# 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
  - Convergence Analysis
  - Extensions

# Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
- General Convex Functions
  - Convergence Analysis
  - Extensions

# General Descent Method

## The Algorithm

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

**Repeat**

1. Determine a descent direction $\Delta x$.
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
  - Convergence Analysis
  - Extensions

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

# 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 $t_{\text{exact}}$ 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

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

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

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

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

# Discussions

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

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

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

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

- $\log\left((f(x^{(0)}) - p^*)/\epsilon\right)$ 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 $\Delta x$ 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 $\Delta x$ for $f$ at $x \in \textbf{dom}\, f$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$

$t := 1$

**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 $t = 1$

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

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

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

- So,

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

## 3. Subtracting $p^*$ from Both Sides

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

# Analysis for Backtracking Line Search

## 4. Combining with Strong Convexity

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

## 5. Applying it Recursively

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

- $c = 1 - \min\left\{2m\alpha, \frac{2\beta\alpha m}{M}\right\} < 1$
- $f\left(x^{(k)}\right)$ 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
  - Convergence Analysis
  - Extensions

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

# 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 $\gamma$ not far from one, convergence is rapid

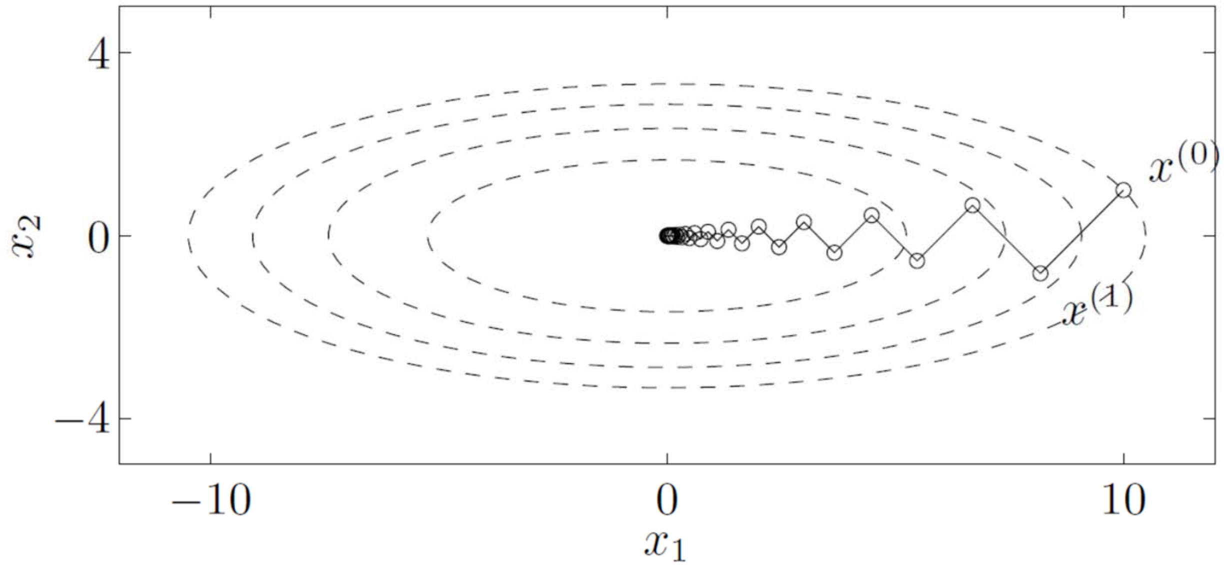


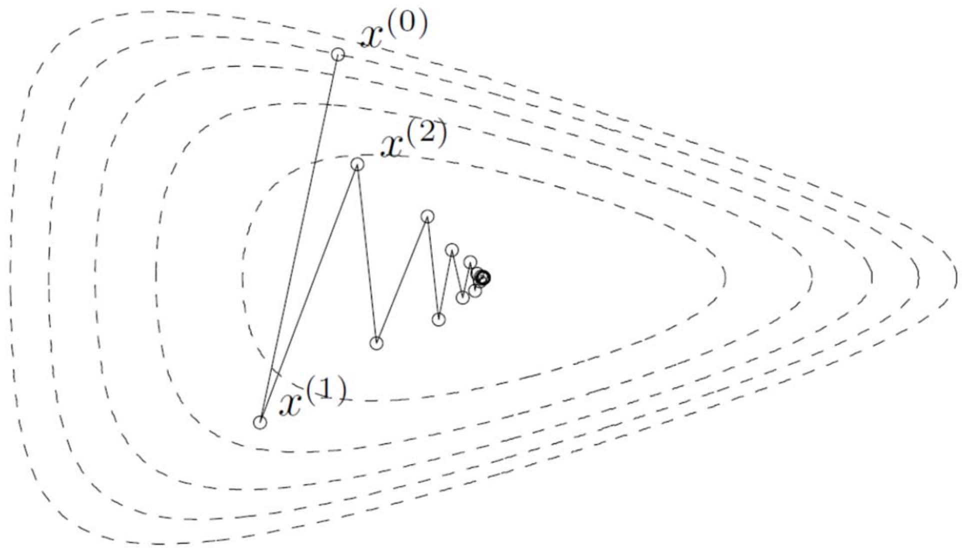

**Figure 9.2** Some contour lines of the function $f(x) = (1/2)(x_1^2 + 10x_2^2)$. The condition number of the sublevel sets, which are ellipsoids, is exactly 10. The figure shows the iterates of the gradient method with exact line search, started at $x^{(0)} = (10, 1)$.

# 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
  - ✓ $\alpha = 0.1, \beta = 0.7$

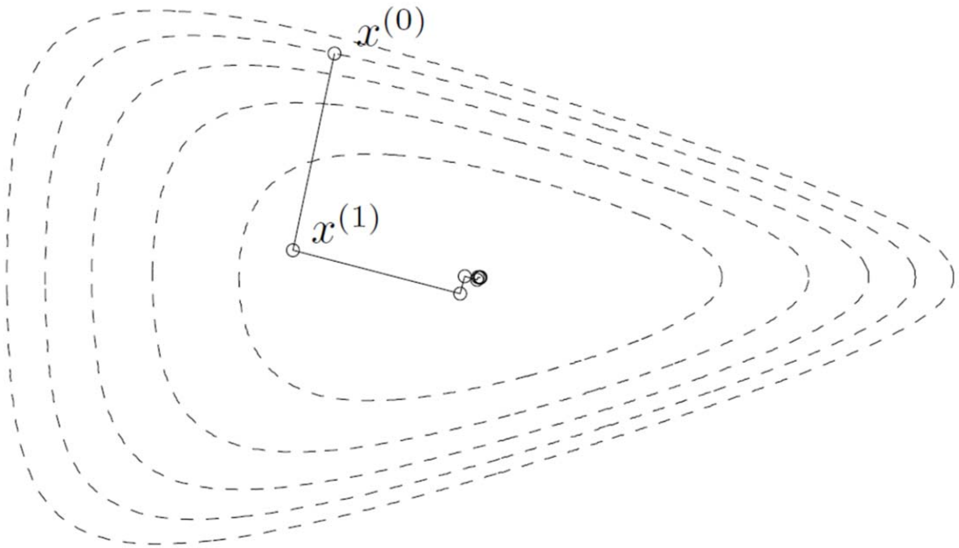


# A Non-Quadratic Problem in $\mathbf{R}^2$

## The Objective Function

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

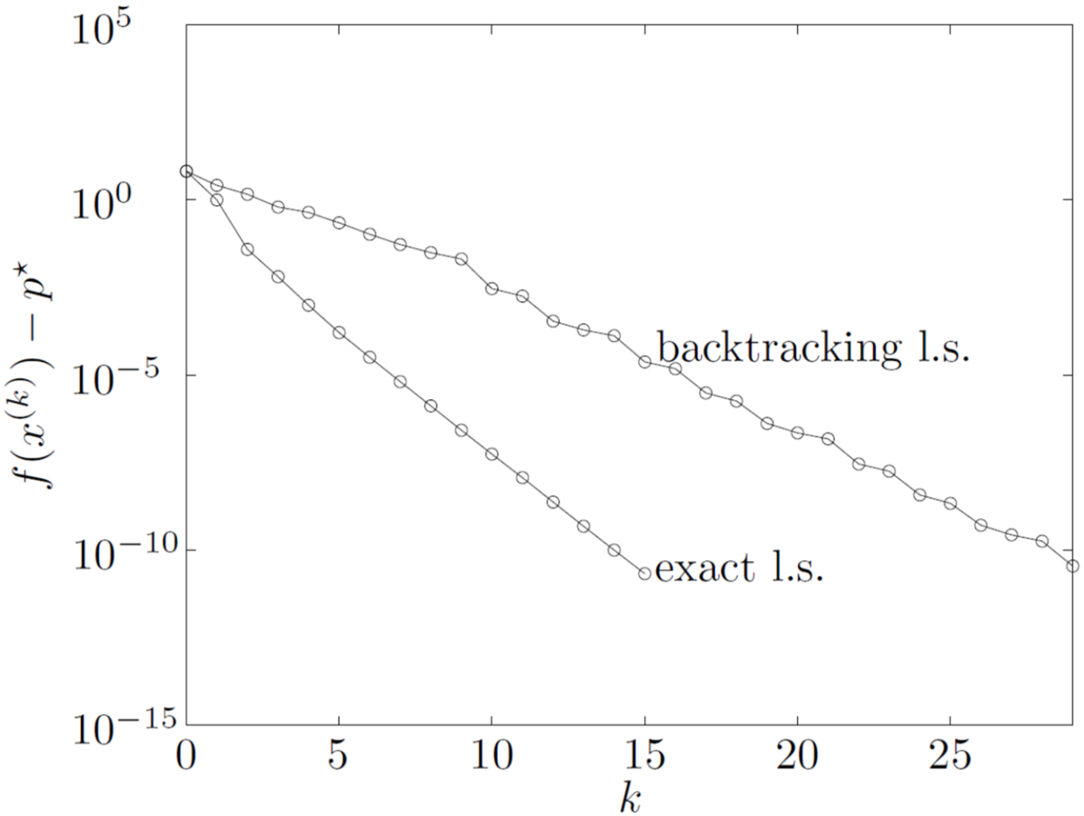
- Gradient descent method with exact line search



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

## ☐ Comparisons

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



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

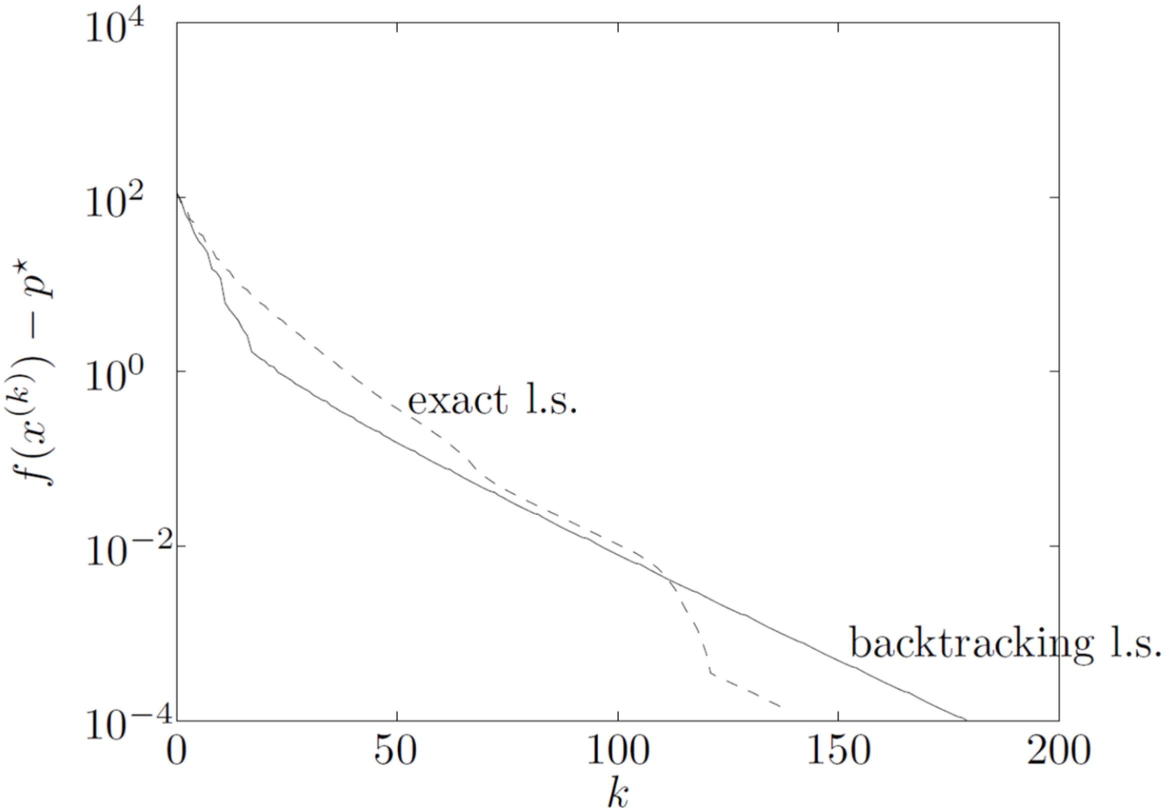
## A Larger Problem

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

- $m = 500$ 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 $T\bar{x}$

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

## ☐ A Family of Optimization Problems

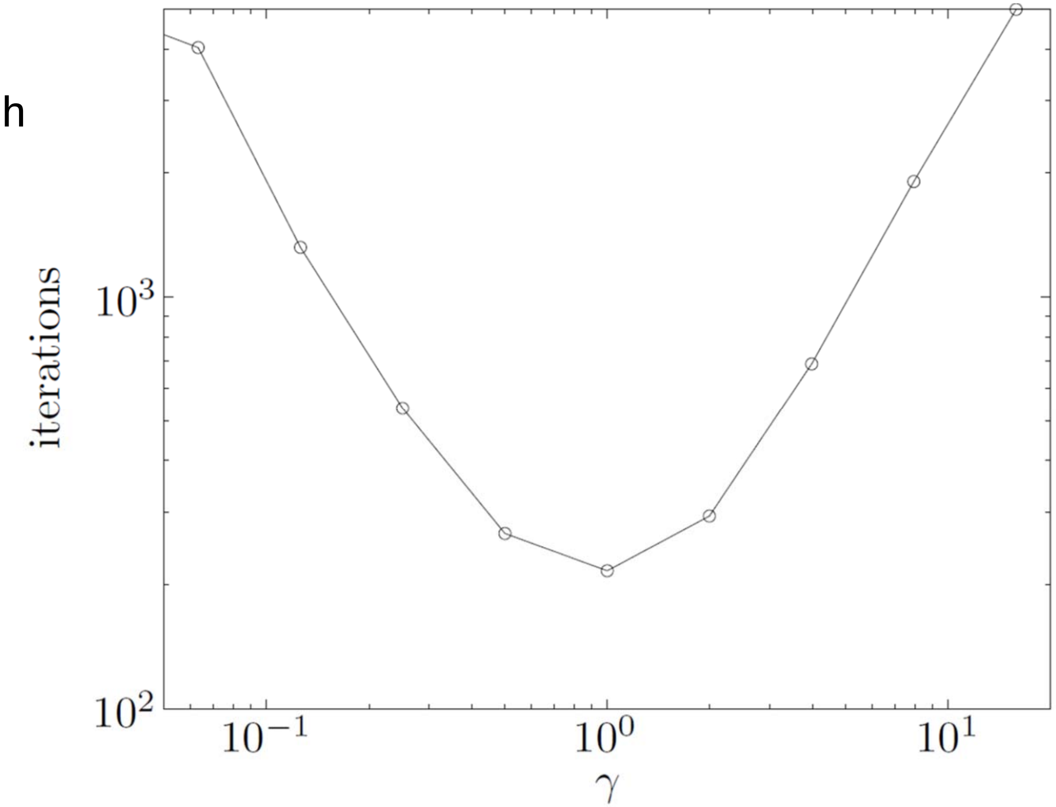
$$\bar{f}(\bar{x}) = c^\top T\bar{x} - \sum_{i=1}^{m} \log(b_i - \alpha_i^\top T\bar{x})$$

- Indexed by $\gamma$

# Gradient Method and Condition Number

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

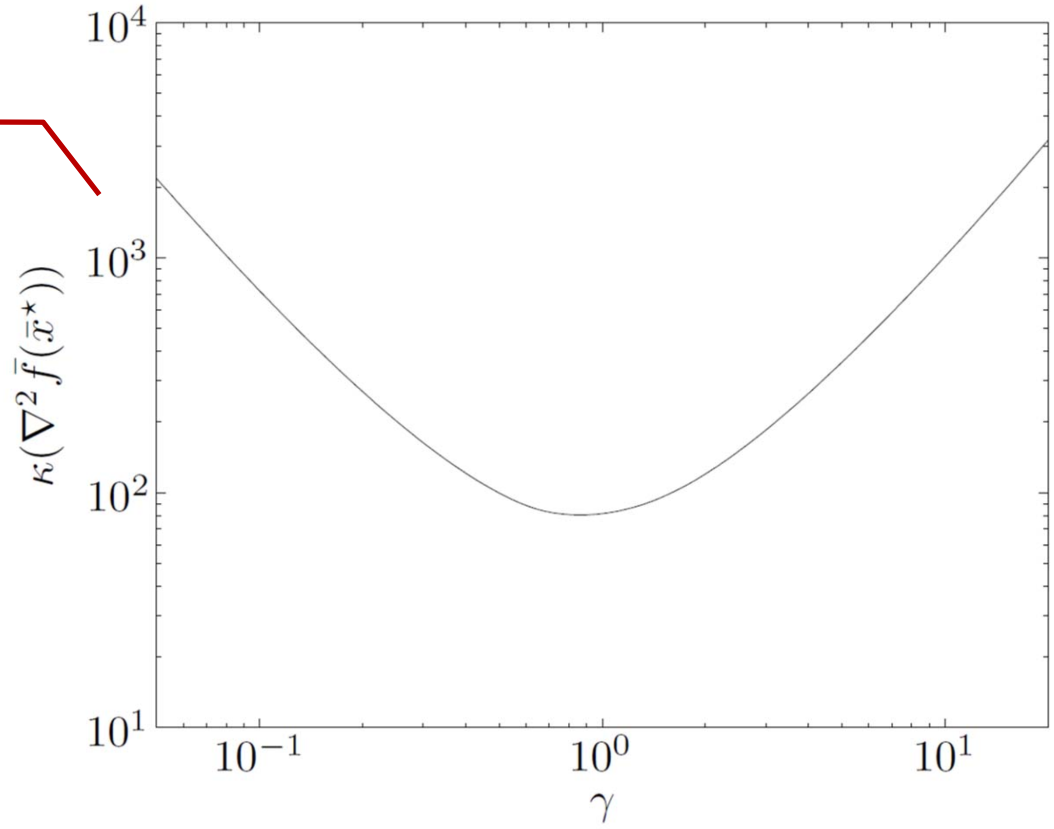
Backtracking line search with $\alpha = 0.3$ and $\beta = 0.7$



# 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 $\alpha, \beta$ has a noticeable but not dramatic effect on the convergence.

# Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
- General Convex Functions
  - Convergence Analysis
  - Extensions

# 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 $x^{(1)} \in \text{dom}\, f$

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

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

**End for**

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

# Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
- General Convex Functions
  - Convergence Analysis
  - Extensions

# Analysis

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

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

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

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

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

# Analysis

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

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

# Analysis

- So,

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

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

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

  - Dividing both sides by $K$

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

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

# Analysis

- By Jensen's Inequality

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

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

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

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

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

# Outline

- Gradient Descent Method
  - Convergence Analysis
  - Examples
- General Convex Functions
  - Convergence Analysis
  - Extensions

# Discussions

- How to Ensure $\|\nabla f(x)\|_2 \leq G$?
- 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

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

# Gradient Descent with Projection

- The Problem

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

- The Algorithm

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

**For** $k = 1,2,\ldots,K$ **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 $D = \left\|x^{(1)} - x^*\right\|_2, x^* \in X$
- Let $t^{(k)} = \eta, k = 1, \dots, K$

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

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

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

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

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

Property of Euclidean Projection

# Summary

- Gradient Descent Method
  - Convergence Analysis
  - Examples

- General Convex Functions
  - Convergence Analysis
  - Extensions