Problem 3

Stochastic gradient descent (SGD) is a simple but widely applicable optimization technique. For example, we can use it to train a Support Vector Machine. The objective function in this case is given by:

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

where $\text{Loss}_h(z)=\max\{0,1-z\}$ is the hinge loss function, $(x^{(i)},y^{(i)})$ with for $i=1,\ldots n$ are the training examples, with $y^{(i)}\in\{1,-1\}$ being the label for the vector $x^{(i)}$.

For simplicity, we ignore the offset parameter θ_0 in all problems on this page.

3. (1)

3.0/3 points (graded)

The stochastic gradient update rule involves the gradient $\nabla_\theta\text{Loss}_h\left(y^{(i)}\theta\cdot x^{(i)}\right)$ of $\text{Loss}_h\left(y^{(i)}\theta\cdot x^{(i)}\right)$ with respect to θ.

Hint:Recall that for a k-dimensional vector $\theta=\begin{bmatrix}\theta_1 & \theta_2 & \cdots & \theta_k\end{bmatrix}^T$, the gradient of $f(\theta)$ w.r.t. θ is $\nabla_\theta f(\theta)=\begin{bmatrix}\frac{\partial f}{\partial\theta_1} & \frac{\partial f}{\partial\theta_2} & \cdots & \frac{\partial f}{\partial\theta_k}\end{bmatrix}^T$.)

Find $\nabla_\theta\text{Loss}_h(y\theta\cdot x)$ in terms of x.

(Enter `lambda` for λ, `y` for y and `x` for the vector x. Use `*` for multiplication between scalars and vectors, or for dot products between vectors. Use `0` for the zero vector.)

For $y\theta\cdot x\leq 1$:

$\nabla_\theta\text{Loss}_h(y\theta\cdot x)=$ -y * x ✓ **Answer:** -y*x

For $y\theta\cdot x>1$:

$\nabla_\theta\text{Loss}_h(y\theta\cdot x)=$ 0 ✓ **Answer:** 0

Let θ be the current parameters. What is the stochastic gradient update rule, where $\eta>0$ is the learning rate? (Choose all that apply.)

$\theta\rightarrow$

☐ $\theta+\eta\nabla_\theta\left[\text{Loss}_h\left(y^{(i)}\theta\cdot x^{(i)}\right)\right]+\eta\lambda\theta$ for random $x^{(i)}$ with label $y^{(i)}$

☑ $\theta-\eta\nabla_\theta\left[\text{Loss}_h\left(y^{(i)}\theta\cdot x^{(i)}\right)\right]-\eta\lambda\theta$ for random $x^{(i)}$ with label $y^{(i)}$ ✓

☐ $\theta+\eta\nabla_\theta\left[\text{Loss}_h\left(y^{(i)}\theta\cdot x^{(i)}\right)\right]+\eta\nabla_\theta\left[\frac{\lambda}{2}\|\theta\|^2\right]$ for random $x^{(i)}$ with label $y^{(i)}$

☑ $\theta-\eta\nabla_\theta\left[\text{Loss}_h\left(y^{(i)}\theta\cdot x^{(i)}\right)\right]-\eta\nabla_\theta\left[\frac{\lambda}{2}\|\theta\|^2\right]$ for random $x^{(i)}$ with label $y^{(i)}$ ✓

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

□ $\theta - \eta \sum_{i=1}^{n} \nabla_\theta \left[\text{Loss}_h\left(y^{(i)}\theta \cdot x^{(i)}\right)\right] - \eta \nabla_\theta \left[\frac{\lambda}{2}\|\theta\|^2\right]$

✓

Correction note: July 29 15:00UTC In the earlier version:

- The conditions for the first answer boxes were written $y\theta \cdot x < 1$ and $y\theta \cdot x \geq 1$ instead of the current $y\theta \cdot x \leq 1$ and $y\theta \cdot x > 1$.
- The summation in the choices had an error: $\sum_{i=1}^{n}$ was written wrongly as $\sum_{i=1^n}$.

Grader is correct: The grader behaves as intended in this problem. If you get an input error, please check your answers carefully. You will also need to complete all parts of the question before the submit button will be un-grayed.

Correction note: July 30 03:00UTC In the earlier version, the question statement did not include the input instruction “ in term of x" nor “Use `0` for the zero vector."

Correction note: July 30 16:00UTC In the earlier version, the confirmation note that the grader is working was not included.

Grading Note: The original problem statement, before the correction, which divides the cases into $y\theta \cdot x < 1$ and $y\theta \cdot x \geq 1$ matches with the definition of hinge loss in lecture. The corrected version assigned the boundary case $y\theta \cdot x = 1$ differently. However, the main property of the hinge loss function is that the loss increases as its argument becomes more negative, and the boundary case if not important. Hence, even with this mismatch, we have decided to proceed with the grading as intended originally. If you had switch the orders of the inputs, the grader would have thrown an error, and the “Grader is correct" note was a reminder to check your answers in this case.

STANDARD NOTATION

Solution:

The hinge loss function is defined as

$$\text{Loss}_h(z) = \begin{cases} 1 - z & \text{if } z < 1 \\ 0 & \text{if } z \geq 1. \end{cases}$$

Hence the gradient $\nabla_\theta \text{Loss}_h(y\theta \cdot x)$ is

$$\nabla_\theta \text{Loss}_h(y\theta \cdot x) = \begin{cases} \nabla_\theta(1 - y\theta \cdot x) = -y \cdot x & \text{if } z < 1 \\ 0 & \text{if } z \geq 1 \end{cases}.$$

The stochastic gradient algorithm update step is

$$\begin{aligned} \theta &\to \theta - \eta \nabla_\theta \left[\text{Loss}_h\left(y^{(i)}\theta \cdot x^{(i)}\right)\right] - \eta \nabla_\theta \left[\frac{\lambda}{2}\|\theta\|^2\right] \\ &= \theta - \eta \nabla_\theta \left[\text{Loss}_h\left(y^{(i)}\theta \cdot x^{(i)}\right)\right] - \eta\lambda\theta \end{aligned}$$

The first and third choices are incorrect because of wrong signs. The final two choices are incorrect: that is the update rule for the true gradient descent algorithm.

Substituting in the gradient, we get the update rule

$$\theta \to \begin{cases} (1 - \eta\lambda)\,\theta + \eta y^{(i)} x^{(i)} & \text{if } y^{(i)}\theta \cdot x^{(i)} \leq 1 \\ (1 - \eta\lambda)\,\theta & \text{if } y^{(i)}\theta \cdot x^{(i)} > 1. \end{cases}$$

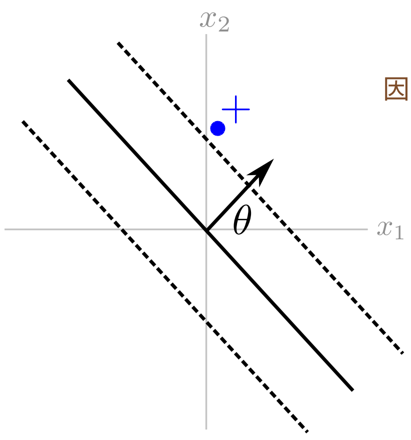
Submit You have used 2 of 3 attempts

Answers are displayed within the problem

3. (2)

1/1 point (graded)

Suppose the current parameter θ is as in the figure below:

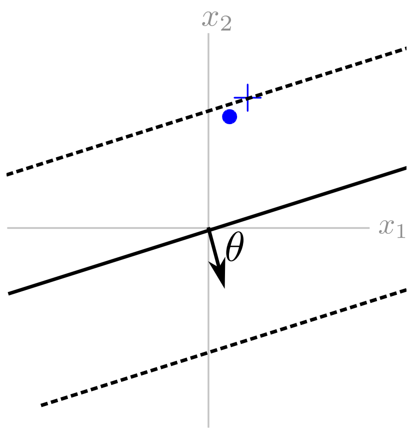


因为判断对了，所以θ不会变方向，regularization term向同方向推。

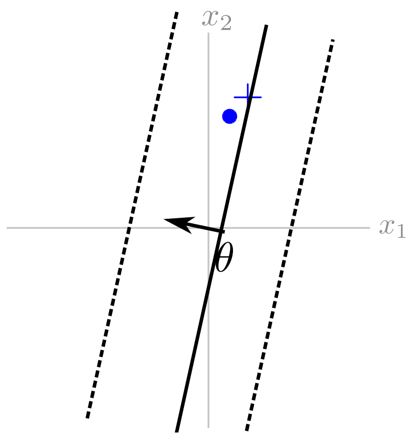
Here, θ is in the direction of the arrow, the solid line represents the classifer defined by $\theta,$ and the dotted lines represent the positive and negative margin boundaries.

For large η (i.e. η close to 1) $0.5 < \eta\lambda < 1$, which of the following figure corresponds to a single SGD update made in response to the point labeled '+' above?

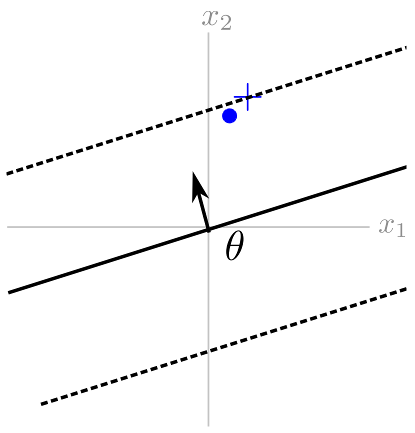
○



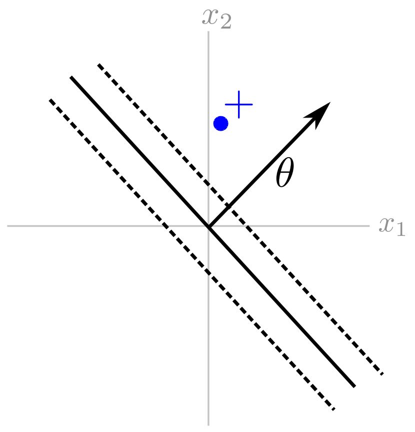
○



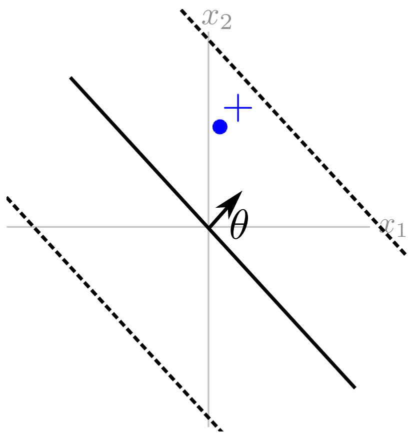
○



○



●



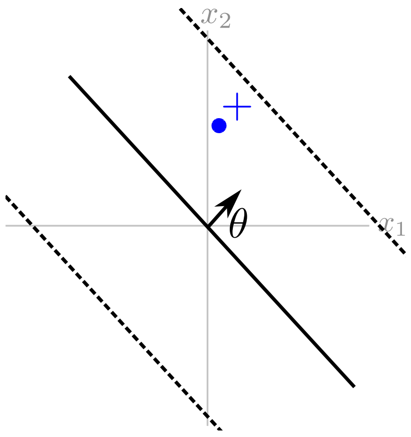
✓

Solution:

For the given θ and given point x with positive label, we have $y\theta \cdot x > 1$. Hence, the update step is as follows and does not depend on x

$$\theta \ \rightarrow \ \theta - \eta\lambda\theta$$

For $\eta\lambda = 0$, the update does not change θ. For $0 < \eta\lambda < 1$, θ is shrunk in length to by factor of $(1 - \eta\lambda)$ but remains in the same direction. As $\eta\theta$ increases from 0, the resulting parameter become shorter, which leads to the margin becoming larger.



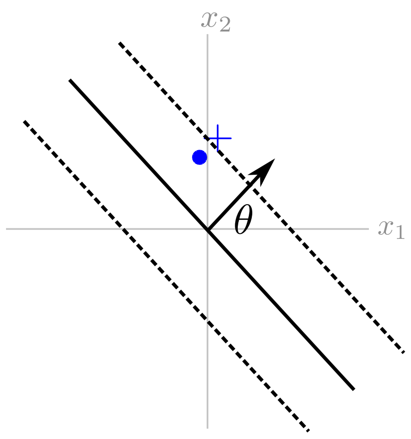
Submit You have used 2 of 3 attempts

Answers are displayed within the problem

3. (3)

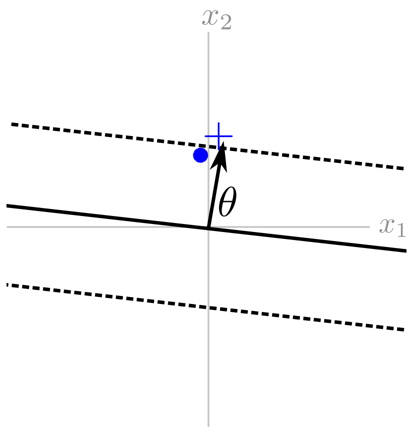
1/1 point (graded)

Again for large η (i.e. η close to 1) and $0.5 < \eta\lambda < 1$, but now we perform a single SGD update made in response to a different point labeled '+', shown below:



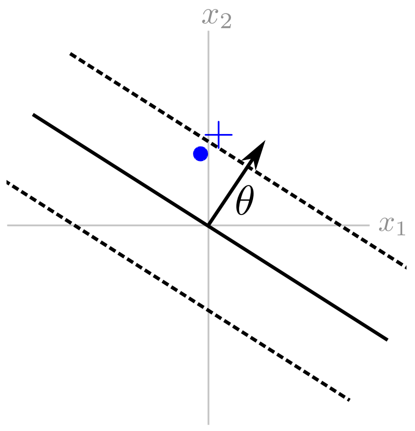
which of the following figure corresponds to a single SGD update made in response to the point labeled '+' above?

●

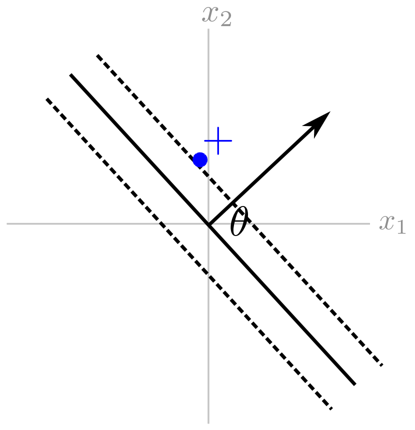


✓

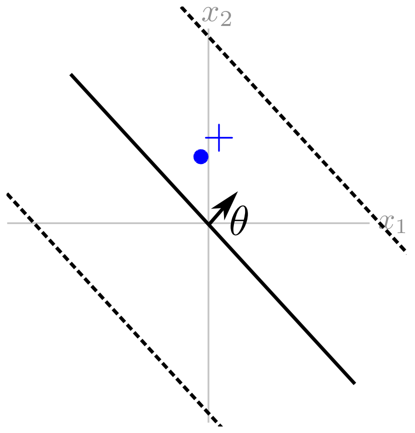
○



○



○

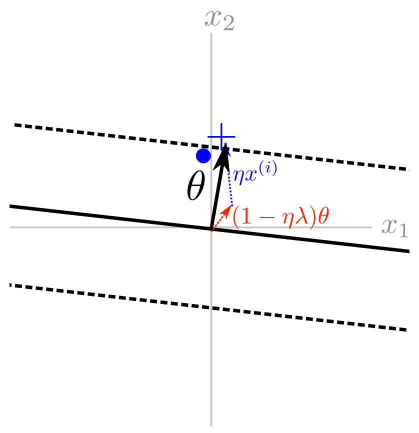


Solution:

In this case, the given positively labeled point x now satisfies $y\theta \cdot x \leq 1$, so the update rule is

$$\theta \;\rightarrow\; (1-\eta\lambda)\,\theta + \eta y^{(i)} x^{(i)}.$$

For large η and $0.5 < \eta\lambda < 1$, the update changes the direction of θ significantly toward $x^{(i)}$, and hence we get



The second is for the case when both η and and λ are small (approach 0), and the update does not alter θ (or the margin) significantly.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Error and Bug Reports/Technical Issues

Show Discussion

Topic: Midterm Exam (1 week):Midterm Exam 1 / Problem 3