EE566 ADAPTATION AND LEARNING **Instructor: Ali H. Sayed**
Homework Solutions #1 **Due: March 3, 2025**

---

1) (**Chapter 1**) Consider an arbitrary matrix $A \in \mathbb{R}^{N\times M}$ and let $A^\dagger$ denote its pseudo inverse, which is defined as the unique matrix satisfying the 4 properties (1.112a)–(1.112d). Show that
   (a) $\mathcal{R}(A^\dagger) = \mathcal{R}(A^\mathsf{T})$ and $\mathcal{N}(A^\dagger) = \mathcal{N}(A^\mathsf{T})$.
   (b) $\mathcal{N}(A) = \mathcal{R}(I - A^\dagger A)$.

Solution:

(a) Verifying $\mathcal{R}(A^\dagger) = \mathcal{R}(A^\mathsf{T})$ is equivalent to showing that $\mathcal{R}(A^\dagger) \perp \mathcal{N}(A)$. Thus, let $x \in \mathcal{N}(A)$ and $y \in \mathcal{R}(A^\dagger)$, i.e., $Ax = 0$ and $y = A^\dagger z$ for some z. It follows that

$$\begin{aligned} y^\mathsf{T} x &= z^\mathsf{T}(A^\dagger)^\mathsf{T} x \\ &= z^\mathsf{T}(A^\dagger A A^\dagger)^\mathsf{T} x \\ &= z^\mathsf{T}(A^\dagger)^\mathsf{T}(A^\dagger A)^\mathsf{T} x \\ &= z^\mathsf{T}(A^\dagger)^\mathsf{T} A^\dagger A x \\ &= 0 \end{aligned}$$

as desired. A similar argument can be used to establish the second result.

(b) Let $x \in \mathcal{R}(I - A^\dagger A)$, i.e., $x = (I - A^\dagger A)z$ for some z. Then,

$$Ax \;=\; A(I - A^\dagger A)z = (A - AA^\dagger A)z = (A - A)z = 0 \implies x \in \mathcal{N}(A)$$

Conversely, let $x \in \mathcal{N}(A)$. By property (a) we know that $x \in \mathcal{N}(A^{\dagger\mathsf{T}})$ so that $x^\mathsf{T} A^\dagger = 0$. Assume $x \notin \mathcal{R}(I - A^\dagger A)$. This means that there exists a vector $z \in \mathcal{N}(I - (A^\dagger A)^\mathsf{T})$ such that $x^\mathsf{T} z \neq 0$. In this case,

$$(I - (A^\dagger A)^\mathsf{T})z = (I - A^\dagger A)z = 0 \implies A^\dagger A z = z$$

and

$$x^\mathsf{T} z = x^\mathsf{T} A^\dagger A z = 0; \quad \text{a contradiction}$$

2) (**Chapter 2**) Consider an $M \times M$ square invertible real matrix X with entries X_{mn}. We know from row 15 in Table 2.1 that $\nabla_X \ln|\det(X)| = X^{-1}$. We further know from part (a) of Prob. 2.10 in the text that $\partial X^{-1}/\partial\alpha = -X^{-1}(\partial X/\partial\alpha)X^{-1}$, for any parameter α. Next, consider a matrix-valued function $G(X)$: $\mathbb{R}^{M\times M} \to \mathbb{R}^{M\times M}$. In a manner similar to (2.26), we use the notation $\nabla_{X^\mathsf{T}} G(X)$ to refer now to the $M^2 \times M^2$ matrix whose individual block entries are the $M \times M$ matrices given by $\partial G(X)/\partial X_{mn}$:

$$\nabla_{X^\mathsf{T}} G(X) \;\stackrel{\Delta}{=}\; \left[\frac{\partial G(X)}{\partial X_{mn}}\right]$$

Show that

$$\nabla_X^2 \ln|\det(X)| = X^{-1} \otimes X^{-1}$$

in terms of the Kronecker product operation.

Solution: We already know from row 15 in Table 2.1 that

$$\nabla_X \ln|\det(X)| = X^{-1}$$

The result is therefore the matrix function $G(X) = X^{-1}$. Next we need to differentiate $G(X)$ relative to X^T to arrive at the desired Hessian matrix for $\ln|\det(X)|$. For each individual entry X_{mn} of X

we know from part (a) of Prob. 2.10 in the text that

$$\begin{aligned}\frac{\partial G(X)}{\partial X_{mn}} &= \frac{\partial X^{-1}}{\partial X_{mn}} \\ &= -X^{-1} \times \frac{\partial X}{\partial X_{mn}} \times X^{-1} \\ &= -X^{-1} \times e_m e_n^{\mathsf{T}} \times X^{-1}\end{aligned}$$

using the basis vectors e_m and e_n with unit entries at locations m and n, respectively. Multiplying X^{-1} by $e_m e_n^{\mathsf{T}}$ from the left extracts the (n,m)th entry.

It follows that the (k,ℓ) entry of the desired gradient to the (m,n) entry of X is given by

$$\left[\frac{\partial X^{-1}}{\partial X_{mn}}\right]_{k\ell} = -[X^{-1}]_{km} \times [X^{-1}]_{n\ell}$$

If we now collect all the partial derivatives $\partial G(X)/\partial X_{mn}$ into a matrix we get

$$\nabla_{X^{\mathsf{T}}} G(X) = -X^{-1} \otimes X^{-1}$$

and consequently

$$\nabla_X^2 \ln|\det(X)| = X^{-1} \otimes X^{-1}$$

□

3) (**Chapter 3**) Consider a nonnegative real random variable $\boldsymbol{x}$ with cdf denoted by $F_{\boldsymbol{x}}(x)$. Show that the mean of $\boldsymbol{x}$ can be recovered from the cdf using the expression

$$\mathbb{E}\boldsymbol{x} = \int_0^{\infty} (1 - F_{\boldsymbol{x}}(t))dt = \int_0^{\infty} \mathbb{P}[\boldsymbol{x} \geq t]dt$$

This result establishes a connection between expectations of random variables and tails of their distributions. Conclude similarly that when $\boldsymbol{x}$ is nonnegative and assumes discrete integer values in $\mathbb{N}$, then

$$\mathbb{E}\boldsymbol{x} = \sum_{n=0}^{\infty} \mathbb{P}[\boldsymbol{x} \geq n]$$

How would you adjust the expressions if the random variables were not necessarily nonnegative?

Solution: Recall first that, by definition,

$$F_{\boldsymbol{x}}(t) = \mathbb{P}[\boldsymbol{x} \leq t] = \int_0^t f_{\boldsymbol{x}}(x)dx$$

and

$$f_{\boldsymbol{x}}(x) = \frac{d}{dt} F_{\boldsymbol{x}}(x)$$

We now use integration by parts, namely, $\int u dv = uv - \int v du$, to evaluate

$$\begin{aligned}\int_0^{\infty} (1 - F_{\boldsymbol{x}}(t))dt &= t[1 - F_{\boldsymbol{x}}(t)]\Big|_0^{\infty} + \int_0^{\infty} t f_{\boldsymbol{x}}(t)dt \\ &= 0 + \int_0^{\infty} t f_{\boldsymbol{x}}(t)dt \\ &= \mathbb{E}\boldsymbol{x}\end{aligned}$$

where we used the fact that $\lim_{t\to+\infty} F_{\boldsymbol{x}}(t) = 1$.

If $\boldsymbol{x}$ is nonnegative, we express it as the combination of two random variables as follows:

$$\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$$

where $\boldsymbol{y} = \boldsymbol{x}\mathbb{I}[\boldsymbol{x} \geq 0] \geq 0$ and $\boldsymbol{z} = \boldsymbol{x}\mathbb{I}[\boldsymbol{x} \leq 0] \leq 0$. It is clear that

$$\mathbb{E}\,\boldsymbol{y} = \int_0^\infty \mathbb{P}[\boldsymbol{x} \geq t]dt$$

On the other hand, the variable $\boldsymbol{z}$ is defined for $\boldsymbol{x} \leq 0$. Note that

$$\begin{aligned}
\mathbb{E}\,\boldsymbol{z} &= \mathbb{E}\Big(\boldsymbol{x}\mathbb{I}[-\boldsymbol{x} \geq 0]\Big) \\
&= -\mathbb{E}\Big(-\boldsymbol{x}\mathbb{I}[-\boldsymbol{x} \geq 0]\Big) \\
&= -\int_0^\infty \mathbb{P}[-\boldsymbol{x} \geq t]dt \\
&= -\int_0^\infty \mathbb{P}[\boldsymbol{x} \leq -t]dt \\
&= \int_{-\infty}^0 \mathbb{P}[\boldsymbol{x} \leq t']dt', \quad \text{using } t' = -t
\end{aligned}$$

We conclude that

$$\mathbb{E}\,\boldsymbol{x} = \int_0^\infty \mathbb{P}[\boldsymbol{x} \geq t]dt + \int_{-\infty}^0 \mathbb{P}[\boldsymbol{x} \leq t]dt$$

When $\boldsymbol{x}$ happens to be discrete and nonnegative, the cdf will have jumps at the integer locations. In particular, it will hold that

$$\begin{aligned}
F_{\boldsymbol{x}}(0) &= \mathbb{P}[\boldsymbol{x} \leq 0] = \mathbb{P}[\boldsymbol{x} = 0] \\
F_{\boldsymbol{x}}(1) &= \mathbb{P}[\boldsymbol{x} \leq 1] = \mathbb{P}[\boldsymbol{x} = 0] + \mathbb{P}[\boldsymbol{x} = 1] \\
F_{\boldsymbol{x}}(2) &= \mathbb{P}[\boldsymbol{x} \leq 2] = \mathbb{P}[\boldsymbol{x} = 0] + \mathbb{P}[\boldsymbol{x} = 1] + \mathbb{P}[\boldsymbol{x} = 2] \\
&\vdots
\end{aligned}$$

and so on, so that

$$\mathbb{P}[\boldsymbol{x} = n] = F_{\boldsymbol{x}}(n) - F_{\boldsymbol{x}}(n-1)$$

Therefore,

$$\begin{aligned}
\mathbb{E}\,\boldsymbol{x} &\triangleq \sum_{n=0}^\infty n\mathbb{P}[\boldsymbol{x} = n] \\
&= \sum_{n=0}^\infty n\Big(F_{\boldsymbol{x}}(n) - F_{\boldsymbol{x}}(n-1)\Big) \\
&= \sum_{n=0}^\infty n\Big([1 - F_{\boldsymbol{x}}(n-1)] - [1 - F_{\boldsymbol{x}}(n)]\Big) \\
&= [1 - F_{\boldsymbol{x}}(0)] - [1 - F_{\boldsymbol{x}}(1)] + 2[1 - F_{\boldsymbol{x}}(1)] - 2[1 - F_{\boldsymbol{x}}(2)] + 3[1 - F_{\boldsymbol{x}}(2)] - 3[1 - F_{\boldsymbol{x}}(3)] + \ldots \\
&= [1 - F_{\boldsymbol{x}}(0)] + [1 - F_{\boldsymbol{x}}(1)] + [1 - F_{\boldsymbol{x}}(2)] + [1 - F_{\boldsymbol{x}}(3)] + \ldots \\
&= \sum_{n=0}^\infty \Big(1 - F_{\boldsymbol{x}}(n)\Big) \\
&= \sum_{n=0}^\infty \mathbb{P}[\boldsymbol{x} \geq n]
\end{aligned}$$

□

4) (**Chapter 8**) Consider the following set defined in terms of the $p-$norm of a vector x for $p > 0$:

$$\mathcal{S}_p = \Big\{x \in \mathbb{R}^M,\ \|x\|_p \leq 1\Big\}$$

For which values of p is this set convex?

Solution: For every $p \geq 1$, the ℓ_p-norm is convex, i.e.,

$$\|\alpha x + (1-\alpha)y\|_p \leq \alpha\|x\|_p + (1-\alpha)\|y\|_p, \quad \alpha \in [0,1]$$

It follows that $\mathcal{S}_p$ will be a convex set for $p \geq 1$. Now consider the case $0 < p < 1$. In this situation, the set $\mathcal{S}_p$ is not convex. Consider the vectors

$$x = e_1, \quad y = e_M$$

We have

$$\begin{aligned} \|x\|_p &= \left(\sum_{m=1}^{M} x_m^p\right)^{1/p} = 1 \\ \|y\|_p &= \left(\sum_{m=1}^{M} y_m^p\right)^{1/p} = 1 \end{aligned}$$

Both points belong to $\mathcal{S}_p$. Next, consider the convex combination

$$z = \frac{1}{2}x + \frac{1}{2}y$$

and note that

$$\|z\|_p = \left(\sum_{m=1}^{M} z_m^p\right)^{1/p} = \Big(\frac{1}{2^p} + \frac{1}{2^p}\Big)^{1/p} = \frac{1}{2}2^{1/p} = 2^{\frac{1-p}{p}}$$

The norm is not bounded by 1 for $0 < p < 1$ and therefore $z \notin \mathcal{S}_p$.

□

5) (**Chapter 11**) Let $P(w) = q(w) + E(w)$ where $w \in \mathbb{R}^M$, $q(w)$ is closed, proper, convex function, and $E(w)$ is also a convex function with $\delta-$Lipschitz gradients. Let $w_2 = \text{prox}_{\mu q}(w - \mu p)$ where $\mu > 0$ and $p \in \mathbb{R}^M$. Show that for any $w_1 \in \mathbb{R}^M$, it holds that

$$q(w_2) \leq q(w_1) + p^{\mathsf{T}}(w_2 - w_1) + \frac{1}{2\mu}\|w - w_1\|^2 - \frac{1}{2\mu}\|w - w_2\|^2 - \frac{1}{2\mu}\|w_2 - w_1\|^2$$

Solution: Since $w_2 = \text{prox}_{\mu q}(w - \mu p)$, we know from (11.13) that

$$\frac{1}{\mu}\Big(w_2 - (w - \mu p)\Big) \in \partial_{w^{\mathsf{T}}} q(w_2)$$

That is,

$$\frac{1}{\mu}(w_2 - w) + p \in \partial_{w^{\mathsf{T}}} q(w_2)$$

Now, since $q(w)$ is convex we have, for any w_1:

$$q(w_1) \geq q(w_2) + \partial_w q(w_2)(w_1 - w_2)$$

That is,

$$\begin{aligned} q(w_2) &\leq q(w_1) - \partial_w q(w_2)(w_1 - w_2) \\ &= q(w_1) - \left(\frac{1}{\mu}(w_2 - w) + p\right)^{\mathsf{T}}(w_1 - w_2) \\ &= q(w_1) + p^{\mathsf{T}}(w_2 - w_1) - \frac{1}{\mu}(w_2 - w)^{\mathsf{T}}(w_1 - w_2) \end{aligned}$$

Expanding the rightmost term gives

$$\begin{aligned} (w_2 - w)^{\mathsf{T}}(w_1 - w_2) &= (w_2 - w_1 + w_1 - w)^{\mathsf{T}}(w_1 - w_2) \\ &= -\|w_2 - w_1\|^2 + (w_1 - w)^{\mathsf{T}}(w_1 - w_2) \\ &= -\|w_2 - w_1\|^2 + (w_1 - w)^{\mathsf{T}}(w_1 - w + w - w_2) \\ &= -\|w_2 - w_1\|^2 + \|w_1 - w\|^2 + (w_1 - w_2 + w_2 - w)^{\mathsf{T}}(w - w_2) \\ &= -\|w_2 - w_1\|^2 + \|w_1 - w\|^2 - \|w_2 - w\|^2 + (w_1 - w_2)^{\mathsf{T}}(w - w_2) \end{aligned}$$

The last term on the RHS coincides with the term on the left hand side (apart from a negative sign). Therefore,

$$2(w_2 - w)^{\mathsf{T}}(w_1 - w_2) = -\|w_2 - w_1\|^2 + \|w_1 - w\|^2 - \|w_2 - w\|^2$$

and we get

$$\begin{aligned} q(w_2) &\leq q(w_1) - \partial_w q(w_2)(w_1 - w_2) \\ &= q(w_1) - \left(\frac{1}{\mu}(w_2 - w) + p\right)^{\mathsf{T}}(w_1 - w_2) \\ &= q(w_1) + p^{\mathsf{T}}(w_2 - w_1) + \frac{1}{2\mu}\|w_1 - w\|^2 - \frac{1}{2\mu}\|w_2 - w_1\|^2 - \frac{1}{2\mu}\|w_2 - w\|^2 \end{aligned}$$

□

6) (**Chapter 12**) Consider a first-order differentiable risk function $P(w) : \mathbb{R}^M \rightarrow \mathbb{R}$. We seek a minimizer $w^\star$ for $P(w)$ by means of the gradient-descent recursion with a constant step size parameter,

$$w_n = w_{n-1} - \mu \nabla_{w^{\mathsf{T}}} P(w_{n-1}), \quad n \geq 0$$

Assume the initial condition w_{-1} is such that $\|\widetilde{w}_{-1}\| \leq W$, where $\widetilde{w}_n = w^\star - w_n$. We focus on the excess risk quantity $\Delta P(n) = P(w_n) - P(w^\star)$. Assume the step-size parameter is small enough to ensure a decaying risk value.

(a) Assume first that $P(w)$ is $\nu-$strongly convex with $\delta-$Lipschitz gradients. Show that the number of iterations necessary for $\Delta P(n) \leq \epsilon$ is $O(\ln(1/\epsilon))$.

(b) Assume next that $P(w)$ is only convex with $\delta-$Lipschitz gradients. Show that the number of iterations necessary for $\Delta P(n) \leq \epsilon$ is $O(1/\epsilon)$.

(c) Assume now that $P(w)$ is convex and $\delta-$Lipschitz itself (rather than its gradients). Show that the number of iterations necessary for $\Delta P(n) \leq \epsilon$ is $O(1/\epsilon^2)$.

Solution:

(a) We know from result (12.43b) in the text that

$$\Delta P_n \leq \frac{\delta}{2} W^2 \lambda^{n+1}$$

where $\lambda = 1 - 2\mu\nu + \mu^2\delta^2 \in [0, 1]$ for $0 < \mu < 2\nu/\delta^2$. Setting $\Delta P_n \leq \epsilon$ gives

$$\frac{\delta}{2} W^2 \lambda^{n+1} \leq \epsilon$$

which leads to $n \geq O(\ln(1/\epsilon))$. It is worth remarking that $P(w_n)$ is nonincreasing as can be seen, for example, from (12.55) for $\mu < 2/\delta$.

(b) We know from Prob. 12.13 part (d) that for $\mu < 1/\delta$,

$$\Delta P_n \leq \frac{1}{2\mu n} W^2$$

Setting $\Delta P_n \leq \epsilon$ gives

$$\frac{1}{2\mu n} W^2 \leq \epsilon$$

which leads to $n \geq O(1/\epsilon)$. Again it is worth remarking that $P(w_n)$ is nonincreasing for $\mu < 1/\delta$. Indeed, using property (10.13) for convex functions with $\delta-$Lipschitz gradients, we get

$$\begin{aligned} P(w_n) &\leq P(w_{n-1}) + \nabla_w P(w_{n-1})(w_n - w_{n-1}) + \frac{\delta}{2}\|w_n - w_{n-1}\|^2 \\ &= P(w_{n-1}) - \mu \nabla_w P(w_{n-1}) \nabla_{w^{\mathsf{T}}} P(w_{n-1}) + \frac{\delta\mu^2}{2}\|\nabla_{w^{\mathsf{T}}} P(w_{n-1})\|^2 \\ &= P(w_n - 1) - \mu\|\nabla_w P(w_{n-1})\|^2 + \frac{\delta\mu^2}{2}\|\nabla_w P(w_{n-1})\|^2 \\ &\leq P(w_{n-1}) - \mu\|\nabla_w P(w_{n-1})\|^2 + \frac{\mu}{2}\|\nabla_w P(w_{n-1})\|^2 \\ &= P(w_{n-1}) - \frac{\mu}{2}\|\nabla_w P(w_{n-1})\|^2 \end{aligned}$$

where the last inequality follows from the condition $\mu < 1/\delta$.

(c) We also note that the risk function is nonincreasing since, by convexity,

$$\begin{aligned} P(w_n) &\leq P(w_{n-1}) + \nabla_w P(w_{n-1})(w_n - w_{n-1}) \\ &= P(w_{n-1}) - \mu\|\nabla_w P(w_{n-1})\|^2 \end{aligned}$$

where we used the gradient descent update in the second equality. Next, we know from (10.41) that the condition of a Lipschitz function $P(w)$ translates into bounded gradients, i.e., $\|\nabla_w P(w)\| \leq \delta$. Now note that

$$\begin{aligned} \|\widetilde{w}_n\|^2 &= \|\widetilde{w}_{n-1} + \mu \nabla_{w^{\mathsf{T}}} P(w_{n-1})\|^2 \\ &= \|\widetilde{w}_{n-1}\|^2 + 2\mu \widetilde{w}_{n-1}^{\mathsf{T}} \nabla_{w^{\mathsf{T}}} P(w_{n-1}) + \mu^2 \|\nabla_{w^{\mathsf{T}}} P(w_{n-1})\|^2 \\ &\leq \|\widetilde{w}_{n-1}\|^2 + 2\mu \widetilde{w}_{n-1}^{\mathsf{T}} \nabla_{w^{\mathsf{T}}} P(w_{n-1}) + \mu^2\delta^2 \end{aligned}$$

From the convexity of $P(w)$ we have

$$P(w^\star) \geq P(w_{n-1}) + \nabla_w P(w_{n-1})(w^\star - w_{n-1})$$

or equivalently

$$\nabla_w P(w_{n-1}) \widetilde{w}_{n-1} \leq P(w^\star) - P(w_{n-1})$$

so that

$$\|\widetilde{w}_n\|^2 \leq \|\widetilde{w}_{n-1}\|^2 + 2\mu\Big(P(w^\star) - P(w_{n-1})\Big) + \mu^2\delta^2$$

We conclude by iterating that

$$0 \leq \|\widetilde{w}_n\|^2 \leq W^2 - 2\sum_{m=0}^{n} \mu(m)\left(P(w_{m-1}) - P(w^\star)\right) + \mu^2\delta^2 n$$

Since $P(w_n)$ is nonincreasing, we know that, for any $0 \leq m \leq n$:

$$P(w_n) - P(w^\star) \leq P(w_{m-1}) - P(w^\star)$$

and we arrive at

$$\Delta P_n = P(w_n) - P(w^\star) \;\leq\; \frac{W^2 + \mu^2\delta^2 n}{2n\mu} \;=\; \frac{W^2}{2n\mu} \;+\; \frac{\mu\delta^2}{2}$$

We can bound each term on the RHS by $\epsilon/2$. Thus, setting $\mu\delta^2/2 \leq \epsilon/2$ gives $\mu < \epsilon/\delta^2$. And setting

$$\frac{W^2}{2n\mu} \leq \frac{\epsilon}{2}$$

gives $n \geq W^2\delta^2/\epsilon^2$.

□