

EdX and its Members use cookies and other tracking technologies for performance, analytics, and marketing purposes. By using this website, you accept this use. Learn more about these technologies in the [Privacy Policy](#).



[Course](#) > [Midter...](#) > [Midter...](#) > [Proble...](#)

## Problem 3

Midterm due Nov 10, 2020 05:29 IST *Completed*

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:

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

where  $\text{Loss}_h(z) = \max\{0, 1 - z\}$  is the hinge loss function,  $(x^{(i)}, y^{(i)})$  with for  $i = 1, \dots, n$  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  $\theta_0$  in all problems on this page.

### 3. (1)

2.0/3 points (graded)

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

*Hint:* Recall that for a  $k$ -dimensional vector  $\theta = [\theta_1 \ \theta_2 \ \dots \ \theta_k]^T$ , the gradient of  $f(\theta)$  w.r.t.  $\theta$  is  $\nabla_{\theta} f(\theta) = \left[ \frac{\partial f}{\partial \theta_1} \ \frac{\partial f}{\partial \theta_2} \ \dots \ \frac{\partial f}{\partial \theta_k} \right]^T$ .

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

(Enter `lambda` for  $\lambda$ , `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  $y\theta \cdot x \leq 1$ :

$$\nabla_{\theta} \text{Loss}_h(y\theta \cdot x) =$$

✓ Answer: -y\*x

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

$$\nabla_{\theta} \text{Loss}_h(y\theta \cdot x) =$$

✓ Answer: 0

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

$\theta \leftarrow$

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

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

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

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

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

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

✗

**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.

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

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

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

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

Submit

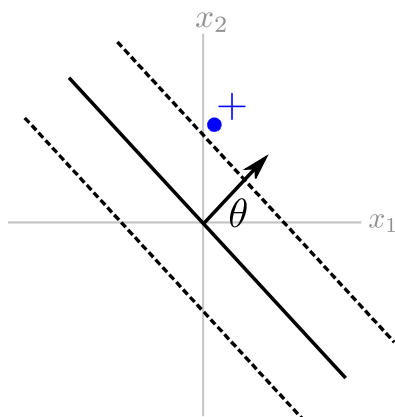
You have used 1 of 3 attempts

**i** Answers are displayed within the problem

3. (2)

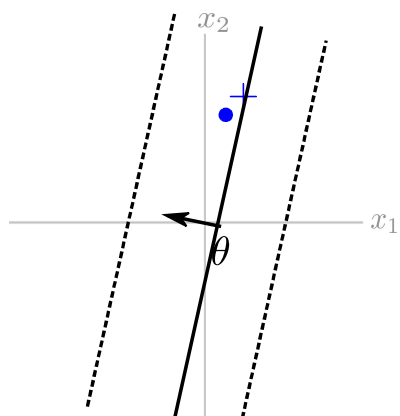
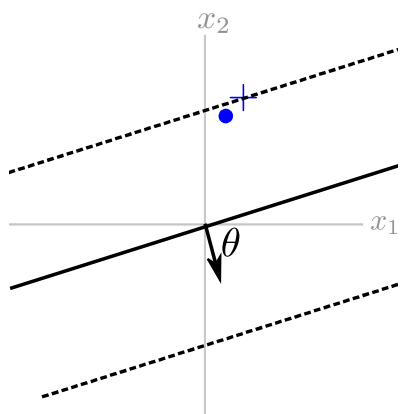
0/1 point (graded)

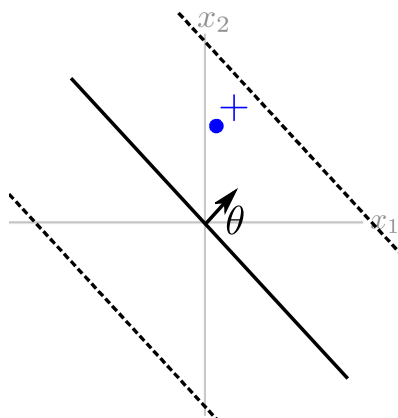
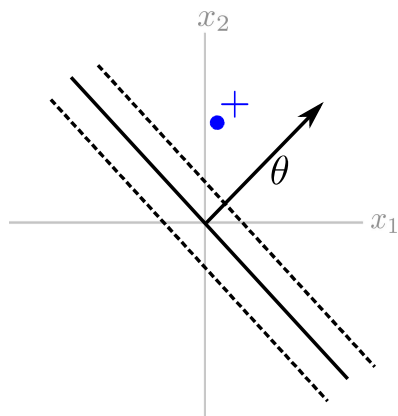
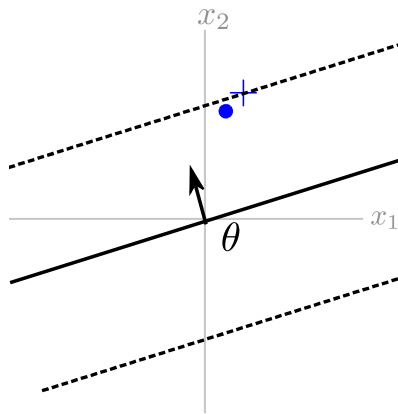
Suppose the current parameter  $\theta$  is as in the figure below:



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

For large  $\eta$  (i.e.  $\eta$  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?



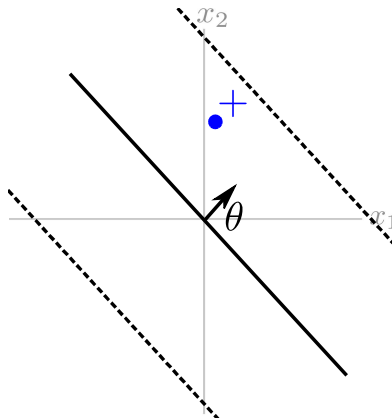


### Solution:

For the given  $\theta$  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  $\theta$ . For  $0 < \eta\lambda < 1$ ,  $\theta$  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.



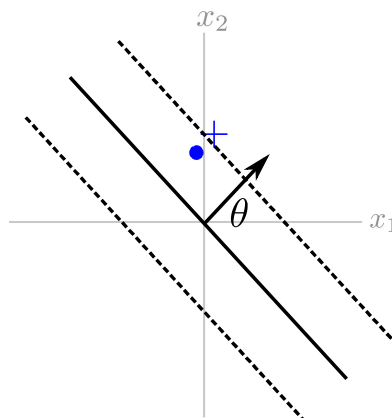

You have used 1 of 3 attempts

**i** Answers are displayed within the problem

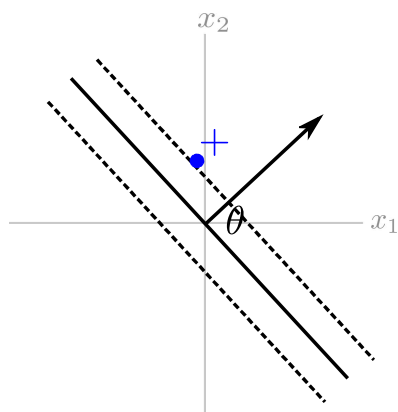
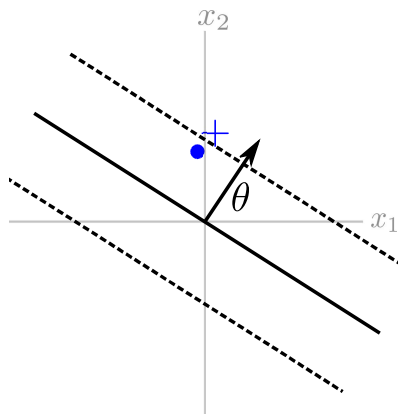
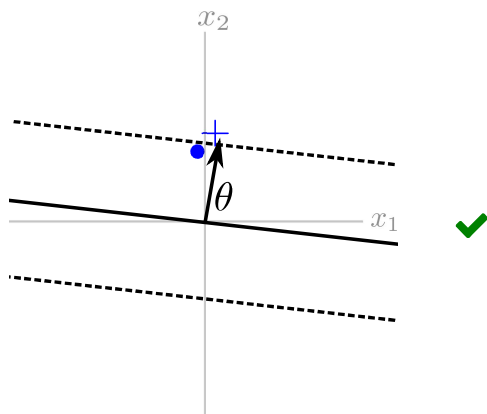
### 3. (3)

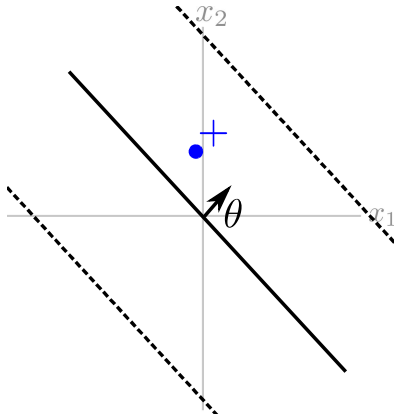
0/1 point (graded)

Again for large  $\eta$  (i.e.  $\eta$  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?



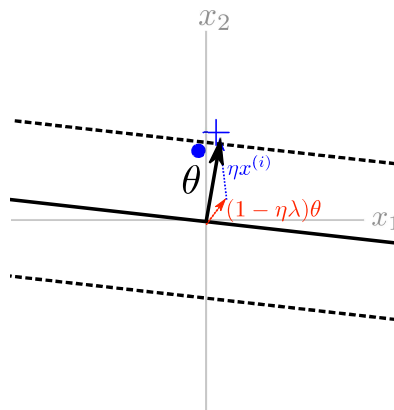


### Solution:

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



The second is for the case when both  $\eta$  and  $\lambda$  are small (approach 0), and the update does not alter  $\theta$  (or the margin) significantly.

Submit

You have used 1 of 3 attempts













 Answers are displayed within the problem

Error and Bug Reports/Technical Issues

Hide Discussion

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

Add a Post

Show all posts	by recent activity
 <u>STAFF: 3-1 grading for getting 1 out of 2</u>	3
<u>I notice that I got 1 out of the 2 in the multiple choice in Problem 3-1 . Surely the Grader should allow...</u>	
 <u>3.(3) Question</u>	19
<u>Can we assume the product <math>\eta \cdot \lambda</math> closer to 1 than to 0.5?</u>	
 <u>Question 3: What were your answers?</u>	3
 <u>[staff] 3.(1). Clarification needed.</u>	4
<u>The problem statement mentions SGD. The question 3.(1) mentions stochastic gradient, without refe...</u>	
 <u>Clarification with figures</u>	2
<u>Could you please clarify : In both the figures there is a single point denoted with a dot? The "+" is just...</u>	
 <u>Clarification question about Theta0-</u>	2
 <u>[STAFF] Isn't the update arrow the wrong way round in Q3.(1)?</u>	3
 <u>Use 'xxx' as zero vector.</u>	3
<u>Question 1, asks us to input 'something similar to phi' as zero vector. What should we actually type i...</u>	
 <u>Question 3(1)</u>	3
<u>I find the phrasing of the question a bit confusing. Should the answer to "Find <math>\nabla_{\theta} \text{Loss}_h(y_{\theta}; x)</math> in terms...</u>	
 <u>Invalid Input: '\lambda\' not permitted in answer as a variable</u>	4
<u>Why I am getting the error while Submitting the answer Invalid Input: '\lambda\' not permitted in an...</u>	

© All Rights Reserved