Gradient descent method

2013.11.10

Sanghyuk Chun

Many contents are from
Large Scale Optimization Lecture 4 & 5 by Caramanis & Sanghavi
Convex Optimization Lecture 10 by Boyd & Vandenberghe
Convex Optimization textbook Chapter 9 by Boyd & Vandenberghe

Contents

Introduction

Unconstraint minimization problem, Description, Pros and Cons

Unconstrained minimization problems

- Recall: Constrained minimization problems
  - From Lecture 1, the formation of a general constrained convex optimization problem is as follows
    - $\min f(x) \ s.t. \ x \in \chi$
    - Where $f: \chi \rightarrow \mathrm{R}$ is convex and smooth
  - From Lecture 1, the formation of an unconstrained optimization problem is as follows
    - $\min f(x)$
    - Where $f: R^n \rightarrow R$ is convex and smooth
    - In this problem, the necessary and sufficient condition for optimal solution x0 is
      - $\nabla f(x) = 0 \ at \ x = x_0$

Unconstrained minimization problems

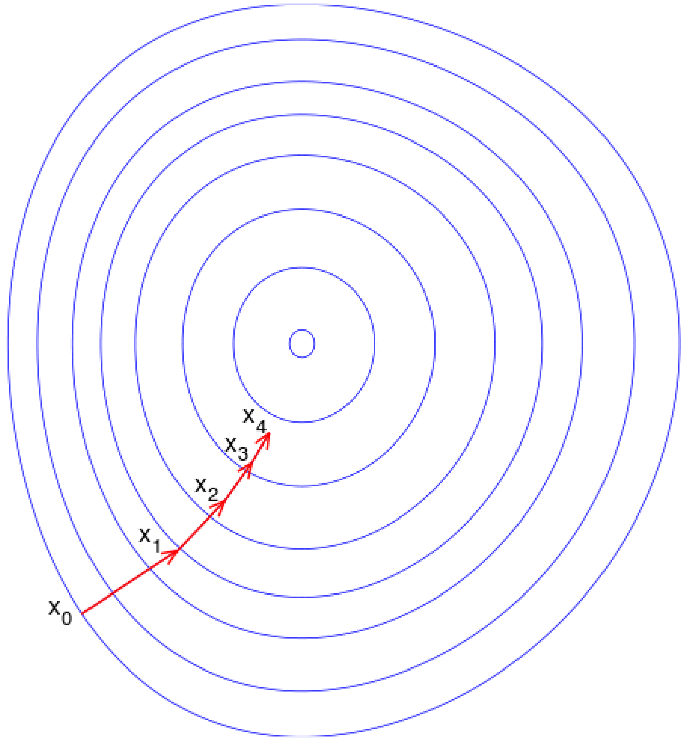
- Minimize f(x)
  - When f is differentiable and convex, a necessary and sufficient condition for a point x^* to be optimal is $\nabla f(x^*) = 0$

- Minimize f(x) is the same as fining solution of $\nabla f(x^*) = 0$
  - Min f(x): Analytically solving the optimality equation
  - $\nabla f(x^*) = 0$: Usually be solved by an iterative algorithm

Description of Gradient Descent Method

- The idea relies on the fact that $-\nabla f(x^{(k)})$ is a descent direction

- $x^{(k+1)} = x^{(k)} - \eta^{(k)} \nabla f(x^{(k)})\ with\ f\left(x^{(k+1)}\right) < f(x^{(k)})$
  - $\Delta x^{(k)}$ is the step, or search direction
  - $\eta^{(k)}$ is the step size, or step length
    - Too small $\eta^{(k)}$ will cause slow convergence
    - Too large $\eta^{(k)}$ could cause overshoot the minima and diverge



Description of Gradient Descent Method

- Algorithm (Gradient Descent Method)
  - **given** a starting point $x \in dom\ f$
  - **repeat**
    1. $\Delta x := -\nabla f(x)$
    2. Line search: Choose step size η via exact or backtracking line search
    3. Update $x := x + \eta \Delta x$
  - **until** stopping criterion is satisfied
- Stopping criterion usually $\|\nabla f(x)\|_2 \leq \epsilon$
- Very simple, but often very slow; rarely used in practice

Pros and Cons

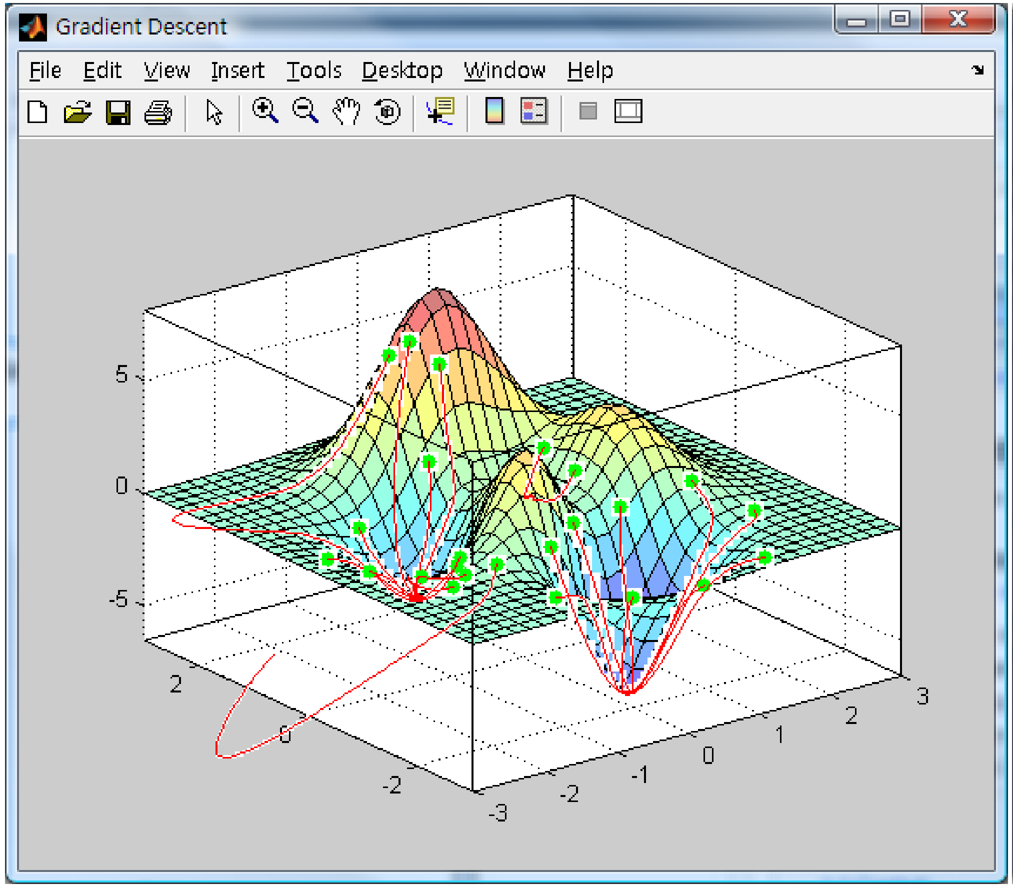
- Pros
  - Can be applied to every dimension and space (even possible to infinite dimension)
  - Easy to implement
- Cons
  - Local optima problem
  - Relatively slow close to minimum
  - For non-differentiable functions, gradient methods are ill-defined

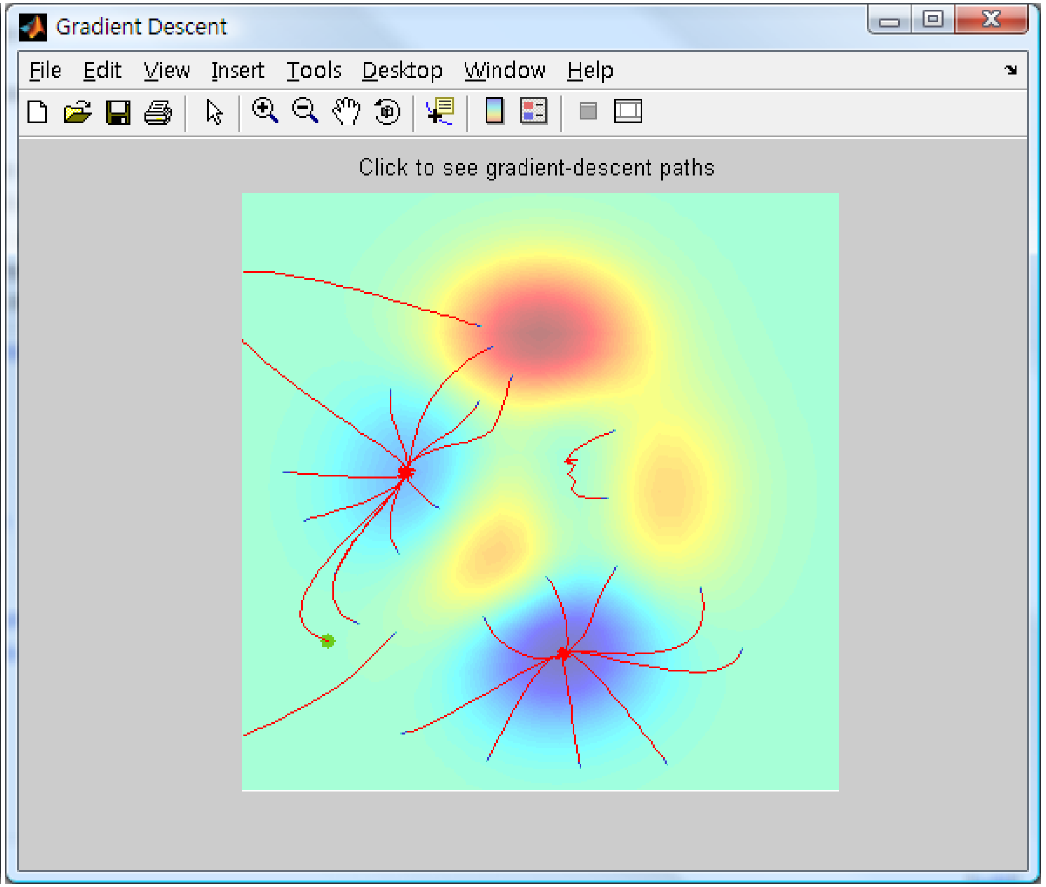
Example Code & Usage

Example Code, Usage, Questions

Gradient Descent Example Code

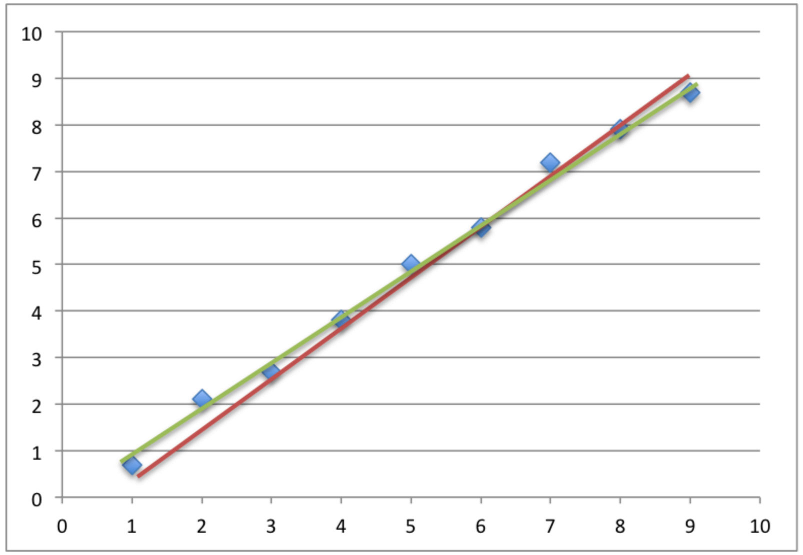
- http://mirlab.org/jang/matlab/toolbox/machineLearning/





Usage of Gradient Descent Method

- Linear Regression
  - Find minimum loss function to choose best hypothesis



Example of Loss function:

$$\sum \left(data_{predict} - data_{observed}\right)^2$$

Find the hypothesis (function) which minimize the loss function

Usage of Gradient Descent Method

- Neural Network
  - Back propagation
- SVM (Support Vector Machine)
- Graphical models
- Least Mean Squared Filter

…and many other applications!

Questions

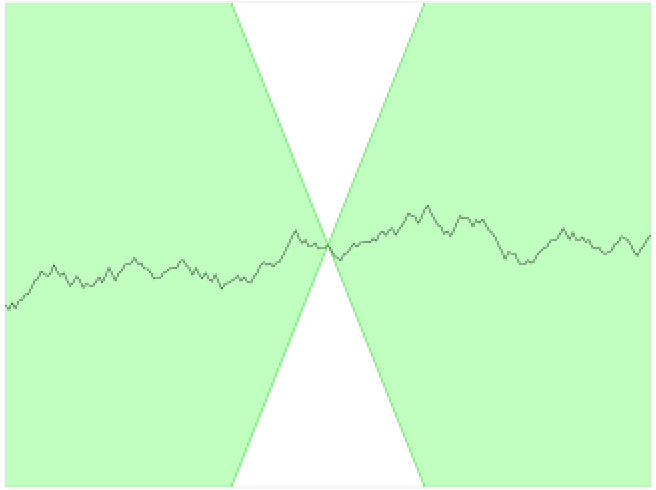
- Does Gradient Descent Method always converge?
- If not, what is condition for convergence?
- How can make Gradient Descent Method faster?
- What is proper value for step size $\eta^{(k)}$

Convergence Conditions

L-Lipschitz function, Strong Convexity, Condition number

L-Lipschitz function

- Definition
  - A function $f\colon R^n \to R$ is called *L-Lipschitz* if and only if
    $$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \forall x, y \in R^n$$
  - We denote this condition by $f \in C_L$, where C_L is class of *L-Lipschitz* functions

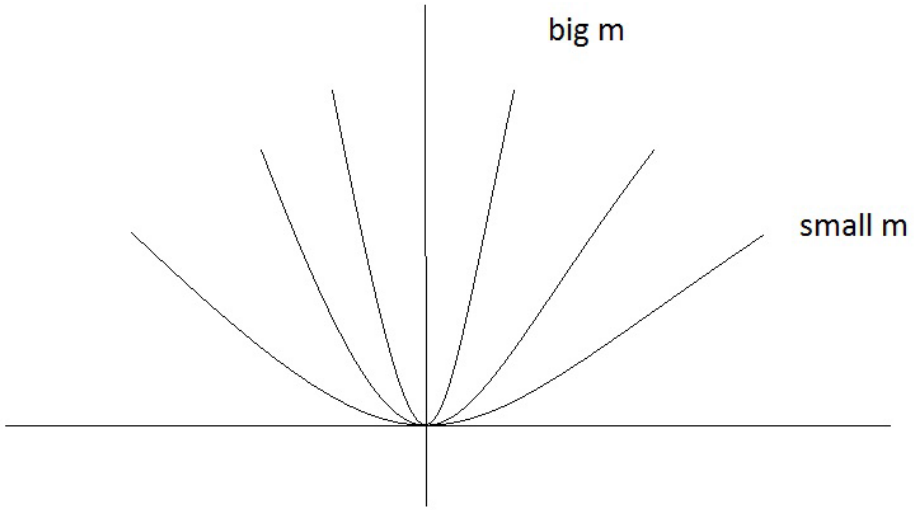


L-Lipschitz function

- Lemma 4.1
  - $If\ f \in C_L, then\ |f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2$
- Theorem 4.2
  - $If\ f \in C_L\ and\ f^* = \min_x f(x) > -\infty, then\ the\ gradient\ descent\ algorithm\ with\ fixed\ step\ size\ statisfying\ \eta < \frac{2}{L}\ will\ converge\ to\ a\ stationary\ point$

Strong Convexity and implications

- Definition
  - If there exist a constant m > 0 such that $\nabla^2 f \succ= mI \ for \ \forall x \in S$, then the function f(x) is strongly convex function on S



Strong Convexity and implications

- Lemma 4.3
  - If f is strongly convex on S, we have the following inequality:
    - $f(y) \geq f(x) + < \nabla f(x), y - x > + \frac{m}{2}\|y - x\|^2 \; for \; \forall x, y \; \in S$
- Proof

For $x, y \in S$, we have

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x)$$

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|^2 \\ &\geq f(x) + \langle \nabla f(x), \tilde{y} - x \rangle + \frac{m}{2}\|\tilde{y} - x\|^2 \qquad (\tilde{y} = x - (1/m)\nabla f(x) \;) \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \end{aligned}$$

$f^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$ for any $y \in S$. useful as stopping criterion (if you know m)

Strong Convexity and implications

Similarly, we can also derive a bound on $\|x - x^*\|_2$

$$\|x - x^*\|_2 \;\leq\; \frac{2}{m}\|\nabla f(x)\|_2^2 \quad \text{where } x^* = \arg\min_x f(x).$$

Proof

$$\begin{aligned} f^* = f(x^*) &\geq f(x) + < \nabla f(x), x^* - x > + \frac{m}{2}\|x^* - x\|_2^2 \\ &\geq f(x) - \|\nabla f(x)\|_2 \|x^* - x\|_2 + \frac{m}{2}\|x^* - x\|_2^2, \end{aligned}$$

$$\text{Since } f^* \leq f(x), \quad -\|\nabla f(x)\|_2 - \|x^* - x\|_2 + \frac{m}{2}\|x^* - x\|_2^2 \leq 0$$

Upper Bound of $\nabla^2 f(x)$

- Lemma 4.3 implies that the sublevel sets contained in S are bounded, so in particular, S is bounded. Therefore the maximum eigenvalue of $\nabla^2 f(x)$ is bounded above on S
  - There exists a constant M such that $\nabla^2 f(x) =\prec MI \ for \ \forall x \ \in S$

- Lemma 4.4
  - $For\ any\ x, y\ \in S, if\ \nabla^2 f(x) =\prec MI\ for\ all\ x\ \in S\ then$

$$f(y) \leq f(x) +< \nabla f(x), y - x > + \frac{M}{2} \|y - x\|^2$$

Condition Number

- From Lemma 4.3 and 4.4 we have

$$mI =\prec \nabla^2 f(x) =\prec MI \; for \; \forall \, x \; \in S, m > 0, M > 0$$

  - The ratio k=M/m is thus an upper bound on the condition number of the matrix $\nabla^2 f(x)$
  - When the ratio is close to 1, we call it *well-conditioned*
  - When the ratio is much larger than 1, we call it *ill-conditioned*
  - When the ratio is exactly 1, it is the best case that only one step will lead to the optimal solution (there is no wrong direction)

Condition Number

- Theorem 4.5
  - Gradient descent for a strongly convex function f and step size η=$\frac{1}{M}$ will converge as
    - $f(x^*) - f^* \leq c^k(f(x^0) - f^*), where\ c \leq 1 - \frac{m}{M}$
    - Rate of convergence c is known as linear convergence

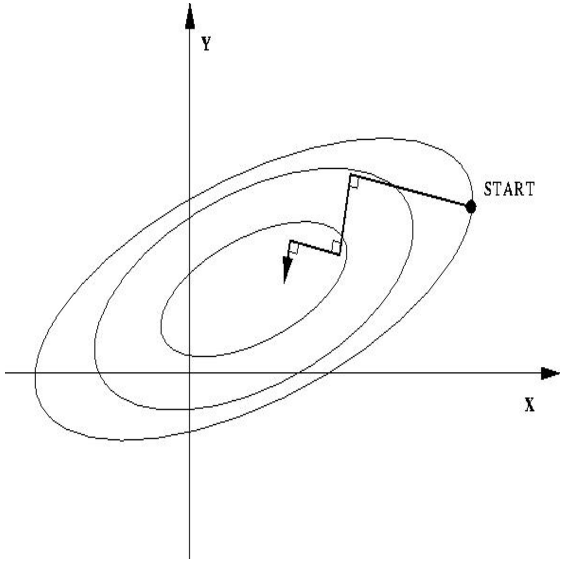
- Since we usually do not know the value of M, we do line search
  - For exact line search, $c = 1 - \frac{m}{M}$
  - For backtracking line search, $c = 1 - \min\left\{2m\alpha, \frac{2\beta\alpha m}{M}\right\} < 1$

Methods & Examples

Exact Line Search, Backtracking Line Search,
Coordinate Descent Method, Steepest Descent Method

Exact Line Search

- The optimal line search method in which η is chosen to minimize f along the ray $\{x - \eta \nabla f(x)\}$, as shown in below



Algorithm (Gradient descent with exact line search)

1. Set iteration counter $k = 0$, and make an initial guess x_0 for the minimum
2. Compute $\nabla f(x^{(k)})$
3. Choose $\eta^{(k)} = \arg\min_{\eta} \{f\left(x^{(k)} - \eta \nabla f(x^{(k)})\right)\}$
4. Update $x^{(k+1)} = x^{(k)} - \eta^{(k)} \nabla f(x^{(k)})$ and $k = k + 1$.
5. Go to 2 until $\|\nabla f(x^{(k)})\| < \epsilon$

- Exact line search is used when the cost of minimization problem with one variable is low compared to the cost of computing the search direction itself.
- It is not very practical

Exact Line Search

- Convergence Analysis
  - $$\begin{aligned} f(x^+) &\leq f(x - \frac{1}{M}\nabla f(x)) \\ &\leq f(x) - \frac{1}{M}\|\nabla f(x)\|_2^2 + \frac{M}{2}(\frac{1}{M})^2\|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2 \end{aligned}$$

    $$\Rightarrow$$

    $$f(x^+) - f^* \leq f(x) - f^* - \frac{1}{2M}\|\nabla f(x)\|_2^2$$

    Recall the analysis for strong convexity: $\|\nabla f(x)\|_2^2 \geq 2m(f(x) - f(y))$

    Thus, the following inequality holds: $f(x^+) - f^* \leq \left(1 - \frac{m}{M}\right)(f(x) - f^*)$
  - $|f(x^{(k)}) - f^*|$ decreases by at least a constant factor in every iteration
  - Converging to 0 geometric fast. (linear convergence)

Backtracking Line Search

- It depends on two constants $\alpha, \beta\ with\ 0 < \alpha < 0.5, 0 < \beta < 1$
- It starts with unit step size and then reduces it by the factor β until the stopping condition

$$f(x - \eta \nabla f(x)) \leq f(x) - \alpha\eta ||\nabla f(x)||^2$$

- Since $-\nabla f(x)$ is a descent direction and $-||\nabla f(x)||^2 < 0$, so for small enough step size η, we have

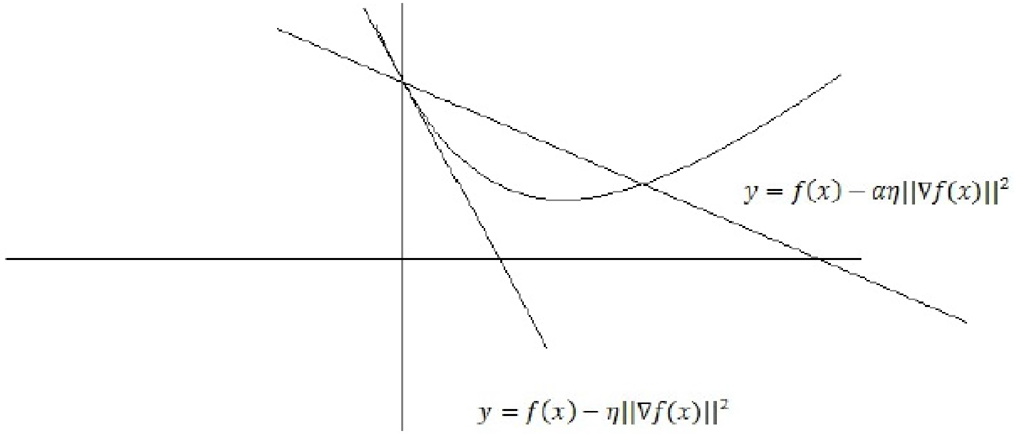
$$f\left(x - \eta \nabla f(x)\right) \approx f(x) - \eta ||\nabla f(x)||^2 < f(x) - \alpha\eta ||\nabla f(x)||^2$$

  - It shows that the backtracking line search eventually terminates
  - α is typically chosen between 0.01 and 0.3
  - β is often chosen to be between 0.1 and 0.8

Backtracking Line Search

- **Algorithm**

  1. Set iteration counter $k = 0$. Make an initial guess x^0 and choose initial $\eta = 1$.
  2. Update $\eta^k = \beta \eta^k$
  3. Go to 2 until $f(x^k - \eta^k \nabla f(x^k)) \leq f(x^k) - \alpha \eta^k \|\nabla f(x^k)\|^2$.
  4. Calculate $x^{k+1} = x^k - \eta^k \nabla f(x^k)$ and update $k = k + 1$.
  5. Go to 1 until $\|\nabla f(x^{(k)})\| < \epsilon$



Backtracking Line Search

- Convergence Analysis
  - Claim: $\eta \le \frac{1}{M}$ always satisfies the stopping condition
  - Proof

$$\text{Recall: } f(x^+) \;\le\; f(x) - \eta\|\nabla f(x)\|^2 + \frac{\eta^2 M}{2}\|\nabla f(x)\|^2$$

With the assumption that $\eta \le \frac{1}{M}$, the inequality implies that:

$$f(x^+) \;\le\; f(x) - \frac{\eta}{2}\|\nabla f(x)\|^2 \Rightarrow \eta \ge \frac{\beta}{M}$$

So overall,

$$\begin{aligned} \eta &\ge \min(1, \frac{\beta}{M}) \\ f(x^+) &\le f(x) - \alpha \min(1, \frac{\beta}{M})|\nabla f(x)\|^2 \end{aligned}$$

Backtracking Line Search

- Proof (cont)

Now, we subtract $f*$ from both sides to get:

$$f(x^+) - f^* \;\; \leq \;\; f(x) - f^* - \alpha \min(1, \frac{\beta}{M}) \|\nabla f(x)\|_2^2,$$

and combines with $\|\nabla f(x)\|_2^2 \geq 2m(f(x) - f^*)$ to obtain:

$$f(x^+) - f^* \;\; \leq \;\; (1 - \alpha \min(1, \frac{\beta}{M})(f(x) - f^*),$$

where

$$c \;\; = \;\; 1 - 2m\alpha \min\{1, \frac{\beta}{M}\} < 1$$

In particular, $f(x^k)$ converges to f^* at least as fast as a geometric series with an exponent that depends (at least in part) on the condition number bound $\frac{M}{m}$. As before with exact line search, the convergence is at least linear (but with a different factor).

Line search types

- Slide from Optimization Lecture 10 by Boyd

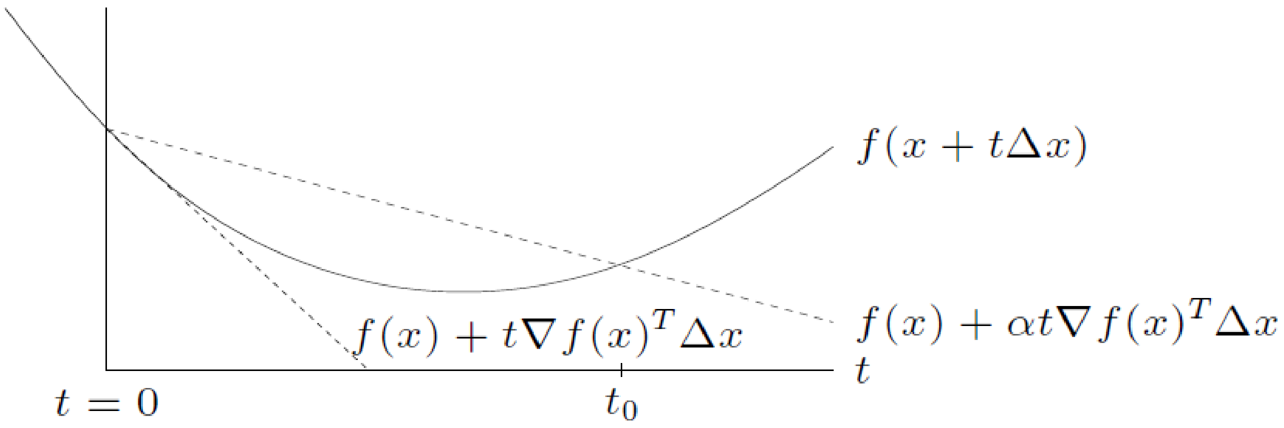
exact line search: $t = \operatorname{argmin}_{t>0} f(x + t\Delta x)$

backtracking line search (with parameters $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$)

- starting at $t = 1$, repeat $t := \beta t$ until

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x$$

- graphical interpretation: backtrack until $t \leq t_0$

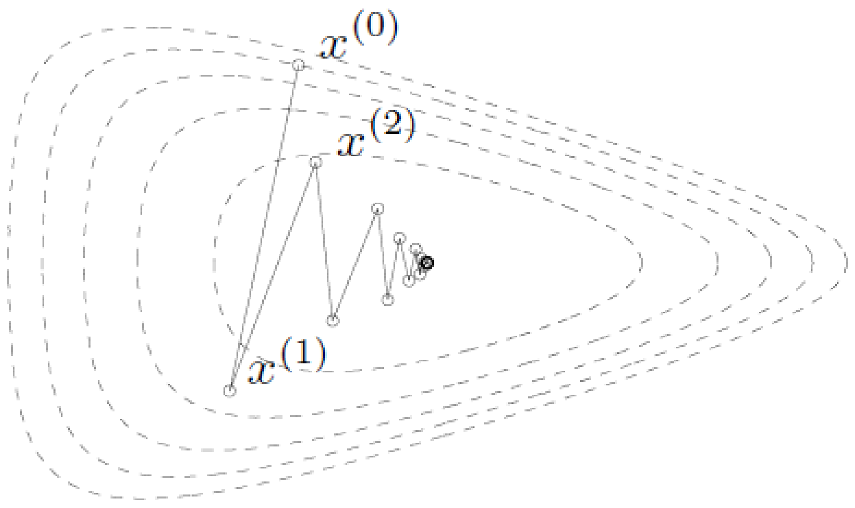


Line search example

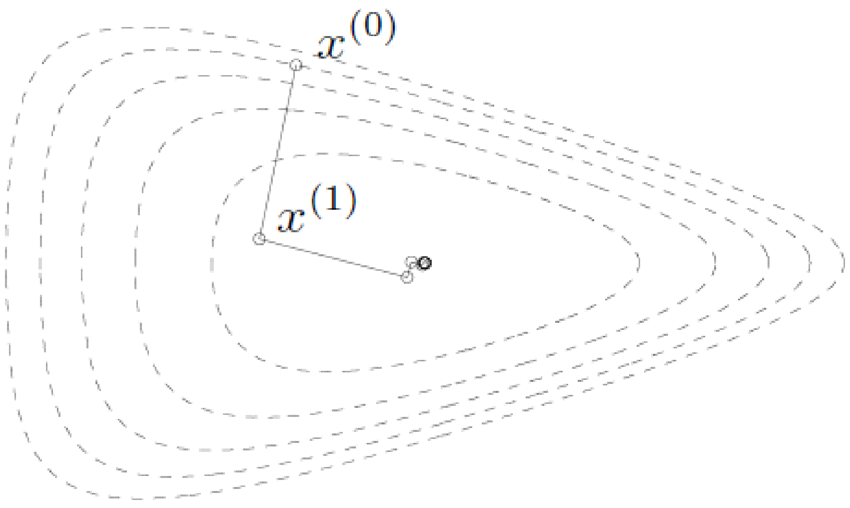
- Slide from Optimization Lecture 10 by Boyd

nonquadratic example

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$



backtracking line search



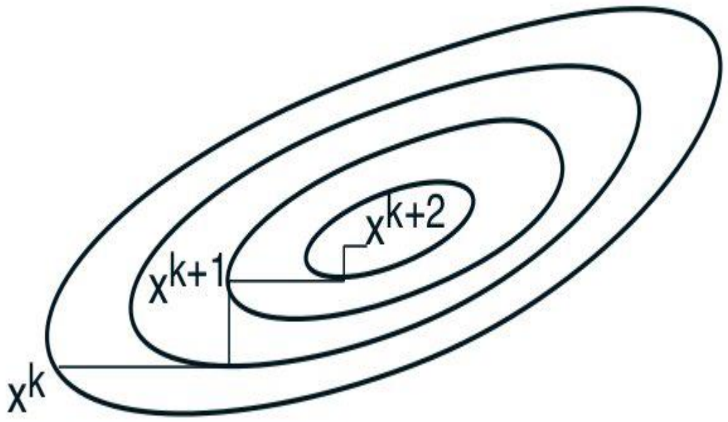
exact line search

Coordinate Descent Method

- Coordinate descent belongs to the class of several non derivative methods used for minimizing differentiable functions.
- Here, cost is minimized in one coordinate direction in each iteration.

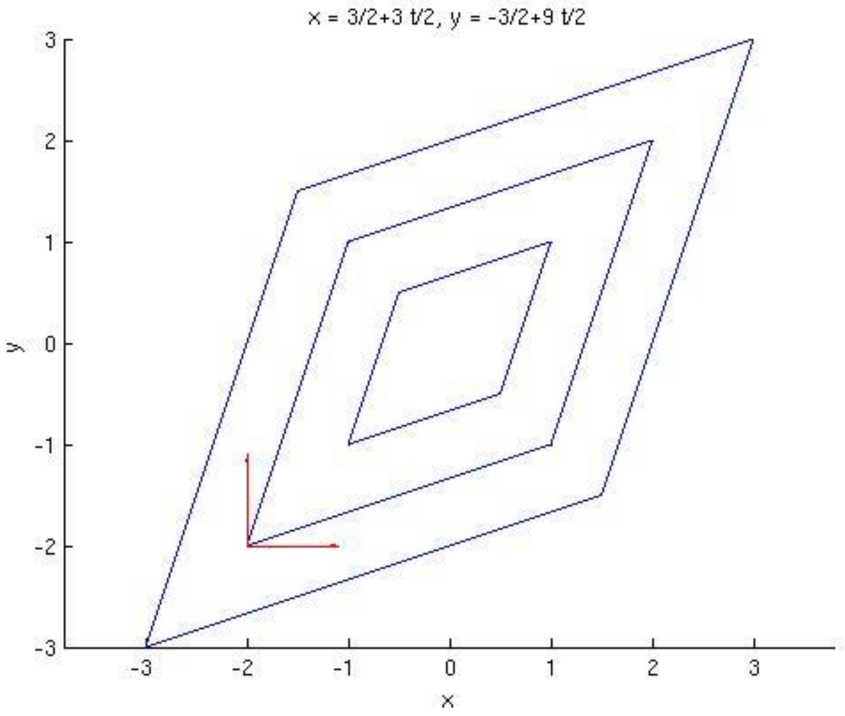
$$
\begin{aligned}
{x_j}^{(k+1)} &= {x_j}^{(k)}, j \neq i \\
{x_i}^{(k+1)} &= \arg\min_{\xi \in \Re} f(x_{\backslash i}^{(k)}, \xi)
\end{aligned}
$$

$$
{x_i}^{(k+1)} = {x_i}^{(k)} - \eta \frac{\partial f}{\partial x_i}(x^{(k)})
$$



Coordinate Descent Method

- Pros
  - It is well suited for parallel computation
- Cons
  - May not reach the local minimum even for convex function



Converge of Coordinate Descent

- Lemma 5.4

Lemma 5.4. *Suppose $\nabla f(x)$ is continuous and for every x and i, $f(x_{\setminus i}, \xi)$ has a unique minimum ξ^*, and is monotonic between x_i and ξ. Then cyclic coordinate descent with exact line search will reach stationary point. (Proposition 2.7.1, Bertsekas).*

Coordinate Descent Method

- Method of selecting the coordinate for next iteration
  - Cyclic Coordinate Descent
  - Greedy Coordinate Descent
  - (Uniform) Random Coordinate Descent

Steepest Descent Method

- The gradient descent method takes many iterations
- Steepest Descent Method aims at choosing the best direction at each iteration
- Normalized steepest descent direction
  - $\Delta x_{nsd} = argmin\{\nabla f(x)^T v \mid \|v\| = 1\}$
  - Interpretation: for small $v, f(x+v) \approx f(x) + \nabla f(x)^T v$ direction Δx_{nsd} is unit-norm step with most negative directional derivative
- Iteratively, the algorithm follows the following steps
  - Calculate direction of descent Δx_{nsd}
  - Calculate step size, t
  - $x_+ = x + t\Delta x_{nsd}$

Steepest Descent for various norms

- The choice of norm used the steepest descent direction can be have dramatic effect on converge rate
- l_2 norm
  - The steepest descent direction is as follows
    - $\Delta x_{nsd} = \frac{-\nabla f(x)}{\|\nabla f(x)\|_2}$
- l_1 norm
  - For $\|x\|_1 = \sum_i |x_i|$, a descent direction is as follows,
    - $\Delta x_{nds} = -sign\left(\frac{\partial f(x)}{\partial x_i^*}\right) e_i^*$
    - $i^* = arg\min_i \left|\frac{\partial f}{\partial x_i}\right|$
- l_∞ norm
  - For $\|x\|_\infty = \underset{i}{\operatorname{argmin}} |x_i|$, a descent direction is as follows
    - $\Delta x_{nds} = -sign(-\nabla f(x))$

Steepest Descent for various norms

Quadratic Norm

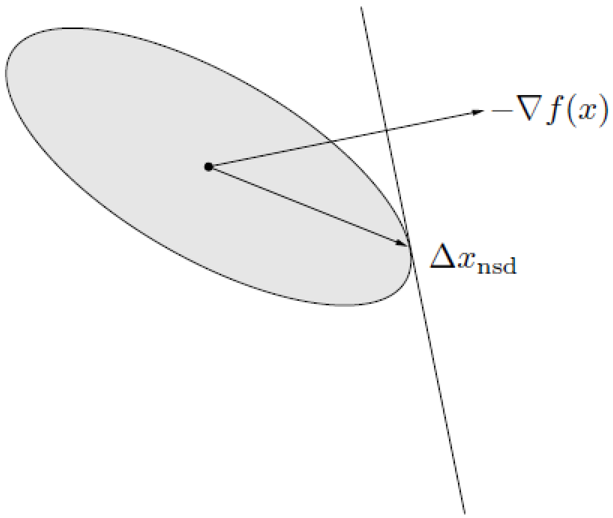


Figure 9.9 Normalized steepest descent direction for a quadratic norm. The ellipsoid shown is the unit ball of the norm, translated to the point x. The normalized steepest descent direction Δx_{nsd} at x extends as far as possible in the direction $-\nabla f(x)$ while staying in the ellipsoid. The gradient and normalized steepest descent directions are shown.

l_1-Norm

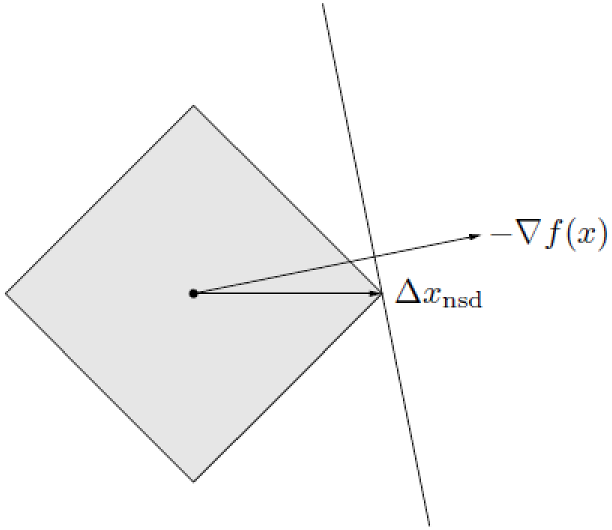


Figure 9.10 Normalized steepest descent direction for the ℓ_1-norm. The diamond is the unit ball of the ℓ_1-norm, translated to the point x. The normalized steepest descent direction can always be chosen in the direction of a standard basis vector; in this example we have $\Delta x_{\mathrm{nsd}} = e_1$.

Steepest Descent for various norms

- Example

$$P_1 = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}.$$

In both cases we use a backtracking line search with $\alpha = 0.1$ and $\beta = 0.7$.

This can be explained by examining the problems after the changes of coordinates $\bar{x} = P_1^{1/2}x$ and $\bar{x} = P_2^{1/2}x$, respectively.

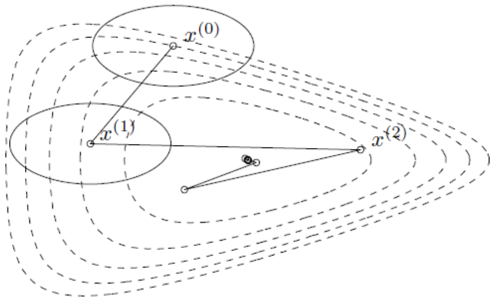


Figure 9.11 Steepest descent method with a quadratic norm $\|\cdot\|_{P_1}$. The ellipses are the boundaries of the norm balls $\{x \mid \|x - x^{(k)}\|_{P_1} \le 1\}$ at $x^{(0)}$ and $x^{(1)}$.

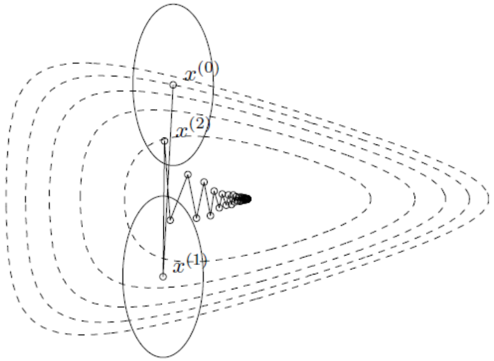


Figure 9.12 Steepest descent method, with quadratic norm $\|\cdot\|_{P_2}$.

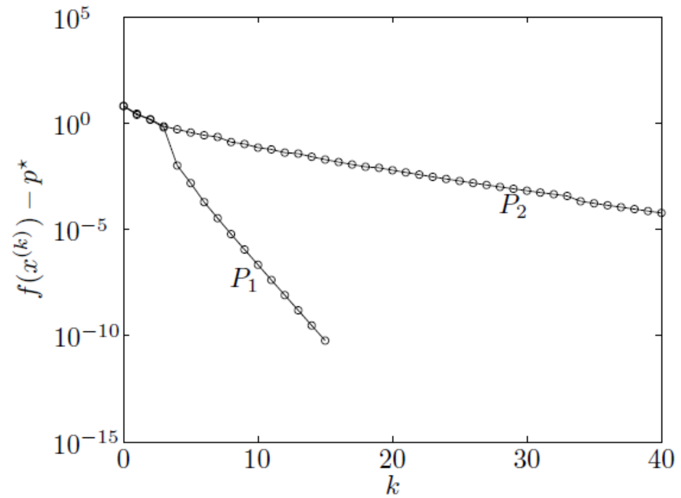


Figure 9.13 Error $f(x^{(k)}) - p^\star$ versus iteration k, for the steepest descent method with the quadratic norm $\|\cdot\|_{P_1}$ and the quadratic norm $\|\cdot\|_{P_2}$. Convergence is rapid for the norm $\|\cdot\|_{P_1}$ and very slow for $\|\cdot\|_{P_2}$.

Steepest Descent Convergence Rate

- Fact: Any norm can be bounded by $\|\cdot\|_2$, i.e., $\exists \gamma, \tilde{\gamma} \in (0,1]$ such that, $\|x\| \geq \gamma \|x\|_2 \; and \|x\|_* \geq \gamma \|x\|_2$
- Theorem 5.5
  - If f is strongly convex with respect to m and M, and $\|\cdot\|_2$ has $\gamma, \tilde{\gamma}$ as above then steepest decent with backtracking line search has linear convergence with rate
    - $c = 1 - 2m\alpha\tilde{\gamma}^2 \min\left\{1, \frac{\beta\gamma}{M}\right\}$
- Proof: Will be proved in the lecture 6

Summary

Summary

- Unconstrained Convex Optimization Problem
- Gradient Descent Method
- Step Size Trade-off between safety and speed
- Convergence Conditions
  - L-Lipschtiz Function
  - Strong Convexity
  - Condition Number

Summary

- Exact Line Search
- Backtracking Line Search
- Coordinate Descent Method
  - Good for parallel computation but not always converge
- Steepest Descent Method
  - The choice of norm is important

END OF DOCUMENT