Optimization Methods in Machine Learning
Lecure 9: Unconstrained optimization, logistic regression

Katya Scheinberg

Lehigh University

Spring 2016

Motivation

- Recall that we have the optimization problem

$$\min f(w) = \frac{1}{m}\sum_i (1 + e^{-y_i x_i^T w}) + \lambda \|w\|_2^2$$

- We want to have an iterative method to approach the solution. For the optimal solution

$$w^* = \arg\min f(w)$$

we want to make steps $\{w^k\}$ so that as $k \to \infty$, $w^k \to w^*$, $f(w^k) \to f(w^*)$, and $\nabla f(w^k) \to 0$.

Gradient Descent Method

$$\min f(x)$$

Assume that $\nabla f(x)$ and possibly $\nabla^2 f(x)$ exist

- We introduce the gradient descent method as a common iterative method for solving optimization problems of this type.
- A typical iteration of this method will take the form:

$$x^{k+1} = x^k + t^k d^k$$

  where t^k is the step size, and d^k is the search direction for the k^{th} step.
- Assume that $(d^k)^T \nabla f(x^k) < 0$.

Gradient Descent Method

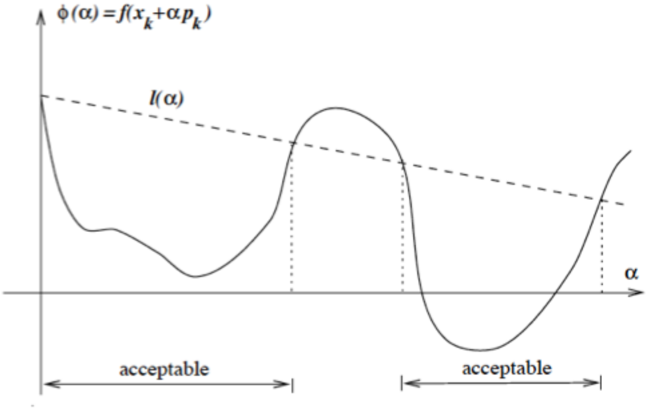
- The Gradient Descent Method implements a conditions to ensure convergence:
  - Sufficient Decrease Condition:

$$f(x^k + t^k d^k) \leq f(x) + \alpha t^k \nabla f(x)^T d^k$$

- For Gradient Descent method, the search direction is often the opposite direction of the gradient: $d^k = -\nabla f(x^k)$.

Sufficient Decrease Condition

- The sufficient decrease condition ensures that each iteration will make sufficient progress in terms of reducing objective value.



Rate of Convergence

- Assume that $f(x)$ is Lipschitz smooth : $f(x+s) \leq f(x) + \nabla f(x)^T s + \frac{L}{2}\|s\|^2$
- Assume that $f(x)$ is strongly convex: $f(x+s) \geq f(x) + \nabla f(x)^T s + \frac{\mu}{2}\|s\|^2$.
- The above can be ensured if

$$\nabla^2 f(x) \succeq \mu I$$

  for some $\mu > 0$.
- The rate of convergence for the gradient descent method with $d^k = \nabla f(x^k)$ is

$$f(x^k) - f(x^*) \leq c^k(f(x^0) - f(x^*)), c \in (0,1) \tag{1}$$

  where $C = \frac{\gamma-1}{\gamma+1}$, and $\gamma = \frac{L}{\mu}$ (the ratio of the largest value of Hessian and smallest value of Hessian)
- So if we want an accuracy of $\epsilon = 10^{-3}$, then $c^k = \frac{1}{1000}$. So $c^k \approx \epsilon$, $k \approx \log_c^\epsilon$.

Convergence of gradient descent

Slides from L. Vandenberghe http://www.ee.ucla.edu/ vandenbe/ee236c.html

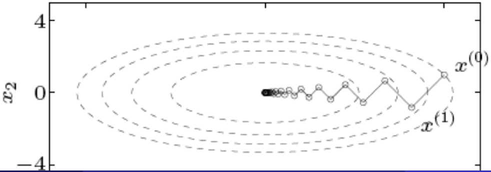
quadratic problem in $\mathbf{R}^2$

$$f(x) = (1/2)(x_1^2 + \gamma x_2^2) \qquad\qquad (\gamma > 0)$$

with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1}\right)^k, \qquad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1}\right)^k$$

- very slow if $\gamma \gg 1$ or $\gamma \ll 1$
- example for $\gamma = 10$:



Steepest descent Vs. Gradient descent

1. Normalized Steepest Descent

$$d_{nsd} = \arg\min\{\nabla f(x)^T v \mid \|v\| = 1\} \tag{2}$$

and Gradient Descent

$$d^k = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|_2} \tag{3}$$

What is the difference?

Steepest descent Vs. Gradient descent

1. Difference is the choice of norm.
   If we use $\|\cdot\|_2$

$$d_{nsd} = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|_2} \tag{4}$$

If we use $\|\cdot\|_1$

$$d_{nsd} = \arg\max |\frac{\partial f(x)}{\partial x_i}| - \text{sign}\frac{\partial f(x)}{\partial x_{i^*}} e_{i^*} \tag{5}$$

If we use $\|v\|_M = v^T M v$ where $M \succeq 0$

$$d_{nsd} = -\frac{M^{-1}\nabla f(x^k)}{\|M^{-1}\nabla f(x^k)\|_2} \tag{6}$$

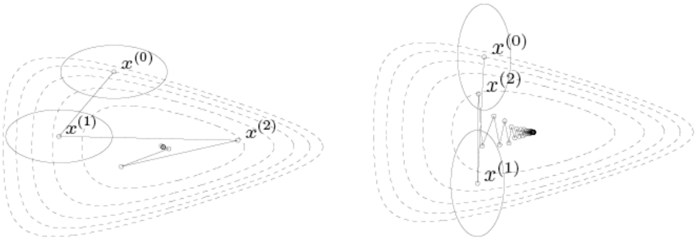
How to pick M?

$$d_{\text{Newton}} = -\nabla^2 f(x)^{-1}\nabla f(x) \tag{7}$$

Convergence of gradient descent

Slides from L. Vandenberghe http://www.ee.ucla.edu/ vandenbe/ee236c.html

choice of norm for steepest descent



- steepest descent with backtracking line search for two quadratic norms
- ellipses show $\{x \mid \|x - x^{(k)}\|_P = 1\}$
- equivalent interpretation of steepest descent with quadratic norm $\|\cdot\|_P$: gradient descent after change of variables $\bar{x} = P^{1/2}x$

shows choice of P has strong effect on speed of convergence

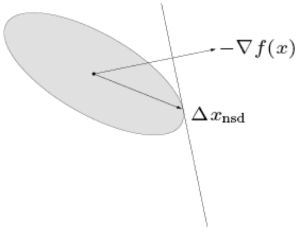
Convergence of gradient descent

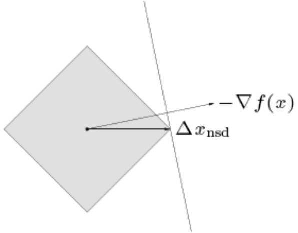
Slides from L. Vandenberghe http://www.ee.ucla.edu/ vandenbe/ee236c.html

examples

- Euclidean norm: $\Delta x_{\rm sd} = -\nabla f(x)$
- quadratic norm $\|x\|_P = (x^T P x)^{1/2}$ ($P \in \mathbf{S}^n_{++}$): $\Delta x_{\rm sd} = -P^{-1}\nabla f(x)$
- ℓ_1-norm: $\Delta x_{\rm sd} = -(\partial f(x)/\partial x_i)e_i$, where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

unit balls and normalized steepest descent directions for a quadratic norm and the ℓ_1-norm:



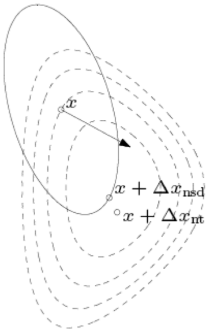


Convergence of gradient descent

Slides from L. Vandenberghe http://www.ee.ucla.edu/ vandenbe/ee236c.html

- $\Delta x_{\rm nt}$ is steepest descent direction at x in local Hessian norm

$$\|u\|_{\nabla^2 f(x)} = \left(u^T \nabla^2 f(x) u\right)^{1/2}$$



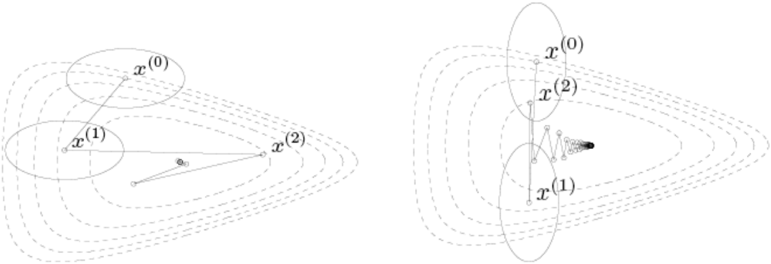
dashed lines are contour lines of f; ellipse is $\{x + v \mid v^T \nabla^2 f(x) v = 1\}$

arrow shows $-\nabla f(x)$

Convergence of gradient descent

Slides from L. Vandenberghe http://www.ee.ucla.edu/ vandenbe/ee236c.html

choice of norm for steepest descent



- steepest descent with backtracking line search for two quadratic norms
- ellipses show $\{x \mid \|x - x^{(k)}\|_P = 1\}$
- equivalent interpretation of steepest descent with quadratic norm $\|\cdot\|_P$: gradient descent after change of variables $\bar{x} = P^{1/2}x$

Slides from L. Vandenberghe http://www.ee.ucla.edu/ vandenbe/ee236c.html

Newton step

$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

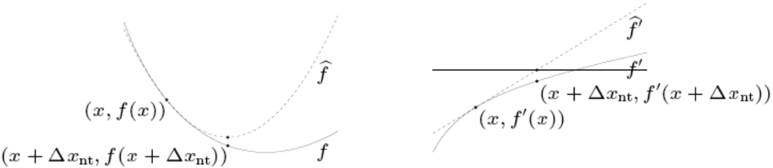
interpretations

- $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation

$$\widehat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

- $x + \Delta x_{\mathrm{nt}}$ solves linearized optimality condition

$$\nabla f(x+v) \approx \nabla \widehat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$

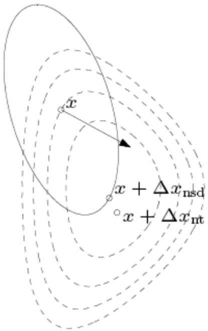


Convergence of gradient descent

Slides from L. Vandenberghe http://www.ee.ucla.edu/ vandenbe/ee236c.html

- Δx_{nt} is steepest descent direction at x in local Hessian norm

$$\|u\|_{\nabla^2 f(x)} = \left(u^T \nabla^2 f(x) u\right)^{1/2}$$



dashed lines are contour lines of f; ellipse is $\{x + v \mid v^T \nabla^2 f(x) v = 1\}$

arrow shows $-\nabla f(x)$

Optimization

We come back to the optimization problem

$$\min f(w) = \frac{1}{m}\sum(1 + e^{-y_i x_i^T w}) + \lambda\|w\|_2^2 \tag{8}$$

And we derive the gradient

$$\nabla f(w) = \frac{1}{m}\sum \frac{1}{e^{-y_i x_i^T w} + 1}(-y_i x_i) + \lambda w \tag{9}$$

And the Hessian

$$\nabla^2 f(w) = \frac{1}{m}\sum \frac{e^{y_i x_i^T w}}{(1 + e^{y_i x_i^T w})^2}(y_i x_i)(y_i x_i)^T + \lambda I \tag{10}$$

Notice that the time complexity for computing the Hessian is $d^2 m$, size of Hessian is $d \times d$, the complexity for computing the inverse of Hessian is d^3. So computation can be very expensive for large Hessian!