Convex Optimization: Part II

CMPT 419/983

Mo Chen

SFU Computing Science

23/09/2018

Textbook

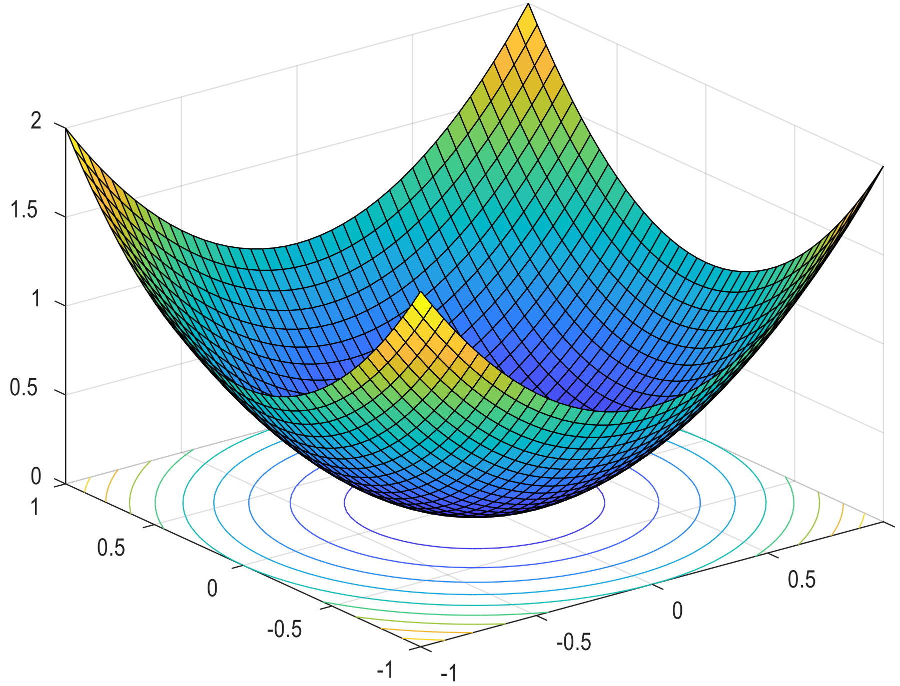
- S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2008.

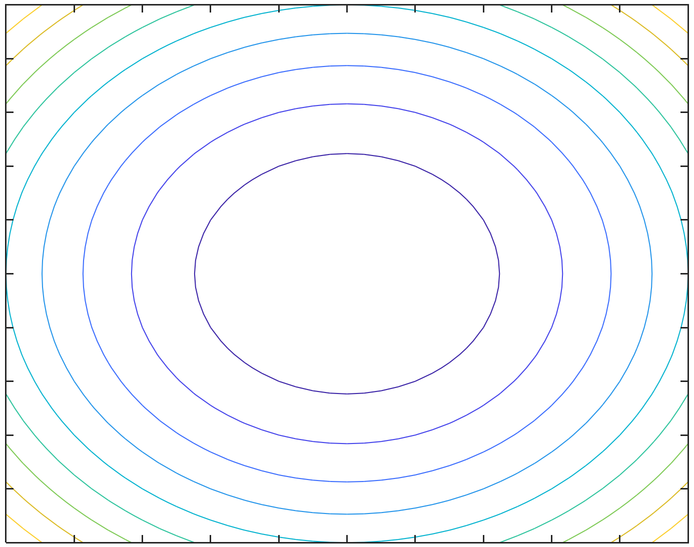
Outline

- Optimization program
  - Examples and classes
- Convex optimization
  - Convex functions
  - Optimality conditions
- Numerical solutions

Optimality Conditions for Convex Programs

- Unconstrained case: $\text{minimize} \quad f(x)$
  - $\nabla f(x) = 0$





Optimality Conditions for Convex Programs

- Inequality constraints only:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g_i(x) \le 0, i = 1, \dots, n \end{aligned}$$

- Penalty view point: penalize constraint violation
  - Lagrangian: $L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x)$, $\lambda_i \ge 0$

- Optimality conditions
  - Stationarity: $\nabla_x L(x^*, \lambda^*) = 0$
  - Primal feasibility: $g_i(x^*) \le 0$
  - Dual feasibility: $\lambda^* \ge 0$
  - Complementary slackness: $\lambda_i^* g_i(x^*) = 0$, $i = 1, \dots, n$

Optimality Conditions for Convex Programs

- Stationarity: $\nabla_x L(x^*, \lambda^*) = 0$
  - Lagrangian:

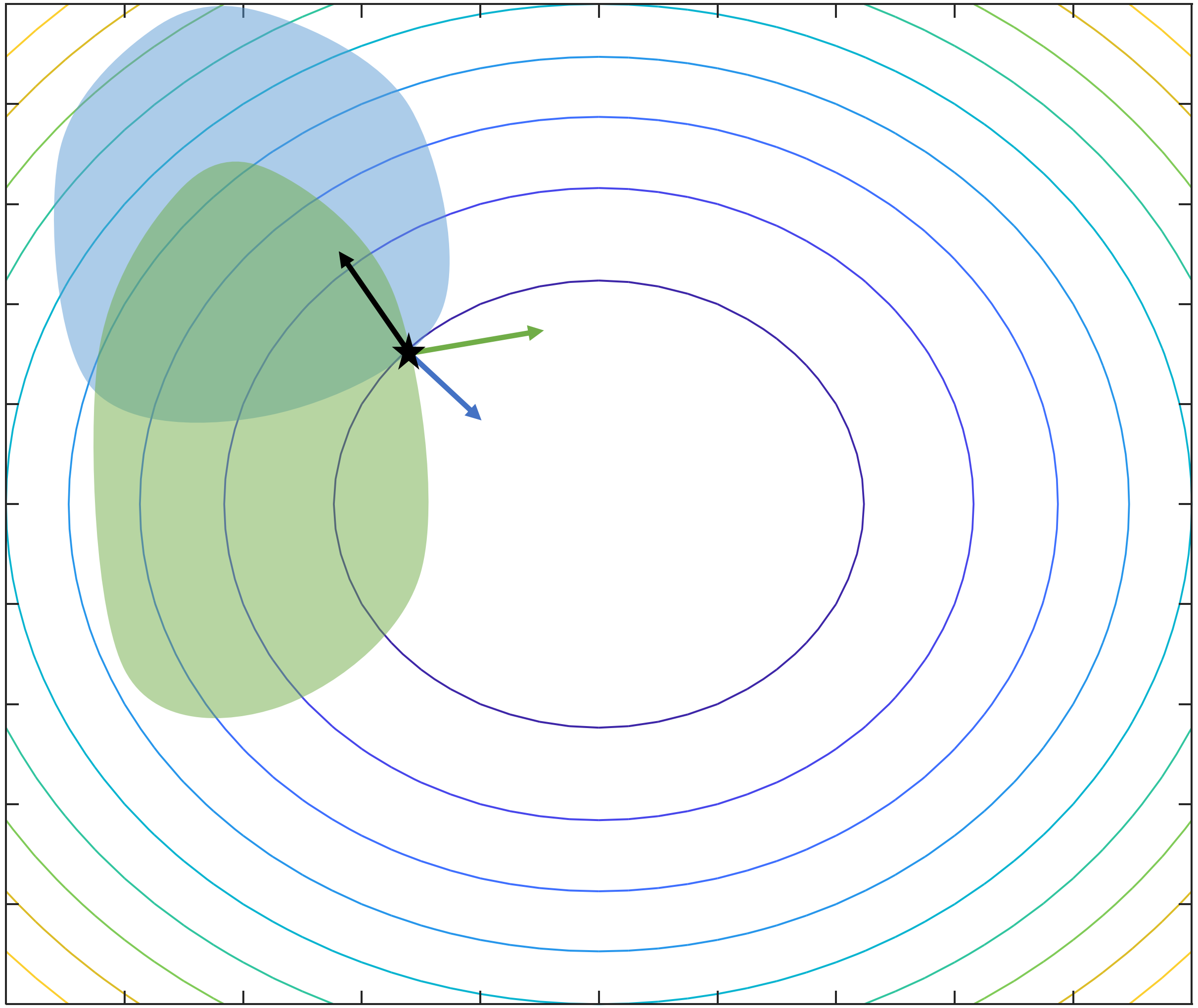
$$L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x), \qquad \lambda_i \geq 0$$

  - Take gradient and set to zero:

$$0 = \nabla f(x) + \sum_{i=1}^{n} \lambda_i \nabla g_i(x)$$

$$\nabla f(x) = -\sum_{i=1}^{n} \lambda_i \nabla g_i(x)$$

  - Since $\lambda_i \geq 0$, gradient of $f(x)$ must point "away" from gradients of active constraint functions



Optimality Conditions for Convex Programs

- Stationarity: $\nabla_x L(x^*, \lambda^*) = 0$
  - Lagrangian:

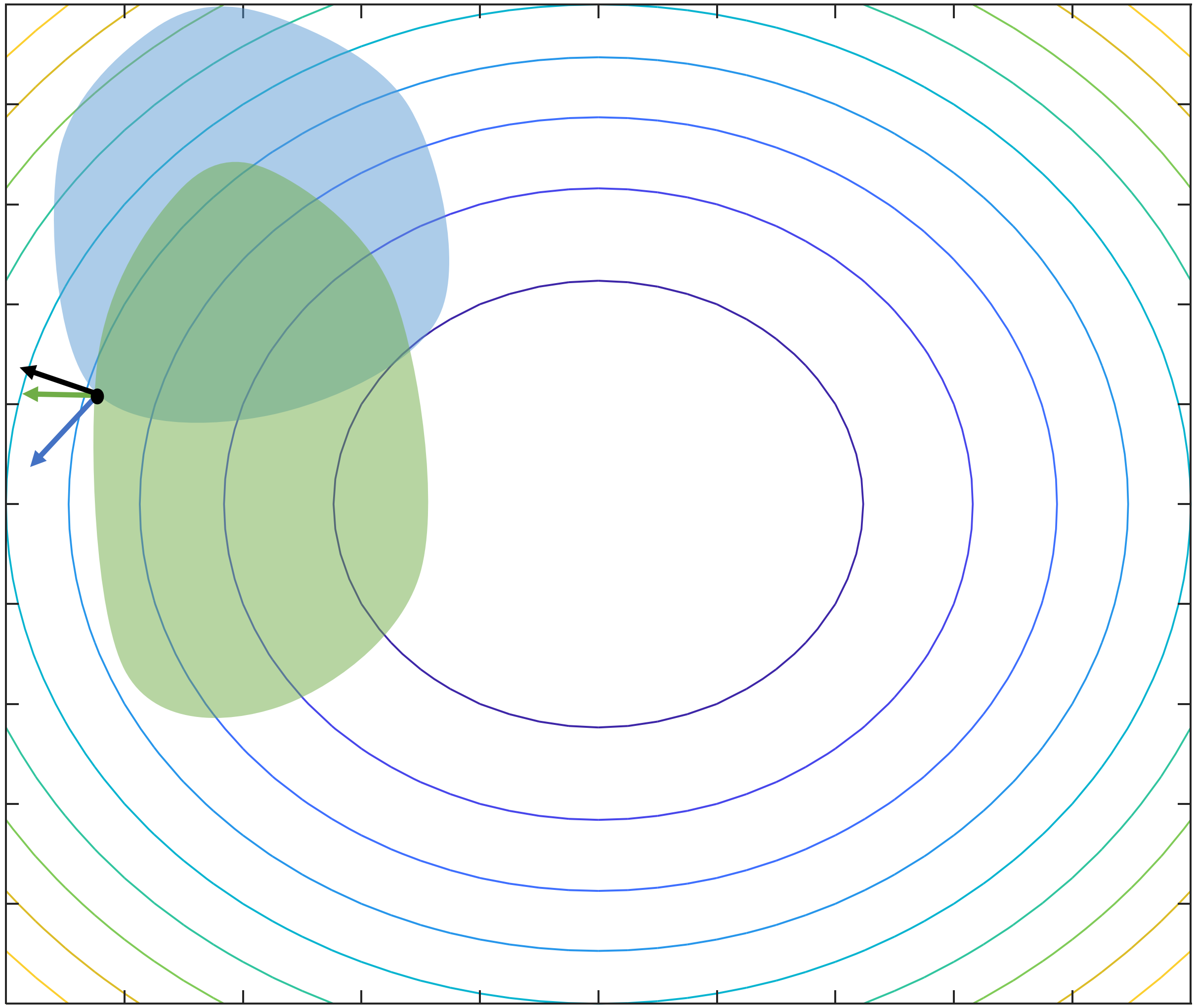
$$L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x), \qquad \lambda_i \geq 0$$

  - Take gradient and set to zero:

$$0 = \nabla f(x) + \sum_{i=1}^{n} \lambda_i \nabla g_i(x)$$

$$\nabla f(x) = -\sum_{i=1}^{n} \lambda_i \nabla g_i(x)$$

  - Since $\lambda_i \geq 0$, gradient of $f(x)$ must point "away" from gradients of active constraint functions



Optimality Conditions for Convex Programs

- Stationarity: $\nabla_x L(x^*, \lambda^*) = 0$
  - Lagrangian:

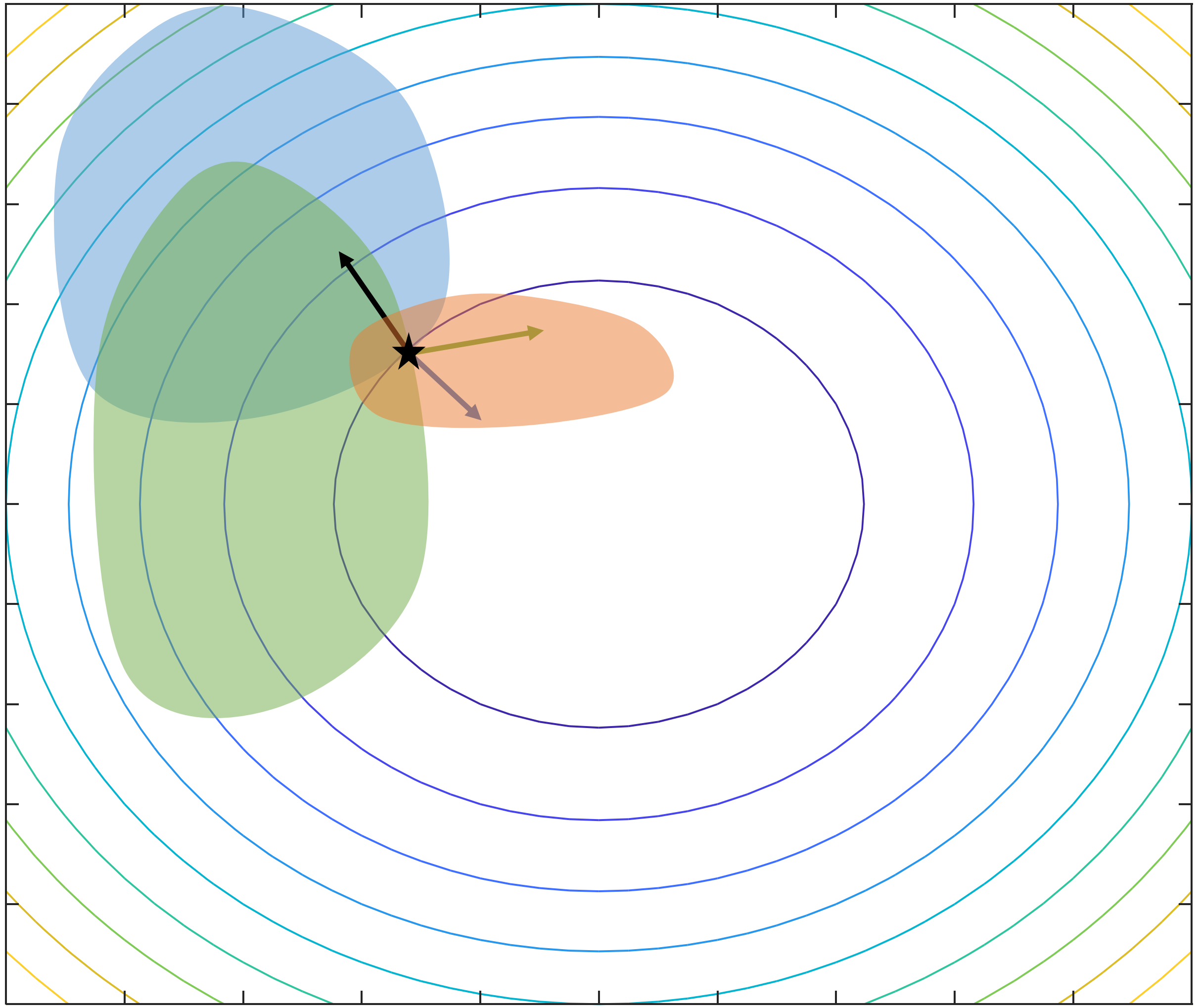
$$L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x), \qquad \lambda_i \geq 0$$

  - Take gradient and set to zero:

$$0 = \nabla f(x) + \sum_{i=1}^{n} \lambda_i \nabla g_i(x)$$

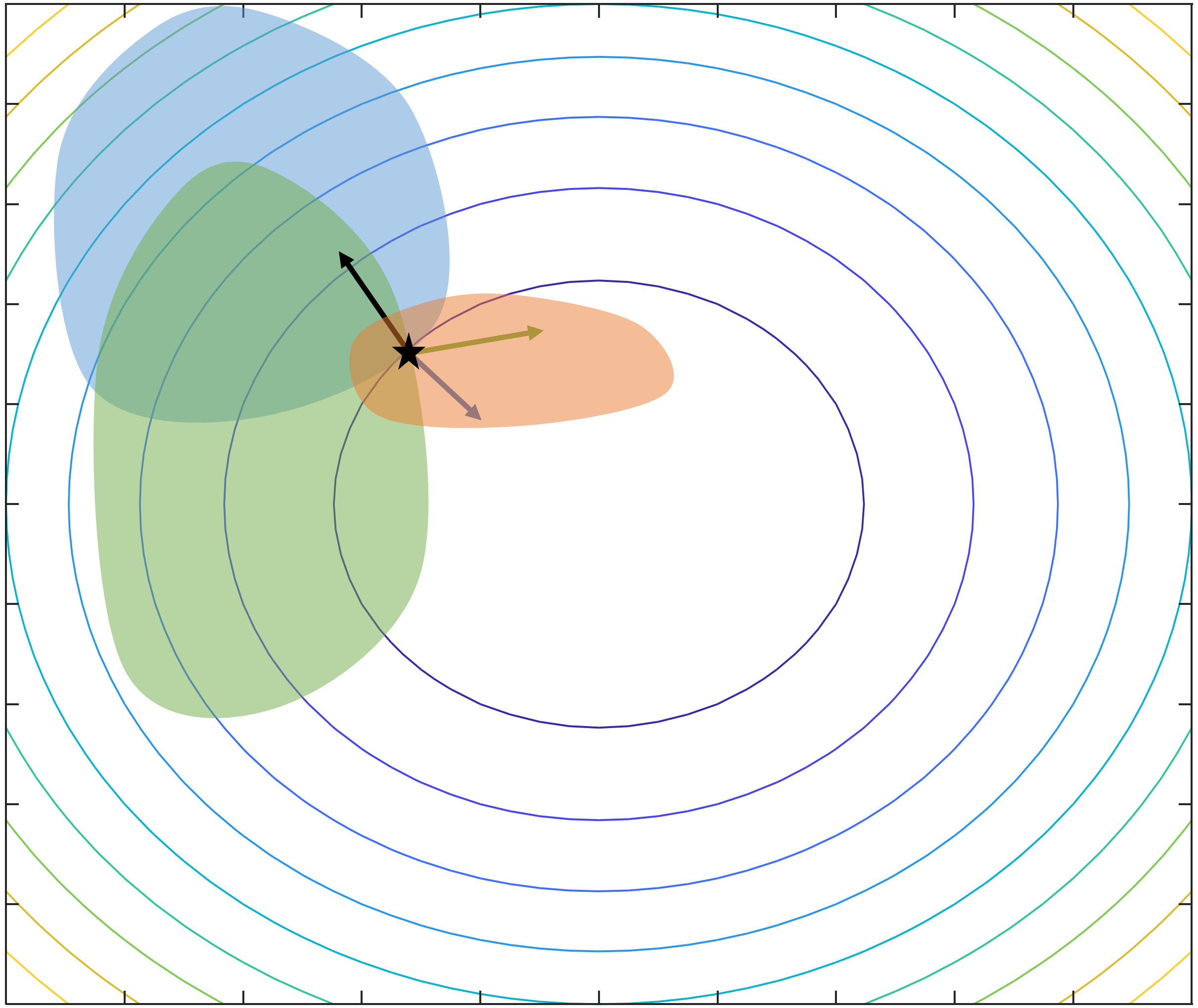
$$\nabla f(x) = -\sum_{i=1}^{n} \lambda_i \nabla g_i(x)$$

  - Since $\lambda_i \geq 0$, gradient of $f(x)$ must point “away” from gradients of **active constraint functions**



Optimality Conditions for Convex Programs

- Primal feasibility: $g_i(x^*) \leq 0$
  - Constraints must be satisfied
- Dual feasibility: $\lambda^* \geq 0$
  - Penalty view point

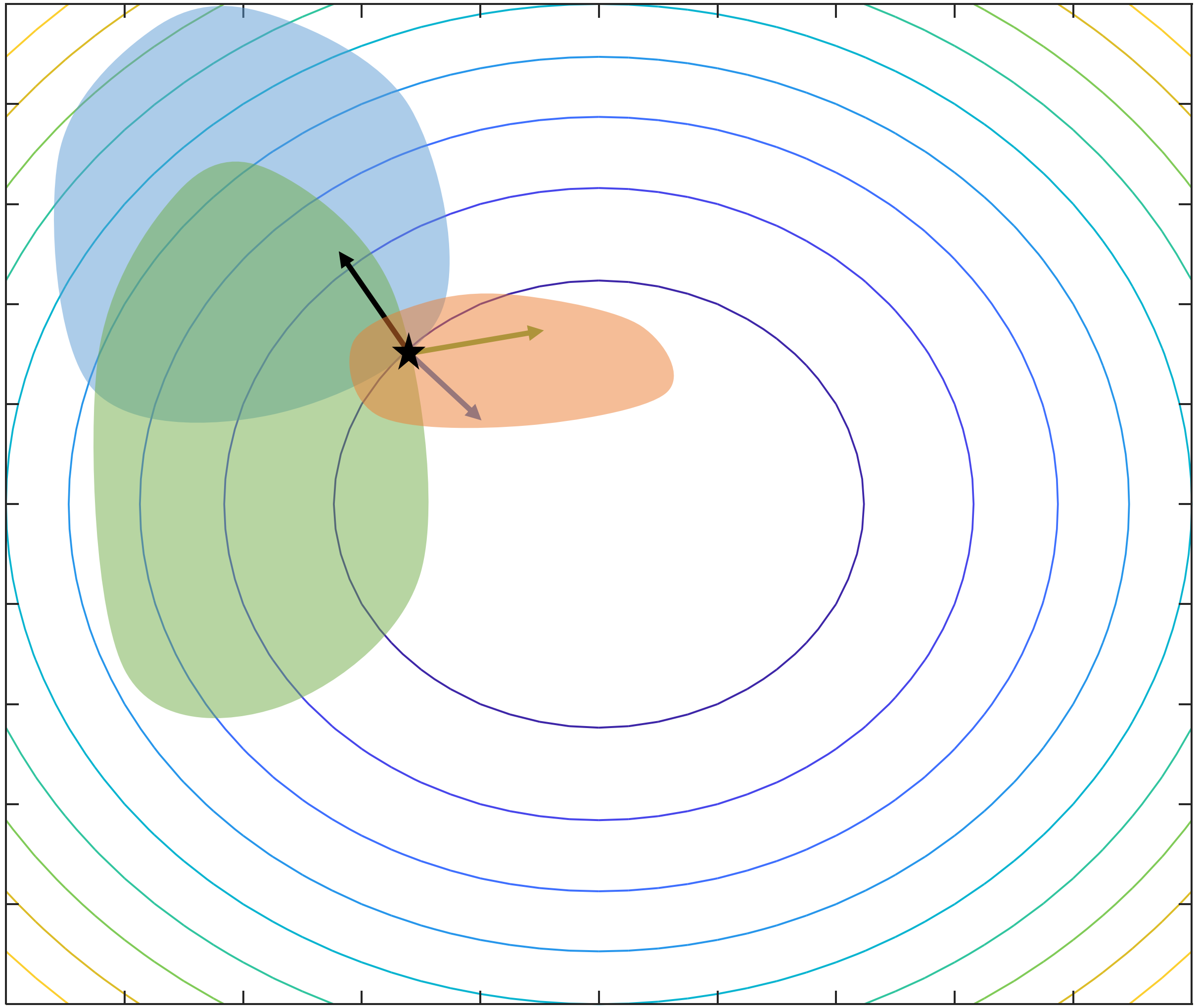


Optimality Conditions for Convex Programs

- Complementary slackness: $\lambda_i^* g_i(x^*) = 0, \ i = 1, \ldots, n$
  - Lagrangian:

$$L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x), \qquad \lambda_i \geq 0$$

  - If $g_i(x^*) < 0$, then the constraint is not active, so λ_i^* is set to 0 to not decrease the Lagrangian
  - If $g_i(x^*) = 0$, then the constraint is active, so λ_i^* is free to be positive



Optimality Conditions for Convex Programs

- Full optimization problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0, i = 1, \dots, n \\ & a_j^\top x = b_j, j = 1, \dots, m \end{array}$$

- Penalty view point:
  - Lagrangian: $L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j \left(a_j^\top x - b_j\right),\ \lambda_i \geq 0$
- Karush-Kuhn-Tucker (KKT) Conditions:
  - Stationarity $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$
  - Primal feasibility: $g_i(x^*) \leq 0,\ a_i^\top x^* - b_i = 0$
  - Dual feasibility: $\lambda^* \geq 0$
  - Complementary slackness: $\lambda_i^* g_i(x^*) = 0,\ i = 1, \dots, n$

- Solve above systems of equations to obtain optimum

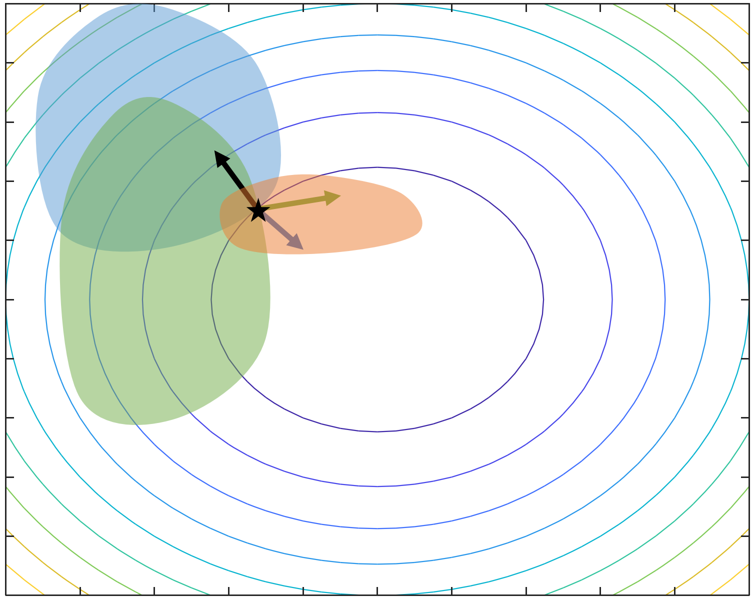
Solving Convex Optimization Problems

- Solve the optimality conditions

- Gradient methods for approximating solutions to convex optimization problems

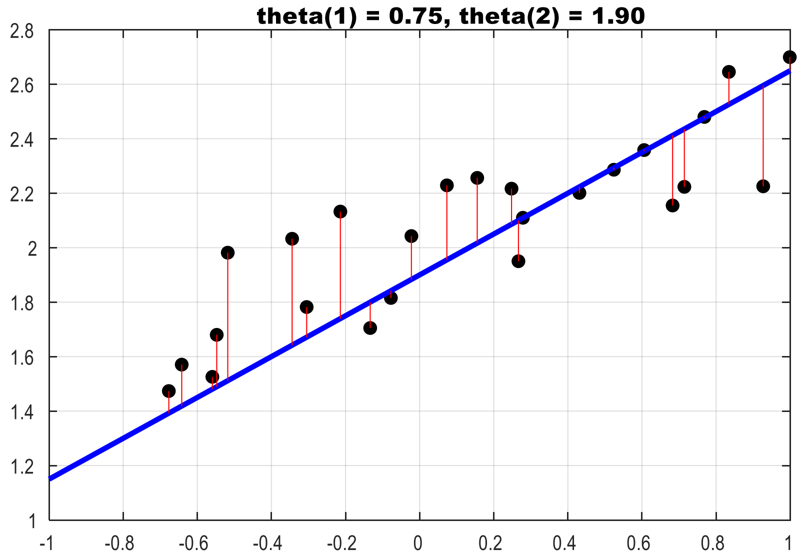
Optimality Conditions for Convex Programs

- Full optimization problem:
$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0, i = 1, \dots, n \\ & a_j^\top x = b_j, j = 1, \dots, m \end{aligned}$$
- Penalty view point:
  - Lagrangian: $L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j \left(a_j^\top x - b_j\right),\ \lambda_i \ge 0$
- Karush-Kuhn-Tucker (KKT) Conditions:
  - Stationarity $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$
  - Primal feasibility: $g_i(x^*) \le 0,\ a_i^\top x^* - b_i = 0$
  - Dual feasibility: $\lambda^* \ge 0$
  - Complementary slackness: $\lambda_i^* g_i(x^*) = 0,\ i = 1, \dots, n$



- Solve above systems of equations to obtain optimum

Example: Least Squares



$$\underset{\theta}{\text{minimize}} \|X\theta - Y\|_2^2$$

- Scalar example:
  - Data: $\{x_i, y_i\}_{i=1}^n$, $x_i, y_i \in \mathbb{R}$
  - Model: $y = mx + b, m, b \in \mathbb{R}$
  - Sum of error of model: $\sum_{i=1}^n (y_i - mx_i - b)^2$
  - No constraints: allow *any* m, b
- Error in matrix form: $e_i = y_i - \begin{bmatrix} x_i & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix}$
  - Stacking the data points: $E_i = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{Y} - \underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} m \\ b \end{bmatrix}}_{\theta}$

Optimality Conditions for Convex Programs

- Full optimization problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \\ \text{subject to} & g_i(x) \leq 0, i = 1, \dots, n \\ & a_j^\top x = b_j, j = 1, \dots, m \end{array}$$

- Penalty view point:
  - Lagrangian: $L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j \left(a_j^\top x - b_j\right),\ \lambda_i \geq 0$
- Karush-Kuhn-Tucker (KKT) Conditions:
  - Stationarity $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$ $\longleftarrow$ $\boldsymbol{\nabla f(x) = 0}$
  - Primal feasibility: $g_i(x^*) \leq 0,\ a_i^\top x^* - b_i = 0$
  - Dual feasibility: $\lambda^* \geq 0$
  - Complementary slackness: $\lambda_i^* g_i(x^*) = 0,\ i = 1, \dots, n$

- Solve above systems of equations to obtain optimum

Example: Least Squares

$$\underset{\theta}{\text{minimize}} \|X\theta - Y\|_2^2$$

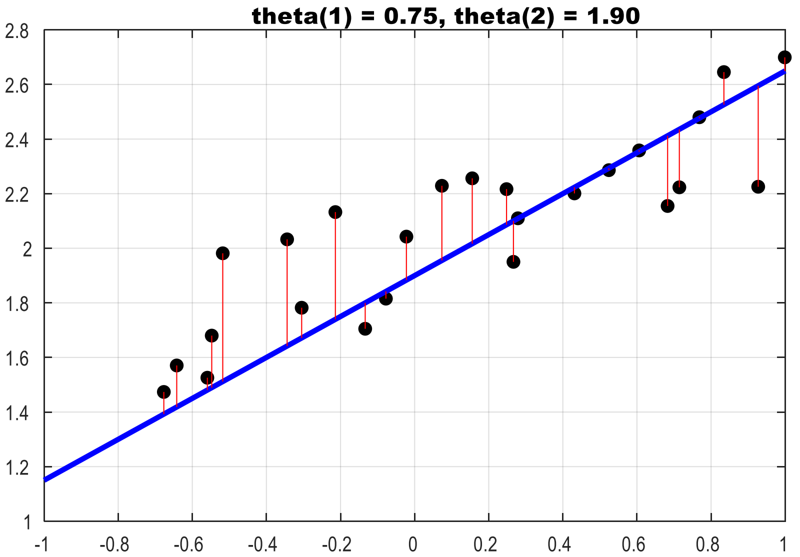
- Analytic solution available!
  - Objective: $f(\theta) = \|X\theta - Y\|_2^2$, set derivative to zero
  - $f(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - $f(\theta) = \theta^\top X^\top X\theta - 2Y^\top X\theta + Y^\top Y$

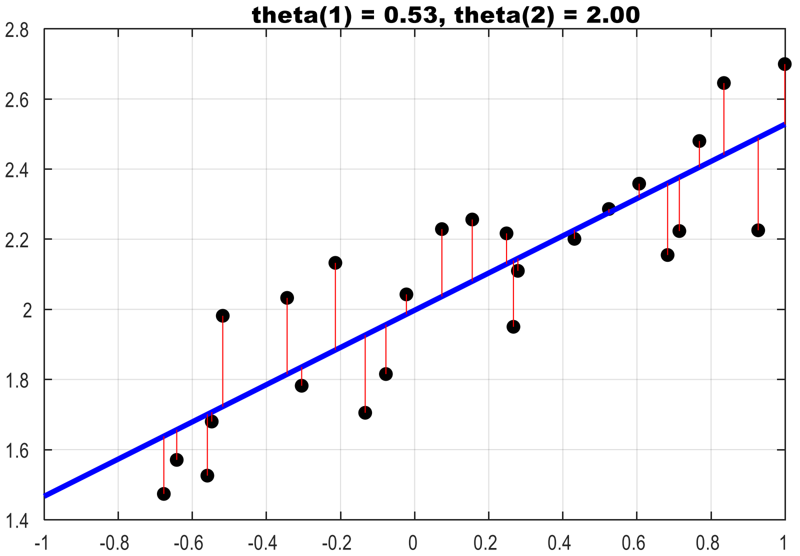
$$\frac{\partial f}{\partial \theta} = 2X^\top X\theta - 2X^\top Y$$

$$0 = 2X^\top X\theta - 2X^\top Y$$

$$X^\top Y = X^\top X\theta$$

$$\theta = (X^\top X)^{-1} X^\top Y$$





Example: Least Squares

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \|X\theta - Y\|_2^2 \\ \text{subject to} \quad & \theta_1^2 + \theta_2^2 \leq 1 \end{aligned}$$

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \|X\theta - Y\|_2^2 \\ \text{subject to} \quad & \|\theta\|_2^2 - 1 \leq 0 \end{aligned}$$

- Lagrangian: $L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x)$

$$L(\theta, \lambda) = \|X\theta - Y\|_2^2 + \lambda(\|\theta\|_2^2 - 1)$$

- Stationarity $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$

$$\nabla_\theta L(\theta, \lambda) = 2X^\top X\theta - 2X^\top Y + 2\lambda\theta$$

$$0 = X^\top X\theta - X^\top Y + \lambda\theta$$

$$X^\top Y = (X^\top X + \lambda I)\theta$$

- Primal feasibility: $g_i(x^*) \leq 0,\ a_i^\top x^* - b_i = 0$
- Dual feasibility: $\lambda^* \geq 0$

$$\|\theta\|_2^2 - 1 \leq 0$$

$$\lambda \geq 0$$

- Complementary slackness: $\lambda_i^* g_i(x^*) = 0,\ i = 1, \dots, n$

$$\lambda(\|\theta\|_2^2 - 1) = 0$$

$\lambda = 0$ or $\|\theta\|_2^2 = 1$

Example: Least Squares

- Case 1: If $\lambda = 0$, then
  - $\lambda \geq 0$ is satisfied automatically
  - $X^\top Y = (X^\top X)\theta \Rightarrow \theta = (X^\top X)^{-1} X^\top Y$
  - If $\|\theta\|_2^2 - 1 \leq 0$ happens to be true, we are done
  - Otherwise, try case 2

KKT conditions:

- $X^\top Y = (X^\top X + \lambda I)\theta$
- $\|\theta\|_2^2 - 1 \leq 0$
- $\lambda \geq 0$
- $\lambda = 0$ or $\|\theta\|_2^2 = 1$

- Case 2: If $\|\theta\|_2^2 = 1$, then
  - $\|\theta\|_2^2 - 1 \leq 0$ is satisfied automatically
  - $X^\top Y = (X^\top X + \lambda I)\theta \Rightarrow \theta = (X^\top X + \lambda I)^{-1} X^\top Y$
  - Solve $\|\theta\|_2^2 = 1$ and $\theta = (X^\top X + \lambda I)^{-1} X^\top Y$ for θ and λ
  - If $\lambda \geq 0$, we are done

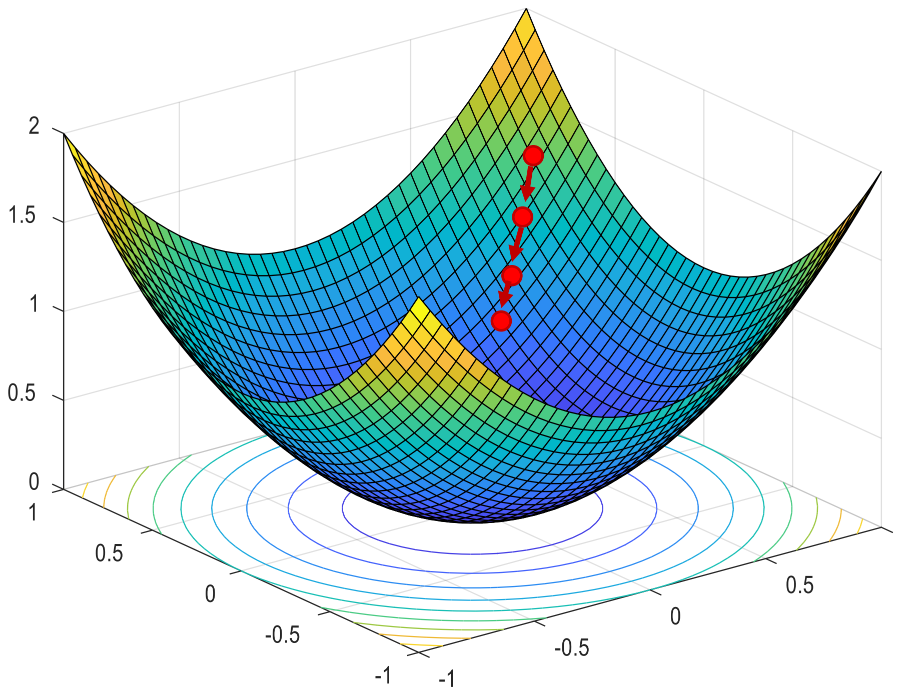
Solving the Optimality Conditions

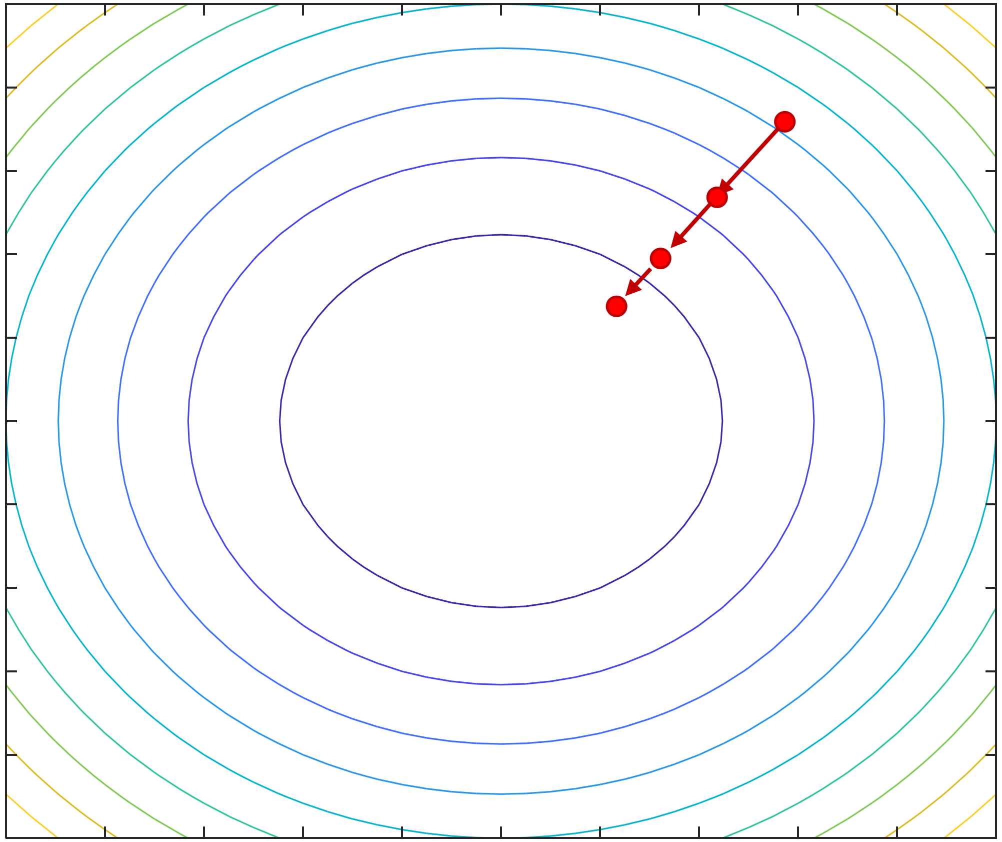
$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, i = 1, \dots, n \\ & a_j^\top x = b_j, j = 1, \dots, m \end{aligned}$$

- Equations to solve: KKT conditions
  - Stationarity $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$
  - Primal feasibility: $g_i(x^*) \leq 0,\ a_i^\top x^* - b_i = 0$
  - Dual feasibility: $\lambda^* \geq 0$
  - Complementary slackness: $\lambda_i^* g_i(x^*) = 0,\ i = 1, \dots, n$
- Use numerical equation solvers, or do it by hand (as much as possible)
- For convex problems, KKT conditions are necessary and sufficient
- For non-convex problems, KKT conditions are just necessary

Numerical Solution: Gradient Methods

- Start from x^0 and construct a sequence x^k such that $x^k \to x^*$
  - Calculate x^{k+1} from x^k by "going down the gradient"
  - Unconstrained case: $x^{k+1} = x^k - \alpha^k \nabla f(x)$, $\alpha^k > 0$



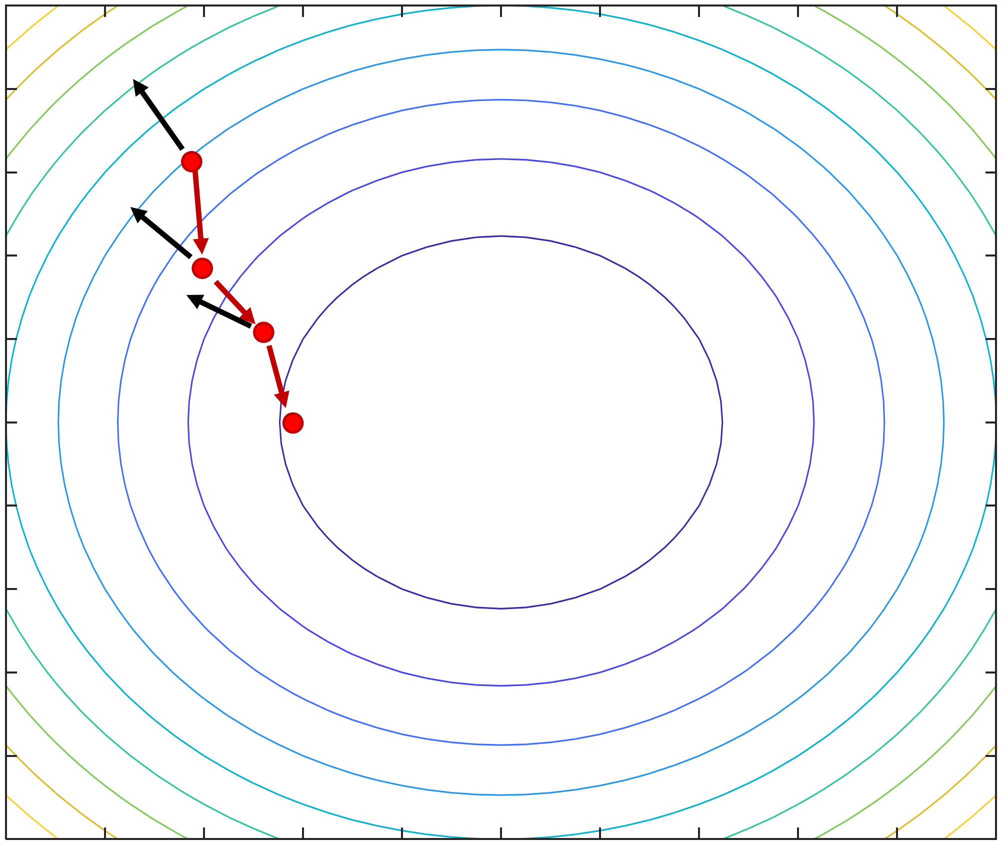


Numerical Solution: Gradient Methods

- Start from x^0 and construct a sequence x^k such that $x^k \to x^*$
  - Calculate x^{k+1} from x^k by "going down the gradient"
  - Unconstrained case: $x^{k+1} = x^k - \alpha^k \nabla f(x)$, $\alpha^k > 0$
- More generally, $x^{k+1} = x^k + \alpha^k d^k$ for some d such that

$$\nabla f\left(x^k\right) \cdot d^k < 0$$

- Tuning parameters: descent direction d^k, and step size α^k



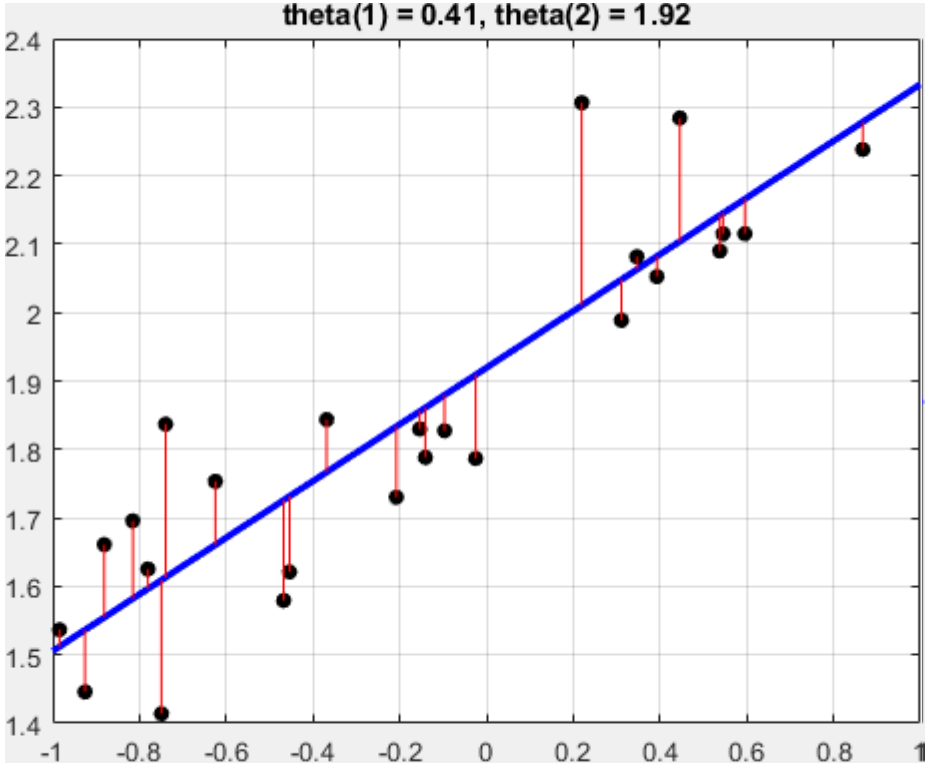
Descent Direction

- Steepest descent: $d^k = -\nabla f\left(x^k\right)$
  - $x^{k+1} = x^k - \alpha^k \nabla f(x)$
  - Simple but sometimes leads to slow convergence

Steepest Descent (Gradient Descent) Example

- Line fitting: $f(\theta) = \|X\theta - Y\|_2^2$
  - $\frac{\partial f}{\partial \theta} = 2X^\top X\theta - 2X^\top Y$

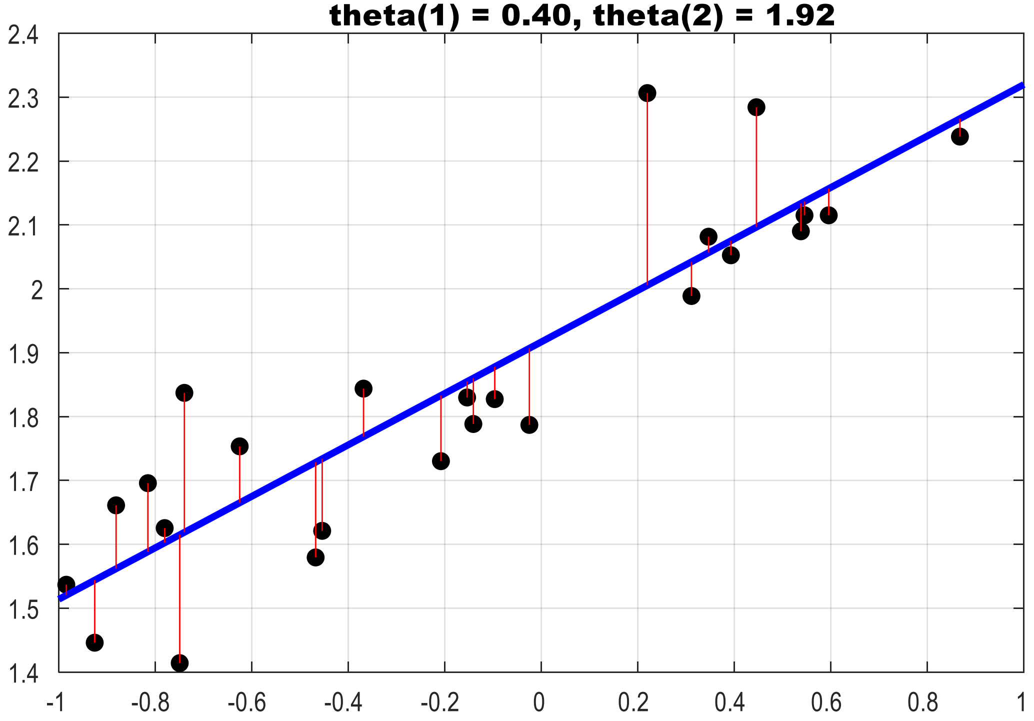
```
theta_last = [-2; -2];
dtheta = inf;
maxIter = 500;

for k = 1:maxIter
    if (norm(dtheta) <= 0.001)
        break;
    end

    alpha = 0.1/k;
    theta = theta_last - alpha*(2*X'*X*theta_last - 2*X'*Y);
    dtheta = theta_last - theta;
    theta_last = theta;
end
```



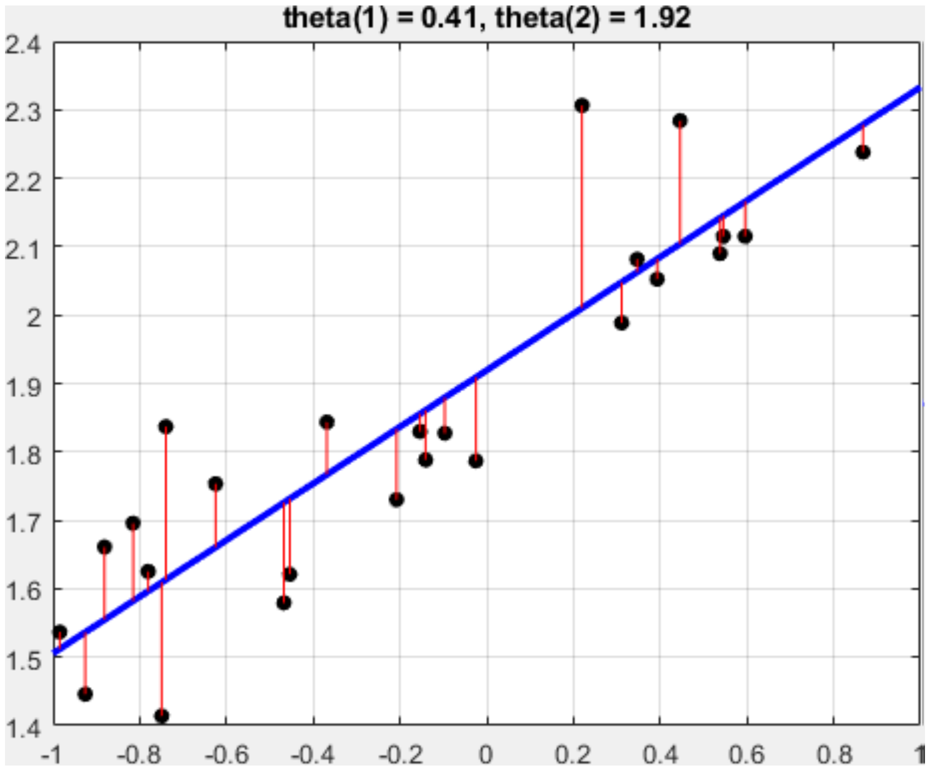
$$\theta^{k+1} = \theta^k - \underbrace{\frac{0.1}{k}}_{\alpha^k}(2X^\top X\theta - 2X^\top Y)$$

Steepest Descent (Gradient Descent) Example

- Line fitting: $f(\theta) = \|X\theta - Y\|_2^2$
  - $\frac{\partial f}{\partial \theta} = 2X^\top X\theta - 2X^\top Y$

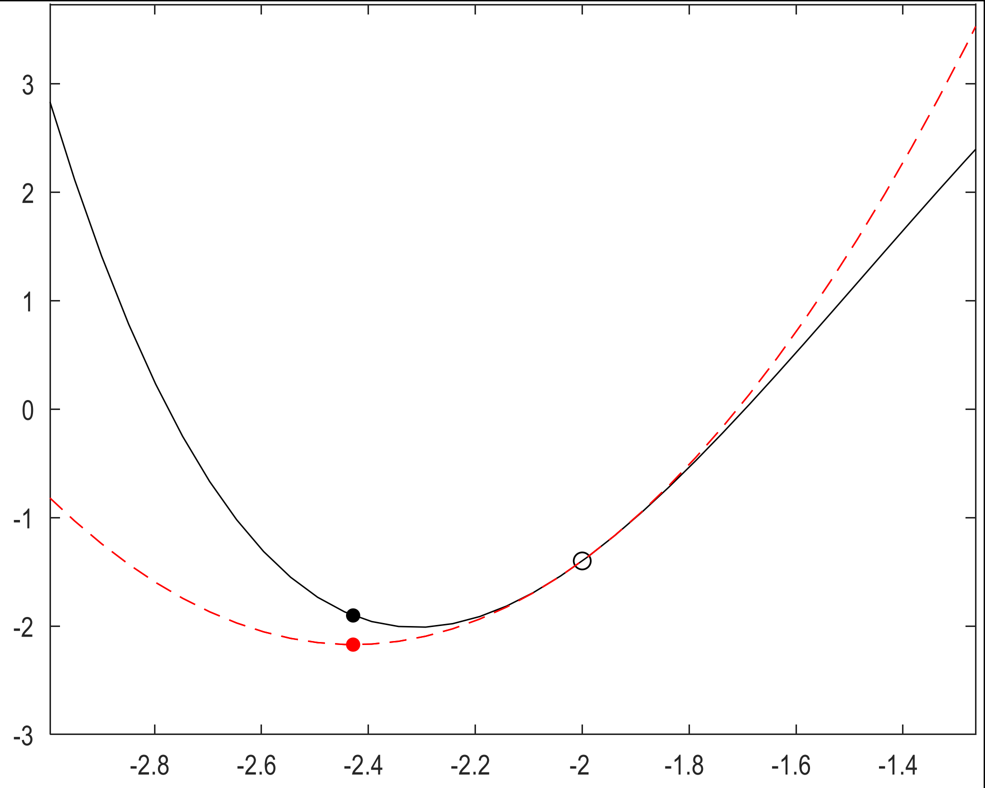


Previously: $\theta = (X^\top X)^{-1} X^\top Y$



$$\theta^{k+1} = \theta^k - \underbrace{\frac{0.1}{k}}_{\alpha^k}(2X^\top X\theta - 2X^\top Y)$$

Descent Direction



- Steepest descent: $d^k = -\nabla f(x^k)$
  - $x^{k+1} = x^k - \alpha^k \nabla f(x)$
  - Simple but sometimes leads to slow convergence

- Newton's method: $d^k = \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$
  - Minimize the quadratic approximation:

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

  - Set gradient to zero to obtain next iterate

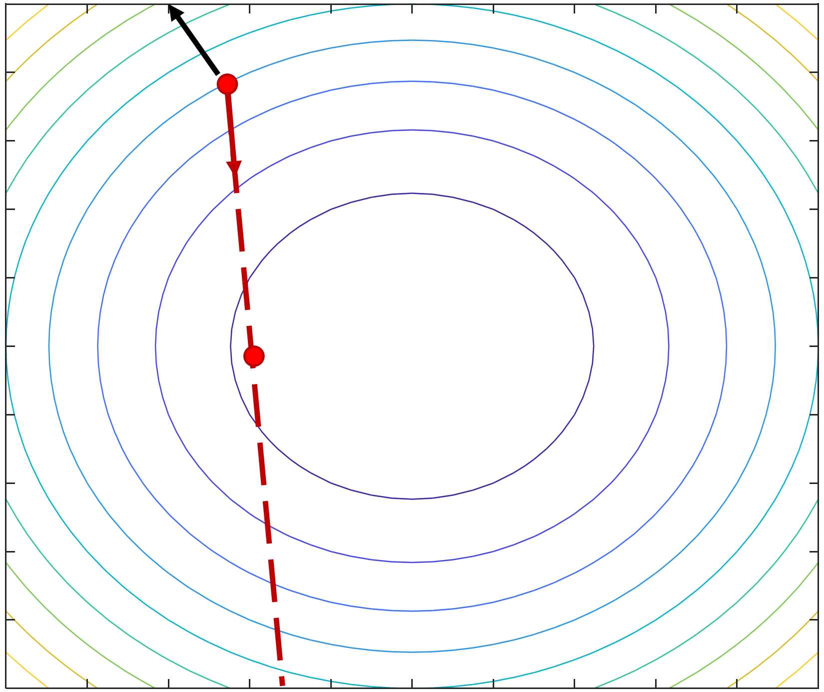
$$\nabla f^k(x) = \nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0$$

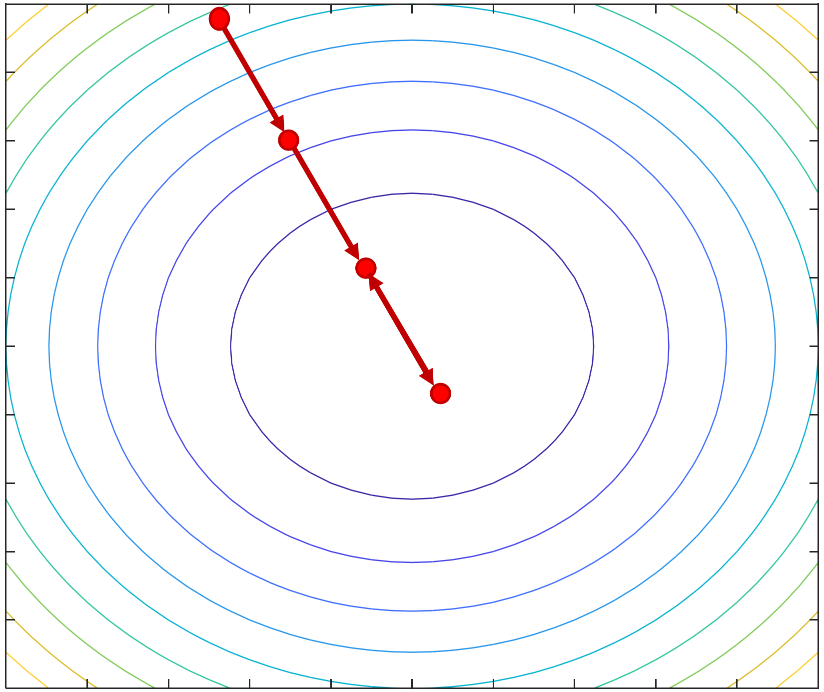
$$\Rightarrow x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

  - Fast convergence, but matrix inverse required
  - Alternatively, use an algorithm to minimize a quadratic function

Step Size

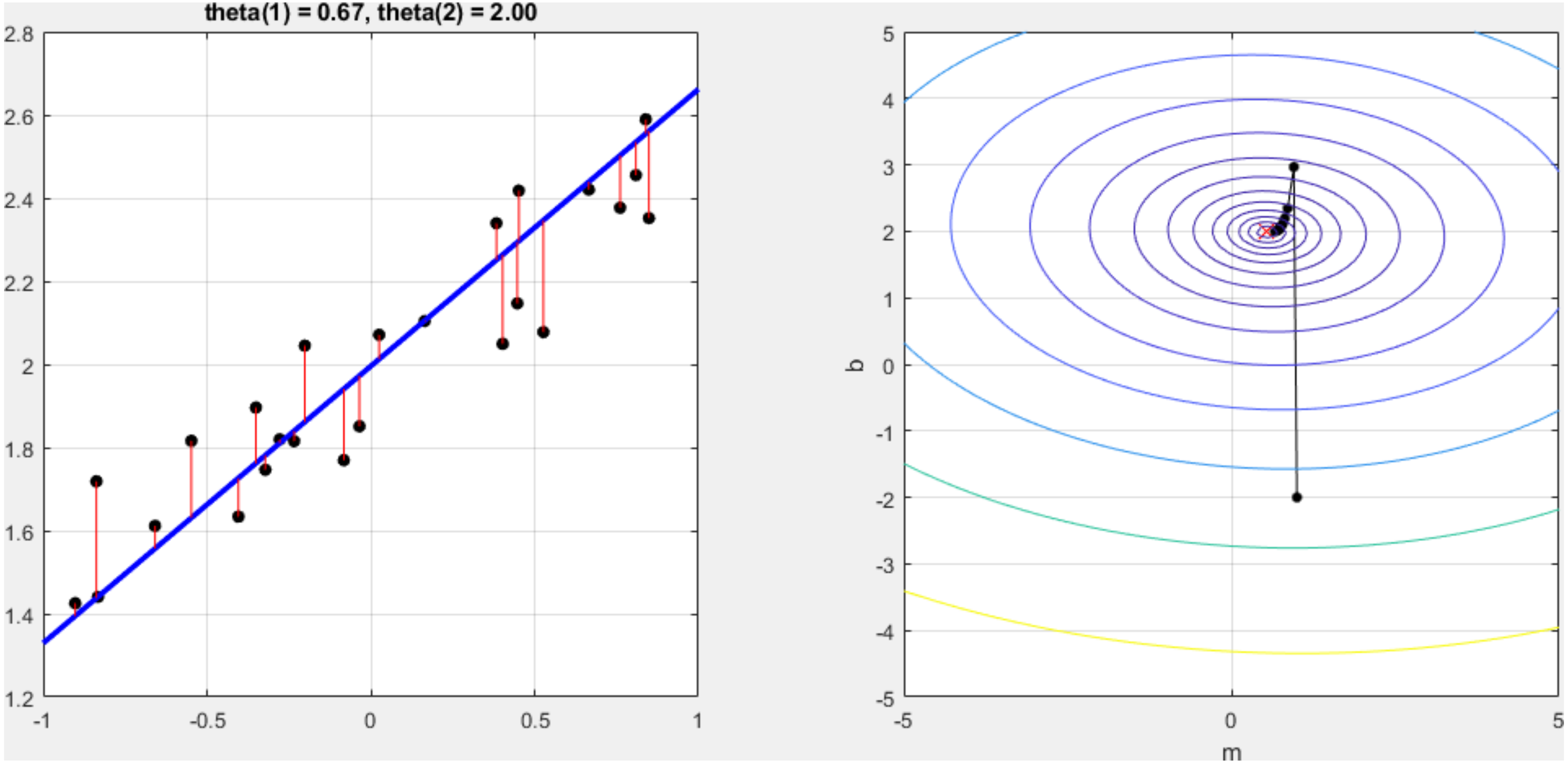
- Recall $x^{k+1} = x^k + \alpha^k d^k$, with $\nabla f(x^k)^{\mathsf{T}} d^k < 0$
- Line search: choose $\alpha^k = \min_{\alpha \geq 0} f(x^k + \alpha^k d^k)$
  - Requires minimization
- Constant step size: $\alpha^k = \alpha$
  - May not converge
- Diminishing step size: $\alpha^k \to 0$
  - Still need to explore all regions $\sum \alpha^k = \infty$
  - For example: $\alpha^k = \frac{\alpha^0}{k}$





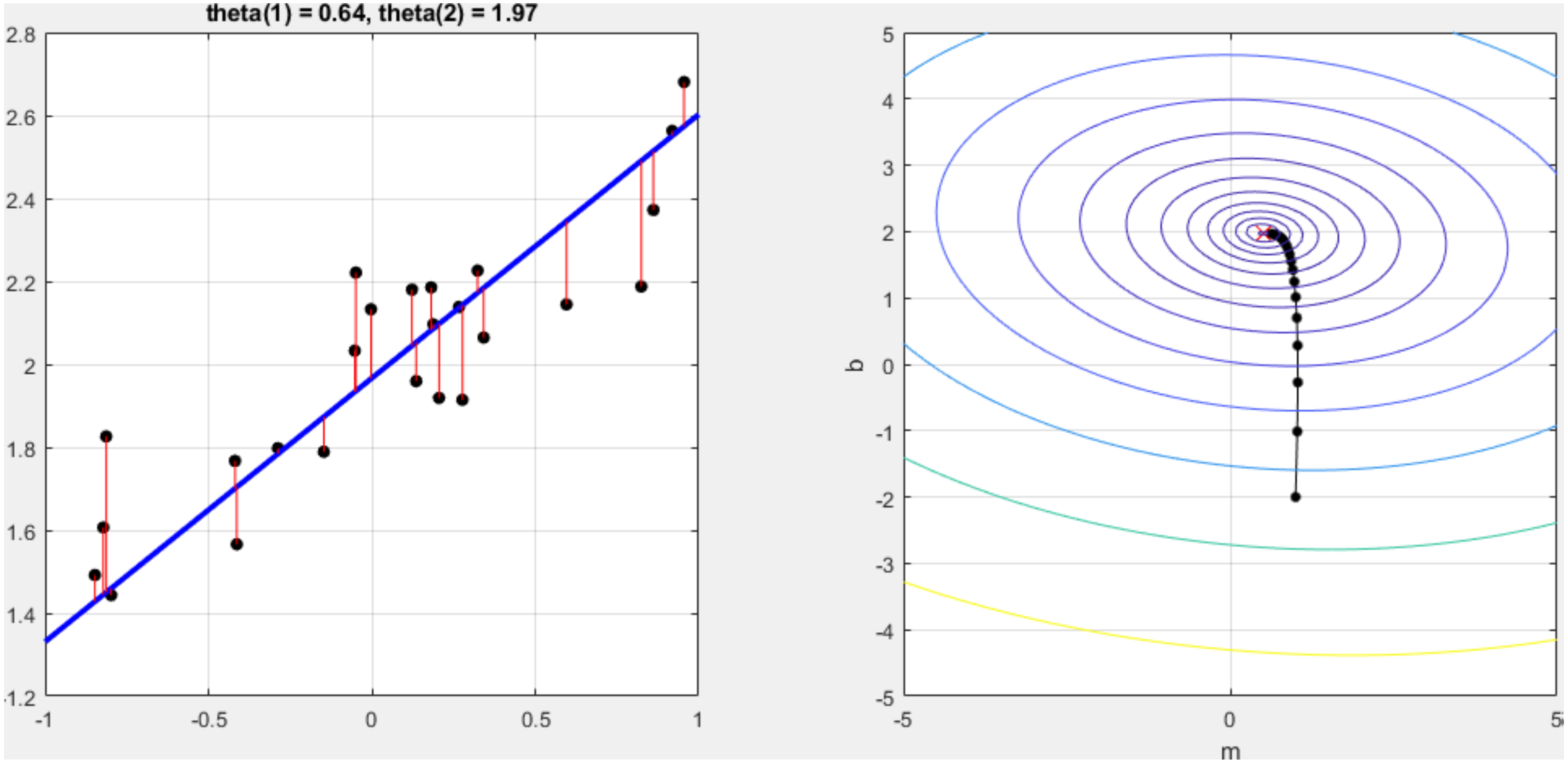
Step Size Example

- Steepest descent, $\alpha^k = \alpha^0 / k$



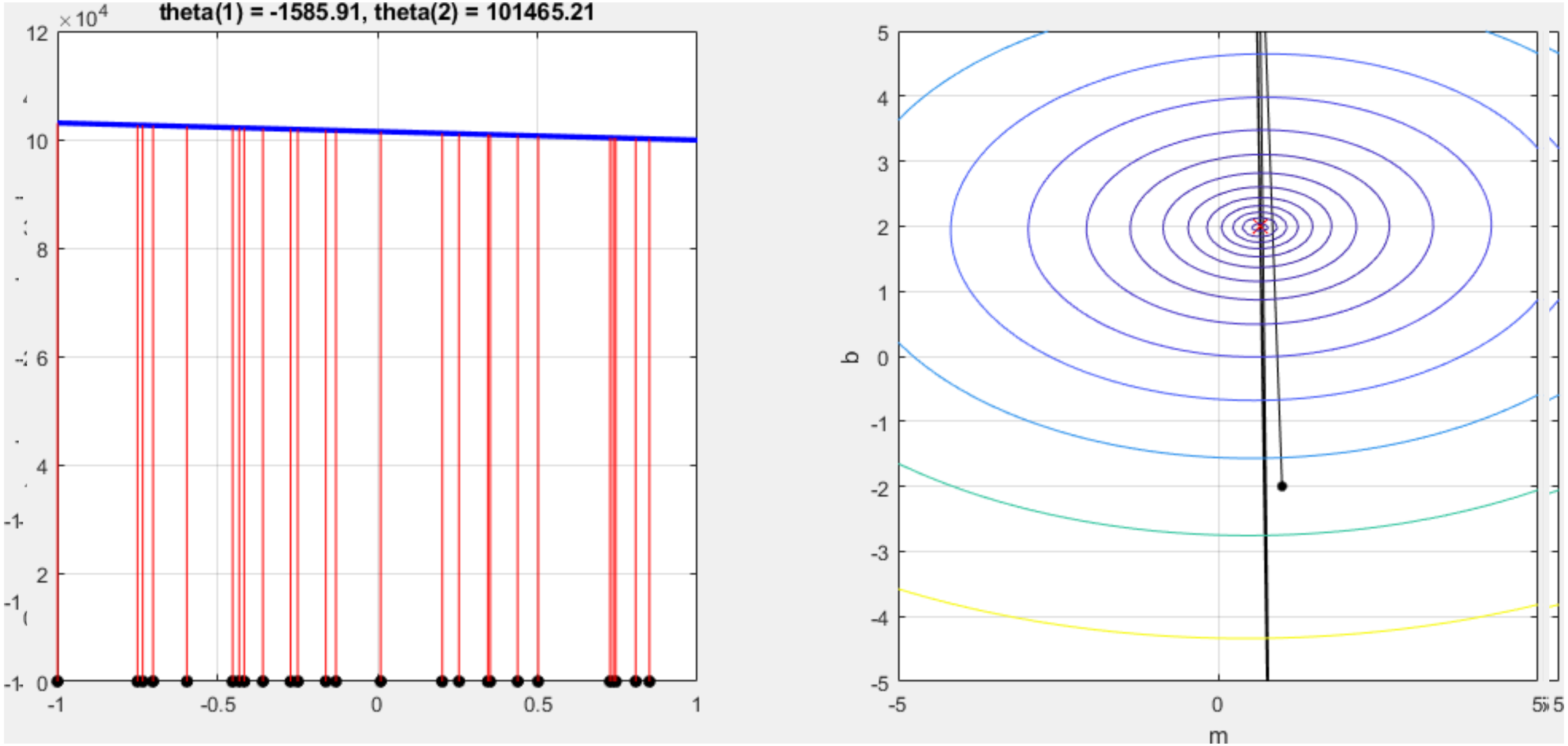
Step Size Example

- Steepest descent, $\alpha^k = \alpha^0$ (small steps)



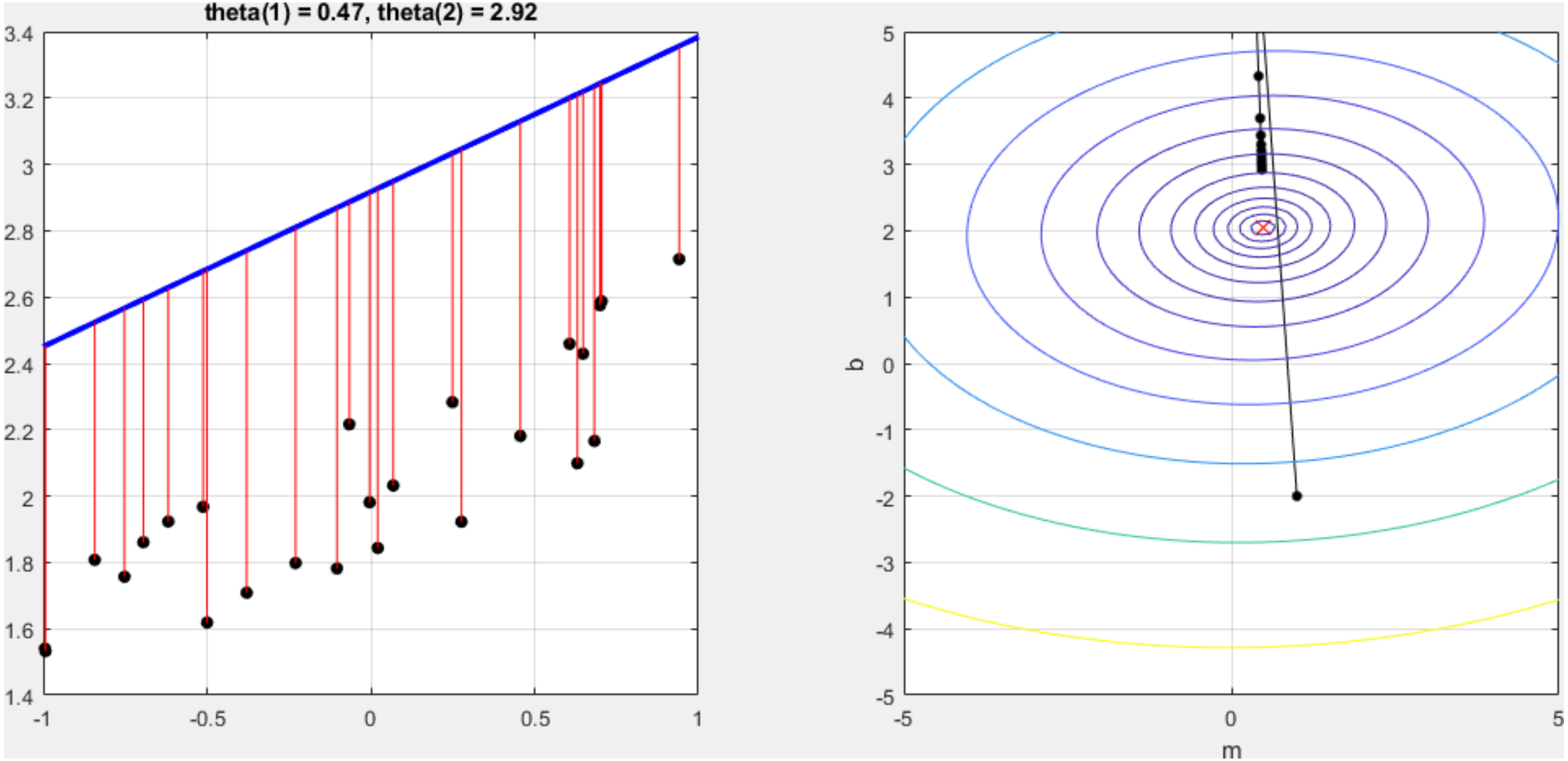
Step Size Example

- Steepest descent, $\alpha^k = \alpha^0$ (large steps)



Step Size Example

- Steepest descent, $\alpha^k = \alpha^0/k^2$ (steps do not sum to ∞: $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$)



Dealing with Constraints

- Idea 1: Apply descent step, and project point to feasible set
  - Proximal gradient methods
  - Difficulty: Computing the projected point

- Idea 2: Set penalty to ∞ for constraint violation
  - Barrier functions