Unconstrained Minimization (II)

Lijun Zhang
zlj@nju.edu.cn
http://cs.nju.edu.cn/zlj

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

General Descent Method

The Algorithm

Given a starting point $x \in \operatorname{dom} f$

Repeat

1. Determine a descent direction Δx.
2. Line search: Choose a step size $t \geq 0$.
3. Update: $x = x + t\Delta x$.

until stopping criterion is satisfied.

Descent Direction

$$\nabla f\left(x^{(k)}\right)^{\top} \Delta x^{(k)} < 0$$

Gradient Descent Method

The Algorithm

Given a starting point $x \in \mathrm{dom}\, f$

Repeat

1. $\Delta x := -\nabla f(x)$.
2. Line search: Choose step size t via exact or backtracking line search.
3. Update: $x := x + t\Delta x$.

until stopping criterion is satisfied.

Stopping Criterion

$$\|\nabla f(x)\|_2 \leq \eta$$

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

Preliminary

- $x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)} \Rightarrow x^{+} = x + t\Delta x$
- $\Delta x = -\nabla f(x)$
- $f(\cdot)$ is both strongly convex and smooth

$$mI \preccurlyeq \nabla^2 f(x) \preccurlyeq MI, \qquad \forall x \in S$$

- Define $\tilde{f}: \mathbf{R} \to \mathbf{R}$ as

$$\tilde{f}(t) = f(x - t\nabla f(x))$$

  - A quadratic upper bound on $\tilde{f}$

$$\tilde{f}(t) \le f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$$

Analysis for Exact Line Search

1. Minimize Both Sides of

$$\tilde{f}(t) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$$

- Left side: $\tilde{f}(t_{\text{exact}})$, where t_{exact} is the step length that minimizes $\tilde{f}$
- Right side: $t = 1/M$ is the solution

$$f(x^+) = \tilde{f}(t_{\text{exact}}) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$$

2. Subtracting p^* from Both Sides

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

Analysis for Exact Line Search

3. $f(\cdot)$ is strongly convex on S

$$\nabla^2 f(x) \succcurlyeq mI, \qquad \forall x \in S$$

$$\Rightarrow \|\nabla f(x)\|_2^2 \geq 2m(f(x) - p^*)$$

4. Combining

$$f(x^+) - p^* \leq (1 - m/M)(f(x) - p^*)$$

5. Applying it Recursively

$$f\left(x^{(k)}\right) - p^* \leq c^k\left(f\left(x^{(0)}\right) - p^*\right)$$

- $c = 1 - m/M < 1$
- $f(x^{(k)})$ coverges to p^* as $k \to \infty$

Discussions

Iteration Complexity

- $f\left(x^{(k)}\right) - p^* \le \epsilon$ after at most

$$\frac{\log\left(\left(f\left(x^{(0)}\right) - p^*\right)/\epsilon\right)}{\log(1/c)} \quad \text{iterations}$$

- $\log\left(\left(f\left(x^{(0)}\right) - p^*\right)/\epsilon\right)$ indicates that initialization is important
- $\log(1/c)$ is a function of the condition number M/m
- When M/m is large

$$\log(1/c) = -\log(1 - m/M) \approx m/M$$

Discussions

- Iteration Complexity
  - $f\left(x^{(k)}\right) - p^* \leq \epsilon$ after at most

$$\frac{\log\left(\left(f\left(x^{(0)}\right) - p^*\right)/\epsilon\right)}{\log(1/c)} \approx \frac{M}{m}\log\left(\left(f\left(x^{(0)}\right) - p^*\right)/\epsilon\right) \text{ iterations}$$

  - $\log\left(\left(f\left(x^{(0)}\right) - p^*\right)/\epsilon\right)$ indicates that initialization is important
  - $\log(1/c)$ is a function of the condition number M/m
  - When M/m is large

$$\log(1/c) = -\log(1 - m/M) \approx m/M$$

Discussions

Iteration Complexity

- $f(x^{(k)}) - p^* \leq \epsilon$ after at most

$$\frac{\log\left((f(x^{(0)}) - p^*)/\epsilon\right)}{\log(1/c)} \text{ iterations}$$

- $\log\left((f(x^{(0)}) - p^*)/\epsilon\right)$ indicates that initialization is important
- $\log(1/c)$ is a function of the condition number M/m
- Linear Convergence
  - ✓ Error lies below a line on a log-linear plot of error versus iteration number

Analysis for Backtracking Line Search

- Backtracking Line Search

given a descent direction Δx for f at $x \in \mathbf{dom}\, f$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$

$t := 1$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^{\top} \Delta x$, $t := \beta t$

1. $\tilde{f}(t) \leq f(x) - \alpha t \|\nabla f(x)\|_2^2$ for all $0 \leq t \leq 1/M$

$$0 \leq t \leq \frac{1}{M} \Rightarrow -t + \frac{Mt^2}{2} \leq -\frac{t}{2}$$

$$\tilde{f}(t) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$$

Analysis for Backtracking Line Search

Backtracking Line Search

given a descent direction Δx for f at $x \in \mathbf{dom}\, f$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$

$t := 1$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^{\top} \Delta x,\ t := \beta t$

1. $\tilde{f}(t) \leq f(x) - \alpha t \|\nabla f(x)\|_2^2$ for all $0 \leq t \leq 1/M$

$$\tilde{f}(t) \leq f(x) - (t/2)\|\nabla f(x)\|_2^2$$

$$\leq f(x) - \alpha t \|\nabla f(x)\|_2^2$$

- $\alpha < 1/2$

Analysis for Backtracking Line Search

2. Backtracking Line Search Terminates

- Either with $t = 1$

$$f(x^+) \leq f(x) - \alpha \|\nabla f(x)\|_2^2$$

- Or with a value $t \geq \beta/M$

$$f(x^+) \leq f(x) - (\beta\alpha/M) \|\nabla f(x)\|_2^2$$

- So,

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

3. Subtracting p^* from Both Sides

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

Analysis for Backtracking Line Search

4. Combining with Strong Convexity

$$f(x^+) - p^* \le \left(1 - \min\left\{2m\alpha, \frac{2\beta\alpha m}{M}\right\}\right)(f(x) - p^*)$$

5. Applying it Recursively

$$f\left(x^{(k)}\right) - p^* \le c^k\left(f\left(x^{(0)}\right) - p^*\right)$$

- $c = 1 - \min\left\{2m\alpha, \frac{2\beta\alpha m}{M}\right\} < 1$
- $f\left(x^{(k)}\right)$ converges to p^* with an exponent that depends on the condition number M/m
- Linear Convergence

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

A Quadratic Problem in $\mathbf{R}^2$

A Quadratic Objective Function

$$f(x) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right), \qquad \gamma > 0$$

- The optimal point $x^* = 0$
- The optimal value is 0
- The Hessian of f is constant and has eigenvalues 1 and γ
- $m = \min\{1, \gamma\}, M = \max\{1, \gamma\}$
- Condition number

$$\frac{\max\{1,\gamma\}}{\min\{1,\gamma\}} = \max\left\{\gamma, \frac{1}{\gamma}\right\}$$

A Quadratic Problem in $\mathbf{R}^2$

- A Quadratic Objective Function

$$f(x) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right), \qquad \gamma > 0$$

- Gradient Descent Method
  - Exact line search starting at $x^{(0)} = (\gamma, 1)$

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

Convergence is exactly linear

$$f\left(x^{(k)}\right) = \frac{\gamma(\gamma+1)}{2}\left(\frac{\gamma-1}{\gamma+1}\right)^{2k} = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f\left(x^{(0)}\right)$$

  - Reduced by the factor $|(\gamma-1)/(\gamma+1)|^2$

A Quadratic Problem in $\mathbf{R}^2$

Comparisons

- $m = \min\{1, \gamma\}, M = \max\{1, \gamma\}$
- From our general analysis, the error is reduced by $1 - \frac{m}{M}$
- From the closed-form solution, the error is reduced by

$$\left(\frac{\gamma - 1}{\gamma + 1}\right)^2 = \left(\frac{1 - m/M}{1 + m/M}\right)^2 = \left(1 - \frac{2m/M}{1 + m/M}\right)^2$$

- When M/m is large, the iteration complexity differs by a factor of 4

A Quadratic Problem in $\mathbf{R}^2$

- Experiments
  - For γ not far from one, convergence is rapid

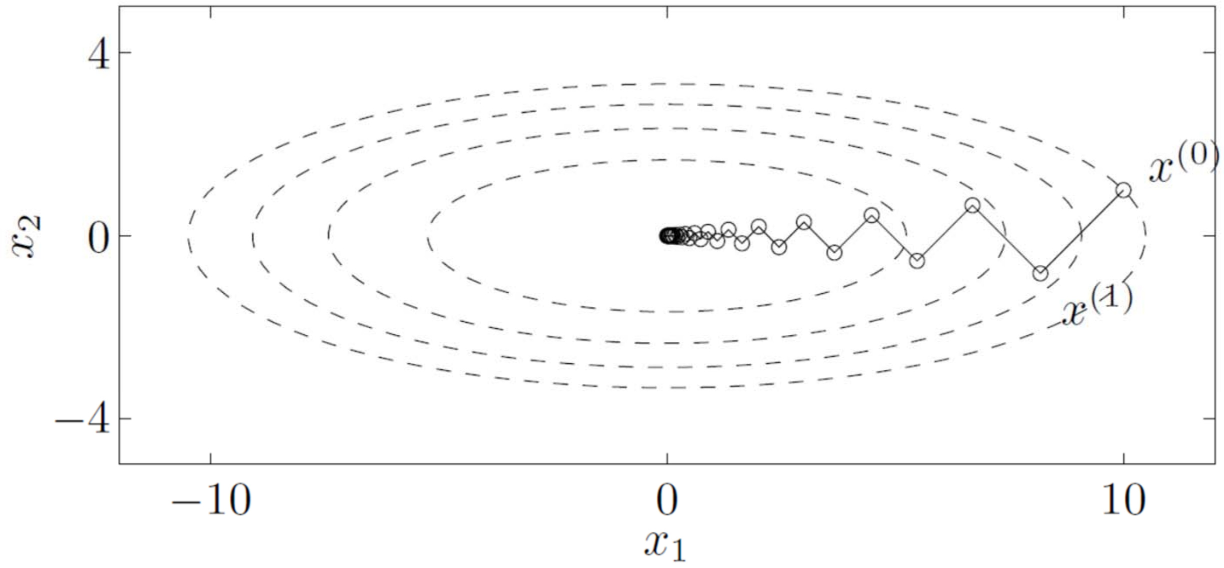


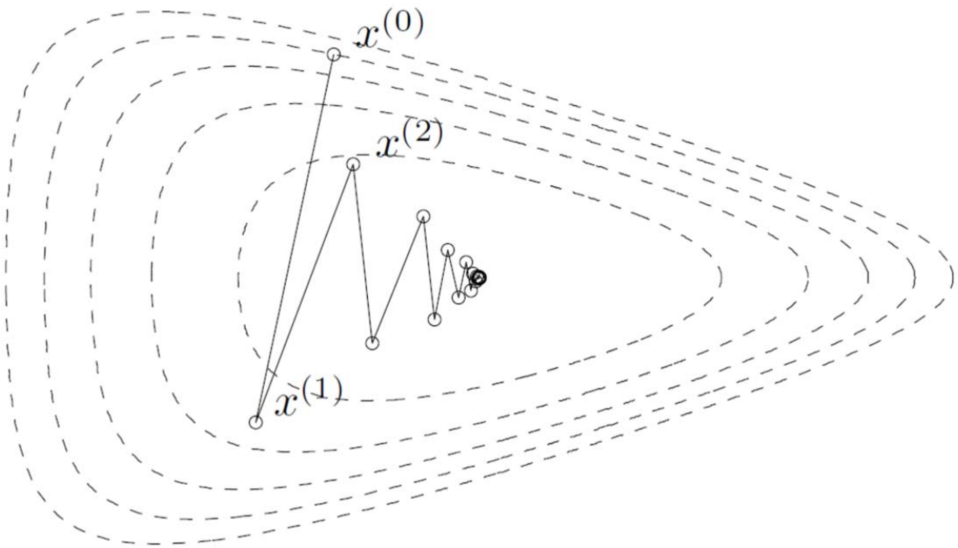
Figure 9.2 Some contour lines of the function $f(x) = (1/2)(x_1^2 + 10x_2^2)$. The condition number of the sublevel sets, which are ellipsoids, is exactly 10. The figure shows the iterates of the gradient method with exact line search, started at $x^{(0)} = (10, 1)$.

A Non-Quadratic Problem in $\mathbf{R}^2$

□ The Objective Function

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

- Gradient descent method with backtracking line search
  - ✓ $\alpha = 0.1, \beta = 0.7$

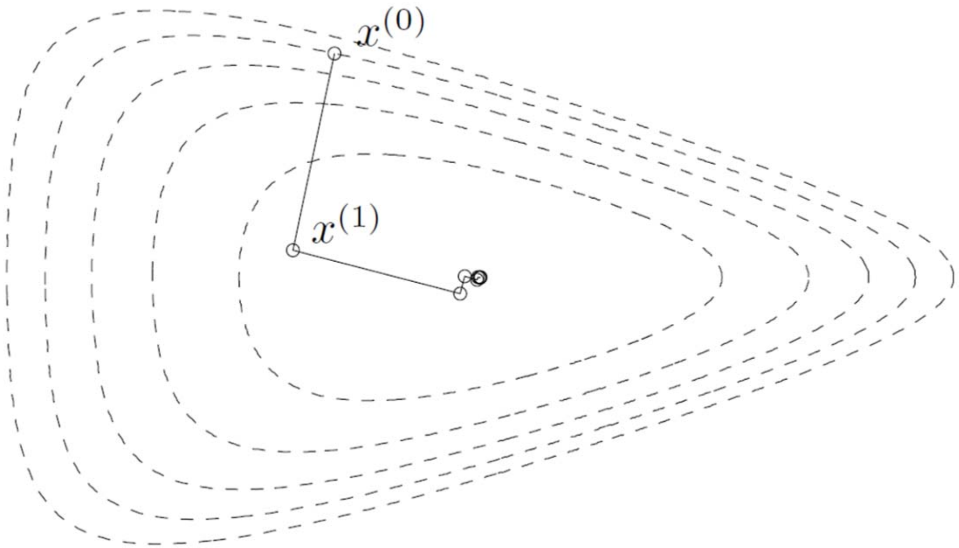


A Non-Quadratic Problem in $\mathbf{R}^2$

The Objective Function

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

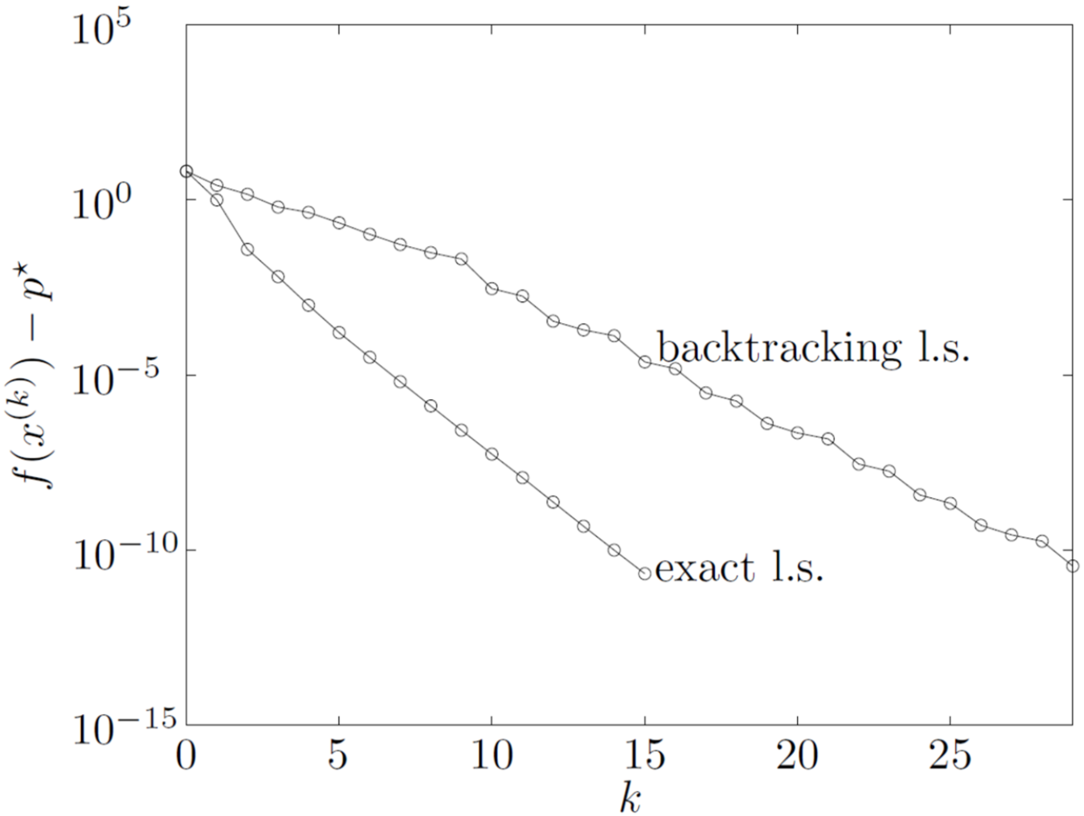
- Gradient descent method with exact line search



A Non-Quadratic Problem in $\mathbf{R}^2$

Comparisons

- Both are linear, and exact l.s. is faster



A Problem in $\mathbf{R}^{100}$

A Larger Problem

$$f(x) = c^\top x - \sum_{i=1}^{m} \log(b_i - a_i^\top x)$$

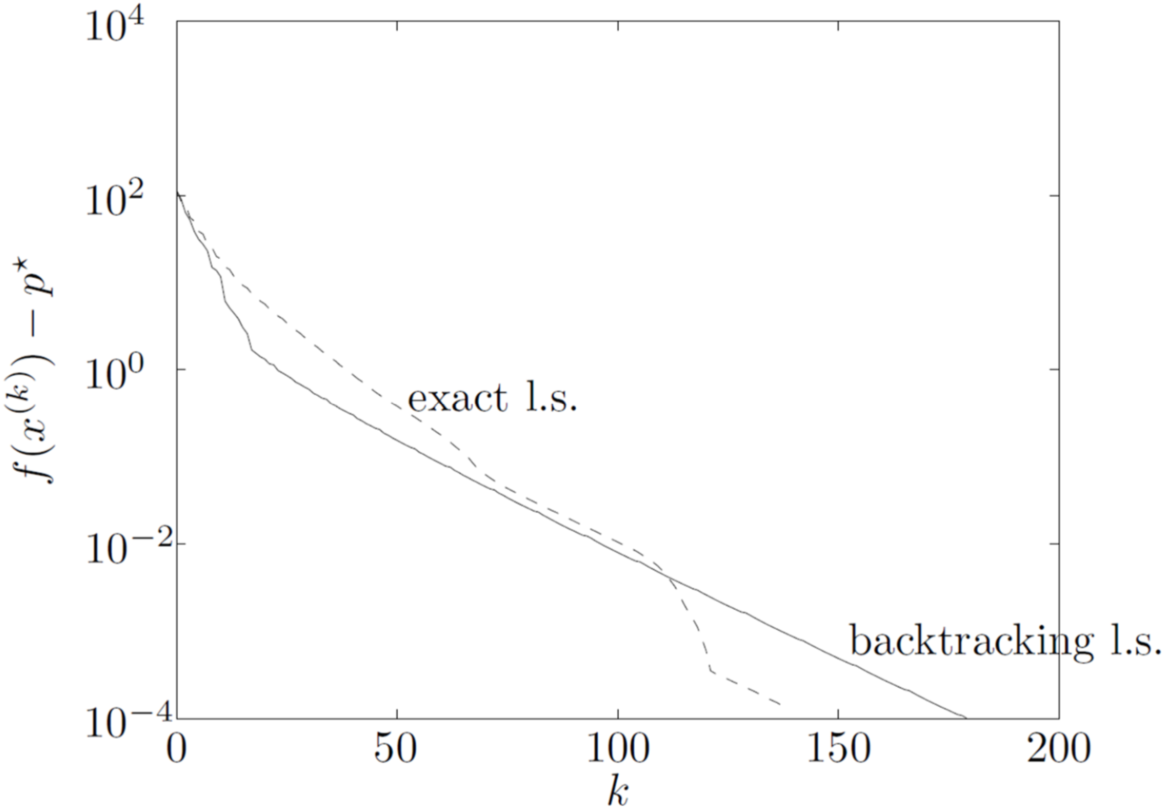
- $m = 500$ and $n = 100$

- Gradient descent method with backtracking line search
  - ✓ $\alpha = 0.1, \beta = 0.5$
- Gradient descent method with exact line search

A Problem in $\mathbf{R}^{100}$

Comparisons

- Both are linear, and exact l.s. is only a bit faster



Gradient Method and Condition Number

☐ A Larger Problem

$$f(x) = c^\top x - \sum_{i=1}^{m} \log(b_i - \alpha_i^\top x)$$

- Replace x by $T\bar{x}$

$$T = \mathrm{diag}\left(1, \gamma^{1/n}, \gamma^{2/n}, \dots, \gamma^{(n-1)/n}\right)$$

☐ A Family of Optimization Problems

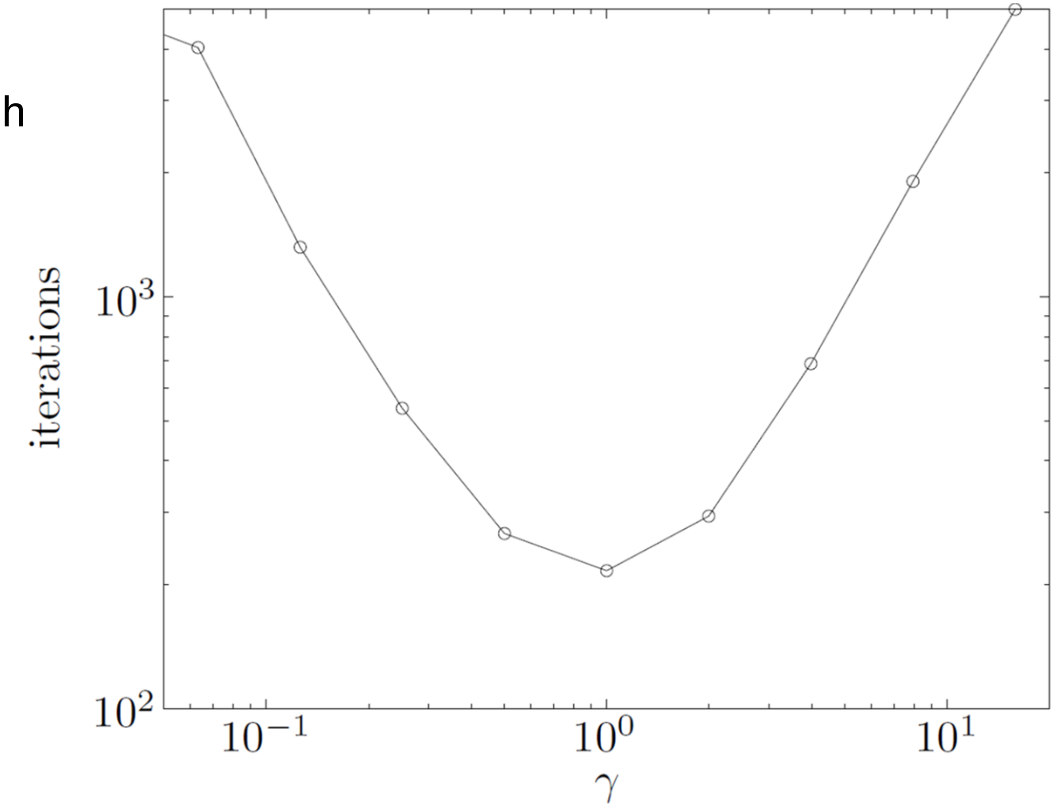
$$\bar{f}(\bar{x}) = c^\top T\bar{x} - \sum_{i=1}^{m} \log(b_i - \alpha_i^\top T\bar{x})$$

- Indexed by γ

Gradient Method and Condition Number

- Number of iterations required to obtain $\bar{f}(\bar{x}^k) - \bar{p}^* < 10^{-5}$

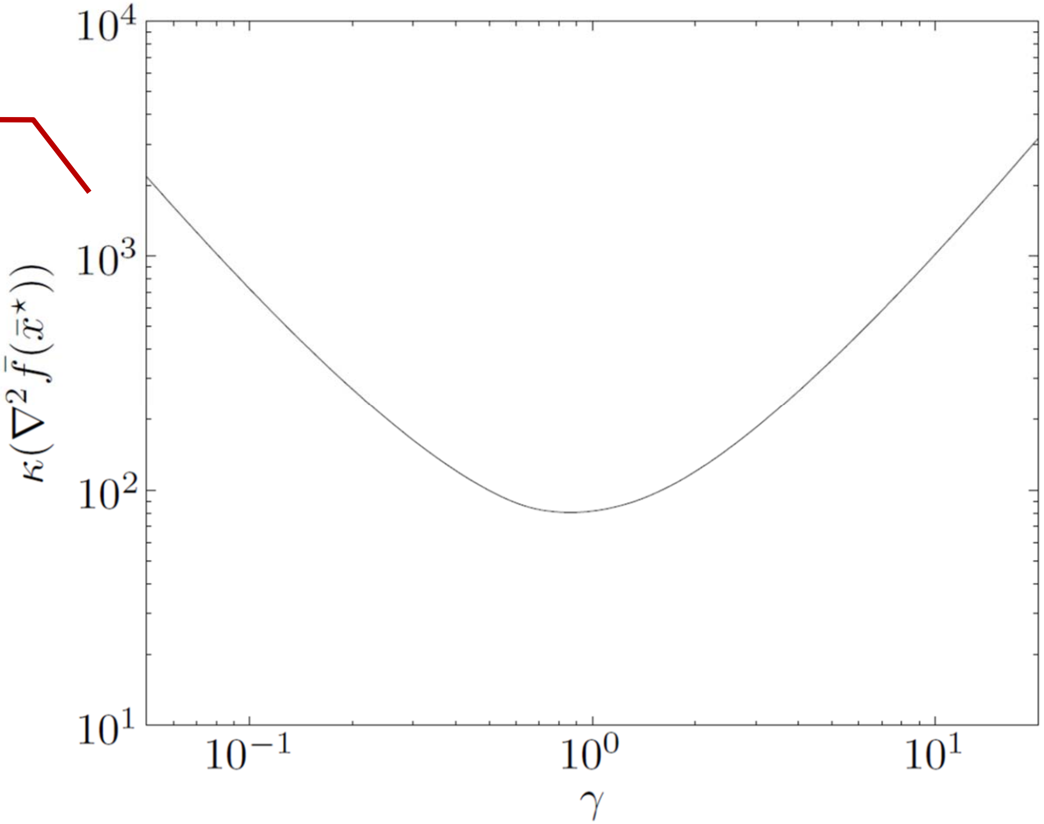
Backtracking line search with $\alpha = 0.3$ and $\beta = 0.7$



Gradient Method and Condition Number

- The condition number of the Hessian $\nabla^2 \bar{f}(\bar{x}^*)$at the optimum

The larger the condition number, the larger the number of iterations



Conclusions

1. The gradient method often exhibits approximately linear convergence.
2. The convergence rate depends greatly on the condition number of the Hessian, or the sublevel sets.
3. An exact line search sometimes improves the convergence of the gradient method, but the effect is not large.
4. The choice of backtracking parameters α, β has a noticeable but not dramatic effect on the convergence.

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

General Convex Functions

- $f(\cdot)$ is convex
- $f(\cdot)$ is Lipschitz continuous

$$\|\nabla f(x)\|_2 \leq G$$

- Gradient Descent Method

Given a starting point $x^{(1)} \in \operatorname{dom} f$

For $k = 1,2,\dots,K$ **do**

Update: $x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f(x^{(k)})$

End for

Return $\bar{x} = \frac{1}{K}\sum_{k=1}^{K} x^{(k)}$

Analysis

- Define $D = \left\|x^{(1)} - x^*\right\|_2$
- Let $t^{(k)} = \eta, k = 1, \ldots, K$

$$f\left(x^{(k)}\right) - f(x^*)$$

$$\leq \nabla f\left(x^{(k)}\right)^\top \left(x^{(k)} - x^*\right)$$

$$= \frac{1}{\eta}\left(x^{(k)} - x^{(k+1)}\right)^\top \left(x^{(k)} - x^*\right)$$

$$= \frac{1}{2\eta}\left(\left\|x^{(k)} - x^*\right\|_2^2 - \left\|x^{(k+1)} - x^*\right\|_2^2 + \left\|x^{(k)} - x^{(k+1)}\right\|_2^2\right)$$

Analysis

- Define $D = \left\|x^{(1)} - x^*\right\|_2$
- Let $t^{(k)} = \eta, k = 1, \dots, K$

$$f\left(x^{(k)}\right) - f(x^*)$$

$$\leq \nabla f\left(x^{(k)}\right)^\top \left(x^{(k)} - x^*\right)$$

$$= \frac{1}{\eta}\left(x^{(k)} - x^{(k+1)}\right)^\top \left(x^{(k)} - x^*\right)$$

$$= \frac{1}{2\eta}\left(\left\|x^{(k)} - x^*\right\|_2^2 - \left\|x^{(k+1)} - x^*\right\|_2^2\right) + \frac{\eta}{2}\left\|\nabla f\left(x^{(k)}\right)\right\|_2^2$$

$$\leq \frac{1}{2\eta}\left(\left\|x^{(k)} - x^*\right\|_2^2 - \left\|x^{(k+1)} - x^*\right\|_2^2\right) + \frac{\eta}{2} G^2$$

Analysis

- So,

$$f\left(x^{(k)}\right) - f(x^*) \le \frac{1}{2\eta}\left(\left\|x^{(k)} - x^*\right\|_2^2 - \left\|x^{(k+1)} - x^*\right\|_2^2\right) + \frac{\eta}{2}G^2$$

- Summing over $k = 1, \ldots, K$

$$\sum_{k=1}^{K} f\left(x^{(k)}\right) - Kf(x^*) \le \frac{1}{2\eta}D^2 + \frac{\eta K}{2}G^2$$

  - Dividing both sides by K

$$\frac{1}{K}\sum_{k=1}^{K} f\left(x^{(k)}\right) - f(x^*) \le \frac{1}{K}\left(\frac{1}{2\eta}D^2 + \frac{\eta K}{2}G^2\right)$$

$$= \frac{D^2}{2\eta K} + \frac{\eta}{2}G^2$$

Analysis

- By Jensen's Inequality

$$f(\bar{x}) - f(x^*) = f\left(\frac{1}{K}\sum_{k=1}^{K} x^{(k)}\right) - f(x^*)$$

$$\leq \frac{1}{K}\sum_{t=1}^{T} f\left(x^{(k)}\right) - f(x^*)$$

$$\leq \frac{D^2}{2\eta K} + \frac{\eta}{2}G^2$$

$$= \frac{GD}{\sqrt{K}}$$

- $\eta = \frac{D}{G\sqrt{K}}$

Discussions

- How to Ensure $\|\nabla f(x)\|_2 \le G$?
- Add a Domain Constraint

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in X \end{aligned}$$

  - Can model any constrained convex optimization problem
- Gradient Descent with Projection

$$\hat{x}^{(k+1)} = x^{(k)} - t^{(k)} \nabla f\left(x^{(k)}\right), \qquad x^{(k+1)} = P_X\left(\hat{x}^{(k+1)}\right)$$

  - Property of Euclidean Projection

$$\left\|x^{(k+1)} - x^*\right\|_2 = \left\|P_X\left(\hat{x}^{(k+1)}\right) - P_X(x^*)\right\|_2 \le \left\|\hat{x}^{(k+1)} - x^*\right\|_2$$

Gradient Descent with Projection

The Problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in X \end{aligned}$$

The Algorithm

Given a starting point $x^{(1)} \in \operatorname{dom} f$

For $k = 1,2, \dots, K$ **do**

Update: $\hat{x}^{(k+1)} = x^{(k)} - t^{(k)} \nabla f\left(x^{(k)}\right)$

Projection: $x^{(k+1)} = P_X(\hat{x}^{(k+1)})$

End for

Return $\bar{x} = \frac{1}{K} \sum_{k=1}^{K} x^{(k)}$

Assumptions $\|\nabla f(x)\|_2 \leq G, \qquad \forall x \in X$

Analysis

- Define $D = \|x^{(1)} - x^*\|_2, x^* = \mathrm{argmin}_{x \in X} f(x)$
- Let $t^{(k)} = \eta, k = 1, \dots, K$

$$f(x^{(k)}) - f(x^*)$$
$$\le \nabla f(x^{(k)})^\top (x^{(k)} - x^*)$$
$$= \frac{1}{\eta}(x^{(k)} - \hat{x}^{(k+1)})^\top (x^{(k)} - x^*)$$
$$\le \frac{1}{2\eta}\left(\|x^{(k)} - x^*\|_2^2 - \|\hat{x}^{(k+1)} - x^*\|_2^2\right) + \frac{\eta}{2}G^2$$
$$\le \frac{1}{2\eta}\left(\|x^{(k)} - x^*\|_2^2 - \|x^{(k+1)} - x^*\|_2^2\right) + \frac{\eta}{2}G^2$$

Property of Euclidean Projection

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

Motivation

- The First-order Taylor Approximation

$$f(x+v) \approx \hat{f}(x+v) = f(x) + \nabla f(x)^\top v$$

  - $\nabla f(x)^\top v$ is the directional derivative of f at x in the direction v
  - It gives the approximate change in f for a small step v
  - v is a descent direction if $\nabla f(x)^\top v$ is negative

- A Good Search Direction v
  - Make $\nabla f(x)^\top v$ as negative as possible

Steepest Descent Method

- Normalized Steepest Descent Direction
  - with respect to the norm $\|\cdot\|$

$$\Delta x_{\text{nsd}} = \text{argmin}\{\nabla f(x)^{\top} v \mid \|v\| = 1\}$$

  - Equivalent to

$$\Delta x_{\text{nsd}} = \text{argmin}\{\nabla f(x)^{\top} v \mid \|v\| \leq 1\}$$

    - ✓ The direction in the unit ball of $\|\cdot\|$ that extends farthest in the direction $-\nabla f(x)$

- Unnormalized Steepest Descent Direction $\Delta x_{\text{sd}} = \|\nabla f(x)\|_* \Delta x_{\text{nsd}}$

$$\nabla f(x)^{\top} \Delta x_{\text{sd}} = \|\nabla f(x)\|_* \nabla f(x)^{\top} \Delta x_{\text{nsd}} = -\|\nabla f(x)\|_*^2$$

Steepest Descent Method

The Algorithm

Given a starting point $x \in \mathrm{dom}\, f$

Repeat

1. Compute steepest descent direction Δx_{sd}.
2. Line search: Choose t via exact or backtracking line search.
3. Update: $x \coloneqq x + t\Delta x_{\mathrm{sd}}$.

until stopping criterion is satisfied.

- When exact line search is used, scale factors in the direction have no effect.

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

Steepest Descent Method

Steepest Descent for Euclidean Norm

$$\Delta x_{\mathrm{nsd}} = \operatorname{argmin}\{\nabla f(x)^\top v \,\big|\, \|v\|_2 \le 1\}$$
$$= -\frac{1}{\|\nabla f(x)\|_2}\nabla f(x)$$

$$\Delta x_{\mathrm{sd}} = \|\nabla f(x)\|_2 \Delta x_{\mathrm{nsd}} = -\nabla f(x)$$

- The steepest descent method coincides with the gradient descent method

Steepest Descent Method

Steepest Descent for Quadratic Norm

- P-quadratic norm, where $P \in \mathbf{S}_{++}^n$

$$\|z\|_P = (z^\top P z)^{1/2} = \left\|P^{1/2} z\right\|_2$$

- The dual norm $\|z\|_* = \|z\|_{P^{-1}} = \left\|P^{-1/2} z\right\|_2$
- Normalized Steepest Descent Direction

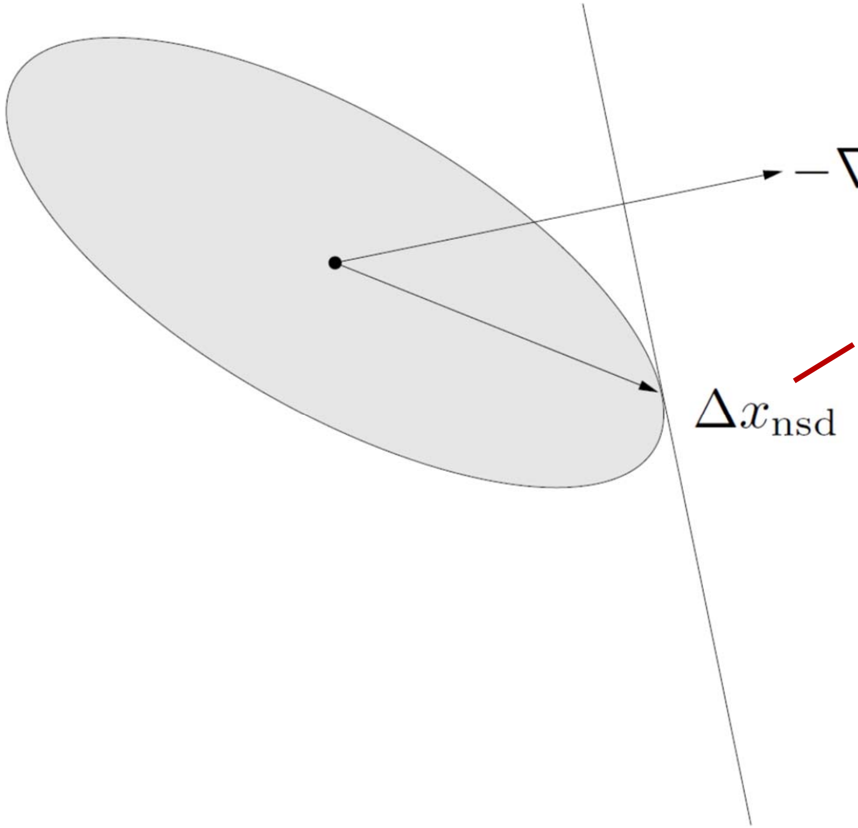
$$\Delta x_{\mathrm{nsd}} = -\left(\nabla f(x)^\top P^{-1} \nabla f(x)\right)^{-1/2} P^{-1} \nabla f(x)$$

- Unnormalized Steepest Descent Direction

$$\Delta x_{\mathrm{sd}} = \|\nabla f(x)\|_* \Delta x_{\mathrm{nsd}} = -P^{-1} \nabla f(x)$$

Steepest Descent Method

Steepest Descent for Quadratic Norm



$\Delta x_{\rm nsd}$ extends as far as possible in the direction $-\nabla f(x)$ while staying in the ellipsoid.

- The ellipsoid is the unit ball of the norm

Steepest Descent Method

Steepest Descent for Quadratic Norm

- Interpretation via Change of Coordinates
- Define $\bar{x} = P^{1/2}x$, so $\|x\|_P = \|\bar{x}\|_2$
- An Equivalent Problem

$$\min\ \bar{f}(\bar{x}) = f\left(P^{-1/2}\bar{x}\right) = f(x)$$

✓ Gradient descent method

$$\Delta\bar{x} = -\nabla\bar{f}(\bar{x}) = -P^{-1/2}\nabla f\left(P^{-1/2}\bar{x}\right) = -P^{-1/2}\nabla f(x)$$

✓ Correspond to the direction

$$\Delta x = P^{-1/2}(-P^{-1/2}\nabla f(x)) = -P^{-1}\nabla f(x)$$

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

Steepest Descent Method

Steepest Descent for ℓ_1-norm

- Normalized Steepest Descent Direction

$$\Delta x_{\text{nsd}} = \operatorname{argmin}\{\nabla f(x)^\top v \mid \|v\|_1 \leq 1\}$$
$$= -\operatorname{sign}\left(\frac{\partial f(x)}{\partial x_i}\right) e_i$$

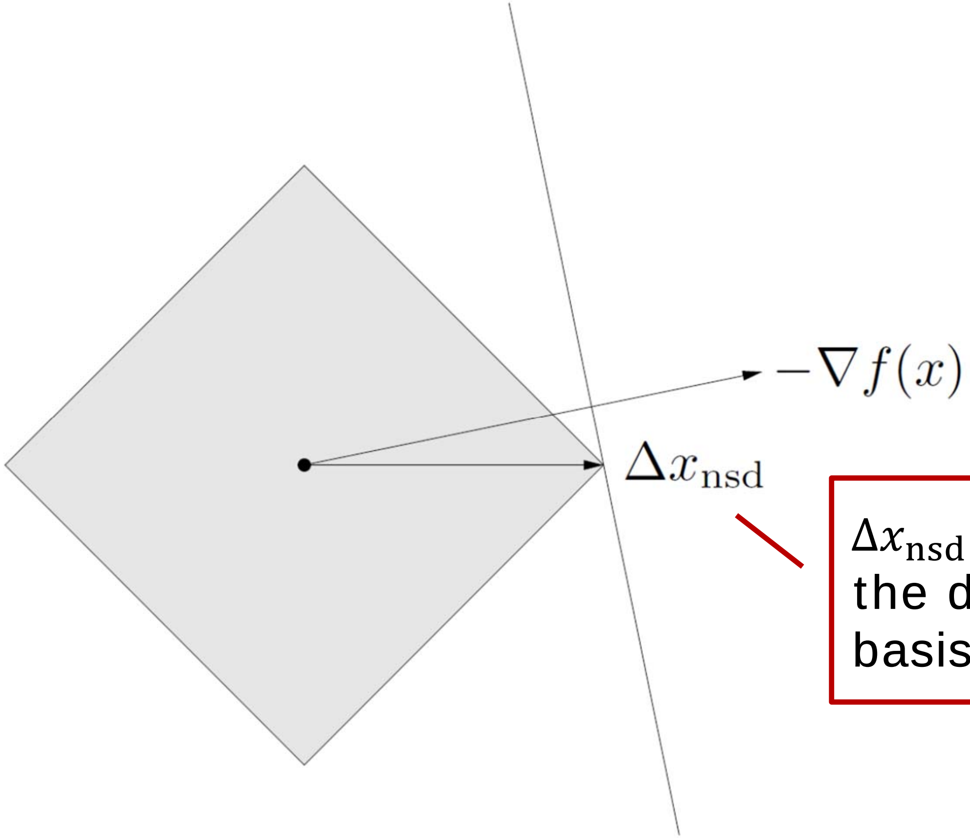
  - ✓ i be any index for which $\|\nabla f(x)\|_\infty = |(\nabla f(x))_i|$
  - ✓ e_i is the i-th standard basis vector

- Unnormalized Steepest Descent Direction

$$\Delta x_{\text{sd}} = \Delta x_{\text{nsd}} \|\nabla f(x)\|_\infty = -\frac{\partial f(x)}{\partial x_i} e_i$$

Steepest Descent Method

□ Steepest Descent for ℓ_1-norm



Δx_{nsd} can always be chosen in the direction of a standard basis vector (or a negative one).

- The diamond is the unit ball of ℓ_1-norm

Steepest Descent Method

- Steepest Descent for ℓ_1-norm

- Coordinate-descent Algorithm
  1. Select a component of $\nabla f(x)$ with maximum absolute value
  2. Decrease or increase the corresponding component of x

- Simplify, or even trivialize, the line search

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

Convergence Analysis

1. Any norm can be bounded in terms of the Euclidean norm
   - Exist $\gamma, \tilde{\gamma} \in (0,1]$

$$\|x\| \geq \gamma \|x\|_2, \qquad \|x\|_* \geq \tilde{\gamma} \|x\|_2$$

2. $f(\cdot)$ is smooth, i.e, $\nabla^2 f(x) \preccurlyeq MI, \forall x \in S$

$$f(x + t\Delta x_{\mathrm{sd}}) \leq f(x) + t\nabla f(x)^\top \Delta x_{\mathrm{sd}} + \frac{M\|\Delta x_{\mathrm{sd}}\|_2^2}{2} t^2$$

$$\leq f(x) + t\nabla f(x)^\top \Delta x_{\mathrm{sd}} + \frac{M\|\Delta x_{\mathrm{sd}}\|^2}{2\gamma^2} t^2$$

$$= f(x) - t\|f(x)\|_*^2 + \frac{M}{2\gamma^2} t^2 \|f(x)\|_*^2$$

Convergence Analysis

3. Exit Condition for the Backtracking Line Search

$$f(x + t\Delta x_{\mathrm{sd}}) \le f(x) + \alpha t \nabla f(x)^{\top} \Delta x_{\mathrm{sd}}, \qquad \forall t \le \gamma^2/M$$

- $a < 1/2$

$$0 \le t \le \frac{\gamma^2}{M} \Rightarrow -t + \frac{Mt^2}{2\gamma^2} \le -\frac{t}{2}$$

$$f(x + t\Delta x_{\mathrm{sd}}) \le f(x) - t\|f(x)\|_*^2 + \frac{M}{2\gamma^2} t^2 \|f(x)\|_*^2$$

$$\Rightarrow f(x + t\Delta x_{\mathrm{sd}}) \le f(x) - \frac{t}{2}\|f(x)\|_*^2$$

$$\Rightarrow f(x + t\Delta x_{\mathrm{sd}}) \le f(x) + \frac{t}{2}\nabla f(x)^{\top}\Delta x_{\mathrm{sd}}$$

Convergence Analysis

3. Exit Condition for the Backtracking Line Search

$$f(x + t\Delta x_{\mathrm{sd}}) \leq f(x) + \alpha t \nabla f(x)^{\top} \Delta x_{\mathrm{sd}}, \qquad \forall t \leq \gamma^2/M$$

- $\alpha < 1/2$
- Backtracking line search terminates

$$t \geq \min\{1, \beta\gamma^2/M\}$$

- So

$$f(x^+) = f(x + t\Delta x_{\mathrm{sd}}) \leq f(x) - \alpha \min\left\{1, \frac{\beta\gamma^2}{M}\right\} \|f(x)\|_*^2$$

$$\leq f(x) - \alpha\, \tilde{\gamma}^2 \min\left\{1, \frac{\beta\gamma^2}{M}\right\} \|f(x)\|_2^2$$

Convergence Analysis

Fail to illustrate the advantage

4. Subtracting p^* from Both Sides

$$f(x^+) - p^* \le f(x) - p^* - \alpha\,\tilde{\gamma}^2 \min\left\{1, \frac{\beta\gamma^2}{M}\right\} \|f(x)\|_2^2$$

5. Combining with Strong Convexity

$$f(x^+) - p^* \le c(f(x) - p^*)$$

- $c = 1 - 2m\alpha\tilde{\gamma}^2 \min\{1, \beta\gamma^2/M\} < 1$

6. Applying it Recursively

$$f\left(x^{(k)}\right) - p^* \le c^k\left(f\left(x^{(0)}\right) - p^*\right)$$

- Linear convergence

Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
  - General Convex Functions
- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis
  - Discussion and Examples

Choice of Norm for Steepest Descent

- Steepest Descent Method with Quadratic P-norm
  - Equivalent to gradient method after the change of coordinates
- Gradient Method Works Well
  - When the condition numbers of the sublevel sets (or Hessian) are moderate
- Steepest Descent Method will Work Well
  - When the sublevel sets, after the change of coordinates, are moderately conditioned

Choice of Norm for Steepest Descent

- Choosing P to make the sublevel sets of $\bar{f}$ are well conditioned
  - If an approximation $\widehat{H}$ of the Hessian at the optimal point $H(x^*)$ were known
  - A good choice of P would be $P = \widehat{H}$
  - The Hessian of $\bar{f}$ at the optimum

$$\widehat{H}^{-1/2}\nabla^2 f(x^*)\widehat{H}^{-1/2} \approx I$$

- Choosing P to make the ellipsoid

$$\mathcal{E} = \{x | x^\top P x \leq 1\}$$

approximate the the sublevel set of f

Example

The Objective Function

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

- Steepest descent method
  - ✓ Using the two quadratic norms

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

- Backtracking line search
  - ✓ $\alpha = 0.1$ and $\beta = 0.7$

Example

□ The Objective Function

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

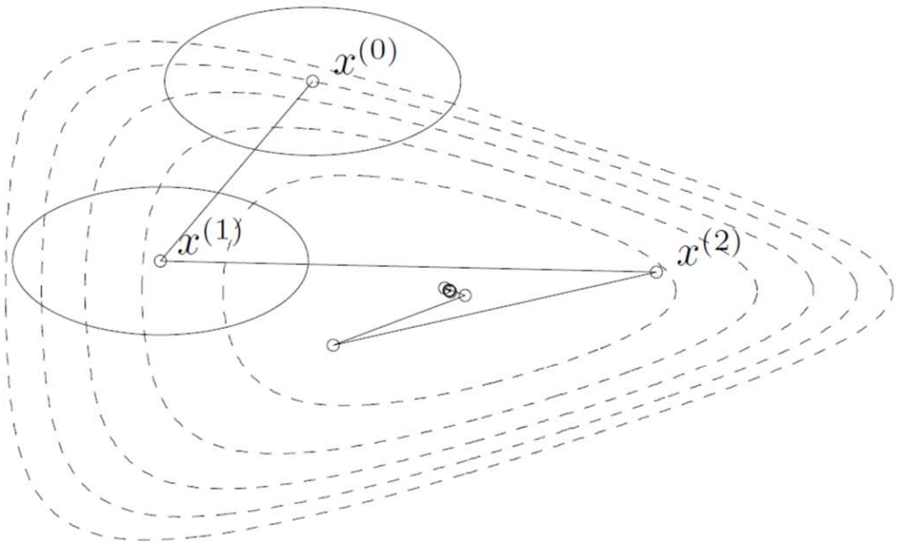


Figure 9.11 Steepest descent method with a quadratic norm $\|\cdot\|_{P_1}$. The ellipses are the boundaries of the norm balls $\{x \mid \|x - x^{(k)}\|_{P_1} \leq 1\}$ at $x^{(0)}$ and $x^{(1)}$.

Example

☐ The Objective Function

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

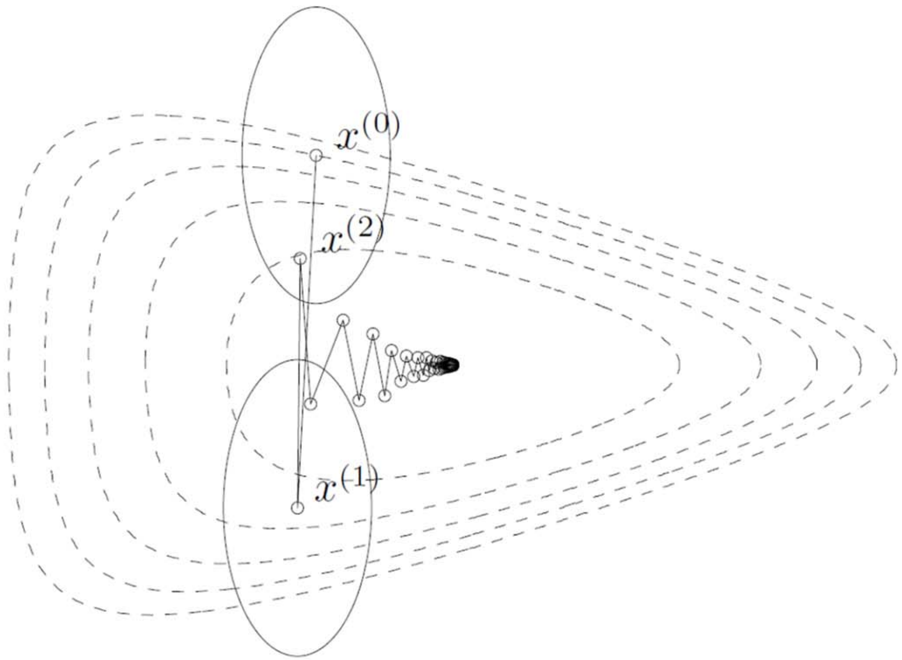


Figure 9.12 Steepest descent method, with quadratic norm $\|\cdot\|_{P_2}$.

Example

□ The Objective Function

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

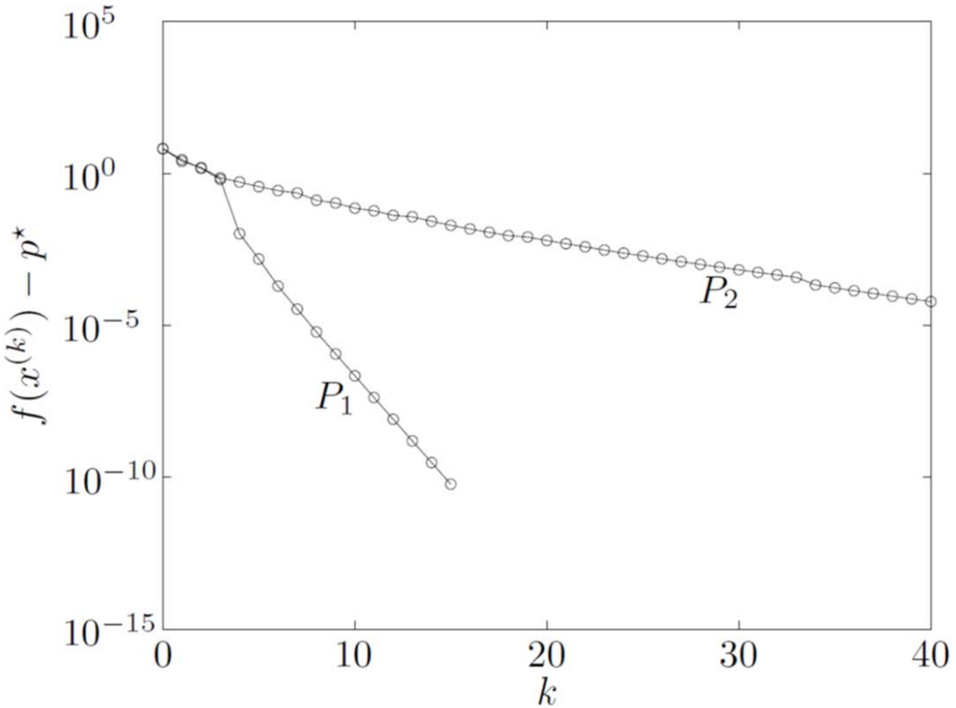
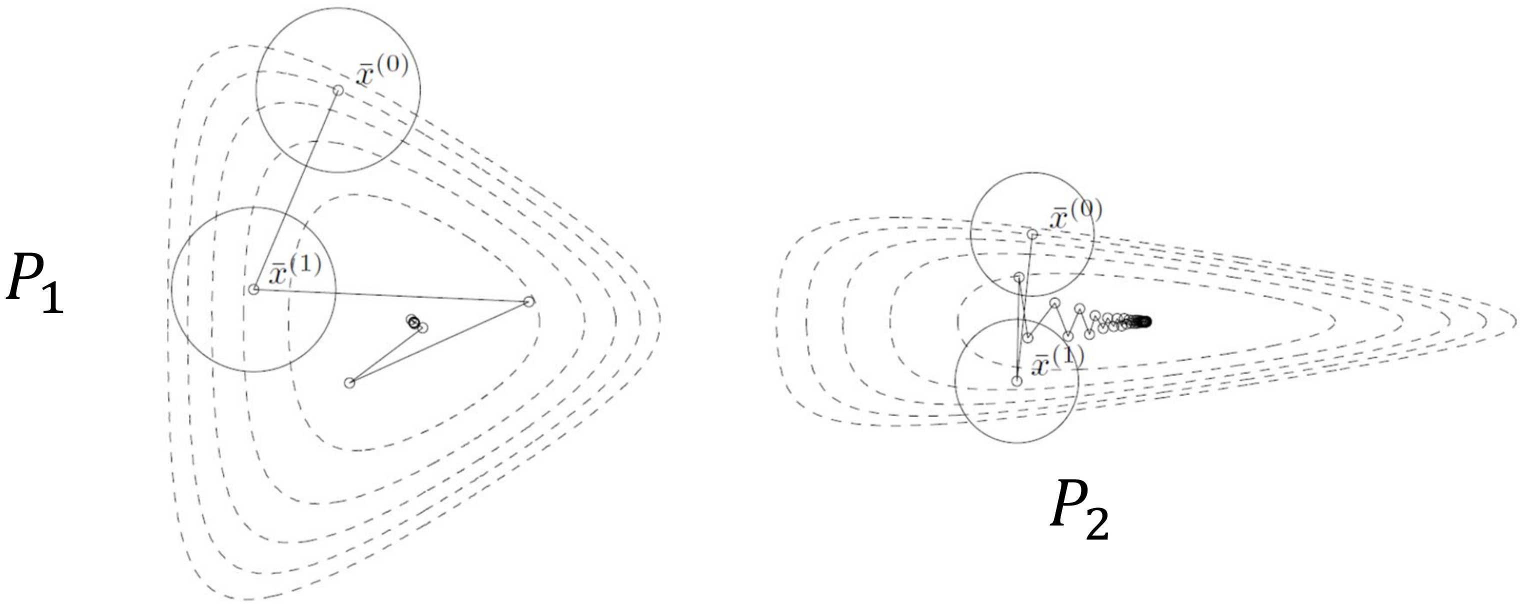


Figure 9.13 Error $f(x^{(k)}) - p^\star$ versus iteration k, for the steepest descent method with the quadratic norm $\|\cdot\|_{P_1}$ and the quadratic norm $\|\cdot\|_{P_2}$. Convergence is rapid for the norm $\|\cdot\|_{P_1}$ and very slow for $\|\cdot\|_{P_2}$.

Example

- Why P_1 is better than P_2?
  - Problems after the changes of coordinates



✓ The change of variables associated with P_1 yields sublevel sets with modest condition number

Summary

- Gradient Descent Method
  - Convergence Analysis
  - General Convex Functions

- Steepest Descent Method
  - Euclidean and Quadratic Norms
  - ℓ_1-norm
  - Convergence Analysis