Comments on Homework Assignments

David Rosenberg

New York University

March 2, 2015
Parameter Tuning

- Can start by trying many different orders of magnitude
  
  \[10^{-5}, 10^{-4}, \ldots, 10^{-1}, 10^0, 10^1, \ldots, 10^4, 10^5\]
  
  \[2^{-10}, 2^{-9}, \ldots, 2^{-1}, 2^0, 2^1, \ldots, 2^9, 2^{10}\]

- See where the action is... and zoom in!

- Keep zooming in until things aren’t improving on validation set.
Parameter Tuning

- If you want to plot all values on one graph, you may want to take logarithms of your axes.
SGD For Total Loss vs Average Loss

- Suppose we write linear regression objective as

\[ J(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 \]

- Then we can do gradient descent using this step direction:

\[ -\nabla J(w) = -\sum_{i=1}^{n} 2 (w^T x_i - y_i) x_i \]

- What about stochastic gradient descent?
- Do we just choose a random \((x_i, y_i)\) and step in direction

\[ -2 (w^T x_i - y_i) x_i? \]
SGD Step and Gradient Step Should have Same Expectation

- Expectation of gradient step is

\[
\mathbb{E} [-\nabla J(w)] = -\mathbb{E} \left[ \sum_{i=1}^{n} 2 (w^T X_i - Y_i) X_i \right]
\]

\[
= - \sum_{i=1}^{n} \mathbb{E} [2 (w^T X_i - Y_i) X_i]
\]

\[
= -n \mathbb{E} [2 (w^T X - Y) X]
\]

- Which is \(n\) times

\[
- \mathbb{E} [2 (w^T X_i - Y_i) X_i] = - \mathbb{E} [2 (w^T X - Y) X]
\]

- Proper SGD step for this objective is

\[
-n \times 2 (w^T X_i - Y_i) X_i
\]

- Alternatively, divide original objective by \(n\).
SGD For Total Loss vs Average Loss

- So we had
  \[ J(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 \]
- Proper SGD step is
  \[ -n \times 2 (w^T X_i - Y_i) X_i \]
- What if we take step
  \[ -2 (w^T X_i - Y_i) X_i \]?
- Then we’re optimizing
  \[ J_1(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \]
- Does it matter?
The objective functions

\[ J(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 \]

\[ J_1(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \]

have the same minimizer \( w^* \).

But they have different minimum values.
The objective functions

\[ J(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|^2 \]

\[ J_1(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|^2 \]

do not have the same minimizer \( w^* \) for the same \( \lambda \).

For the same \( \lambda \), which objective has the minimizer with smaller “complexity” \( \|w\|^2 \)?
Definition

A directional derivative of $f$ at $x$ in the direction $\delta x$ is

$$f'(x; \delta x) = \lim_{h \downarrow 0} \frac{f(x + h\delta x) - f(x)}{h},$$

and it can be $\pm \infty$ (e.g. for discontinuous functions).

- If $f$ is convex and finite near $x$, then $f'(x; \delta x)$ exists.
- $f$ is differentiable at $x$ iff for some $g(= \nabla f(x))$ and all $\delta x$,

$$f'(x; \delta x) = g^T \delta x.$$
Descent Directions and Optimality

Definition

$\delta x$ is a **descent direction** for $f$ at $x$ if $f'(x; \delta x) < 0$.

- For differentiable $f$, if $\nabla f(x) \neq 0$, then $\delta x = -\nabla f(x)$ is a descent direction.
- We have a nice characterization for a minimum in terms of directional derivative:

Theorem

*If $f$ is convex and finite near $x$, then either*

- $x$ minimizes $f$, or
- there is a descent direction for $f$ at $x$.  

---

Boyd EE364b: Subgradients Slides
Lasso objective

\[ J_\lambda(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda |w|_1 \]

Is there a \( \lambda_{\text{max}} \) such that \( \lambda \geq \lambda_{\text{max}} \) implies \( \text{arg min}_w J_\lambda(w) = 0 \)?

Suppose yes.

Then \( w = 0 \) is a minimum of \( J_\lambda(w) \).

Let’s see what that means in terms of our directional derivative characterization.
Consider a step direction $\nu$. For convenience, take $\nu$ s.t. $|\nu| = 1$.

Then directional derivative at $w = 0$ in direction $\nu$ is

$$J'_\lambda(0; \nu) = \lim_{h \downarrow 0} \frac{J(h\nu) - J(0)}{h}.$$ 

For $w = 0$ to be a minimizer, need to have $J'_\lambda(0; \nu) \geq 0$ for every direction $\nu$.

Can find $\lambda_{\text{max}}$ by finding conditions on $\lambda$ for this to be the case.