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 $\varepsilon$ (expected) error:

<table>
<thead>
<tr>
<th>Method</th>
<th>$\rho$-Lipschitz + $\lambda$ strongly convex</th>
<th>$\beta$-smooth + $\lambda$ strongly convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM w/ gradient descent</td>
<td>$\frac{B^2 \rho^2}{\varepsilon^2}$</td>
<td>$\frac{B\beta}{2\varepsilon}$</td>
</tr>
<tr>
<td>(One-pass) SGD</td>
<td>$\frac{B^2 \rho^2}{\varepsilon^2}$ (avg iterate)</td>
<td>$12\frac{B^2 \beta}{2\varepsilon^2}$ (avg)</td>
</tr>
</tbody>
</table>

$B$, $\rho$, $\beta$, $\varepsilon$, and $\lambda$ are parameters related to the optimization problem.
Last time: convex optimization

steps to reach $\varepsilon$ (expected) error:

<table>
<thead>
<tr>
<th>$\rho$-Lipschitz</th>
<th>$\rho$-Lipschitz + $\lambda$ strongly convex</th>
<th>$\beta$-smooth + $\lambda$ strongly convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM w/ gradient descent</td>
<td>$\frac{B^2 \rho^2}{\varepsilon^2}$</td>
<td>$\Theta \left( \frac{1}{\varepsilon} \right)$</td>
</tr>
<tr>
<td>(One-pass) SGD</td>
<td>$\frac{B^2 \rho^2}{\varepsilon^2}$</td>
<td>$\Theta \left( \frac{1}{\varepsilon} \right)$</td>
</tr>
</tbody>
</table>

(avg iterate) (avg)
Last time: convex optimization

steps to reach \( \varepsilon \) (expected) error:

<table>
<thead>
<tr>
<th></th>
<th>( \rho )-Lipschitz</th>
<th>( \rho )-Lipschitz + ( \lambda ) strongly convex</th>
<th>( \beta )-smooth</th>
<th>( \beta )-smooth + ( \lambda ) strongly convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM w/ gradient descent</td>
<td>( \frac{B^2 \rho^2}{\varepsilon^2} )</td>
<td>( \Theta \left( \frac{1}{\varepsilon} \right) )</td>
<td>( \frac{B\beta}{2\varepsilon} )</td>
<td></td>
</tr>
<tr>
<td>(One-pass) SGD</td>
<td>( \frac{B^2 \rho^2}{\varepsilon^2} )</td>
<td>( \frac{15 \rho^2}{\lambda \varepsilon} )</td>
<td>( \frac{12 B^2 \beta}{2 \varepsilon^2} )</td>
<td></td>
</tr>
</tbody>
</table>

(avg iterate) (tail avg) (avg)
Last time: convex optimization

steps to reach $\varepsilon$ (expected) error:

<table>
<thead>
<tr>
<th>$\rho$-Lipschitz</th>
<th>$\rho$-Lipschitz + $\lambda$ strongly convex</th>
<th>$\beta$-smooth + $\lambda$ strongly convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{B^2 \rho^2}{\varepsilon^2}$</td>
<td>$\beta \left( \frac{1}{\varepsilon} \right)$</td>
<td>$\frac{\beta}{\lambda \log \frac{\beta B}{2\varepsilon}}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ERM w/ gradient descent</th>
<th>(avg iterate)</th>
<th>(tail avg)</th>
<th>(avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{B^2 \rho^2}{\varepsilon^2}$</td>
<td>$15\frac{\rho^2}{\lambda \varepsilon}$</td>
<td>$12\frac{B^2 \beta}{2\varepsilon^2}$</td>
<td>$12\frac{B^2 \beta}{2\varepsilon^2}$</td>
</tr>
</tbody>
</table>
Last time: convex optimization

steps to reach $\varepsilon$ (expected) error:

<table>
<thead>
<tr>
<th>$\rho$-Lipschitz</th>
<th>$\rho$-Lipschitz + $\lambda$ strongly convex</th>
<th>$\beta$-smooth</th>
<th>$\beta$-smooth + $\lambda$ strongly convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{B^2 \rho^2}{\varepsilon^2}$</td>
<td>$\Theta \left( \frac{1}{\varepsilon} \right)$</td>
<td>$\frac{B\beta}{2\varepsilon}$</td>
<td>$\frac{\beta}{\lambda} \log \frac{\beta B}{2\varepsilon}$</td>
</tr>
</tbody>
</table>

ERM w/ gradient descent

(One-pass) SGD

<table>
<thead>
<tr>
<th>$\frac{B^2 \rho^2}{\varepsilon^2}$</th>
<th>$15 \frac{\rho^2}{\lambda \varepsilon}$</th>
<th>$12 \frac{B^2 \beta}{2\varepsilon^2}$</th>
<th>$2 \frac{\beta \rho^2}{\lambda^2 \varepsilon}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(avg iterate)</td>
<td>(tail avg)</td>
<td>(avg)</td>
<td>(last)</td>
</tr>
</tbody>
</table>

(Also $\rho$-Lipschitz)
Last time: convex optimization

steps to reach $\varepsilon$ (expected) error:

<table>
<thead>
<tr>
<th></th>
<th>$\rho$-Lipschitz</th>
<th>$\rho$-Lipschitz + $\lambda$ strongly convex</th>
<th>$\beta$-smooth</th>
<th>$\beta$-smooth + $\lambda$ strongly convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM w/ gradient descent</td>
<td>( \frac{B^2 \rho^2}{\varepsilon^2} )</td>
<td>( \Theta \left( \frac{1}{\varepsilon} \right) )</td>
<td>( \frac{B\beta}{2\varepsilon} )</td>
<td>( \frac{\beta}{\lambda} \log \frac{\beta B}{2\varepsilon} )</td>
</tr>
<tr>
<td>(One-pass) SGD</td>
<td>( \frac{B^2 \rho^2}{\varepsilon^2} )</td>
<td>( 15\frac{\rho^2}{\lambda \varepsilon} )</td>
<td>( 12\frac{B^2 \beta}{2\varepsilon^2} )</td>
<td>( 2\frac{\beta \rho^2}{\lambda^2 \varepsilon} )</td>
</tr>
<tr>
<td></td>
<td>(avg iterate)</td>
<td>(tail avg)</td>
<td>(avg)</td>
<td>(last)</td>
</tr>
</tbody>
</table>

Better rates with strong convexity!
L2 / Tikhonov Regularization

• Some problems just aren’t strongly convex (or \( \lambda \) is very small)
L2 / Tikhonov Regularization

- Some problems just aren’t strongly convex (or $\lambda$ is very small)
- If $f$ is convex and $g$ is $\lambda$-strongly convex, then $f + g$ is also $\lambda$-strongly convex
L2 / Tikhonov Regularization

- Some problems just aren’t strongly convex (or \( \lambda \) is very small)
- If \( f \) is convex and \( g \) is \( \lambda \)-strongly convex, then \( f + g \) is also \( \lambda \)-strongly convex
- \( g(w) = \frac{\lambda}{2}\|w\|^2 \) is \( \lambda \)-strongly convex
L2 / Tikhonov Regularization

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

- **Regularized loss minimization (RLM):** $\arg\min_w L_S(w) + R(w)$
L2 / Tikhonov Regularization

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

**Regularized loss minimization (RLM):** $\arg\min_w L_S(w) + R(w)$
- Recall SRM: $\arg\min_h L_S(h) + \epsilon_{k_h}(n, \delta w_{k_h})$
Ridge regression: $H = \{x \mapsto w^T x\}$, $\ell(h, (x, y)) = \frac{1}{2}(h(x) - y)^2$, $R(w) = \frac{\lambda}{2} \|w\|^2$

\[
\begin{align*}
L_S(w) &= \sum \frac{1}{2} (w^T x_i - y_i)^2 = \frac{1}{2} \|Xw - y\|^2 \\
R(w) &= \frac{\lambda}{2} \|w\|^2 = w^T \left( \frac{\lambda}{2} I \right) w = \frac{1}{2} w^T X^T X w - y^T X w + \frac{1}{2} \|y\|^2 \\
\text{argmin } L_S(w) + R(w) &= \text{argmin } \frac{1}{2} w^T (X^T X + \frac{\lambda}{2} I) w - y^T X w \\
\n\n\end{align*}
\]

\[
\begin{align*}
\nabla = (X^T X + \frac{\lambda}{2} I)^T \hat{w} - X^T y &= 0 \\
\hat{w} &= (X^T X + \frac{\lambda}{2} I)^{-1} X^T y \\
\end{align*}
\]

5
Regularized loss minimization with SGD

• L2 regularizer is also called \textit{weight decay}:
Regularized loss minimization with SGD

• L2 regularizer is also called **weight decay**:
  
  \[ \nabla \frac{\lambda}{2} \| w \|^2 = \lambda w, \text{ so use } w_{t+1} = w_t - \eta (\lambda w_t + \hat{g}_t) \]
Regularized loss minimization with SGD

- L2 regularizer is also called **weight decay**: 
  \[ \nabla \frac{\lambda}{2} \| w \|^2 = \lambda w, \text{ so use } w_{t+1} = w_t - \eta (\lambda w_t + \hat{g}_t) \]
- With \( \rho \)-Lipschitz loss, SGD from 0 with \( \eta_t = 1/(\lambda t) \) has step norm at most \( 2\rho \)
  \[
  f(w) \quad f_\lambda(w) = f(w) + \frac{1}{2} \lambda \| w \|^2
  \]
  \[
  w_{t+1} = w_t - \eta_t (\lambda w_t + \hat{g}_t)
  \]
  \[
  = w_t - \frac{1}{\lambda} \cdot \frac{1}{t} \cdot w_t - \frac{1}{\lambda t} \hat{g}_t
  \]
  \[
  = \frac{t-1}{t} \cdot w_t - \frac{1}{\lambda t} \hat{g}_t
  \]
  \[
  = \frac{t-1}{t} \left( \frac{t-2}{t-1} \cdot w_{t-1} - \frac{1}{\lambda (t-1)} \hat{g}_{t-1} \right) - \frac{1}{\lambda t} \hat{g}_t
  \]
  \[
  = \frac{1}{\lambda} \cdot \frac{1}{t} \sum_{k=1}^{t} \frac{1}{\lambda k} \hat{g}_k
  \]
Regularized loss minimization with SGD

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

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

$$\mathbb{E}f(w_T) - f(w^*) = \mathbb{E}f(w_T) - f(w^*_\lambda) + f(w^*_\lambda) - f(w^*)$$
Regularized loss minimization with SGD

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

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

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

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

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

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

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

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

  \[
  = \frac{\lambda}{2}\mathbb{E}\left[\left(\|w_\lambda^*\| + \|w_T\|\right)\left(\|w_\lambda^*\| - \|w_T\|\right)\right] + \frac{2\beta\rho^2}{\lambda^2 T}
  \]
Regularized loss minimization with SGD

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

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

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

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

$$= \frac{\lambda}{2} \mathbb{E}\left[\left(\|w^*_\lambda\| + \|w^*_T\|\right) \left(\|w^*_\lambda\| - \|w^*_T\|\right)\right] + \frac{2\beta \rho^2}{\lambda^2 T}$$

$$\leq \lambda B \mathbb{E}\left[\|w^*_\lambda - w^*_T\|\right] + \frac{2\beta \rho^2}{\lambda^2 T}$$
Regularized loss minimization with SGD

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

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

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

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

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

\[
\leq \lambda B\mathbb{E}\left[\|w_*^* - w_T\|\right] + \frac{2\beta\rho^2}{\lambda^2 T} \leq \frac{2B\rho}{\sqrt{T}} + \frac{2\beta\rho^2}{\lambda^2 T}
\]
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 $S = (z_1, \ldots, z_n) \sim \mathcal{D}^n$, $z' \sim \mathcal{D}$, $i \sim \text{Uniform}([n])$ be independent.
  Let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_n)$. 
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 $S = (z_1, \ldots, z_n) \sim \mathcal{D}^n$, $z' \sim \mathcal{D}$, $i \sim \text{Uniform}([n])$ be independent.
  Let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_n)$.
  • **Theorem:** $\mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))] = \mathbb{E}_{S,z',i}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)]$.
  • A is on-average-replace-one-stable with rate $\varepsilon(n)$ if for all $\mathcal{D}$,
    $\mathbb{E}_{S,z',i}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)] \leq \varepsilon(n)$. 


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

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

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

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

$$f_S(A(S^{(i)}) - f_S(A(S)) \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{n} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{n}$$
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 $\lambda$-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 \( f_S(w) = L_S(w) + \frac{1}{2} \lambda \|w\|^2 \)

\[
f_S(v) - f_S(u) = L_S(v) + \frac{1}{2} \lambda \|v\|^2 - L_S(u) - \frac{1}{2} \lambda \|u\|^2 = L_{S(i)}(v) + \frac{\lambda}{2} \|v\|^2 - L_{S(i)}(u) - \frac{\lambda}{2} \|u\|^2 + \frac{\ell(v, z_i) - \ell(u, z_i)}{n} + \frac{\ell(u, z') - \ell(v, z')}{n}
\]

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

\[
\frac{\lambda}{2} \|A(S^{(i)}) - A(S)\|^2 \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{n} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{n}
\]
On-average replace-one stability for Lipschitz RLM

\[ \frac{\lambda}{2} \|A(S^{(i)}) - A(S)\|^2 \leq \frac{\ell(A(S^{(i)}, z_i) - \ell(A(S), z_i)}{n} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{n} \]
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 \( \rho \)-Lipschitz:
On-average replace-one stability for Lipschitz RLM

$$\begin{align*}
\frac{\lambda}{2} \|A(S^{(i)}) - A(S)\|^2 &\leq \frac{\ell(A(S^{(i)}, z_i) - \ell(A(S), z_i)}{n} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}, z')}{n} \\
\text{If } \ell(\cdot, z_i) \text{ is } \rho\text{-Lipschitz:} \\
\frac{\lambda}{2} \|A(S^{(i)}) - A(S)\|^2 &\leq \frac{\rho \|A(S^{(i)}) - A(S)\|}{n} + \frac{\rho \|A(S) - A(S^{(i)})\|}{n}
\end{align*}$$
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 \( \rho \)-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 $\rho$-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 \( \rho \)-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 \( \rho \)-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 \( \rho \)-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 \( \rho \)-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

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

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

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

and \(\mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \leq \frac{4\rho^2}{\lambda n}\)
On-average replace-one stability for Smooth RLM

If $\ell(\cdot, z_i)$ is $\beta$-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 $\mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \leq \frac{96\beta}{\lambda n} \mathbb{E}_S[L_S(A(S))]$
On-average replace-one stability for Smooth RLM

If $\ell(\cdot, z_i)$ is $\beta$-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 $\mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \leq \frac{96\beta}{\lambda n} \mathbb{E}_S[L_S(A(S))]$

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

\[ \mathbb{E}_S[L_{\mathcal{D}}(A(S))] = \mathbb{E}_S[L_S(A(S))] + \mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \]
Fitting-Stability Tradeoff

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

- Second term is the on-average replace-one stability
Fitting-Stability Tradeoff

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

- Second term is the on-average replace-one stability
  - Bigger \( \lambda \) means more stable
Fitting-Stability Tradeoff

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

- Second term is the on-average replace-one stability
  - Bigger \( \lambda \) means more stable
- First term is how well the algorithm fits the training data
Fitting-Stability Tradeoff

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

- Second term is the on-average replace-one stability
  - Bigger \( \lambda \) means more stable
- First term is how well the algorithm fits the training data
  - Bigger \( \lambda \) 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 $w^*$

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

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

For any fixed vector $w^*$

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

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

\[
\mathbb{E}_s [L_D(A(S))] \leq L_D(w^*) + \frac{\lambda}{2} \|w^*\|^2 + \mathbb{E}_S[L_D(A(S)) - L_S(A(S))]
\]
Training error of RLM

For any fixed vector $w^*$

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

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

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

So, for convex $\rho$-Lipschitz loss, 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 $\mathbb{E}_S[L_\mathcal{D}(A(S)) \leq L_\mathcal{D}(w^*) + \frac{\lambda}{2} \|w^*\|^2 + \frac{4\rho^2}{\lambda n}$
RLM learns any convex-Lipschitz-bounded problem

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

- Taking $\lambda = \frac{\rho}{B} \sqrt{\frac{8}{n}}$ gives $\mathbb{E}_S[L_\mathcal{D}(A(S))] \leq L_\mathcal{D}(w^*) + \rho B \sqrt{\frac{8}{n}}$
RLM learns any convex-Lipschitz-bounded problem

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

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

- So, for $n \geq 8\rho^2 B^2 / \varepsilon^2$, $\mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq \inf_{w \in \mathcal{H}} L_{\mathcal{D}}(w) + \varepsilon$
RLM learns any convex-Lipschitz-bounded problem

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

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

- So, for $n \geq \frac{8\rho^2 B^2}{\epsilon^2}$, $\mathbb{E}_S[L_D(A(S))] \leq \inf_{w \in \mathcal{H}} L_D(w) + \epsilon$

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

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

- So, for $n \geq 8\rho^2 B^2/\varepsilon^2$, $\mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq \inf_{w \in \mathcal{H}} L_{\mathcal{D}}(w) + \varepsilon$

- 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 $\gamma = \frac{4\rho^2}{\lambda n}$ in our proof for $\rho$-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 $\gamma = \frac{4\rho^2}{\lambda n}$ in our proof for $\rho$-Lipschitz $f$ with $\frac{\lambda}{2} \|w\|^2$
- Bousquet and Elisseef (2002) (also in MRT chap 14): if $\ell \in [0, C]$
  \[ L_D(A(S)) - L_S(A(S)) \leq \text{const} \left( \sqrt{n\gamma} + C/\sqrt{n} \right) \sqrt{\log \frac{1}{\delta}} \]
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 $\gamma = \frac{4\rho^2}{\lambda n}$ in our proof for $\rho$-Lipschitz $f$ with $\frac{\lambda}{2} ||w||^2$

- Bousquet and Elisseef (2002) (also in MRT chap 14): if $\ell \in [0,C]$
  \[ L_\mathcal{D}(A(S)) - L_S(A(S)) \leq \text{const} \left( \sqrt{n\gamma} + \frac{C}{\sqrt{n}} \right) \sqrt{\log \frac{1}{\delta}} \]

- Bousquet et al. (2019), building off Feldman and Vondrák (2018, 19), show
  \[ L_\mathcal{D}(A(S)) - L_S(A(S)) \leq \text{const} \left( \gamma \log(n) \log \left( \frac{1}{\delta} \right) + \frac{C}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}} \right) \]
SGD is uniformly stable

• Hardt, Recht, and Singer (ICML-16):
  • For convex, $\beta$-smooth, $\rho$-Lipschitz losses, $T$-step multi-pass SGD with

$$\eta_t \leq \frac{2}{\beta} \text{ gives uniform stability of } \frac{2\rho^2}{n} \sum_{t=1}^{T} \eta_t$$
SGD is uniformly stable

- Hardt, Recht, and Singer (ICML-16):
  - For convex, $\beta$-smooth, $\rho$-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 $\lambda$-strongly convex, $\beta$-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

- Hardt, Recht, and Singer (ICML-16):
  - For convex, $\beta$-smooth, $\rho$-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 $\lambda$-strongly convex, $\beta$-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
SGD is uniformly stable

• Hardt, Recht, and Singer (ICML-16):
  • For convex, $\beta$-smooth, $\rho$-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 $\lambda$-strongly convex, $\beta$-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