\mathbf{w}) + \lambda\|\mathbf{w}\|^2\right)$, and $f_S(\mathbf{w}) = L_S(\mathbf{w}) + \lambda\|\mathbf{w}\|^2$.Then ($f_S(\mathbf{w})$ is $2\lambda - \textit{strongly}$ convex.)

$$f_S(\mathbf{v}) - f_S(A(S)) \geq \lambda\|\mathbf{v} - A(S)\|^2 \tag{13.6}$$

We also have:

$$\begin{aligned} f_S(\mathbf{v}) - f_S(\mathbf{u}) =& L_S(\mathbf{v}) + \lambda\|\mathbf{v}\|^2 - (L_S(\mathbf{u}) + \lambda\|\mathbf{u}\|^2) \\ =& L_{S^{(i)}}(\mathbf{v}) + \lambda\|\mathbf{v}\|^2 - (L_{S^{(i)}}(\mathbf{u}) + \lambda\|\mathbf{u}\|^2) \\ &+ \frac{l(\mathbf{v}, z_i) - l(\mathbf{u}, z_i)}{m} + \frac{l(\mathbf{u}, z') - l(\mathbf{v}, z')}{m} \end{aligned} \tag{13.7}$$

which means:

$$f_S(A(S^{(i)}))-f_S(A(S)) \le \frac{l(A(S^{(i)}),z_i)-l(A(S),z_i)}{m}+\frac{l(A(S),z')-l(A(S^{(i)}),z')}{m} \tag{13.8}$$

Combining this with Equation 13.6, we obtain that:

$$\lambda\|A(S^{(i)})-A(S)\|^2 \le \frac{l(A(S^{(i)}),z_i)-l(A(S),z_i)}{m}+\frac{l(A(S),z')-l(A(S^{(i)}),z')}{m} \tag{13.9}$$

13.3.1 Lipschitz Loss

Let loss function $l(\cdot, z_i)$ be $\rho - Lipschitz$, then:

$$l(A(S^{(i)}),z_i)-l(A(S),z_i) \le \rho\|A(S^{(i)})-A(S)\|$$
$$l(A(S),z')-l(A(S^{(i)}),z') \le \rho\|A(S^{(i)})-A(S)\|$$
$$\lambda\|A(S^{(i)})-A(S)\|^2 \le \frac{2\rho\|A(S^{(i)})-A(S)\|}{m}$$
$$l(A(S^{(i)}),z_i)-l(A(S),z_i) \le \frac{2\rho^2}{\lambda m}$$

Finally, we get

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))-L_S(A(S))] \le \frac{2\rho^2}{\lambda m} \tag{13.10}$$

Theorem 13.3. *Assume that the loss function is convex and $\rho - Lipschitz$. Then, the RLM rule with the regularizer $\lambda\|\mathbf{w}\|^2$ is* on-average-replace-one-stable *with rate $\frac{2\rho^2}{\lambda m}$.*

13.3.2 Smooth and Nonnegative Loss

If the loss is $\beta-smooth$ and nonnegative then it is also self-bounded: $\|\nabla f(\mathbf{w})\|^2 \le 2\beta f(\mathbf{w})$.

$$\begin{aligned}
&l(A(S^{(i)}),z_i)-l(A(S),z_i)\\
\le&\|\nabla l(A(S),z_i)\|\|A(S^{(i)})-A(S)\|+\frac{\beta}{2}\|A(S^{(i)})-A(S)\|^2\\
\le&\sqrt{2\beta l(A(S),z_i)}\|A(S^{(i)})-A(S)\|+\frac{\beta}{2}\|A(S^{(i)})-A(S)\|^2
\end{aligned} \tag{13.11}$$

We also have:

$$l(A(S),z')-l(A(S^{(i)}),z') \le \sqrt{2\beta l(A(S^{(i)}),z')}\|A(S^{(i)})-A(S)\|+\frac{\beta}{2}\|A(S^{(i)})-A(S)\|^2 \tag{13.12}$$

Put these two equation into Equation 13.9, we can get:

$$\|A(S^{(i)})-A(S)\| \le \frac{\sqrt{2\beta}}{\lambda m-\beta}\left(\sqrt{l(A(S),z_i)}+\sqrt{l(A(S^{(i)}),z')}\right)$$

We assume $\lambda \geq 2\beta/m$, we have

$$\|A(S^{(i)}) - A(S)\| \leq \frac{\sqrt{8\beta}}{\lambda m}\left(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S^{(i)}), z')}\right)$$

Combining the preceding with Equation 13.11, we have

$$\begin{aligned}
l(A(S^{(i)}), z_i) - l(A(S), z_i) \leq& \sqrt{2\beta l(A(S), z_i)}\|A(S^{(i)}) - A(S)\| + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2 \\
\leq& \left(\frac{4\beta}{\lambda m} + \frac{4\beta^2}{(\lambda m)^2}\right)\left(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S^{(i)}), z')}\right)^2 \\
\leq& \frac{6\beta}{\lambda m}\left(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S^{(i)}), z')}\right)^2 \\
\leq& \frac{12\beta}{\lambda m}\Big(l(A(S), z_i) + l(A(S^{(i)}), z')\Big)
\end{aligned} \tag{13.13}$$

This proves the following theorem.

Theorem 13.4.

$$\mathbb{E}[l(A(S^{(i)}), z_i) - l(A(S), z_i)] \leq \frac{24\beta}{\lambda m}\mathbb{E}[L_S(A(S))] \tag{13.14}$$

If $\forall z, l(\mathbf{0}, z) \leq C$, then we have $L_S(A(S)) \leq L_S(\mathbf{0}) \leq C$, which means

$$\mathbb{E}[l(A(S^{(i)}), z_i) - l(A(S), z_i)] \leq \frac{24\beta C}{\lambda m}$$

13.4 CONTROLLING THE FITTING-STABLITY TRADE-OFF

$$\mathbb{E}_S[L_{\mathcal{D}}(A(S))] = \mathbb{E}_S[L_S(A(S))] + \mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \tag{13.15}$$

- The first term is empirical risks of $A(S)$.
- The second term is the stability of $A(S)$.
- There is trade-off between these two terms.

Then we derive bounds on the empirical risk term for the RLM rule.

$$L_S(A(S)) \leq L_S(A(S)) + \lambda\|A(S)\|^2 \leq L_S(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2$$

Taking expectation of both sides w.r.t. S, we obtain that

$$\mathbb{E}_S[L_S(A(S))] \leq L_{\mathcal{D}}(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2$$

Theorem 13.5.

$$\forall \mathbf{w}, \mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq L_{\mathcal{D}}(\mathbf{w}) + \lambda\|\mathbf{w}\|^2 + \frac{2\rho^2}{\lambda m}$$

In practice, we usually do not know the norm of $\mathbf{w}^*$, we usually tune λ on the basis of a validation set, as described in Chapter 11.

If $\forall \mathbf{w}, \|\mathbf{w}\| \leq B$, we have

$$\forall \mathbf{w}, \mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \rho B\sqrt{\frac{8}{m}} \quad \left(\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}\right)$$

Now we consider the loss function is smooth and nonnegative, then we get

$$\forall \mathbf{w}, \mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq \left(1 + \frac{24\beta}{\lambda m}\right)\mathbb{E}_S[L_S(A(S))] \leq \left(1 + \frac{24\beta}{\lambda m}\right)(L_{\mathcal{D}}(\mathbf{w}) + \lambda\|\mathbf{w}\|^2)$$

Let us play with this equation:

$$\begin{aligned}
\mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq & \left(1 + \frac{24\beta}{\lambda m}\right)\left(L_{\mathcal{D}}(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2\right) \\
=& L_{\mathcal{D}}(\mathbf{w}^*) + \frac{24\beta L_{\mathcal{D}}(\mathbf{w}^*)}{\lambda m} + \lambda\|\mathbf{w}^*\|^2 + \frac{24\beta\|\mathbf{w}^*\|^2}{m} \\
\leq& L_{\mathcal{D}}(\mathbf{w}^*) + \frac{24\beta L_{\mathcal{D}}(\mathbf{w}^*)}{\lambda m} + \lambda B^2 + \frac{24\beta B^2}{m} \\
\leq& L_{\mathcal{D}}(\mathbf{w}^*) + \frac{24\beta C}{\lambda m} + \lambda B^2 + \frac{24\beta B^2}{m} \quad (L_{\mathcal{D}}(\mathbf{w}^*) \leq L_{\mathcal{D}}(\vec{0}) = C) \\
\leq& L_{\mathcal{D}}(\mathbf{w}^*) + \frac{24\beta C B^2}{\alpha\epsilon m} + \alpha\epsilon + \frac{24\beta B^2}{m} \quad \left(\lambda = \frac{\alpha\epsilon}{B^2}, \alpha \in (0,1)\right)
\end{aligned}$$

If we want to get $\mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon$, we need

$$m \geq \frac{C + \alpha\epsilon}{(1-\alpha)\alpha\epsilon^2} \cdot 24\beta B^2 \quad \textit{or} \quad m \geq \frac{2C + \epsilon}{\epsilon^2} \cdot 48\beta B^2 \quad (\alpha = 1/2)$$

$$\left(\lambda \geq \frac{2\beta}{m}, \lambda = \frac{\alpha\epsilon}{B^2}\right)$$