ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Information Theory and Signal Processing	Assignment date: January 17th, 2019, 16:15
Fall 2018	Due date: January 17th, 2019, 19:15

Final Exam – CE1106

There are six problems. We do not presume that you will finish all of them. Choose the ones you find easiest and collect as many points as possible. Good luck!

Name: ______________________________

Problem 1	/ 10
Problem 2	/ 20
Problem 3	/ 15
Problem 4	/ 15
Problem 5	/ 20
Problem 6	/ 20
Total	/100

Problem 1. *(Canonical Correlations)*

[10pts] Let $\mathbf{X}$ and $\mathbf{Y}$ be zero-mean real-valued random vectors with covariance matrices $R_{\mathbf{X}}$ and $R_{\mathbf{Y}}$, respectively. Moreover, let $R_{\mathbf{XY}} = \mathbb{E}[\mathbf{X}\mathbf{Y}^T]$. Our goal is to find vectors $\mathbf{u}$ and $\mathbf{v}$ such as to maximize the correlation between $\mathbf{u}^T\mathbf{X}$ and $\mathbf{v}^T\mathbf{Y}$, that is,

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathbb{E}[\mathbf{u}^T\mathbf{X}\mathbf{Y}^T\mathbf{v}]}{\sqrt{\mathbb{E}[|\mathbf{u}^T\mathbf{X}|^2]}\sqrt{\mathbb{E}[|\mathbf{v}^T\mathbf{Y}|^2]}}. \tag{1}$$

Show how we can find the optimizing choices of the vectors $\mathbf{u}$ and $\mathbf{v}$ from the problem parameters $R_{\mathbf{X}}, R_{\mathbf{Y}}$, and $R_{\mathbf{XY}}$.

Hint: Recall that we have seen in class that

$$\max_{\mathbf{v}} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} = \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\| = \sigma_1(A), \tag{2}$$

where $\sigma_1(A)$ denotes the maximum singular value of the matrix A. The corresponding maximizer is the right singular vector $\mathbf{v}_1$ (i.e., eigenvector of A^TA) corresponding to $\sigma_1(A)$.

Solution: Observe that

$$\begin{align}
\mathbb{E}[\mathbf{u}^T\mathbf{X}\mathbf{Y}^T\mathbf{v}] &= \mathbf{u}^T R_{\mathbf{XY}}\mathbf{v} \tag{3}\\
\mathbb{E}[|\mathbf{u}^T\mathbf{X}|^2] &= \mathbf{u}^T R_{\mathbf{X}}\mathbf{u} \tag{4}\\
\mathbb{E}[|\mathbf{v}^T\mathbf{Y}|^2] &= \mathbf{v}^T R_{\mathbf{Y}}\mathbf{v} \tag{5}\\
& \tag{6}
\end{align}$$

So, we can express our problem as

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^T R_{\mathbf{XY}}\mathbf{v}}{\sqrt{\mathbf{u}^T R_{\mathbf{X}}\mathbf{u}}\sqrt{\mathbf{v}^T R_{\mathbf{Y}}\mathbf{v}}} \tag{7}$$

Changing coordinates as $\mathbf{a} = R_{\mathbf{X}}^{1/2}\mathbf{u}$ and $\mathbf{b} = R_{\mathbf{Y}}^{1/2}\mathbf{v}$, we obtain

$$\max_{\mathbf{a},\mathbf{b}} \frac{\mathbf{a}^T \left(R_{\mathbf{X}}^{-1/2}\right)^T R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}, \tag{8}$$

and since $R_{\mathbf{X}}^{-1/2}$ is a symmetric matrix,

$$\max_{\mathbf{a},\mathbf{b}} \frac{\mathbf{a}^T R_{\mathbf{X}}^{-1/2} R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}. \tag{9}$$

Note that this expression is invariant to scaling of the vectors $\mathbf{a}$ and $\mathbf{b}$. Hence, equivalently,

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1} \mathbf{a}^T R_{\mathbf{X}}^{-1/2} R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2}\mathbf{b}. \tag{10}$$

For a fixed choice of $\mathbf{a}$, this is the inner product of the fixed vector $\left(R_{\mathbf{X}}^{-1/2} R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2}\right)^T \mathbf{a}$ with the vector $\mathbf{b}$. We can argue for example via Cauchy-Schwarz that the maximizing choice of $\mathbf{b}$ is exactly equal to that fixed vector (normalized), that is,

$$\mathbf{b} = \frac{\left(R_{\mathbf{X}}^{-1/2} R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2}\right)^T \mathbf{a}}{\left\| \left(R_{\mathbf{X}}^{-1/2} R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2}\right)^T \mathbf{a} \right\|}. \tag{11}$$

Plugging this in, we find

$$\max_{\|\mathbf{a}\|=1} \frac{\mathbf{a}^T R_{\mathbf{X}}^{-1/2} R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2} \left(R_{\mathbf{X}}^{-1/2} R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2}\right)^T \mathbf{a}}{\left\| \left(R_{\mathbf{X}}^{-1/2} R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2}\right)^T \mathbf{a} \right\|}, \tag{12}$$

or, equivalently,

$$\max_{\|\mathbf{a}\|=1} \left\| \left(R_{\mathbf{X}}^{-1/2} R_{\mathbf{XY}} R_{\mathbf{Y}}^{-1/2}\right)^T \mathbf{a} \right\|, \tag{13}$$

or perhaps better

$$\max_{\|\mathbf{a}\|=1} \left\| R_{\mathbf{Y}}^{-1/2} R_{\mathbf{XY}}^T R_{\mathbf{X}}^{-1/2} \mathbf{a} \right\|, \tag{14}$$

Grading Notes:

- Correctly remove the expectations and express in terms of the covariance matrices: 2 Pts
- Express in eigenbasis: +1 Pt
- Idea of treating $\mathbf{u}^T R_{\mathbf{XY}} \mathbf{v}$ as an inner product and upper bounding by Cauchy-Schwarz: 6 Pt.

Problem 2. *(Growth of Expected Capital vs Expected Growth of Capital)*

Suppose $U_1, U_2, \dots$ are i.i.d. random variables taking values on a finite alphabet $\mathcal{U}$; let $P(u) = \Pr(U_1 = u)$ denote their common distribution. As in class let $\hat{P}_n$ denote the empirical distribution of U^n.

Suppose $f : \mathcal{U} \to [0, \infty)$ is a non-negative real valued function defined on $\mathcal{U}$. Define now the random variables $X_0, X_1, \dots$ as $X_0 = 1$, $X_n = f(U_n)X_{n-1}$, $\forall n \geq 1$. In other words

$$X_n = \prod_{i=1}^{n} f(U_i).$$

One refers to the value $R_n = \frac{1}{n} \log X_n$ as the (exponential) *rate of growth* of X_n. (The terminology is motivated by the relationship $X_n = \exp(nR_n)$).

Fix $\alpha = \sum_u P(u) \log f(u) = E[\log f(U)]$, and for a given $\epsilon > 0$, let

$$A = \bigl\{Q \in \Pi : \bigl|\textstyle\sum_u Q(u) \log f(u) - \alpha\bigr| < \epsilon\bigr\}.$$

Let $D^* = \min_{Q \notin A} D(Q\|P)$. Observe that $D^* > 0$.

(a) [5pts] What can you say about $\Pr(|R_n - \alpha| \geq \epsilon)$ as n gets large? *Hint:* How are the events $\{|R_n - \alpha| \geq \epsilon\}$ and $\{\hat{P}_n \notin A\}$ related?

(b) [5pts] Let $\beta = \log E[f(U)]$. What is the relationship between $e_n = \frac{1}{n} \log E[X_n]$ and β? Which one of α and β is larger?

In a casino a game of chance is played. The outcome of the game is a random variable U, and if the outcome is u, the money bet on that outcome is multipled by a factor $\phi(u)$. The money bet on other outcomes is lost. The game can be played successively with independent, identically distributed outcomes.

We allocate our capital among the outcomes by placing a fraction $q(u)$ of it on outcome u. Clearly $q(u) \geq 0$ and $Q = \sum_u q(u) \leq 1$. (The fraction $1 - Q$ is the fraction of our capital not bet on the game and kept in reserve.) Observe that $f(u) = (1 - Q) + q(u)\phi(u)$ is the factor our capital is multipled by if the outcome of the game is u.

Let $X_0 = 1$ be our initial capital, and let X_n, $n = 1, 2, \dots$ denote our capital as we play the game repeatedly with a fixed allocation strategy q.

(c) [5pts] Suppose $\mathcal{U} = \{0, 1\}$, $P(0) = 1/4$, $P(1) = 3/4$, $\phi(0) = \phi(1) = 2$. What is the allocation q that maximizes the value of β in (b)?

(d) [5pts] Continuing with (c) and the allocation you just found, what is the value of α? What will happen to our capital X_n in the long run if we repeatedly play the game?

Solution:

(a) As $R_n = \frac{1}{n} \log X_n$ and $X_n = \Pi_{i=1}^n f(U_i)$, we have

$$R_n = \frac{1}{n} \log \Pi_{i=1}^n f(U_i) = \frac{1}{n} \sum_{i=1}^n \log f(U_i) = \sum_u \hat{P}_n(u) \log f(u) \tag{15}$$

Thus, the event $\{|R_n - \alpha| \geq \epsilon\}$ is the same as $\{\hat{P}_n \notin A\}$, which means $\Pr(|R_n - \alpha| \geq \epsilon) = \Pr(\hat{P}_n \notin A)$.

Let B denote the complement set of A, then $B = \{Q \in \Pi : \left|\sum_u Q(u) \log f(u) - \alpha\right| \geq \epsilon\}$ which is a closed set. Thus we have $\Pr(\hat{P}_n \notin A) = \Pr(\hat{P}_n \in B)$. Moreover, the closure of B is equal to the closure of the interior of B. According to Theorem 2.13 in the lecture notes, we have

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr(\hat{P}_n \in B) = -\inf_{Q\in B} D(Q\|P) = -\min_{Q\in B} D(Q\|P) = -D^*. \tag{16}$$

As $D^* > 0$, the probability $\Pr(\hat{P}_n \in B)$ must converge to 0 as n goes to infinity.

Therefore, we have $Pr(|R_n - \alpha| \geq \epsilon)$ goes to 0 as n gets large.

(b) As $X_n = \Pi_{i=1}^n f(U_i)$ and $U_i, \ldots, U_n$ are i.i.d, we have

$$e_n = \frac{1}{n} \log E[X_n] = \frac{1}{n} \log E[\Pi_{i=1}^n f(U_i)] = \frac{1}{n} \sum_{i=1}^n \log E[f(U_i)] = \log E(f(U)) \tag{17}$$

Hence, $e_n = \beta$ for all n.

By Jensen's inequality we have

$$\beta = \log E[f(U)] \geq E[\log f(U)] = \alpha. \tag{18}$$

(c) Since $f(u) = (1 - Q) + q(u)\phi(u)$ and $Q = \sum_u q(u)$, we have

$$\begin{aligned}
\arg \max_{q(0),q(1)} \beta &= \arg \max_{q(0),q(1)} E[f(U)] && (19)\\
&= \arg \max_{q(0),q(1)} P(0)f(0) + p(1)f(1) && (20)\\
&= \arg \max_{q(0),q(1)} P(0)(1 - Q + q(0)\phi(0)) + P(1)(1 - Q + q(1)\phi(1)) && (21)\\
&= \arg \max_{q(0),q(1)} 1 - Q + P(0)q(0)\phi(0) + P(1)q(1)\phi(1) && (22)\\
&= \arg \max_{q(0),q(1)} (\underbrace{P(0)\phi(0) - 1}_{<0})q(0) + (\underbrace{P(1)\phi(1) - 1}_{>0})q(1) && (23)
\end{aligned}$$

It is obvious that $\{q(0) = 0, q(1) = 1\}$ maximizes β.

(d) With the strategy in (c), we have $f(0) = 0$ and $f(1) = 2$. Hence,

$$\alpha = E[\log U] = P(0) \log f(0) + P(1) \log f(1) = -\infty \tag{24}$$

The probability that we lose all money is actually goes to 1 as we repeatedly play the game.

$$\Pr(X_n = 0) = 1 - \Pr(X_n \neq 0) = 1 - (3/4)^n \to 1. \tag{25}$$

In other words, the strategy of maximizing $E[X_n]$ will surely ruin us in the long term. A more reasonable strategy would be to maximize α, which can be found to be $\{q(0) = 1/4, q(1) = 3/4\}$. It is easy to show that q=P maximizes alpha in general. This choice of q is known in portfolio theory as Kelly betting.

Problem 3. *(Hypothesis Testing and Exponential Families)*

Let P denote the zero-mean and unit-variance Gaussian distribution. Assume that you are given N iid samples distributed according to P and let $\hat{P}_N$ be the empirical distribution.

Let Π denote the set of distributions with second moment $\mathbb{E}[X^2] = 2$. We are interested in

$$\lim_{N\to\infty} \frac{1}{N} \log \Pr\{\hat{P_N} \in \Pi\} = - \inf_{Q\in\Pi} D(Q\|P).$$

1. [10pts] Determine $-\text{arginf}_{Q\in\Pi} D(Q\|P)$, i.e., determine the element Q for which the infinum is taken on.

2. [5pts] Determine $-\inf_{Q\in\Pi} D(Q\|P)$.

Solution: We are looking for the I-projection of P onto Π, call the result Q. Since Π is a linear family with a single constraint on the expected value of x^2 we know that the density of the minimizing distribution has the form

$$q(x) = p(x)e^{\theta x^2 - A(\theta)}.$$

If we insert $p(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ this gives us

$$q(x) = e^{-\frac{x^2}{2}+\theta x^2 - \tilde{A}(\theta)}.$$

We recognize the right-hand side to be the density of a zero-mean Gaussian distribution and by assumption this distribution has second moment 2. Hence, the solution is a zero-mean Gaussian distribution with variance 2, i.e., $q(x) = \frac{1}{\sqrt{4\pi}}e^{-\frac{x^2}{4}}$. The asymptotic exponent is given by the KL distance between these two distributions. We have

$$\begin{aligned} D(q\|p) &= \int \frac{1}{\sqrt{4\pi}}e^{-\frac{x^2}{4}} \log \frac{\frac{1}{\sqrt{4\pi}}e^{-\frac{x^2}{4}}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}} dx \\ &= \frac{1}{2}\log\frac{1}{2} + \int \frac{1}{\sqrt{4\pi}}e^{-\frac{x^2}{4}}[-\frac{x^2}{4} + \frac{x^2}{2}]dx \\ &= \frac{1}{2}(\log\frac{1}{2} + 1) = \frac{1}{2}(-\log 2 + 1) \sim 0.153426. \end{aligned}$$

Alternatively, we can use the formula in problem 1 of homework 5 to compute $D(q\|p)$.

To summarize

1. $-\text{arginf}_{Q\in\Pi} D(Q\|P)$ is given by $q(x) = \frac{1}{\sqrt{4\pi}}e^{-\frac{x^2}{4}}$.

2. $-\inf_{Q\in\Pi} D(Q\|P) = -0.153426$.

Grading Notes:

- Correct expression for $p(x)$: 2 pts.
- Realize it is I-projection 2 pts. Know that $q(x)$ is $p(x)\times$ exponential family: 2 pts.
- Find $q(x)$ is zero-mean Gaussian with variance equal 2: 3 pts.
- Correct formula for divergence 3 pts. Correct answer 2 pts.

Problem 4. *(Choose the Shortest Description)*

Suppose $\mathcal{C}_0 : \mathcal{U} \to \{0,1\}^*$ and $\mathcal{C}_1 : \mathcal{U} \to \{0,1\}^*$ are two prefix-free codes for the alphabet $\mathcal{U}$. Consider the code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ defined by

$$\mathcal{C}(u) = \begin{cases} 0\mathcal{C}_0(u) & \text{if length}\mathcal{C}_0(u) \leq \text{length}\mathcal{C}_1(u) \\ 1\mathcal{C}_1(u) & \text{else.} \end{cases}$$

Observe that $\text{length}(\mathcal{C}(u)) = 1 + \min\{\text{length}(\mathcal{C}_0(u)), \text{length}(\mathcal{C}_1(u))\}$.

(a) [5pts] Is $\mathcal{C}$ a prefix-free code? Explain.

(b) [5pts] Suppose $\mathcal{C}_0, \ldots, \mathcal{C}_{K-1}$ are K prefix-free codes for the alphabet $\mathcal{U}$. Show that there is a prefix-free code $\mathcal{C}$ with

$$\text{length}(\mathcal{C}(u)) = \lceil \log_2 K \rceil + \min_{0 \leq k < K-1} \text{length}(\mathcal{C}_k(u)).$$

(c) [5pts] Suppose we are told that U is a random variable taking values in $\mathcal{U}$, and we are also told that the distribution p of U is one of K distributions $p_0, \ldots, p_{K-1}$, but we do not know which. Using (b) describe how to construct a prefix-free code $\mathcal{C}$ such that

$$E[\text{length}(\mathcal{C}(U))] \leq \lceil \log_2 K \rceil + 1 + H(U).$$

[Hint: From class we know that for each k there is a prefix-free code $\mathcal{C}_k$ that descibes each letter u with at most $\lceil -\log_2 p_k(u) \rceil$ bits.]

Solution:

(a) Yes, $\mathcal{C}$ is a prefix-free code. We can prove it by contradiction. Suppose there exist $u, v \in \mathcal{U}$ such that $\mathcal{C}(u)$ is a prefix of $\mathcal{C}(v)$. Then they must start with the same bit. Without loss of generality, let us assume they start with 0, then we have $\mathcal{C}(u) = 0\mathcal{C}_0(u)$ is a prefix of $\mathcal{C}(v) = 0\mathcal{C}_0(v)$. This requires $\mathcal{C}_0(u)$ is a prefix of $\mathcal{C}_0(v)$ which contradicts to $\mathcal{C}_0$ is prefix free code.

(b) Generalizing the given construction, we can construct the code $\mathcal{C}(u)$ for any $u \in \mathcal{U}$ as follows.

$$\mathcal{C}(u) = \text{Bin}(i^*)\mathcal{C}_{i^*}(u) \tag{26}$$

where $i^* = \arg\min_{0 \leq k \leq K-1\}} \text{length}\mathcal{C}_i(u)$ and $\text{Bin}(i^*)$ is the binary representation of number i^*. The length of such code is exactly the given expression and by the same reason in (a), we can show that it is prefix-free.

(c) As the hint suggests, we can use prefix free code $\mathcal{C}_k$ such that length$(\mathcal{C}_k) \leq \lceil -\log_2 p_k(u) \rceil$ and construct the prefix-free code $\mathcal{C}$ as in [b]. Then we have

$$\begin{aligned} \text{length}(\mathcal{C}(u)) &= \lceil \log_2 K \rceil + \min_{0 \leq k < K-1} \text{length}(\mathcal{C}_k(u)) && (27) \\ &\leq \lceil \log_2 K \rceil + 1 - \min_{0 \leq k < K-1} \log_2 p_k(u) && (28) \\ &\leq \lceil \log_2 K \rceil + 1 - \log_2 p(u) && (29) \end{aligned}$$

Taking expectation at both sides, we get that

$$E[\text{length}(\mathcal{C}(U))] \leq \lceil \log_2 K \rceil + 1 + H(U). \tag{30}$$

Problem 5. *(Inner Products)*

Consider the standard n-dimensional vector space $\mathbb{R}^n$.

1. [5pts] Characterize the set of matrices W for which $\mathbf{y}^T W\mathbf{x}$ is a valid inner product for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

2. [5pts] Prove that *every* inner product $\langle \mathbf{x}, \mathbf{y}\rangle$ on $\mathbb{R}^n$ can be expressed as $\mathbf{y}^T W\mathbf{x}$ for an approriately chosen matrix W.

3. [10pts] For a subspace of dimension $k < n$, spanned by the basis $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_k \in \mathbb{R}^n$, express the orthogonal projection operator (matrix) with respect to the general inner product $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{y}^T W\mathbf{x}$. *Hint:* For any vector $\mathbf{x} \in \mathbb{R}^n$, express its projection as $\widehat{\mathbf{x}} = \sum_{j=1}^k \alpha_j \mathbf{b}_j$.

Solution:

1. Looking at the lecture notes, Section 7.3, an inner product must satisfy linearity properties, which clearly hold for all matrices W. The symmetry property $\langle \mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{y}, \mathbf{x}\rangle$ only holds if the matrix W is *symmetric,* i.e., $W^T = W$. The crucial requirement is the last property, namely, $\langle \mathbf{x}, \mathbf{x}\rangle \geq 0$, with equality if and only if $\mathbf{x} = 0$. To tackle this, note that W has to be symmetric, so it has a spectral decomposition $W = U\Lambda U^H$. Hence, it is a clever idea to express the vectors $\mathbf{x}$ and $\mathbf{y}$ in terms of the eigenvectors of W. Then, clearly, if all eigenvalues of W are strictly positive, then the property is satisfied. Conversely, if there is a eigenvalue equal to zero, or a negative eigenvalue, then there exists a choice $\mathbf{x} \neq 0$ for which $\langle \mathbf{x}, \mathbf{x}\rangle = 0$. In conclusion, $\mathbf{y}^T W\mathbf{x}$ is a valid inner product if and only if W is a symmetric and positive definite.

2. To prove this, use the standard basis vectors to express $\mathbf{x} = x_1\mathbf{e}_1 + \ldots + x_n\mathbf{e}_n$, and likewise for $\mathbf{y}$. Then, using the properties of the inner product, we find ...

3. As we have seen in class, the error $\mathbf{x} - \widehat{\mathbf{x}}$ must be orthogonal to the estimate $\widehat{\mathbf{x}}$, or, equivalently, orthogonal to all of the basis vectors $\mathbf{b}_i$. That is,

$$\langle \mathbf{x} - \widehat{\mathbf{x}}, \mathbf{b}_i\rangle = 0. \tag{31}$$

Plugging in the hint $\widehat{\mathbf{x}} = \sum_{j=1}^k \alpha_j \mathbf{b}_j$, we get

$$\langle \mathbf{x} - \sum_{j=1}^k \alpha_j \mathbf{b}_j, \mathbf{b}_i\rangle = 0, \tag{32}$$

and using the standard properties of the inner product,

$$\langle \mathbf{x}, \mathbf{b}_i\rangle - \sum_{j=1}^k \alpha_j \langle \mathbf{b}_j, \mathbf{b}_i\rangle = 0. \tag{33}$$

Defining the $n \times k$ matrix

$$B = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_k), \tag{34}$$

we can collect all k conditions (for $i = 1, 2, \ldots, k$) into

$$B^H W\mathbf{x} - B^H W B\alpha = 0, \tag{35}$$

where α denotes the column vector of all the coefficients α_i. Hence,

$$\alpha = \left(B^H W B\right)^{-1} B^H W\mathbf{x}, \tag{36}$$

where we note that $B^H W B$ is invertible since the vectors $\mathbf{b}_j$ constitute a basis. Finally, we observe that we can write

$$\widehat{\mathbf{x}} = B\alpha = B\left(B^H W B\right)^{-1} B^H W\mathbf{x}, \tag{37}$$

which is thus the desired projection matrix.

Grading Notes:

- Part 1: W symmetric: 2 pts. Positive definite: 3 pts.
  - Properties of inner products: 2 pts. Spectral decomposition or eigenvalue decomposition: 1 pts.
- Part 2: correct W: 5 pts.
- Part 3: Orthogonal projection property: 3 pts. Replace inner product with $B^T W x$: 3 pts. Correct expression for α: 2 pts. Correct projection matrix: 2 pts.

Problem 6. *(Thompson Sampling with Bernoulli Losses)*

This problem deals with a Bayesian approach to multi-arm bandits. Although we will not pursue this facet in the current problem, the Bayesian approach is useful since within this framework it is relatively easy to incorporate prior information into the algorithm.

Assume that we have K bandits, and that bandit k outputs a $\{0,1\}$-valued Bernoulli random variable with parameter $\theta_k \in [0,1]$. Let π be the uniform prior on $[0,1]^K$, i.e., the uniform prior on the set of all parameters $\theta = (\theta_1, \cdots, \theta_K)$. Let

$$\begin{aligned} T_k^1(t) &= |\{\tau \leq t : A_\tau = k; Y_\tau = 1\}|, \\ T_k^0(t) &= |\{\tau \leq t : A_\tau = k; Y_\tau = 0\}|. \end{aligned}$$

In words, $T_k^1(t)$ is the number of times up to and including time t that we have chosen action k and the output of arm k was 1 and similarly $T_k^0(t)$ is the number of times up to and including time t that we have choses action k and the output of the arm k was 0.

The goal is to find the arm with the highest parameter, i.e., the goal is to determine

$$k^* = \text{argmax}_k \theta_k.$$

In the Bayesian approach we proceed as follows. At time time t:

1. Compute for each arm k the distribution $p(\theta_k(t) | T_k^1(t-1), T_k^0(t-1))$.
2. Generate samples of these parameters according to their distributions.
3. Pick the arm j with the largest sample.
4. Observe the output of the j-th arm, call it $Y_j(t)$, and update the counters T_j^1 and T_j^0 accordingly.

Show that this algorithm "works" in the sense that eventually it will pick the best arm. More precisely, show the following two claims.

1. [10pts] Show that $p(\theta_k(t) | T_k^1(t-1), T_k^0(t-1))$ is a Beta distributed and determine α and β.
2. [10pts] Show that as t tends to infinity the probability that we choose the correct arm tends to 1. [HINT: To simplify your life, you can assume that for every arm k, $T_k^1(t-1) + T_k^0(t-1) \overset{t \to \infty}{\to} \infty$.]

NOTE: Recall that the density of the Beta distribution on $[0,1]$ with parameters α and β is equal to

$$f(x; \alpha, \beta) = \text{constant } x^{\alpha-1}(1-x)^{\beta-1}.$$

Further, the expected value of $f(x; \alpha, \beta)$ is $\frac{\alpha}{\alpha+\beta}$ and its variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Solution:

1. A quick calculation shows that $p(\theta_k(t)|T_k^1(t-1), T_k^0(t-1)) = f(x; 1 + T_k^1(t-1), 1 + T_k^0(t-1))$. Note that this is the same calculation that we did when we showed that the Beta distribution is the conjugate prior to the Binomial distribution. Explicity, and dropping the time index as well as the index indicating the arm, we have

$$\begin{aligned} p(\theta \mid T^1, T^0) &\sim p(\theta) p(T^1, T^0 \mid \theta) \\ &\sim \theta^{T^1} (1-\theta)^{T^0} \\ &= f(\theta; 1 + T^1, 1 + T^0). \end{aligned}$$

2. According to the hint and our computation above, the expected value at time t is equal to

$$\frac{1 + T_k^1(t-1)}{2 + T_k^1(t-1) + T_k^0(t-1)}.$$

By assumption $T_k^1(t-1) + T_k^0(t-1) \overset{t \to \infty}{\to} \infty$ and by the law of larger numbers $T_k^1(t-1)/(T_k^1(t-1) + T_k^0(t-1))$ and hence also $(1 + T_k^1(t-1))/(2 + T_k^1(t-1) + T_k^0(t-1))$, converges to θ_k almost surely. Therefore, our estimates for all means converge to the correct values almost surely. Further, all variances tend to 0 and hence the probability that we choose the correct arm will tend to 1 as t tends to infinity.