EECS 127/227AT Optimization Models in Engineering UC Berkeley Spring 2024

Homework 6

Self grades are due at 11 PM on March 1, 2024.

1. Condition Number

In lecture, we examined the sensitivity of solutions to linear system $A\vec{x} = \vec{y}$ (for nonsingular/invertible square matrix A) to perturbations in our measurements $\vec{y}$. Specifically, we showed that if we model measurement noise $\Delta\vec{y}$ as a linear perturbation on $\vec{y}$, resulting in a linear perturbation $\Delta\vec{x}$ on $\vec{x}$ — i.e., $A(\vec{x}+\Delta\vec{x}) = \vec{y}+\Delta\vec{y}$ — we can bound the magnitude of the solution perturbations $\Delta\vec{x}$ as

$$\frac{\|\Delta\vec{x}\|_2}{\|\vec{x}\|_2} \leq \kappa(A)\frac{\|\Delta\vec{y}\|_2}{\|\vec{y}\|_2}, \tag{1}$$

where $\kappa(A) = \frac{\sigma_{\max}\{A\}}{\sigma_{\min}\{A\}} = \|A\|_2 \left\|A^{-1}\right\|_2$ is the condition number of A, or the ratio of A's maximum and minimum singular values. In this problem, we will establish a similar bound for perturbations on A.

(a) Consider the linear system $A\vec{x} = \vec{y}$ above, where $A \in \mathbb{R}^{n\times n}$ is invertible (i.e., square and nonsingular). Let $\Delta A \in \mathbb{R}^{n\times n}$ denote a linear perturbation on matrix A generating a corresponding linear perturbation $\Delta\vec{x}$ in solution $\vec{x}$, i.e.,

$$(A+\Delta A)(\vec{x}+\Delta\vec{x}) = \vec{y}. \tag{2}$$

Show that

$$\frac{\|\Delta\vec{x}\|_2}{\|\vec{x}+\Delta\vec{x}\|_2} \leq \kappa(A)\frac{\|\Delta A\|_2}{\|A\|_2}. \tag{3}$$

Solution: Rearranging the given linear system equation, we have

$$\begin{aligned}
(A+\Delta A)(\vec{x}+\Delta\vec{x}) = \vec{y} &\implies A\vec{x} + A\Delta\vec{x} + \Delta A\vec{x} + \Delta A\Delta\vec{x} = \vec{y} && (4)\\
&\implies A\Delta\vec{x} + \Delta A\vec{x} + \Delta A\Delta\vec{x} = \vec{0} \qquad (A\vec{x} = \vec{y}) && (5)\\
&\implies A\Delta\vec{x} = -\Delta A(\vec{x}+\Delta\vec{x}) && (6)\\
&\implies \Delta\vec{x} = -A^{-1}\Delta A(\vec{x}+\Delta\vec{x}) && (7)\\
&\implies \|\Delta\vec{x}\|_2 = \left\|A^{-1}\Delta A(\vec{x}+\Delta\vec{x})\right\|_2 \leq \left\|A^{-1}\right\|_2 \|\Delta A\|_2 \|\vec{x}+\Delta\vec{x}\|_2 && (8)\\
&\implies \|\Delta\vec{x}\|_2 \leq \left\|A^{-1}\right\|_2 \|\Delta A\|_2 \|\vec{x}+\Delta\vec{x}\|_2 \frac{\|A\|_2}{\|A\|_2} && (9)\\
&\implies \frac{\|\Delta\vec{x}\|_2}{\|\vec{x}+\Delta\vec{x}\|_2} \leq \|A\|_2 \left\|A^{-1}\right\|_2 \frac{\|\Delta A\|_2}{\|A\|_2} && (10)\\
&\implies \frac{\|\Delta\vec{x}\|_2}{\|\vec{x}+\Delta\vec{x}\|_2} \leq \sigma_{\max}\{A\}\frac{1}{\sigma_{\min}\{A\}}\frac{\|\Delta A\|_2}{\|A\|_2} && (11)\\
&\implies \frac{\|\Delta\vec{x}\|_2}{\|\vec{x}+\Delta\vec{x}\|_2} \leq \kappa(A)\frac{\|\Delta A\|_2}{\|A\|_2} && (12)
\end{aligned}$$

as desired.

(b) Note that Equations (1) and (3) above bound two slightly different quantities: $\frac{\|\Delta\vec{x}\|_2}{\|\vec{x}\|_2}$ and $\frac{\|\Delta\vec{x}\|_2}{\|\vec{x}+\Delta\vec{x}\|_2}$, respectively. In general, we wish to establish these bounds because we want to characterize the size of $\Delta\vec{x}$

under different sizes of perturbation. Which of these two bounds better serves this purpose? Consider the following two cases. (i) $\Delta\vec{x} \gg \vec{x}$ (ii) $\vec{x} \gg \Delta\vec{x}$ and answer whether Equation (1) or Equation (3) is better.

Solution: When $\Delta\vec{x}$ is small relative to $\vec{x}$, both bounds are almost equivalent, since $\frac{\|\Delta\vec{x}\|_2}{\|\vec{x}+\Delta\vec{x}\|_2} \sim \frac{\|\Delta\vec{x}\|_2}{\|\vec{x}\|_2}$ for small $\Delta\vec{x}$. However, when $\Delta\vec{x}$ is very large relative to $\vec{x}$, $\frac{\|\Delta\vec{x}\|_2}{\|\vec{x}+\Delta\vec{x}\|_2} \sim \frac{\|\Delta\vec{x}\|_2}{\|\Delta\vec{x}\|_2} = 1$ regardless of the value of $\Delta\vec{x}$, so our bound in Equation (3) tells us nothing about $\Delta\vec{x}$'s size. Our bound on solution error for perturbations in $\vec{y}$ in Equation (1) is therefore much more useful for characterizing $\Delta\vec{x}$ over a wider range of perturbations than our bound on solution error for perturbations in A in Equation (3).

2. Ridge Regression for Bounded Output Perturbation

We will first solve the ridge regression problem in the case where our output measurements $\vec{y}$ are perturbed and we have some bounds on this perturbation, as well as some specific knowledge about data matrix A.

Let square matrix $A \in \mathbb{R}^{n\times n}$ have the singular value decomposition $A = U\Sigma V^\top$, and let its smallest singular value be $\sigma_{\min}\{A\} > 0$.

(a) Is A invertible? If so, write the singular value decomposition of A^{-1}.

Solution: Since A is a square matrix and all of its singular values are positive, it is invertible, and the SVD of this inverse is

$$A^{-1} = V\Sigma^{-1}U^\top. \tag{13}$$

(b) Consider the linear equation $A\vec{x} = \vec{y}_p$, where $\vec{y}_p \in \mathbb{R}^n$ is a perturbed measurement satisfying

$$\|\vec{y}_p - \vec{y}\|_2 \leq r \tag{14}$$

for some vector $\vec{y} \in \mathbb{R}^n$ and $r > 0$. Let $\vec{x}^\star(\vec{y})$ denote the solution of $A\vec{x} = \vec{y}$.

Show that

$$\max_{\vec{y}_p:\|\vec{y}_p-\vec{y}\|_2\leq r} \|\vec{x}^\star(\vec{y}_p) - \vec{x}^\star(\vec{y})\|_2 = \frac{r}{\sigma_{\min}\{A\}} \tag{15}$$

Solution: Since A is invertible, for any $\vec{y} \in \mathbb{R}^n$, we have

$$\vec{x}^\star(\vec{y}) = A^{-1}\vec{y} = V\Sigma^{-1}U^\top\vec{y}. \tag{16}$$

Note that

$$\{\vec{y}_p \in \mathbb{R}^n : \|\vec{y}_p - \vec{y}\|_2 \leq r\} = \{\vec{y} + \vec{u} : \vec{u} \in \mathbb{R}^n, \|\vec{u}\|_2 \leq r\}, \tag{17}$$

and therefore,

$$\max_{\vec{y}_p:\|\vec{y}_p-\vec{y}\|_2\leq r} \|\vec{x}^\star(\vec{y}_p) - \vec{x}^\star(\vec{y})\|_= \max_{\vec{u}:\|\vec{u}\|_2\leq r} \|\vec{x}^\star(\vec{y}+\vec{u}) - \vec{x}^\star(\vec{y})\|_2 . \tag{18}$$

Since we can write the differences between the estimates as

$$\vec{x}^\star(\vec{y}+\vec{u}) - \vec{x}^\star(\vec{y}) = A^{-1}(\vec{y}+\vec{u}) - A^{-1}\vec{y} = A^{-1}\vec{u} = V\Sigma^{-1}U^\top\vec{u}, \tag{19}$$

we obtain

$$\max_{\vec{u}:\|\vec{u}\|_2\leq r} \|\vec{x}^\star(\vec{y}+\vec{u}) - \vec{x}^\star(\vec{y})\|_2 = \max_{\vec{u}:\|\vec{u}\|_2\leq r} \left\|V\Sigma^{-1}U^\top\vec{u}\right\|_2 \tag{20}$$

$$= \max_{\vec{u}:\|\vec{u}\|_2\leq r} \left\|\Sigma^{-1}\vec{u}\right\|_2 \tag{21}$$

where the last equality follows from the fact that U and V are orthonormal matrices. The matrix Σ^{-1} is a diagonal matrix of entries are the inverse of those in Σ, and thus

$$\max_{\vec{u}:\|\vec{u}\|_2\leq r} \left\|\Sigma^{-1}\vec{u}\right\|_2 = r\sigma_{\max}\{\Sigma^{-1}\} = \frac{r}{\sigma_{\min}\{\Sigma\}} = \frac{r}{\sigma_{\min}\{A\}} \tag{22}$$

as desired.

(c) What happens if the smallest singular value of A is very close to zero? Why is this problematic for finding our solution vector $\vec{x}^\star$?

Solution: In part (b), we showed that a perturbation of magnitude r on the measurement can change our estimate $\vec{x}^\star$ by up to $\frac{r}{\sigma_{\min}\{A\}}$. If $\sigma_{\min}\{A\}$ is very small, the estimate can change by a large amount even if the measurements are only slightly perturbed. We say in this instance that our solution is very "sensitive" to perturbations in $\vec{y}$.

(d) Now assume that we find optimal value $\vec{x}^\star$ via ridge regression, i.e., we compute

$$\vec{x}^\star_\lambda(\vec{y}_p) = \operatorname*{argmin}_{\vec{x}} \left\{ \|A\vec{x} - \vec{y}_p\|_2^2 + \lambda \|\vec{x}\|_2^2 \right\} \tag{23}$$

for some chosen value $\lambda \geq 0$. Compute $\vec{x}^\star_\lambda(\vec{y}_p)$, our optimal solution vector (now parameterized by λ), by solving this optimization problem. You may use the solution from class for this part.

Solution: By setting the gradient of $\langle A\vec{x} - \vec{y}_p,\ A\vec{x} - \vec{y}_p\rangle + \lambda \langle \vec{x},\ \vec{x}\rangle$ to zero, we obtain:

$$2A^\top(A\vec{x} - \vec{y}_p) + 2\lambda\vec{x} = 2\left[(A^\top A + \lambda I)\vec{x} - A^\top \vec{y}_p\right] = 0, \tag{24}$$

and since this is the point at which $\vec{x} = \vec{x}^\star_\lambda(\vec{y}_p)$, we have

$$\vec{x}^\star_\lambda(\vec{y}_p) = (A^\top A + \lambda I)^{-1} A^\top \vec{y}_p. \tag{25}$$

Alternatively, we can note that

$$\|A\vec{x} - \vec{y}_p\|_2^2 + \lambda \|\vec{x}\|_2^2 = \left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} \vec{x} - \begin{bmatrix} \vec{y}_p \\ \vec{0} \end{bmatrix} \right\|_2^2, \tag{26}$$

then using the traditional least squares solution, we have

$$\vec{x}^\star_\lambda(\vec{y}_p) = \left(\begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}^\top \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} \right)^{-1} \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}^\top \begin{bmatrix} \vec{y}_p \\ \vec{0} \end{bmatrix} = (A^\top A + \lambda I)^{-1} A^\top \vec{y}_p. \tag{27}$$

(e) Show that for all $\lambda > 0$,

$$\max_{\vec{y}_p : \|\vec{y}_p - \vec{y}\|_2 \leq r} \|\vec{x}^\star_\lambda(\vec{y}_p) - \vec{x}^\star_\lambda(\vec{y})\|_2 \leq \frac{r}{2\sqrt{\lambda}}. \tag{28}$$

How does the value of λ affect the sensitivity of your solution $\vec{x}^\star_\lambda(\vec{y})$ to the perturbation level in $\vec{y}$? *HINT: For every $\lambda > 0$, we have*

$$\max_{\sigma > 0} \frac{\sigma}{\sigma^2 + \lambda} = \frac{1}{2\sqrt{\lambda}}. \tag{29}$$

You need not show this; this optimization can be solved by setting the derivative of the objective function to 0 and solving for σ.

Solution:

Using our solution from part (d), we have that

$$\vec{x}^\star_\lambda(\vec{y}_p) - \vec{x}^\star_\lambda(\vec{y}) = (A^\top A + \lambda I)^{-1} A^\top (\vec{y}_p - \vec{y}) \tag{30}$$

and thus

$$\|\vec{x}^\star_\lambda(\vec{y}_p) - \vec{x}^\star_\lambda(\vec{y})\|_2 = \left\|(A^\top A + \lambda I)^{-1} A^\top (\vec{y}_p - \vec{y})\right\|_2 \tag{31}$$

$$\leq \left\|(A^\top A+\lambda I)^{-1}A^\top\right\|_2 \|\vec{y}_p-\vec{y}\|_2 \tag{32}$$

$$= \sigma_{\max}\left\{(A^\top A+\lambda I)^{-1}A^\top\right\}\|\vec{y}_p-\vec{y}\|_2 . \tag{33}$$

This value is maximized when $\|\vec{y}_p-\vec{y}\|_2 = r$, so we can write new bound

$$\max_{\vec{y}_p:\|\vec{y}_p-\vec{y}\|_2\leq r} \|\vec{x}^\star_\lambda(\vec{y}_p)-\vec{x}^\star_\lambda(\vec{y})\|_2 \leq \sigma_{\max}\left\{(A^\top A+\lambda I)^{-1}A^\top\right\} r. \tag{34}$$

To simplify this bound further, we need to compute the largest singular value of $(A^\top A+\lambda I)^{-1}A^\top$. Plugging in our decomposition $A = U\Sigma V^\top$, we first compute

$$A^\top A+\lambda I = V\Sigma U^\top U\Sigma V^\top+\lambda I = V\Sigma^2V^\top+\lambda I = V\Sigma^2V^\top+\lambda VV^\top = V(\Sigma^2+\lambda I)V^\top, \tag{35}$$

which leads to

$$(A^\top A+\lambda I)^{-1}A^\top = V(\Sigma^2+\lambda I)^{-1}V^\top V\Sigma U^\top = V(\Sigma^2+\lambda I)^{-1}\Sigma U^\top. \tag{36}$$

This is the SVD decomposition of $(A^\top A+\lambda I)^{-1}A^\top$, and its singular values are the diagonal elements of $(\Sigma^2+\lambda I)^{-1}\Sigma$. Given the singular values of A, $\{\sigma_i\}_{i=1}^n$, the singular values of $(\Sigma^2+\lambda I)^{-1}\Sigma$ are

$$\left\{\frac{\sigma_i}{\sigma_i^2+\lambda}\right\}_{i=1}^n, \tag{37}$$

all of which we know are smaller than $\frac{1}{2\sqrt{\lambda}}$ from the given hint. We can then rewrite our bound above as

$$\max_{\vec{y}_p:\|\vec{y}_p-\vec{y}\|_2\leq r} \|\vec{x}^\star_\lambda(\vec{y}_p)-\vec{x}^\star_\lambda(\vec{y})\|_2 \leq \sigma_{\max}\left\{(A^\top A+\lambda I)^{-1}A^\top\right\} r \leq \frac{r}{2\sqrt{\lambda}} \tag{38}$$

as desired.

The larger we choose our λ, the tighter our bound on the deviation of our perturbed solution $\vec{x}^\star_\lambda(\vec{y}_p)$ from our true solution $\vec{x}^\star_\lambda(\vec{y})$. In other words, if the regularization parameter λ is large enough, **small perturbations in the measurement cannot change the estimate by a large amount**.

3. Linear Regression with Weights

In this problem, we discuss multiple interpretations of weighted linear regression.

Let $A \in \mathbb{R}^{m \times n}$ be a data matrix whose data points are the m rows $\vec{a}_1^\top, \ldots, \vec{a}_m^\top \in \mathbb{R}^n$. Suppose $m \geq n$ and A has full column rank. Let $\vec{y} \in \mathbb{R}^m$ be a vector of outputs, each corresponding to a data point. Let $\vec{w} \in \mathbb{R}^m_{++}$ be a vector of *positive* real numbers, also called weights, each corresponding to a data point—output pair. We are interested in the following least-squares type optimization problem:

$$\min_{\vec{x} \in \mathbb{R}^n} \sum_{i=1}^{m} w_i (\vec{a}_i^\top \vec{x} - y_i)^2. \tag{39}$$

In general, assigning a high weight w_i means that we want our learned linear predictor $\vec{a}_i^\top \vec{x}$ to achieve a close value to y_i; that is, we believe this data point is significant or important to get right.

(a) Show that the problem in Equation (39) is equivalent to the problem:

$$\min_{\vec{x} \in \mathbb{R}^n} \left\| W^{1/2} (A\vec{x} - \vec{y}) \right\|_2^2 \tag{40}$$

where $W \doteq \operatorname{diag}(\vec{w}) \in \mathbb{R}^{m \times m}$ is a diagonal matrix whose diagonal entries are the entries of $\vec{w}$.

Solution: We have

$$\sum_{i=1}^{m} w_i (\vec{a}_i^\top \vec{x} - y_i)^2 = \left\| \begin{bmatrix} w_1^{1/2} (\vec{a}_1^\top \vec{x} - y_1) \\ \vdots \\ w_m^{1/2} (\vec{a}_1^\top \vec{x} - y_m) \end{bmatrix} \right\|_2^2 \tag{41}$$

$$= \left\| \begin{bmatrix} w_1^{1/2} & & \\ & \ddots & \\ & & w_n^{1/2} \end{bmatrix} \begin{bmatrix} \vec{a}_1^\top \vec{x} - y_1 \\ \vdots \\ \vec{a}_1^\top \vec{x} - y_m \end{bmatrix} \right\|_2^2 \tag{42}$$

$$= \left\| W^{1/2} \begin{bmatrix} \vec{a}_1^\top \vec{x} - y_1 \\ \vdots \\ \vec{a}_1^\top \vec{x} - y_m \end{bmatrix} \right\|_2^2 \tag{43}$$

$$= \left\| W^{1/2} \left(\begin{bmatrix} \vec{a}_1^\top \vec{x} \\ \vdots \\ \vec{a}_m^\top \vec{x} \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \right) \right\|_2^2 \tag{44}$$

$$= \left\| W^{1/2} \left(\begin{bmatrix} \vec{a}_1^\top \vec{x} \\ \vdots \\ \vec{a}_m^\top \vec{x} \end{bmatrix} - \vec{y} \right) \right\|_2^2 \tag{45}$$

$$= \left\| W^{1/2} \left(\begin{bmatrix} \vec{a}_1^\top \\ \vdots \\ \vec{a}_m^\top \end{bmatrix} \vec{x} - \vec{y} \right) \right\|_2^2 \tag{46}$$

$$= \left\| W^{1/2} \left(A\vec{x} - \vec{y} \right) \right\|_2^2. \tag{47}$$

(b) Using Equation (40), compute the gradient (with respect to $\vec{x}$) of the objective function

$$f(\vec{x}) \doteq \left\|W^{1/2}(A\vec{x}-\vec{y})\right\|_2^2. \tag{48}$$

Solution: We have

$$\begin{aligned}
f(\vec{x}) &= \left\|W^{1/2}(A\vec{x}-\vec{y})\right\|_2^2 &(49)\\
&= (A\vec{x}-\vec{y})^\top W(A\vec{x}-\vec{y}) &(50)\\
&= (A\vec{x}-\vec{y})^\top (WA\vec{x}-W\vec{y}) &(51)\\
&= \vec{x}^\top A^\top WA\vec{x} - \vec{y}^\top WA\vec{x} - \vec{x}^\top A^\top W\vec{y} + \vec{y}^\top W\vec{y} &(52)\\
&= \vec{x}^\top A^\top WA\vec{x} - 2\vec{x}^\top A^\top W\vec{y} + \vec{y}^\top W\vec{y}. &(53)\\
\implies \nabla_{\vec{x}} f(\vec{x}) &= \nabla_{\vec{x}}\left\{\vec{x}^\top A^\top WA\vec{x} - 2\vec{x}^\top A^\top W\vec{y} + \vec{y}^\top W\vec{y}\right\} &(54)\\
&= 2A^\top WA\vec{x} - 2A^\top W\vec{y}. &(55)
\end{aligned}$$

(c) Show that the optimal solution to Equation (40) is given by

$$\vec{x}^\star_{\text{WLR}} \doteq (A^\top WA)^{-1}A^\top W\vec{y}. \tag{56}$$

HINT: There are multiple ways to do this problem; one uses the gradient that you just computed, and another finds the least squares solution of a particular linear system.

HINT: If using the gradient method, you may assume that f is minimized at any $\vec{x}^\star$ such that $\nabla_{\vec{x}} f(\vec{x}^\star) = \vec{0}$; this is because f is convex, as we will see a little later in the course.

Solution:

Approach 1: Using the gradient.

Since $W_{ii} = w_i > 0$, W is positive definite. Thus, the Hessian of the objective function, i.e.,

$$\nabla^2_{\vec{x}} f(\vec{x}) = 2A^\top WA \tag{57}$$

is everywhere positive definite (one can show this formally by the matrix's Rayleigh coefficient). Thus f is convex and to find the optimal solution it suffices to set the gradient to $\vec{0}$. We have

$$\begin{aligned}
\vec{0} &\overset{\text{set}}{=} \nabla_{\vec{x}} f(\vec{x}^\star_{\text{WLR}}) &(58)\\
&= 2A^\top WA\vec{x}^\star_{\text{WLR}} - 2A^\top W\vec{y} &(59)\\
\implies A^\top WA\vec{x}^\star_{\text{WLR}} &= A^\top W\vec{y} &(60)\\
\implies \vec{x}^\star_{\text{WLR}} &= (A^\top WA)^{-1}A^\top W\vec{y}. &(61)
\end{aligned}$$

Approach 2: Using linear regression formula.

We have

$$f(\vec{x}) = \left\|W^{1/2}(A\vec{x}-\vec{y})\right\|_2^2 = \left\|W^{1/2}A\vec{x} - W^{1/2}\vec{y}\right\|_2^2. \tag{62}$$

Since A has full rank and W is a diagonal matrix of positive entries, $W^{1/2}A$ has full rank. Thus one can solve this system using the least squares formula, obtaining

$$\vec{x}^\star_{\text{WLR}} = ((W^{1/2}A)^\top(W^{1/2}A))^{-1}(W^{1/2}A)^\top(W^{1/2}\vec{y}) = (A^\top WA)^{-1}A^\top W\vec{y}. \tag{63}$$

(d) Now we will look at this problem from a probabilistic interpretation. Suppose our output value $\vec{y}$ is noisy, and in particular there is some $\vec{x}_0$ such that for every i we have $y_i = \vec{a}_i^\top \vec{x}_0 + u_i$, where u_i is a random variable. Here we assume the u_i are independent but *not* identically distributed. In particular, we assume that for each i we have that u_i is distributed according to a Gaussian $\mathcal{N}(0, \sigma_i^2)$ where $\sigma_i > 0$ is a known noise parameter for each data point i.

To recover $\vec{x}_0$ given data A and $\vec{y}$, as well as the σ_i^2, we want to compute the maximum likelihood estimator (MLE). Show that the maximum likelihood problem

$$\operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} p(\vec{y} \mid A, \vec{x}) \tag{64}$$

is equivalent to the weighted linear regression problem:

$$\operatorname*{argmin}_{\vec{x}\in\mathbb{R}^n} \sum_{i=1}^{m} w_i(\vec{a}_i^\top \vec{x} - y_i)^2. \tag{65}$$

for some choice of $\vec{w}$. What choice of $\vec{w}$ makes them equivalent?

Note: Refer to Sec.4.5 of the course reader for a discussion on MLE.

Solution: We have

$$\begin{aligned}
\operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} p(\vec{y} \mid A, \vec{x}) &= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \prod_{i=1}^{m} p(y_i \mid \vec{a}_i, \vec{x}) &(66)\\
&= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \log\left(\prod_{i=1}^{m} p(y_i \mid \vec{a}_i, \vec{x})\right) &(67)\\
&= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \sum_{i=1}^{m} \log(p(y_i \mid \vec{a}_i, \vec{x})) &(68)\\
&= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \mathrm{e}^{-(y_i - \vec{a}_i^\top \vec{x})^2/(2\sigma_i^2)}\right) &(69)\\
&= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \sum_{i=1}^{m} \left\{\log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) + \log\left(\mathrm{e}^{-(y_i - \vec{a}_i^\top \vec{x})^2/(2\sigma_i^2)}\right)\right\} &(70)\\
&= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \left\{\sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) + \sum_{i=1}^{m} \log\left(\mathrm{e}^{-(y_i - \vec{a}_i^\top \vec{x})^2/(2\sigma_i^2)}\right)\right\} &(71)\\
&= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \sum_{i=1}^{m} \log\left(\mathrm{e}^{-(y_i - \vec{a}_i^\top \vec{x})^2/(2\sigma_i^2)}\right) &(72)\\
&= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \sum_{i=1}^{m} -\frac{(y_i - \vec{a}_i^\top \vec{x})^2}{2\sigma_i^2} &(73)\\
&= \operatorname*{argmin}_{\vec{x}\in\mathbb{R}^n} \sum_{i=1}^{m} \frac{(y_i - \vec{a}_i^\top \vec{x})^2}{2\sigma_i^2}. &(74)
\end{aligned}$$

This is the same format as the original weighted regression problem, with weights $\vec{w}$ given by $w_i = \frac{1}{2\sigma_i^2}$.

(e) In addition to assigning weights to data points to place higher importance on getting those points right, sometimes we want to make sure that our learned $\vec{x}$ is close to some value $\vec{z} \in \mathbb{R}^n$. This could be true, for example, if we had prior information that said $\vec{x}$ is close to $\vec{z}$.

In particular, we want to make sure that the quantity

$$\sum_{i=1}^{n} s_i(x_i - z_i)^2 \tag{75}$$

is small, where $\vec{s} \in \mathbb{R}^n_{++}$ is a vector of *positive* weights. We may add this term to the weighted least squares objective function, with a regularization parameter $\lambda \geq 0$, to create a modified objective function

$$g(\vec{x}) \doteq \left\|W^{1/2}(A\vec{x} - \vec{y})\right\|_2^2 + \lambda \left\|S^{1/2}(\vec{x} - \vec{z})\right\|_2^2 \tag{76}$$

where $S \doteq \operatorname{diag}(\vec{s}) \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose entries are the entries of $\vec{s}$. This formulation is called *Tikhonov regression*.

Compute the gradient (with respect to $\vec{x}$) of g, and show that the optimal solution is

$$\vec{x}^\star_{\mathrm{TR}} \doteq (A^\top WA + \lambda S)^{-1}(A^\top W\vec{y} + \lambda S\vec{z}). \tag{77}$$

HINT: You may assume that g is minimized at any $\vec{x}^\star$ such that $\nabla_{\vec{x}} g(\vec{x}^\star) = \vec{0}$; this is because g is convex, as we will see a little later in the course.

Solution: We know that

$$\begin{aligned} g(\vec{x}) &= f(\vec{x}) + \lambda \left\|S^{1/2}(\vec{x} - \vec{z})\right\|_2^2 && (78)\\ \implies \nabla_{\vec{x}} g(\vec{x}) &= \nabla_{\vec{x}} f(\vec{x}) + \lambda \nabla_{\vec{x}} \left\|S^{1/2}(\vec{x} - \vec{z})\right\|_2^2. && (79) \end{aligned}$$

Now we can compute this rightmost term to get

$$\begin{aligned} \nabla_{\vec{x}} \left\|S^{1/2}(\vec{x} - \vec{z})\right\|_2^2 &= \nabla_{\vec{x}} (\vec{x} - \vec{z})^\top S(\vec{x} - \vec{z}) && (80)\\ &= \nabla_{\vec{x}} (\vec{x} - \vec{z})^\top (S\vec{x} - S\vec{z}) && (81)\\ &= \nabla_{\vec{x}} (\vec{x}^\top S\vec{x} - \vec{z}^\top S\vec{x} - \vec{x}^\top S\vec{z} + \vec{z}^\top \vec{z}) && (82)\\ &= \nabla_{\vec{x}} (\vec{x}^\top S\vec{x} - 2\vec{x}^\top S\vec{z} + \vec{z}^\top \vec{z}) && (83)\\ &= 2S\vec{x} - 2S\vec{z}. && (84) \end{aligned}$$

Putting it all together, we have

$$\begin{aligned} \nabla_{\vec{x}} g(\vec{x}) &= \nabla_{\vec{x}} f(\vec{x}) + \lambda \nabla_{\vec{x}} \left\|S^{1/2}(\vec{x} - \vec{z})\right\|_2^2 && (85)\\ &= 2A^\top WA\vec{x} - 2A^\top W\vec{y} + 2\lambda S\vec{x} - 2\lambda S\vec{z}. && (86) \end{aligned}$$

Now to find the optimal solution, we set the gradient to $\vec{0}$ and get

$$\begin{aligned} \vec{0} &\overset{\text{set}}{=} \nabla_{\vec{x}} g(\vec{x}^\star_{\mathrm{TR}}) && (87)\\ &= 2A^\top WA\vec{x}^\star_{\mathrm{TR}} - 2A^\top W\vec{y} + 2\lambda S\vec{x}^\star_{\mathrm{TR}} - 2\lambda S\vec{z} && (88)\\ \implies (A^\top WA + \lambda S)\vec{x}^\star_{\mathrm{TR}} &= A^\top W\vec{y} + \lambda S\vec{z} && (89)\\ \implies \vec{x}^\star_{\mathrm{TR}} &= (A^\top WA + \lambda S)^{-1}(A^\top W\vec{y} + \lambda S\vec{z}). && (90) \end{aligned}$$

4. Quadratics and Least Squares

In this question, we will see that every least squares problem can be considered as minimization of a quadratic cost function; whereas not every quadratic minimization problem corresponds to a least-squares problem. To begin with, consider the quadratic function, $f : \mathbb{R}^2 \to \mathbb{R}$ given by:

$$f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top \vec{w} + c \tag{91}$$

where $A \in \mathbb{S}^2_+$ (set of symmetric positive semidefinite matrices in $\mathbb{R}^{2\times 2}$), $\vec{b} \in \mathbb{R}^2$ and $c \in \mathbb{R}$.

(a) Assume $c = 0$, and assume that setting $\nabla f(\vec{w}) = 0$ allows us to find the unique minimizer. Give a concrete example of a matrix $A \succ 0$ and a vector $\vec{b}$ such that the point $\vec{w}^\star = \begin{bmatrix}-1 & 1\end{bmatrix}^\top$ is the unique minimizer of the quadratic function $f(\vec{w})$.

Solution: First, let $A \succ 0$. Now, by taking the gradient of $f(\vec{w})$ and setting it to zero, we get:

$$\nabla f(\vec{w}^\star) = 2A\vec{w}^\star - 2b = 0. \tag{92}$$

Since A is positive definite, it is invertible and therefore, the above minimizer is unique. Concretely, let $A = I$. By setting the gradient to zero, we obtain

$$\nabla f(\vec{w}^\star) = (A + A^\top)\vec{w}^\star - 2\vec{b} = 0 \implies \vec{w}^\star = \vec{b}. \tag{93}$$

Then $\vec{w}^\star = [-1\ 1]^\top$ is the unique minimizer if $\vec{b} = [-1\ 1]^\top$ and $A = I$.

(b) Assume $c = 0$. Give a concrete example of a matrix $A \succeq 0$, and a vector $\vec{b}$ such that the quadratic function $f(\vec{w})$ has infinitely many minimizers and all of them lie on the line $w_1 + w_2 = 0$.

HINT: Take the gradient of the expression and set it to zero. What needs to be true for there to be infinitely many solutions to the equation?

Solution: Since $A \in \mathbb{R}^{2\times 2}$ is positive semidefinite, setting gradient to zero shows us that each minimizer $\vec{w}^\star$ satisfies

$$\nabla f(\vec{w}^\star) = (A + A^\top)\vec{w}^\star - 2\vec{b} = 2A\vec{w}^\star - 2\vec{b} = 0. \tag{94}$$

In order to have infinitely many solutions, the positive semidefinite matrix A cannot have full rank. Since $A \in \mathbb{S}^2_+$, this amounts to A having rank at most 1. In other words, there must exist a vector $\vec{v} \in \mathbb{R}^2$ such that $A = \vec{v}\vec{v}^\top$. By setting $A = \vec{v}\vec{v}^\top$, each minimizer $\vec{w}^\star$ should satisfy

$$A\vec{w}^\star - \vec{b} = \vec{v}\vec{v}^\top\vec{w}^\star - \vec{b} = 0. \tag{95}$$

Note that each point on the line

$$\mathcal{L} = \{\vec{w} \in \mathbb{R}^2 : w_1 + w_2 = 0\} = \left\{\alpha \begin{bmatrix}1\\-1\end{bmatrix} : \alpha \in \mathbb{R}\right\} \tag{96}$$

is a minimizer of f. Along with (95), this implies that

$$\vec{v}\vec{v}^\top\left(\alpha\begin{bmatrix}1\\-1\end{bmatrix}\right) - \vec{b} = 0 \quad \forall \alpha \in \mathbb{R}. \tag{97}$$

This is satisfied only when $\vec{b} = 0$ and $\vec{v} \perp [-1\ 1]^\top$. Choosing $\vec{v} = [1\ 1]^\top$, we have

$$A = \beta\begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix}, \quad \vec{b} = 0 \quad \text{for some } \beta > 0. \tag{98}$$

(c) Assume $c = 0$. Let $\vec{w} = \begin{bmatrix}1 & 0\end{bmatrix}^\top$. Give a concrete example of a **non-zero** matrix $A \succeq 0$ and a vector $\vec{b}$ such that the quadratic function $f(\alpha\vec{w})$ tends to $-\infty$ as $\alpha \to \infty$. *HINT: Use the eigenvalue decomposition to write $A = \sigma_1\vec{u}_1\vec{u}_1^\top + \sigma_2\vec{u}_2\vec{u}_2^\top$ and express $\vec{w}$ in the basis formed by $\vec{u}_1, \vec{u}_2$.*

Solution: Let $\vec{w} = \begin{bmatrix}1 & 0\end{bmatrix}^\top$. We first expand $f(\alpha\vec{w})$ as follows:

$$f(\alpha\vec{w}) = (\vec{w}^\top A\vec{w})\alpha^2 - 2b_1\alpha + c. \tag{99}$$

Note that since A is PSD, $\vec{w}^\top A\vec{w} \geq 0$. Therefore, for $f(\alpha\vec{w})$ to tend to $-\infty$ as $\alpha \to \infty$, we must have $\vec{w}^\top A\vec{w} = 0$ and $b_1 > 0$.

Using the spectral theorem since A is symmetric positive semidefinite it can be written as $A = \sigma_1\vec{u}_1\vec{u}_1^\top + \sigma_2\vec{u}_2\vec{u}_2^\top$ with $\sigma_1 \geq \sigma_2 \geq 0$ and $\vec{u}_1, \vec{u}_2$ orthonormal vectors that form a basis for $\mathbb{R}^2$. Further $\sigma_1 > 0$ since A is not the zero matrix.

Thus we can write $\vec{w} = \beta_1\vec{u}_1 + \beta_2\vec{u}_2$. Substituting this we have, $\vec{w}^\top A\vec{w} = \sigma_1\beta_1^2 + \sigma_2\beta_2^2$. For this to be zero, we must have $\beta_1 = 0$ since $\sigma_1 > 0$.

This implies $\vec{w} = \beta_2\vec{u}_2$ and we must have $\vec{u}_2 = \pm\vec{w} = \pm[1, 0]^\top$ and $\beta_2 = \pm 1$ since both $\vec{w}$ and $\vec{u}$ are unit norm. Using the fact that $\beta_2^2 = 1$ we require $\sigma_2 = 0$ for $\vec{w}^\top A\vec{w}$ to be zero. Further since $\vec{u}_1$ and $\vec{u}_2$ are orthonormal we have $\vec{u}_1 = \pm[0, 1]^\top$.

Putting everything together we can construct one example of A and b as $A = (1)\begin{bmatrix}0 \\ 1\end{bmatrix}\begin{bmatrix}0 & 1\end{bmatrix} = \begin{bmatrix}0 & 0 \\ 0 & 1\end{bmatrix}$, and $\vec{b} = \begin{bmatrix}1 \\ 0\end{bmatrix}$.

(d) Say that we have the data set $\{(\vec{x}_i, y_i)\}_{i=1,\ldots,n}$ of data points $\vec{x}_i \in \mathbb{R}^d$ and values $y_i \in \mathbb{R}$. Define $X = \begin{bmatrix}\vec{x}_1 & \ldots & \vec{x}_n\end{bmatrix}^\top$ and $\vec{y} = \begin{bmatrix}y_1 & \ldots & y_n\end{bmatrix}^\top$. In terms of X and $\vec{y}$, find a matrix A, a vector $\vec{b} \in \mathbb{R}^d$ and a scalar c, so that we can express the sum of the square losses $\sum_{i=1}^{n}(\vec{w}^\top\vec{x}_i - y_i)^2$ as the quadratic function $f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top\vec{w} + c$.

Solution:

$$\sum_{i=1}^{n}(\vec{w}^\top\vec{x}_i - y_i)^2 = \sum_{i=1}^{n}\left(\vec{w}^\top\vec{x}_i(\vec{x}_i)^\top\vec{w} - 2\vec{w}^\top(y_i\vec{x}_i) + (y_i)^2\right) \tag{100}$$

Rearranging terms, we have

$$A = \sum_{i=1}^{n}\vec{x}_i\,(\vec{x}_i)^\top = X^\top X, \quad \vec{b} = \sum_{i=1}^{n}y_i\vec{x}_i = X^\top\vec{y}, \quad c = \sum_{i=1}^{n}(y_i)^2 = \vec{y}^\top\vec{y} = \|\vec{y}\|_2^2\,. \tag{101}$$

(e) Here are three statements with regards to the minimization of a quadratic loss function:

i. It can have a unique minimizer.

ii. It can have infinitely many minimizers.

iii. It can be unbounded from below, i.e. there is some direction, $\vec{w}$ so that $f(\alpha\vec{w})$ goes to $-\infty$ as $\alpha \to \infty$.

All three statements apply to general minimization of a quadratic cost function. Parts (a), (b) and (c) give concrete examples of quadratic cost functions where (i), (ii) and (iii) apply respectively. However, notice that statement (iii) cannot apply to the least squares problem as the objective is always positive. The least-squares problem can have infinitely many minimizers though. How? Consider the gradient of the least squares problem in part (d) at an optimal solution $\vec{w}^\star$:

$$\nabla f(\vec{w}^\star) = 2X^\top X\vec{w}^\star - 2\vec{b} = 0. \tag{102}$$

Therefore, the least squares problem only has multiple solutions if $X^\top X$ is not full rank. This means that $\operatorname{rank}\left(X^\top X\right) = \operatorname{rank}(X) < d$. Finally, the rank of X is less than d when the data points $\{\vec{x}_i\}_{i=1}^n$ do not span $\mathbb{R}^d$. This can happen when the number of data points n is less than d or when $\{\vec{f}_i\}_{i=1}^d$ are linearly dependent where $\vec{f}_i$ are the columns of X, i.e., the features.

We will see soon that these cases correspond to the *convexity* of the function: if the function is strictly convex, then it has a unique minimizer; and if it is just convex, then it can have multiple minimizers; and in both cases, it can have no minimizers. We will see soon how to prove that the quadratic objective functions we discuss in this problem are convex, strictly convex, or even non-convex.

Indicate below that you have read and understood the discussion above.

Solution: Any answer is fine.

5. **Neural Networks and Backpropagation**

Neural networks are parametric functions that have been widely used to fit complex patterns in vision and natural languages. Given some training data of the form $(\vec{x}_i, y_i)$, a neural network $\mathcal{N}$ is trained to minimize a *loss function* on the data. This is often done using *gradient descent*, an optimization method we will cover later in this class. Gradient descent requires us to compute the gradients of the loss function with respect to the parameters of the neural network. In practice, computational frameworks for neural networks compute the gradients automatically and efficiently via *back-propagation*, which uses the chain rule to recursively compute the gradients of the loss function. In this problem, we study a toy neural network trained on a single data point $(\vec{x}, y)$.

In particular, consider the following simplified three-layer neural network $\mathcal{N}$, representing a map from $\mathbb{R}^d$ to $\mathbb{R}$ whose parameters are $(\vec{w}_1, w_2, w_3) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$:

$$p_1 = \vec{w}_1^\top \vec{x} \tag{103}$$

$$h_1 = \sigma(p_1) \tag{104}$$

$$p_2 = w_2 h_1 \tag{105}$$

$$h_2 = \sigma(p_2) \tag{106}$$

$$z = w_3 h_2, \tag{107}$$

where $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a nonlinear function (also called the "activation function"), whose derivative is denoted by $\sigma'\colon \mathbb{R} \to \mathbb{R}$.

We want the output of the network $z = \mathcal{N}(\vec{x})$ to match the true label y. A natural choice of the loss function encouraging this behavior is the squared loss:

$$L(y, z) \doteq \frac{1}{2}(y - z)^2. \tag{108}$$

In the parts that follow, we will compute the derivative of L with respect to the parameters $\vec{w}_1, w_2, w_3$.

(a) Compute the following gradients and partial derivatives sequentially from left-to-right:

$$\frac{\partial L}{\partial z}, \quad \frac{\partial L}{\partial w_3}, \quad \frac{\partial L}{\partial h_2}, \quad \frac{\partial L}{\partial p_2}, \quad \frac{\partial L}{\partial w_2}, \quad \frac{\partial L}{\partial h_1}, \quad \frac{\partial L}{\partial p_1}, \quad \nabla_{\vec{w}_1} L, \quad \nabla_{\vec{x}} L. \tag{109}$$

Here $\nabla_{\vec{x}} L$ is the gradient whose entries are the derivatives of L with respect to the entries of $\vec{x}$, etc. We compute the first 4 derivatives for you.

$$\frac{\partial L}{\partial z} = z - y, \quad \frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial z} h_2, \quad \frac{\partial L}{\partial h_2} = \frac{\partial L}{\partial z} w_3, \quad \frac{\partial L}{\partial p_2} = \frac{\partial L}{\partial h_2} \sigma'(p_2) \tag{110}$$

Note how $\frac{\partial L}{\partial w_3}$ can be calculated using $\frac{\partial L}{\partial z}$ and $\frac{\partial L}{\partial p_2}$ can be calculated using $\frac{\partial L}{\partial h_2}$. In numerical computation, the result of $\frac{\partial L}{\partial z}$ and $\frac{\partial L}{\partial h_2}$ can thus be reused. This technique of saving computations for calculating the derivatives of a neural network is called *back-propagation*.

Use the chain rule to calculate the 5 remaining derivatives $\frac{\partial L}{\partial w_2}$, $\frac{\partial L}{\partial h_1}$, $\frac{\partial L}{\partial p_1}$,$\nabla_{\vec{w}_1} L$, and $\nabla_{\vec{x}} L$.

Solution: We compute each term via chain rule.

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial p_2} h_1 \tag{111}$$

$$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial p_2} w_2 \tag{112}$$

$$\frac{\partial L}{\partial p_1} = \frac{\partial L}{\partial h_1}\sigma'(p_1) \tag{113}$$

$$\nabla_{\vec{w}_1} L = \frac{\partial L}{\partial p_1}\vec{x} \tag{114}$$

$$\nabla_{\vec{x}} L = \frac{\partial L}{\partial p_1}\vec{w}_1 \tag{115}$$

(b) **(OPTIONAL)** Now suppose that Equation (107) is written as

$$z' = w_3 h_2 + h_1. \tag{116}$$

That is, we define a new neural network $\mathcal{N}'$ with parameters $(\vec{w}_1, w_2, w_3) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$ as follows.

$$p_1 = \vec{w}_1^\top \vec{x} \tag{117}$$

$$h_1 = \sigma(p_1) \tag{118}$$

$$p_2 = w_2 h_1 \tag{119}$$

$$h_2 = \sigma(p_2) \tag{120}$$

$$z' = w_3 h_2 + h_1. \tag{121}$$

This introduces a change in the network architecture, called the *skip connection*.

Again, compute the following gradients and partial derivatives with respect to the loss function $L(y, z') = \frac{1}{2}(y - z')^2$:

$$\frac{\partial L}{\partial z'}, \quad \frac{\partial L}{\partial w_3}, \quad \frac{\partial L}{\partial h_2}, \quad \frac{\partial L}{\partial p_2}, \quad \frac{\partial L}{\partial w_2}, \quad \frac{\partial L}{\partial h_1}, \quad \frac{\partial L}{\partial p_1}, \quad \nabla_{\vec{w}_1} L, \quad \nabla_{\vec{x}} L. \tag{122}$$

We compute the first 4 derivatives for you.

$$\frac{\partial L}{\partial z'} = z' - y, \quad \frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial z'} h_2, \quad \frac{\partial L}{\partial h_2} = \frac{\partial L}{\partial z'} w_3, \quad \frac{\partial L}{\partial p_2} = \frac{\partial L}{\partial h_2}\sigma'(p_2). \tag{123}$$

Use the chain rule to calculate the 5 remaining derivatives $\frac{\partial L}{\partial w_2}, \frac{\partial L}{\partial h_1}, \frac{\partial L}{\partial p_1}, \nabla_{\vec{w}_1} L$, and $\nabla_{\vec{x}} L$.

Solution:

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial p_2} h_1 \tag{124}$$

$$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial z'}\frac{\partial z'}{\partial h_1} = \frac{\partial L}{\partial z'}\left(1 + w_3 \frac{\partial h_2}{\partial h_1}\right) = \frac{\partial L}{\partial z'}\left(1 + w_3 \frac{\partial h_2}{\partial p_2}\frac{\partial p_2}{\partial h_1}\right) = \frac{\partial L}{\partial z'}(1 + w_3\sigma'(p_2)w_2) \tag{125}$$

$$\frac{\partial L}{\partial p_1} = \frac{\partial L}{\partial z'}\frac{\partial z'}{\partial p_1} = \frac{\partial L}{\partial z'}\left(w_3 \frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial p_1} + \frac{\partial h_1}{\partial p_1}\right) = \frac{\partial L}{\partial z'}(w_3\sigma'(p_2)w_2\sigma'(p_1) + \sigma'(p_1)) \tag{126}$$

$$\nabla_{\vec{w}_1} L = \frac{\partial L}{\partial p_1}\vec{x} \tag{127}$$

$$\nabla_{\vec{x}} L = \frac{\partial L}{\partial p_1}\vec{w}_1. \tag{128}$$

(c) **(OPTIONAL)** In optimizing a neural network using gradient descent, we need the gradient of the loss function with respect to the parameters of the network. Please express $\frac{\partial L}{\partial w_3}, \frac{\partial L}{\partial w_2}$, and $\nabla_{\vec{w}_1} L$ for $\mathcal{N}$ and $\mathcal{N}'$ respectively with no dependence on partial derivatives of other variables. We compute $\frac{\partial L}{\partial w_3}$ for you, as follows.

- For $\mathcal{N}$, we have $\dfrac{\partial L}{\partial w_3} = \dfrac{\partial L}{\partial z} h_2 = (z - y)h_2$.
- For $\mathcal{N}'$, we have $\dfrac{\partial L}{\partial w_3} = \dfrac{\partial L}{\partial z'} h_2 = (z' - y)h_2$.

Express the remaining derivatives $\dfrac{\partial L}{\partial w_2}$ and $\nabla_{\vec{w}_1} L$ within $\mathcal{N}$ and $\mathcal{N}'$.

Solution: For $\mathcal{N}$, we have

$$\begin{aligned}
\frac{\partial L}{\partial w_2} &= \frac{\partial L}{\partial p_2} h_1 && (129)\\
&= \frac{\partial L}{\partial h_2} \sigma'(p_2) h_1 && (130)\\
&= \frac{\partial L}{\partial z} w_3 \sigma'(p_2) h_1 && (131)\\
&= (z - y) w_3 \sigma'(p_2) h_1 && (132)\\
\nabla_{\vec{w}_1} L &= \frac{\partial L}{\partial p_1} \vec{x} && (133)\\
&= \frac{\partial L}{\partial h_1} \sigma'(p_1) \vec{x} && (134)\\
&= \frac{\partial L}{\partial p_2} w_2 \sigma'(p_1) \vec{x} && (135)\\
&= \frac{\partial L}{\partial h_2} \sigma'(p_2) w_2 \sigma'(p_1) \vec{x} && (136)\\
&= \frac{\partial L}{\partial z} w_3 \sigma'(p_2) w_2 \sigma'(p_1) \vec{x} && (137)\\
&= (z - y) w_3 \sigma'(p_2) w_2 \sigma'(p_1) \vec{x}. && (138)
\end{aligned}$$

For $\mathcal{N}'$, we have

$$\begin{aligned}
\frac{\partial L}{\partial w_2} &= \frac{\partial L}{\partial p_2} h_1 && (139)\\
&= \frac{\partial L}{\partial h_2} \sigma'(p_2) h_1 && (140)\\
&= \frac{\partial L}{\partial z'} w_3 \sigma'(p_2) h_1 && (141)\\
&= (z' - y) w_3 \sigma'(p_2) h_1 && (142)\\
\nabla_{\vec{w}_1} L &= \frac{\partial L}{\partial p_1} \vec{x} && (143)\\
&= \frac{\partial L}{\partial z'} (w_3 \sigma'(p_2) w_2 \sigma'(p_1) + \sigma'(p_1)) \vec{x} && (144)\\
&= (z' - y)(w_3 \sigma'(p_2) w_2 \sigma'(p_1) + \sigma'(p_1)) \vec{x} && (145)\\
&= (z' - y)(w_3 \sigma'(p_2) w_2 + 1) \sigma'(p_1) \vec{x}. && (146)
\end{aligned}$$

(d) **(OPTIONAL)** Many activation functions σ have the property that $\sigma' \leq 1$. For example, the sigmoid function $\sigma(p) = \frac{1}{1+e^{-p}}$ is sometimes used as an activation function. Its derivative, $\sigma'(p) = \frac{e^{-p}}{(1+e^{-p})^2}$ has the range $(0, 1/4]$. That is, $\sigma' < 1$. Consider the case when σ' is much smaller than 1, such that $\sigma'(p)\sigma'(q) \approx 0$ for any p, q, but $\sigma'(p) \not\approx 0$ for any p. Consider the derivatives $\dfrac{\partial L}{\partial w_3}$, $\dfrac{\partial L}{\partial w_2}$, and $\nabla_{\vec{w}_1} L$ within the neural network $\mathcal{N}$; with the above approximations, which of them will approximately be zero? Also answer this question for the neural network $\mathcal{N}'$.

NOTE: Some of the above gradients will indeed be approximately zero, and this is called the *vanishing gradient* problem in deep learning.

Solution: Yes, $\nabla_{\vec{w}_1} L$ for $\mathcal{N}$ involves a product of σ's and thus approximately vanishes. But all other derivatives don't go to zero for both $\mathcal{N}$ and $\mathcal{N}'$, including $\nabla_{\vec{w}_1} L$ for $\mathcal{N}'$.

Notably, $\nabla_{\vec{w}_1} L$ does not go to zero for $\mathcal{N}'$ because the network architecture of $\mathcal{N}'$ includes a skip connection that adds h_1 to the last layer z', adding a $+1$ in the gradient expression.

6. Homework Process

With whom did you work on this homework? List the names and SIDs of your group members.

NOTE: If you didn't work with anyone, you can put "none" as your answer.