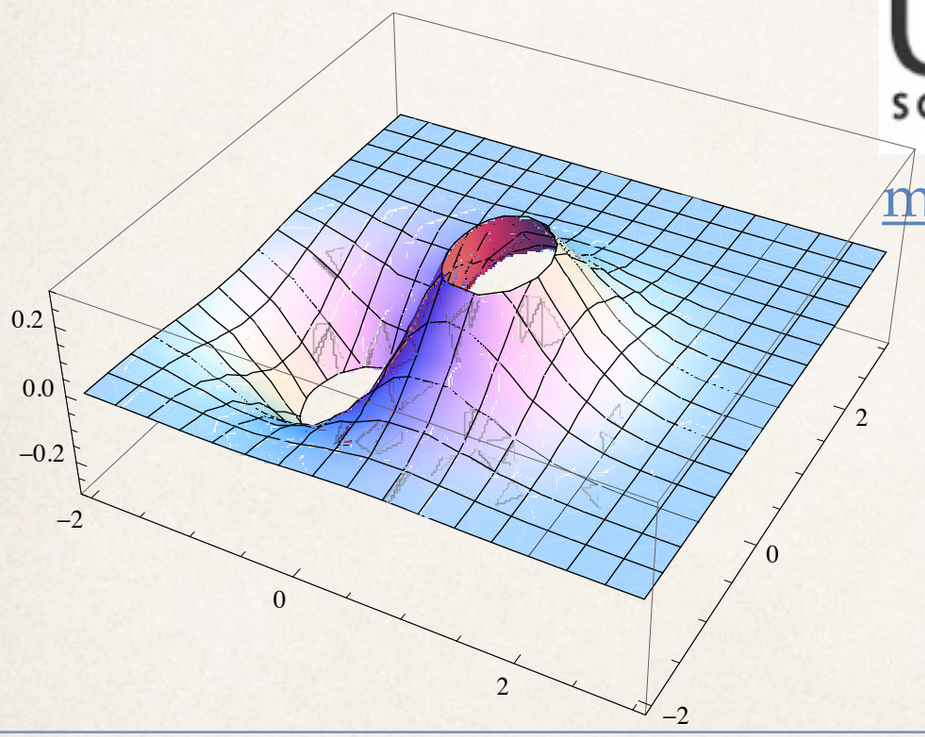
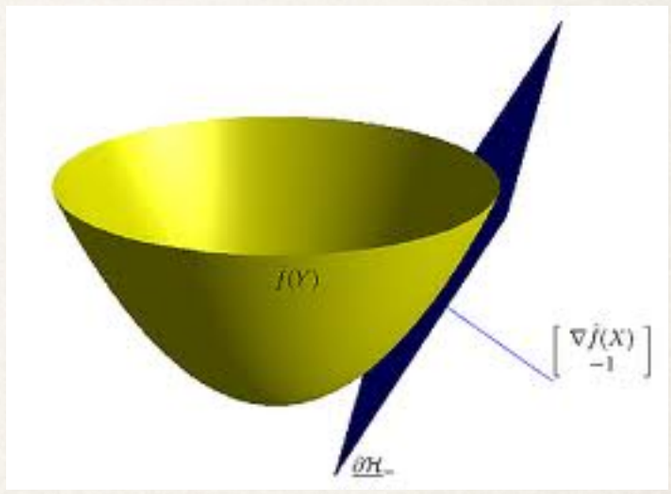
Optimization for CS: Duality and Unconstrained Optimization

Sridhar Mahadevan

UMASSCS
SCHOOL OF COMPUTER SCIENCE

mahadeva@cs.umass.edu





Lecture 4

Transfer Learning

Computer vision

Transfer Learning

Athens is to Greece as Baghdad is to ?

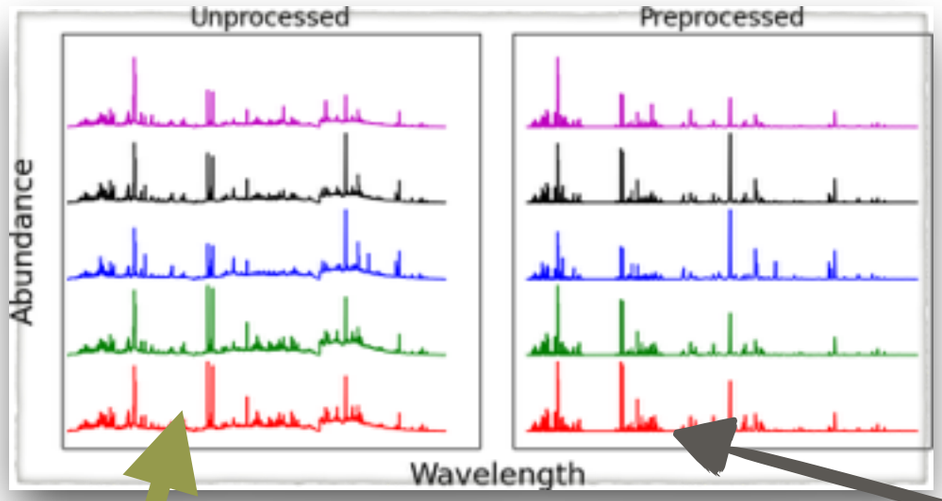
he is to she as grandpa is to X?

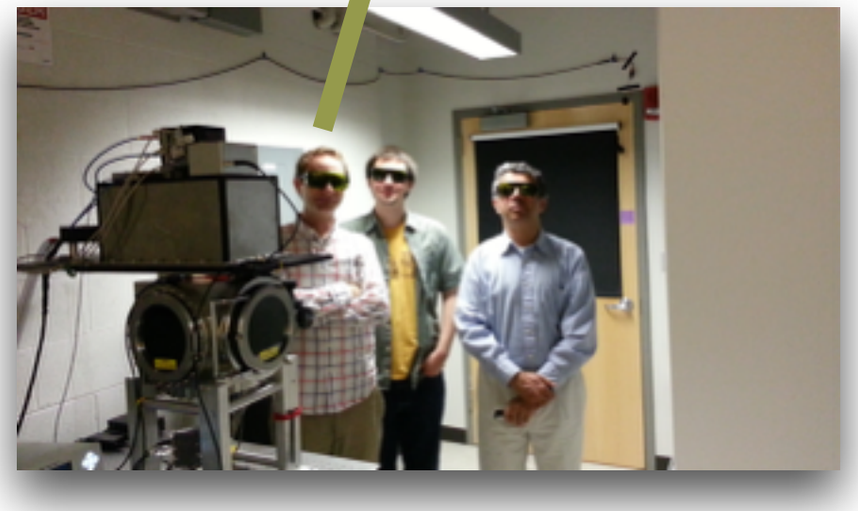
cheap is to cheaper as high is to X?

Europe is to euro as Vietnam is to X?

NLP

Transfer Learning on Mars





Same laser
on Earth
as on Mars

Curiosity zapping a rock with a laser

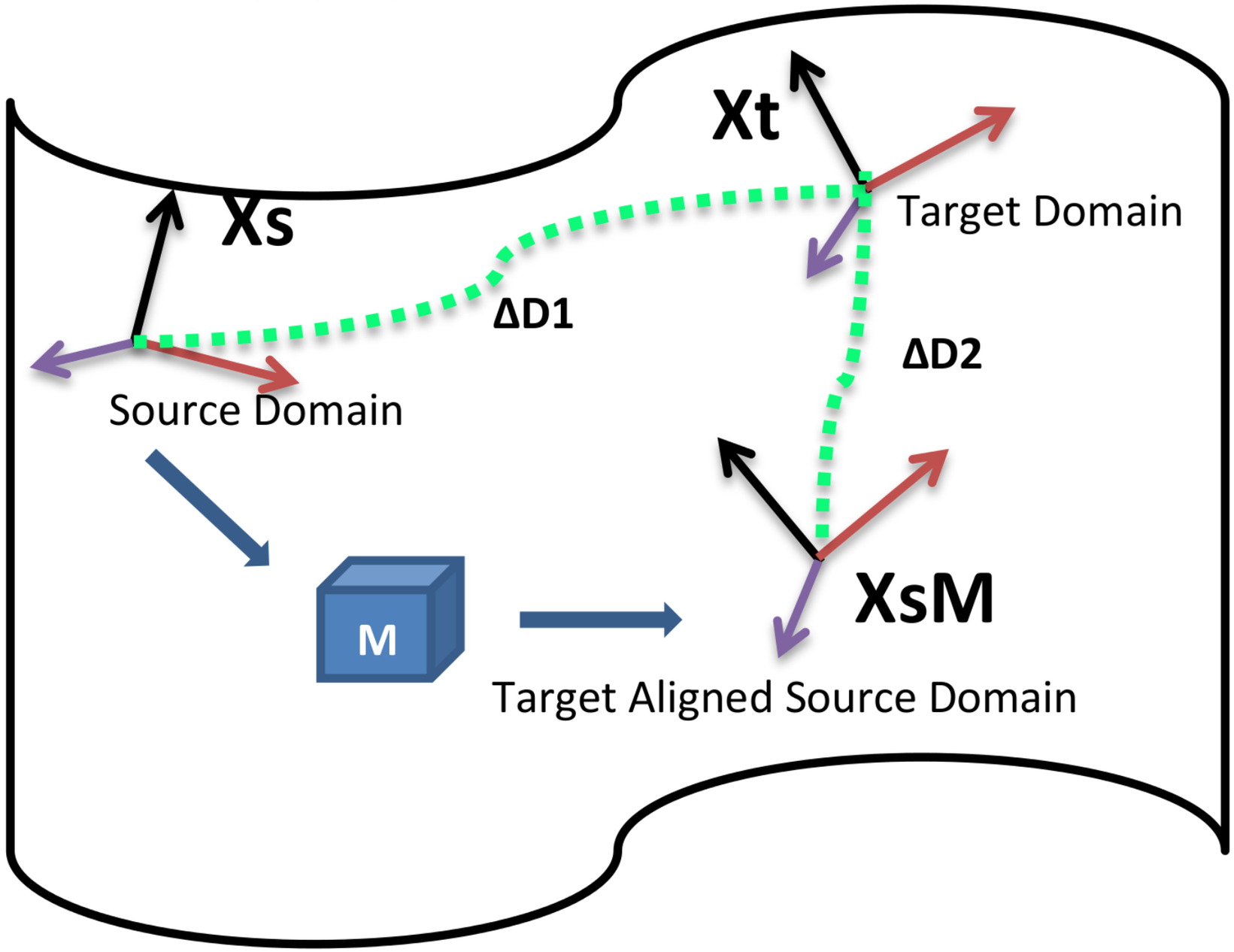
Domain Adaptation

- DA is a framework for transfer learning
  - Training and test data may have different distributions
- How to train a classifier on the source domain, and then adapt to the target domain with a different distribution?

$$F(M) = ||X_S M - X_T||_F^2$$

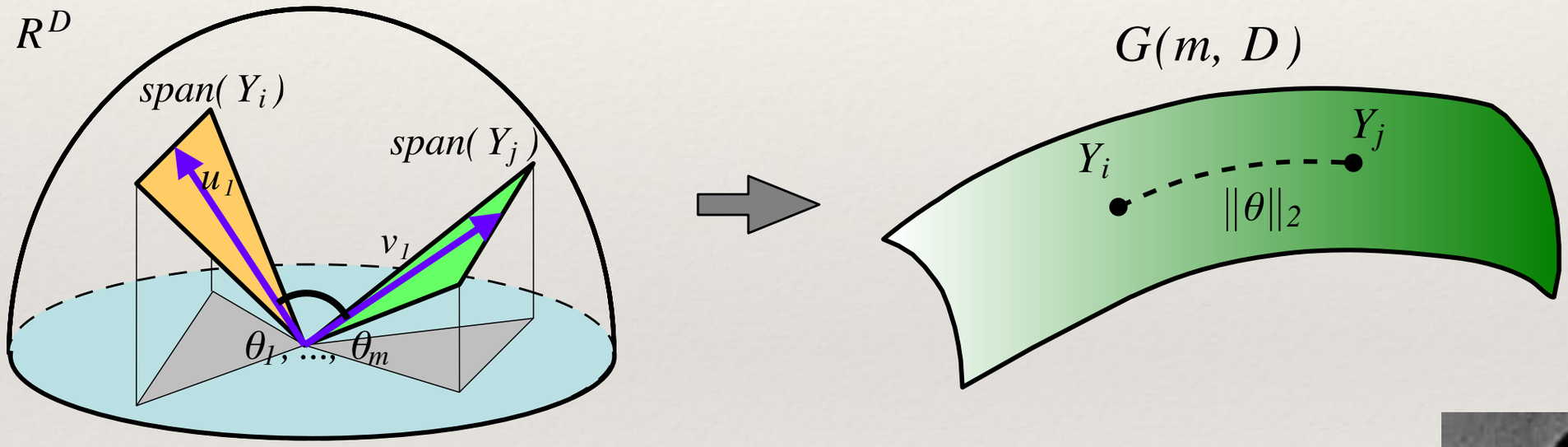
$$\begin{aligned} M^* &= argmin_M ||X_S' X_S M - X_S' X_T||_F^2 \\ &= argmin_M ||M - X_S' X_T||_F^2. \end{aligned}$$

$$M^* = argmin_M(F(M))$$



Subspace Alignment (Fernando et al., CVPR 2014)

Grassmannian Manifolds



How to optimize subspace alignment by choosing the source and target subspaces adaptively?

1809-1877

Outline

- Duality in LP
- Lagrangian Duality and Kuhn Tucker theorem
- Fundamentals of Unconstrained Optimization
  - Steepest descent
  - Conjugate gradient descent
  - Newton's method

Duality in Linear Programming

Primal LP Problem : $\min_{x \in \mathbb{R}^n} c'x$

such that $Ax = b, x \geq 0$

Dual LP Problem: $\max_{p \in \mathbb{R}^m} p'b$

such that $p'A \leq c'$

Example of Dual LP problem

- Primal problem:

Minimize $x_1 + x_2$ such that

$x_1 + 2x_2 - x_3 = 2$

$x_1 - x_4 = 1$

$x_1, x_2, x_3, x_4 >= 0$

- Dual Problem:

Maximize $2p_1 + p_2$ such that

$p_1 + p_2 <= 1$

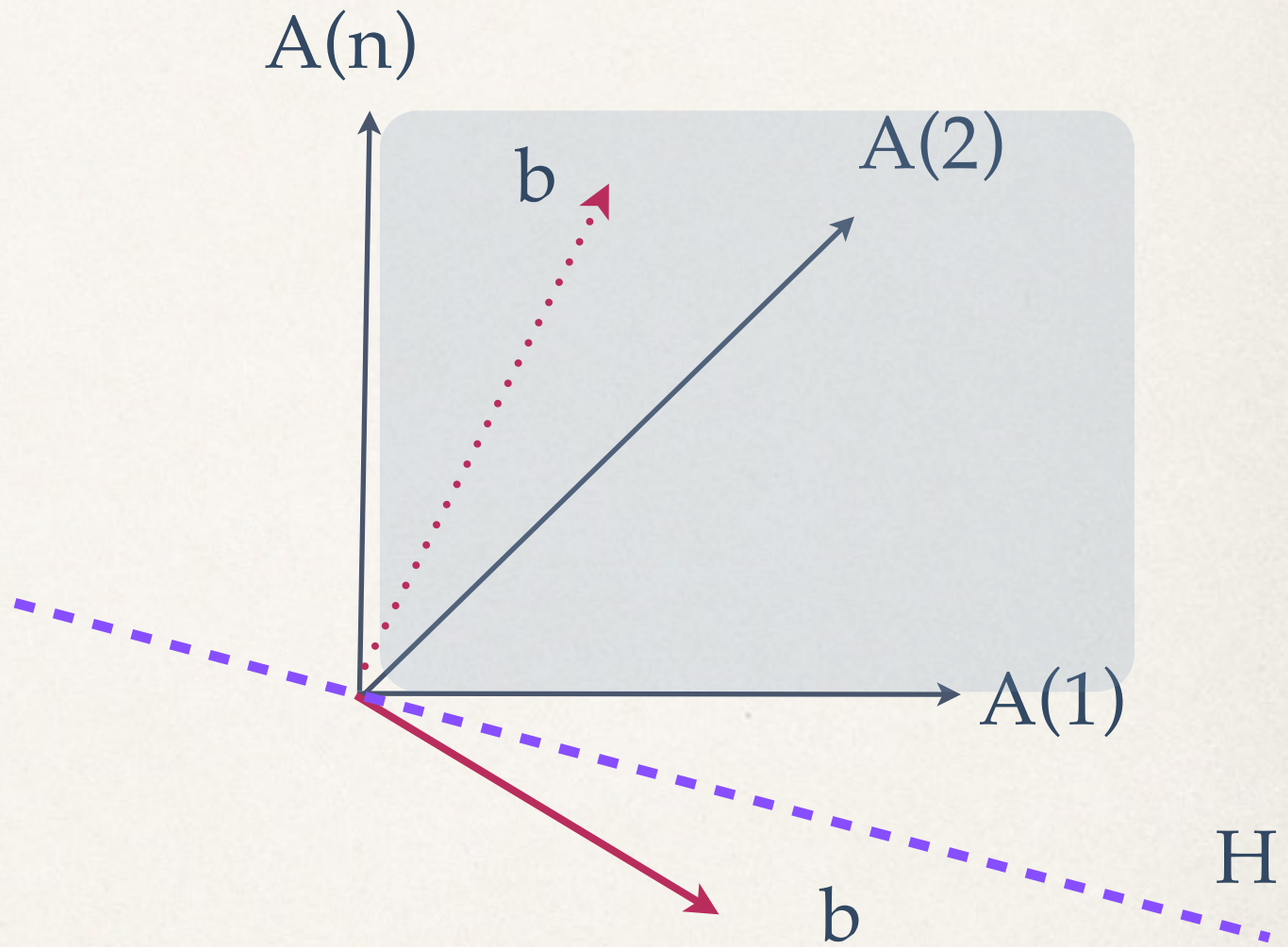
$2p_1 <= 1$

$p_1, p_2 >= 0$

Farkas Lemma and LP

- Farkas' lemma states that either a vector is in a given convex cone, or there is a hyperplane separating the vector from the cone

- Fundamental result that forms the basis for understanding duality in LP



Farkas Lemma

- Let A be a matrix of size m x n.
- Let b be a column vector of size m
- Then, only one of the following holds:
  1. $Ax = b$ has a nonnegative solution ($x >= 0$)
  2. $y' A >= 0$ and $y' b < 0$ has a solution
- Proof: Follows from Hahn-Banach theorem

Proof of Farkas Lemma

- **Proof**: Suppose both conditions are true. Then, a contradiction ensues:

$$0 <= (y' A)x = y'(Ax) = y' b < 0!$$

- Let C be the convex set of non-negative vectors Ax, x>=0.
- If b is in C, then condition 1 holds. Otherwise, b is separable from C by a hyperplane. Specifically, there is a vector p such that p'b < p'y for all y in C
- Since $A_i\delta$ is in C for $\delta >= 0$, it follows $p'b < \delta p' A_i$.
- This implies $(1/\delta)p'b < p' A_i$. As δ -> infinity, $p' A_i >= 0$, so $p' A >= 0$

Duality in LP: Alternate Form

- Primal:
  - Minimize c'x such that Ax >= b
- Dual form:
  - Maximize p'b such that
  - p'A = c'
  - p >= 0

Alternative Form of Farkas' Lemma

- **Lemma**: Let $\mathbf{A_1}$, ..., $\mathbf{A_n}$ and **b** be given vectors, and suppose that any vector **p** that satisfies $\mathbf{p'A_i} >= 0$, i = 1, ..., n must also satisfy **p'b** $> =0$.

- Then **b** can be expressed as a nonnegative combination of the vectors $\mathbf{A_1,...A_n}$

- **Proof:** This follows easily from the previous statement of Farkas' lemma (since this is saying condition 2 does not hold, so condition 1 must hold)

Financial Planning using Farkas Lemma

- Consider a market in which **n** different assets are traded
- During a trading period, the market can be in one of **m** possible states
- Investing $1 in asset **i** provides a payoff of $\mathbf{r_{si}}$ when market is in state **s**
- Define the payoff matrix as **R** (m x n)
- A portfolio of assets is a vector **x** (positive or negative)
- Negative assets indicate ``short'' positions (you agree to sell $|\mathbf{x_i}|$

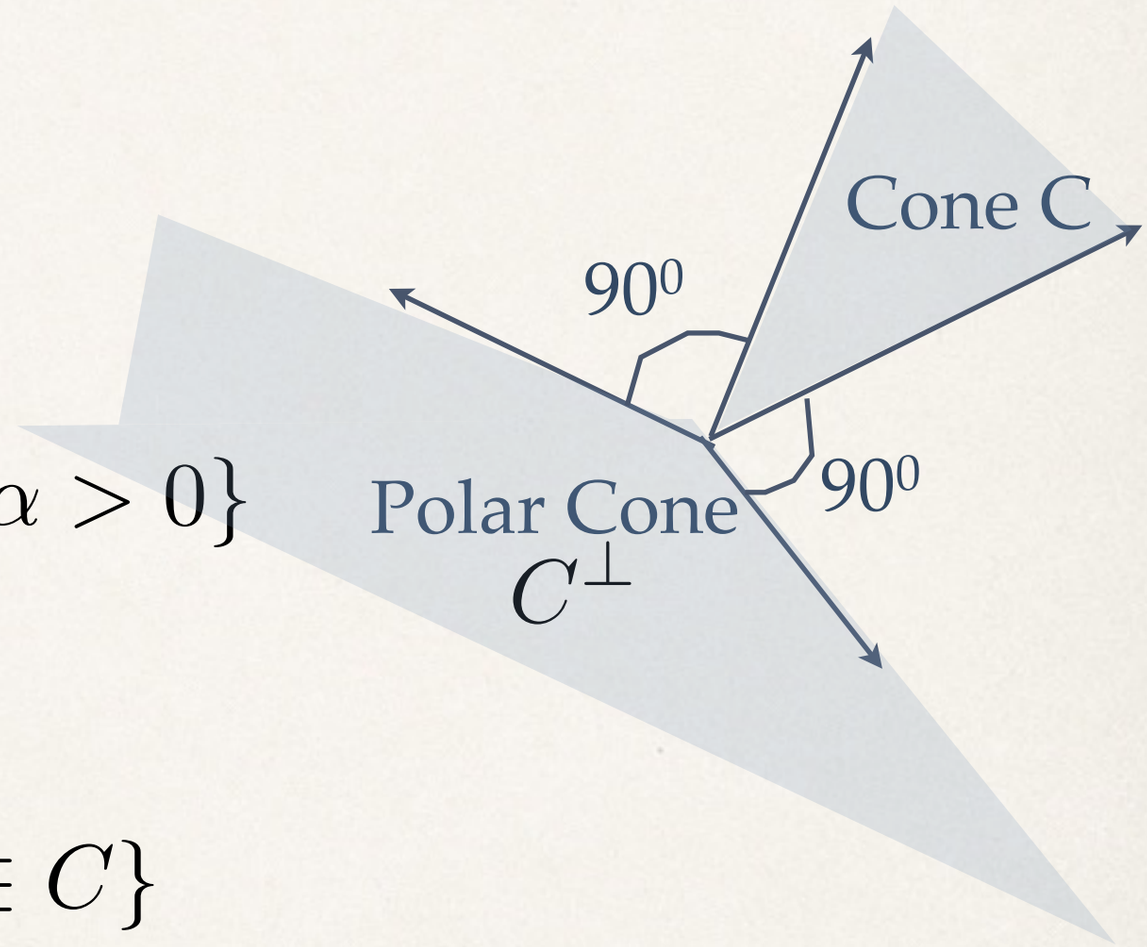
Financial Planning example (cont)

- The wealth **w** that results from a portfolio **x** is given by **w = R x**
- Let **p** be a vector of asset prices, and **p'x** be the cost of acquiring portfolio **x**
- Absence of arbitrage condition: if **Rx >= 0** then **p'x >=0**
  - Any portfolio that has a nonnegative payoff must have a nonnegative cost
- **Theorem**: Absence of arbitrage condition holds if and only if there exists a nonnegative vector $\mathbf{q} = (q_1,\ldots,q_m)$ s.t. price $p_i = \Sigma_s\, q_s\, r_{si}$
- **Proof:** Straightforward consequence of Farkas lemma

Polar Cone Theorem

- Recall the definition of a cone:

$$C = \{y \in X : y = \alpha x, x \in X, \alpha > 0\}$$



- The **polar cone** is defined as:

$$C^\perp = \{y \in X : \langle y, x \rangle \leq 0, x \in C\}$$

- The polar cone theorem states:

$$(C^\perp)^\perp = C$$

System of Inequalities

Let C be a convex set and $f_1, \ldots, f_k$ be convex functions such that dom $f_i \supset$ ri C. Let $g_1, \ldots, g_l$ be affine functions such that the system

$$g_1(x) \leq 0, \ldots, g_l(x) \leq 0$$

has at least one solution in the ri C. Then, only one of the following alternatives holds:

- There exists some $x \in C$ such that

$$f_1(x) < 0, \ldots, f_k(x) < 0 \qquad g_1(x) \leq 0, \ldots g_l(x) \leq 0$$

- There exists non-negative real numbers $\lambda_1, \ldots, \lambda_k, \xi_1, \ldots, \xi_l$ such that at least one of the λ_i are non-zero and

$$\lambda_1 f_1(x) + \ldots + \lambda_k f_k(x) + \xi_1 g_1(x) + \ldots + \xi_l g_l(x) \geq 0, \quad \forall x \in C$$

Conjugate Duality

- Given a convex function f on a convex set C, the conjugate set and conjugate function are defined as:

$$C^* = \{x^* \in X^* : \sup_{x \in C} [\langle x, x^* \rangle - f(x)] < \infty\}$$

$$f^*(x^*) = \sup_{x \in C} [\langle x, x^* \rangle - f(x)]$$

Note: for R^n, x^* and x are both n-tuples of real numbers

General Constrained Optimization

- Consider the constrained minimization problem

$$\min_w f(w) \qquad \text{such that} \quad g_i(w) \leq 0, i = 1, \ldots, k$$
$$\text{and} \qquad h_i(w) = 0, i = 1, \ldots, l$$

- We now review the framework of Lagrange Duality and the Karush Kuhn Tucker theorem

Lagrange Dual Formulation

- The *primal* optimization problem is

$$\min_w f(w) \qquad \text{such that} \quad g_i(w) \leq 0, i = 1, \ldots, k$$

$$\text{and} \qquad h_i(w) = 0, i = 1, \ldots, l$$

- The *dual* problem can be formulated using Lagrange multipliers as $\max_{\alpha,\beta:\alpha\geq 0} (L_D(\alpha, \beta))$:

$$L_D(\alpha, \beta) = \min_w \left(f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w) \right)$$

Duality Theorems

- *Weak Duality Theorem:* The dual formulation always produces a solution that is upper bounded by the solution to the primal problem.
- *Strong Duality Theorem:* The solution to the Lagrange dual is exactly the same as the primal solution, assuming that the function $f(w)$ and the constraints $g_i(w)$ are convex, and $h_i(w)$ is an affine set

Weak Duality Theorem

- Suppose w is a feasible solution to the primal problem, and that α and β constitute a solution to the dual problem.

$$\begin{aligned} L_D(\alpha, \beta) &= \min_u L(u, \alpha, \beta) \\ &\leq L(w, \alpha, \beta) \\ &= f(w) + \sum_i \alpha_i g_i(w) + \sum_i \beta_i h_i(w) \leq f(w) \end{aligned}$$

- This implies the following condition:

$$\max_{\alpha,\beta:\alpha\geq 0} L_D(\alpha, \beta) \leq \min_w \{f(w) : g_i(w) \leq 0, h_i(w) = 0\}$$

Weak Duality Theorem for LP

- **Lemma**: if there exists x and p such that c'x = p'b, then x = x* is the optimal primal solution and p = p* is the optimal dual solution

- **Proof**: Note that since any dual solution forms a lower bound on the primal solution, it follows that:

  - c'x = p'b <= c'y (for any y)

  - Consequently, x must be optimal

Strong Duality Theorem (for LP): Proof

- Let $\mathbf{I = \{i \mid a'_i x^* = b_i\}}$ where $\mathbf{x^*}$ is the optimal primal solution
- Any vector **d** such that $\mathbf{a_i'd >= 0}$ for any i in I also satisfies $\mathbf{c'd>=0}$
  - This holds since $\mathbf{a'_i(x^* + \varepsilon d) >= b_i}$ for all i. Hence $\mathbf{c'(x^* + \varepsilon d)>=c'x^*}$ implying $\mathbf{c'd >=0}$
- By Farkas lemma, c can be written as $\mathbf{c = \Sigma_{i\varepsilon I}\, p_i\, a_i}$. Define $\mathbf{p_i=0}$ for i not in I. This implies $\mathbf{p'A = c'}$.
- Also, $\mathbf{p'b = \Sigma_{i\varepsilon I}\, p_i\, b_i = \Sigma_{i\varepsilon I}\, p_i\, a'_i\, x^* = c'x^*}$.
- This means dual cost = primal optimal cost, and both are optimal!

Sparsity of Parameters

- **Corollary:** Let w^* be a weight vector that satisfies the primal constraints and α^*, β^* be the Lagrangian variables that satisfies both the dual constraints.

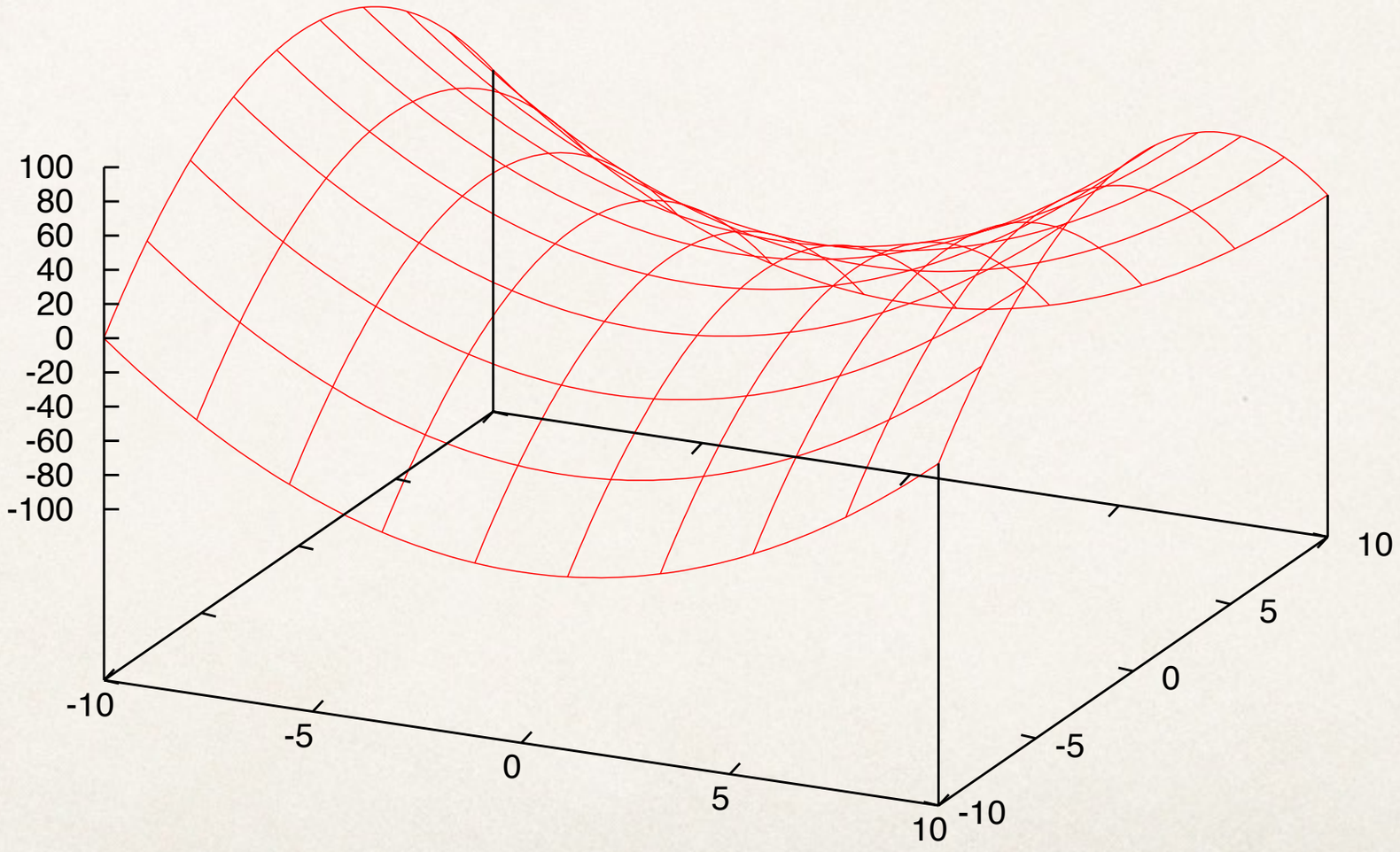
  $f(w^*) = L_D(\alpha^*, \beta^*)$ where $\alpha_i^* \geq 0$ and $g_i(w^*) \leq 0, h_i(w^*) = 0$

- Then, $\alpha_i^* g_i(w^*) = 0$ for $i = 1, \ldots, k$.
- The proof follows easily by noting that the inequality

$$f(w^*) + \sum_i \alpha_i^* g_i(w^*) + \sum_i \beta_i^* h_i(w^*) \leq f(w^*)$$

becomes an equality only when $\alpha_i^* g_i(w^*) = 0$ for $i = 1, \ldots, k$.

Saddle Point Function



Duality Gap and Saddle Points

- Define a *saddle point* as the triple w^*, α^*, β^*, where $w^* \in \Omega, \alpha^* \geq 0$, and

$$L(w^*, \alpha, \beta) \leq L(w^*, \alpha^*, \beta^*) \leq L(w, \alpha^*, \beta^*)$$

- *Theorem:* The triple w^*, α^*, β^* is a saddle point if and only if w^* is a solution to the primal problem, and α^*, β^* is a solution to the dual problem, and there is no duality gap, so $f(w^*) = L_D(\alpha^*, \beta^*)$.
- *Strong Duality Theorem:* If $f(w)$ is convex, and $w \in \Omega$, where Ω is a convex set, and g_i, h_i are affine functions, the duality gap is 0.

General KKT Conditions

- Let w^* and α^*, β^* the optimal primal and dual solutions with zero duality gap.

- Since x^* minimizes $L(x, \alpha^*, \beta^*)$, it follows that its gradient must vanish. The general KKT conditions are given as:

$$h_i(x^*) = 0, i = 1, \ldots, l$$
$$g_i(x^*) \leq 0, i = 1, \ldots, k$$
$$\alpha_i^* \geq 0, i = 1, \ldots, k$$
$$\alpha_i^* g_i(x^*) = 0, i = 1, \ldots, k$$
$$\nabla f_0(x^*) + \sum_{i=1}^{k} \alpha_i^* g_i(x^*) + \sum_{i=1}^{l} \beta_i^* h_i(x^*) = 0$$

Least Squares Estimation

- Consider the problem: minimize $x^T x$ such that $Ax = b$
- **Lagrange Dual form:**

$$L(x, \nu) = x^T x + \nu^T (Ax - b)$$

- Applying KKT theorem:

$$\nabla_x L(x, \nu) = 2x + A^T \nu = 0 \Rightarrow x = -\frac{1}{2} A^T \nu$$

Dual problem: $\max_{\nu} L(\nu) = -\frac{1}{4} \nu^T A A^T \nu - b^T \nu$

Equality Constrained Convex Quadratic Minimization

Primal Problem:

$$\text{Minimize } \frac{1}{2}x^T Px + q^T x + r \text{ such that } Ax = b$$

Lagrange Dual:

$$L(x, \beta) = \frac{1}{2}x^T Px + q^T x + r + \beta^T (Ax - b)$$

At optimal solution x^*:

$$Ax^* = b, \quad Px^* + q + A^T \beta^* = 0$$

KKT Matrix:

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \beta^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

Example: Optimal Margin Classifiers

Optimal Margin Classification

- Consider the problem of finding a set of weights w that produces a hyperplane with the maximum geometric margin.

$$\max_{\gamma,w,b} \quad \gamma \quad \text{such that}$$

$$y_i(\langle w, x_i \rangle + b) \geq \gamma, i = 1, \ldots, m$$

$$\|w\| = 1$$

- We eliminate the non-convex constraint $\|w\| = 1$ as follows:

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ such that}$$

$$y_i(\langle w, x_i \rangle + b) \geq 1, i = 1, \ldots, m$$

Support Vectors

- We can formulate the optimal margin classification problem as:

$$\min_w \frac{1}{2}\|w\|^2 \text{ such that}$$

$$g_i(w) = -y_i(\langle w, x_i \rangle + b) + 1 \leq 0, \quad i = 1, \ldots, m$$

- KKT implies instances for which $\alpha_i > 0$ are those which have functional margins exactly $= 1$ (because then $g_i(w) = 0$).
- The functional margin is the smallest of all the margins, which implies that we will only have nonzero α_i for the points closest to the decision boundary! These are called the *support vectors*.

Dual Form

- We can write the Lagrangian for our optimal margin classifier as

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left(y_i(\langle w, x_i \rangle + b) - 1\right)$$

- To solve the dual form, we first minimize with respect to w and b, and then maximize w.r.t. α

$$\nabla_w L(w, b, a) = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

$$\nabla_b L(w, b, a) = \sum_{i=1}^{m} \alpha_i y_i = 0$$

Support Vectors

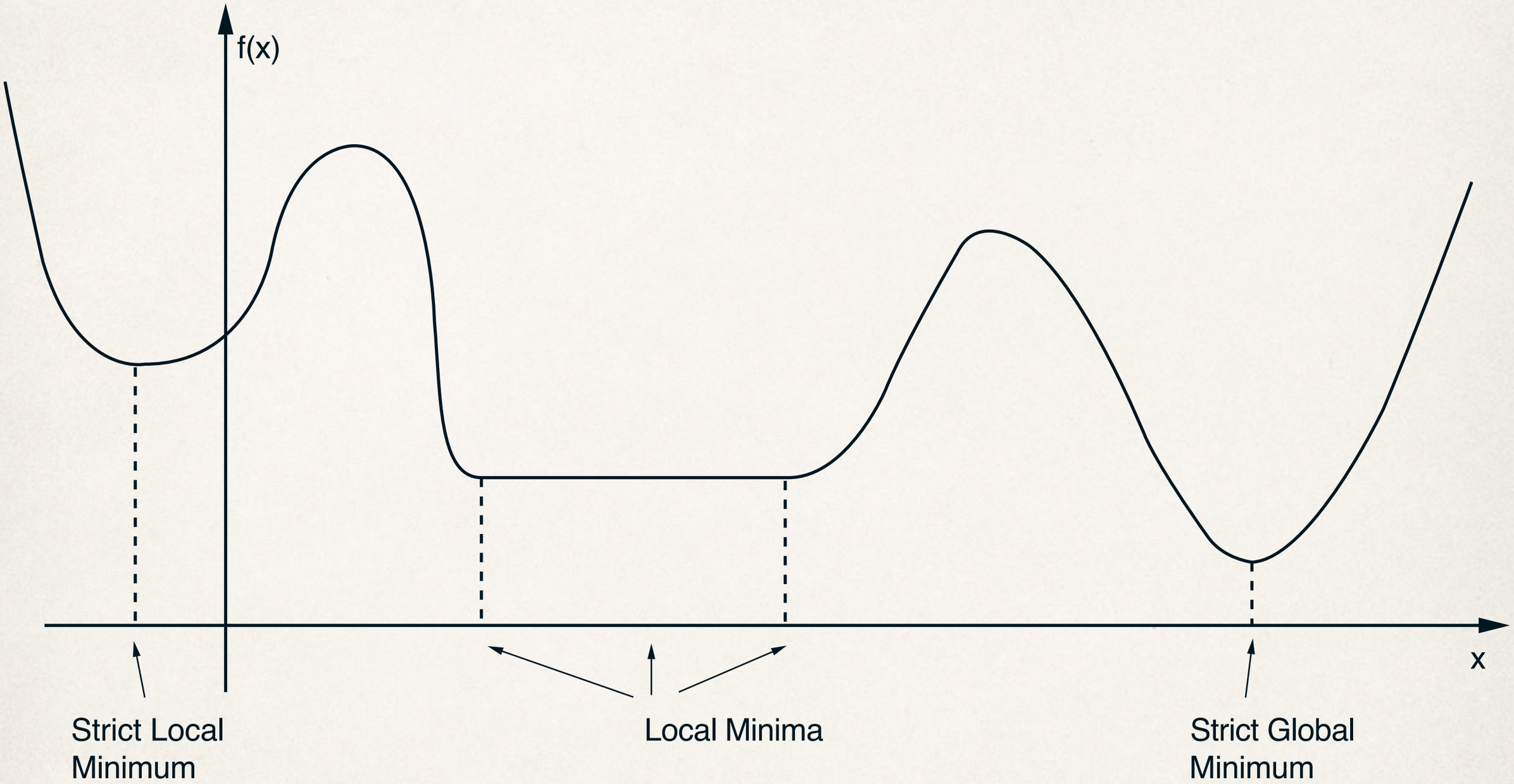
- We can simplify the Lagrangian into the following form:

$$\max_{\alpha} \left(\sum_{i}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle \right)$$

s.t. $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$

Unconstrained Optimization

- We will study one of the simplest and most basic of questions
  - How to minimize (or maximize) a function over Euclidean space?
- We will explore a variety of basic concepts
  - Local vs. global optima
  - Different search algorithms: line search, trust region, conjugate gradients
- The basic concepts will be useful in more complex settings

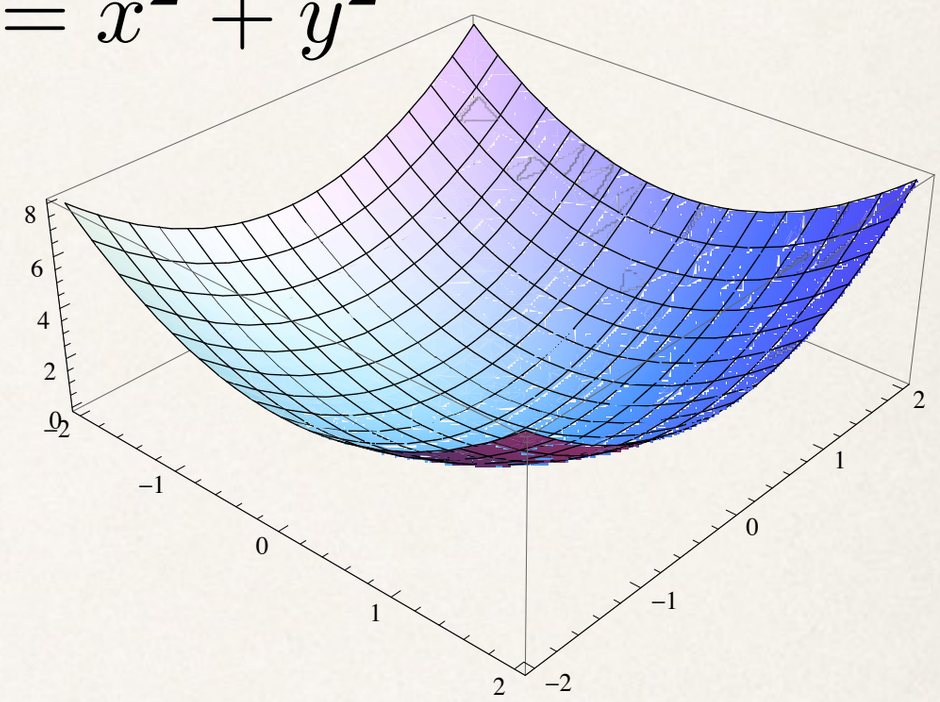

f(x)
x
Strict Local Minimum
Local Minima
Strict Global Minimum

Well-posed optimization problem

- Given an optimization problem, we need to understand whether it is well-posed
- For example:
  - Is there a unique solution?
  - Is there a solution at all?
  - Are there many solutions?

What is a solution?

$$f(x,y) = x^2 + y^2$$



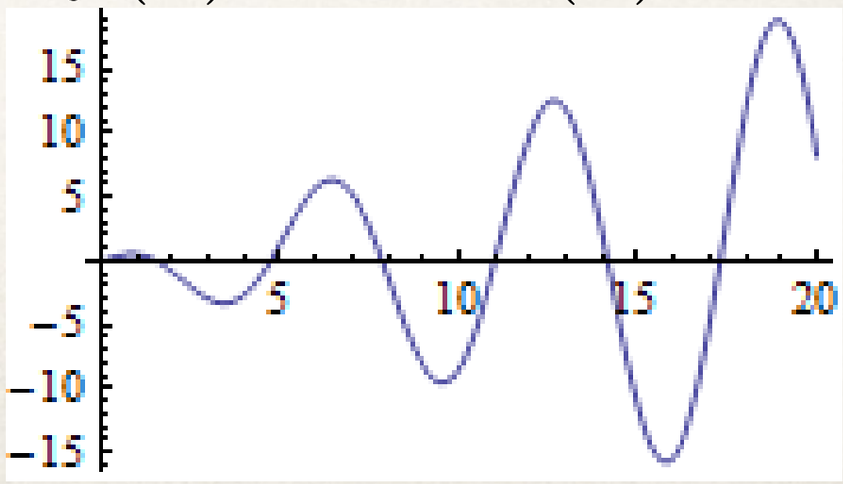
- A point x* is a global minimizer of a function if and only if

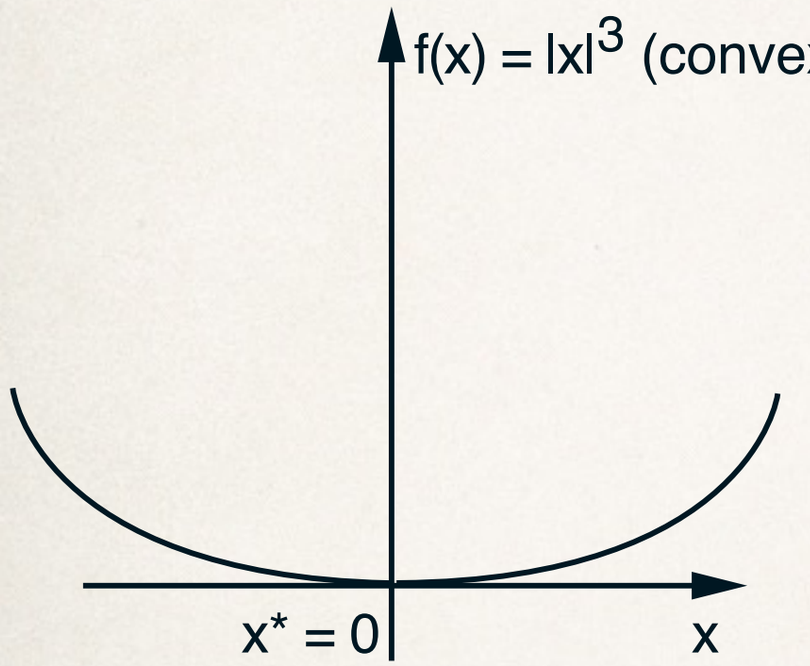
$$f(x^*) \leq f(x) \quad \forall x \in \mathbb{R}^n$$

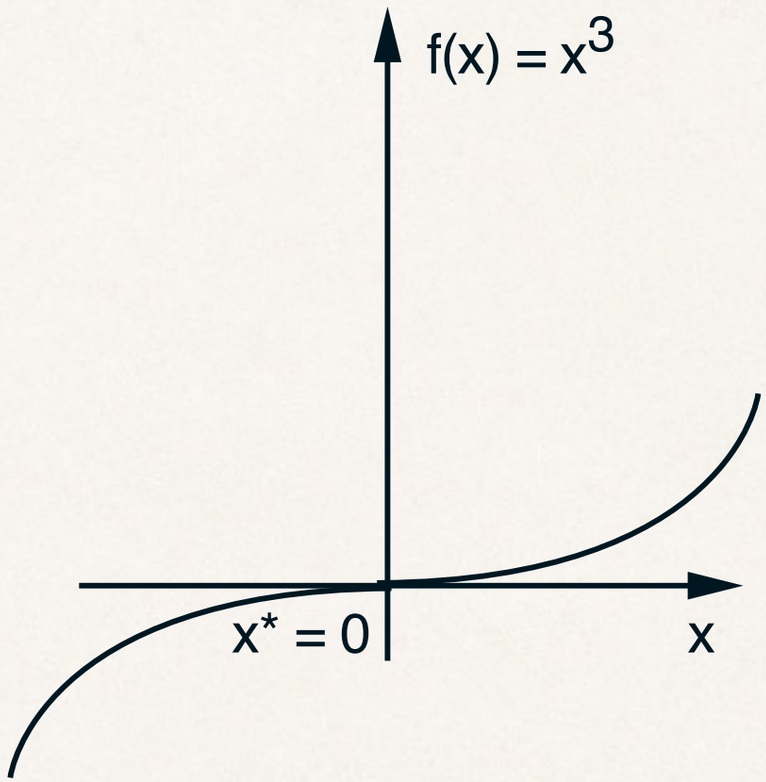
- A point x* is a local minimizer of a function if there is a neighborhood where:

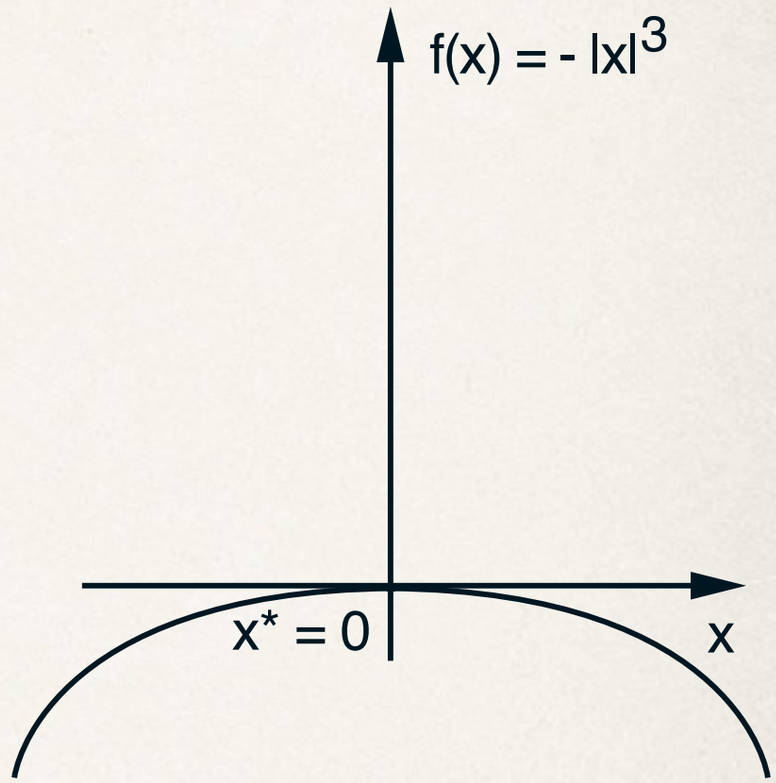
$$f(x^*) \leq f(x), \quad \forall x \in \mathcal{N}$$

$$f(x) = x\cos(x)$$



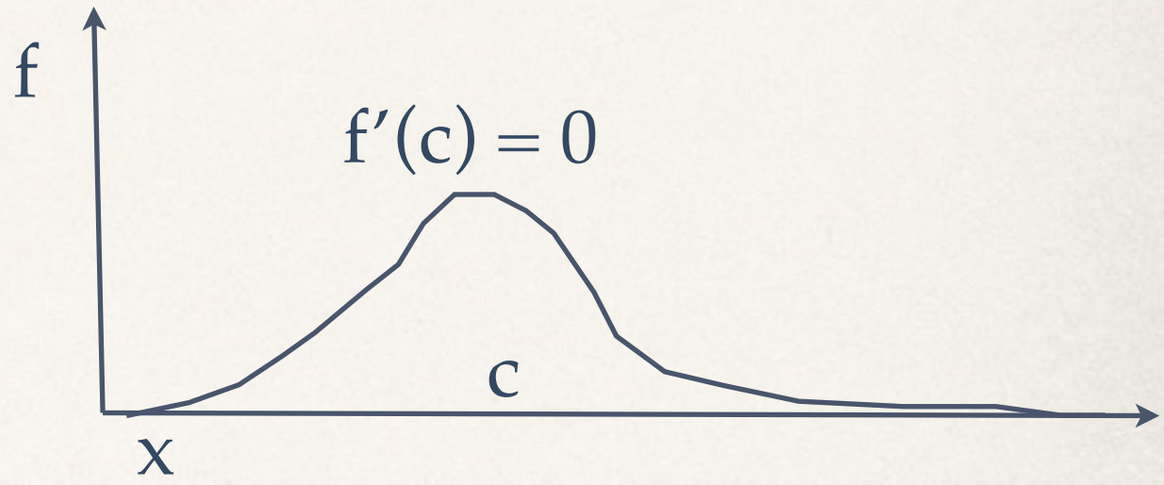

f(x) = |x|3 (convex)
x* = 0
x


f(x) = x3
x* = 0
x

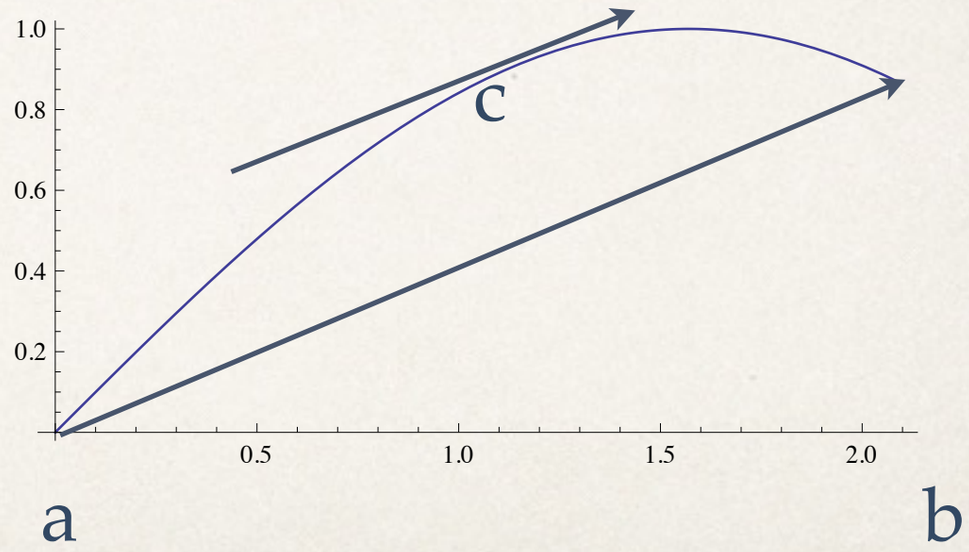

f(x) = - |x|3
x* = 0
x

Basic Theorems on Differentiable Functions

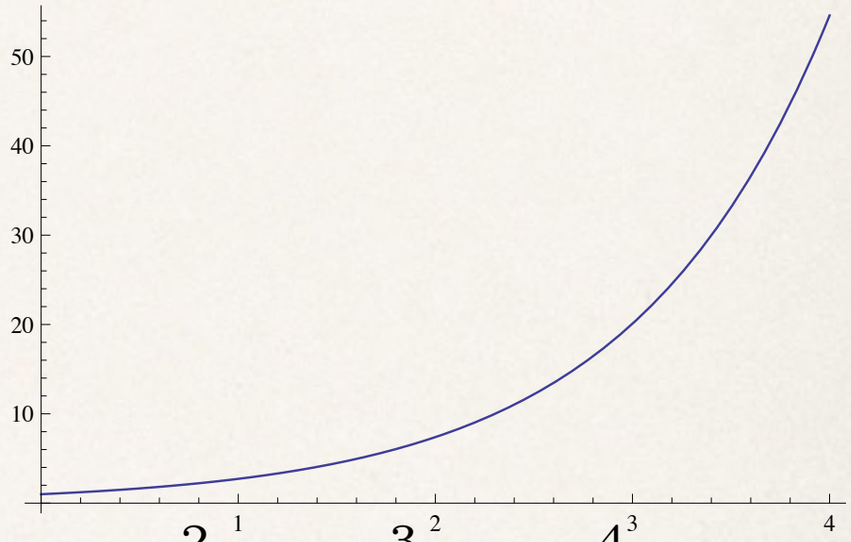
- Rolle's theorem: If f(x) is continuous and differentiable on [a,b], and f(a)=f(b) = 0, then exists c s.t. f'(c) = 0



- Lagrange's (mean-value) theorem: If f(x) is differentiable and continuous on [a,b], then f(b) - f(a) = f'(c) (b - a)



Taylor Expansions



- Any smooth function f(x) that has derivatives up to order n can be expressed as

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + O[x]^5$$

$$f(x) = f(a) + \frac{(x-1}{1!} f'(a) + \frac{(x-a)^2}{2!} f\text{“}(a) + \ldots + \frac{(x-a)^n}{n!} f^n(a) + R_{n+1}$$

Frechet Differential

- For a multivariate function f on Euclidean space, we need to generalize the concept of a derivative

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Jacobian

- If there exists a vector g such that

$$\lim_{\|y\| \to 0} \frac{f(x+y) - f(y) - g^T y}{\|y\|} = 0$$

- we say f is Frechet differentiable at x

Hessian

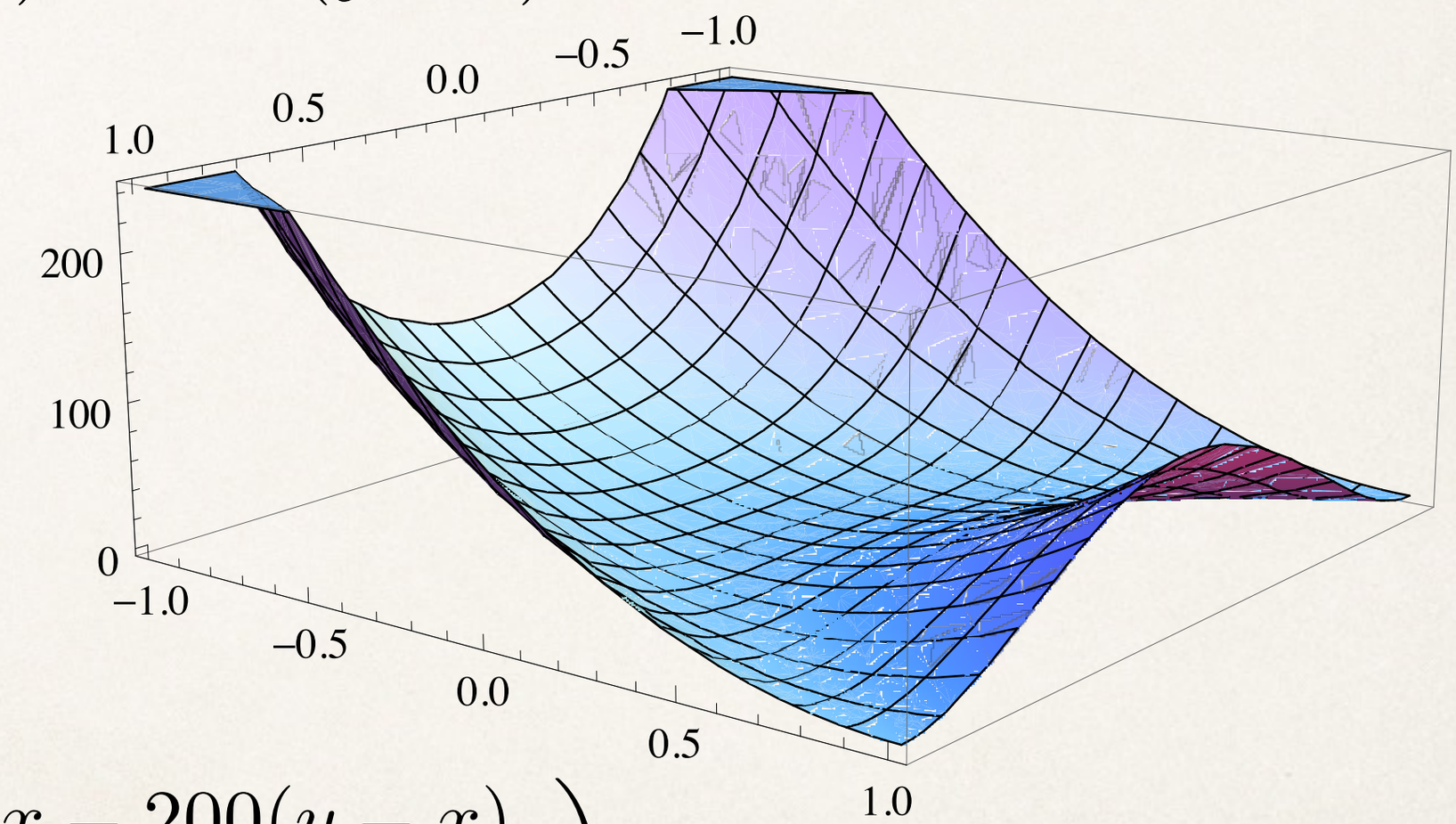
- Many unconstrained optimization methods use the second order derivative of a smooth function

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

- Newton's method
- Modified Newton methods

Rosenbrock Problem

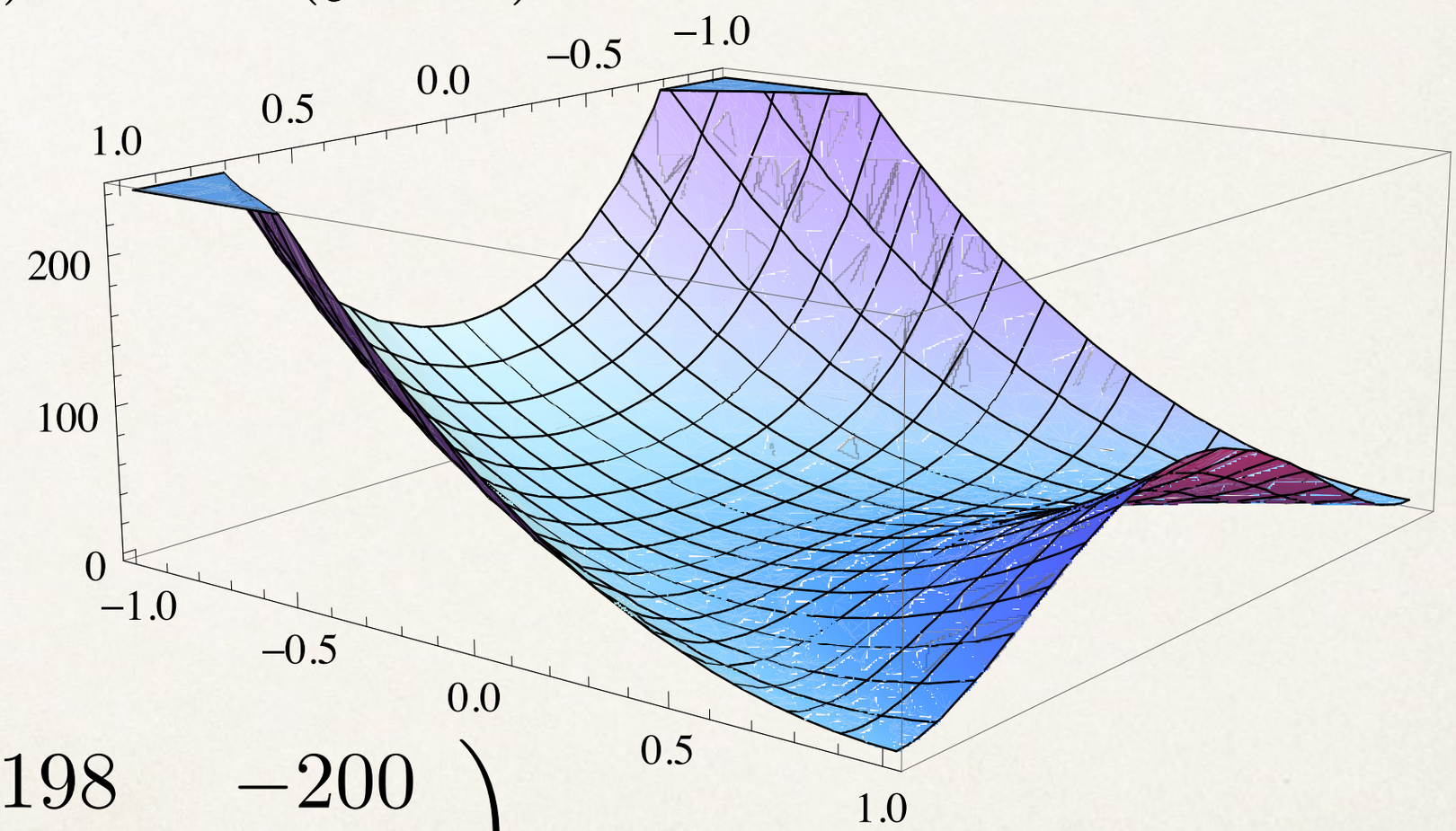
$$f(x,y) = (1-x)^2 + 100(y-x)^2$$



$$\nabla f(x,y) = \begin{pmatrix} -2x - 200(y-x) \\ 200(y-x) \end{pmatrix}$$

Rosenbrock Problem

$$f(x,y) = (1-x)^2 + 100(y-x)^2$$



$$\nabla^2 f(x,y) = \begin{pmatrix} 198 & -200 \\ -200 & 200 \end{pmatrix}$$

Vector-valued functions

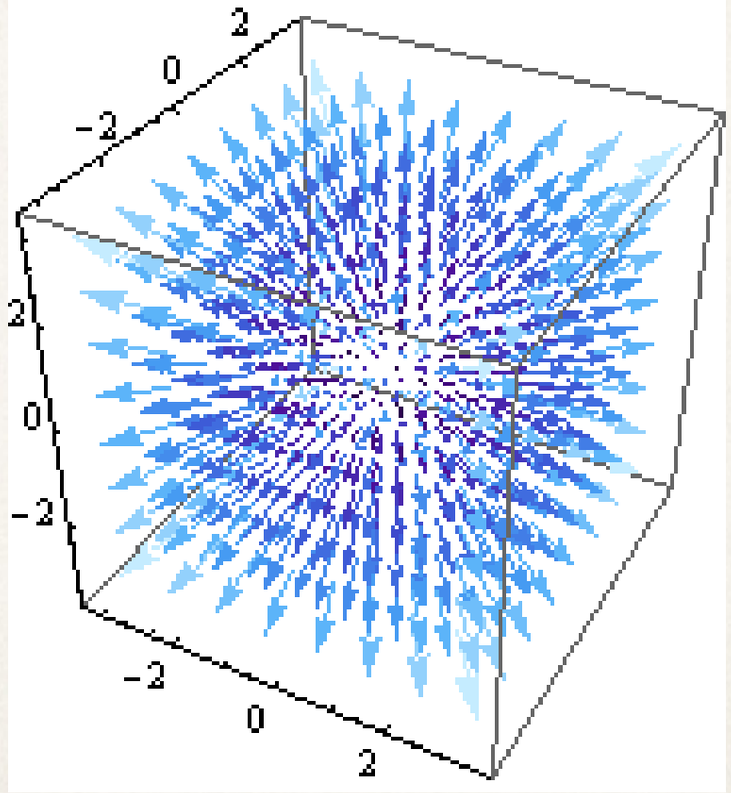
- A vector-valued function is a vector of 1-dimensional functions

$$\mathbf{f}(\mathbf{x}) = (f_1(x), \ldots, f_m(x))$$

- Its Jacobian is an m x n matrix

$$\nabla \mathbf{f}(\mathbf{x}) = [\frac{\partial f_i(x)}{\partial x_j}]$$

- Its Hessian is a tensor



Example: Solving linear equations

- Recall the system of linear equations example
- Using differentiation by parts, we can easily show that

$$\nabla f(x) = \frac{\partial f}{\partial x} = \frac{1}{2}A^T x + \frac{1}{2}Ax - b$$

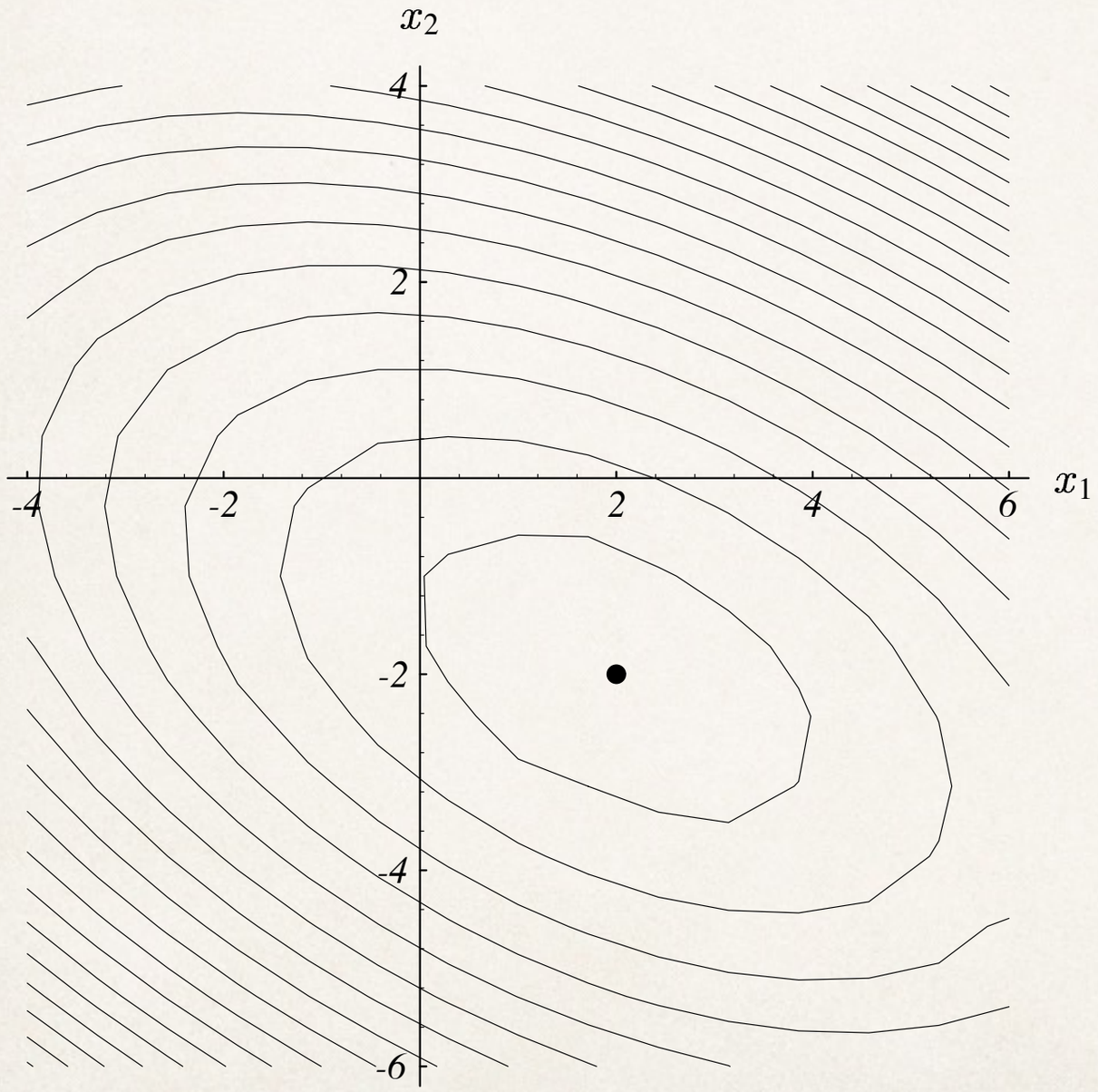
- Note its Hessian is just A!

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c$$

If A is symmetric

$$\frac{\partial f}{\partial x} = Ax - b$$

Quadratic Form: Linear Equation Solving



$$f(x) = \frac{1}{2}x^T Ax - b^T x + c$$

Introduction to
the conjugate gradient
method without
the agonizing pain,
Shewchuck,
CMU CS TR, 1994

Positive Definite Matrices

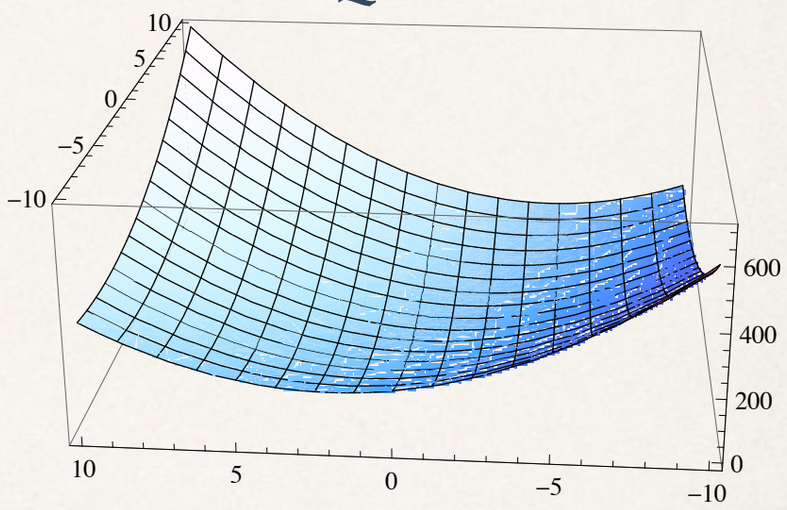
- A matrix A is positive definite if for all non-zero vectors, x:

$$x^T Ax > 0$$

- This implies all its eigenvalues are positive and real

- Eigenvalues of A: 7, 2

Quadratic form



$$\begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}$$

Taylor's Theorem: n-dimensions

- Let f be a continuously differentiable function. Then:

$$f(x+p) = f(x) + \nabla f(x+tp)^T p, \;\; t \in \{0,1\}$$

- Also, for the Jacobian:

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp) p dt$$

- Interpolation rule:

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

Consequences of Taylor's Theorem

- First-order necessary conditions:
  - If x^* is a local minimizer, then the Jacobian of f at $x^* = 0$
- Second-order necessary conditions:
  - If x^* is a local minimizer, then the Jacobian of f at $x^*=0$ and the Hessian at x^* is PSD
- Second-order sufficient conditions:
  - If the Jacobian at $x^* = 0$ and the Hessian at x^* is positive definite, then x^* is a strict local minimizer

Algorithms: Unconstrained Minimization

- Line search
  - Pick a search direction p
  - Pick t s.t. f(x + pt) is minimized
- Trust region
  - Pick a region size and build a quadratic local model
- Find the optimal direction
- Conjugate gradient search
  - Pick a set of orthogonal directions to search
  - Move in each direction once!
  - Search converges in n steps

Line Search

- Line search picks a direction p to move in
- What is the best direction? $$p = \frac{-\nabla f(x_k)}{\|\nabla f(x_k)\|}$$
- Use Taylor's theorem

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f(x_k) + \frac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + tp)p, \ t \in \{0, \alpha\}$$

- This implies minimizing $\min_p p^T \nabla f(x_k), \ s.t. \|p\| = 1$

$$p^T \nabla f(x_k) = \|p\| \|\nabla f(x_k)\| \cos(\theta)$$

Line Search: Method of Steepest Descent

- The line search algorithm of choice typically moves in the direction of steepest descent

- Recall for solving the system of linear equations Ax = b
  - The gradient $\nabla f(x) = \frac{\partial f}{\partial x} = \frac{1}{2}A^T x + \frac{1}{2}Ax - b$
  - When A is symmetric, this equals Ax - b
  - Steepest descent direction = -f'(x(k)) = b - A x(k)

Step size: Line Search

- Once the direction is chosen (steepest descent), the next issue is step size selection

- Once again, we illustrate step size selection for solving Ax = b

- Consider the same example: $\begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -8 \end{pmatrix}$

- Suppose we start at x(0) = [-2, -2]’

- How to choose step size t such that x(1) = x(0) - t f’(x(0)) = x(0) + t r(0)

- Define r(0) = b - Ax(0) (the initial residual error)

Step Size: Line Search

- We need to compute the gradient of f(x(1)) w.r.t step size t
- We use the chain rule to determine t

$$\frac{\partial f(x_1)}{\partial t} = \frac{\partial f(x_1)}{\partial x_1}\frac{\partial x_1}{\partial t} = f'(x_1)^T r_0 = -r^T r_0$$

- In other words, set step size t such that the new residual r(1) is orthogonal to the original residual r(0)

Step Size Computation: Line Search

$$r_1^T r_0 = 0$$

$$(b - Ax_1)^T r_0 = 0$$

$$(b - A(x_0 + tr_0))^T r_0 = 0$$

$$(b - Ax_0)^T r_0 - t(Ar_0)^T r_0$$

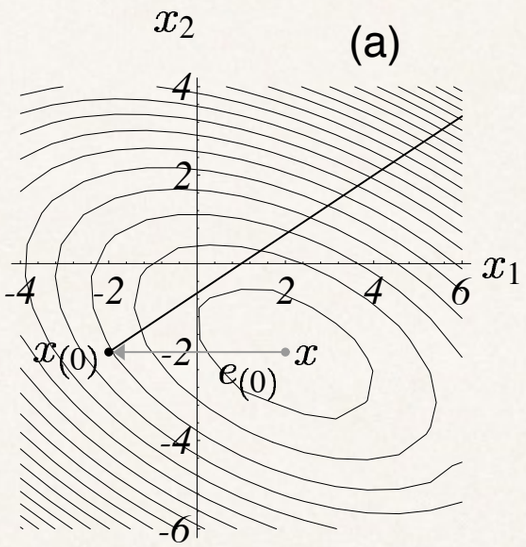
$$t = \frac{r_0^T r_0}{r_0^T A r_0}$$

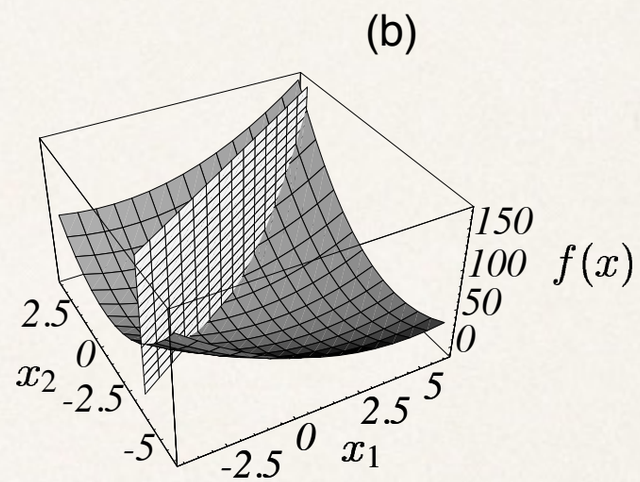
Steepest Descent Algorithm

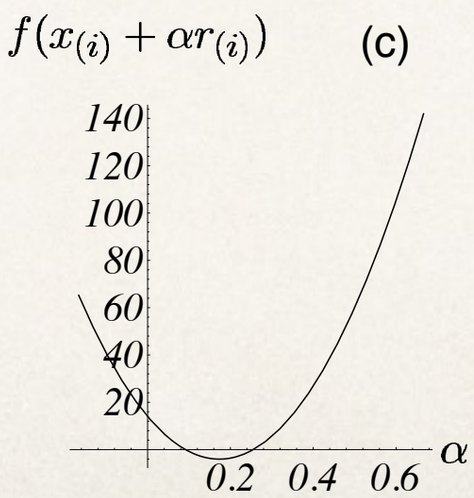
$$r_i = b - Ax_i$$

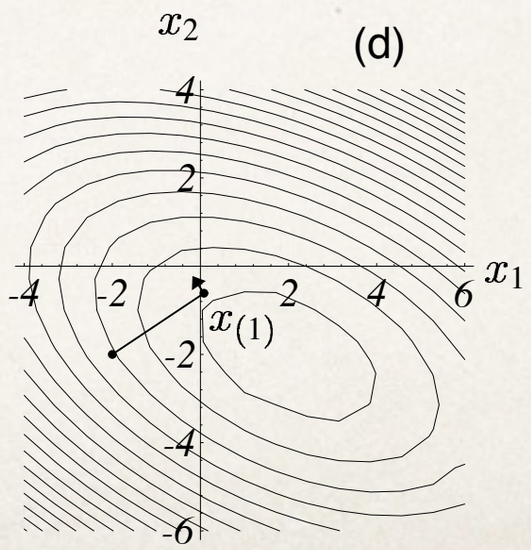
$$t_i = \frac{r_i^T r_i}{r_i^T A r_i}$$

$$x_{i+1} = x_i + t_i r_i$$

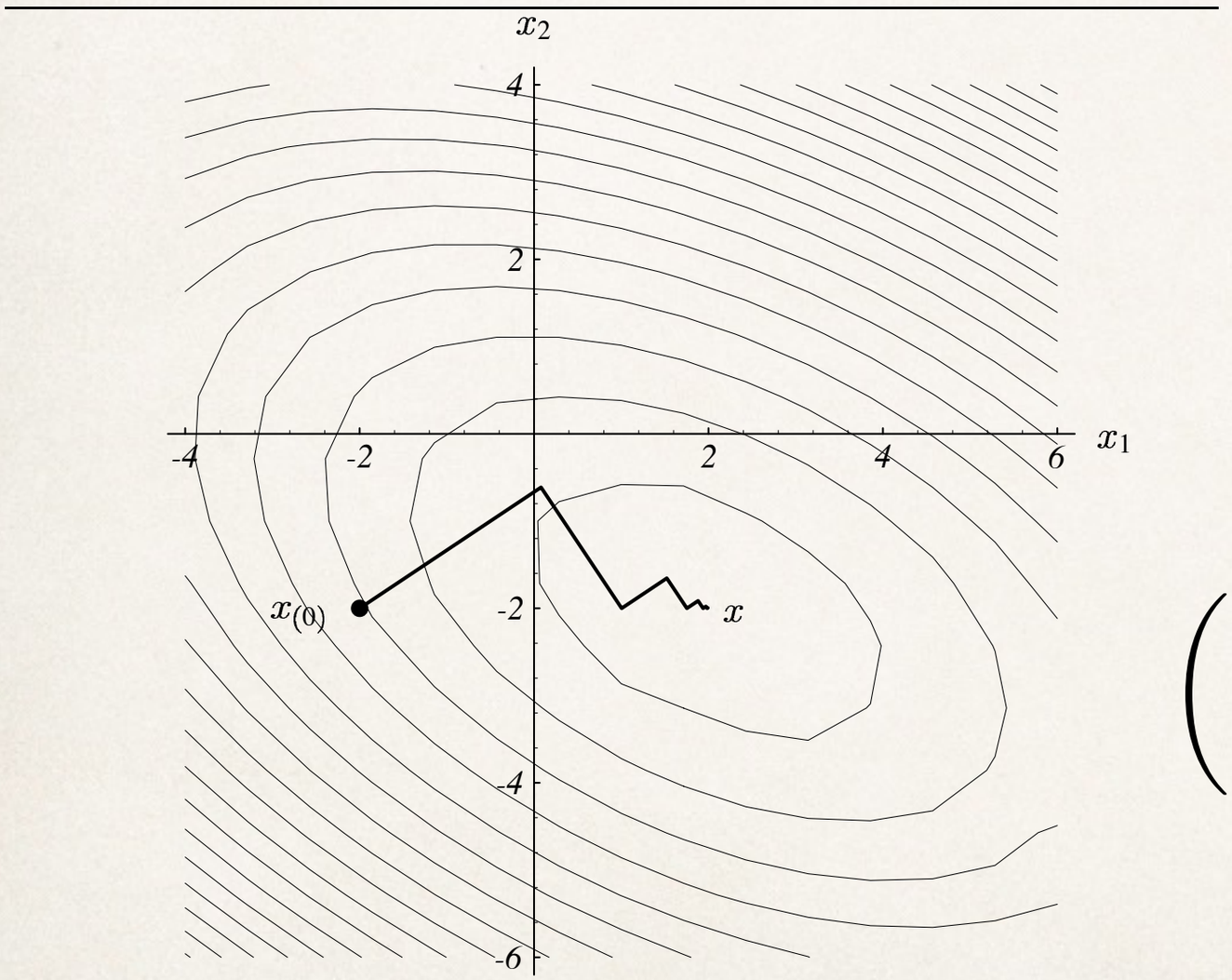








Steepest Descent Algorithm



$$\begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -8 \end{pmatrix}$$

Convergence Analysis: Steepest Descent for solving Ax=bs

- Define the error e(i) = x(i) - x (where x is the true solution)
- Recall the residual r(i) = b - A x(i) = -A e(i)
- Express the error in terms of the eigenvector basis
- Then, we have $e_i = \sum_{i=1}^{n} \xi_i v_i$

Convergence Analysis: Steepest Descent

$$x_{i+1} = x_i + t_i r_i$$

$$x_{i+1} - x = x_i - x + t_i r_i$$

$$e_{i+1} = e_i + t_i r_i$$

$$e_{i+1} = e_i + \frac{r_i^T r_i}{r_i^T A r_i} r_i$$

Convergence Analysis: Steepest Descent

$$r_i = -Ae_i = -\sum_i \xi_i \lambda_i v_i$$

$$e_i^T e_i = \sum_i \xi_i^2$$

$$e_i^T A e_i = (\sum_i \xi_i v_i^T)(\sum_j \xi_j \lambda_j v_j) = \sum_i \xi_i^2 \lambda_j$$

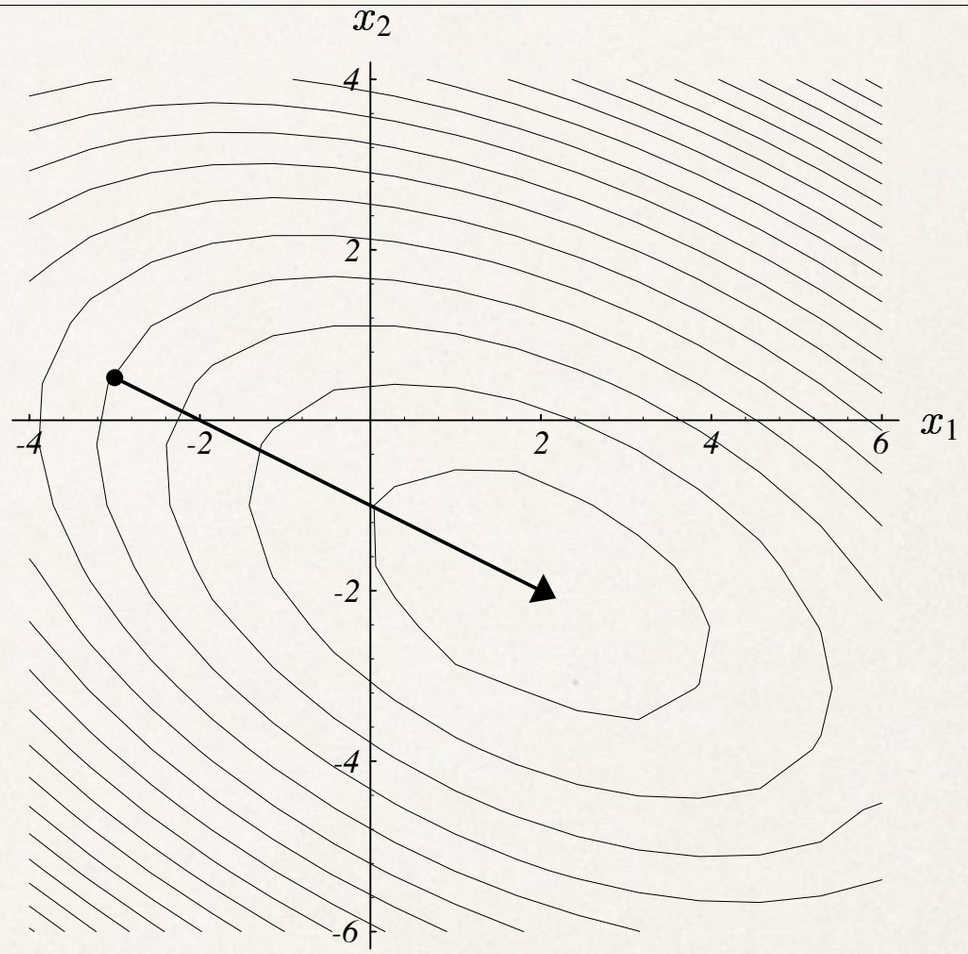
$$\|r_i\|^2 = r_i^T r_i = \sum_i \xi_j^2 \lambda_j^2$$

$$r_i^T A r_i = \sum_i \xi_i^2 \lambda_i^3$$

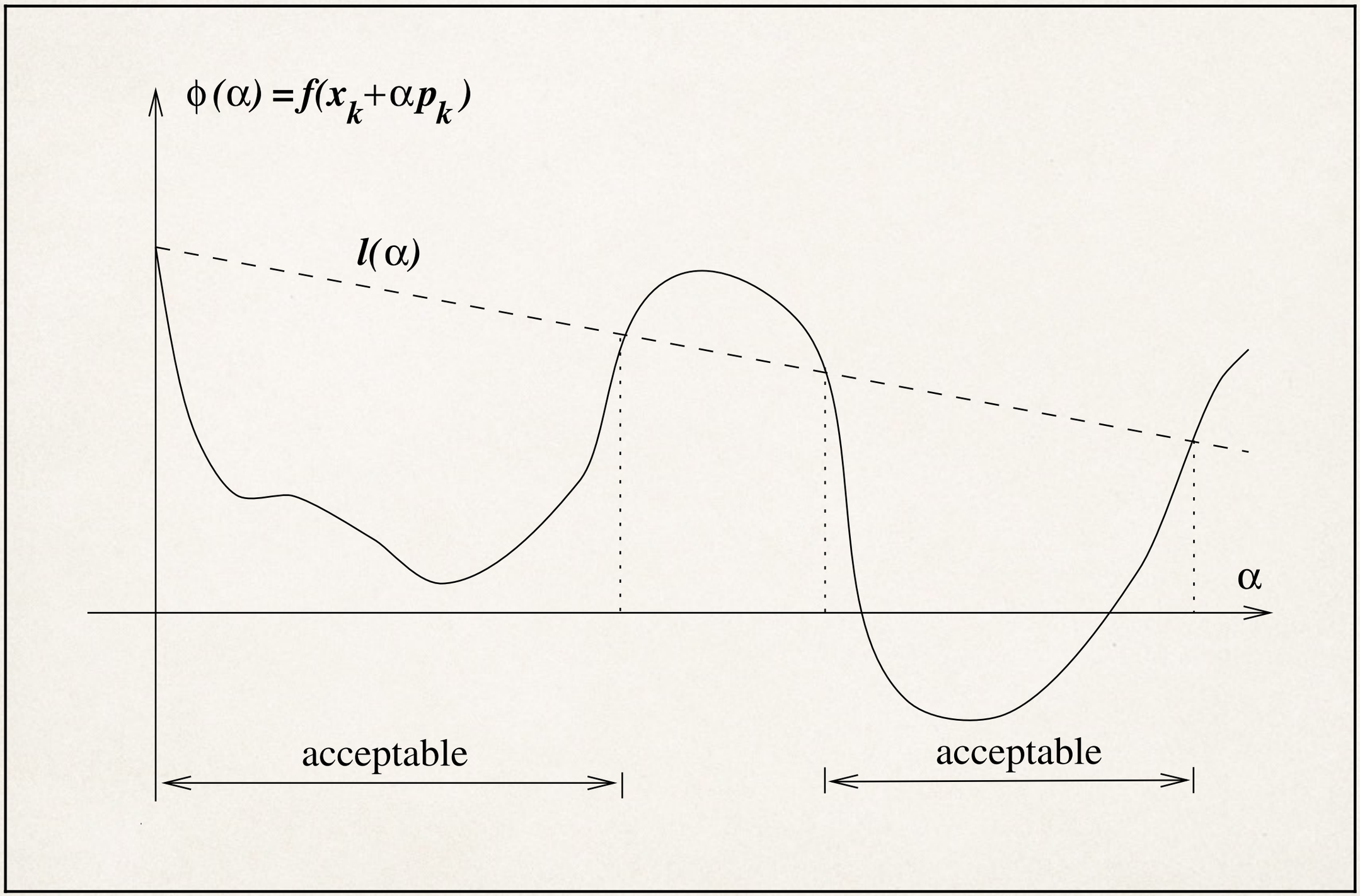
Convergence Analysis: Summary

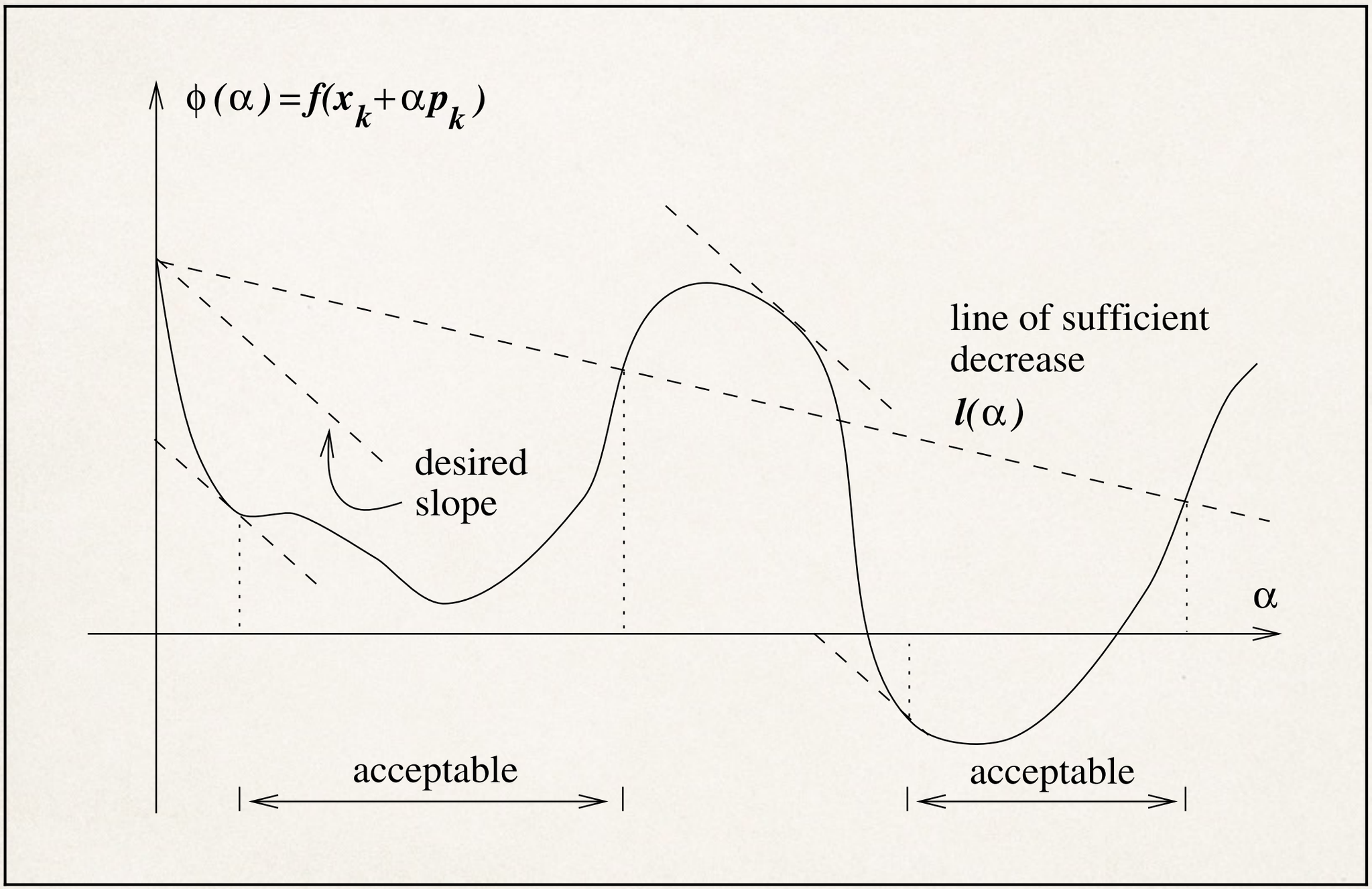
$$
\begin{aligned}
e_{i+1} &= e_i + \frac{r_i^T r_i}{r_i^T A r_i} r_i \\
&= e_i + \frac{\sum_i \xi_i^2 \lambda_i^2}{\sum_i \xi_i^2 \lambda_i^3} r_i
\end{aligned}
$$

In general, the error will not be aligned with the eigenvectors.

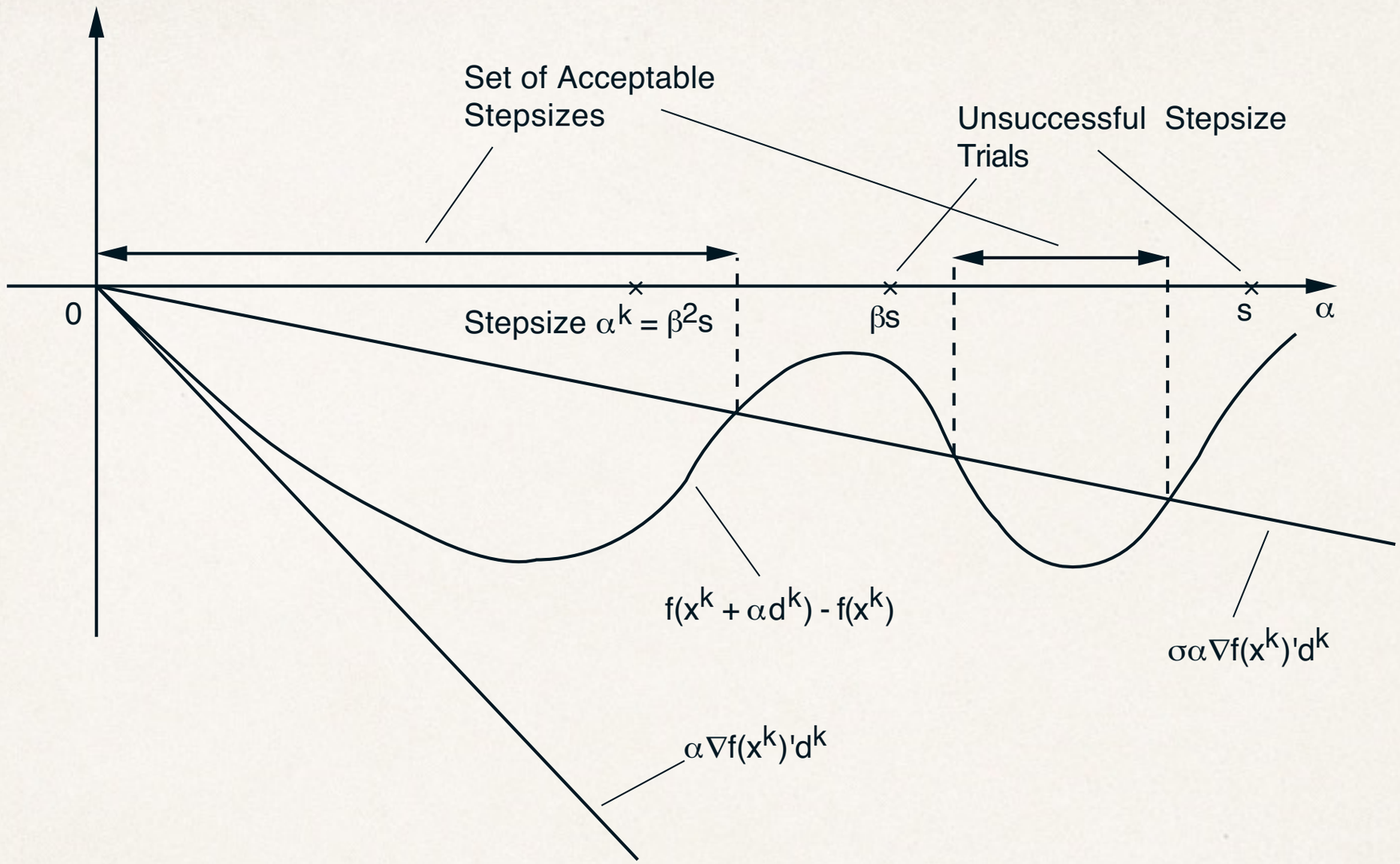


Steepest descent converges in one step if the error term is an eigenvector!


$\phi(\alpha) = f(x_k + \alpha p_k)$
$l(\alpha)$
α
acceptable
acceptable


$\phi(\alpha) = f(x_k + \alpha p_k)$
line of sufficient
decrease
$l(\alpha)$
desired
slope
α
acceptable
acceptable

- Armijo rule:



Start with s and continue with $\beta s, \beta^2 s, ...,$ until $\beta^m s$ falls within the set of α with

$$f(x^k) - f(x^k + \alpha d^k) \geq -\sigma\alpha\nabla f(x^k)' d^k .$$

Wolfe Conditions

reduction in f should be commensurate with step length

$$f(x_k + tp_k) \leq f(x_k) + c_1 t \nabla f(x_k)^T p_k$$

Armijo condition

Step length should be dictated by curvature

$$\nabla f(x_k + t_k p_k)^T p_k \geq c_2 \nabla f(x_k)^T p_k$$

Curvature condition

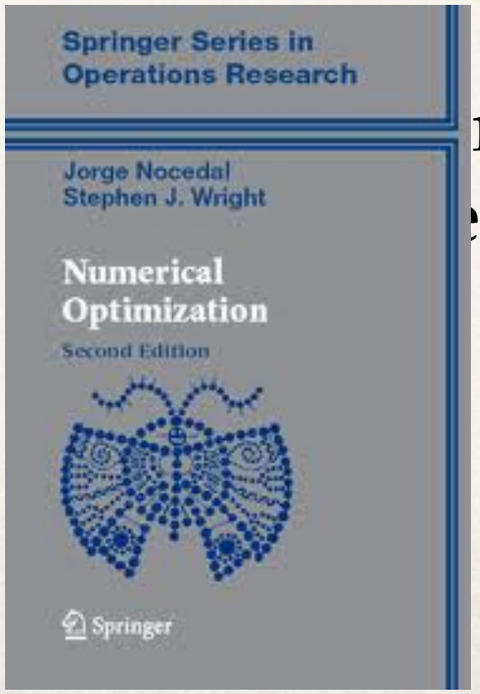
Summary

- Unconstrained optimization provides a foundation for more advanced methods
- Optimization algorithms
- ✓ Line search
  - Trust region
  - Conjugate gradient

Reading Assignment

- Chapter 2 and 3 in Nocedal and Wright covers unconstrained optimization and line search
- Sections 9.1-9.4 in Boyd and Vanderberghe's book covers similar material
- Suggested programming assignment
- Implement a line search method for solving Ax=b
- Next lecture
- Newton's method

njug
ethoc



Reading Assignment

- Conjugate Duality is covered in Chapter 3 of Boyd and Vandenberghe

- Lagrange Duality is covered in Chapter 5 of Boyd and Vandenberghe

- Read article on Hahn-Banach theorem on Moodle

- Work out all the examples in this lecture