Notes: Stochastic Gradient Descent I

1 Pre-lecture Work

None. Get plenty of sleep and do well on all your midterms :)

2 Regularized Loss Minimization

Recall that in *empirical risk minimization* (ERM), we select a hypothesis according to the rule

$$\hat{h} = \arg\min_{h \in \mathcal{H}} L_S(h) \tag{1}$$

where

$$L_S(h) = \tfrac{1}{m} \sum_{z \in S} \ell(h, z). \tag{2}$$

In *regularized loss minimization* (RLM), we modify Eq 1 into

$$\hat{h} = \arg\min_{h \in \mathcal{H}} L_S(h) + \lambda R(h) \tag{3}$$

where R is called a regularization function and λ is called the regularization strength.

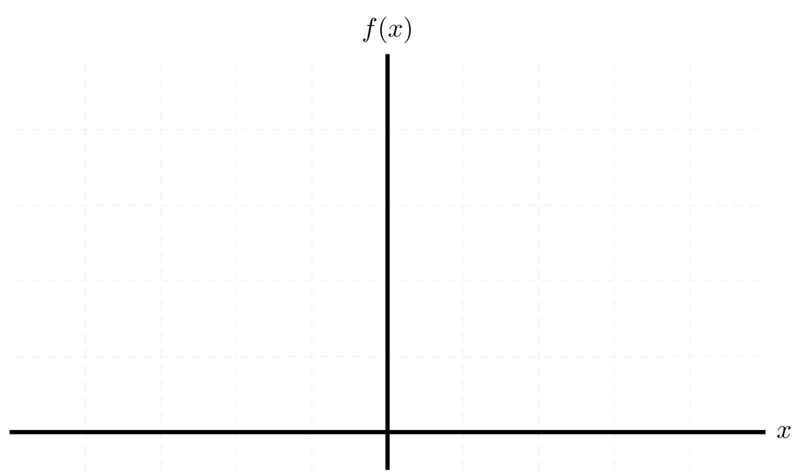
Problem 1. Regression loss functions are typically defined by the formula

$$\ell(h, (\mathbf{x}, y)) = f(h(\mathbf{x}) - y) \tag{4}$$

for some function f. The following table lists commonly used f functions and their properties.

Loss Name	$f(x)$	Convex	Strongly Convex	Lipschitz	Smooth
squared loss	$f(x) = \frac{1}{2}x^2$				
absolute loss	$f(x) = \|x\|$				
Huber loss	$f(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } \|x\| < 1 \\ \|x\| - \frac{1}{2} & \text{otherwise} \end{cases}$				

Plot each of the f functions below.



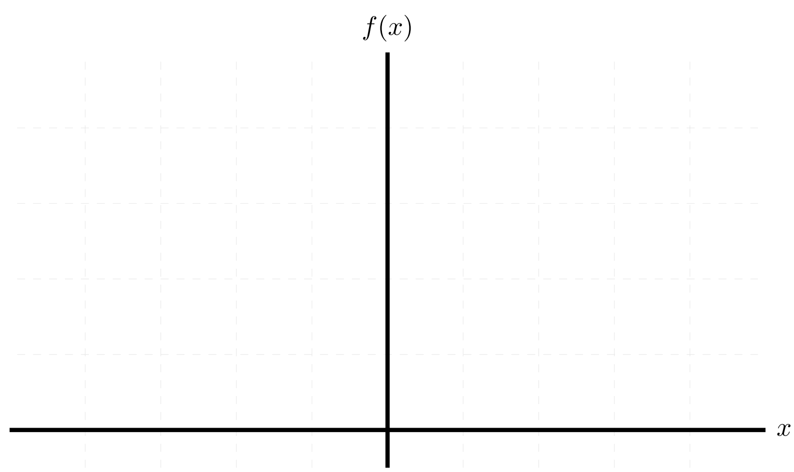
Problem 2. Binary classification loss functions are typically defined by the formula

$$\ell(\mathbf{w}, (\mathbf{x}, y)) = f(y\mathbf{w}^T\mathbf{x}) \tag{5}$$

for some function f. Notice that this formula does not mention a hypothesis h anywhere; instead, the vector $\mathbf{w}$ acts as the hypothesis. The following table lists commonly used f functions and their properties.

Loss Name	$f(x)$	Convex	Strongly Convex	Lipschitz	Smooth
0-1 loss	$f(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$				
exponential loss	$f(x) = \exp(-x)$				
logistic loss	$f(x) = \log(1 + \exp(-x))$				
hinge loss	$f(x) = \begin{cases} -x + 1 & \text{if } x < 1 \\ 0 & \text{otherwise} \end{cases}$				
sigmoid loss	$f(x) = \dfrac{1}{1 + \exp(x)}$				

Plot each of the f functions below.



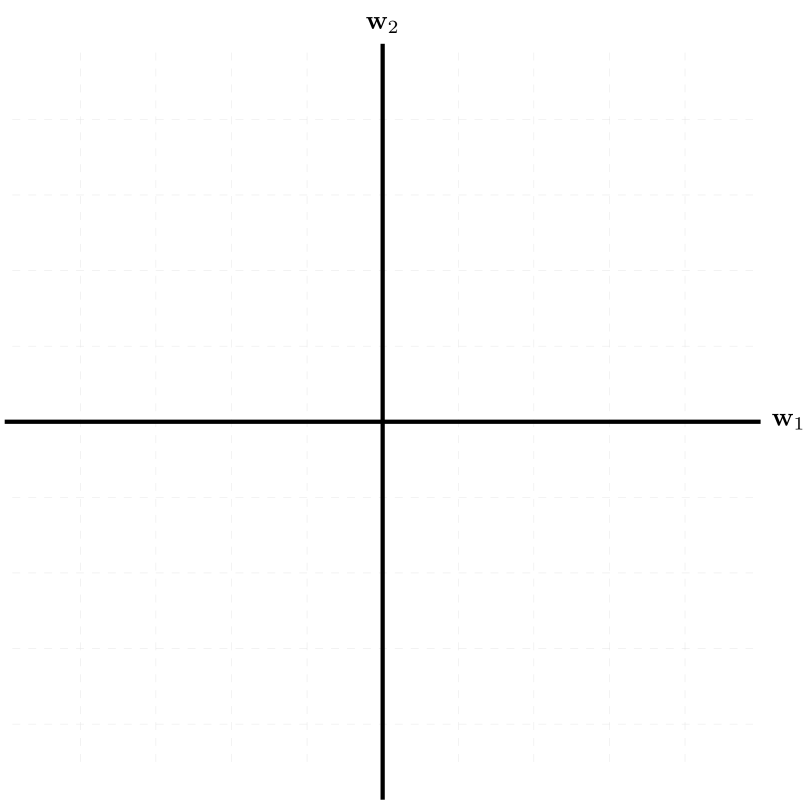
Problem 3. Regularization functions are typically defined by the formula

$$\ell(\mathbf{w}, (\mathbf{x}, y)) = f(y\mathbf{w}^T\mathbf{x}) \tag{6}$$

for some function f. Notice that this formula does not mention a hypothesis h anywhere; instead, the vector $\mathbf{w}$ acts as the hypothesis. The following table lists commonly used f functions and their properties.

$R(x)$	Convex	Strongly Convex	Lipschitz	Smooth
$R(\mathbf{w}) = \|\mathbf{w}\|_2^2$				
$R(\mathbf{w}) = \|\mathbf{w}\|_1$				
$R(\mathbf{w}) = \|\mathbf{w}\|_0$				
$R(\mathbf{w}) = (1-\alpha)\|\mathbf{w}\|_1 + \alpha\|\mathbf{w}\|_2^2$				

Plot each of the R functions below.



Problem 4. Convexity.

1. Definition 12.1 (Convex Set)

2. Definition 12.2 (Convex Function)

3. Lemma 12.3 (equivalent definitions of convex functions)

4. Claim 12.4

5. Claim 12.5

Problem 5. Strong convexity.

1. Definition 13.4 (strongly convex function)

2. Lemma 13.5

Problem 6. Lipschitzness.

1. Definition 12.6 (Lipschitz function)

2. Claim 12.7

Problem 7. Smoothness.

1. Definition 12.8 (Smooth functions)

2. Claim 12.9

3. Subgradients (Section 14.2)