ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Foundations of Data Science
Fall 2024

Assignment date: Friday, January 29, 2025, 9:15
Due date: Friday, January 29, 2025, 12:15

Final Exam – INJ 218

This exam is open book. No electronic devices of any kind are allowed. There are 5 problems. Choose the ones you find easiest and collect as many points as possible. We do not necessarily expect you to finish all of them. Good luck!

Name: _____

Problem 1	/ 12
Problem 2	/ 10
Problem 3	/ 9
Problem 4	/ 12
Problem 5	/ 9
Total	/52

Problem 1 (Donsker-Varadhan to Pinsker inequality – 12 pts)**.** In this problem, we further explore information measures.

Remark: If you refer to class materials, be precise (Theorem or equation numbers, Homework problem identifiers and so on.) Your overall argument must be complete.

Let Z be an arbitrary random variable and let $f(z)$ be an arbitrary function satisfying $0 \leq f(z) \leq b$ for all values z.

(i) [3 pts] Prove that for any distribution Q, we have (recall $0 \leq f(z) \leq b$)

$$\log \mathbb{E}_Q \left[e^{f(Z)} \right] \leq \mathbb{E}_Q \left[f(Z) \right] + \frac{1}{8} b^2, \tag{1}$$

where as in class, the notation $\mathbb{E}_Q \left[e^{f(Z)} \right]$ means that the expectation is taken assuming that Z is distributed according to Q.

HINT: Observe that irrespective of the distribution of Z, we know that the random variable $f(Z) \in [0, b]$. Like in class, use this information to bound the moment generating function of the random variable $f(Z)$.

(ii) [3 pts] Prove that for any distributions P and Q, we have (recall $0 \leq f(z) \leq b$)

$$\mathbb{E}_P[f(Z)] - \mathbb{E}_Q \left[f(Z) \right] \leq D(P \| Q) + \frac{1}{8} b^2. \tag{2}$$

where the KL divergence is computed with respect to the natural logarithm.

(iii) [3 pts] Prove that for arbitary distributions P and Q,

$$\max_f \mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)] = \frac{b}{2} \|P - Q\|_1, \tag{3}$$

where the maximum is over all functions $f(z)$ satisfying $0 \leq f(z) \leq b$ for all values z.

(iv) [3 pts] Using Parts (ii) and (iii), prove the Pinsker inequality (Example 4.1 in the lecture notes). That is, prove that for arbitary distributions P and Q, we have

$$\|P - Q\|_1 \leq \sqrt{2 D(P \| Q)}, \tag{4}$$

where the KL divergence is computed with respect to the natural logarithm.

Solution 1. We take up the items in turn:

(i) For the first item, observe that we must have (a.s.)

$$-\mathbb{E}_Q \left[f(Z) \right] \leq f(Z) - \mathbb{E}_Q \left[f(Z) \right] \leq b - \mathbb{E}_Q \left[f(Z) \right], \tag{5}$$

meaning that the random variable $f(Z) - \mathbb{E}_Q[f(Z)]$ is supported on an interval of length b. But then, we know from Lemma 2.4 that

$$\mathbb{E}_Q\left[e^{\lambda\left(f(Z)-\mathbb{E}_Q[f(Z)]\right)}\right] \leq e^{\lambda^2 b^2/8}. \tag{6}$$

Then, since $\mathbb{E}_Q[f(Z)]$ is constant, we have:

$$\mathbb{E}_Q\left[e^{\lambda f(Z)}\right] \leq e^{\lambda \mathbb{E}_Q[f(Z)]} e^{\lambda^2 b^2/8}. \tag{7}$$

Plugging in $\lambda = 1$ and taking (natural) logarithms on both sides gives the result.

(ii) This is directly Donsker-Varadhan, combined with Part (i):

$$D(P\|Q) \geq \mathbb{E}_P[f(Z)] - \log \mathbb{E}_Q[e^{f(Z)}] \tag{8}$$

$$\geq \mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)] - \frac{1}{8}b^2. \tag{9}$$

Recall that here, we are assuming that the KL divergence is computed with respect to the natural logarithm. (Otherwise, there would be a correction factor.)

(iii) This is implied by Lemma 11.2 in the lecture notes. Simply re-scale the function. Namely,

$$\max_{f:0\leq f(z)\leq b} \mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)] = b \max_{f:0\leq f(z)\leq 1} \mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)] \tag{10}$$

$$= b\left(\frac{\|P-Q\|_1}{2}\right), \tag{11}$$

where the second step is Lemma 11.2 from the lecture notes.

(iv) Since Part (ii) holds for any function $f(z)$, it specifically also holds for the maximizing function in Part (iii), call this function $f^*(z)$. In this sense, combining Parts (ii) and (iii), we have

$$\frac{b}{2}\|P-Q\|_1 = \mathbb{E}_P[f^*(Z)] - \mathbb{E}_Q[f^*(Z)] \leq D(P\|Q) + \frac{1}{8}b^2. \tag{12}$$

Let us read this as

$$\|P-Q\|_1 \leq \frac{2}{b}D(P\|Q) + \frac{b}{4}. \tag{13}$$

This bounds holds for every choice of $b > 0$. Selecting $b = \sqrt{8D(P\|Q)}$ gives the claimed bound. ... well, that's if you happen to guess this choice! Rather than guessing, better to just optimize the bound: simply take derivatives to obtain:

$$\frac{d}{db}\left(\frac{2}{b}D(P\|Q) + \frac{b}{4}\right) = -\frac{2D(P\|Q)}{b^2} + \frac{1}{4}. \tag{14}$$

Setting this to zero gives the claimed $b = \sqrt{8D(P\|Q)}$.

Problem 2 (χ^2 Divergence Distance Measure – 10 pts)**.** In class we defined the ℓ_1 distance, the ℓ_2 distance, as well as the KL divergence measure. But there are other distance measures that are important and used in practice.

One of those is the χ^2 divergence distance measure. It is defined as

$$\chi^2(p,q) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i}.$$

Recall the definition of the min-max loss for a given distance measure L, an alphabet size of k and assuming we have n iid samples:

$$r_{k,n}^{L} = \min_{q} \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q(X^n))].$$

(i) [2 pt] Show that, alternatively, $\chi^2(p,q) = \sum_{i=1}^{k} \frac{p_i^2}{q_i} - 1$.

(ii) [8 pts] Show that for $k \geq 2$ and $n \geq 1$, $r_{k,n}^{\chi^2} \leq \frac{k-1}{n+1}$. [Note: There is also a corresponding lower bound that is fairly close to this upper bound showing that this upper bound is relatively tight, but we will only be concerned with the upper bound.]

HINT: We are looking for an upper bound on the min-max loss. Hence, we are free to consider any particular estimator. The add+1 estimator is your friend. Also, remember that $\frac{\binom{n}{t}}{t+1} = \frac{\binom{n+1}{t+1}}{n+1}$.

Solution 2.

(i) Since $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1$, we have:

$$\chi^2(p,q) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i} = -1 + \sum_{i=1}^{k} \frac{p_i^2}{q_i}.$$

(ii) As suggested by the hint, we will use the add+1 estimator. Hence we have:

$$\begin{aligned}
r_{k,n}^{L} &= \min_{q} \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q(X^n))] \\
&\leq \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}\left[\chi^2(p, q^{+1}(X^n))\right] \\
&= \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}\left[-1 + \sum_{i=1}^{k} \frac{p_i^2}{\frac{T_i(X^n)+1}{n+k}}\right] \\
&= \max_{p \in \Delta_k} -1 + (n+k) \sum_{i=1}^{k} p_i^2 \mathbb{E}_{X^n \sim p}\left[\frac{1}{T_i(X^n)+1}\right].
\end{aligned}$$

Now note that

$$\begin{aligned} p_i^2 \mathbb{E}_{X^n \sim p}\left[\frac{1}{T_i(X^n)+1}\right] &= p_i^2 \sum_{t=0}^{n} \frac{1}{t+1}\binom{n}{t} p_i^t (1-p_i)^{n-t} \\ &= \frac{p_i}{n+1} \sum_{t=0}^{n} \binom{n+1}{t+1} p_i^{t+1} (1-p_i)^{n-t} \\ &= \frac{p_i}{n+1} \sum_{t=1}^{n+1} \binom{n+1}{t} p_i^{t} (1-p_i)^{n+1-t} \\ &= \frac{p_i(1-(1-p_i)^{n+1})}{n+1} \leq \frac{p_i}{n+1}, \end{aligned}$$

where the first equality is true since $T_i(X^n)$ is distributed according to a binomial random variable with parameter p_i and hence the probability of $T_i(X^n)$ being t is equal to $\binom{n}{t} p_i^t (1-p_i)^{n-t}$.

Therefore,

$$r_{k,n}^L \leq \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[\chi^2(p, q^{+1}(X^n))] \leq -1 + (n+k) \sum_{i=1}^{k} \frac{p_i}{n+1} = \frac{k-1}{n+1}.$$

The right-hand side is independent of p and hence this expression is also the desired upper bound.

Problem 3 (Exponential Families and Conjugate Priors – 9 pts)**.** Let $p_\theta(x) = h(x)e^{\langle \phi(x),\theta\rangle - A(\theta)}$ denote a generic exponential family with sufficient statistics $\phi(x)$ and parameter θ.

Assume that we receive iid samples from this family, call them $\{x_i\}_{i=1}^n$. From these samples, we want to infer the unknown parameter θ via a maximum a-posteriori (MAP) procedure.

In order to apply a MAP procedure we need to define a prior distribution on the parameter θ. Consider the family of prior distributions $q_{\mu,\lambda}(\theta) = K(\mu,\lambda)e^{\langle \theta,\mu\rangle - \lambda A(\theta)}$, parametrized by (μ,λ). Note that this is also an exponential family. However, we have written it in a slightly non-standard form, where $K(\mu,\lambda)$ denotes the normalization constant which is a function of the parameters (μ,λ).

(i) [3 pts] Write down the posterior distribution $p_{\mu,\lambda}(\theta \mid x_1,\cdots,x_n)$ for a fixed set of parameters (μ,λ).

(ii) [3 pts] If you have not already done so in part (i), write the posterior as explicitly and compactly as you can. Justify why we called $q_{\mu,\lambda}(\theta)$ a conjugate prior.

(iii) [3 pts] Derive the MAP estimator of the parameter θ given the samples $\{x_i\}_{i=1}^n$ starting with the posterior derived in (ii). When will the estimate be unique?

Solution 3.

(i)/(ii) We have (where in the following Z denotes a normalization constant, not necessarily always the same):

$$\begin{aligned}
p_{\mu,\lambda}(\theta \mid x_1,\cdots,x_n) &= \frac{p_{\mu,\lambda}(\theta)p(x_1,\cdots,x_n \mid \theta)}{p(x_1,\cdots,x_n)} \\
&= \frac{1}{Z}K(\mu,\lambda)e^{\langle \theta,\mu\rangle - \lambda A(\theta)} \prod_{i=1}^n h(x_i)e^{\langle \phi(x_i),\theta\rangle - A(\theta)} \\
&= \frac{1}{Z}e^{\langle \theta,\mu+\sum_{i=1}^n \phi(x_n)\rangle - (\lambda+n)A(\theta)} \\
&= K(\mu+\sum_{i=1}^n \phi(x_n),\lambda+n)e^{\langle \theta,\mu+\sum_{i=1}^n \phi(x_n)\rangle - (\lambda+n)A(\theta)} \\
&= q_{\mu+\sum_{i=1}^n \phi(x_n),\lambda+n}(\theta)
\end{aligned}$$

In the first step we used Bayes rule. In the second step we plugged in the various expressions, keeping in mind that $p(x_1,\cdots,x_n)$ only influences the normalization and can therefore be omitted. In the third step we consolidated the expression. In the fourth and firths step we take into account the resulting expression has the same form as the prior but just with different parameters.

The chosen prior is a conjugate prior since the posterior is again a member of the exponential family.

(iii) In order to find the MAP estimate we have to find the θ that maximes $p_{\mu,\lambda}(\theta \mid x_1, \cdots, x_n)$. Let $\tilde{\mu} = \mu + \sum_{i=1}^{n} \phi(x_n)$ and $\tilde{\lambda} = \lambda + n$. Taking the gradient wrt to the parameter θ and setting the result to 0 we arrive at

$$\nabla_\theta e^{\langle \theta, \tilde{\mu} \rangle - \tilde{\lambda} A(\theta)} = e^{\langle \theta, \tilde{\mu} \rangle - \tilde{\lambda} A(\theta)} (\tilde{\mu} - \tilde{\lambda} \nabla_\theta A(\theta)) = 0.$$

The solution is therefore a θ^* so that $\nabla_\theta A(\theta^*) = \mathbb{E}_{X \sim p_{\theta^*}(x)}[\phi(X)] = \frac{\tilde{\mu}}{\tilde{\lambda}} = \frac{\mu + \sum_{i=1}^{n} \phi(x_n)}{\lambda + n}$. If the family $p_\theta(x)$ is minimal, there will be a unique such value θ^*.

Problem 4 (Fano method – 12 pts)**.** In this problem, we will develop a framework to find lower bounds on the estimation error of the minimax distribution estimator. We will use Fano's inequality, which we saw in class. First, recall the minimax distribution estimation problem:

$$r_{k,n}^L = \min_q \sup_{p \in \Delta_k} \mathbb{E}_{X^n \sim p^n}[L(p, q(X^n))].$$

Assume that the loss L is symmetric in its arguments, satisfies the triangle inequality, and that $L(x, x) = 0 \ \forall x$.

(i) [3 pts] Let $\mathcal{P} := \{P_1, \ldots, P_m\}$ be a collection of distributions such that $L(P_i, P_j) \geq \delta > 0$ for $i \neq j$.

Show that

$$\sup_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q(X^n))] \geq \frac{\delta}{2} \max_j \mathbb{P}_{X^n \sim P_j^n} \left[L(P_j, q(X^n)) \geq \frac{\delta}{2} \right].$$

HINT: For a non-negative random variable X we have $\mathbb{E}[X] \geq \epsilon \, \mathbb{P}[X \geq \epsilon]$.

(ii) [3 pts] Now let $V, X^n \sim P_{V,X^n}$ be jointly distributed such that V is uniformly distributed over $[1 : m]$ and $\mathbb{P}[X^n = x^n | V = j] = P_j^n(x^n)$. Define $Z := \arg\min_j L(q(X^n), P_j)$. Show that

$$\mathbb{P}[Z \neq V] \leq \max_j \mathbb{P}_{X^n \sim P_j^n} \left[L(q(X^n), P_j) \geq \frac{\delta}{2} \right].$$

(iii) [3 pts] Use Fano's inequality to show that

$$\max_j \mathbb{P}_{X^n \sim P_j^n} \left[L(p, q(X^n)) \geq \frac{\delta}{2} \right] \geq 1 - \frac{I(X^n; V) + \log 2}{\log m}.$$

HINT: Recall $I(Y; W) = H(Y) - H(Y|W)$.

(iv) [3 pts] Show that

$$I(X^n; V) \leq \frac{1}{m^2} \sum_{i,j \in [1:m]} D(P_i^n \| P_j^n) \leq n \max_{i,j} D(P_i \| P_j),$$

and thus

$$r_{k,n}^L \geq \frac{\delta}{2} \left(1 - \frac{n \max_{i,j} D(P_i \| P_j) - \log 2}{\log m} \right).$$

HINT: $I(W; Y) = D(P_{W|Y} \| P_W | P_Y)$, and the KL divergence is a convex function.

Solution 4. (i) The triangle inequality along with symmetry and $L(x,x)=0$ implies the nonnegativity of the loss function. Using the hint,

$$\mathbb{E}[L(p,q(X^n))] \geq \frac{\delta}{2}\mathbb{P}\left[L(p,q(X^n)) \geq \frac{\delta}{2}\right].$$

Taking the supremum on both sides of the inequality,

$$\begin{aligned}\sup_{p\in\Delta_k}\mathbb{E}[L(p,q(X^n))] &\geq \sup_{p\in\Delta_k}\frac{\delta}{2}\mathbb{P}\left[L(p,q(X^n)) \geq \frac{\delta}{2}\right] \\ &\geq \frac{\delta}{2}\sup_{p\in\mathcal{P}}\mathbb{P}\left[L(p,q(X^n)) \geq \frac{\delta}{2}\right] = \frac{\delta}{2}\max_j \mathbb{P}_{X^n\sim P_j^n}\left[L(P_j,q(X^n)) \geq \frac{\delta}{2}\right].\end{aligned}$$

where the second inequality is due to the expression being maximized over a smaller set.

(ii)

$$\begin{aligned}\mathbb{P}[Z\neq V] &= \frac{1}{m}\sum_j \mathbb{P}[Z\neq V|V=j] \\ &= \frac{1}{m}\sum_j \mathbb{P}_{X^n\sim P_j^n}\left[\min_i L(q(X^n),P_i)\neq j\right]\end{aligned}$$

Now, since $\min_{i\neq j} L(P_i,P_j)\geq\delta$, the triangle inequality requires $\min_j L(q(X^n),P_j)\neq i \implies L(q(X^n),P_i)\geq\frac{\delta}{2}$. Consequently,

$$\begin{aligned}\mathbb{P}[Z\neq V] &\leq \frac{1}{m}\sum_j \mathbb{P}_{X^n\sim P_j^n}\left[L(q(X^n),P_j)\geq\frac{\delta}{2}\right] \\ &\leq \max_j \mathbb{P}_{X^n\sim P_j^n}\left[L(q(X^n),P_j)\geq\frac{\delta}{2}\right].\end{aligned}$$

(iii) Notice that Z is a prediction about V, based on X^n. Directly applying Fano's inequality from the lecture notes,

$$\begin{aligned}\log m\cdot\mathbb{P}[Z\neq V] &\geq H(V|X^n) - h_2(\mathbb{P}[Z\neq V]) \\ &= \log m - I(V;X^n) - h_2(\mathbb{P}[Z\neq V]) \\ &\geq \log m - I(V;X^n) - \log 2.\end{aligned}$$

(iv) Using the hint,

$$I(X^n;V) = D(P_{X^n|V}\|P_{X^n}|P_V).$$

The marginal distribution $\bar{P}[x^n] := P_{X^n}[x^n] = \frac{1}{m}\sum_j P_j^n[x^n]$, and thus

$$\begin{aligned} I(X^n;V) =& \mathbb{E}_V[D(P_V\|\bar{P})] = \frac{1}{m}\sum_j D\left(P_j^n \middle\| \frac{1}{m}\sum_i P_i^n\right) \\ \leq& \frac{1}{m^2}\sum_{i,j} D(P_j^n\|P_i^n) \leq n \max_{i,j} D(P_j\|P_i) \end{aligned}$$

where the first inequality follows from the convexity of the KL divergence.

Problem 5 (Dual Basis – 9 pts)**.** Consider a Hilbert space H. Let $G \subseteq H$ be a (Hilbert) subspace of H, exactly like in class. We are given a basis $\{\mathbf{g}_n\}_{n=1}^N$ that spans G but is *not* orthonormal.

(i) [3 pts] Show that there exists a so-called dual basis $\{\tilde{\mathbf{g}}_n\}_{n=1}^N$ that also spans G and has the property that

$$\langle \mathbf{g}_n, \tilde{\mathbf{g}}_m \rangle = \begin{cases} 1, & \text{for } m = n, \\ 0, & \text{for } m \neq n. \end{cases} \tag{15}$$

HINT: Start by considering $\tilde{\mathbf{g}}_1 = \alpha(\mathbf{g}_1 - \sum_{n=2}^N \beta_n \mathbf{g}_n)$. Argue that we can select α and β_n appropriately. No need to give explicit formulas for these coefficients. The more "rigorous" your argument, the more points you will get.

(ii) [3 pts] Show that for any $\mathbf{x} \in H$, the minimum of $\|\mathbf{y} - \mathbf{x}\|$ over all $\mathbf{y} \in G$ is attained by the selection

$$\mathbf{y}^* = \sum_{n=1}^N \langle \mathbf{x}, \tilde{\mathbf{g}}_n \rangle \mathbf{g}_n. \tag{16}$$

HINT: As in class, here $\|\cdot\|$ denotes the Hilbert space norm induced by the inner product.

(iii) [3 pts] Show that for any $\mathbf{x} \in H$, we have

$$\sum_{n=1}^N \langle \mathbf{x}, \tilde{\mathbf{g}}_n \rangle \mathbf{g}_n = \sum_{m=1}^N \langle \mathbf{x}, \mathbf{g}_m \rangle \tilde{\mathbf{g}}_m. \tag{17}$$

Solution 5. We take up the items in turn:

(i) Show that there exists a so-called dual basis $\{\tilde{\mathbf{g}}_n\}_{n=1}^N$ that also spans G and has the property that

$$\langle \mathbf{g}_n, \tilde{\mathbf{g}}_m \rangle = \begin{cases} 1, & \text{for } m = n, \\ 0, & \text{for } m \neq n. \end{cases} \tag{18}$$

This can be proved in a number of ways. Following the hint, we may think about constructions. What comes to mind here is, of course, the Gram-Schmidt procedure. In this spirit, we first construct $\tilde{\mathbf{g}}_1$. It needs to be orthogonal to $\mathbf{g}_2, \mathbf{g}_3, \ldots, \mathbf{g}_N$. The Gram-Schmidt idea is to take $\mathbf{g}_1$ and subtract a linear combination of the vectors $\{\mathbf{g}_n\}_{n=2}^N$. This can be written as

$$\tilde{\mathbf{g}}_1' = \mathbf{g}_1 - \sum_{n=2}^N \beta_n \mathbf{g}_n. \tag{19}$$

First, let us observe that $\tilde{\mathbf{g}}_1'$ cannot be the all-zero vector. This follows by contradiction: If it were, it would imply that $\mathbf{g}_1$ lies in the span of $\mathbf{g}_2, \mathbf{g}_3, \ldots, \mathbf{g}_N$ contradicting the assumption that $\{\mathbf{g}_n\}_{n=2}^N$. is a basis. We will now argue that we can choose the coefficients β_n to satisfy the required orthogonality condition. To see that this is indeed possible, write out, for $m = 2, 3, \ldots, N$,

$$\langle \tilde{\mathbf{g}}_1', \mathbf{g}_m \rangle = \left\langle \mathbf{g}_1 - \sum_{n=2}^{N} \beta_n \mathbf{g}_n, \mathbf{g}_m \right\rangle \tag{20}$$

$$= \langle \mathbf{g}_1, \mathbf{g}_m \rangle - \left\langle \sum_{n=2}^{N} \beta_n \mathbf{g}_n, \mathbf{g}_m \right\rangle \tag{21}$$

$$= \langle \mathbf{g}_1, \mathbf{g}_m \rangle - \sum_{n=2}^{N} \beta_n \langle \mathbf{g}_n, \mathbf{g}_m \rangle , \tag{22}$$

which is thus a system of $N-1$ linear equations in the $N-1$ variables $\beta_2, \beta_3, \ldots, \beta_N$. The coefficient matrix of this system of linear equations is precisely the Gram matrix of the vectors $\mathbf{g}_2, \mathbf{g}_3, \ldots, \mathbf{g}_N$. Since these vectors are a basis, we know that they are linearly independent. This directly implies that the Gram matrix is positive definite,[1] meaning that there is indeed a unique solution for the variables $\beta_2, \beta_3, \ldots, \beta_N$. This completes the proof. Finally, just like in the Gram-Schmidt procedure, we normalize accordingly. That is, we set $\tilde{\mathbf{g}}_1 = \alpha \tilde{\mathbf{g}}_1'$, where we select the scalar α such that $\langle \tilde{\mathbf{g}}_1, \mathbf{g}_1 \rangle = 1$. It is clear that this is possible.

We can construct the remaining $\tilde{\mathbf{g}}_m$ in exactly the same fashion.

(ii) In class, we have seen that in complete generality, in a Hilbert space, $\mathbf{y}$ minimizes $\|\mathbf{y} - \mathbf{x}\|$ if and only if the error is orthogonal to every element in the subspace G. Since $\{\tilde{\mathbf{g}}_n\}_{n \in \mathbb{Z}}$ is a basis for G, this is the same as requiring that

$$\langle \mathbf{y} - \mathbf{x}, \tilde{\mathbf{g}}_m \rangle = 0 \tag{23}$$

[1] Just for completeness: Call the Gram matrix G, with entries $G_{nm} = \langle \mathbf{g}_n, \mathbf{g}_m \rangle$. Now consider $\mathbf{x}^H G \mathbf{x}$ for any $\mathbf{x} \in \mathbb{C}^{N-1}$, except $\mathbf{x} = \mathbf{0}$. If we can show that $\mathbf{x}^H G \mathbf{x} > 0$, then this proves that G is positive definite. To see that this is indeed true, write $\mathbf{x}^H G \mathbf{x} = \sum_{n=1}^{N-1} \sum_{m=1}^{N-1} x_n x_m^* G_{nm} = \sum_{n=1}^{N-1} \sum_{m=1}^{N-1} x_n x_m^* \langle \mathbf{g}_n, \mathbf{g}_m \rangle = \sum_{n=1}^{N-1} \sum_{m=1}^{N-1} \langle x_n \mathbf{g}_n, x_m \mathbf{g}_m \rangle$, by the bi-linearity of the inner product. Using the bi-linearity again, we can write this as $\mathbf{x}^H G \mathbf{x} = \left\langle \sum_{n=1}^{N-1} x_n \mathbf{g}_n, \sum_{m=1}^{N-1} x_m \mathbf{g}_m \right\rangle = \| \sum_{n=1}^{N-1} x_n \mathbf{g}_n \|^2$. But since $\mathbf{g}_2, \mathbf{g}_3, \ldots, \mathbf{g}_N$ are a basis, this is zero if and only if all of the x_n are zero, which thus completes the proof.

for all m. Let us check if this indeed holds for the claimed formula. Namely,

$$\left\langle \sum_n \langle \mathbf{x}, \tilde{\mathbf{g}}_n \rangle \mathbf{g}_n - \mathbf{x}, \tilde{\mathbf{g}}_m \right\rangle = \left\langle \sum_n \langle \mathbf{x}, \tilde{\mathbf{g}}_n \rangle \mathbf{g}_n, \tilde{\mathbf{g}}_m \right\rangle - \langle \mathbf{x}, \tilde{\mathbf{g}}_m \rangle \tag{24}$$

$$= \sum_n \langle \mathbf{x}, \tilde{\mathbf{g}}_n \rangle \langle \mathbf{g}_n, \tilde{\mathbf{g}}_m \rangle - \langle \mathbf{x}, \tilde{\mathbf{g}}_m \rangle \tag{25}$$

$$= \underbrace{\sum_n \langle \mathbf{x}, \tilde{\mathbf{g}}_n \rangle \delta(n-m)}_{= \langle \mathbf{x}, \tilde{\mathbf{g}}_m \rangle} - \langle \mathbf{x}, \tilde{\mathbf{g}}_m \rangle \tag{26}$$

where we use the Kronecker delta function $\delta(n)$. This completes the proof.

(iii) This can be proved in a number of ways. For example, by the answer to Part (ii), it should be clear that both expressions are minimizers of $\|\mathbf{y} - \mathbf{x}\|$. But we have seen in class that in a Hilbert space, the minimizer is unique. Hence, the two expressions must be equal.