13 Regularization and Stability

From chapter 12 we are familiar with convex-Lipschitz-bounded and convex-smooth-bounded learning problems. Now, we wish to show that those two families of problems are learnable. *Regularized Loss Minimization* (RLM) is a learning paradigm which minimizes the sum of the empirical risk and a regularization function. Intuitively, regularization captures the complexity of the hypothesis, as well as the stability of the procedure. Stability will be more formally defined, intuitively it reflects that the output should vary proportionally to the input.

13.1. Regularized loss minimization

RLM is a learning rule where the empirical risk and a regularization function are jointly minimized for $R : \mathbb{R}^d \to \mathbb{R}$:

$$argmin_w(L_S(w) + R(w))$$

A simple regularization function is $R(w) = \lambda||w||^2$, a combination of a scalar and norm, which yields the rule often called Tikhonov regularization

$$argmin_w(L_S(w) + \lambda||w||^2)$$

13.1.1 Ridge Regression

If we apply Tikhonov regularization to linear regression, we get

$$argmin_w(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\langle w, x_i\rangle - y_i)^2 + \lambda||w||^2)$$

Minimizing linear regression using Tikhonov regularization is called ridge regression. We minimize it with respect to w and get

$$w(2\lambda mI + \sum_{i=1}^{m} x_i x_i^T) = \sum_{i=1}^{m} y_i x_i$$

where we substitute $A = \sum_{i=1}^{m} x_i x_i^T$ and $b = \sum_{i=1}^{m} y_i x_i$. A is a positive semidefinite matrix, the expression next to w has eigenvalues bounded by $2\lambda m$, implying that the matrix is invertible.

Theorem 13.1.

Let D be a distribution over $\mathcal{X} \times [-1, 1]$, ($\mathcal{X}$ bounded by norm of 1) and $\mathcal{H} = \{w \in \mathbb{R}^d : ||w|| \leq B\}$ and $\epsilon \in (0, 1)$, $m \geq 150B^2/\epsilon^2$, then applying ridge regression with $\lambda = \epsilon/(3B^2)$, we get:

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

This tells us how many examples we need that the expected value of the risk is within the minimum true error and error rate.

13.2. Stable Rules Do Not Overfit

Now, we define stability. We consider an algorithm to overfit if the difference between the empirical and true risk is large. We will observe this difference through expectation.

Small change of input is replacing one input instance with another. After the change, we measure the performance of the algorithm, which should behave worse with the unseen instance, if the difference is very large, the algorithm might be overfitting.

Theorem 13.2. Let D be a distribution and $S = (z_1, .., z_m)$ a i.i.d. sequence of examples with z' another i.i.d. example. $U(m)$ is the uniform distribution over m, then:

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

When the right side is small, we declare an algorithm to be stable.

Definition 13.3 Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be a monotonically defined decreasing function. A learning algorithm A is on-average-replace-one-stable with rate $\epsilon(m)$ if for every distribution D:

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

However, an algorithm that does not overfit is not necessarily a good one, as one might underfit. λ parameter can control the complexity-overfitting trade-off.

13.3 Tikhonov Regularization as a Stabilizer

Applying RLM with Tikhonov leads to a stable algorithm. Now, we assume convexity of the loss function, and make the loss function either Lipschitz or smooth.

Defintion 13.4 (Strongly Convex Functions) A function f is λ-strongly convex if for all $w, u, \alpha \in (0, 1)$:

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

Every convex function is 0-strongly convex.

Now, we turn back to $\lambda||w||^2$ which is 2λ-strongly convex. Proof is here

Lemma 13.5 If f is λ-strongly convex and g is convex, then f + g is λ-strongly convex. If f is λ-strongly convex and u is a minimizer of f, then for any w:

$$f(w) - f(u) \geq \frac{\lambda}{2}||w-u||^2$$

Now, look at the stability of RLM. We know that the rule is 2λ strongly convex. From rule 3 of convexity:

$$f_S(v) - f_S(A(S)) \geq \lambda||v - A(S)||^2$$

- TODO fill up

13.3.1. Lipschitz Loss

If the loss function is ρ-Lipschitz, then:

$$l(A(S^{(i)}), z_i) - l(A(S), z_i) \leq \rho||A(S^{(i)}) - A(S)||$$

which we now combine with the stability condition (for any loss function) to get

$$\lambda||(A(S^{(i)}) - A(S)||^2 \leq \frac{2\rho||A(S^{(i)} - A(S)||}{m}$$
$$||A(S^{(i)}) - A(S)|| \leq \frac{2\rho}{\lambda m}$$

Corollary 13.6 Assume that loss function is convex and ρ-Lipschitz. Then, the RLM rule with the regularizer $\lambda||w||^2$ is on-average-replace-one-stable with rate $\frac{2\rho^2}{\lambda m}$.

13.3.2 Smooth and Nonnegative Loss

If loss is β-smooth and nonnegative it is also self-bounded.

13.4. Controlling the Fitting-Stability Trade-Off

We rewrite the risk as the sum of the empirical risk and the expectation of the difference between the true and empirical risk. We difference corresponds to the stability factor.