EECS 445

Introduction to Machine Learning

Stochastic Gradient Descent
Support Vector Machines

Prof. Kutty

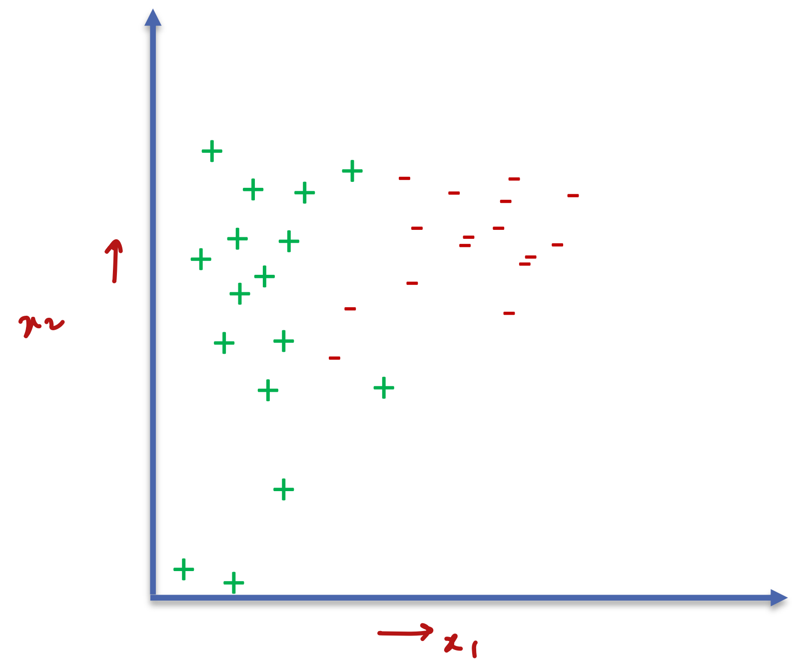
Today's Agenda

- Recap
  - Loss functions and Gradient Descent
- Section 1
  - Stochastic Gradient Descent
- Section 2
  - Support Vector Machines
  - hard margin SVM
- And later…
  - Soft Margin SVMs and feature maps

https://forms.gle/ffiBvNbPjHF8ghi77

Datasets that are not linearly separable

(in the original feature space)



Idea: minimize empirical risk with hinge loss using gradient descent

In other words, find $\bar{\boldsymbol{\theta}}$ that minimizes

empirical risk

$$R_n(\bar{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_{\text{hinge}}(y^{(i)}\bar{\boldsymbol{\theta}}.\bar{\boldsymbol{x}}^{(i)})$$

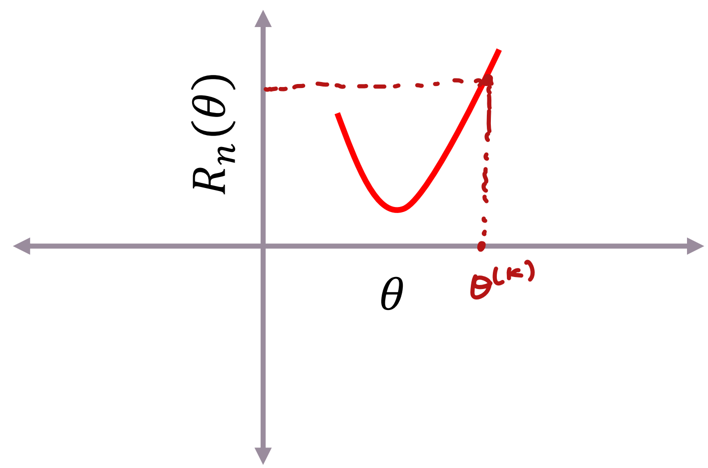
training data

parameter

Gradient Descent (GD)

Gradient Descent (GD) Idea: take a small step in the opposite direction to which the gradient points

$\eta_k = \frac{1}{k+1}$

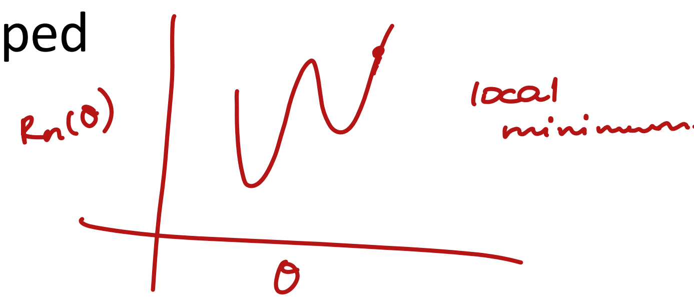


$$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} - \eta_k \nabla_{\bar{\theta}} R_n(\bar{\theta}) \Big|_{\bar{\theta} = \bar{\theta}^{(k)}}$$

informally, a convex function is characteristically 'bowl'-shaped

$\bar{\theta}^{(0)} \rightarrow$ initial guess

$\bar{\theta}^{(k)} \rightarrow$ current guess



Gradient Descent (GD)

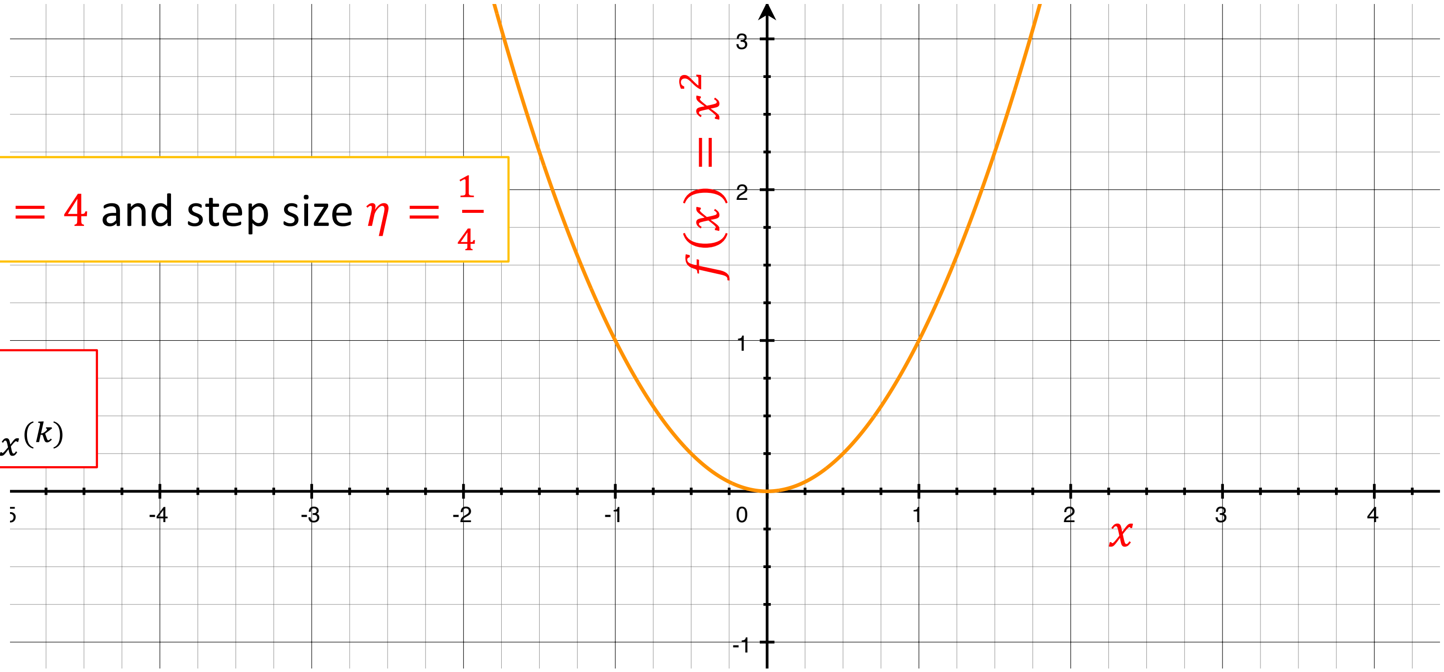
Idea via a simple example

Goal: Find value for variable x that minimizes the convex function $f(x)$.

Initial guess on minimizer $x^{(0)} = 4$ and step size $\eta = \frac{1}{4}$

Update guess on x:

$$x^{(k+1)} = x^{(k)} - \eta \nabla_x f(x) \Big|_{x=x^{(k)}}$$



https://forms.gle/ffiBvNbPjHF8ghi77

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

A. $x^{(k)}$

B. $\frac{x^{(k)}}{2}$

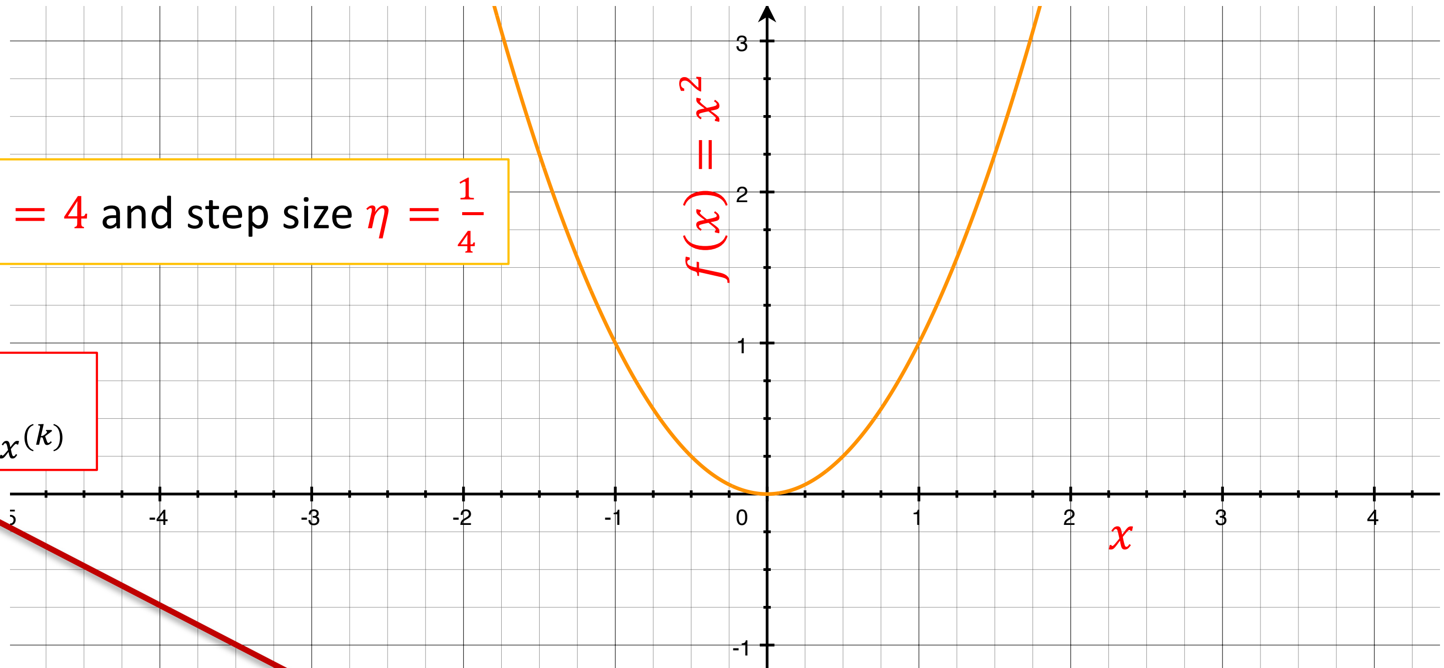
C. 0

D. unsure

Gradient Descent (GD)

Idea via a simple example

Goal: Find value for variable x that minimizes the convex function $f(x)$.



Initial guess on minimizer $x^{(0)} = 4$ and step size $\eta = \frac{1}{4}$

Update guess on x:

$$\boxed{x^{(k+1)} = x^{(k)} - \eta \nabla_x f(x) \Big|_{x=x^{(k)}}}$$

$$x^{(k+1)} = x^{(k)} - \frac{1}{4} 2x^{(k)}$$

$$= \frac{x^{(k)}}{2}$$

$$\nabla_x f(x)|_{x=x^{(k)}} = \nabla_x \, x^2|_{x=x^{(k)}} = 2\, x^{(k)}$$

$x^{(0)} = 4$

$x^{(1)} = 2$

$x^{(2)} = 1$

$x^{(3)} = 0.5$

$\vdots$

Gradient Descent (GD)

Goal:

Given $S_n = \{\bar{x}^{(i)}, y^{(i)}\}_{i=1}^{n}$

Find the value of parameter $\bar{\boldsymbol{\theta}}$ that minimizes empirical risk $R_n(\bar{\boldsymbol{\theta}})$

$$R_n(\bar{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\max\{1 - y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)}), 0\}$$

$k = 0, \bar{\theta}^{(0)} = \bar{0}$

while convergence criteria is not met

$$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} - \eta\nabla_{\bar{\theta}}R_n(\bar{\theta})|_{\bar{\theta}=\bar{\theta}^k}$$

k++

(variable or fixed) step size η

$$\nabla_{\bar{\theta}}R_n(\bar{\theta}) = \left[\frac{\partial R_n(\bar{\theta})}{\partial\theta_1}, \dots, \frac{\partial R_n(\bar{\theta})}{\partial\theta_d}\right]^{\mathrm{T}}$$

Gradient Descent (GD)

$$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} - \eta \nabla_{\bar{\theta}} R_n(\bar{\theta})|_{\bar{\theta}=\bar{\theta}^k}$$

$$\nabla_{\bar{\theta}} R_n(\bar{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\bar{\theta}} \max\{1 - y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}), 0\}$$

Bad news:

Due to the summation involved in calculating the gradient, in order to make a single update, you have to look at *every training example*

If we have a lot of training examples, this will be *slow*

Stochastic Gradient Descent

or How to speed things up!

Stochastic Gradient Descent (SGD)

Idea: Reminder: Hinge loss $Loss_h(z) = max\{1-z, 0\}$

Instead of looping over all examples before update, update based on a single point

$$R_n(\bar{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \max\{1 - y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)}), 0\}$$

k = 0, $\bar{\theta}^{(0)} = \bar{0}$

GD update step

$$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} - \eta\nabla_{\bar{\theta}} R_n(\bar{\theta})|_{\bar{\theta}=\bar{\theta}^k}$$

while convergence criteria is not met

randomly shuffle points

pick a datapoint uniformly at random without replacement from the dataset

for i = 1, …,n

$$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} - \eta\boxed{\nabla_{\bar{\theta}}\text{Loss}_h(y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)}))|_{\bar{\theta}=\bar{\theta}^k}}$$

k++

$$\nabla_{\bar{\theta}} \max\{1 - y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)}), 0\}$$

Stochastic Gradient Descent

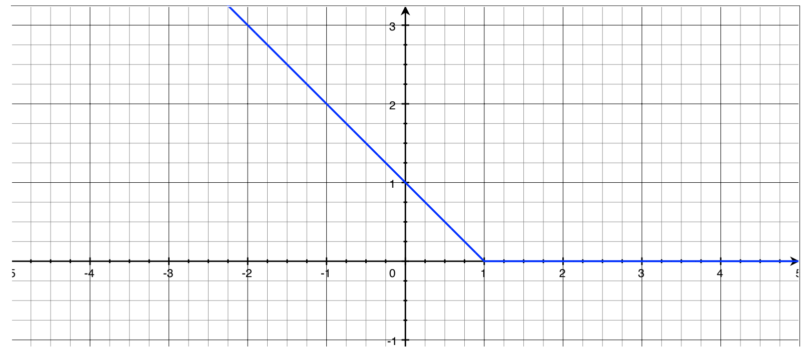
$$\nabla_{\bar{\theta}} \max\{1 - y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}), 0\}$$

Case 1:

$$y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)} > 1$$

$$\Rightarrow \quad 1 - y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)} < 0$$

$$\Rightarrow \max\{1 - y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)}, 0\} = 0$$



Case 2:

$$y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)} < 1$$

No update to $\bar{\theta}$ is made if

A. $y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}) > 1$

B. $y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}) < 1$

C. *unsure*

https://forms.gle/ffiBvNbPjHF8ghi77

$$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} - \eta \nabla_{\bar{\theta}} \max\{1 - y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)}, 0\}$$

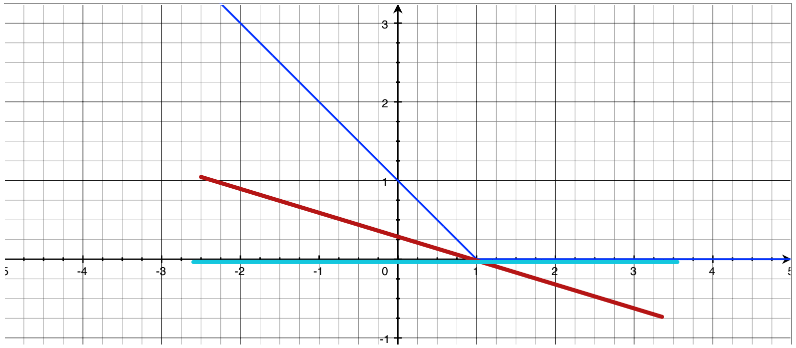
Stochastic Gradient Descent

$$\nabla_{\bar{\theta}} \max\{1 - y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}), 0\}$$

Case 1:

$$y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)} \geq 1$$

- Loss is 0
- Gradient is 0
- No update is made



subgradient method

Case 2:

$$y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)} < 1$$

$$\Rightarrow \max\{1 - y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)}, 0\} = 1 - y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)}$$

$$\nabla_{\bar{\theta}}(1 - y^{(i)}\bar{\theta} \cdot \bar{x}^{(i)}) = -y^{(i)}\bar{x}^{(i)}$$

$$\Rightarrow \bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} + \eta\, y^{(i)}\bar{x}^{(i)}$$

Stochastic Gradient Descent

$k = 0, \bar{\theta}^{(0)} = \bar{0}$

while convergence criteria* are not met

randomly shuffle points

for $i = 1, \ldots, n$

if $y^{(i)} \left(\bar{\theta}^{(k)} \cdot \bar{x}^{(i)}\right) < 1$

$$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} + \eta y^{(i)} \, \bar{x}^{(i)}$$

typically use variable step size

$k++$

* could check this in every update or every k' updates

Looks a lot like the perceptron algorithm!

Differences?

Convergence criteria

1. Keep track of $R_n(\bar{\theta})$

stop when less than some amount

2. Keep track of $\nabla_{\bar{\theta}} R_n(\bar{\theta})$

stop when less than some amount

don't do this too often; or defeats the purpose of SGD!

3. Keep track of $\bar{\theta}$

stop when doesn't change by much

4. Keep track of number of iterations

stop when max # iterations reached

Stochastic Gradient Descent

with appropriate learning rate, since $R_n(\bar{\theta})$ is convex, will almost surely converge to global minimum

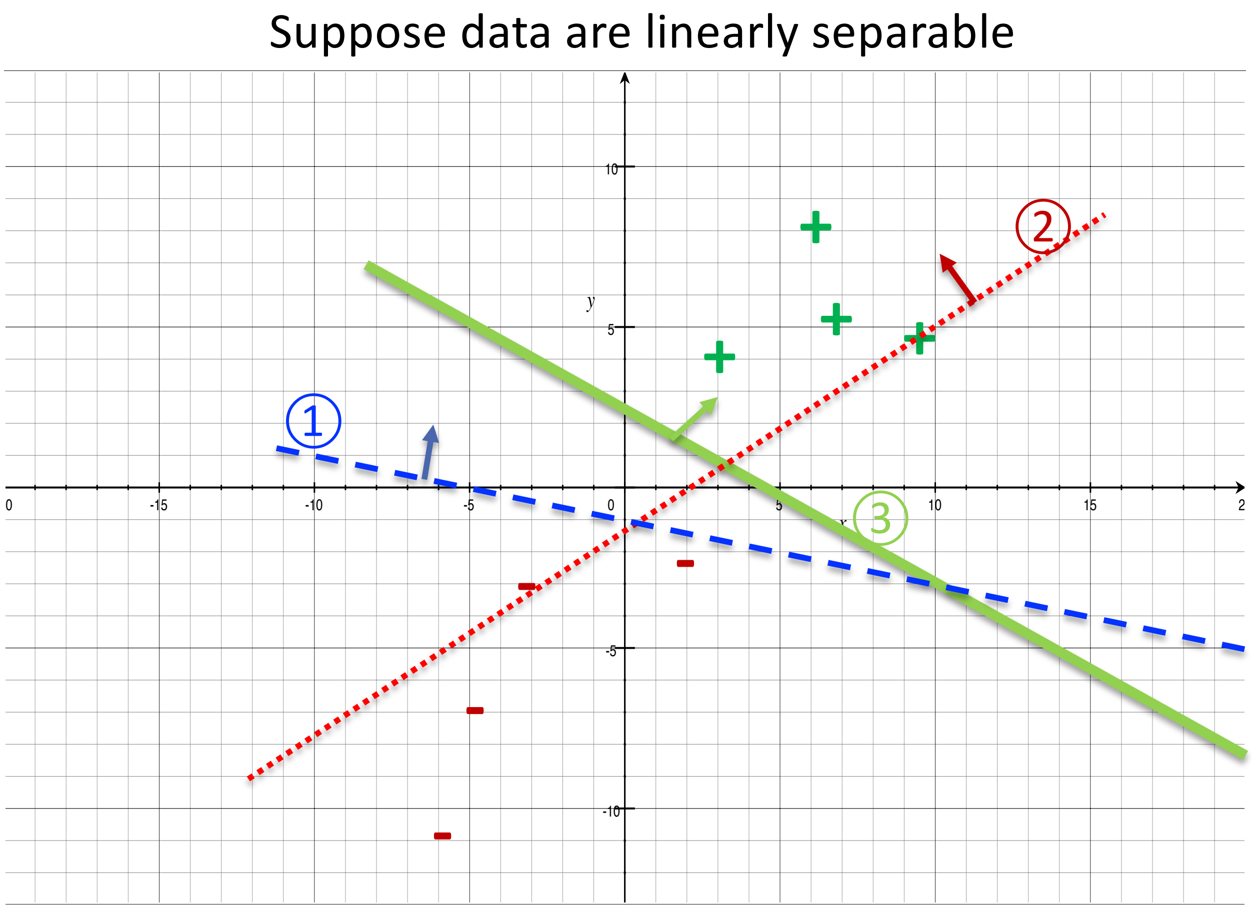
SGD often gets close to minimum faster than GD

Note: can be applied to non-convex functions

compared to GD, SGD more sensitive to step size
may never “converge” to the minimum
   parameters may keep oscillating around the minimum
   in practice most of the values near the minimum will be reasonably good approximations to true minimum

Support Vector Machines

Picking a decision boundary

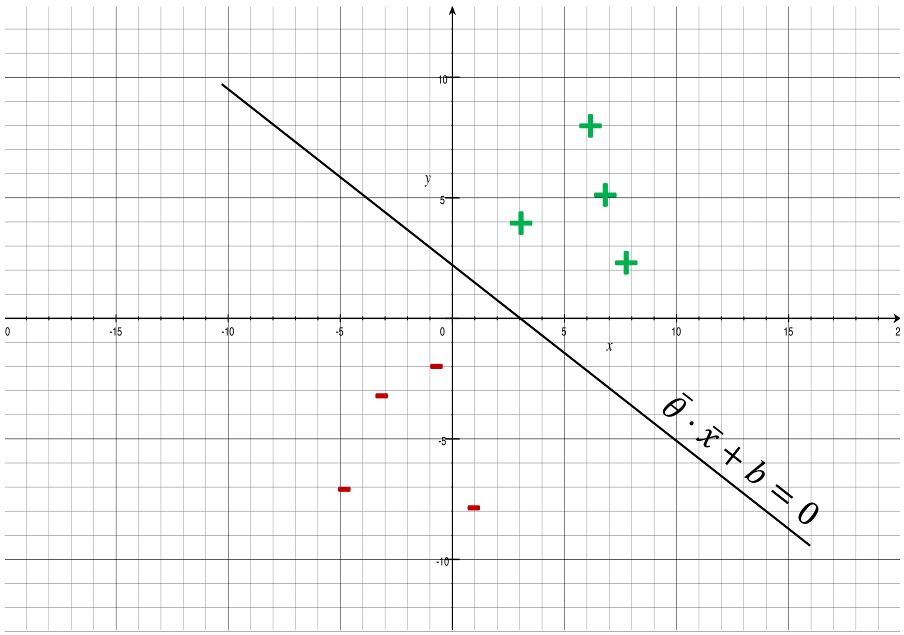


Want:

- Boundary that classifies the training set correctly and,
- That is maximally removed from training examples closest to the decision boundary

Maximum Margin Separator as an optimization problem

Assume data are linearly separable



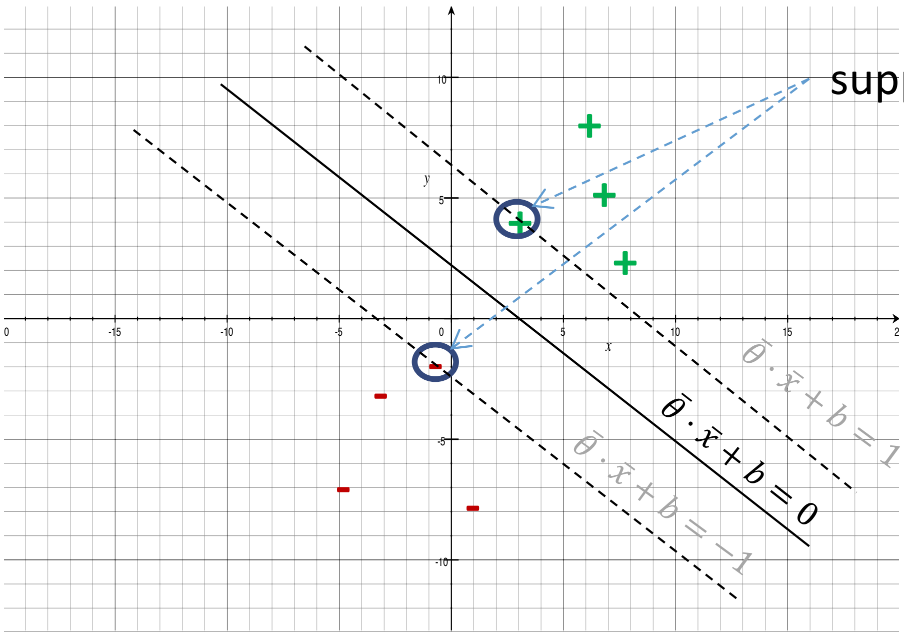
Want:

- Boundary that classifies the training set correctly and,
- That is maximally removed from training examples closest to the decision boundary

Want $\bar{\theta}$, b such that $$y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)} + b) \geq 1$$ for $i \in \{1, \dots, n\}$

Maximum Margin Separator as an optimization problem

Assume data are linearly separable



support vectors lie on margin boundaries

$\gamma^{(i)}(\bar{\theta}, b)$

distance between datapoint i, $\bar{x}^{(i)}$ and the hyperplane $\bar{\theta}\cdot\bar{x}+b=0$

$$\min_i \gamma^{(i)}(\bar{\theta}, b)$$

$$= \min_i \frac{y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)}+b)}{\|\bar{\theta}\|}$$

$$= \frac{1}{\|\bar{\theta}\|}$$

$$\max_{\bar{\theta},b} \min_i \gamma^{(i)}(\bar{\theta}, b) \text{ subject to } y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)}+b) \geq 1$$

maximizes the minimum distance to any of the datapoints in my training dataset

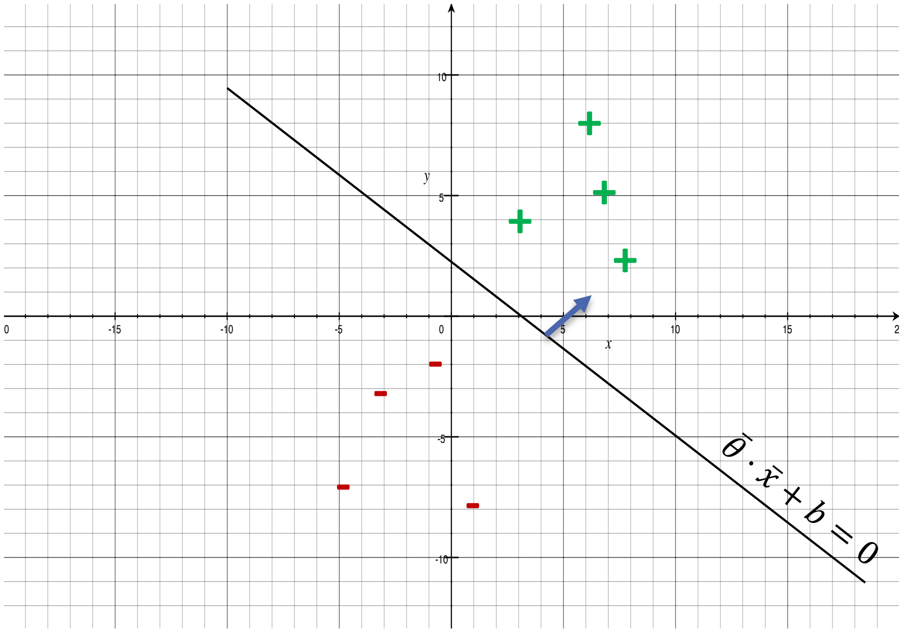
for $i \in \{1, \dots, n\}$

$$\max_{\bar{\theta},b} \frac{1}{\|\bar{\theta}\|} \quad \text{s.t.} \quad y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)}+b) \geq 1 \ \forall i$$

$$\min_{\bar{\theta},b} \frac{\|\bar{\theta}\|^2}{2} \quad \text{s.t.} \quad y^{(i)}\, \bar{\theta}\cdot\bar{x}^{(i)}+b \geq 1 \ \forall i$$

Hard Margin SVM

Assuming data are linearly separable



Linear classifier output by this QP: $sign(\bar{\theta} \cdot \bar{x} + b)$

Quadratic Program

$$\min_{\bar{\theta}, b} \frac{\|\bar{\theta}\|^2}{2} \text{ subject to } y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)} + b\right) \geq 1 \text{ for } i \in \{1, \dots, n\}$$

Linear Separability

What if data are not linearly separable?

How can we handle such cases?

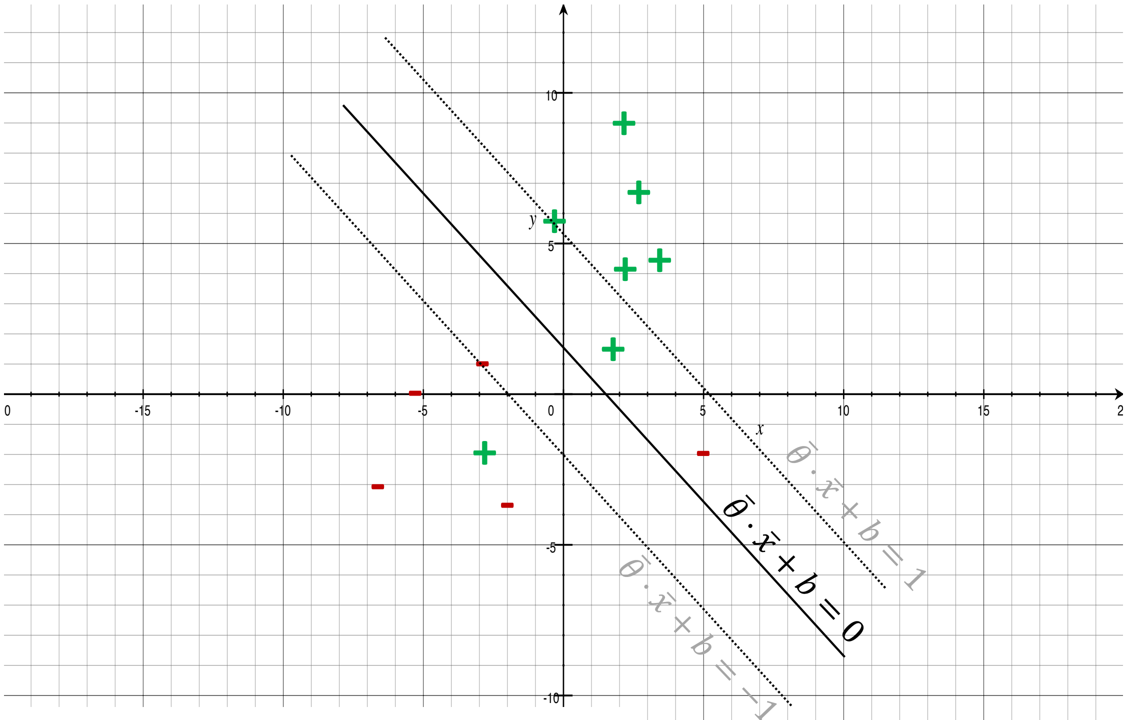
1. Constraints seem too restrictive
   – Fix the constraints: Soft-Margin SVMs

2. Map to a higher dimensional space

Section 3: Soft Margin SVMs

Soft-Margin SVM

Suppose data are *not* linearly separable



Hard margin SVM formulation:

$$\min_{\bar{\theta}, b} \frac{\|\bar{\theta}\|^2}{2}$$

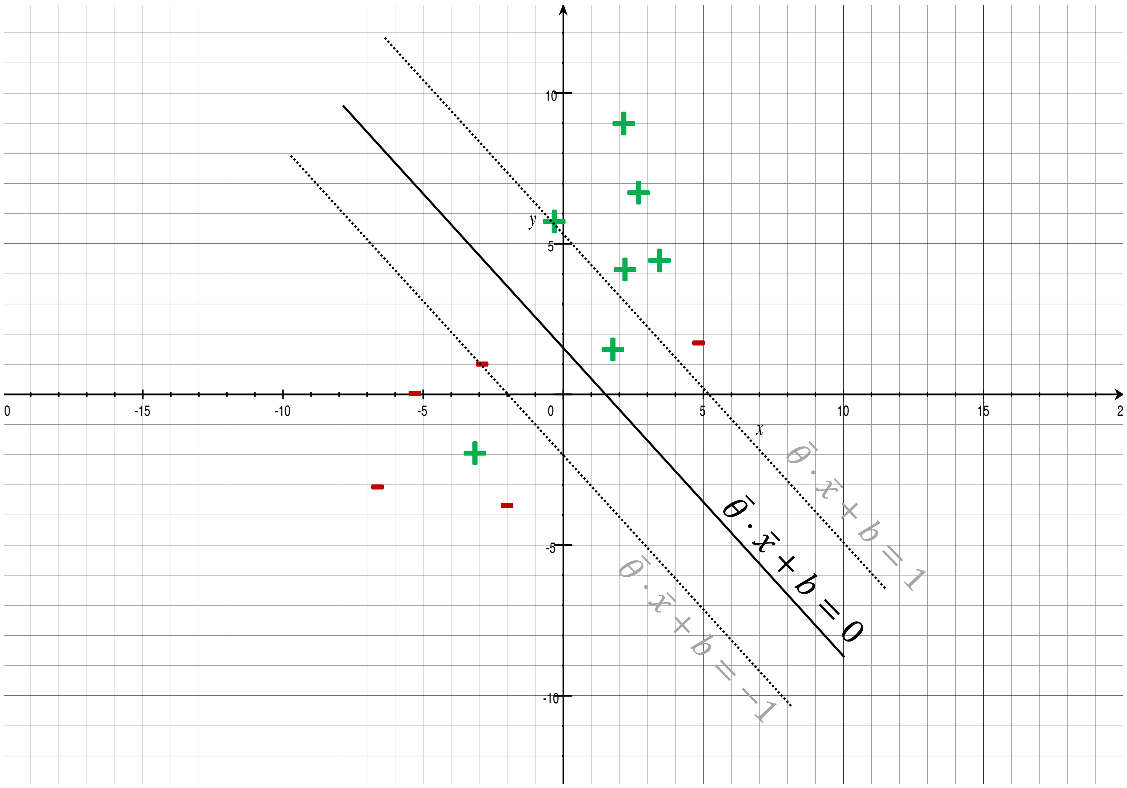
subject to $y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)} + b\right) \geq 1$

for $i \in \{1, \dots, n\}$

What goes wrong?

Soft-Margin SVM

Suppose data are *not* linearly separable



$$\min_{\bar{\theta},b,\bar{\xi}} \frac{\|\bar{\theta}\|^2}{2} + C\sum_{i=1}^{n} \boxed{\xi_i}$$ slack variables

$\bar{\theta} \in \mathbb{R}^d; b \in \mathbb{R}$

$\bar{\xi} \in$

subject to $\boxed{y^{(i)}\left(\bar{\theta}\cdot\bar{x}^{(i)} + b\right) \geq 1 - \xi_i}$ soft constraints

for $i \in \{1, \ldots, n\}$ and