More Basic Principles

- $\begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} \vec{w} \\ \vec{f} \end{bmatrix} = \vec{y}$ $\Rightarrow$ psuedo-inverse $\Rightarrow \hat{w} = X^T (XX^T + \lambda I)^{-1} \vec{y}$

- Using SVD: $Xw \Rightarrow U \Sigma \underbrace{V^T \vec{w}}_{\tilde{w}} \approx \vec{y} \Rightarrow \Sigma \tilde{w} \approx \underbrace{U^T \vec{y}}_{\tilde{y}}$

$\hookrightarrow \sigma_i \tilde{w}[i] \approx \tilde{y}[i] \Rightarrow \tilde{w}[i] = \frac{1}{\sigma} \tilde{y}[i]$

$\uparrow$ if small, $\tilde{w}[i]$ can blow up in OLS

Ridge: $(V \Sigma^T \Sigma V^T + \lambda I)^{-1} V \Sigma^T \underbrace{U^T y}_{\tilde{y}}$

$V(\Sigma^T \Sigma + \lambda I)^{-1} V^T V \Sigma^T \tilde{y} \Rightarrow \tilde{w} = (\Sigma^T \Sigma + \lambda I)^{-1} \Sigma^T \tilde{y}$

$\hookrightarrow \tilde{w}[i] = \frac{\sigma_i}{\sigma_i^2 + \lambda} \hat{y}[i]$ for very small $\sigma_i^2 << \lambda$, behaves as $\frac{1}{\lambda}$, doesn't blow up.

Implicit regularization

$\hookrightarrow$ optimizer does regularization for you

GD for OLS (SVD coord)

$\tilde{w}_{t+1} = \tilde{w}_t + 2\eta \Sigma^T (\tilde{y} - \Sigma \tilde{w}_t)$

$\tilde{w}[i]_{t+1} = \tilde{w}[i]_t + 2\eta \sigma_i (\tilde{y}[i] - \sigma_i \tilde{w}_t[i])$

GD moves slowly on ill conditioned points, so running limited iters limits growth

"early stopping"

Trade-offs between sources of error

$(Y - f_{\hat\theta}(x))^2$

3 error sources: irreducible, approximation, estimation

features in deep learning