Machine Learning
Lecture 4

Outline

- Understanding optimization view of learning
  - large margin linear classification
  - regularization, generalization
- Optimization algorithms
  - preface: gradient descent optimization
  - stochastic gradient descent
  - quadratic program

Recall: learning as optimization

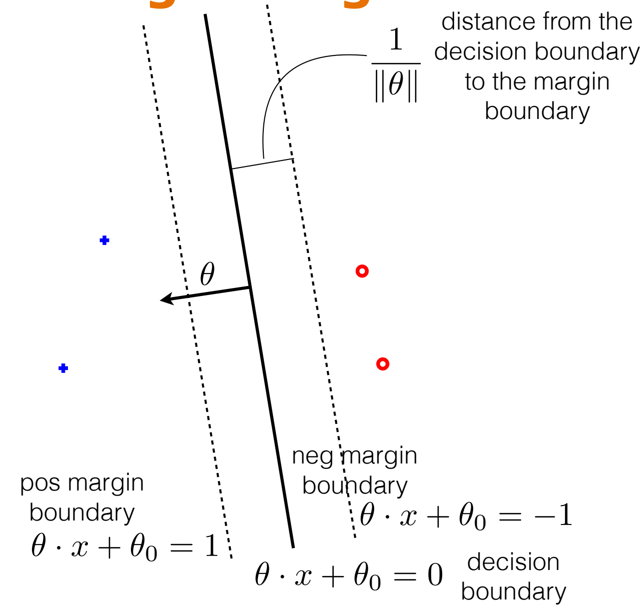
- Machine learning problems are often cast as optimization problems

  objective function = average loss + regularization

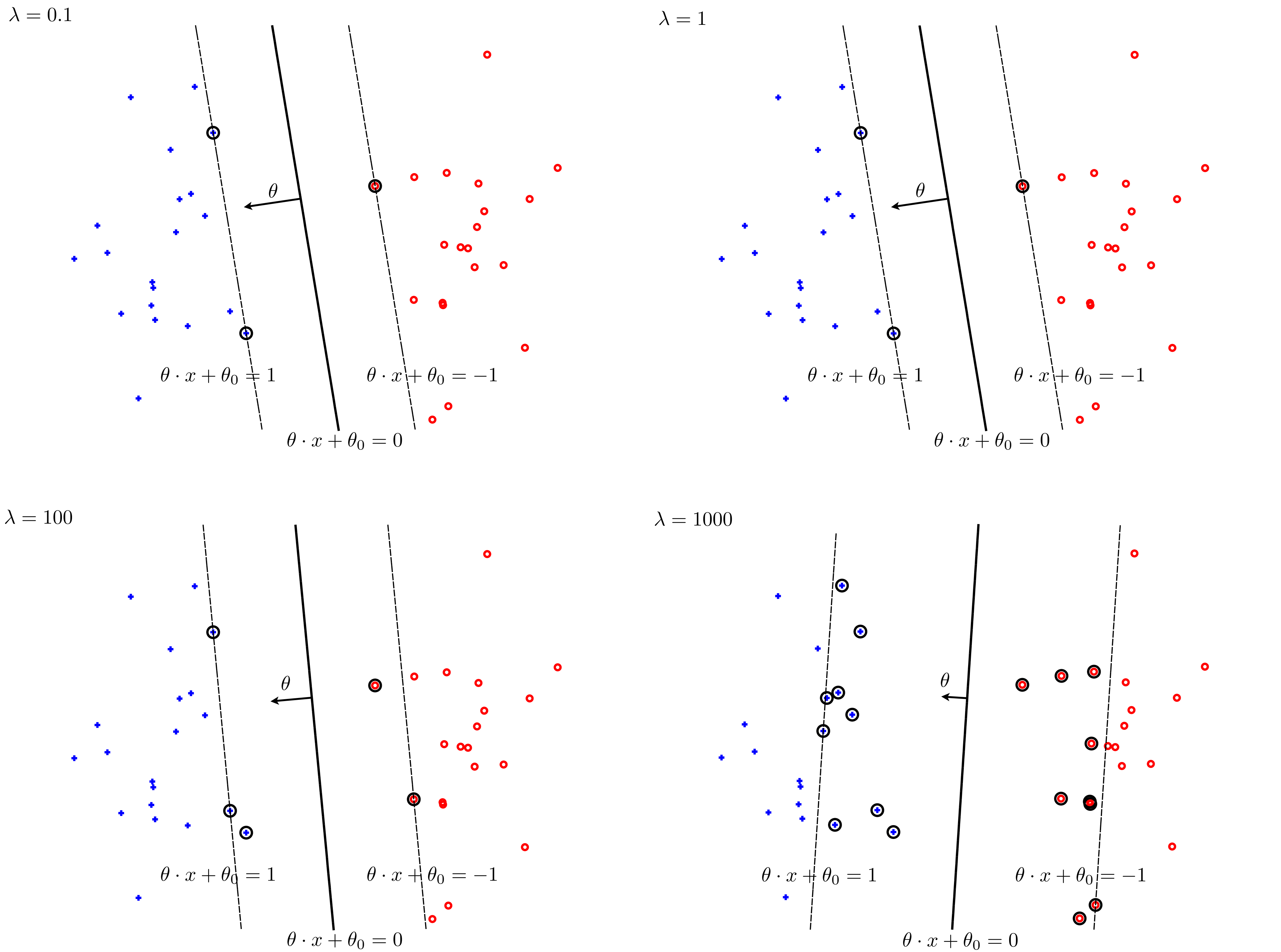
- Large margin linear classification as optimization (Support Vector Machine)

$$J(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_h\big(y^{(i)}(\theta \cdot x^{(i)} + \theta_0)\big) + \frac{\lambda}{2}\|\theta\|^2$$

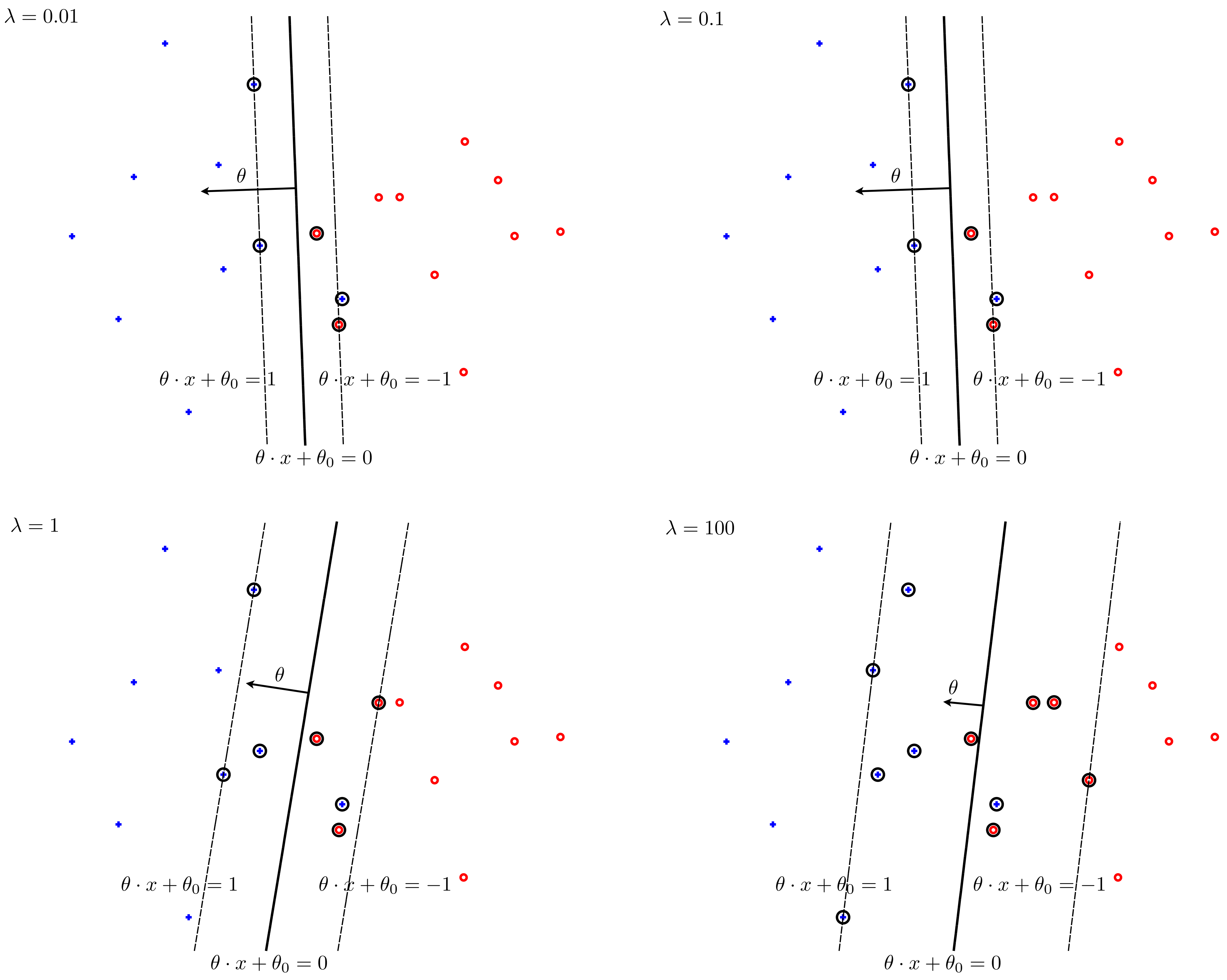
Recall: large margin classifier



$$J(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_h\big(y^{(i)}(\theta \cdot x^{(i)} + \theta_0)\big) + \frac{\lambda}{2}\|\theta\|^2$$



$\lambda = 0.1$
θ
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = -1$
$\theta \cdot x + \theta_0 = 0$
$\lambda = 1$
θ
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = -1$
$\theta \cdot x + \theta_0 = 0$
$\lambda = 100$
θ
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = -1$
$\theta \cdot x + \theta_0 = 0$
$\lambda = 1000$
θ
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = -1$
$\theta \cdot x + \theta_0 = 0$



$\lambda = 0.01$
θ
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = -1$
$\theta \cdot x + \theta_0 = 0$
$\lambda = 0.1$
θ
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = -1$
$\theta \cdot x + \theta_0 = 0$
$\lambda = 1$
θ
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = -1$
$\theta \cdot x + \theta_0 = 0$
$\lambda = 100$
θ
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = -1$
$\theta \cdot x + \theta_0 = 0$

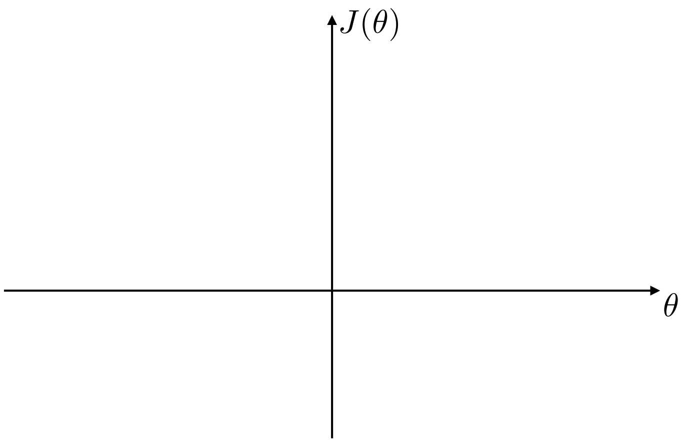
Regularization, generalization

$$J(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_h\big(y^{(i)}(\theta \cdot x^{(i)} + \theta_0)\big) + \frac{\lambda}{2}\|\theta\|^2$$

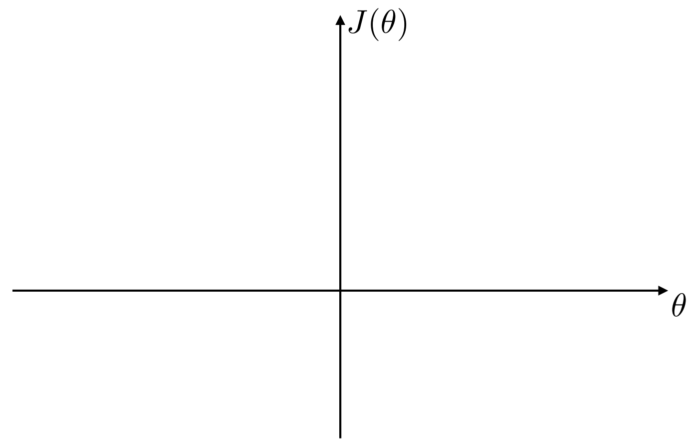
Outline

- Understanding optimization view of learning
  - large margin linear classification
  - regularization, generalization
- Optimization algorithms
  - preface: gradient descent optimization
  - stochastic gradient descent
  - quadratic program

Preface: Gradient descent



Preface: Gradient descent



Stochastic gradient descent

$$J(\theta,\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Loss}_h\big(y^{(i)}(\theta\cdot x^{(i)}+\theta_0)\big) + \frac{\lambda}{2}\|\theta\|^2$$
$$= \frac{1}{n}\sum_{i=1}^{n} \left[\mathrm{Loss}_h\big(y^{(i)}(\theta\cdot x^{(i)}+\theta_0)\big) + \frac{\lambda}{2}\|\theta\|^2\right]$$

Stochastic gradient descent

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} \left[\mathrm{Loss}_h(y^{(i)}\theta\cdot x^{(i)}) + \frac{\lambda}{2}\|\theta\|^2\right]$$

Stochastic gradient descent

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} \left[\mathrm{Loss}_h(y^{(i)}\theta\cdot x^{(i)}) + \frac{\lambda}{2}\|\theta\|^2\right]$$

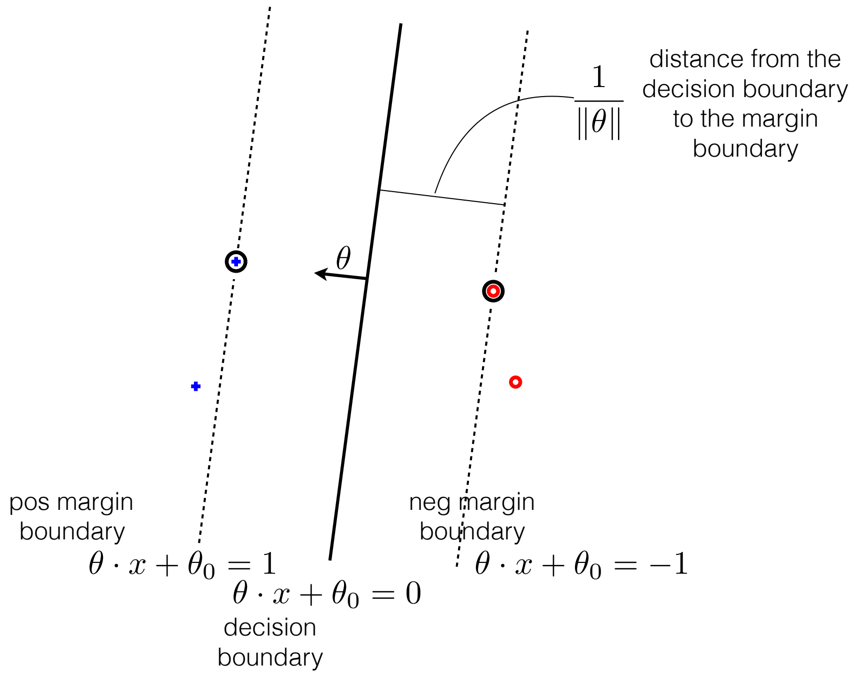
Select $i \in \{1,\ldots,n\}$ at random

$$\theta \leftarrow \theta - \eta_t \nabla_\theta \left[\mathrm{Loss}_h(y^{(i)}\theta\cdot x^{(i)}) + \frac{\lambda}{2}\|\theta\|^2\right]$$

Support Vector Machine

- Support Vector Machine finds the maximum margin linear separator by solving the quadratic program that corresponds to $J(\theta,\theta_0)$
- In the realizable case, if we disallow any margin violations, the quadratic program we have to solve is

$$\begin{aligned}&\text{Find } \theta,\ \theta_0 \text{ that}\\ &\quad \text{minimize } \frac{1}{2}\|\theta\|^2 \ \text{ subject to}\\ &\qquad y^{(i)}(\theta\cdot x^{(i)}+\theta_0) \ge 1,\ \ i=1,\ldots,n\end{aligned}$$



Summary

- Learning problems can be formulated as optimization problems of the form: loss + regularization
- Linear, large margin classification, along with many other learning problems, can be solved with stochastic gradient descent algorithms
- Large margin linear classifier can be also obtained via solving a quadratic program (Support Vector Machine)