CSE 512: Homework 2 — Due Wednesday September 25

Extra practice problems, ungraded

1. *Gradients.* Compute the gradients of the following functions. Give the exact dimension of the output.

   (a) *Linear regression.* $f(x) = \frac{1}{40}\|Ax - b\|_2^2$, $A \in \mathbb{R}^{20\times 10}$
   Ans. Actually, the best way to do this is to invoke the chain rule, which you will prove in the first graded problem. Write $g(v) = \frac{1}{40}\|v-b\|_2^2$. Then since $b \in \mathbb{R}^{20}$,

$$\nabla g(v) = \nabla_v \left(\frac{1}{40}\sum_{i=1}^{20}(v[i]-b[i])^2\right) \overset{\text{linearity}}{=} \frac{1}{40}\sum_{i=1}^{20}\nabla_v\left((v[i]-b[i])^2\right).$$

   Note that

$$\nabla_v (v[i]-b[i])^2 = \begin{bmatrix} \frac{\partial}{\partial v[1]}(v[i]-b[i])^2 \\ \frac{\partial}{\partial v[2]}(v[i]-b[i])^2 \\ \vdots \\ \frac{\partial}{\partial v[20]}(v[i]-b[i])^2 \end{bmatrix}$$

   and

$$\frac{\partial}{\partial v[k]}(v[i]-b[i])^2 = \begin{cases} 2(v[i]-b[i]) & \text{if } i = k \\ 0 & \text{else.} \end{cases}$$

   So,

$$\sum_{i=1}^{20}\nabla_v (v[i]-b[i])^2 = 2\begin{bmatrix} (v[1]-b[1]) \\ (v[2]-b[2]) \\ \vdots \\ (v[20]-b[20]) \end{bmatrix} = 2(v-b).$$

   and $\nabla g(v) = \frac{1}{20}(v-b)$.
   Now, we invoke the chain rule. (Note that f and g are flipped as to their position in 1.(b).) Then

$$\nabla f(x) = A^T \nabla g(Ax) = A^T(\frac{1}{20}(Ax-b)) = \frac{1}{20}A^T(Ax-b).$$

   To get the dimension, you can do this in two ways. One, you notice that A has 10 columns, so A^T has 10 rows. Two, you notice that the gradient $\nabla f(x)$ should always have the same number of elements as x, which is 10. In either case, $\nabla f(x) \in \mathbb{R}^{10}$.

   (b) *Sigmoid.* $f(x) = \sigma(c^T x)$, $c \in \mathbb{R}^5$, $\sigma(s) = \frac{1}{1+\exp(-x)}$. Hint: Start by showing that $\sigma'(s) = \sigma(s)(1-\sigma(s))$.
   Ans. We start with the hint, noting that

$$\sigma'(s) = \frac{\exp(-x)}{(1+\exp(-x))^2} = \frac{1}{1+\exp(-x)}\cdot\left(1 - \frac{1}{1+\exp(-x)}\right) = \sigma(s)(1-\sigma(s)).$$

   Then using chain rule, (where $A = c^T$) we can get

$$\nabla f(x) = \sigma'(c^T x)c = \sigma(c^T x)(1-\sigma(c^T x))c \in \mathbb{R}^5.$$

Main assignment, graded

1. **(1 pts, 0.5 pts each)** *Linearity.* A function $f : \mathbb{R}^n \to \mathbb{R}$ is *linear* if for any x and y in the domain of f, and any scalar α and β,

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y).$$

   Are the following functions linear? Justify your answer.

   (a) $f(x) = \|x\|_2^2$

   (b) $f(x) = c^T x + b^T A x$

2. **(1 pt, 0.5 each)** A function $f : \mathbb{R}^n \to \mathbb{R}$ is a norm if it satisfies three properties:

   - Nonnegativity: $f(x) \geq 0$ for all x and $f(x) = 0$ only when $x = 0$
   - Positive homogeneity $f(\alpha x) = \alpha f(x)$ whenever $\alpha \geq 0$
   - Triangle inequality $f(x+y) \leq f(x) + f(y)$.

   Using the properties of norms, verify that the following are norms, or prove that they are not norms by finding a counterexample.

   (a) *Sum of square roots, squared.* $f : \mathbb{R}^d \to \mathbb{R}$, $f(x) = \left(\sum_{k=1}^d \sqrt{|x[k]|}\right)^2$

   (b) *Weighted 2-norm.* $f : \mathbb{R}^d \to \mathbb{R}$, $f(x) = \sqrt{\sum_{k=1}^d \frac{|x[k]|^2}{k}}$

3. *Gradient properties.* **(1 pt, 0.5 pts each.)** Prove the following two properties of gradients:

   (a) *Linearity.* If $h(x) = \alpha f(x) + \beta g(x)$, then $\nabla h(x) = \alpha \nabla f(x) + \beta \nabla g(x)$.

   (b) *Chain rule.* Show that if $g(v) = f(Av)$, then $\nabla g(v) = A^T \nabla f(Av)$.

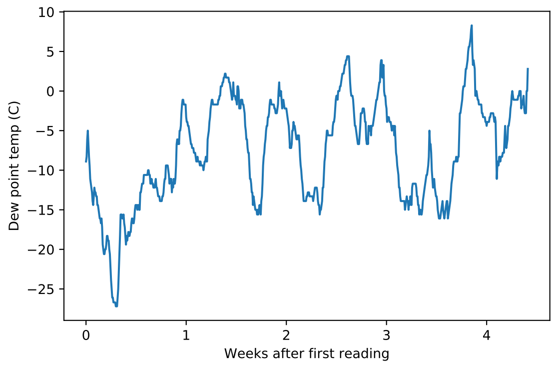
4. *Gradients.* **(2 pts, 1 pt each.)** Compute the gradients of the following functions. Give the exact dimension of the output.

   (a) *Quadratic function.* $f(x) = \frac{1}{2} x^T Q x + p^T x + r$, $Q \in \mathbb{R}^{12 \times 12}$ and Q is symmetric ($Q[i,j] = Q[j,i]$).

   (b) *Softmax function.* $f(x) = \frac{1}{\mu} \log(\sum_{i=1}^8 \exp(\mu x[i]))$, $x \in \mathbb{R}^8$, μ is a positive scalar

5. *Polyfit via linear regression.* **(3 pts)**

- Download weatherDewTmp.mat. Plot the data. It should look like the following



- We want to form a polynomial regression of this data. That is, given w = weeks and d = dew readings, we want to find $\theta_1, ..., \theta_p$ as the solution to

$$\underset{\theta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{m}(\theta_1 + \theta_2 w_i + \theta_3 w_i^2 + \cdots + \theta_p w_i^{p-1} - d_i)^2. \tag{1}$$

  Form X and y such that (1) is equivalent to the least squares problem

$$\underset{\theta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|X\theta - y\|_2^2. \tag{2}$$

  That is, for w the vector containing the week number, and y containing the dew data, form

$$X = \begin{bmatrix} 1 & w_1 & w_1^2 & w_1^3 & \cdots & w_1^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 1 & w_m & w_m^2 & w_m^3 & \cdots & w_m^{p-1} \end{bmatrix}.$$

(a) *Linear regression.* **(1pt)**

  i. Write down the normal equations for problem (2).

  ii. Fill in the code to solve the normal equations for θ, and use it to build a predictor. To verify your code is running correctly, the number after `check number` should be 1.759 (implemented correctly) or 1.341 (also accepted).

  iii. Implement a polynomial fit of orders $p = 1, 2, 3, 10, 100$, for the weather data provided. Include a figure that plots the original signal, overlaid with each polynomial fit. Comment on the "goodness of fit" for each value of p.

(b) *Ridge regression.* **(0.5pt)** Oftentimes, it is helpful to add a *regularization term* to (2), to improve stability. In other words, we solve

$$\underset{\theta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|X\theta - y\|_2^2 + \frac{\rho}{2}\|\theta\|_2^2. \tag{3}$$

  for some $\rho > 0$.

  i. Again, write down the normal equations for (3). Your equation should be of form $A\theta = b$ for some matrix A and vector b that you specify.

  ii. Write the code for solving the ridge regression problem and run it. To verify your code is running correctly, the number after `check number` should be *Checknumber* : 1.636 (implemented correctly) or 1.206 (also accepted).

  iii. Using $\rho = 1.0$, plot the weather data with overlaying polynomial fits with ridge regression. Provide these plots for $p = 1, 2, 3, 10, 100$. Comment on the "goodness of fit" and the stability of the fit, and also compare with the plots generated without using the extra penalty term.

(c) *Conditioning.* **(1pt)**

i. An *unconstrained quadratic problem* is any problem that can be written as

$$\underset{\theta}{\text{minimize}} \quad \frac{1}{2}\theta^T Q\theta + c^T\theta + r \tag{4}$$

for some symmetric positive semidefinite matrix Q, and some vector c and some scalar r. Show that the ridge regression problem (3) is an unconstrained quadratic problem by writing down Q, c, and r in terms of X and y such that (4) is equivalent to (3). Show that the Q you picked is positive semidefinite.

ii. In your code, write a function that takes in X and y, constructs Q as specified in the previous problem, and returns the condition number of Q. Report the condition number $\kappa(Q)$ for varying values of p and ρ, by filling in the following table. Here, $m = 742$ is the total number of data samples. Report at least 2 significant digits. Comment on how much ridge regression is needed to affect conditioning.

p	$\rho = 0$	$\rho = m$	$\rho = 10m$	$\rho = 100m$
1				
2				
5				
10				

iii. Under the *same experimental parameters* as the previous question, run ridge regression for each choice of p and ρ, and fill in the table with the mean squared error of the fit:

$$\textbf{mean squared error} = \frac{1}{m}\sum_{i=1}^{m}(x_i^T\theta - y[i])^2$$

where x_i is the ith row of X. Comment on the tradeoff between using larger ρ to improve conditioning vs its affect on the final performance.

p	$\rho = 0$	$\rho = m$	$\rho = 10m$	$\rho = 100m$
1				
2				
5				
10				

(d) *Forcasting.* **(0.5pt)** Picking your favorite set of hyperparameters (p, ρ), forecast the next week's dew point temperature. Plot the forecasted data over the current observations. Do you believe your forecast? Why?

6. **PAC learning. (2pts)** Consider the following hypothesis class in $\mathbb{R}^2$:

$$\mathcal{H} = \left\{ h_a : [-2,2]^2 \to \mathbb{R} : h_a(x) = \begin{cases} 1 & \text{if } |x[1] - x[2]| \leq a \\ 0 & \text{else.} \end{cases}, \quad 0 \leq a \leq 1. \right\}$$

The notation $h_a : [-2,2]^2 \to \mathbb{R}$ means that the inputs x are restricted in the two-dimensional domain

$$\begin{bmatrix} -2 \\ -2 \end{bmatrix} \leq x \leq \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

We now consider a scenario where the true function $y = f(x)$ is *realizable*, e.g. $f \in \mathcal{H}$. We draw samples $\mathcal{S} = \{(x_1, y_1), ..., (x_m, y_m)\}$ where $y_i = f(x_i)$ and compute an ERM

$$h_{\mathcal{S}} = \underset{h \in \mathcal{H}_{\text{in}}}{\text{argmin}}\, \mathcal{L}_{\mathcal{S}}(h)$$

where $\mathcal{L}_{\mathcal{S}}$ is the empirical risk.

(a) **(0.25 pts)** Draw a picture of one possible hypothesis in $\mathcal{H}$. That is, draw the 2-D region where the area for x where $h_a(x) = 1$ is shaded, and $h_a(x) = 0$ is not shaded, for some plausible a.

(b) **(0.25 pts)** Propose a training sampling strategy (e.g. a distribution $\mathcal{D}$ where we draw $x_i \sim \mathcal{D}$) that guarantees PAC learning.

.

(c) **(0.25 pts)** On the image above, indicate the region where no samples in $\mathcal{S}$ exist in order for the ERM estimate of $\hat{a}$ to be wrong by β, e.g. $|a - \hat{a}| = \beta$. Calculate the area of that region. (Your answer will be in terms of a.)

(d) **(0.25 pts)** Next, suppose that $\hat{a} = a + \beta$. What is $\mathcal{L}_{\mathcal{D}}(h_{\hat{a}})$? (Your answer will be in terms of a.)

(e) **(0.25 pts)** Next, suppose that $\hat{a} = a - \beta$. What is $\mathcal{L}_{\mathcal{D}}(h_{\hat{a}})$? (Your answer will be in terms of a.)

(f) **(0.75 pts)** Put the pieces together to prove that $\mathcal{H}$ is PAC-learnable by computing the number of samples m needed such that

$$\mathbf{Pr}(\mathcal{L}_{\mathcal{D}}(h_{\mathcal{S}}) \geq \epsilon) \leq \delta$$

for general $0 \leq (\delta, \epsilon) \leq 1$. At this point your answer should *not* depend on a, so you need to find the most extreme value of a such that your bound holds tight.
Hint: use $(1-x)^m \leq \exp(-xm)$ and $x - \sqrt{4x+9} + 3 \geq x/3$ for $x \geq 0$.