Statistical Signal Processing – Practice Exam 2025/26

Kapteyn Learning Community

Exam instructions:

Use vector/matrix notation exclusively.
Give motivation and/or derivations for your answers when asked to do so.
If you do not know how to solve a question mathematically, describe what you would do with words for partial credit.
Illegible handwriting will not be graded.
The exam is divided between a theoretical and a calculation part.

Question 1

a) While designing an estimator $\hat{\theta}_N$, there are certain properties we desire in that estimator; what are they? Sometimes not all properties are achievable, what should a good estimator *at least* have?

b) Sometimes we cannot use MLE and MVU, and we have to resort to use the least squares estimation. When is it appropriate to use least squares and what is the most important feature of it compared to MLE and MVU

c) Newton-Raphson method is used to find the MLE by iterating

$$\theta_{k+1} = \theta_k - \left(\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2} \right)^{-1} \left. \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta} \right|_{\theta=\theta_k}.$$

What is important to keep in mind when using Newton-Raphson?

Question 2

a) In Bayesian approach we maximize the probability of the parameter vector given the data:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}$$

Describe each term in the formula above.

b) Let us consider two competing models A and B with parameters $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ respectively, for some experimental data $\mathbf{x}$. To compare these two models, the posterior ratio is calculated as,

$$\frac{p(A|\mathbf{x})}{p(B|\mathbf{x})} = \frac{p(\mathbf{x}|A)}{p(\mathbf{x}|B)} \frac{p(A)}{p(B)},$$

from which embodies the Occam's razor. Describe what the Occam's razor and provide the formula for it.

c) Any real, symmetric $M \times N$ matrix $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^\top.$$

The main advantage to decomposing $\mathbf{A}$ using SVD is to find the pseduoinverse of a matrix (more specifically the Moore–Penrose pseudoinverse), denoted $\mathbf{A}^+$. What is a pseudomatrix, how is it defined and what are its properties?

Question 3

You choose a resistor from the pile of resistors that are meant to have the same $R = 100\,\Omega$ resistance within manufacturing tolerances and use a multimeter to measure its resistance. Unfortunately, the physics lab multimeters are not very good and result in a measurement error that can be assumed to be drawn from $\mathcal{N} \sim (0, 1\,\Omega)$ for each measurement, so you have to take N measurements of each resistor. You also know that the manufacturing tolerances are drawn from $\epsilon \sim \mathcal{N}(0, 0.011\,\Omega)$.

a) How many measurements of a single resistor do you need to make to ensure that your estimate of the resistance R is correct to $0.1\,\Omega$ on average?

b) How many measurements do you need if you didn't have any prior knowledge of the manufacturing tolerances?

You may find the following identity useful: given two Gaussians $\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$,

$$\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1) \times \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2) = Z\,\mathcal{N}(\mathbf{m}_c, \mathbf{\Sigma}_c),$$

where

$$m_c = (\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 + \mathbf{\Sigma}_2^{-1}\mathbf{m}_2)$$

and

$$\mathbf{\Sigma}_c = (\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}.$$

Hint: write down a data model for $\mathbf{x}$, then its likelihood function in terms of the resistance R. What is the prior in this problem? Then think about the form of the posterior given these two probabilities — what is its variance? How does this connect to the information you are given?

Question 4

If there is some constant signal embedded in white Gaussian noise with unknown variance σ, $\mathbf{x} = A + \mathbf{w}$, with

$$\mathbf{x} = (x_0, \ldots, x_{N-1})^\top, \qquad \mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^N.$$

The PDF becomes

$$p(\mathbf{x}|A) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - A\mathbf{1})^\top(\mathbf{x} - A\mathbf{1})\right).$$

a) What does the Fisher information matrix encode?

b) Find the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = [A, \sigma^2]^\top$.

c) From $\mathbf{I}(\boldsymbol{\theta})$, read off the CRLB of $\text{var}\{\hat{A}\}$ and $\text{var}\{\hat{\sigma^2}\}$.

d) Let's assume we want $\hat{\psi}(A)$ instead of $\hat{A}$, with $\psi(A) = 2A^4 + A$, what would the CRLB of $\text{var}\{\hat{\psi}(A)\}$ be?

Question 5

a) Assume two variables $\mathbf{x} = [x_1, x_2]^\top$ with zero mean and the following covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2/2 \\ \sigma^2/2 & \sigma^2 \end{bmatrix}.$$

The variables follow a bivariate Gaussian distribution:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^2 \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}\right),$$

derive $p(x_1|\boldsymbol{\theta})$.

b) Let

$$\mathbf{x} = (x_1, \ldots, x_n)^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where the covariance matrix is diagonal,

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2).$$

Derive $p(x_1|\boldsymbol{\theta})$.

You may find the following identities useful:

$$\text{For} \quad \mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$\int_{-\infty}^{\infty} e^{-a(x-b)^2} dx = \sqrt{\frac{\pi}{a}}, \quad a > 0.$$

$$x_2^2 - x_1 x_2 = \left(x_2 - \frac{x_1}{2}\right)^2 - \frac{x_1^2}{4}.$$

Question 6

Let $\mathbf{x}$ be independent and identically distributed measurements which are integers drawn from the Poisson distribution:

$$p(\mathbf{x}|\theta) = \frac{\theta^{\mathbf{1}^\top \mathbf{x}}}{\mathbf{x}!} e^{-n\theta}$$

where θ is a positive real number that is equal to the mean and variance of the distribution.

a) Write the expression for the likelihood function, $p(\mathbf{x}|\theta)$ and calculate the MLE for θ.

b) Show that the MLE is unbiased.

Now assume that the parameter θ is a random variable with a prior of an exponential distribution:

$$p(\theta) = \begin{cases} \lambda e^{-\lambda\theta} & \text{if } \theta > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where λ is a given constant.

c) What is the log-posterior PDF $\ln p(\theta|\mathbf{x}, \lambda)$?

d) Find the MAP estimator of θ and discuss its difference with respect to the MLE solution.

Question 7

Consider the linear model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w},$$

where $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{H} \in \mathbb{R}^{N \times p}$, and $\boldsymbol{\theta} \in \mathbb{R}^p$. Assume that $\boldsymbol{\theta}$ is subject to $r < p$ *independent linear constraints*

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b},$$

where $\mathbf{A} \in \mathbb{R}^{r \times p}$ and $\mathbf{b} \in \mathbb{R}^r$.

a) Formulate the constrained least squares problem and write down the corresponding Lagrangian cost function.

b) Compute the gradient of the Lagrangian with respect to $\boldsymbol{\theta}$ and derive the expression for the constrained least squares estimator $\hat{\boldsymbol{\theta}}_c$ in terms of the Lagrange multipliers.

c) Use the constraint $\mathbf{A}\hat{\boldsymbol{\theta}}_c = \mathbf{b}$ to solve for the Lagrange multipliers and show that the constrained estimator can be written as

$$\hat{\boldsymbol{\theta}}_c = \hat{\boldsymbol{\theta}} - (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{A}^\top\left[\mathbf{A}(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{A}^\top\right]^{-1}\left(\mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{b}\right)$$

where the unconstrained least squares estimator is

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{x}.$$

You will need following matrix differential identities are used in the derivation:

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left(\boldsymbol{\theta}^\top\mathbf{M}\boldsymbol{\theta}\right) = (\mathbf{M} + \mathbf{M}^\top)\boldsymbol{\theta},$$

and if $\mathbf{M} = \mathbf{M}^\top$,

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left(\boldsymbol{\theta}^\top\mathbf{M}\boldsymbol{\theta}\right) = 2\mathbf{M}\boldsymbol{\theta}.$$

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left(\mathbf{c}^\top\boldsymbol{\theta}\right) = \mathbf{c}.$$

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) = \mathbf{x}^\top\mathbf{x} - 2\boldsymbol{\theta}^\top\mathbf{H}^\top\mathbf{x} + \boldsymbol{\theta}^\top\mathbf{H}^\top\mathbf{H}\boldsymbol{\theta}.$$

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left[(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})\right] = -2\mathbf{H}^\top\mathbf{x} + 2\mathbf{H}^\top\mathbf{H}\boldsymbol{\theta}.$$

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left(\boldsymbol{\lambda}^\top\mathbf{A}\boldsymbol{\theta}\right) = \mathbf{A}^\top\boldsymbol{\lambda}.$$

Answers

Question 1

a) The estimator is to be **unbiased** over some range $a < \theta < b$ with $E\{\hat{\theta}_N\} = \theta$, **consistent** that when more data is collected the estimator approaches the true value, and **efficient** with respect to another estimator $\hat{\theta}_H$ with $\text{var}\{\hat{\theta}_N\} < \text{var}\{\hat{\theta}_H\}$. A good estimator should at least be unbiased and show the *minimum possible variance* between the estimator and the true values(s).

b) Least squares does not look for an optimal or nearly optimal estimator, but it has to be used in overdetermined systems; where there are many more data points than unknowns. An important feature of least squares is that *no* probabilistic assumptions about the data are made and only a single model is assumed. Due to this, assumptions on the system when using least squares are very limited making the interpretations of the system very limited.

c) When using NR the iteration may not converge, especially when the second derivative of the log-likelihood is small making the correction term to fluctuate wildly at each iteration. Even if the iteration converges, it may not be the global maximum. To combat this it is best to try several starting points and choose the one that yields the maximum of the log-likelihood function. Generally if the initial guess is close to the global maximum, NR should find it.

Question 2

a) $p(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood function, $p(\boldsymbol{\theta})$ is the prior probability, $p(\mathbf{x})$ the evidence, and $p(\boldsymbol{\theta}|\mathbf{x})$ the posterior probability.

b) Occam's razor is given by the *ratio of evidences*,

$$\frac{p(\mathbf{x}|A)}{p(\mathbf{x}|B)}.$$

According to the results of Occam's razor, a theoretical model that includes more **new assumptions** is **less likely** to have produced your experimental data. Thus, a simpler model is generally preferred since **less new assumptions** have been made, subsequently the **probability is higher** that this model produced the data.

c) A pesudoinverse of a matrix is a generalization of the inverse for singular or non-square matrices. For a matrix $\mathbf{A}$ we define its pseudoinverse as

$$\mathbf{A}^+ = \mathbf{V}\mathbf{W}^+\mathbf{U}^\top = \mathbf{V}\,\text{diag}(w_0^{-1}, ..., w_{r-1}^{-1}, 0, ..., 0)\mathbf{U}^\top,$$

with r being the matrix rank. The pseudoinverse satisfies the following:

$$\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}, \quad \mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+, \quad (\mathbf{A}\mathbf{A}^+)^\top = \mathbf{A}\mathbf{A}^+, \quad (\mathbf{A}^+\mathbf{A})^\top = \mathbf{A}^+\mathbf{A}$$

Question 3

a) Let's start by realizing that the data model is the simple constant-plus-noise model we used as an example throughout the course:

$$\mathbf{x} = R + \mathbf{w},$$

where $\mathbf{w} \sim \mathcal{N}(0, 0.1)$ and $p(x[n]|R) \sim \mathcal{N}(100, 0.011)$ for an individual measurement $x[n]$. The likelihood function therefore (up to a normalization constant) looks like

$$p(\mathbf{x}|R) \propto \exp\left[-\frac{\sum_{n=0}^{N-1}(x[n] - R)^2}{2\sigma^2}\right] \sim \mathcal{N}(\bar{x}, \sigma^2/N),$$

where $\sigma^2 = 1^2 = 1$. The prior is the manufacturing tolerances,

$$p(R) \propto \exp\left[-\frac{(R-\mu_R)^2}{2\sigma_R^2}\right] \sim \mathcal{N}(\mu_r, \sigma_R^2),$$

where μ_R is the true mean value of the resistance and $\sigma_R^2 = 0.011$ is the quoted tolerance. These are obviously both Gaussians. The posterior, by Bayes' theorem, is then just

$$p(R|\mathbf{x}) \propto p(\mathbf{x}|R)p(R) \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2).$$

What we really care about here is $\tilde{\sigma}^2$, which can be found from the identity below as

$$\tilde{\sigma}^2 = [(\sigma^2/N)^{-1} + (\sigma_R^2)^{-1}]^{-1} = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_R^2}}.$$

We want to solve this for N, as we know $\tilde{\sigma}^2 = 0.1^2 = 0.01$, $\sigma^2 = 1$, and $\sigma_R^2 = 0.011$:

$$\begin{aligned}
\frac{1}{\tilde{\sigma}^2} &= \frac{N}{\sigma^2} + \frac{1}{\sigma_R^2} \\
N &= \sigma^2\left(\frac{1}{\tilde{\sigma}^2} - \frac{1}{\sigma_R^2}\right) = 1 \times \left(\frac{1}{0.01} - \frac{1}{0.011}\right) \\
&= 100 - 90.91 = 9.09,
\end{aligned}$$

so you need to measure it at least 9 (ok, 10) times *if you have prior knowledge.*

b) If you *don't* have prior knowledge, then $\sigma_R^2 \to \infty$ and

$$N = \frac{\sigma^2}{\tilde{\sigma}^2} = 1/0.01 = 100,$$

so you need to measure it 100 times – a lot more work!

Question 4

a) FIM quantifies how much information the observed data $\mathbf{x}$ carries about the unknown parameter vector $\boldsymbol{\theta}$. It is defined as

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\Big[(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}|\boldsymbol{\theta}))\,(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}|\boldsymbol{\theta}))^\top\Big] = -\mathbb{E}_{\boldsymbol{\theta}}\big[\nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{x}|\boldsymbol{\theta})\big]\,.$$

Its inverse provides a lower bound (the Cramér–Rao lower bound, CRLB) on the covariance matrix of any unbiased estimator of $\boldsymbol{\theta}$.

b) The log-likelihood is

$$\ln p(A, \sigma^2|\boldsymbol{\theta}) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{x} - A\mathbf{1})^\top(\mathbf{x} - A\mathbf{1}).$$

For A,

$$\frac{\partial^2 \ln p(A, \sigma^2|\boldsymbol{\theta})}{\partial A^2} = -\frac{1}{\sigma^2}\mathbf{1}^\top\mathbf{1} = -\frac{N}{\sigma^2}, \qquad I_{AA} = \frac{N}{\sigma^2}.$$

For σ^2,

$$\frac{\partial^2 \ln p(A, \sigma^2|\boldsymbol{\theta})}{\partial(\sigma^2)^2} = \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{x} - A\mathbf{1})^\top(\mathbf{x} - A\mathbf{1}),$$

and compute

$$\mathbb{E}\big[(\mathbf{x} - A\mathbf{1})^\top(\mathbf{x} - A\mathbf{1})\big]\,.$$

Since

$$\mathbf{x} = A\mathbf{1} + \mathbf{w}, \qquad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N),$$

we have

$$\mathbf{x} - A\mathbf{1} = \mathbf{w}.$$

Hence

$$(\mathbf{x} - A\mathbf{1})^\top (\mathbf{x} - A\mathbf{1}) = \mathbf{w}^\top \mathbf{w} = \sum_{i=0}^{N-1} w_i^2.$$

Taking expectations,

$$\mathbb{E}\big[(\mathbf{x} - A\mathbf{1})^\top (\mathbf{x} - A\mathbf{1})\big] = \sum_{i=0}^{N-1} \mathbb{E}[w_i^2].$$

Each component satisfies

$$w_i \sim \mathcal{N}(0, \sigma^2), \qquad \mathbb{E}[w_i^2] = \mathrm{var}(w_i) = \sigma^2,$$

and the components are independent. Therefore,

$$\mathbb{E}\big[(\mathbf{x} - A\mathbf{1})^\top (\mathbf{x} - A\mathbf{1})\big] = \sum_{i=0}^{N-1} \sigma^2 = N\sigma^2.$$

So,

$$I_{\sigma^2\sigma^2} = \frac{N}{2\sigma^4}.$$

The cross term vanishes,

$$I_{A\sigma^2} = 0.$$

Hence

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{N}{\sigma^2} & 0 \\ 0 & \dfrac{N}{2\sigma^4} \end{pmatrix}.$$

c) Since $\mathbf{I}(\boldsymbol{\theta})$ is diagonal,

$$\mathrm{var}\{\hat{A}\} \geq \frac{\sigma^2}{N}, \qquad \mathrm{var}\{\widehat{\sigma^2}\} \geq \frac{2\sigma^4}{N}.$$

d) For a scalar function $\psi(A)$,

$$\mathrm{var}\{\hat{\psi}(A)\} \geq \left(\frac{\partial \psi}{\partial A}\right)^2 \mathrm{var}\{\hat{A}\}.$$

Since

$$\psi'(A) = 8A^3 + 1,$$

the CRLB is

$$\mathrm{var}\{\hat{\psi}(A)\} \geq (8A^3 + 1)^2 \frac{\sigma^2}{N}.$$

Question 5

a)

$$\det(\boldsymbol{\Sigma}) = \sigma^4 - \left(\frac{\sigma^2}{2}\right)^2 = \frac{3}{4}\sigma^4.$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\det(\boldsymbol{\Sigma})} \begin{bmatrix} \sigma^2 & -\sigma^2/2 \\ -\sigma^2/2 & \sigma^2 \end{bmatrix} = \frac{4}{3\sigma^2} \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}.$$

The normalization constant is

$$\sqrt{(2\pi)^2 \det(\boldsymbol{\Sigma})} = 2\pi\sigma^2\sqrt{\frac{3}{4}} = \pi\sigma^2\sqrt{3}.$$

Hence

$$p(x_1, x_2) = \frac{1}{\pi\sigma^2\sqrt{3}} \exp\left[-\frac{2}{3\sigma^2}\left(x_1^2 - x_1x_2 + x_2^2\right)\right].$$

$$x_2^2 - x_1x_2 = \left(x_2 - \frac{x_1}{2}\right)^2 - \frac{x_1^2}{4}.$$

Thus the exponent becomes

$$-\frac{2}{3\sigma^2}\left(x_2 - \frac{x_1}{2}\right)^2 - \frac{x_1^2}{2\sigma^2}.$$

$$p(x_1) = \frac{1}{\pi\sigma^2\sqrt{3}} \exp\left(-\frac{x_1^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} \exp\left[-\frac{2}{3\sigma^2}\left(x_2 - \frac{x_1}{2}\right)^2\right] dx_2.$$

$$\int_{-\infty}^{\infty} \exp\left[-\frac{2}{3\sigma^2}\left(x_2 - \frac{x_1}{2}\right)^2\right] dx_2 = \sqrt{\frac{3\pi\sigma^2}{2}}.$$

Finally,

$$\begin{aligned} p(x_1) &= \frac{1}{\pi\sigma^2\sqrt{3}}\sqrt{\frac{3\pi\sigma^2}{2}} \exp\left(-\frac{x_1^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x_1^2}{2\sigma^2}\right) \\ &= \mathcal{N}(0, \sigma^2) \end{aligned}$$

b)

$$\boldsymbol{\Sigma}^{-1} = \operatorname{diag}\left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2}\right), \qquad \det(\boldsymbol{\Sigma}) = \prod_{i=1}^{n} \sigma_i^2.$$

The joint probability density function is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2}\right).$$

The marginal distribution of x_1 is obtained by integrating out $x_2, \ldots, x_n$:

$$p(x_1) = \int_{\mathbb{R}^{n-1}} p(x_1, x_2, \ldots, x_n)\, dx_2 \cdots dx_n.$$

Substituting the joint density,

$$p(x_1) = \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i} \exp\left(-\frac{x_1^2}{2\sigma_1^2}\right) \prod_{i=2}^{n} \int_{-\infty}^{\infty} \exp\left(-\frac{x_i^2}{2\sigma_i^2}\right) dx_i.$$

Using

$$\int_{-\infty}^{\infty} \exp\left(-\frac{x_i^2}{2\sigma_i^2}\right) dx_i = \sqrt{2\pi}\,\sigma_i,$$

we obtain

$$p(x_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left(-\frac{x_1^2}{2\sigma_1^2}\right).$$

Question 6

a) The likelihood function is

$$L(\theta|\mathbf{x}) = p(\mathbf{x}|\theta) = \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}\, e^{-n\theta}.$$

The log-likelihood is

$$\ln L(\theta|\mathbf{x}) = \left(\sum_{i=1}^n x_i\right) \ln\theta - n\theta - \sum_{i=1}^n \ln(x_i!).$$

Differentiating with respect to θ and setting to zero:

$$\frac{\partial \ln L(\theta|\mathbf{x})}{\partial\theta} = \frac{\sum_{i=1}^n x_i}{\theta} - n = 0 \quad \Rightarrow \quad \hat{\theta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}.$$

b)

$$\mathbb{E}[\hat{\theta}_{\text{MLE}}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n x_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_i] = \frac{1}{n}\cdot n\theta = \theta.$$

Hence, the MLE is unbiased.

c) The posterior is proportional to likelihood $\times$ prior:

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)\, p(\theta) = \theta^{\sum_{i=1}^n x_i} e^{-n\theta} \cdot \lambda e^{-\lambda\theta} \propto \theta^{\sum_{i=1}^n x_i} e^{-(n+\lambda)\theta}.$$

Hence, the log-posterior is

$$\ln p(\theta|\mathbf{x},\lambda) = \left(\sum_{i=1}^n x_i\right)\ln\theta - (n+\lambda)\theta + \text{const.}$$

d) To find the MAP, differentiate the log-posterior with respect to θ and set to zero:

$$\frac{\partial}{\partial\theta}\ln p(\theta|\mathbf{x},\lambda) = \frac{\sum_{i=1}^n x_i}{\theta} - (n+\lambda) = 0.$$

Solving for θ:

$$\hat{\theta}_{\text{MAP}} = \frac{\sum_{i=1}^n x_i}{n+\lambda} = \frac{n}{n+\lambda}\bar{x}.$$

The MAP estimator shrinks the MLE toward zero due to the exponential prior. As $n \to \infty$, the effect of the prior diminishes and $\hat{\theta}_{\text{MAP}} \to \hat{\theta}_{\text{MLE}}$.

Question 7

We wish to solve the constrained least squares problem

$$\min_{\boldsymbol{\theta}} \ (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \quad \text{subject to} \quad \mathbf{A}\boldsymbol{\theta} = \mathbf{b}.$$

a) Introduce a vector of Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^r$. The Lagrangian cost function is

$$J_c(\boldsymbol{\theta}, \boldsymbol{\lambda}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) + \boldsymbol{\lambda}^\top (\mathbf{A}\boldsymbol{\theta} - \mathbf{b}).$$

Using standard algebra,

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) = \mathbf{x}^\top \mathbf{x} - 2\boldsymbol{\theta}^\top \mathbf{H}^\top \mathbf{x} + \boldsymbol{\theta}^\top \mathbf{H}^\top \mathbf{H}\boldsymbol{\theta}.$$

Substituting into J_c,

$$J_c = \mathbf{x}^\top \mathbf{x} - 2\boldsymbol{\theta}^\top \mathbf{H}^\top \mathbf{x} + \boldsymbol{\theta}^\top \mathbf{H}^\top \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\lambda}^\top \mathbf{A}\boldsymbol{\theta} - \boldsymbol{\lambda}^\top \mathbf{b}.$$

b) We differentiate term by term.

- $\frac{\partial}{\partial \boldsymbol{\theta}}(\mathbf{x}^\top \mathbf{x}) = 0$ since it does not depend on $\boldsymbol{\theta}$.
- $\frac{\partial}{\partial \boldsymbol{\theta}}(-2\boldsymbol{\theta}^\top \mathbf{H}^\top \mathbf{x}) = -2\mathbf{H}^\top \mathbf{x}$.
- Since $\mathbf{H}^\top \mathbf{H}$ is symmetric,

$$\frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^\top \mathbf{H}^\top \mathbf{H}\boldsymbol{\theta}) = 2\mathbf{H}^\top \mathbf{H}\boldsymbol{\theta}.$$

- $\frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\lambda}^\top \mathbf{A}\boldsymbol{\theta}) = \mathbf{A}^\top \boldsymbol{\lambda}$.

Hence,

$$\frac{\partial J_c}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^\top \mathbf{x} + 2\mathbf{H}^\top \mathbf{H}\boldsymbol{\theta} + \mathbf{A}^\top \boldsymbol{\lambda}.$$

Setting the gradient equal to zero,

$$2\mathbf{H}^\top \mathbf{H}\boldsymbol{\theta} = 2\mathbf{H}^\top \mathbf{x} - \mathbf{A}^\top \boldsymbol{\lambda}.$$

Assuming $\mathbf{H}^\top \mathbf{H}$ is invertible, we obtain

$$\boldsymbol{\theta} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x} - \frac{1}{2}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top \boldsymbol{\lambda}.$$

Define the unconstrained least squares estimator

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x}.$$

Then

$$\hat{\boldsymbol{\theta}}_c = \hat{\boldsymbol{\theta}} - \frac{1}{2}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top \boldsymbol{\lambda}.$$

c) Impose $\mathbf{A}\hat{\boldsymbol{\theta}}_c = \mathbf{b}$:

$$\mathbf{A}\hat{\boldsymbol{\theta}} - \frac{1}{2}\mathbf{A}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top \boldsymbol{\lambda} = \mathbf{b}.$$

Rearranging,

$$\frac{1}{2}\mathbf{A}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top \boldsymbol{\lambda} = \mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{b}.$$

Since the constraints are independent, $\mathbf{A}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top$ is invertible, and thus

$$\frac{\boldsymbol{\lambda}}{2} = \left[\mathbf{A}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top\right]^{-1} (\mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{b}).$$

Substituting this result into the expression for $\hat{\boldsymbol{\theta}}_c$,

$$\hat{\boldsymbol{\theta}}_c = \hat{\boldsymbol{\theta}} - (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top \left[\mathbf{A}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top\right]^{-1} (\mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{b}).$$

The constrained estimator is obtained by projecting the unconstrained least squares solution $\hat{\boldsymbol{\theta}}$ onto the affine subspace defined by $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$, with respect to the metric induced by $\mathbf{H}^\top \mathbf{H}$.