Lecture 17: Iterative Solvers (continued)

Least squares: \[ \hat{w} = (X^TX)^{-1} X^T y \]

Gradient descent:

start with initial guess \( \hat{w}^{(0)} \)

for \( k = 0, 1, 2, \ldots \)

\[ \hat{w}^{(k+1)} = \hat{w}^{(k)} - \tau X^T (X \hat{w}^{(k)} - y) \]

minimize \[ \|y - Xw\|_2^2 \]

\[ f(w) \] is convex

if \( \tau < \frac{2}{\|X\|_{op}^2} \), then \[ \hat{w}^{(k)} \xrightarrow{k \to \infty} (X^TX)^{-1} X^Ty \]
Regularized Regression

\[
\hat{w} = \arg\min_w \| y - Xw \|^2_2 + \lambda \ r(w)
\]

- Regularizer
  - \( \lambda > 0 \)
  - \( r(w) = \| w \|^2_2 \) (Ridge)
  - \( r(w) = \| w \|^1_1 \) (Lasso)

Squared error loss measures data fit
Let \( f(w) = \|y - Xw\|_2^2 + \lambda r(w) \)

\[
= \left\| y - X\hat{w}^{(k)} + X\hat{w}^{(k)} - Xw \right\|_2^2 + \lambda r(w) \\
= \left\| y - X\hat{w}^{(k)} \right\|_2^2 + \| X(\hat{w}^{(k)} - w) \|_2^2 + \lambda r(w)
\]

does not depend on \( w \)

is Constant \( C \) with respect to \( w \)

\[
f(w) \leq C + \| X \|_{op} \| \hat{w}^{(k)} - w \|_2^2 + 2(y - X\hat{w}^{(k)})^T X(\hat{w}^{(k)} - w) + \lambda r(w)
\]

let \( \tau > 0 \) be a step size, and assume \( \tau < \frac{1}{\| X \|_{op}} \)
\[ f(w) \leq C + 2(y - \hat{w}^{(k)})^T X (\hat{w}^{(k)} - w) + \tau^{-1} \| \hat{w}^{(k)} - w \|^2 + \lambda r(w) \]

\[ f(w) \leq g_k(w) \]

\[ f(\hat{w}^{(k)}) = g_k(\hat{w}^{(k)}) \]
choose \( \hat{w}^{(k+1)} \) to minimize \( \frac{\partial}{\partial \hat{w}} g_k(w) \)

\[
\hat{w}^{(k+1)} = \arg \min_w 2 (y - X \hat{w}^{(k)})^T X (\hat{w}^{(k)} - w) + \tau^{-1} \| \hat{w}^{(k)} - w \|_2^2 + \lambda r(w)
\]

\[= \arg \min_w \tau (y - X \hat{w}^{(k)})^T X (\hat{w}^{(k)} - w) + \| \hat{w}^{(k)} - w \|_2^2 + \lambda \tau r(w)
\]

\[= v^T
\]

let \( v = \tau X^T (y - X \hat{w}^{(k)}) \) \( (v \text{ does not depend on } w) \)

\[\hat{w}^{(k+1)} = \arg \min_w 2v^T (\hat{w}^{(k)} - w) + \| \hat{w}^{(k)} - w \|_2^2 + \lambda \tau r(w)
\]

Complete square

\[= \arg \min_w \| v + \hat{w}^{(k)} - w \|_2^2 - \| v \|_2^2 + \lambda \tau r(w)
\]

\[= \arg \min_w \| v + \hat{w}^{(k)} - w \|_2^2 + \lambda \tau r(w)
\]
Let $\hat{z}^{(k)} = v + \hat{w}^{(k)}$

$= \hat{w}^{(k)} + \tau X^T(y - X\hat{w}^{(k)})$

$\underbrace{= \hat{w}^{(k)} - \tau X^T(X\hat{w}^{(k)} - y)}$

= Landweber iteration/step
= Grad. descent on squared error

$\hat{w}^{(k+1)} = \arg\min_w \|\hat{z}^{(k)} - w\|_2^2 + \lambda \tau r(w)$

Algorithm (Proximal Gradient Descent)
- choose $\hat{w}^{(0)}$ (initial guess)
- for $k = 0, 1, 2, \ldots$
  $\hat{z}^{(k)} = \hat{w}^{(k)} - \tau X^T(X\hat{w}^{(k)} - y)$ (GD)
  $\hat{w}^{(k+1)} = \arg\min_w \|\hat{z}^{(k)} - w\|_2^2 + \lambda \tau r(w)$ (regularize)
- check convergence: if $\|\hat{w}^{(k)} - \hat{w}^{(k+1)}\| < \varepsilon$, break
Ex: \[ r(w) = \|w\|^2 \]

\[
\hat{w}^{(k+1)} = \arg\min_w \| \hat{z}^{(k)} - w \|_2^2 + \lambda \tau \|w\|_2^2
\]

\[
\nabla_w g = -2(\hat{z}^{(k)} - w) + 2\lambda \tau w = 0
\]

\[
\Rightarrow (2 + 2\lambda \tau)w = 2\hat{z}^{(k)}
\]

\[
\hat{W}^{(k+1)} = \frac{\hat{z}^{(k)}}{1 + \lambda \tau}
\]

\[
\Rightarrow \hat{w}^{(k+1)} = \frac{1}{1 + \lambda \tau} \left( \hat{w}^{(k)} - \tau X^T(Xw^{(k)} - y) \right)
\]