Solutions to Homework Set Two
ECE 271A
Electrical and Computer Engineering
University of California San Diego

1.

a) Start with a vector $\mathbf{a}$ and

$$\nabla_\mu \mathbf{a}^T \mu = \nabla_\mu \sum_i a_i \mu_i.$$

It should not be too difficult to see that this is just

$$\nabla_\mu \mathbf{a}^T \mu = \mathbf{a}.$$

Next, consider a matrix $\mathbf{C}$ and

$$\nabla_\mu \mathbf{C}\mu = \nabla_\mu \begin{bmatrix} \mathbf{c}_1\mu \\ \vdots \\ \mathbf{c}_n\mu \end{bmatrix}$$

where $\mathbf{c}_i$ is the i^{th} row of vector. Applying the result above it should not be too hard to see that

$$\nabla_\mu \mathbf{C}\mu = \mathbf{C}^T.$$

It follows that

$$\begin{aligned} \nabla_\mu (\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) &= \nabla_\mu [(\mathbf{x}-\mu)^T \times \mathbf{a}|_{\mathbf{a}=\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}] + \nabla_\mu [\mathbf{a}^T|_{\mathbf{a}=\mathbf{x}-\mu} \times \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)] \\ &= \nabla_\mu [(\mathbf{x}-\mu)] \times \mathbf{a}|_{\mathbf{a}=\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)} + \nabla_\mu [\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)] \times \mathbf{a}|_{\mathbf{a}=\mathbf{x}-\mu} \\ &= -\mathbf{I} \times \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) - \mathbf{\Sigma}^{-1} \times (\mathbf{x}-\mu)] \\ &= -2\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \end{aligned}$$

b) Consider the line l that passes through μ and whose direction is given by the unit vector $\mathbf{l}$. Then

$$l = \{\mathbf{x} | \mathbf{x} = \mu + \alpha \mathbf{l}, \alpha \in R\} \tag{1}$$

and, $\forall \mathbf{p} \in l, \mathbf{p} - \mu = \alpha \mathbf{l}$, for some $\alpha \in R$. It follows that, $\forall \mathbf{p} \in l$,

$$\begin{aligned} \nabla_r^2 &= 2\mathbf{\Sigma}^{-1}(\mathbf{p}-\mu) \\ &= 2\alpha \mathbf{\Sigma}^{-1}\mathbf{l} \end{aligned}$$

i.e. the gradient always points in the direction $\mathbf{\Sigma}^{-1}\mathbf{l}$. This direction is not parallel to the line unless $\mathbf{\Sigma} = \mathbf{I}$.

c) If $\mathbf{p}$ is a point in the line from μ_1 to μ_2 then

$$\exists \alpha \in [0,1] \text{ such that } \mathbf{p} = \alpha\mu_1 + (1-\alpha)\mu_2.$$

It follows that

$$\begin{aligned} \nabla r_1^2 &= 2\mathbf{\Sigma}^{-1}[(\alpha-1)\mu_1 + (1-\alpha)\mu_2] \\ &= 2(1-\alpha)\mathbf{\Sigma}^{-1}[\mu_2 - \mu_1] \end{aligned}$$

and

$$\begin{aligned} \nabla r_2^2 &= 2\boldsymbol{\Sigma}^{-1}[\alpha\mu_1 - \alpha\mu_2] \\ &= 2\alpha\boldsymbol{\Sigma}^{-1}[\mu_1 - \mu_2] \end{aligned}$$

Hence, the gradients point in opposite directions.

d) Let $\mathbf{p}$ be the point where the optimal separating hyperplane cuts the line from μ_1 to μ_2. From **c)** we know that there is an $\alpha \in [0, 1]$ such that, at $\mathbf{p}$,

$$\nabla r_1^2 = -2(1-\alpha)\boldsymbol{\Sigma}^{-1}[\mu_1 - \mu_2] \quad (2)$$

$$\nabla r_2^2 = 2\alpha\boldsymbol{\Sigma}^{-1}[\mu_1 - \mu_2] \quad (3)$$

Recalling that the gradient at a point is perpendicular to the constant probability density ellipsoid at that point, it follows that, at $\mathbf{p}$, this perpendicular direction is $\boldsymbol{\Sigma}^{-1}[\mu_1 - \mu_2]$ for one of the Gaussians and $-\boldsymbol{\Sigma}^{-1}[\mu_1 - \mu_2]$ for the other. Given that the optimal hyperplane has normal vector $\boldsymbol{\Sigma}^{-1}[\mu_1 - \mu_2]$ (equation 63 in DHS), its normal is obviously parallel to the the two perpendicular directions above. It follows that the hyperplane is tangent to the hyperellipsoids.

e) False. From equation 59 in DHS we know that, when the prior is uniform, the Bayes decision boundary consists of the points of equal Mahalanobis distance from the two *class means*. However, there is no guarantee that the *sample means* will be equal to the true *class means*.

2.

a) The key property of this problem is that the π_j are probabilities and, therefore, we need to solve

$$\Pi^* = \arg \max_{\pi_1,\ldots,\pi_N} \log P_{C_1,\ldots,C_N}(c_1,\ldots,c_N)$$

subject to the constraint

$$\sum_j \pi_j = 1.$$

There are two ways of doing this

i) A Lagrangian formulation, where instead of maximizing only the likelihood, we maximize the Lagrangian

$$\begin{aligned} L &= \log P_{C_1,\ldots,C_N}(c_1,\ldots,c_N) + \lambda(\sum_j \pi_j - 1) \\ &= \sum_j c_j \log \pi_j + \log(n!) - \sum_k \log(c_k!) + \lambda(\sum_j \pi_j - 1) \end{aligned}$$

Taking the gradient with respect to both the π_j and λ and setting to zero, we obtain

$$\begin{aligned} \frac{\partial L}{\partial \pi_j} &= \frac{c_j}{\pi_j} + \lambda = 0 \\ \frac{\partial L}{\partial \lambda} &= \sum_j \pi_j - 1 = 0 \end{aligned}$$

Multiplying the first equation by π_j on both sides, summing over j, and using the second equation

$$\lambda = -\sum_j c_j = -n$$

from which it follows (still from the first equation) that

$$\pi_j^* = \frac{c_j}{n}.$$

Computing the Hessian

$$\frac{\partial^2 L}{\partial \pi_j^2} = -\frac{c_j}{\pi_j^2} \qquad\qquad \frac{\partial^2 L}{\partial \pi_j \pi_k} = 0$$

it follows that $\nabla^2_{\mathbf{\Pi}} L = -diag(c_j/\pi_j^2)$. This is clearly negative definite since, for a diagonal matrix, the eigenvalues are just the entries in the main diagonal. Hence we have a maximum.

Note: to verify that we have a maximum, $\nabla^2_{\mathbf{\Pi}} L$ actually only needs to be negative definite on the *subspace of first order feasible variations*, i.e. the subspace

$$V(\mathbf{\Pi}) = \left\{ \mathbf{y} | \left[\nabla_{\mathbf{\Pi}} \left(\sum_j \pi_j - 1 \right) \right]^T \mathbf{y} = 0 \right\}$$

In this problem, this did not make a difference since the matrix is negative definite everywhere, i.e. the function is concave. But, in the general case, you need to consider what happens in this subspace only. See *Nonlinear Programming, D. Bertsekas, Athena Scientific, 1995, pp 255.*

ii) The second way of solving this problem is to include the constraint directly in the cost function, i.e. to solve

$$\Pi^* = \arg \max_{\pi_1,\ldots,\pi_{N-1}} \log f(\pi_1, \ldots, \pi_{N-1})$$

where

$$f(\pi_1, \ldots, \pi_{N-1}) = \frac{n!}{\prod_k c_k!} \prod_{j=1}^{N-1} \pi_j^{c_j} \times \left(1 - \sum_{i=1}^{N-1} \pi_i\right)^{c_N}$$

and

$$\log f(\pi_1, \ldots, \pi_{N-1}) = \log(n!) - \sum_k \log c_k! + \sum_{j=1}^{N-1} c_j \log \pi_j + c_N \log \left(1 - \sum_{i=1}^{N-1} \pi_i\right).$$

Taking the gradient with respect to $\pi_j, j = 1, \ldots, N-1$ and setting to zero

$$\frac{\partial \log f}{\partial \pi_j} = \frac{c_j}{\pi_j} - \frac{c_N}{\left(1 - \sum_{i=1}^{N-1} \pi_i\right)} = 0$$

it follows that

$$\pi_j = \frac{c_j}{c_N} \left(1 - \sum_{i=1}^{N-1} \pi_i\right)$$

and summing over all $j = 1, \ldots, N-1$

$$\begin{aligned}
\sum_{i=1}^{N-1} \pi_i &= \left(1 - \sum_{i=1}^{N-1} \pi_i\right) \frac{1}{c_N} \sum_{i=1}^{N-1} c_i \\
&= \left(1 - \sum_{i=1}^{N-1} \pi_i\right) \frac{n - c_N}{c_N} \\
\left(1 + \frac{c_N}{n - c_N}\right) \sum_{i=1}^{N-1} \pi_i &= 1 \\
\sum_{i=1}^{N-1} \pi_i &= \frac{n - c_N}{n}
\end{aligned}$$

Hence,

$$1 - \sum_{i=1}^{N-1} \pi_i = \frac{c_N}{n}$$

and

$$\pi_j^* = \frac{c_j}{n}$$

Next we compute the Hessian,

$$\frac{\partial^2 \log f}{\partial \pi_j \pi_k} = \begin{cases} -\frac{c_N}{\left(1 - \sum_{i=1}^{N-1} \pi_i\right)^2}, & \text{if } k = j \\ -\frac{c_j}{\pi_j} - \frac{c_N}{\left(1 - \sum_{i=1}^{N-1} \pi_i\right)^2}, & \text{if } k \neq j \end{cases}$$

which can therefore be written as

$$\nabla^2_{\mathbf{\Pi}} f = -\left(\mathbf{\Lambda} + \frac{c_N}{\left(1 - \sum_{i=1}^{N-1} \pi_i\right)^2} \mathbf{1}\right)$$

where $\mathbf{\Lambda}$ is a diagonal matrix with entries c_j/π_j and $\mathbf{1}$ is the matrix of all ones. $\mathbf{\Lambda}$ is clearly positive definite (all its eigenvalues, which are just the diagonal entries, are positive). $\mathbf{1}$ has only one non-zero eigenvalue which is equal to its trace (the sum of the eigenvalues of matrix is equal to its trace), i.e. equal to $N-1$, and is therefore positive semidefinite. It follows that $\nabla^2_{\mathbf{\Pi}} f$ is negative definite and we have a maximum.

b) The mean of the estimator $\hat{\pi}_i$ is

$$E[\hat{\pi}_i] = E\Big[\frac{C_i}{n}\Big] = \frac{1}{n}E[C_i] = \frac{1}{n}\cdot n\pi_i = \pi_i.$$

Hence it's an unbiased estimator. Its variance is

$$Var[\hat{\pi}_i] = Var\Big[\frac{C_i}{n}\Big] = \frac{1}{n^2}Var[C_i] = \frac{1}{n^2}\cdot n\pi_i(1-\pi_i) = \frac{\pi_i(1-\pi_i)}{n}.$$

Hence $\hat{\pi}_i$ will converge to π_i as the number of obervations n goes to infinity, which makes $\hat{\pi}_i$ a desiable esimator.

3. Problem 3.2.8 in DHS.

The goal of this problem is to emphasize the point, discussed in class, that the Bayesian decision rule is optimal only insofar as the probability models on which it based are correct.

a) In the limit of infinite data, the sample mean and variance converge to the true class mean and variance. Hence

$$\begin{aligned}
\mu_1 &= (1-k) - kX \\
\sigma_1^2 &= (1-k)(1-\mu_1)^2 + k(X+\mu_1)^2 \\
\mu_2 &= -(1-k) + kX = -\mu_1 \\
\sigma_2^2 &= (1-k)(1+\mu_2)^2 + k(X-\mu_2)^2 \\
&= (1-k)(1-1\mu_1)^2 + k(X+\mu_1)^2 = \sigma_1^2
\end{aligned}$$

and we have two classes with the same variance. This means that, from (2.58) in DHS, the class boundary is the average of the means, which is $x = 0$.

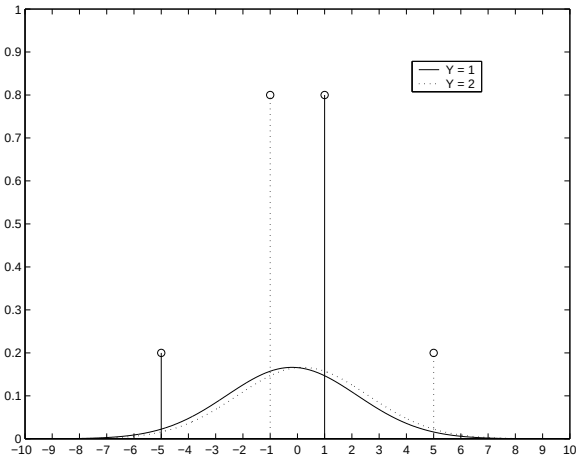
b) There is a switch of the means if

$$\begin{aligned}
\hat{\mu}_1 &\geq \hat{\mu}_2 \\
(1-k) - kX &\geq -(1-k) + kX \\
2(1-k) &\geq 2kX \\
X &\leq \frac{1-k}{k}
\end{aligned}$$

and, therefore, the switching point is

$$X = \frac{1-k}{k}$$

c) When $k = .2$ and $X = 5$ we have $\mu_1 = -\mu_2 = -.2$ and $\sigma_1^2 = \sigma_2^2 = 5.76$. The true distributions and Gaussian estimates are shown in the figure below.



The probability of error is

$$P_{X,Y}(g(x) \neq y) = P_{X|Y}(g(x) = 2|y = 1)P_Y(1) + P_{X|Y}(g(x) = 1|y = 2)P_Y(2)$$

$$
\begin{aligned}
&= P_{X|Y}(x \geq 0|y=1)P_Y(1) + P_{X|Y}(x \leq 0|y=2)P_Y(2) \\
&= (1-k)[P_Y(1) + P_Y(2)] \\
&= (1-k) = 0.8
\end{aligned}
$$

d)

$$\hat{\mu}_1 < 0 \Leftrightarrow X > \frac{1-k}{k}.$$

e) We have already answered this in **c)**. The probability of error is

$$P_{X,Y}(g(x) \neq y) = (1-k)$$

f) No. As long the two variances have the same value, the boundary point will be the mid-point between the means and will therefore not depend on this value. It follows that, assuming that the condition derived in **d)** continues to hold, the probability of error will also not change.

g) Under these models, as $k \to 0$ and as long as

$$X(k) > \frac{1-k}{k}$$

we have $P_{X,Y}(g(x) \neq y) \to 1$. This has nothing to do with local minima, but is due to the fact that the probability models are mismatched with the true class densities. It is a clever example of how the Bayes decision rule can give the worse possible solution (twice as worse as the 50% probability of error of random guessing!) if one is not careful in the modeling of the probabilities involved.

4. Problem 3.2.10 in DHS

a) When $\hat{\mu} = \mathbf{x}_1$

$$E_{\mathbf{X}_1,\ldots,\mathbf{X}_n}[\hat{\mu}] = E_{\mathbf{X}_1}[\mathbf{x}_1] = \mu$$

which shows that $\hat{\mu}$ is unbiased.

b) For the covariance of this estimator we have

$$var(\hat{\mu}) = E_{\mathbf{X}_1}[(\mathbf{x}_1 - \mu)(\mathbf{x}_1 - \mu)^T] = \mathbf{\Sigma}.$$

This means that the covariance is constant and independent of n, i.e. the confidence in the estimate does not improve as the sample size grows. This is not surprising since we are simply discarding all sample points after $\mathbf{x}_1$, but is the sign of a very poor estimator. The point of this problem is to show that concentrating on bias alone is not a very smart thing to do. Here we have an estimator that is clearly worse than the sample mean (whose covariance is $1/n\mathbf{\Sigma}$), even though it is also unbiased.

5.

a) Problem 3.4.13. in DHS

i) We actually prove a stronger result than this. Consider two matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$. Then

$$\begin{aligned} tr(\mathbf{AB}) &= \sum_i (\mathbf{AB})_{ii} \\ &= \sum_{il} a_{il} b_{li} \\ &= \sum_{li} b_{li} a_{il} \\ &= \sum_l (\mathbf{BA})_{ll} \\ &= tr(\mathbf{BA}) \end{aligned}$$

This is a really neat result. It shows that, even though matrix multiplication is not commutative, the trace of the multiplication is. Can be very useful, as we will see in **iii)** below. Of course, by making $\mathbf{A} = \mathbf{a}^T$ and $\mathbf{B} = \mathbf{Aa}$ it follows that

$$tr(\mathbf{a}^T\mathbf{Aa}) = tr(\mathbf{Aaa}^T).$$

Since $\mathbf{a}^T\mathbf{Aa}$ is a scalar, it is equal to its trace and

$$\mathbf{a}^T\mathbf{Aa} = tr(\mathbf{Aaa}^T).$$

ii) The likelihood function for an iid sample $\mathbf{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is

$$\begin{aligned} P_{\mathbf{X}}(\mathbf{D}; \mu, \Sigma) &= \prod_i P_{\mathbf{X}}(\mathbf{x}_i; \mu, \Sigma) \\ &= \left(\frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}|}} \right)^n exp\left\{ -\frac{1}{2} \sum_i (\mathbf{x}_i - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mu) \right\} \\ &= \frac{1}{(2\pi)^{nd/2} |\mathbf{\Sigma}|^{n/2}} exp\left\{ -\frac{1}{2} \sum_i tr\left[\mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T \right] \right\} \\ &= \frac{1}{(2\pi)^{nd/2} |\mathbf{\Sigma}|^{n/2}} exp\left\{ -\frac{1}{2} tr\left[\mathbf{\Sigma}^{-1} \sum_i (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T \right] \right\}. \end{aligned}$$

iii) Defining $\mathbf{A} = \mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}$ this can be written as

$$P_{\mathbf{X}}(\mathbf{D}; \mu, \Sigma) = \frac{1}{(2\pi)^{nd/2} |\hat{\mathbf{\Sigma}}\mathbf{A}^{-1}|^{n/2}} exp\left\{ -\frac{n}{2} tr\left[\mathbf{A}\right] \right\}.$$

We can then use the fact that, for any square matrix $\mathbf{A}$,

$$trace(\mathbf{A}) = \sum_i \lambda_i \tag{4}$$

$$|\mathbf{A}| = \prod_i \lambda_i \tag{5}$$

where λ_i are the eigenvalues of $\mathbf{A}$, and for any two matrices of appropriate dimensions

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|. \tag{6}$$

It follows from $\mathbf{A} = \mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}$ that

$$\begin{aligned} |\mathbf{\Sigma}| &= \frac{|\hat{\mathbf{\Sigma}}|}{|\mathbf{A}|} & (7) \\ &= |\hat{\mathbf{\Sigma}}| / \prod_i \lambda_i & (8) \end{aligned}$$

and

$$P_{\mathbf{X}}(\mathbf{D}; \mu, \Sigma) = \frac{1}{(2\pi)^{nd/2}|\hat{\mathbf{\Sigma}}|^{n/2}} \left(\prod_i \lambda_i \right)^{n/2} exp\left\{ -\frac{n}{2} \sum_i \lambda_i \right\}.$$

iv) The maximum likelihood estimate is then given by

$$\begin{aligned} \mathbf{\Lambda}^* &= \arg \max_{\lambda_1,\ldots,\lambda_d} \mathbf{l}(\lambda_1, \ldots, \lambda_d) \\ &= \arg \max_{\lambda_1,\ldots,\lambda_d} \log P_{\mathbf{X}}(\mathbf{D}; \mu, \Sigma) \\ &= \arg \max_{\lambda_1,\ldots,\lambda_d} \left\{ -\frac{n}{2} \sum_i \lambda_i + \frac{n}{2} \sum_i \log \lambda_i - \log[(2\pi)^{nd/2}|\hat{\mathbf{\Sigma}}|^{n/2}] \right\}. \end{aligned}$$

Taking the gradient of the left-hand side

$$\frac{\partial \mathbf{l}(\lambda_1, \ldots, \lambda_d)}{\partial \lambda_i} = \frac{n}{2}\left(-1 + \frac{1}{\lambda_i} \right)$$

and setting to zero, leads to

$$\lambda_i = 1, \, \forall i. \tag{9}$$

Furthermore the Hessian is

$$\frac{\partial^2 \mathbf{l}(\lambda_1, \ldots, \lambda_d)}{\partial \lambda_i \lambda_j} = \begin{cases} -\frac{1}{\lambda_i^2}, & i = j \\ 0, & i \neq j \end{cases}$$

and clearly negative definite, from which we have a maximum. Assuming that $\mathbf{A}$ can be diagonalized, we can apply the eigenvector/eigenvalue decomposition

$$\mathbf{A} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T$$

where $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_d)$ is the diagonal matrix of eigenvalues of $\mathbf{A}$ and $\mathbf{\Phi}$ the orthonormal matrix whose columns are its eigenvectors. Since, the maximum occurs when $\Lambda = \mathbf{I}$, it follows that $\mathbf{A} = \mathbf{\Phi}\mathbf{\Phi}^T = \mathbf{I}$ and the maximum likelihood estimate for $\mathbf{\Sigma}$ is therefore $\hat{\mathbf{\Sigma}}$.

b) We can obtain the same result by simply finding the maximum of the likelihood

$$\mathbf{\Sigma}^* = \arg \max_{\mathbf{\Sigma}} \log P_{\mathbf{X}}(\mathbf{D}; \mu, \Sigma).$$

In these types of problems it is usually simpler to work with the *precision matrix* $\mathbf{P} = \mathbf{\Sigma}^{-1}$. Since there is a one to one correspondence between the two matrices this does not affect the location of the maximum, i.e.

$$\mathbf{\Sigma}^* = (\mathbf{P}^*)^{-1},$$

but the computation can be significantly simplified. In particular,

$$\begin{aligned}
\mathbf{P}^* &= \arg\max_{\mathbf{P}} \mathbf{l}(\mathbf{P}) \\
&= \arg\max_{\mathbf{P}} \log P_{\mathbf{X}}(\mathbf{D}; \mu, \mathbf{P}) \\
&= \arg\max_{\mathbf{P}} -\frac{1}{2}\sum_i (\mathbf{x}_i - \mu)^T \mathbf{P}(\mathbf{x}_i - \mu) + \frac{n}{2}\log\left[\frac{1}{(2\pi)^d}|\mathbf{P}|\right].
\end{aligned}$$

Using the results

$$\nabla_{\mathbf{P}}(\mathbf{x}^T\mathbf{P}\mathbf{x}) = \mathbf{x}\mathbf{x}^T \quad \text{and} \quad \nabla_{\mathbf{P}}|\mathbf{P}| = |\mathbf{P}|\mathbf{P}^{-T}$$

it follows that

$$\nabla_{\mathbf{P}}\mathbf{l}(\mathbf{P}) = -\frac{1}{2}\sum_i (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T + \frac{n}{2}\mathbf{P}^{-1}.$$

Setting the gradient to zero then leads to

$$\mathbf{\Sigma}^* = (\mathbf{P}^*)^{-1} = \frac{1}{n}\sum_i (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T = \hat{\mathbf{\Sigma}}.$$

Next, we compute the Hessian

$$\begin{aligned}
\nabla^2_{\mathbf{P}}\mathbf{l}(\mathbf{P}) &= \frac{n}{2}\nabla_P(\mathbf{P}^{-1}) \\
&= -\mathbf{P}^{-1}\mathbf{P}^{-1}
\end{aligned}$$

and apply the fact that a symmetric matrix $\mathbf{A}$ is positive definite if and only if there is a non-singular matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{R}^T\mathbf{R}$. (See e.g. *Linear Algebra and its Applications* by G. Strang, Harcourt Brace Jovanovic, 1988.) In this case, we have

$$\begin{aligned}
\nabla^2_{\mathbf{P}}\mathbf{l}(\mathbf{P}^*) &= -(\mathbf{P}*)^{-1}(\mathbf{P}*)^{-1} \\
&= -\mathbf{\Sigma}^*\mathbf{\Sigma}^* \\
&= -(\mathbf{\Sigma}^*)^T\mathbf{\Sigma}^*.
\end{aligned}$$

Since $\mathbf{\Sigma}^*$ is non-singular, it follows from the theorem that $\nabla^2_{\mathbf{P}}\mathbf{l}(\mathbf{P}^*)$ is negative definite and we have a maximum.