Linear Algebra & Convex Optimization Review

CS 534: Machine Learning

Slides adapted from Lee Cooper

Probability Review: Recap

Probability Theory

- Random variables
- Joint PDF, CDF
- Marginal & conditional distribution
- Expectation (mean and variance)
- Bayes rule
- Independence, covariance, correlation

Linear Algebra

Notation

- Vector: $\mathbf{x} \in \mathbb{R}^n$

Hastie et al. book notation

$$\mathbf{x} = X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- Matrix: $\mathbf{A} \in \mathbb{R}^{m \times n}$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} \mathbf{a_1} & \mathbf{a_2} & \cdots & \mathbf{a_n} \end{bmatrix}$$

Special Matrices

- Identity Matrix:

$$\mathbf{I} \in \mathbb{R}^{n \times n}, \text{ where } I_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$$

$$\mathbf{AI} = \mathbf{A} = \mathbf{IA}$$

- Diagonal Matrix:

$$\mathbf{D} = \text{diag}(d_1, d_2, \cdots, d_n) \text{ with } D_{ij} = \begin{cases} d_i, \ i = j \\ 0, \ i \neq j \end{cases}$$

Matrix Multiplication

$$\text{If } \mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times p},$$

$$\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times p}, \text{ where } C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

- Properties
  - Associative

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$$

  - Distributive

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$$

Generally not commutative so AB =/= BA

Transpose

- "Flip" rows and columns of a matrix

$$(A^\top)_{ij} = A_{ji}$$

- Properties
  - $(\mathbf{A}^\top)^\top = A$
  - $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$
  - $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

Trace

- Sum of the diagonal elements in a square matrix

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$$

- Properties
  - $\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathbf{A}^\top)$
  - $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n},\ \mathrm{Tr}(\mathbf{A} + \mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B})$
  - $\mathbf{A} \in \mathbb{R}^{n \times n},\ t \in \mathbb{R},\ \mathrm{Tr}(t\mathbf{A}) = t\mathrm{Tr}(\mathbf{A})$

    $\mathbf{AB} \in \mathbb{R}^{n \times n},\ \mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$

Norms

- Norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that satisfies 4 properties:
  - Non-negativity

$$\text{For all } \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) \geq 0$$

  - Definiteness

$$f(\mathbf{x}) = 0 \text{ if and only if } \mathbf{x} = 0$$

  - Homogeneity

$$\text{For all } \mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}, f(t\mathbf{x}) = |t| f(x)$$

  - Triangle Inequality

$$\text{For all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$$

Common Vector Norms

- Euclidean (ℓ_2) norm

$$||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

- ℓ_1 norm

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$$

- ℓ_∞ norm

$$||\mathbf{x}||_\infty = \max_{x_i} |x_i|$$

- ℓ_p norm

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Common Matrix Norms

- Frobenius norm

$$||\mathbf{A}||_F = \sqrt{\sum_{ij} |A_{ij}|^2} = \sqrt{\mathrm{Tr}(\mathbf{A}^\top \mathbf{A})}$$

- 1-norm

$$||\mathbf{A}||_1 = \max_j \sum_i |A_{ij}|$$

- 2-norm

$$||\mathbf{A}||_2 = \sqrt{\max \mathrm{eig}(\mathbf{A}^\top \mathbf{A})}$$

- p-norm

$$||\mathbf{A}||_p = (\max_{||\mathbf{x}||_p = 1} ||\mathbf{A}\mathbf{x}||_p)^{1/p}$$

Linear Independence

- Set of vectors are linearly independent if no vector can be represented as a linear combination of the remaining vectors

- Linearly dependent vector:

$$\mathbf{x}_n = \sum_{i=1}^{n-1} \alpha_i \mathbf{x}_i$$

Rank

- Column rank: size of largest subset of columns of A such that constitute a linearly dependent set
- Row rank: largest number of rows of A that constitute a linearly independent set
- For any matrix in real space, column rank = row rank

Rank Properties

$\mathbf{A} \in \mathbb{R}^{m \times n}$

- Rank vs dimension

$$\text{rank}(\mathbf{A}) \leq \min(m, n)$$

- Full rank

$$\text{rank}(\mathbf{A}) = \min(m, n)$$

- Rank of transpose

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^\top)$$

Rank Properties (2)

- Multiplication of two matrices

$$\mathrm{rank}(\mathbf{AB}) \leq \min(\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B}))$$

- Addition of two same sized matrices

$$\mathrm{rank}(\mathbf{A} + \mathbf{B}) \leq \mathrm{rank}(\mathbf{A}) + \mathrm{rank}(\mathbf{B})$$

Matrix Inverse

- Unique matrix such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$$

- A is invertible and non-singular if inverse exists
- A is singular if not invertible
- A must be full rank to have an inverse

Matrix Inverse Properties

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}^{-1})^{\top} = (\mathbf{A}^{\top})^{-1}$

Pseudo Inverse (Moore-Penrose)

- Generalization of inverse for non-square but full rank
- Criteria:
  - $\mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A}$
  - $\mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}$
  - $(\mathbf{A}\mathbf{A}^{\dagger})^{\top} = \mathbf{A}\mathbf{A}^{\dagger}$
  - $(\mathbf{A}^{\dagger}\mathbf{A})^{\top} = \mathbf{A}^{\dagger}\mathbf{A}$

Orthogonal Matrices

- Orthogonal vectors x, y:

$$\mathbf{x}^\top \mathbf{y} = 0$$

- Normalized vector:

$$||\mathbf{x}||_2 = 1$$

- Orthogonal square matrix if all columns are orthogonal to one another

- Orthonormal square matrix if orthogonal matrix and all columns are normalized

Orthogonal Properties

- Inverse of orthogonal matrix is its transpose

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^\top$$

- Vector operation will not change its Euclidean norm

$$||\mathbf{U}\mathbf{x}||_2 = ||\mathbf{x}||_2$$

Range and Nullspace

- Span of a set of vectors is all the vectors that can expressed as linear combination of these vectors

$$\text{span}(\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}) = \left\{ \mathbf{v} : \mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i \right\}$$

- Range (columnspace) is the span of the columns of the matrix

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{v} \in \mathbb{R}^m : \mathbf{v} = \mathbf{A}\mathbf{x}, \mathbf{x} \in \mathbb{R}^n\}$$

- Nullspace is the set of all vectors that equal 0 when multiple by matrix

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = 0\}$$

Fundamental Subspaces

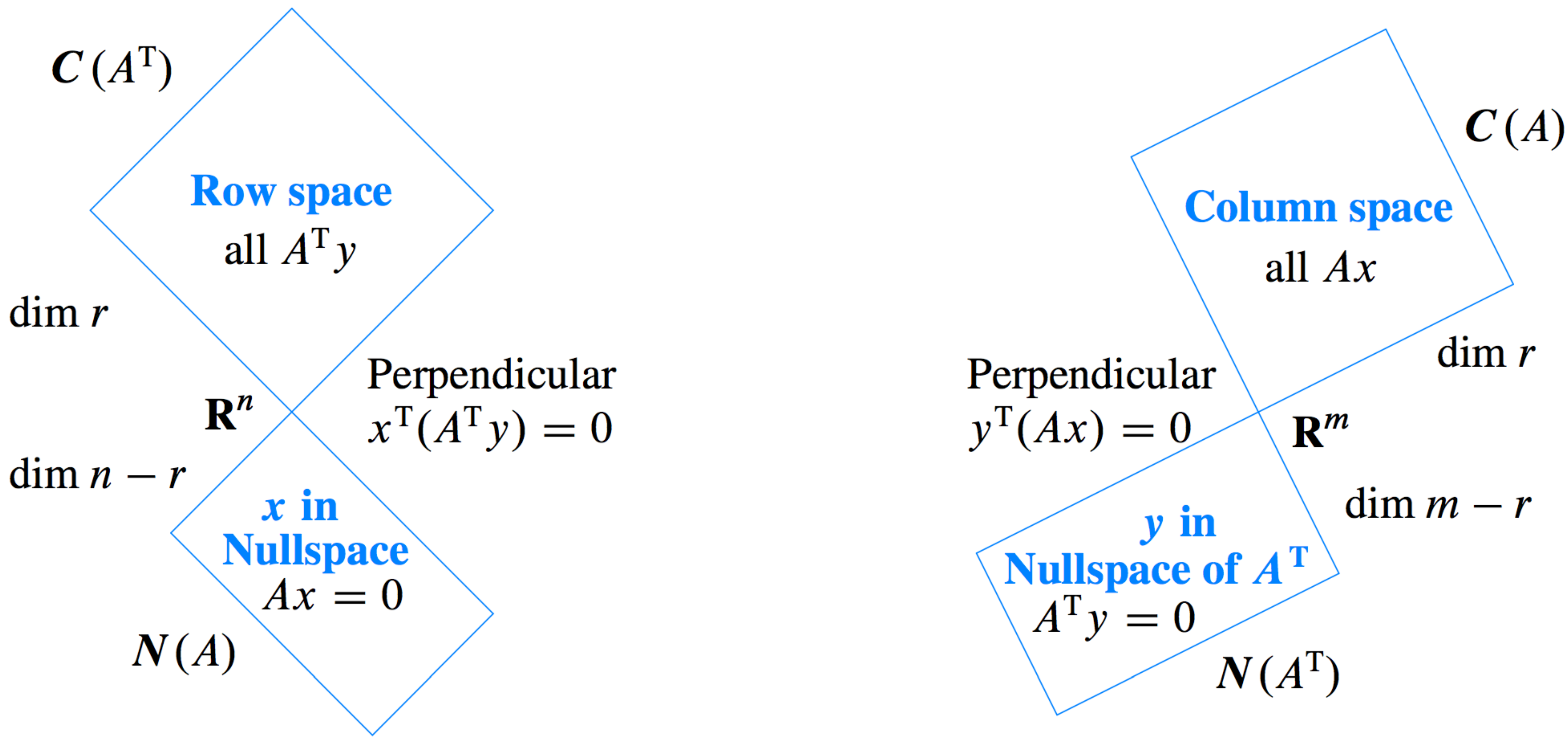


Figure 1: Dimensions and orthogonality for any m by n matrix A of rank r.

http://web.mit.edu/18.06/www/Essays/newpaper_ver3.pdf

Eigenvalues and Eigenvectors

- Instrumental to systems

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

- Analogy: Matrix is a gust of wind (invisible force with visible result)
  - Eigenvector is like a weathervane which tells you the direction the wind is blowing in
  - Eigenvalue is just the scalar coefficient

https://deeplearning4j.org/eigenvector

Eigenvalue Properties

- Trace of a matrix is sum of its eigenvalues

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$$

- Determinant of matrix is equal to product of its eigenvalues

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i$$

- Rank of matrix is the number of non-zero eigenvalues

- If eigenvectors of matrix are linearly independent, then the matrix is invertible

$$\mathbf{A} = \mathbf{X}\Lambda\mathbf{X}^{-1}$$

Symmetric Matrix & Eigenvectors

- Two remarkable properties from a symmetric matrix
  - Eigenvalues of the matrix are real
  - Eigenvectors of the matrix are orthonormal

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top$$

- Eigenvalues are positive —> positive definite
- Eigenvalues are non-negative —> positive semidefinite

Convex Optimization Review

Optimization Problem

- Minimize a function subject to some constraints

$$\begin{aligned} \min_x \;& f_0(x) \\ \text{s.t. } & f_k(x) \leq 0, k = 1, 2, \cdots, K \\ & h_j(x) = 0, j = 1, 2, \cdots, J \end{aligned}$$

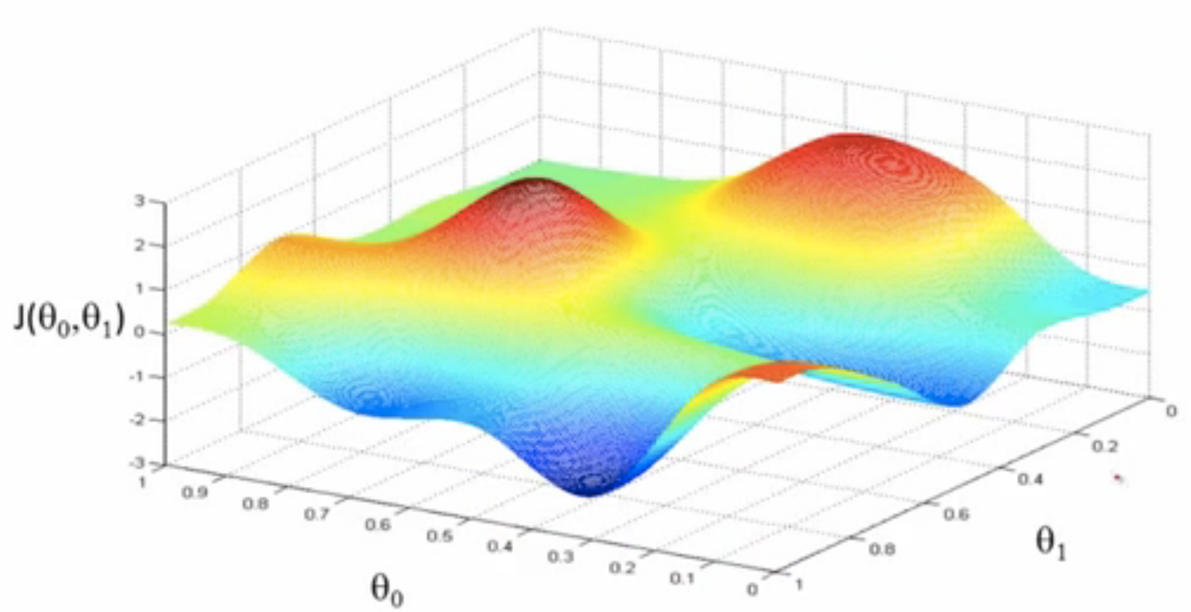
- Example: Minimize the variance of your returns while earning at least $100 in the stock market.

Machine Learning and Optimization

- Linear regression $$\min_w \; ||Xw - y||^2$$
- Logistic regression $$\min_w \; \sum_i \log(1 + \exp(-y_i x_i^\top w))$$
- SVM $$\min_w \; ||w||^2 + C \sum_i \xi_i$$ $$\text{s.t. } \xi_i \geq 1 - y_i x_i^\top w$$ $$\xi_i \geq 0$$
- And many more ...

Non-Convex Problems are Everywhere

- Local (non-global) minima
- All kinds of constraints



No easy solution for these problems

Consider convex problems

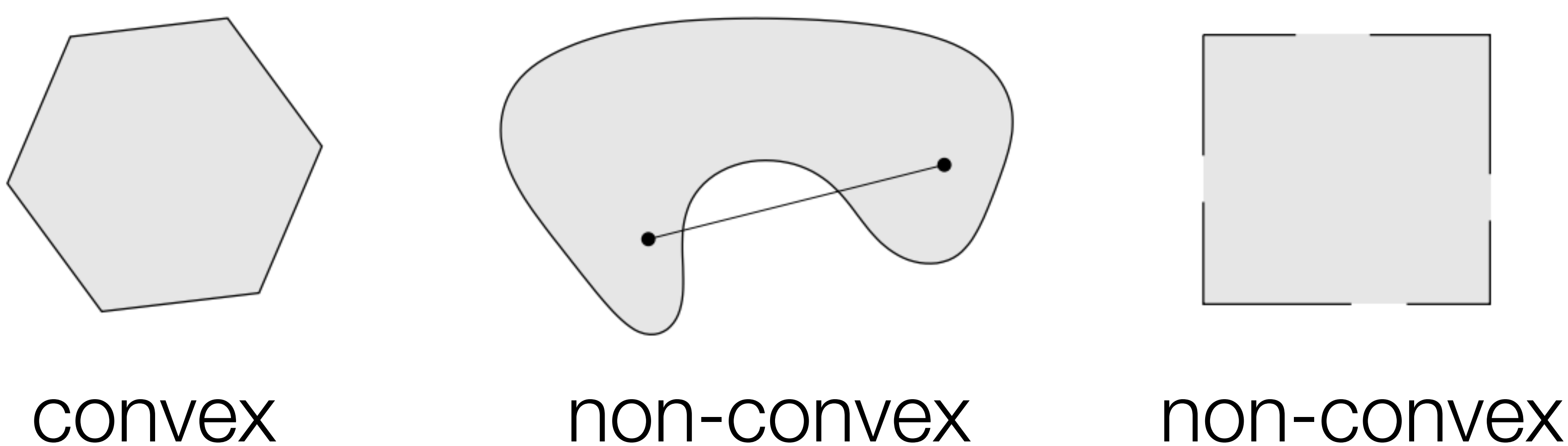
Why Convex Optimization?

- Achieves global minimum, no local traps
- Highly efficient software available
- Can be solved by polynomial time complexity algorithms
- Dividing line between “easy” and “difficult” problems

Convex Sets

Any line segment joining any two elements lies entirely in set

$$x_1, x_2 \in C, 0 \leq \theta \leq 1 \implies \theta x_1 + (1-\theta)x_2 \in C$$

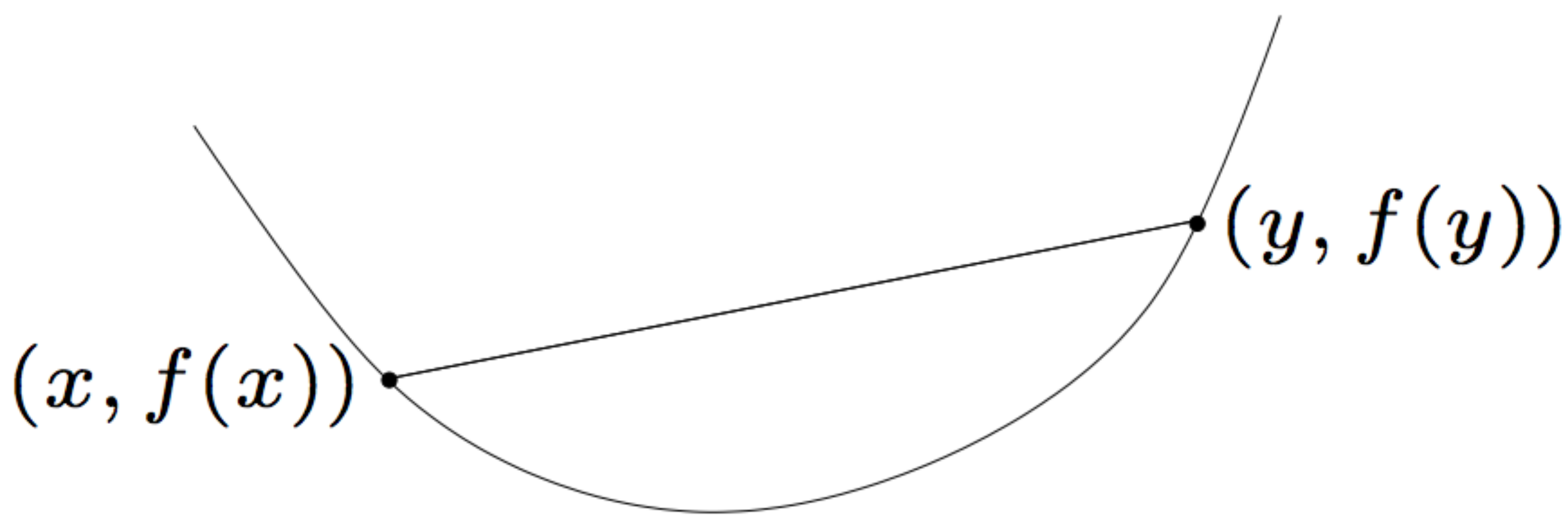


Convex Function

$f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathbf{dom}\ f$ is a convex set and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

for all $x, y \in \mathbf{dom}\ f, 0 \leq \theta \leq 1$



f lies below the line segment joining f(x), f(y)

Properties of Convex Functions

- Convexity over all lines

$$f(x) \text{ is convex } \implies f(x_0 + th) \text{ is convex in } t \text{ for all } x_0, h$$

- Positive multiple

$$f(x) \text{ is convex } \implies \alpha f(x) \text{ is convex for all } \alpha \geq 0$$

- Sum of convex functions

$$f_1(x), f_2(x) \text{ convex } \implies f_1(x) + f_2(x) \text{ is convex}$$

- Pointwise maximum

$$f_1(x), f_2(x) \text{ convex } \implies \max\{f_1(x), f_2(x)\} \text{ is convex}$$

- Affine transformation of domain

$$f(x) \text{ is convex } \implies f(Ax + b) \text{ is convex}$$

Convex Optimization Problem

Definition:
An optimization problem is **convex** if its objective is a convex function, the inequality constraints are convex, and the equality constraints are affine

$$\min_x \; f_0(x)$$ convex function

$$\text{s.t. } f_k(x) \le 0, k = 1, 2, \cdots, K$$ convex sets

$$h_j(x) = 0, j = 1, 2, \cdots, J$$ affine constraints

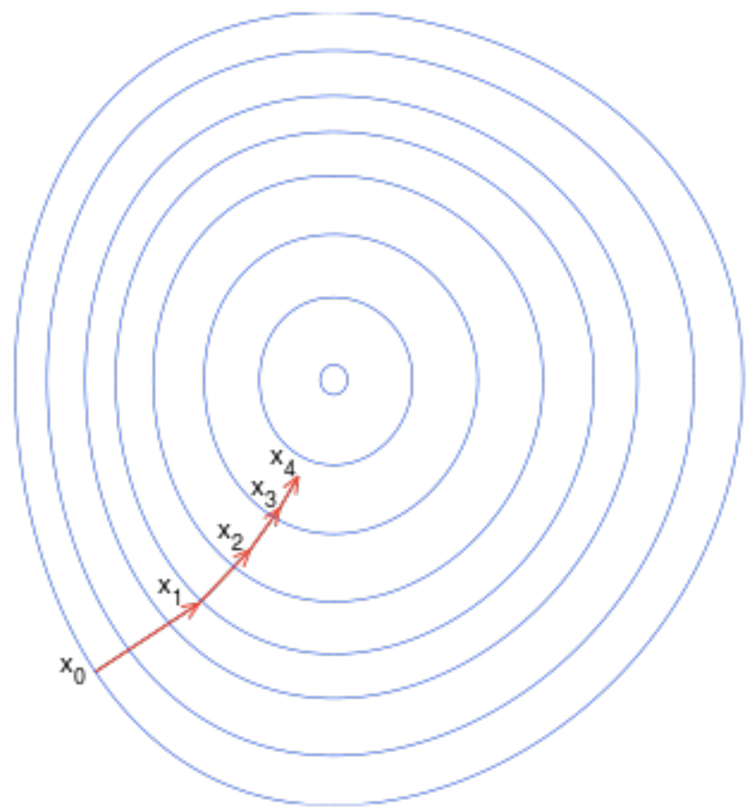
Benefits of Convexity

- Theorem: If x is a local minimizer of a convex optimization problem, it is a **global** minimizer
- Theorem: If the gradient at c is zero, then c is the global minimum of f(x)

$$\nabla f(c) = 0 \iff c = x^*$$

Gradient Descent (Steepest Descent)

- Simplest and extremely popular
- Main Idea: take a step proportional to the negative of the gradient
- Easy to implement
- Each iteration is relatively cheap
- Can be slow to converge



Gradient Descent Algorithm

Algorithm 1: Gradient Descent

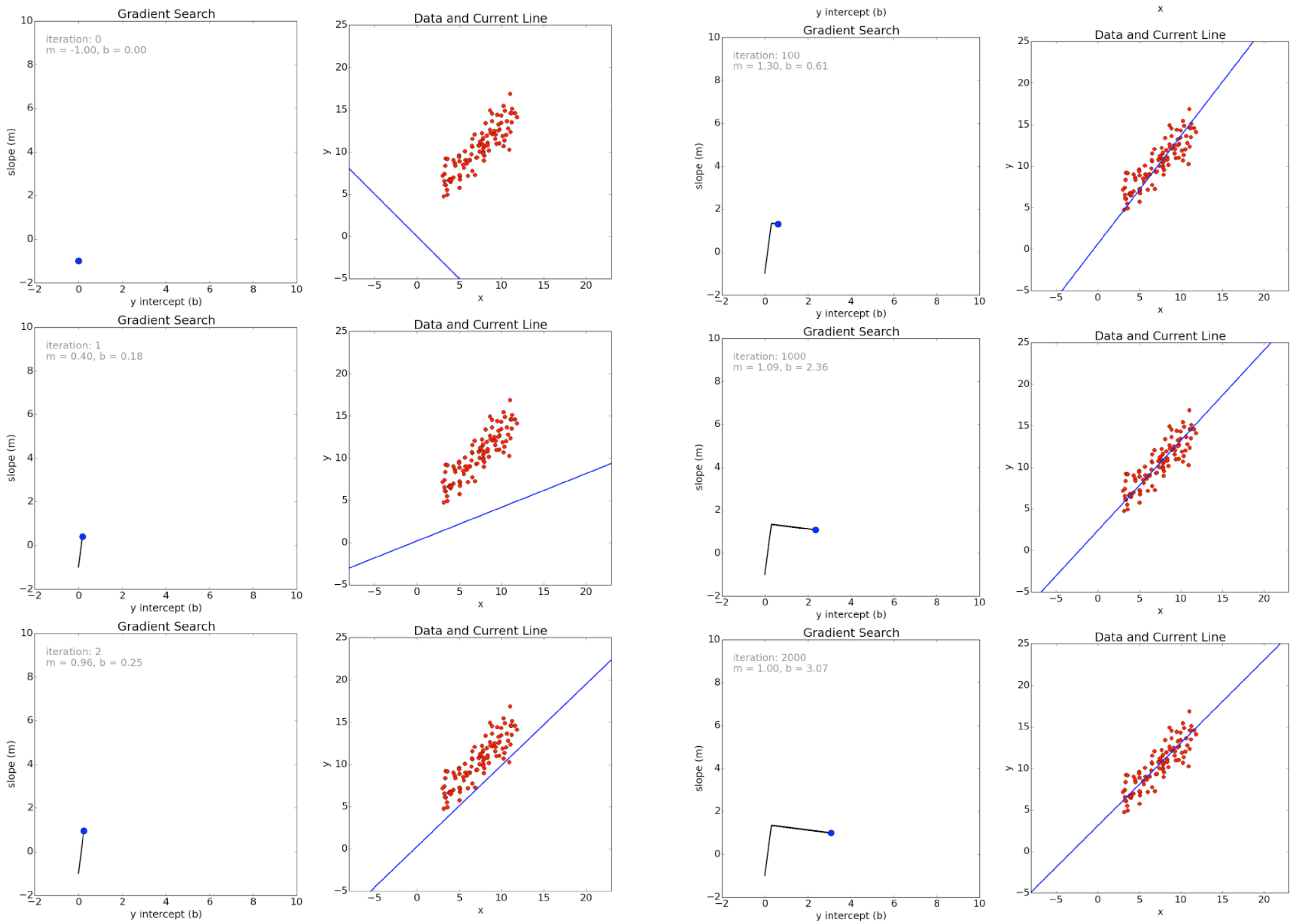
while *Not Converged* **do**

$$x^{(k+1)} = x^{(k)} - \eta^{(k)} \nabla f(x)$$

end

return $x^{(k+1)}$

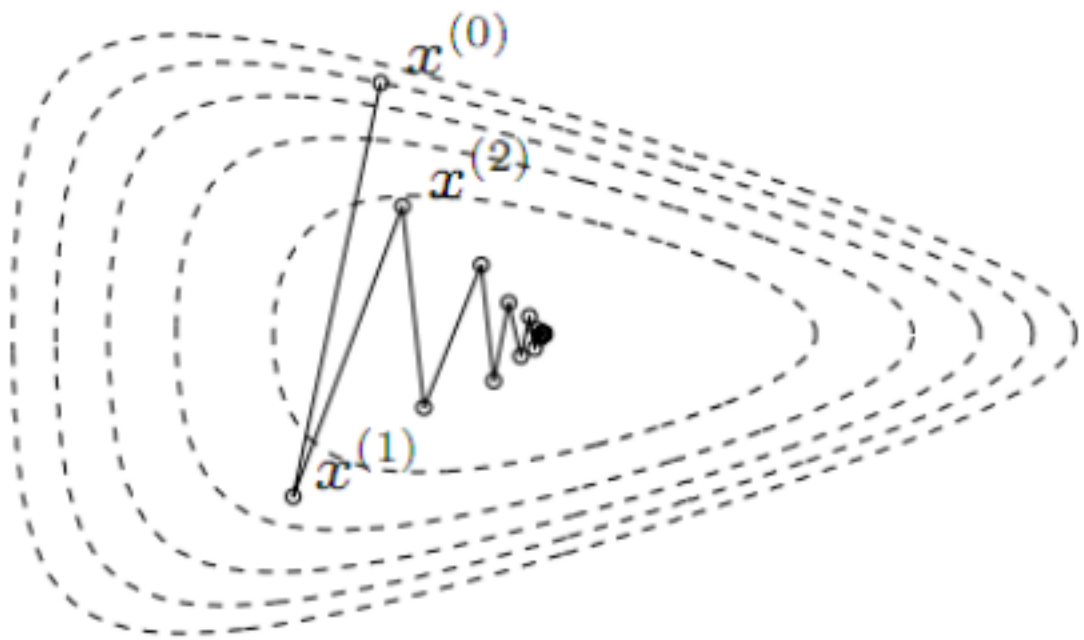
Gradient Descent: Linear Regression



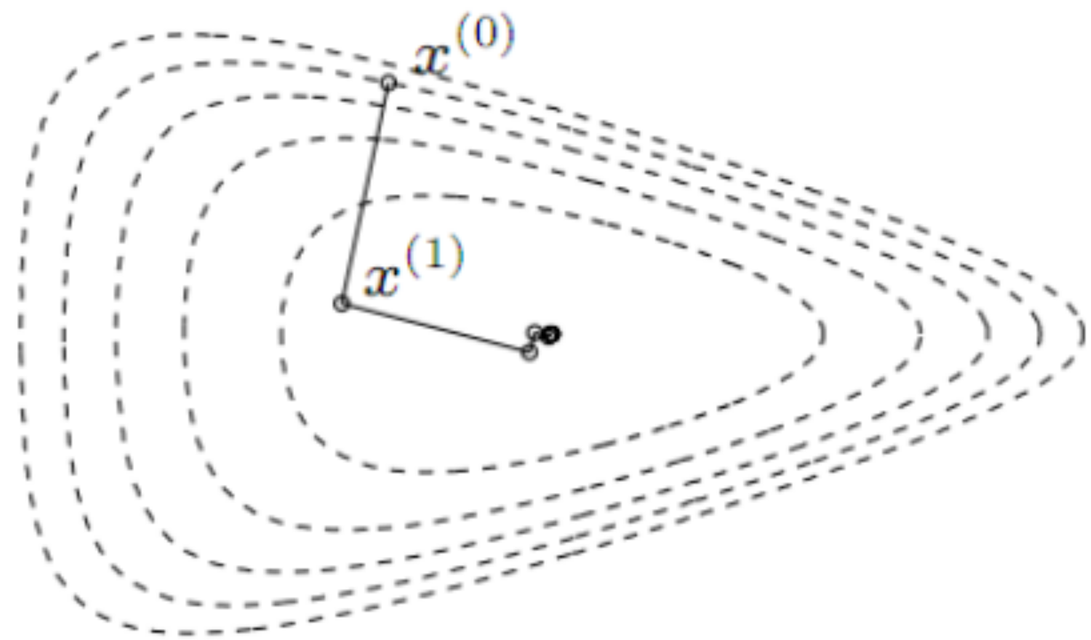
http://spin.atomicobject.com/2014/06/24/gradient-descent-linear-regression/

Gradient Descent: Example 2

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



backtracking line search



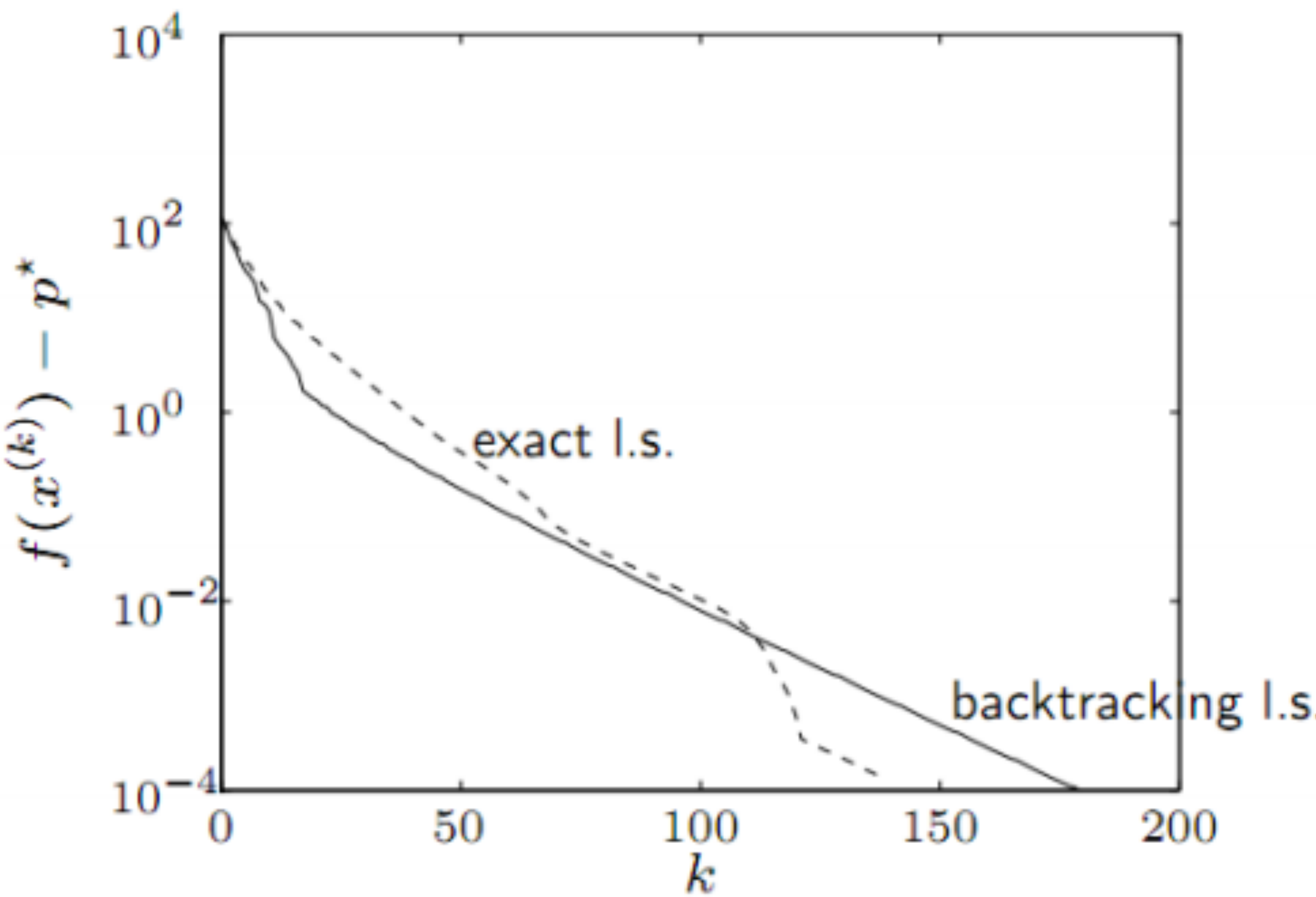
exact line search

Boyd & Landenberghe's Book on Convex Optimization

Gradient Descent: Example 3

$$f(x) = c^T x - \sum_{i=1}^{500} \log(b_i - a_i^T x)$$

A problem in R100



Boyd & Landenberghe's Book on Convex Optimization

Limitations of Gradient Descent

- Step size search may be expensive
- Convergence is slow for ill-conditioned problems
- Convergence speed depends on initial starting position
- Does not work for non differentiable or constrained problems

Constrained Optimization

$$\min_x \; f_0(x)$$
$$\text{s.t } f_k(x) \leq 0, \; k = 1, \cdots, K$$

Lagrange Duality

- Bound or solve an optimization problem via a different optimization problem
- Optimization problems (even non-convex) can be transformed to their dual problems
- Purpose of the dual problem is to determine the lower bounds for the optimal value of the original problem
- Under certain conditions, solutions of both problems are equal and the dual problem often offers easier and analytical way to the solution

Reasons Why Dual is Easier

- Dual problem is unconstrained or has simple constraints
- Dual objective is differentiable or has a simple non differentiable term
- Exploit separable structure in the decomposition for easier algorithm

Construct the Dual

Original optimization problem or primal problem

$$\min_x \; f_0(x)$$

$$\text{s.t. } f_k(x) \leq 0, k = 1, 2, \cdots, K$$

$$h_j(x) = 0, j = 1, 2, \cdots, J$$

Lagrangian

$$L(x, \lambda, v) = f_0(x) + \sum_k \lambda_k f_k(x) + \sum_j v_j h_j(x)$$

Lagrange multipliers or dual variables

Constructing the Dual

Original optimization problem or primal problem

$$\min_x \; f_0(x)$$

$$\text{s.t. } f_k(x) \leq 0, k = 1, 2, \cdots, K$$

$$h_j(x) = 0, j = 1, 2, \cdots, J$$

infimum is the element that is smallest or equal to all elements in the set

Dual problem

$$\max g(\lambda, v) = \inf_x L(x, \lambda, v)$$

$$\text{subject to } \lambda \geq 0$$

dual function is always lower bound for optimal value of original function

$$g(\lambda, v) \leq L(\tilde{x}, \lambda, v) \leq f_0(\tilde{x})$$

Lagrange Dual: Separable Example

$$\begin{aligned} \min \ & f_1(x_1) + f_2(x_2) \\ \text{subject to } & A_1 x_1 + A_2 x_2 \leq b \end{aligned}$$

coupling constraint in primal problem

$$\begin{aligned} \max \ & -f_1^*(-A_1^\top z) - f_2^*(-A_2^\top z) - b^\top z \\ \text{subject to } & z \geq 0 \end{aligned}$$

dual problem can be easily solved by gradient projection

Some Resources for Convex Optimization

- Boyd & Landenberghe's Book on Convex Optimization
https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf

- Stephen Boyd's Class at Stanford
http://stanford.edu/class/ee364a/

- Vandenberghe's Class at UCLA
http://www.seas.ucla.edu/~vandenbe/ee236b/ee236b.html

- Ben-Tai & Nemirovski Lectures on Modern Convex Optimization
http://epubs.siam.org/doi/book/10.1137/1.9780898718829