Ch 13: Regularization and Stability
APPM 7400 Theory of Machine Learning, Spring 2020

Stephen Becker

March 16 2020

1 Intro

Intro to regularization and stability **Regularized Loss Minimization** (RLM) just means ERM plus a regularizer; the regularizer makes it stable to slight changes in input

$$\operatorname*{argmin}_{\mathbf{w}} \widehat{L}_S(\mathbf{w}) + R(\mathbf{w}) \tag{RLM}$$

Ideas behind regularization: (1) penalizes complexity (often imperfectly), (2) stabilizes problem

We'll show that if a loss function is (1) convex, (2) Lipschitz or smooth, (3) and bounded $\mathcal{H}$, then by adding a strongly convex regularizer, we can get PAC learning bounds

We focus on $\boxed{R(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2}$ (write $\|\cdot\|$ for $\|\cdot\|_2$ now)

In particular, **ridge regression** for least-squares

$$\min_{\mathbf{w}} f(\mathbf{w}) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(\langle \mathbf{x}_i, \mathbf{w}\rangle - y_i\right)^2 + \lambda\|\mathbf{w}\|^2$$

Ridge Regression **Ridge regression** objective is

$$\begin{aligned} f(\mathbf{w}) &\stackrel{\text{def}}{=} \underbrace{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(\langle \mathbf{x}_i, \mathbf{w}\rangle - y_i\right)^2}_{\widehat{L}_S(\mathbf{w})} + \underbrace{\lambda\|\mathbf{w}\|^2}_{R(\mathbf{w})} \\ &= \frac{1}{m}\frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2 \end{aligned}$$

To find solution, we can solve the normal equations (in practice, for large systems and/or ill-conditioned, there are many alternatives, such as SGD, conjugate gradient, etc.). We derive this by solving $\nabla f(\mathbf{w}) = 0$ (in this case, a necessary and sufficient condition for optimality).

$$\begin{aligned} 0 = \nabla f(\mathbf{w}) &= \frac{1}{m}\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y}) + 2\lambda\mathbf{w} \\ \implies \quad (\underbrace{\mathbf{X}^\top\mathbf{X}}_{A} + 2\lambda m I)\mathbf{w} &= \underbrace{\mathbf{X}^\top\mathbf{y}}_{b} \end{aligned} \tag{Normal Eq'n}$$

2 Analysis Setup

Analysis Framework, 1 Recall we've already talked about the traditional **bias-variance** decomposition

$$L_{\mathcal{D}}(h) = \left(\underbrace{L_{\mathcal{D}}(h) - \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')}_{\text{variance}}\right) + \underbrace{\min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')}_{\text{bias}}$$

and most of our existing analysis has been controlling the variance, e.g., via uniform convergence to get $\left|L_{\mathcal{D}}(h) - \widehat{L}_S(h)\right| < \epsilon/2$ and $\widehat{L}_S(H)$ small if $h \in \text{ERM}$.

Now, instead of uniform convergence, introduce **average** or **expected risk**

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathtt{A}(S))] \quad \text{instead of} \quad (\forall h) L_{\mathcal{D}}(h) \leq \dots$$

where we are acknowledging that the classifier h (or $\mathbf{w}$) is chosen by an algorithm $\mathtt{A}$ based on the data S.

Exercise 13.1 shows how expected risk can be used to get an agnostic PAC learning bound.

Analysis Framework, 2 Our goal is a bound like

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

and we'll get there in to parts: just like the bias-variance tradeoff, we'll do

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathtt{A}(S))] = \underbrace{\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathtt{A}(S)) - \widehat{L}_S(\mathtt{A}(S))]}_{\text{I}} + \underbrace{\mathbb{E}_{S\sim\mathcal{D}^m}[\widehat{L}_S(\mathtt{A}(S))]}_{\text{II}}$$

3 Analysis of I (stability)

Analysis Framework, 2 Our notion of stability is that if we take

$$\begin{aligned} S &= (\mathbf{z}_1, \dots, \mathbf{z}_{i-1}, \mathbf{z}_i, \mathbf{z}_{i+1}, \dots, \mathbf{z}_m) \quad \text{and replace it with} \\ S^{(i)} &= (\mathbf{z}_1, \dots, \mathbf{z}_{i-1}, \mathbf{z}', \mathbf{z}_{i+1}, \dots, \mathbf{z}_m) \end{aligned}$$

then $\mathtt{A}(S) \approx \mathtt{A}(S^{(i)})$. We'll want

$$\underbrace{0 \leq}_{\text{usually}} \ell(\mathtt{A}(S^{(i)}), \mathbf{z}_i) - \ell(\mathtt{A}(S), \mathbf{z}_i) \underbrace{\leq \epsilon}_{\text{hopefully}}$$

This relates to our error I via this theorem:

Theorem 3.1 (Thm. 13.2 in Shalev-Shwartz and Ben-David)**.** *If $S \overset{iid}{\sim} \mathcal{D}^m$, $\mathbf{z}' \sim \mathcal{D}$ (independent of S), $i \sim$ Uniform($[m]$), then $\forall$ algorithms $\mathtt{A}$*

$$\text{I} \overset{def}{=} \mathop{\mathbb{E}}_{S}\left[L_{\mathcal{D}}(\mathtt{A}(S)) - \widehat{L}_S(\mathtt{A}(S))\right] = \mathop{\mathbb{E}}_{\substack{S\\ \mathbf{z}',i}}\left[\ell(\mathtt{A}(S^{(i)}), \mathbf{z}_i) - \ell(\mathtt{A}(S), \mathbf{z}_i)\right].$$

Proof. We'll show the left terms on both sides equal, then the right terms. For the left terms,

$$\mathop{\mathbb{E}}_{\substack{S\\ \mathbf{z}',i}}\left[\ell(\mathtt{A}(S^{(i)}), \mathbf{z}_i)\right] = \mathop{\mathbb{E}}_{\substack{S\\ \mathbf{z}'}}[\ell(\mathtt{A}(S), \mathbf{z}')] = \mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(\mathtt{A}(S))]$$

and similarly for the right terms

$$\mathop{\mathbb{E}}_{\substack{S\\ i}}[\ell(\mathtt{A}(S), \mathbf{z}_i)] = \mathop{\mathbb{E}}_{S}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\mathtt{A}(S), \mathbf{z}_i)\right] = \mathbb{E}_S\left[\widehat{L}_S(\mathtt{A}(S))\right]$$

□

Analysis Framework, 4 Informally, if Ⓘ is small, the algorithm $\mathtt{A}$ is **stable**

Formally, say $\mathtt{A}$ is **(on-average-replacement) stable** with rate $\epsilon(m)$ (non-increasing in m) if $\forall\mathcal{D}$, Ⓘ $\le \epsilon(m)$.

We'll investigate how to prove an algorithm is stable, using our theorem to characterize Ⓘ. We'll assume regularizer is 2λ-strongly convex, e.g., $R(\mathbf{w}) = \lambda\|\mathbf{w}\|^2$.

Recall if f is μ-strongly convex and non-negative, then if $\mathbf{u} \in \operatorname{argmin} f$,

$$(\forall \mathbf{v} \in \mathbb{R}^d)\ \frac{\mu}{2}\|\mathbf{v}-\mathbf{u}\|^2 \le f(\mathbf{v}) - f(\mathbf{u}) \le f(\mathbf{v}) \qquad \text{(self-boundedness)}$$

Showing stability Write $f_S(\mathbf{w}) = \widehat{L}_S(\mathbf{w}) + \lambda\|\mathbf{w}\|^2$, or just $f(\mathbf{w})$ when S is clear from context. This is 2λ strongly convex. Our algorithm $\mathtt{A}$ is RLM, so $\boxed{\mathbf{u} = \mathtt{A}(S)} \stackrel{\text{def}}{=} \operatorname{argmin}_{\mathbf{w}} f(\mathbf{w})$

Similarly, define $\boxed{\mathbf{v} = \mathtt{A}(S^{(i)})}$.

$$\begin{aligned} f(\mathbf{v}) - f(\mathbf{u}) &\stackrel{\text{def}}{=} \widehat{L}_S(\mathbf{v}) + \lambda\|\mathbf{v}\|^2 - \left(\widehat{L}_S(\mathbf{u}) + \lambda\|\mathbf{u}\|^2\right) \\ &= \underbrace{\widehat{L}_{S^{(i)}}(\mathbf{v}) + \lambda\|\mathbf{v}\|^2}_{ⓐ} - \underbrace{\widehat{L}_{S^{(i)}}(\mathbf{u}) + \lambda\|\mathbf{u}\|^2}_{ⓑ} + \frac{1}{m}\left(\ell(\mathbf{v},\mathbf{z}_i) - \ell(\mathbf{u},\mathbf{z}_i)\right) + \frac{1}{m}\left(\ell(\mathbf{u},\mathbf{z}') - \ell(\mathbf{v},\mathbf{z}')\right) \end{aligned}$$

Because we chose $\mathbf{v}$ as above, it minimizes $\widehat{L}_{S^{(i)}}(\mathbf{v}) + \lambda\|\mathbf{v}\|^2$, so ⓐ $\le$ ⓑ, hence

$$f(\mathbf{v}) - f(\mathbf{u}) \le \frac{1}{m}\left(\ell(\mathbf{v},\mathbf{z}_i) - \ell(\mathbf{u},\mathbf{z}_i)\right) + \frac{1}{m}\left(\ell(\mathbf{u},\mathbf{z}') - \ell(\mathbf{v},\mathbf{z}')\right)$$

By self-boundedness, $f(\mathbf{v}) - f(\mathbf{u}) \ge \lambda\|\mathbf{v}-\mathbf{u}\|^2$, so combining this with above,

$$\lambda\|\mathbf{v}-\mathbf{u}\|^2 \le \frac{1}{m}\Big(\ell(\mathbf{v},\mathbf{z}_i) - \ell(\mathbf{u},\mathbf{z}_i)\Big) + \frac{1}{m}\Big(\ell(\mathbf{u},\mathbf{z}') - \ell(\mathbf{v},\mathbf{z}')\Big) \tag{1}$$

From here, there are two ways to proceed:

Case 1: assuming $(\forall \mathbf{z})$, $\mathbf{w} \mapsto \ell(\mathbf{w},\mathbf{z})$ is ρ-Lipschitz

So, directly from Lipschitz property,

$$\begin{aligned} \ell(\mathbf{v},\mathbf{z}_i) - \ell(\mathbf{u},\mathbf{z}_i) &\le \rho\|\mathbf{v}-\mathbf{u}\| \\ \ell(\mathbf{u},\mathbf{z}') - \ell(\mathbf{v},\mathbf{z}') &\le \rho\|\mathbf{v}-\mathbf{u}\| \end{aligned} \tag{2}$$

so substitute this into Eq. (1) gives

$$\lambda\|\mathbf{v}-\mathbf{u}\|^2 \le \frac{1}{m}\rho\|\mathbf{v}-\mathbf{u}\| + \frac{1}{m}\rho\|\mathbf{v}-\mathbf{u}\|$$

and either $\mathbf{v} = \mathbf{u}$ or $\|\mathbf{v}-\mathbf{u}\| > 0$ and then we can divide by it; either way,

$$\|\mathbf{v}-\mathbf{u}\| \le \frac{2\rho}{\lambda m}$$

and put this back into Eq. (2) to get

$$\ell(\underbrace{\mathbf{v}}_{\mathtt{A}(S^{(i)})},\mathbf{z}_i) - \ell(\underbrace{\mathbf{u}}_{\mathtt{A}(S)},\mathbf{z}_i) \le \rho\frac{2\rho}{\lambda m}$$

and thus by Thm. 3.1

$$Ⓘ = \mathop{\mathbb{E}}_{\substack{S \\ \mathbf{z}',i}}\left[\ell(\mathtt{A}(S^{(i)}),\mathbf{z}_i) - \ell(\mathtt{A}(S),\mathbf{z}_i)\right] \le \frac{2\rho^2}{\lambda m} \stackrel{\text{def}}{=} \epsilon(m)$$

leading to **Corollary 13.6** which states that if ℓ is uniformly ρ-Lipschitz and strongly convex with parameter $\mu > 0$ then it is (on-average-replace-one)**stable** with rate $\epsilon(m) = \frac{4\rho^2}{\mu m}$.

Note that we did not need to assume $\mathbf{x}$ or $\mathbf{w}$ was bounded.

Case 2: assuming $(\forall \mathbf{z})$, $\mathbf{w} \mapsto \ell(\mathbf{w}, \mathbf{z})$ is β-smooth

(And assume ℓ is non-negative, but we already made that assumption; of course, all we really need is that it is bounded below, with a known bound, since there is nothing special about 0).

When $\nabla \ell$ is β-Lipschitz, we have

$$(\forall \mathbf{w})(\forall \mathbf{z})\ \|\nabla \ell(\mathbf{w}, \mathbf{z})\|^2 \leq 2\beta\left(\ell(\mathbf{w}, \mathbf{z}) - \ell(\mathbf{w}^\star, \mathbf{z})\right) \leq 2\beta \ell(\mathbf{w}, \mathbf{z}) \tag{3}$$

where $\mathbf{w}^\star \in \operatorname{argmin}_{\mathbf{w}}\ \ell(\mathbf{w}, \mathbf{z})$ and the 2nd inequality follows by assuming non-negativity. Also, using the quadratic upper bound property of strongly smooth functions,

$$\begin{aligned}
\ell(\underbrace{\mathbf{v}}_{\mathtt{A}(S^{(i)})}, \mathbf{z}_i) - \ell(\underbrace{\mathbf{u}}_{\mathtt{A}(S)}, \mathbf{z}_i) &\leq \langle \nabla \ell(\mathbf{u}, \mathbf{z}_i), \mathbf{v} - \mathbf{u}\rangle + \frac{\beta}{2}\|\mathbf{v} - \mathbf{u}\|^2 \\
&\leq \|\nabla \ell(\mathbf{u}, \mathbf{z}_i)\| \cdot \|\mathbf{v} - \mathbf{u}\| + \frac{\beta}{2}\|\mathbf{v} - \mathbf{u}\|^2 \quad \text{(Cauchy-Schwarz)} \\
&\leq \sqrt{2\beta \ell(\mathbf{u}, \mathbf{z})} \cdot \|\mathbf{v} - \mathbf{u}\| + \frac{\beta}{2}\|\mathbf{v} - \mathbf{u}\|^2 \quad \text{via Eq. (3)} \qquad (4)
\end{aligned}$$

and an analogous result holds for $\ell(\mathbf{v}, \mathbf{z}') - \ell(\mathbf{u}, \mathbf{z}')$. Plug these results into Eq. (1) and divide by $\lambda\|\mathbf{v} - \mathbf{u}\|$ and re-arrange to get

$$\|\mathbf{v} - \mathbf{u}\| \leq \frac{\sqrt{2\beta}}{\lambda m - \beta}\left(\sqrt{\ell(\mathbf{u}, \mathbf{z}_i)} + \sqrt{\ell(\mathbf{v}, \mathbf{z}')}\right)$$

We can choose the value of λ, so pick it such that $\beta \leq \frac{\lambda m}{2}$ so then

$$\|\mathbf{v} - \mathbf{u}\| \leq \frac{\sqrt{8\beta}}{\lambda m}\left(\sqrt{\ell(\mathbf{u}, \mathbf{z}_i)} + \sqrt{\ell(\mathbf{v}, \mathbf{z}')}\right)$$

Now, as before, we go back to an earlier bound: plug the above eq into Eq. (4) to get (skipping a few steps)

$$\begin{aligned}
\ell(\underbrace{\mathbf{v}}_{\mathtt{A}(S^{(i)})}, \mathbf{z}_i) - \ell(\underbrace{\mathbf{u}}_{\mathtt{A}(S)}, \mathbf{z}_i) &\leq \left(\frac{4\beta}{\lambda m} + \frac{8\beta^2}{(\lambda m)^2}\right)\left(\sqrt{\ell(\mathbf{u}, \mathbf{z}_i)} + \sqrt{\ell(\mathbf{v}, \mathbf{z}')}\right)^2 \quad \text{and bound } \frac{8\beta^2}{(\lambda m)^2} \geq 0 \\
&\leq \frac{24\beta}{\lambda m}\left(\ell(\mathbf{u}, \mathbf{z}_i) + \ell(\mathbf{v}, \mathbf{z}')\right)^2 \quad \text{since } (a+b)^2 \leq 3(a^2 + b^2)
\end{aligned}$$

thus via Thm. 3.1

$$\begin{aligned}
\text{I} = \mathop{\mathbb{E}}_{\substack{S \\ \mathbf{z}', i}}\left[\ell(\mathtt{A}(S^{(i)}), \mathbf{z}_i) - \ell(\mathtt{A}(S), \mathbf{z}_i)\right] &\leq \frac{24\beta}{\lambda m} \mathop{\mathbb{E}}_{\substack{S \\ \mathbf{z}', i}}\left[\ell(\mathbf{u}, \mathbf{z}_i) + \ell(\mathbf{v}, \mathbf{z}')\right] \\
&= \frac{48\beta}{\lambda m} \mathop{\mathbb{E}}_{S} \widehat{L}_S(\mathtt{A}(S))
\end{aligned}$$

and typically the loss function is bounded for all $\mathbf{z}$, e.g., $\ell(0, \mathbf{z}) \leq c$, hence $\mathbb{E}_S\, \widehat{L}_S(\mathtt{A}(S))$, so in this case, we have **Corollary 13.7** which is that if ℓ is uniformly β-smooth and strongly convex with parameter $\mu > 0$ then it is (on-average-replace-one)**stable** with rate $\epsilon(m) = \frac{96\beta c}{\mu m}$.

4 Analysis of II (bias/underfitting)