CSE 250: Machine Learning Theory Spring 2017

Lecture 7 - Regularization and Stability

Instructor: Raef Bassily Scribe: Andrew Leverentz

7.1 Regularized Loss Minimization (RLM)

Definition 7.1 (True risk and empirical risk)**.** Let D denote the distribution over $\mathcal{Z}$. For a given loss function $\ell : C \times \mathcal{Z} \to \mathbb{R}_+$, the true risk of $w \in C \subset \mathbb{R}^d$ is

$$L(w; D) = E_{z \sim D}[\ell(w, z)].$$

Also define the empirical risk over a dataset:

$$\hat{L}(w; S) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i),$$

where $S = (z_1, \ldots, z_n) \sim D^n$ is the training set.

In RLM, instead of minimizing $\hat{L}(w; S)$ over $w \in C$, we will minimize the sum

$$\hat{L}(w; S) + R(w).$$

Here, $R : \mathbb{R}^d \to \mathbb{R}_+$ is a "regularizer" and only depends on the hypothesis. Intuitively, the regularizer captures some notion of the "complexity" of a hypothesis. This effectively penalizes complicated hypotheses, which helps us manage the tradeoff between empirical risk and generalization error which is essentially the same as the tradeoff between bias and complexity we discussed before. In a sense, this is much like the goal behind boosting, even though the algorithmic approach is very different.

We will focus on *Tikhonov regularization*, which uses the regularizer

$$R(w) = \Lambda \|w\|^2,$$

for some $\Lambda > 0$ (the "regularization parameter").

7.2 Stability

Recall that

$$L(w; D) = \underbrace{L(w; D) - \hat{L}(w; S)}_{\text{(signed) generalization error}} + \underbrace{\hat{L}(w; S)}_{\text{empirical risk}}.$$

A hypothesis that has high generalization error despite low empirical risk is said to "overfit" the training data.

Stability, roughly, means that small changes in the input (i.e., in the training data) lead to small changes in the output (i.e., the trained hypothesis).

We will view regularization as a "stabilizer" for the learning algorithm, and we will see that stable algorithms typically do not overfit.

Definition 7.2 (Notation)**.** Given the training set $S = (z_1, \ldots, z_n)$ and an extra example z', let

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

This is a dataset in which the i^{th} data point has been replaced by z'.

Also, let A be a learning algorithm. We will denote the output hypothesis of A given S by $A(S)$.

Consider the difference

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i).$$

Intuitively, this difference is ≥ 0 for a good learner. However, if this difference is very large, this indicates that the learner is not very stable and does not generalize well to new examples.

Definition 7.3 (On-Average Replace-One (OARO) Stability)**.** Let $\tau : \mathbb{N} \to \mathbb{R}_+$ be a decreasing function. A learner A is said to be τ-OARO stable if for all distributions D, we have

$$\underset{\substack{(S,z')\sim D^{n+1} \\ i\sim\text{Unif}([n])}}{E} [\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)] \leq \tau(n).$$

Note: if we rewrite the expectation over i explicitly, this becomes

$$\underset{(S,z')\sim D^{n+1}}{E} \left[\frac{1}{n} \sum_{i=1}^{n} (\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)) \right] \leq \tau(n).$$

Theorem 7.4. Let D be a distribution, let $S \sim D^n$, and let z' be another independent example from D. For any learner A,

$$\underset{S\sim D^n}{E} [L(A(S); D) - \hat{L}(A(S); S)] = \underset{\substack{(S,z')\sim D^{n+1} \\ i\sim\text{Unif}([n])}}{E} [\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)].$$

Proof. Observe that for all $i \in [n]$,

$$\begin{aligned} E_{S\sim D^n}[L(A(S); D)] &= E_{(S,z')\sim D^{n+1}}[\ell(A(S), z')] \\ &= E_{(S,z')\sim D^{n+1}}[\ell(A(S^{(i)}), z_i)], \end{aligned}$$

because S and z' are i.i.d. On the other hand,

$$E_{S\sim D^n}[\hat{L}(A(S); S)] = \underset{\substack{S\sim D^n \\ i\sim\text{Unif}([n])}}{E} [\ell(A(S), z_i)].$$

Combining these facts yields the desired result. □

Corollary 7.5. Let A be a learner. If A is τ-OARO stable, then for all distributions D,

$$\underset{S\sim D^n}{E} [L(A(S); D) - \hat{L}(A(S); S)] \leq \tau(n).$$

7.3 Stability of Regularized Loss Minimization (RLM)

The stability of RLM relies on a property called "strong convexity."

Definition 7.6 (Strongly convex function)**.** Let $\Lambda > 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is Λ-strongly convex if for all $u, v \in \mathbb{R}^d$ and for all $\alpha \in (0, 1)$ we have

$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha)f(v) - \frac{\Lambda}{2}\alpha(1-\alpha)\left\|u - v\right\|^2.$$

When f is differentiable, an equivalent definition is the following: For all $u, v \in \mathbb{R}^d$,

$$f(u) \geq f(v) + \langle \nabla f(v), u - v\rangle + \frac{\Lambda}{2}\left\|u - v\right\|^2.$$

For non-differentiable functions, we simply use a subgradient instead of the gradient.

Lemma 7.7. Strongly convex functions have the following properties:

1. $f(w) = \Lambda\left\|w\right\|^2$ is 2Λ-strongly convex.
2. If f is Λ-strongly convex and g is convex, then $f + g$ is Λ-strongly convex.
3. If f is Λ-strongly convex and u is a (constrained) minimizer of f over a convex set $C \subseteq \mathbb{R}^d$, then for all $w \in C$,
$$f(w) - f(u) \geq \frac{\Lambda}{2}\left\|w - u\right\|^2.$$

Proof of property 3. For simplicity, let's assume f is differentiable. We have for all $w \in C$,

$$f(w) - f(u) \geq \langle \nabla f(u), w - u\rangle + \frac{\Lambda}{2}\left\|w - u\right\|^2.$$

We want to show that the inner product in the above expression is non-negative.

If the minimum is in the interior, then the gradient will be zero, and we're done.

In one dimension, if the minimizer u occurs at the boundary, then we can see that the inner product will be positive by considering the two possible cases (either u is at the left boundary or u is at the right boundary).

More generally, we can prove this in higher dimensions. Fix $w \in C$, and let $\alpha \in (0, 1)$. Define

$$\begin{aligned} w_\alpha &= u + \alpha(w - u) \\ &= \alpha w + (1-\alpha)u. \end{aligned}$$

Therefore $w_\alpha \in C$ by the convexity of C. Define $g(\alpha) = f(w_\alpha) = f(u + \alpha(w - u))$. By the Taylor expansion of g,

$$g(\alpha) = g(0) + \alpha g'(0) \pm \alpha^2 O(\left\|w - u\right\|^2),$$

and so

$$\begin{aligned} \alpha g'(0) &= \overbrace{g(\alpha) - g(0)}^{\geq 0} \pm \alpha^2 O(\left\|w - u\right\|^2) \\ &\geq \pm\alpha^2 O(\left\|w - u\right\|^2). \end{aligned}$$

The quantity $g(\alpha)-g(0)$ is non-negative because u is a minimizer of f and therefore 0 is a minimizer of g. Next, divide by α and take the limit as $\alpha \to 0$; from this, we see $g'(0) \geq 0$.

Note that

$$\begin{aligned} g'(0) &= \left.\frac{\partial g(\alpha)}{\partial \alpha}\right|_{\alpha=0} \\ &= \left.\langle \nabla f(w_\alpha), \frac{\partial w_\alpha}{\partial \alpha}\rangle\right|_{\alpha=0} \\ &= \left.\langle \nabla f(w_\alpha), w-u\rangle\right|_{\alpha=0} \end{aligned}$$

Hence,

$$\langle \nabla f(u), w-u\rangle = g'(0) \geq 0.$$

□

Theorem 7.8. Let $C \subseteq \mathbb{R}^d$ be a convex set. Let $\ell : C \times \mathcal{Z} \to \mathbb{R}_+$ be a loss function which is convex and ρ-Lipschitz (with respect to its first argument, for all possible values of its second argument). Then the RLM learner with regularizer $\Lambda \|w\|^2$ (denoted by algorithm A) is $2\rho^2/(\Lambda n)$-OARO stable.

Moreover, using a corollary from the previous lecture, we have, for all distributions D,

$$\mathop{E}_{S\sim D^n}\left[L(A(S);D) - \hat{L}(A(S);S)\right] \leq \frac{2\rho^2}{\Lambda n}.$$

Proof. Let $S = (z_1, \ldots, z_n)$, let z' be an extra example, and let

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

We know that

$$A(S) = \mathop{\mathrm{argmin}}_{w\in C}(\hat{L}(w;S) + \Lambda \|w\|^2).$$

Define $f_S(w) = \hat{L}(w;S) + \Lambda \|w\|^2$. Note that f_S is 2Λ-strongly convex. From property 3 of the previous lemma, for all $v \in C$,

$$f_S(v) - f_S(A(S)) \geq \Lambda \|v - A(S)\|^2. \tag{1}$$

On the other hand, for all $u, v \in C$, and all indices $i \in [n]$,

$$\begin{aligned} f_S(v) - f_S(u) &= \hat{L}(v;S) + \Lambda\|v\|^2 - (\hat{L}(u;S) + \Lambda\|u\|^2) \\ &= \hat{L}(v;S^{(i)}) + \Lambda\|v\|^2 - (\hat{L}(u;S^{(i)}) + \Lambda\|u\|^2) \\ &\quad + \frac{\ell(v,z_i) - \ell(v,z')}{n} \\ &\quad + \frac{\ell(u,z') - \ell(u,z_i)}{n}. \end{aligned}$$

Choose $v = A(S^{(i)})$ and $u = A(S)$. Then,

$$f_S(A(S^{(i)})) - f_S(A(S)) \leq \frac{\ell(A(S^{(i)}),z_i) - \ell(A(S),z_i)}{n} + \frac{\ell(A(S),z') - \ell(A(S^{(i)}),z')}{n}. \tag{2}$$

Now, combining (1) and (2),

$$\Lambda \left\| A(S^{(i)}) - A(S) \right\|^2 \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{n} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{n}. \tag{3}$$

Since ℓ is ρ-Lipschitz,

$$\ell(A(S^{(i)}, z_i) - \ell(A(S), z_i) \leq \rho \left\| A(S^{(i)}) - A(S) \right\|. \tag{4}$$

A similar bound holds for the second term involving z'. Hence,

$$\Lambda \left\| A(S^{(i)}) - A(S) \right\|^2 \leq \frac{2\rho \left\| A(S^{(i)}) - A(S) \right\|}{n},$$

and so

$$\left\| A(S^{(i)}) - A(S) \right\| \leq \frac{2\rho}{\Lambda n}.$$

Substituting this into (4), we have

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \frac{2\rho^2}{\Lambda n}.$$

Note that we have not made any restrictions on S, and so we can conclude this bound holds in expectation, thus completing the proof.

□

7.4 RLM analysis of excess risk for convex Lipschitz loss (continued)

We have showed that the stability rate of RLM for convex Lipschitz loss is at most $\frac{2\rho^2}{\Lambda n}$.

This implies

$$\underset{S\sim D^n}{E}\left[L(A(S); D) - \widehat{L}(A(S); S) \right] \leq \frac{2\rho^2}{\Lambda n}.$$

Furthermore,

$$\underset{S\sim D^n}{E}\left[L(A(S); D) \right] = \underset{S\sim D^n}{E}\left[\widehat{L}(A(S); S) \right] + \underbrace{\underset{S\sim D^n}{E}\left[L(A(S); D) - \widehat{L}(A(S); S) \right]}_{\leq 2\rho^2/(\Lambda n)}. \tag{*}$$

Recall also that

$$A(S) = \underset{w\in C}{\operatorname{argmin}} \left(\widehat{L}(A(S); S) + \Lambda \left\| w \right\|^2 \right).$$

As Λ increases, the expected generalization error decreases, but the empirical risk starts to increase since the minimizer of the regularized risk (the sum of the empirical risk and regularization function) starts moving away from the minimizer of the empirical risk. When $\Lambda = 0$ (i.e., no regularization), the empirical risk is minimized, but we no longer have control on the generalization error (risk of overfitting). On the other hand, in the limit as $\Lambda \to \infty$, the algorithm will no longer depend on the data since it will effectively try to minimize the $\Lambda \left\| w \right\|^2$ term only (hence, when Λ is too large, we are at risk of underfitting). We need to tune Λ to balance the tradeoff between empricial risk and generalization error (or, equivalently, the tradeoff between overfitting and underfitting, or, equivalently, bias and complexity).

Theorem 7.9. As before, suppose $C \subseteq \mathbb{R}^d$ is a closed convex set and $\ell : C \times \mathcal{Z} \to \mathbb{R}_+$ is convex ρ-Lipschitz. Let D be any distribution on $\mathcal{Z}$. Then, the RLM learner with regularizer $\Lambda \|w\|^2$ (denoted by A) satisfies the following for all $\tilde{w} \in C$:

$$\mathop{E}_{S \sim D^n} [L(A(S); D)] \leq L(\tilde{w}; D) + \Lambda \|\tilde{w}\|^2 + \frac{2\rho^2}{\Lambda n}.$$

Note that this theorem can give us a bound on the excess risk of $A(S)$. It also captures the tradeoff involved in changing Λ, since one term increases with Λ and the other decreases with Λ.

Proof. Fix any $\tilde{w} \in C$. Then

$$\begin{aligned} \widehat{L}(A(S); S) &\leq \widehat{L}(A(S); S) + \Lambda \|A(S)\|^2 \\ &\leq \widehat{L}(\tilde{w}; S) + \Lambda \|\tilde{w}\|^2 . \end{aligned}$$

Now take the expectation of both sides:

$$\begin{aligned} \mathop{E}_{S \sim D^n} \left[\widehat{L}(A(S); S)\right] &\leq \mathop{E}_{S \sim D^n} \left[\widehat{L}(\tilde{w}; S)\right] + \Lambda \|\tilde{w}\|^2 \\ &= L(\tilde{w}; D) + \Lambda \|\tilde{w}\|^2 . \end{aligned}$$

Plugging this into (*) and using a theorem from the previous lecture, we get the desired result. □

Corollary 7.10. Consider a convex-Lipschitz-bounded model where the parameter set $C \subset \mathbb{R}^d$ is M-bounded and the loss function $\ell : C \times \mathcal{Z} \to \mathbb{R}_+$ is convex and ρ-Lipschitz. Let D be any distribution. If we set $\Lambda = \frac{\rho}{M}\sqrt{\frac{2}{n}}$, then the RLM A satisfies

$$\mathop{E}_{S \sim D^n} [L(A(S); D)] \leq \min_{w \in C} L(w; D) + \rho M \sqrt{\frac{8}{n}}.$$

In particular, for any $\varepsilon > 0$, if $n \geq 8M^2\rho^2/\varepsilon^2$, then

$$\mathop{E}_{S \sim D} [L(A(S); D)] \leq \min_{w \in C} L(w; D) + \varepsilon.$$

Proof. Choose $\tilde{w}$ in Theorem 7.9 to be $w^* = \operatorname{argmin}_{w \in C} L(w; D)$. Then, pick Λ to give the best bound:

$$\Lambda = \frac{\rho}{M}\sqrt{\frac{2}{n}}.$$

□

Remark 7.11. We can get a more "agnostic-PAC-like" bound by turning the expectation bound into a probability bound (using Markov's inequality). If the loss function is bounded, one can further amplify the confidence using repeated training/validation sets (similar to an earlier homework problem).