Q1
1 Point

Consider stochastic gradient descent used to minimize empirical risk on a data set of size n. True or False: if the batch size equals n, stochastic gradient descent is equivalent to gradient descent.

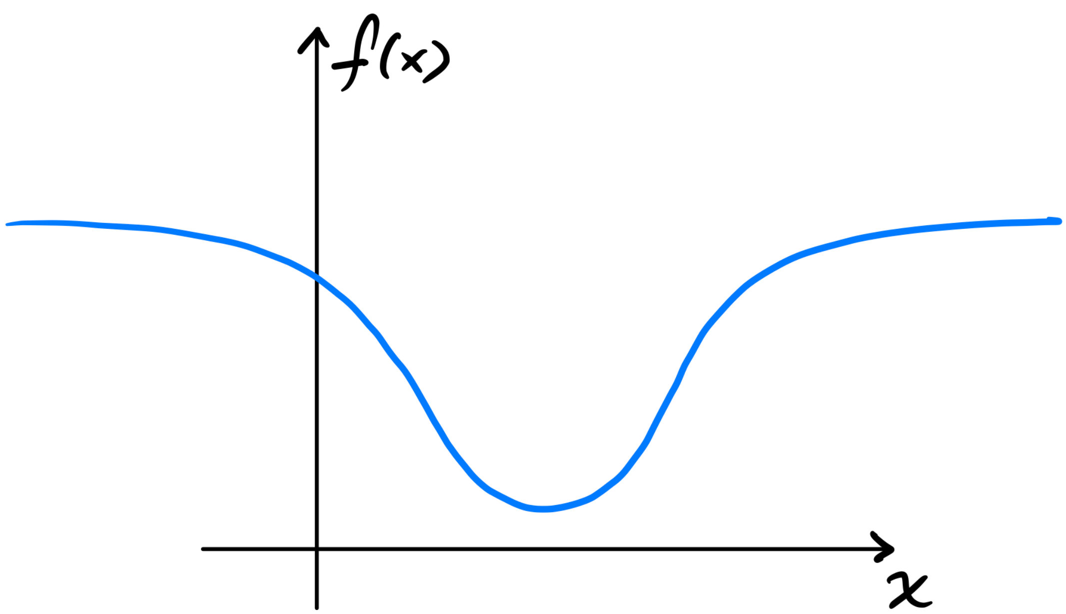
You may assume that the batch of points is sampled uniformly at random *without* replacement.

True

False

Q2
1 Point

Is the function shown below convex or non-convex?



Convex

Non-convex

Q3

1 Point

Define $f(x) = \sum_{i=1}^{k} \alpha_i e^{\beta_i(x-\mu)}$, where k is a postive integer, μ is a real number, and α_i (for $i \in \{1, 2, \ldots, k\}$) is a *positive* real number, and β_i (for $i \in \{1, 2, \ldots, k\}$ is a (possibly-negative) real number.

Is f convex (no matter how α_i, β_i, μ are chosen) or is it possibly non-convex?

Convex

Non-convex

Q4

1 Point

Recall the 0-1 loss from lecture. Define $R_{01}(\vec{w})$ to be the risk of a linear predictor $\vec{w} \cdot \vec{\text{Aug}}(x)$ on a data set. Is R_{01} convex or non-convex as a function of $\vec{w}$?

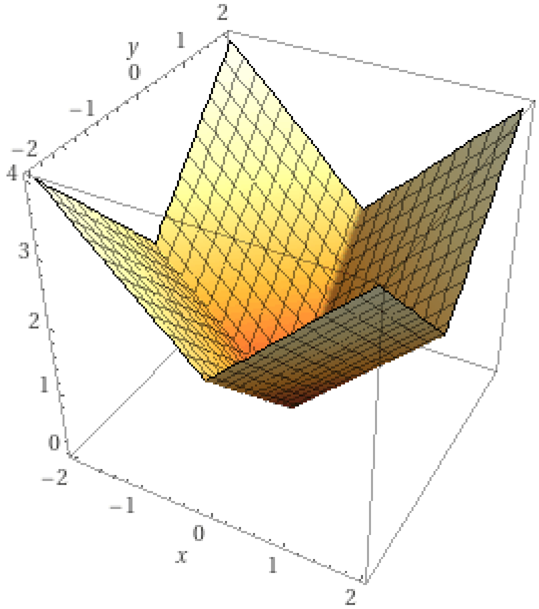
Convex

Non-convex

Q5

1 Point

Consider the function $f(x, y) = |x| + |y|$ from everyone's favorite problem on Lab 02. This function is plotted below for your convenience:



Consider the point $(2, 0)$. Which of the below are valid subgradients at this point? Mark all which apply.

- [ ] $(2, 1)^T$
- [ ] $(0, 1)^T$
- [ ] $(-1/2, 1)^T$
- [ ] $(-1, 1)^T$
- [ ] $(1, 2)^T$
- [x] $(1, 0)^T$
- [x] $(1, -1/2)^T$
- [x] $(1, -1)^T$

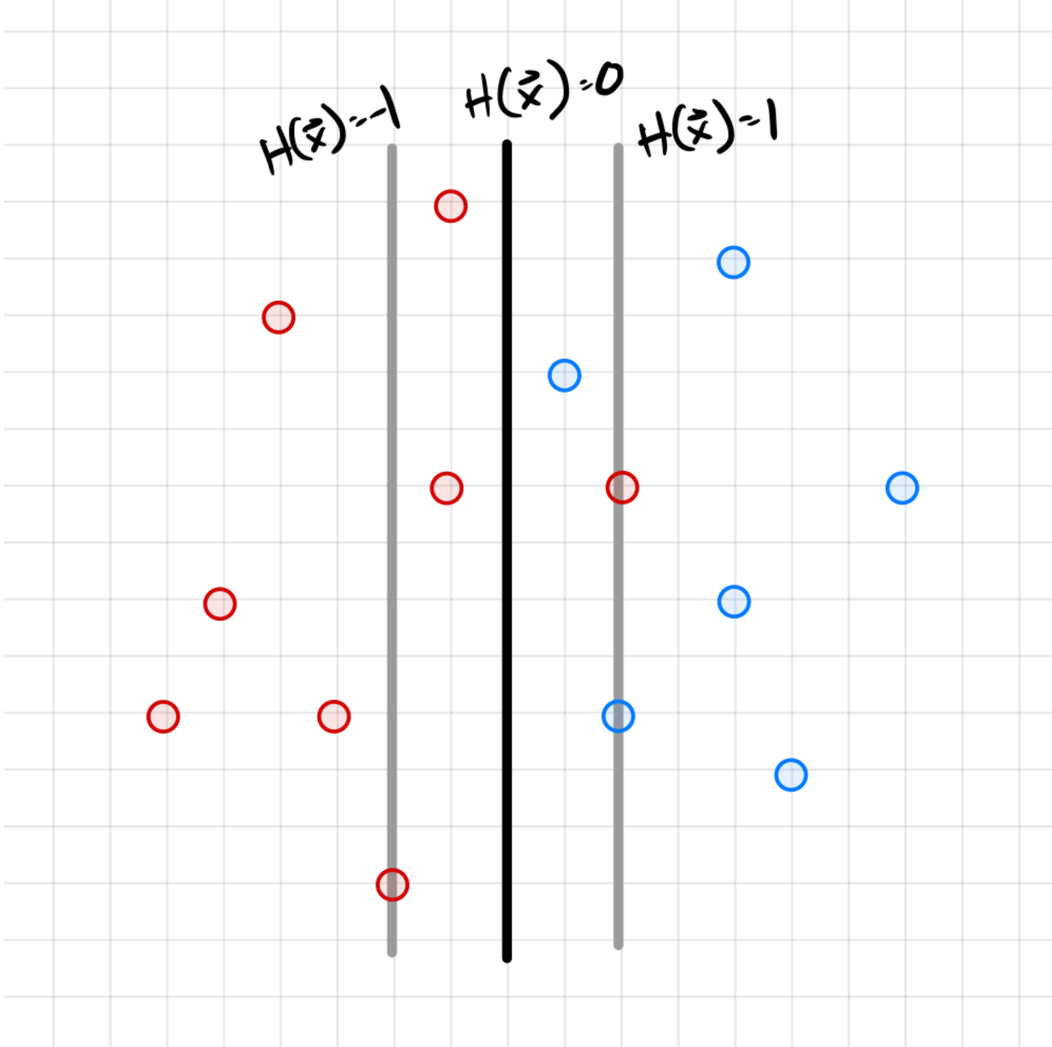
Explanation

In the x direction, the slope of f is well-defined: it's 1. So any valid subgradient needs to have 1 in its x component. In the y direction, the slope of the function is undefined. For a plane attached to the surface at $(2, 0)$ to stay below the surface itself, the plane's slope in the y direction can be anything between 1 and -1. This means that (1,2) is not a valid subgradient, because the plane defined by that subgradient would go above the plot of f.

Q6

1 Point

Consider the linear classifier depicted below. Shown is its decision boundary, as well as the lines where the output of the linear prediction function is 1 and -1. You can assume that the data and lines are placed precisely on the grid; this can be used to calculate distances. You may assume that the red points have correct label of -1, while the blue points have a correct label of +1.



What is the empirical risk of this classifier with respect to the hinge loss? Report your answer as a decimal number with two decimal places of precision.

=0.25+-.02

Explanation

The hinge loss will only penalize points that are 1) misclassified or 2) too close to the decision boundary (|H(x)| < 1). In this case this is only four points: the two red points immediately to the left of the decision boundary (they are too close to the boundary), the blue point closest to the decision boundary (also too close), and the red point that has been misclassified.

First, what is the loss of the two red points to the left of the boundary? On these two points, $H = -1/2$. Therefore, the loss is $1 - (-1)(-1/2) = 1/2$.

The blue point to the right of the boundary incurs a loss of 1/2 as well, since $H = 1/2$ on this point, and the hinge loss is therefore $1 - (1)(1/2) = 1/2$.

The misclassified red point incurs a loss of $1 - (-1)(1) = 2$, since $H(x) = 1$ for that point.

The total loss is therefore $2(1/2) + (1/2) + 2 = 3.5$. Since there are 14 points in total, the risk is 3.5 / 14 = 0.25

Q7

1 Point

True or False: a Soft-SVM trained on a linearly-separable training data set must achieve 100% training accuracy (it will correctly predict the label of all training points).

True

False