Regularization + Stability

CPSC 532S: Modern Statistical Learning Theory

28 February 2022

cs.ubc.ca/~dsuth/532S/22/

Admin

- Hope your break was good!

- Office hour times have moved:
  - **Monday 12-1pm** (and still Thursdays 4-5pm)
  - Both are available both on Zoom or in person (ICICS X563)

- A1 grades are up on Gradescope
- A2 solutions are posted
- A3 will be posted tonight or tomorrow, due in ~2 weeks
  - If you don’t have a group and want one, post on Piazza (asap)

Last time: convex optimization

steps to reach ε (expected) error:

	ρ-Lipschitz	ρ-Lipschitz + λ strongly convex	β-smooth	β-smooth + λ strongly convex
ERM w/ gradient descent	$\frac{B^2\rho^2}{\varepsilon^2}$		$\frac{B\beta}{2\varepsilon}$	
(One-pass) SGD	$\frac{B^2\rho^2}{\varepsilon^2}$ (avg iterate)		$12\frac{B^2\beta}{2\varepsilon^2}$ (avg)	

Last time: convex optimization

steps to reach ε (expected) error:

	ρ-Lipschitz	ρ-Lipschitz + λ strongly convex	β-smooth	β-smooth + λ strongly convex
ERM w/ gradient descent	$\frac{B^2\rho^2}{\varepsilon^2}$	$\mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\frac{B\beta}{2\varepsilon}$	
(One-pass) SGD	$\frac{B^2\rho^2}{\varepsilon^2}$ (avg iterate)		$12\frac{B^2\beta}{2\varepsilon^2}$ (avg)	

Last time: convex optimization

steps to reach ε (expected) error:

	ρ-Lipschitz	ρ-Lipschitz + λ strongly convex	β-smooth	β-smooth + λ strongly convex
ERM w/ gradient descent	$\frac{B^2\rho^2}{\varepsilon^2}$	$\mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\frac{B\beta}{2\varepsilon}$	
(One-pass) SGD	$\frac{B^2\rho^2}{\varepsilon^2}$ (avg iterate)	$15\frac{\rho^2}{\lambda\varepsilon}$ (tail avg)	$12\frac{B^2\beta}{2\varepsilon^2}$ (avg)	

Last time: convex optimization

steps to reach ε (expected) error:

	ρ-Lipschitz	ρ-Lipschitz + λ strongly convex	β-smooth	β-smooth + λ strongly convex
ERM w/ gradient descent	$\frac{B^2\rho^2}{\varepsilon^2}$	$\mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\frac{B\beta}{2\varepsilon}$	$\frac{\beta}{\lambda}\log\frac{\beta B}{2\varepsilon}$
(One-pass) SGD	$\frac{B^2\rho^2}{\varepsilon^2}$ (avg iterate)	$15\frac{\rho^2}{\lambda\varepsilon}$ (tail avg)	$12\frac{B^2\beta}{2\varepsilon^2}$ (avg)	

Last time: convex optimization

steps to reach ε (expected) error:

	ρ-Lipschitz	ρ-Lipschitz + λ strongly convex	β-smooth	β-smooth + λ strongly convex
ERM w/ gradient descent	$\frac{B^2\rho^2}{\varepsilon^2}$	$\mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\frac{B\beta}{2\varepsilon}$	$\frac{\beta}{\lambda}\log\frac{\beta B}{2\varepsilon}$
(One-pass) SGD	$\frac{B^2\rho^2}{\varepsilon^2}$ (avg iterate)	$15\frac{\rho^2}{\lambda\varepsilon}$ (tail avg)	$12\frac{B^2\beta}{2\varepsilon^2}$ (avg)	$2\frac{\beta\rho^2}{\lambda^2\varepsilon}$ (last)

(also ρ-Lipschitz)

Last time: convex optimization

steps to reach ε (expected) error:

	ρ-Lipschitz	ρ-Lipschitz + λ strongly convex	β-smooth	β-smooth + λ strongly convex
ERM w/ gradient descent	$\frac{B^2\rho^2}{\varepsilon^2}$	$\mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\frac{B\beta}{2\varepsilon}$	$\frac{\beta}{\lambda}\log\frac{\beta B}{2\varepsilon}$
(One-pass) SGD	$\frac{B^2\rho^2}{\varepsilon^2}$ (avg iterate)	$15\frac{\rho^2}{\lambda\varepsilon}$ (tail avg)	$12\frac{B^2\beta}{2\varepsilon^2}$ (avg)	$2\frac{\beta\rho^2}{\lambda^2\varepsilon}$ (last)

(also ρ-Lipschitz)

Better rates with strong convexity!

L2 / Tikhonov Regularization

- Some problems just aren't strongly convex (or λ is very small)

L2 / Tikhonov Regularization

- Some problems just aren't strongly convex (or λ is very small)
- If f is convex and g is λ-strongly convex, then $f + g$ is also λ-strongly convex

L2 / Tikhonov Regularization

- Some problems just aren't strongly convex (or λ is very small)
- If f is convex and g is λ-strongly convex, then $f + g$ is also λ-strongly convex
- $g(w) = \frac{\lambda}{2}||w||^2$ is λ-strongly convex

L2 / Tikhonov Regularization

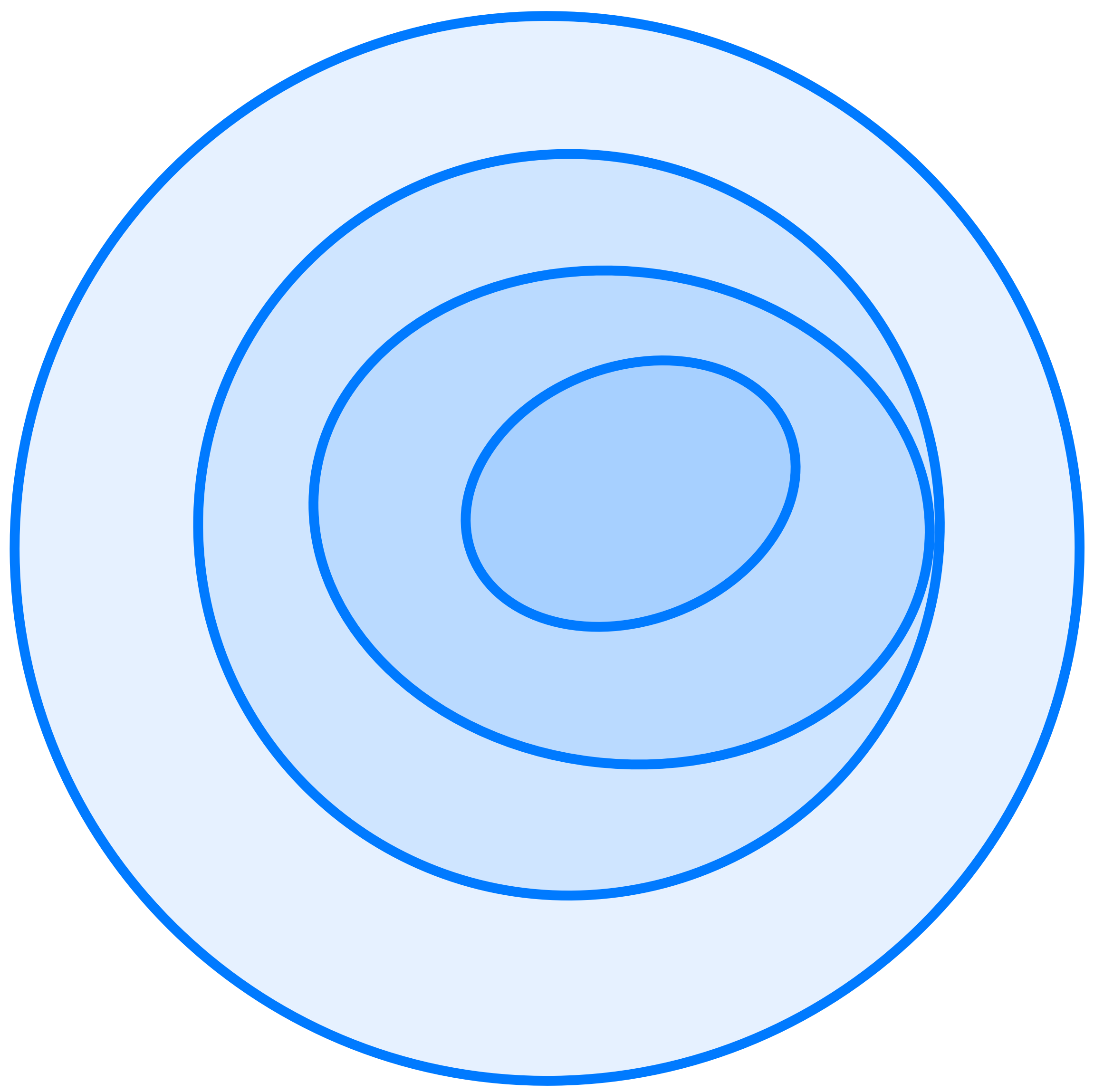
- Some problems just aren't strongly convex (or λ is very small)
- If f is convex and g is λ-strongly convex, then $f + g$ is also λ-strongly convex
- $g(w) = \frac{\lambda}{2}||w||^2$ is λ-strongly convex

- **Regularized loss minimization (RLM)**: $\text{argmin}_w \, L_S(w) + R(w)$

L2 / Tikhonov Regularization

- Some problems just aren't strongly convex (or λ is very small)
- If f is convex and g is λ-strongly convex, then $f + g$ is also λ-strongly convex
- $g(w) = \frac{\lambda}{2}||w||^2$ is λ-strongly convex

- **Regularized loss minimization (RLM)**: $\text{argmin}_w L_S(w) + R(w)$
  - Recall SRM: $\text{argmin}_h L_S(h) + \varepsilon_{k_h}(n, \delta w_{k_h})$



Ridge regression

Ridge regression: $\mathscr{H} = \{x \mapsto w^\top x\}, \ell(h, (x, y)) = \frac{1}{2}(h(x) - y)^2, R(w) = \frac{\lambda}{2}\|w\|^2$

$$L_S(w) = \sum_i \frac{1}{2}(w^\top x_i - y_i)^2 = \frac{1}{2}\|Xw - y\|^2$$

$$= \frac{1}{2} w^\top X^\top X w - y^\top X w + \underbrace{\frac{1}{2}\|y\|^2}$$

$$R(w) = \frac{\lambda}{2}\|w\|^2 = w^\top\left(\frac{\lambda}{2} I\right) w$$

$$\arg\min L_S(w) + R(w) = \arg\min_w \frac{1}{2} w^\top (X^\top X + \lambda I) w - y^\top X w$$

$$\nabla = (\underbrace{X^\top X} + \underbrace{\lambda I})\hat{w} - X^\top y = 0$$

$$\hat{w} = (X^\top X + \lambda I)^{-1} X^\top y$$

Regularized loss minimization with SGD

- L2 regularizer is also called **weight decay**:

Regularized loss minimization with SGD

- L2 regularizer is also called **weight decay**:
  - $\nabla \frac{\lambda}{2}\|w\|^2 = \lambda w$, so use $w_{t+1} = w_t - \eta(\lambda w_t + \hat{g}_t)$

Regularized loss minimization with SGD

- L2 regularizer is also called **weight decay**:
  - $\nabla \frac{\lambda}{2}\|w\|^2 = \lambda w$, so use $w_{t+1} = w_t - \eta(\lambda w_t + \hat{g}_t)$
- With ρ-Lipschitz loss, SGD from 0 with $\eta_t = 1/(\lambda t)$ has step norm at most 2ρ

$f(w)$ $\qquad f_\lambda(w) = f(w) + \frac{1}{2}\lambda\|w\|^2$

$$
\begin{aligned}
w_{t+1} &= w_t - \eta_t(\lambda w_t + \hat{g}_t) \\
&= w_t - \frac{1}{t} w_t - \frac{1}{\lambda t}\hat{g}_t \\
&= \frac{t-1}{t} w_t - \frac{1}{\lambda t}\hat{g}_t \\
&= \frac{t-1}{t}\left(\frac{t-2}{t-1} w_{t-1} - \frac{1}{\lambda(t-1)}\hat{g}_{t-1}\right) - \frac{1}{\lambda t}\hat{g}_t \\
&= -\frac{1}{\lambda} \cdot \frac{1}{t}\sum_{k=1}^{t} \hat{g}_k
\end{aligned}
$$

$\|w_{t+1}\| \le \frac{1}{\lambda}\rho$

$\|\lambda w_t + \hat{g}_t\| \le \lambda \cdot \frac{1}{\lambda} \cdot \rho + \underbrace{\|\hat{g}_t\|}_{\le \rho}$

$\le 2\rho$

Regularized loss minimization with SGD

- For ρ-Lipschitz, β-smooth, convex f, $f_\lambda(w) = f(w) + \frac{1}{2}\lambda\|w\|^2$, $\eta_t = 1/(\lambda t)$:

$$\mathbb{E}[f_\lambda(w_T)] - f_\lambda(w_\lambda^*) \leq \frac{2\beta\rho^2}{\lambda^2 T} \quad \text{b/c} \quad \mathbb{E}\left[\|w_T - w_\lambda^*\|^2\right] \leq \frac{4\rho^2}{\lambda^2 T}$$

$$\mathbb{E}f(w_T) - f(w^*) = \mathbb{E}f(w_T) - f(w_\lambda^*) + f(w_\lambda^*) - f(w^*)$$

Regularized loss minimization with SGD

- For ρ-Lipschitz, β-smooth, convex f, $f_\lambda(w) = f(w) + \frac{1}{2}\lambda\|w\|^2$, $\eta_t = 1/(\lambda t)$:

$$\mathbb{E}[f_\lambda(w_T)] - f_\lambda(w_\lambda^*) \leq \frac{2\beta\rho^2}{\lambda^2 T} \quad \text{b/c} \quad \mathbb{E}\left[\|w_T - w_\lambda^*\|^2\right] \leq \frac{4\rho^2}{\lambda^2 T}$$

$$\mathbb{E}f(w_T) - f(w^*) = \underbrace{\mathbb{E}f(w_T) - f(w_\lambda^*)} + \underbrace{f(w_\lambda^*) - f(w^*)}$$

$$\mathbb{E}f(w_T) - f(w_\lambda^*) \leq \frac{\lambda}{2}\mathbb{E}\left[\|w_\lambda^*\|^2 - \|w_T\|^2\right] + \frac{2\beta\rho^2}{\lambda^2 T}$$

Regularized loss minimization with SGD

- For ρ-Lipschitz, β-smooth, convex f, $f_\lambda(w) = f(w) + \frac{1}{2}\lambda\|w\|^2$, $\eta_t = 1/(\lambda t)$:

$$\mathbb{E}[f_\lambda(w_T)] - f_\lambda(w_\lambda^*) \leq \frac{2\beta\rho^2}{\lambda^2 T} \quad \text{b/c} \quad \mathbb{E}\left[\|w_T - w_\lambda^*\|^2\right] \leq \frac{4\rho^2}{\lambda^2 T}$$

$$\mathbb{E}f(w_T) - f(w^*) = \mathbb{E}f(w_T) - f(w_\lambda^*) + f(w_\lambda^*) - f(w^*)$$

$$\begin{aligned}\mathbb{E}f(w_T) - f(w_\lambda^*) &\leq \frac{\lambda}{2}\mathbb{E}\left[\|w_\lambda^*\|^2 - \|w_T\|^2\right] + \frac{2\beta\rho^2}{\lambda^2 T} \\ &= \frac{\lambda}{2}\mathbb{E}\left[\left(\|w_\lambda^*\| + \|w_T\|\right)\left(\|w_\lambda^*\| - \|w_T\|\right)\right] + \frac{2\beta\rho^2}{\lambda^2 T}\end{aligned}$$

Regularized loss minimization with SGD

- For ρ-Lipschitz, β-smooth, convex f, $f_\lambda(w) = f(w) + \frac{1}{2}\lambda\|w\|^2$, $\eta_t = 1/(\lambda t)$:

$$\mathbb{E}[f_\lambda(w_T)] - f_\lambda(w_\lambda^*) \leq \frac{2\beta\rho^2}{\lambda^2 T} \quad \text{b/c} \quad \mathbb{E}\left[\|w_T - w_\lambda^*\|^2\right] \leq \frac{4\rho^2}{\lambda^2 T}$$

$$\mathbb{E}f(w_T) - f(w^*) = \mathbb{E}f(w_T) - f(w_\lambda^*) + f(w_\lambda^*) - f(w^*)$$

$$\begin{aligned}
\mathbb{E}f(w_T) - f(w_\lambda^*) &\leq \frac{\lambda}{2}\mathbb{E}\left[\|w_\lambda^*\|^2 - \|w_T\|^2\right] + \frac{2\beta\rho^2}{\lambda^2 T} \\
&= \frac{\lambda}{2}\mathbb{E}\left[\left(\|w_\lambda^*\| + \|w_T\|\right)\left(\|w_\lambda^*\| - \|w_T\|\right)\right] + \frac{2\beta\rho^2}{\lambda^2 T} \\
&\leq \lambda B \mathbb{E}\left[\|w_\lambda^* - w_T\|\right] + \frac{2\beta\rho^2}{\lambda^2 T}
\end{aligned}$$

Regularized loss minimization with SGD

- For ρ-Lipschitz, β-smooth, convex f, $f_\lambda(w) = f(w) + \frac{1}{2}\lambda\|w\|^2$, $\eta_t = 1/(\lambda t)$:

$$\mathbb{E}[f_\lambda(w_T)] - f_\lambda(w^*_\lambda) \le \frac{2\beta\rho^2}{\lambda^2 T} \quad \text{b/c} \quad \mathbb{E}\left[\|w_T - w^*_\lambda\|^2\right] \le \frac{4\rho^2}{\lambda^2 T}$$

$$\mathbb{E}f(w_T) - f(w^*) = \mathbb{E}f(w_T) - f(w^*_\lambda) + f(w^*_\lambda) - f(w^*)$$

$$\begin{aligned}
\mathbb{E}f(w_T) - f(w^*_\lambda) &\le \frac{\lambda}{2}\mathbb{E}\left[\|w^*_\lambda\|^2 - \|w_T\|^2\right] + \frac{2\beta\rho^2}{\lambda^2 T} \\
&= \frac{\lambda}{2}\mathbb{E}\left[\left(\|w^*_\lambda\| + \|w_T\|\right)\left(\|w^*_\lambda\| - \|w_T\|\right)\right] + \frac{2\beta\rho^2}{\lambda^2 T} \\
&\le \lambda B \mathbb{E}\left[\|w^*_\lambda - w_T\|\right] + \frac{2\beta\rho^2}{\lambda^2 T} \le \frac{2B\rho}{\sqrt{T}} + \frac{2\beta\rho^2}{\lambda^2 T}
\end{aligned}$$

Stability

- So, that attempt didn't work here (though it can in some specific settings)

Stability

- So, that attempt didn't work here (though it can in some specific settings)
- Another approach which will work generically: **stability**

Stability

- So, that attempt didn't work here (though it can in some specific settings)
- Another approach which will work generically: **stability**
  - "If we don't change S much, the predictor $A(S)$ doesn't change much"

Stability

- So, that attempt didn't work here (though it can in some specific settings)
- Another approach which will work generically: **stability**
    - "If we don't change S much, the predictor $A(S)$ doesn't change much"
    - Regularizer $R(w) = \frac{1}{2}||w||^2$ **stabilizes** the algorithm

Stability

- So, that attempt didn't work here (though it can in some specific settings)
- Another approach which will work generically: **stability**
  - "If we don't change S much, the predictor $A(S)$ doesn't change much"
  - Regularizer $R(w) = \frac{1}{2}||w||^2$ **stabilizes** the algorithm

- One variant is **on-average replace-one stability**:
  Let $S = (z_1, \ldots, z_n) \sim \mathscr{D}^n, z' \sim \mathscr{D},\ i \sim \mathrm{Uniform}([n])$ be independent.
  Let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_n)$.

Stability

- So, that attempt didn't work here (though it can in some specific settings)
- Another approach which will work generically: **stability**
  - "If we don't change S much, the predictor $A(S)$ doesn't change much"
  - Regularizer $R(w) = \frac{1}{2}||w||^2$ **stabilizes** the algorithm

- One variant is **on-average replace-one stability**:
  Let $S = (z_1, \ldots, z_n) \sim \mathcal{D}^n, z' \sim \mathcal{D},\ i \sim \text{Uniform}([n])$ be independent.
  Let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_n)$.
  - **Theorem**: $\mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))] = \mathbb{E}_{S,z',i}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)]$.

Stability

- So, that attempt didn't work here (though it can in some specific settings)
- Another approach which will work generically: **stability**
  - "If we don't change S much, the predictor $A(S)$ doesn't change much"
  - Regularizer $R(w) = \frac{1}{2}\|w\|^2$ **stabilizes** the algorithm

- One variant is **on-average replace-one stability**:
  Let $S = (z_1, \ldots, z_n) \sim \mathscr{D}^n$, $z' \sim \mathscr{D}$, $i \sim \mathrm{Uniform}([n])$ be independent.
  Let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_n)$.
  - **Theorem**: $\mathbb{E}_S[L_{\mathscr{D}}(A(S)) - L_S(A(S))] = \mathbb{E}_{S,z',i}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)]$.
  - A is on-average-replace-one-stable with rate $\varepsilon(n)$ if for all $\mathscr{D}$,
    $\mathbb{E}_{S,z',i}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)] \leq \varepsilon(n)$.

On-average replace-one stability of RLM

Let $f_S(w) = L_S(w) + \frac{1}{2}\lambda \|w\|^2$

On-average replace-one stability of RLM

Let $f_S(w) = L_S(w)+\frac{1}{2}\lambda\|w\|^2$

$$f_S(v) - f_S(u) = L_S(v)+\frac{\lambda}{2}\|v\|^2 - L_S(u)-\frac{\lambda}{2}\|u\|^2$$

On-average replace-one stability of RLM

Let $f_S(w) = L_S(w) + \frac{1}{2}\lambda\|w\|^2$

$$f_S(v) - f_S(u) = L_S(v) + \frac{\lambda}{2}\|v\|^2 - L_S(u) - \frac{\lambda}{2}\|u\|^2$$
$$= L_{S^{(i)}}(v) + \frac{\lambda}{2}\|v\|^2 - L_{S^{(i)}}(u) - \frac{\lambda}{2}\|u\|^2 + \frac{\ell(v, z_i) - \ell(u, z_i)}{n} + \frac{\ell(u, z') - \ell(v, z')}{n}$$

On-average replace-one stability of RLM

Let $f_S(w) = L_S(w) + \frac{1}{2}\lambda\|w\|^2$

$$f_S(v) - f_S(u) = L_S(v) + \frac{\lambda}{2}\|v\|^2 - L_S(u) - \frac{\lambda}{2}\|u\|^2$$

$$= \underbrace{L_{S^{(i)}}(v) + \frac{\lambda}{2}\|v\|^2}_{f_{S^{(i)}}(v)} - \underbrace{L_{S^{(i)}}(u) - \frac{\lambda}{2}\|u\|^2}_{f_{S^{(i)}}(u)} + \frac{\ell(v, z_i) - \ell(u, z_i)}{n} + \frac{\ell(u, z') - \ell(v, z')}{n}$$

Plug in $v = A(S^{(i)})$, $u = A(S)$; note that v minimizes $f_{S^{(i)}}$:

On-average replace-one stability of RLM

Let $f_S(w) = L_S(w) + \frac{1}{2}\lambda\|w\|^2$

$$f_S(v) - f_S(u) = L_S(v) + \frac{\lambda}{2}\|v\|^2 - L_S(u) - \frac{\lambda}{2}\|u\|^2$$
$$= L_{S^{(i)}}(v) + \frac{\lambda}{2}\|v\|^2 - L_{S^{(i)}}(u) - \frac{\lambda}{2}\|u\|^2 + \frac{\ell(v, z_i) - \ell(u, z_i)}{n} + \frac{\ell(u, z') - \ell(v, z')}{n}$$

Plug in $v = A(S^{(i)})$, $u = A(S)$; note that v minimizes $f_{S^{(i)}}$:

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

On-average replace-one stability of RLM

Let $f_S(w) = L_S(w) + \frac{1}{2}\lambda\|w\|^2$

$$f_S(v) - f_S(u) = L_S(v) + \frac{\lambda}{2}\|v\|^2 - L_S(u) - \frac{\lambda}{2}\|u\|^2$$
$$= L_{S^{(i)}}(v) + \frac{\lambda}{2}\|v\|^2 - L_{S^{(i)}}(u) - \frac{\lambda}{2}\|u\|^2 + \frac{\ell(v, z_i) - \ell(u, z_i)}{n} + \frac{\ell(u, z') - \ell(v, z')}{n}$$

Plug in $v = A(S^{(i)})$, $u = A(S)$; note that v minimizes $f_{S^{(i)}}$:

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

f_S is λ-strongly convex, so $f_S(v) - f_S(A(S)) \geq \frac{\lambda}{2}\|v - A(S)\|^2$

On-average replace-one stability of RLM

Let $f_S(w) = L_S(w) + \frac{1}{2}\lambda\|w\|^2$

$$f_S(v) - f_S(u) = L_S(v) + \frac{\lambda}{2}\|v\|^2 - L_S(u) - \frac{\lambda}{2}\|u\|^2$$
$$= L_{S^{(i)}}(v) + \frac{\lambda}{2}\|v\|^2 - L_{S^{(i)}}(u) - \frac{\lambda}{2}\|u\|^2 + \frac{\ell(v, z_i) - \ell(u, z_i)}{n} + \frac{\ell(u, z') - \ell(v, z')}{n}$$

Plug in $v = A(S^{(i)})$, $u = A(S)$; note that v minimizes $f_{S^{(i)}}$:

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

f_S is λ-strongly convex, so $f_S(v) - f_S(A(S)) \ge \frac{\lambda}{2}\|v - A(S)\|^2$

$$\frac{\lambda}{2}\|A(S^{(i)}) - A(S)\|^2 \le \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{n} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{n}$$

On-average replace-one stability for Lipschitz RLM

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

On-average replace-one stability for Lipschitz RLM

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

If $\ell(\,\cdot\,, z_i)$ is ρ-Lipschitz:

On-average replace-one stability for Lipschitz RLM

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

If $\ell(\cdot\,, z_i)$ is ρ-Lipschitz:

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

On-average replace-one stability for Lipschitz RLM

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

If $\ell(\cdot, z_i)$ is ρ-Lipschitz:

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

On-average replace-one stability for Lipschitz RLM

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

If $\ell(\cdot, z_i)$ is ρ-Lipschitz:

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

$$\|A(S^{(i)}) - A(S)\| \leq \frac{4\rho}{\lambda n}$$

On-average replace-one stability for Lipschitz RLM

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

If $\ell(\cdot, z_i)$ is ρ-Lipschitz:

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

$$\|A(S^{(i)}) - A(S)\| \leq \frac{4\rho}{\lambda n}$$

So, because $\ell(\cdot, z_i)$ is ρ-Lipschitz:

On-average replace-one stability for Lipschitz RLM

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

If $\ell(\cdot, z_i)$ is ρ-Lipschitz:

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

$$\|A(S^{(i)}) - A(S)\| \leq \frac{4\rho}{\lambda n}$$

So, because $\ell(\cdot, z_i)$ is ρ-Lipschitz:

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \frac{4\rho^2}{\lambda n}$$

On-average replace-one stability for Lipschitz RLM

$$\frac{\lambda}{2}\|A(S^{(i)}) - A(S)\|^2 \le \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{n} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{n}$$

If $\ell(\cdot, z_i)$ is ρ-Lipschitz:

$$\frac{\lambda}{2}\|A(S^{(i)}) - A(S)\|^2 \le \frac{\rho\|A(S^{(i)}) - A(S)\|}{n} + \frac{\rho\|A(S) - A(S^{(i)})\|}{n} = \frac{2\rho\|A(S^{(i)}) - A(S)\|}{n}$$

$$\|A(S^{(i)}) - A(S)\| \le \frac{4\rho}{\lambda n}$$

So, because $\ell(\cdot, z_i)$ is ρ-Lipschitz:

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \le \frac{4\rho^2}{\lambda n}$$

and $\mathbb{E}_S[L_{\mathscr{D}}(A(S)) - L_S(A(S))] \le \dfrac{4\rho^2}{\lambda n}$

On-average replace-one stability for Smooth RLM

If $\ell(\cdot, z_i)$ is β-smooth and nonnegative, can show (SSBD section 13.3.2):

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \frac{48\beta}{\lambda n}\left(\ell(A(S), z_i) + \ell(A(S^{(i)}), z')\right)$$

if $\lambda \geq \beta/n$

which implies $\mathbb{E}_S[L_{\mathscr{D}}(A(S)) - L_S(A(S))] \leq \dfrac{96\beta}{\lambda n}\mathbb{E}_S[L_S(A(S))]$

On-average replace-one stability for Smooth RLM

If $\ell(\,\cdot\,, z_i)$ is β-smooth and nonnegative, can show (SSBD section 13.3.2):

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \frac{48\beta}{\lambda n}\left(\ell(A(S), z_i) + \ell(A(S^{(i)}), z')\right)$$

if $\lambda \geq \beta/n$

which implies $\mathbb{E}_S[L_{\mathscr{D}}(A(S)) - L_S(A(S))] \leq \dfrac{96\beta}{\lambda n}\mathbb{E}_S[L_S(A(S))]$

(and note, e.g., $\mathbb{E}_S[L_S(\cancel{A(S))] \leq \mathbb{E}_S[L_S(0)] = L_{\mathscr{D}}(0)}$, often bounded by a constant)

oops

Fitting-Stability Tradeoff

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

Fitting-Stability Tradeoff

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

- Second term is the on-average replace-one stability

Fitting-Stability Tradeoff

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

- Second term is the on-average replace-one stability
    - Bigger λ means more stable

Fitting-Stability Tradeoff

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

- Second term is the on-average replace-one stability
  - Bigger λ means more stable
- First term is how well the algorithm fits the training data

Fitting-Stability Tradeoff

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

- Second term is the on-average replace-one stability
    - Bigger λ means more stable
- First term is how well the algorithm fits the training data
    - Bigger λ means worse fit

Training error of RLM

$$L_S(A(S)) \leq L_S(A(S)) + \frac{\lambda}{2}\|A(S)\|^2$$

Training error of RLM

For any fixed vector w^*

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

Training error of RLM

For any fixed vector $w*$

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

$$\mathbb{E}_S L_S(A(S)) \leq \mathbb{E}_S L_S(w*) + \frac{\lambda}{2}\|w*\|^2$$

Training error of RLM

For any fixed vector $w*$

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

$$\mathbb{E}_S L_S(A(S)) \le \mathbb{E}_S L_S(w*) + \frac{\lambda}{2}\|w*\|^2 = L_{\mathscr{D}}(w*) + \frac{\lambda}{2}\|w*\|^2$$

Training error of RLM

For any fixed vector $w*$

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

$$\mathbb{E}_S L_S(A(S)) \leq \mathbb{E}_S L_S(w*) + \frac{\lambda}{2}\|w*\|^2 = L_{\mathscr{D}}(w*) + \frac{\lambda}{2}\|w*\|^2$$

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

Training error of RLM

For any fixed vector w^*

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

$$\mathbb{E}_S L_S(A(S)) \le \mathbb{E}_S L_S(w^*) + \frac{\lambda}{2}\|w^*\|^2 = L_{\mathscr{D}}(w^*) + \frac{\lambda}{2}\|w^*\|^2$$

$$\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \le L_{\mathscr{D}}(w^*) + \frac{\lambda}{2}\|w^*\|^2 + \mathbb{E}_S[L_{\mathscr{D}}(A(S)) - L_S(A(S))]$$

So, for convex ρ-Lipschitz loss, RLM with regularizer $\frac{\lambda}{2}\|w\|^2$ has

$$\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \le L_{\mathscr{D}}(w^*) + \frac{\lambda}{2}\|w^*\|^2 + \frac{4\rho^2}{\lambda n}$$

RLM learns any convex-Lipschitz-bounded problem

- RLM with regularizer $\frac{\lambda}{2}\|w\|^2$ has $\mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq L_{\mathcal{D}}(w^*) + \frac{\lambda}{2}\|w^*\|^2 + \frac{4\rho^2}{\lambda n}$

RLM learns any convex-Lipschitz-bounded problem

- RLM with regularizer $\frac{\lambda}{2}\|w\|^2$ has $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq L_{\mathscr{D}}(w^*) + \frac{\lambda}{2}\|w^*\|^2 + \frac{4\rho^2}{\lambda n}$

- Taking $\lambda = \frac{\rho}{B}\sqrt{\frac{8}{n}}$ gives $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq L_{\mathscr{D}}(w^*) + \rho B\sqrt{\frac{8}{n}}$

RLM learns any convex-Lipschitz-bounded problem

- RLM with regularizer $\frac{\lambda}{2}\|w\|^2$ has $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq L_{\mathscr{D}}(w^*) + \frac{\lambda}{2}\|w^*\|^2 + \dfrac{4\rho^2}{\lambda n}$
- Taking $\lambda = \dfrac{\rho}{B}\sqrt{\dfrac{8}{n}}$ gives $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq L_{\mathscr{D}}(w^*) + \rho B\sqrt{\dfrac{8}{n}}$
- So, for $n \geq 8\rho^2 B^2/\varepsilon^2$, $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq \inf_{w \in \mathscr{H}} L_{\mathscr{D}}(w) + \varepsilon$

RLM learns any convex-Lipschitz-bounded problem

- RLM with regularizer $\frac{\lambda}{2}\|w\|^2$ has $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq L_{\mathscr{D}}(w^*) + \frac{\lambda}{2}\|w^*\|^2 + \frac{4\rho^2}{\lambda n}$
- Taking $\lambda = \frac{\rho}{B}\sqrt{\frac{8}{n}}$ gives $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq L_{\mathscr{D}}(w^*) + \rho B\sqrt{\frac{8}{n}}$
- So, for $n \geq 8\rho^2 B^2/\varepsilon^2$, $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq \inf_{w \in \mathscr{H}} L_{\mathscr{D}}(w) + \varepsilon$
- Similar result for convex-smooth-bounded (SSBD Corollaries 13.10, 13.11)

RLM learns any convex-Lipschitz-bounded problem

- RLM with regularizer $\frac{\lambda}{2}\|w\|^2$ has $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq L_{\mathscr{D}}(w^*) + \frac{\lambda}{2}\|w^*\|^2 + \dfrac{4\rho^2}{\lambda n}$
- Taking $\lambda = \dfrac{\rho}{B}\sqrt{\dfrac{8}{n}}$ gives $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq L_{\mathscr{D}}(w^*) + \rho B\sqrt{\dfrac{8}{n}}$
- So, for $n \geq 8\rho^2 B^2/\varepsilon^2$, $\mathbb{E}_S[L_{\mathscr{D}}(A(S))] \leq \inf_{w \in \mathscr{H}} L_{\mathscr{D}}(w) + \varepsilon$
- Similar result for convex-smooth-bounded (SSBD Corollaries 13.10, 13.11)
- Can convert these expectation bounds into high-probability: SSBD exercise 13.1

Uniform stability

- **Uniform stability** instead says $|\ell(A(S), z) - \ell(A(S^{(i)}, z)| \leq \gamma$ for **all** possible training sets S, slightly different sets $S^{(i)}$, test points z

Uniform stability

- **Uniform stability** instead says $|\ell(A(S), z) - \ell(A(S^{(i)}, z)| \leq \gamma$ for **all** possible training sets S, slightly different sets $S^{(i)}$, test points z
- Stronger condition: implies on-average replace-one stability

Uniform stability

- **Uniform stability** instead says $|\ell(A(S), z) - \ell(A(S^{(i)}, z)| \leq \gamma$ for **all** possible training sets S, slightly different sets $S^{(i)}$, test points z
- Stronger condition: implies on-average replace-one stability
  - We (*almost*) showed $\gamma = \dfrac{4\rho^2}{\lambda n}$ in our proof for ρ-Lipschitz f with $\frac{\lambda}{2}||w||^2$

Uniform stability

- **Uniform stability** instead says $|\ell(A(S), z) - \ell(A(S^{(i)}, z)| \leq \gamma$ for **all** possible training sets S, slightly different sets $S^{(i)}$, test points z
- Stronger condition: implies on-average replace-one stability
  - We (*almost*) showed $\gamma = \dfrac{4\rho^2}{\lambda n}$ in our proof for ρ-Lipschitz f with $\frac{\lambda}{2}\|w\|^2$
- Bousquet and Elisseef (2002) (also in MRT chap 14): if $\ell \in [0,C]$,

$$L_{\mathscr{D}}(A(S)) - L_S(A(S)) \leq \text{const}\left(\sqrt{n}\gamma + C/\sqrt{n}\right)\sqrt{\log \tfrac{1}{\delta}}$$

Uniform stability

- **Uniform stability** instead says $|\ell(A(S), z) - \ell(A(S^{(i)}, z)| \leq \gamma$ for **all** possible training sets S, slightly different sets $S^{(i)}$, test points z
- Stronger condition: implies on-average replace-one stability
  - We (*almost*) showed $\gamma = \dfrac{4\rho^2}{\lambda n}$ in our proof for ρ-Lipschitz f with $\frac{\lambda}{2}\|w\|^2$
- Bousquet and Elisseef (2002) (also in MRT chap 14): if $\ell \in [0,C]$,
$$L_{\mathscr{D}}(A(S)) - L_S(A(S)) \leq \text{const}\left(\sqrt{n}\gamma + C/\sqrt{n}\right)\sqrt{\log\frac{1}{\delta}}$$
- Bousquet et al. (2019), building off Feldman and Vondrák (2018, 19), show
$$L_{\mathscr{D}}(A(S)) - L_S(A(S)) \leq \text{const}\left(\gamma\log(n)\,\log\left(\frac{1}{\delta}\right) + \frac{C}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}\right)$$

SGD is uniformly stable

- Hardt, Recht, and Singer (ICML-16):
  - For convex, β-smooth, ρ-Lipschitz losses, T-step **multi-pass** SGD with $\eta_t \leq \frac{2}{\beta}$ gives uniform stability of $\frac{2\rho^2}{n} \sum_{t=1}^{T} \eta_t$

SGD is uniformly stable

- Hardt, Recht, and Singer (ICML-16):
  - For convex, β-smooth, ρ-Lipschitz losses, T-step **multi-pass** SGD with $\eta_t \leq \frac{2}{\beta}$ gives uniform stability of $\frac{2\rho^2}{n} \sum_{t=1}^{T} \eta_t$
  - For λ-strongly convex, β-smooth losses, T-step **multi-pass** projected SGD with constant $\eta \leq 1/\beta$ has uniform stability $\frac{2L^2}{\lambda n}$, **independent of** T

SGD is uniformly stable

- <u>Hardt, Recht, and Singer (ICML-16)</u>:
  - For convex, β-smooth, ρ-Lipschitz losses, T-step **multi-pass** SGD with $\eta_t \leq \frac{2}{\beta}$ gives uniform stability of $\dfrac{2\rho^2}{n}\displaystyle\sum_{t=1}^{T}\eta_t$
  - For λ-strongly convex, β-smooth losses, T-step **multi-pass** projected SGD with constant $\eta \leq 1/\beta$ has uniform stability $\dfrac{2L^2}{\lambda n}$, **independent of** T
  - Some results even for nonconvex case

SGD is uniformly stable

- Hardt, Recht, and Singer (ICML-16):
  - For convex, β-smooth, ρ-Lipschitz losses, T-step **multi-pass** SGD with $\eta_t \leq \frac{2}{\beta}$ gives uniform stability of $\frac{2\rho^2}{n}\sum_{t=1}^{T}\eta_t$
  - For λ-strongly convex, β-smooth losses, T-step **multi-pass** projected SGD with constant $\eta \leq 1/\beta$ has uniform stability $\frac{2L^2}{\lambda n}$, **independent of** T
  - Some results even for nonconvex case
- Chen, Jin, Yu (2018) and Feldman and Vondrák (2019) extend to full-batch gradient descent, other variants

Summary

- Adding an L2 regularizer makes things strongly convex, which is nicer
  - Ridge regression / weight decay
  - Regularized loss minimization (RLM), the regularized version of ERM
    - Analogous to SRM

Summary

- Adding an L2 regularizer makes things strongly convex, which is nicer
  - Ridge regression / weight decay
  - Regularized loss minimization (RLM), the regularized version of ERM
    - Analogous to SRM

- Can analyze via stability
  - On-average replace-one stability characterizes learnability
    - Holds for convex-Lipschitz-bounded / convex-smooth-bounded
  - Amount of regularization trades off between fitting and stability

Summary

- Adding an L2 regularizer makes things strongly convex, which is nicer
  - Ridge regression / weight decay
  - Regularized loss minimization (RLM), the regularized version of ERM
    - Analogous to SRM

- Can analyze via stability
  - On-average replace-one stability characterizes learnability
    - Holds for convex-Lipschitz-bounded / convex-smooth-bounded
  - Amount of regularization trades off between fitting and stability

- Uniform stability: stronger notion that can give better bounds
  - Multi-pass SGD / GD / … are uniformly stable