Optimization for Machine Learning
Optimization for machine learning

• Many engineering disciplines cannot survive without optimization
• Including machine learning
• The generic ERM + regularization minimization:

$$\min_{\mathbf{w}} \sum_{i}^{n} L(f_{\mathbf{w}}(x_i), y_i) + \Omega(\mathbf{w})$$

Minimize a sum of loss function on every training example
How to solve optimization problems

\[ \min_w f(w) \]

\[ \nabla f(w) = \left[ \frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \ldots, \frac{\partial f}{\partial w_d} \right]^T \]

• First-order condition (Stationarity):

\[ \nabla f (w) = 0 \]

• *Necessary* for optimality
  • Not *sufficient*!
  • Sufficient when \( f (w) \) convex (will talk about later)
Try take some gradients

• The gradient of \( \mathbf{w}^T \mathbf{x} \) w.r.t. to \( \mathbf{w} \)?
• The gradient of \( (\mathbf{w}^T \mathbf{x} - y)^2 \) w.r.t. to \( \mathbf{w} \)?
Gradient Descent

\[ \min_w f(w) \]

while \( \|\nabla w\| > \epsilon \)
\[ w = w - \alpha \nabla f(w) \]

\( \alpha \): Step size (Learning rate)
Line search, step size

- One needs the correct step size to converge faster
- In traditional optimization, in order to decide step-size, line search was often used on the descent direction
  - Satisfy certain conditions (e.g. Armijo-Goldstein, Frank-Wolfe)
Gradient direction can be bad

• $\min_{w_1,w_2} (w_1 - 1)^2 + 100(w_2 - 1)^2$

• What is $\nabla f (\mathbf{w})$?
• What is $\nabla f (\mathbf{w})$ at (0,0)?

• What is a good step size?

• That’s why usually need second-order information
  • Curiously deep learning does not often use second-order information
Hessian

• The Hessian: \( \mathbf{H} = \nabla^2 f = \left( \begin{array}{ccc} \frac{\partial^2 f}{\partial w_1^2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_d \partial w_1} & \cdots & \frac{\partial^2 f}{\partial w_d^2} \end{array} \right) \)

• A second-order Taylor expansion:

\[
f(w) = f(a) + \nabla_w f(a)(w - a) + \frac{1}{2}(w - a)^T \mathbf{H}(a)(w - a) + o(||w - a||^2)
\]
Newton direction

\[ d = [\nabla^2 f(w)]^{-1} \nabla f(w) \]

- e.g. \[ \min_{w_1, w_2} (w_1 - 1)^2 + 100(w_2 - 1)^2 \]
- Algorithm:

\[
\begin{align*}
\text{while } & \| \nabla f(w) \| > \epsilon \\
& w = w - \alpha d
\end{align*}
\]

- Other variants of Newton-type methods:
  - Quasi-Newton (e.g. BFGS, use an approximation of Hessian)
  - Limited Memory Quasi-Newton (use a low-rank Hessian)
  - Barzilai-Borwein (use diagonal of Hessian)
Convexity

• F is convex if

\[ \forall x_1, x_2 \in X, \forall t \in [0,1], \]
\[ f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2) \]

• First order condition:

\[ f(y) \geq f(x) + \nabla f(x)^T (y - x). \]

• Second order condition:

\[ \nabla^2 f(x) \geq 0 \]
Positive semi-definiteness review

• Important concept in linear algebra

• $\mathbf{M}$ p. s. d. $\iff \mathbf{z}^T \mathbf{M} \mathbf{z} \geq 0$

• All eigenvalues of $\mathbf{M}$ are nonnegative

• All principal minors are nonnegative

• Positive-definiteness:
  • (Change $\geq 0$ to $> 0$)
Saddle Point

- Stationarity doesn’t necessarily mean local optimum
  - Simple example: $z = x^2 - y^2$
  - $x = 0, y = 0$

- Definition of local optimum
  - Locally convex
Nonconvexity