

## Chapter 2 Probability Distributions

Yue Yu

### 2.1 Binary Variables

(2.3)

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \sum_x xp(x) \\ &= \sum_x x\mu^x(1-\mu)^{1-x} \\ &= 1 \cdot \mu^1(1-\mu)^0 + 0 \cdot \mu^0(1-\mu)^1 \\ &= \mu.\end{aligned}$$

(2.4)

$$\begin{aligned}\text{var}[\mathbf{x}] &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 \\ &= \sum_x x^2 p(x) - \mu^2 \\ &= \sum_x x^2 \mu^x(1-\mu)^{1-x} - \mu^2 \\ &= 1^2 \cdot \mu^1(1-\mu)^0 + 0^2 \cdot \mu^0(1-\mu)^1 - \mu^2 \\ &= \mu(1-\mu).\end{aligned}$$

(2.7)

$$\begin{aligned}\mu_{\text{ML}} &= \arg_{\mu} \left( \frac{\partial}{\partial \mu} \ln p(\mathcal{D}|\mu) = 0 \right) \\ &= \arg_{\mu} \left( \frac{\partial}{\partial \mu} \sum_{n=1}^N \{x_n \ln \mu + (1-x_n) \ln(1-\mu)\} = 0 \right) \\ &= \arg_{\mu} \left( \sum_{n=1}^N \frac{x_n}{\mu} - \sum_{n=1}^N \frac{1-x_n}{1-\mu} = 0 \right) \\ &= \frac{1}{N} \sum_{n=1}^N x_n.\end{aligned}$$

**(2.11) and (2.12)**

We prove these two equations through problem 2.3 and 2.4. Firstly, notice that

$$\begin{aligned}
 \binom{N}{m} + \binom{N}{m-1} &= \frac{N!}{(N-m)!m!} + \frac{N!}{(N-m+1)!(m-1)!} \\
 &= \frac{N!(N-m+1)}{(N-m+1)!m!} + \frac{N!m}{(N-m+1)!m!} \\
 &= \frac{(N+1)!}{(N+1-m)!m!} \\
 &= \binom{N+1}{m}.
 \end{aligned} \tag{*}$$

Next, we prove by induction the *binomial theorem* that is given by

$$(1+x)^N = \sum_{m=0}^N \binom{N}{m} x^m.$$

When  $N = 0$ , we have

$$\sum_{m=0}^0 \binom{0}{m} x^m = 1 = (1+x)^0.$$

If the equation is correct for any integer  $N > 0$ , then for  $N+1$ , we have

$$\begin{aligned}
 (1+x)^{N+1} &= (1+x)^N (1+x) \\
 &= \sum_{m=0}^N \binom{N}{m} x^m (1+x) \\
 &= \sum_{m=0}^N \binom{N}{m} x^m + \sum_{m=0}^N \binom{N}{m} x^{m+1} \\
 &= \binom{N}{0} x^0 + \sum_{m=1}^N \binom{N}{m} x^m + \sum_{m=0}^{N-1} \binom{N}{m} x^{m+1} + \binom{N}{N} x^{N+1} \\
 &= \binom{N+1}{0} x^0 + \left[ \sum_{m=1}^N \binom{N}{m} x^m + \sum_{m=1}^N \binom{N}{m-1} x^m \right] + \binom{N+1}{N+1} x^{N+1} \\
 &= \binom{N+1}{0} x^0 + \sum_{m=1}^N \binom{N+1}{m} x^m + \binom{N+1}{N+1} x^{N+1} \\
 &= \sum_{m=0}^{N+1} \binom{N+1}{m} x^m,
 \end{aligned}$$

where the fifth step used (\*). Hence, the binomial theorem holds.

Then, we prove the binomial distribution is normalized. Specifically,

$$\begin{aligned}
 \sum_{m=0}^N \text{Bin}(m|N, \mu) &= \sum_{m=0}^N \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
 &= (1-\mu)^N \sum_{m=0}^N \binom{N}{m} \left( \frac{\mu}{1-\mu} \right)^m \\
 &= (1-\mu)^N \left( 1 + \frac{\mu}{1-\mu} \right)^N \\
 &= 1,
 \end{aligned} \tag{*}$$

where the second last step used the binomial theorem that we just proved.

Differentiating both sides of (\*) with respect to  $\mu$ , we have

$$\begin{aligned}
\frac{d}{d\mu} \sum_{m=0}^N \text{Bin}(m|N, \mu) &= \sum_{m=0}^N \binom{N}{m} \left( m\mu^{m-1}(1-\mu)^{N-m} - (N-m)\mu^m(1-\mu)^{N-m-1} \right) \\
&= \sum_{m=0}^N \binom{N}{m} \left( \frac{m}{\mu} \mu^m(1-\mu)^{N-m} - \frac{N-m}{1-\mu} \mu^m(1-\mu)^{N-m} \right) \\
&= \frac{1}{\mu(1-\mu)} \sum_{m=0}^N (m - N\mu) \binom{N}{m} \mu^m(1-\mu)^{N-m} \\
&= \frac{1}{\mu(1-\mu)} \sum_{m=0}^N (m - N\mu) \text{Bin}(m|N, \mu) \\
&= 0.
\end{aligned}$$

Rearranging the equation, we obtain

$$\begin{aligned}
\mathbb{E}[m] &= \sum_{m=0}^N m \text{Bin}(m|N, \mu) \\
&= N\mu \sum_{m=0}^N \text{Bin}(m|N, \mu) \\
&= N\mu,
\end{aligned}$$

where we used the fact we just proved that the binomial distribution is normalized.

To compute the variance, we further differentiate both sides of the above equation with respect to  $\mu$ ,

$$\begin{aligned}
\frac{d}{d\mu} \mathbb{E}[m] &= \sum_{m=0}^N m \binom{N}{m} (m\mu^{m-1}(1-\mu)^{N-m} - (N-m)\mu^m(1-\mu)^{N-m-1}) \\
&= \frac{1}{\mu(1-\mu)} \sum_{m=0}^N (m^2 - mN\mu) \binom{N}{m} \mu^m(1-\mu)^{N-m} \\
&= \frac{1}{\mu(1-\mu)} \sum_{m=0}^N (m^2 - mN\mu) \text{Bin}(m|N, \mu) \\
&= \frac{1}{\mu(1-\mu)} \left\{ \sum_{m=0}^N m^2 \text{Bin}(m|N, \mu) - N\mu \sum_{m=0}^N m \text{Bin}(m|N, \mu) \right\} \\
&= \frac{1}{\mu(1-\mu)} (\mathbb{E}[m^2] - \mathbb{E}[m]^2) \\
&= \frac{1}{\mu(1-\mu)} \text{var}[m] \\
&= N.
\end{aligned}$$

Therefore,

$$\text{var}[m] = N\mu(1-\mu).$$

## (2.14)

From the definition of the gamma function

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du,$$

we have

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty \exp(-x)x^{a-1} dx \int_0^\infty \exp(-y)y^{b-1} dy \\ &= \int_0^\infty \int_0^\infty \exp(-(x+y))x^{a-1}y^{b-1} dy dx.\end{aligned}$$

Substituting  $t = x + y$ , we have

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty \int_x^\infty \exp(-t)x^{a-1}(t-x)^{b-1} \left| \frac{dy}{dt} \right| dt dx \\ &= \int_0^\infty \int_x^\infty \exp(-t)x^{a-1}(t-x)^{b-1} dt dx \\ &= \int_0^\infty \int_0^t \exp(-t)x^{a-1}(t-x)^{b-1} dx dt.\end{aligned}$$

We further substitute  $x = t\mu$ , which gives

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty \int_0^1 \exp(-t)(t\mu)^{a-1}(t-t\mu)^{b-1} \left| \frac{dx}{d\mu} \right| d\mu dt \\ &= \int_0^\infty \int_0^1 \exp(-t)(t\mu)^{a-1}(t-t\mu)^{b-1} t d\mu dt \\ &= \int_0^\infty \exp(-t)t^{a+b-1} dt \int_0^1 \mu^{a-1}(1-\mu)^{b-1} d\mu \\ &= \Gamma(a+b) \int_0^1 \mu^{a-1}(1-\mu)^{b-1} d\mu.\end{aligned}$$

Therefore,

$$\begin{aligned}\int_0^1 \text{Beta}(\mu|a, b) d\mu &= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} d\mu \\ &= 1.\end{aligned}$$

**(2.15)**

$$\begin{aligned}\mathbb{E}[\mu] &= \int_0^1 \mu \text{Beta}(\mu|a, b) d\mu \\ &= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^a(1-\mu)^{b-1} d\mu \\ &= \frac{\Gamma(a+b)\Gamma(a+1)}{\Gamma(a+b+1)\Gamma(a)} \int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} \mu^a(1-\mu)^{b-1} d\mu \\ &= \frac{\Gamma(a+b)\Gamma(a+1)}{\Gamma(a+b+1)\Gamma(a)} \\ &= \frac{a\Gamma(a+b)\Gamma(a)}{(a+b)\Gamma(a+b)\Gamma(a)} \\ &= \frac{a}{a+b},\end{aligned}$$

where the third step used the fact that the gamma distribution is normalized, and the second last step used the property  $\Gamma(x+1) = x\Gamma(x)$ .

(2.16)

$$\begin{aligned}
\text{var}[\mu] &= \mathbb{E}[\mu^2] - \mathbb{E}[\mu]^2 \\
&= \int_0^1 \mu^2 \text{Beta}(\mu|a, b) \, d\mu - \left(\frac{a}{a+b}\right)^2 \\
&= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a+1} (1-\mu)^{b-1} \, d\mu - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{\Gamma(a+b)\Gamma(a+2)}{\Gamma(a+b+2)\Gamma(a)} \int_0^1 \frac{\Gamma(a+b+2)}{\Gamma(a+2)\Gamma(b)} \mu^{a+1} (1-\mu)^{b-1} \, d\mu - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{\Gamma(a+b)\Gamma(a+2)}{\Gamma(a+b+2)\Gamma(a)} - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{a(a+1)\Gamma(a+b)\Gamma(a)}{(a+b)(a+b+1)\Gamma(a+b)\Gamma(a)} - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{a(a+1)}{(a+b)(a+b+1)} - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{ab}{(a+b)^2(a+b+1)},
\end{aligned}$$

where, again, in the fifth step, we used the fact that the gamma distribution is normalized, and in the sixth step, we used the property  $\Gamma(x+1) = x\Gamma(x)$ .

(2.19)

$$\begin{aligned}
p(x=1|\mathcal{D}) &= \int_0^1 p(x=1, \mu|\mathcal{D}) \, d\mu \\
&= \int_0^1 p(x=1|\mu, \mathcal{D}) p(\mu|\mathcal{D}) \, d\mu \\
&= \int_0^1 p(x=1|\mu) p(\mu|\mathcal{D}) \, d\mu \\
&= \int_0^1 \mu p(\mu|\mathcal{D}) \, d\mu \\
&= \mathbb{E}[\mu|\mathcal{D}],
\end{aligned}$$

which is the expected value of  $\mu$  after observing the dataset  $\mathcal{D}$ .

The third step omitted  $\mathcal{D}$  by the i.i.d assumption such that the probability of  $x=1$  given  $\mu$  does not depend on the observed data. In the second last step, we assumed that  $x \sim \text{Bern}(\mu)$ , and hence,  $p(x=1|\mu) = \mu$ .

## 2.2 Multinomial Variables

(2.29)

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^N p(\mathbf{x}_n|\boldsymbol{\mu}) = \prod_{n=1}^N \prod_{k=1}^K \mu_k^{x_{nk}} = \prod_{k=1}^K \left( \prod_{n=1}^N \mu_k^{x_{nk}} \right) = \prod_{k=1}^K \mu_k^{\sum_n x_{nk}} = \prod_{k=1}^K \mu_k^{m_k}.$$

## 2.3 The Gaussian Distribution

(2.45)

To see that the matrix  $\Sigma$  can be taken to be symmetric, for precision matrix  $\Lambda$ , denote symmetric matrix  $\Lambda^S = (\Lambda + \Lambda^T)/2$ , and anti-symmetric matrix  $\Lambda^A = (\Lambda - \Lambda^T)/2$ , then

$$\begin{aligned}\Delta^2 &= \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Lambda (\mathbf{x} - \boldsymbol{\mu}) \\ &= \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T (\Lambda^S + \Lambda^A) (\mathbf{x} - \boldsymbol{\mu}) \\ &= \frac{1}{2} \sum_{i=1}^D \sum_{j=1}^D (x_i - \mu_i) \Lambda_{ij}^S (x_j - \mu_j) + \frac{1}{2} \sum_{i=1}^D \sum_{j=1}^D (x_i - \mu_i) \Lambda_{ij}^A (x_j - \mu_j).\end{aligned}$$

In the last step, for each  $\Lambda_{ij}^A$ , there is a corresponding  $\Lambda_{ji}^A = -\Lambda_{ij}^A$ . Hence, the second term vanishes, which implies that the covariance matrix can be chosen to be symmetric.

(2.46)

To prove that all the eigenvalues of a real symmetric matrix are real, suppose that for

$$\Sigma \mathbf{u} = \lambda \mathbf{u}, \tag{*}$$

$\lambda = a + bi$  and  $\bar{\lambda} = a - bi$ . Taking conjugates of (\*), we have

$$\Sigma \bar{\mathbf{u}} = \bar{\lambda} \bar{\mathbf{u}},$$

transposing both sides, we have

$$\bar{\mathbf{u}}^T \Sigma = \bar{\mathbf{u}}^T \bar{\lambda}. \tag{**}$$

Multiplying  $\bar{\mathbf{u}}^T$  to the left of (\*) on both sides gives

$$\bar{\mathbf{u}}^T \Sigma \mathbf{u} = \bar{\mathbf{u}}^T \lambda \mathbf{u}.$$

Similarly, by multiplying  $\mathbf{u}$  to the right of (\*\*) on both sides, we obtain

$$\bar{\mathbf{u}}^T \Sigma \mathbf{u} = \bar{\mathbf{u}}^T \bar{\lambda} \mathbf{u}$$

Therefore,

$$\bar{\mathbf{u}}^T \lambda \mathbf{u} = \bar{\mathbf{u}}^T \bar{\lambda} \mathbf{u}.$$

Since  $\bar{\mathbf{u}}^T \mathbf{u} \neq 0$ , we must have  $\lambda = \bar{\lambda}$ , that is,  $a + bi = a - bi$ . Hence,  $b = 0$ , which implies that  $\lambda$  is real.

For any pair of  $\mathbf{u}_i$  and  $\mathbf{u}_j$  where  $i \neq j$ , we have

$$\begin{aligned}\mathbf{u}_i^T \lambda_i \mathbf{u}_j &= (\lambda_i \mathbf{u}_i)^T \mathbf{u}_j \\ &= (\Sigma \mathbf{u}_i)^T \mathbf{u}_j \\ &= \mathbf{u}_i^T \Sigma^T \mathbf{u}_j \\ &= \mathbf{u}_i^T \Sigma \mathbf{u}_j \\ &= \mathbf{u}_i^T \lambda_j \mathbf{u}_j.\end{aligned}$$

Since  $\lambda_i \neq \lambda_j$ , we obtain  $\lambda_i \mathbf{u}_i^T \mathbf{u}_j = \lambda_j \mathbf{u}_i^T \mathbf{u}_j$ , which means that  $\mathbf{u}_i^T \mathbf{u}_j = 0$ , that is,  $\mathbf{u}_i \perp \mathbf{u}_j$ . Therefore,

$$\Sigma = \mathbf{U} \Lambda \mathbf{U}^{-1} = \mathbf{U} \Lambda \mathbf{U}^T,$$

indicating that the set of eigenvectors can be chosen to be orthonormal.

(2.48)

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_D] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_D^T \end{bmatrix} = \sum_{i=1}^D \lambda_i \mathbf{u}_i \mathbf{u}_i^T.$$

(2.49)

$$\mathbf{\Sigma}^{-1} = (\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)^{-1} = \mathbf{U}^{-T}\mathbf{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T = \sum_{i=1}^D \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T.$$

(2.60)

According to (2.52),

$$\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{U}\mathbf{z},$$

we obtain

$$\mathbf{z} = \mathbf{U}^T \mathbf{y} = [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_D] \begin{bmatrix} y_1 \\ \vdots \\ y_D \end{bmatrix} = \sum_{j=1}^D y_j \mathbf{u}_j,$$

where  $y_j$  is defined by (2.51).

(2.120)

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{\Sigma}) &= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{n=1}^N (\boldsymbol{\mu}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu} - 2\mathbf{x}_n^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}) \\ &= -\sum_{n=1}^N (\mathbf{\Sigma}^{-1} \boldsymbol{\mu} - \mathbf{\Sigma}^{-1} \mathbf{x}_n) \\ &= \sum_{n=1}^N \mathbf{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}). \end{aligned}$$

(2.122)

Setting the derivative of likelihood function with respect to  $\mathbf{\Sigma}^{-1}$  to  $\mathbf{0}$ , we have

$$\begin{aligned} \frac{\partial}{\partial \mathbf{\Sigma}^{-1}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{\Sigma}) &= \frac{\partial}{\partial \mathbf{\Sigma}^{-1}} \left( -\frac{N}{2} \ln |\mathbf{\Sigma}| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \right) \\ &= \frac{\partial}{\partial \mathbf{\Sigma}^{-1}} \left( -\frac{N}{2} \ln |\mathbf{\Sigma}| - \frac{1}{2} \sum_{n=1}^N \text{tr} \left[ (\mathbf{x}_n - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \right] \right) \\ &= \frac{\partial}{\partial \mathbf{\Sigma}^{-1}} \left( -\frac{N}{2} \ln |\mathbf{\Sigma}| - \frac{1}{2} \sum_{n=1}^N \text{tr} \left[ \mathbf{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) (\mathbf{x}_n - \boldsymbol{\mu})^T \right] \right) \\ &= \frac{N}{2} \mathbf{\Sigma} - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}) (\mathbf{x}_n - \boldsymbol{\mu})^T \\ &= \mathbf{0}, \end{aligned}$$

which implies that

$$\mathbf{\Sigma} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}) (\mathbf{x}_n - \boldsymbol{\mu})^T.$$

**(2.123)**

This is a multivariate generalization of (1.56).

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}_{\text{ML}}] &= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^N \mathbf{x}_n\right] \\
&= \frac{1}{N} \mathbb{E}\left[\sum_{n=1}^N \mathbf{x}_n\right] \\
&= \frac{1}{N} N \boldsymbol{\mu} \\
&= \boldsymbol{\mu},
\end{aligned}$$

where in the second last step, we took advantage of the i.i.d assumption such that  $\mathbb{E}[\mathbf{x}_n] = \boldsymbol{\mu}$  for any  $n \in \{1, \dots, N\}$ .

**(2.124)**

This is a multivariate generalization of (1.57), see (1.57) in *Chapter 1 Introduction* for details.

**(2.126)**

$$\begin{aligned}
\boldsymbol{\mu}_{\text{ML}}^{(N)} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \\
&= \frac{1}{N} \mathbf{x}_N + \frac{1}{N} \sum_{n=1}^{N-1} \mathbf{x}_n \\
&= \frac{1}{N} \mathbf{x}_N + \frac{N-1}{N} \frac{1}{N-1} \sum_{n=1}^{N-1} \mathbf{x}_n \\
&= \frac{1}{N} \mathbf{x}_N + \frac{N-1}{N} \boldsymbol{\mu}_{\text{ML}}^{(N-1)} \\
&= \boldsymbol{\mu}_{\text{ML}}^{(N-1)} + \frac{1}{N} (\mathbf{x}_N - \boldsymbol{\mu}_{\text{ML}}^{(N-1)}).
\end{aligned}$$

**(2.135)**

The observed value of  $z$  is with respect to  $x_N$  where we have the estimated  $\theta$  based on the previous  $N-1$  observations.

**(2.136)**

$$\begin{aligned}
z &= \frac{\partial}{\partial \mu_{\text{ML}}} [-\ln p(x|\mu_{\text{ML}}, \sigma^2)] \\
&= \frac{\partial}{\partial \mu_{\text{ML}}} \left[ \frac{(x - \mu_{\text{ML}})^2}{2\sigma^2} \right] \\
&= -\frac{1}{\sigma^2} (x - \mu_{\text{ML}}).
\end{aligned}$$

Substituting back into (2.135), we have

$$\begin{aligned}
\mu_{\text{ML}}^{(N)} &= \mu_{\text{ML}}^{(N-1)} - a_{N-1} \frac{\partial}{\partial \mu_{\text{ML}}^{(N-1)}} [-\ln p(x_N|\mu_{\text{ML}}^{(N-1)})] \\
&= \mu_{\text{ML}}^{(N-1)} + a_{N-1} \frac{1}{\sigma^2} (x_N - \mu_{\text{ML}}^{(N-1)}).
\end{aligned}$$



Comparing with (2.126), we obtain

$$a_N = \frac{\sigma^2}{N}.$$

### (2.141) and (2.142)

Recall

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu)$$

where

$$\begin{aligned} p(\mu|\mathbf{x}) &= \mathcal{N}(\mu|\mu_N, \sigma_N^2) \\ p(\mu) &= \mathcal{N}(\mu|\mu_0, \sigma_0^2). \end{aligned}$$

By completing the square in the exponent, specifically, comparing the coefficients of the quadratic and linear term, we have

$$-\frac{1}{2\sigma_N^2} = -\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma^2} \sum_{n=1}^N 1 \quad (*)$$

$$\frac{1}{\sigma_N^2} \mu_N = \frac{1}{\sigma_0^2} \mu_0 + \frac{1}{\sigma^2} \sum_{n=1}^N x_n. \quad (**)$$

Solving for  $\sigma_N$  using (\*\*) and substituting back into (\*) for  $\mu_N$ , we obtain

$$\begin{aligned} \mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{\text{ML}} \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}, \end{aligned}$$

where

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n.$$

### (2.146)

To see the gamma distribution is normalized, consider

$$\begin{aligned} \int_0^\infty \text{Gam}(\lambda|a, b) d\lambda &= \int_0^\infty \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) d\lambda \\ &= \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a-1} \exp(-b\lambda) d\lambda. \end{aligned}$$

Substituting  $u = b\lambda$ , we have

$$\begin{aligned} \int_0^\infty \text{Gam}(\lambda|a, b) d\lambda &= \frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{u}{b}\right)^{a-1} \exp(-u) \left|\frac{d\lambda}{du}\right| du \\ &= \frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{u}{b}\right)^{a-1} \exp(-u) \frac{1}{b} du \\ &= \frac{1}{\Gamma(a)} \int_0^\infty u^{a-1} \exp(-u) du \\ &= \frac{1}{\Gamma(a)} \Gamma(a) \\ &= 1. \end{aligned}$$

(2.147)

$$\begin{aligned}
\mathbb{E}[\lambda] &= \int_0^\infty \text{Gam}(\lambda|a, b) \lambda \, d\lambda \\
&= \int_0^\infty \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \lambda \, d\lambda \\
&= \frac{1}{\Gamma(a)} \int_0^\infty (b\lambda)^a \exp(-b\lambda) \, d\lambda.
\end{aligned}$$

Substituting  $u = b\lambda$ , we have

$$\begin{aligned}
\mathbb{E}[\lambda] &= \frac{1}{\Gamma(a)} \int_0^\infty u^a \exp(-u) \left| \frac{d\lambda}{du} \right| du \\
&= \frac{a}{b\Gamma(a)} \int_0^\infty u^{a-1} \exp(-u) \, du \\
&= \frac{a}{b\Gamma(a)} \Gamma(a) \\
&= \frac{a}{b}.
\end{aligned}$$

(2.148)

$$\begin{aligned}
\text{var}[\lambda] &= \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda]^2 \\
&= \int_0^\infty \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \lambda^2 \, d\lambda - \left( \frac{a}{b} \right)^2 \\
&= \frac{1}{b\Gamma(a)} \int_0^\infty (b\lambda)^{a+1} \exp(-b\lambda) \, d\lambda - \left( \frac{a}{b} \right)^2.
\end{aligned}$$

Substituting  $u = b\lambda$ , we have

$$\begin{aligned}
\text{var}[\lambda] &= \frac{1}{b\Gamma(a)} \int_0^\infty u^{a+1} \exp(-u) \left| \frac{d\lambda}{du} \right| du - \left( \frac{a}{b} \right)^2 \\
&= \frac{1}{b^2\Gamma(a)} \int_0^\infty u^{a+1} \exp(-u) \, du - \left( \frac{a}{b} \right)^2 \\
&= \frac{1}{b^2\Gamma(a)} \Gamma(a+2) - \left( \frac{a}{b} \right)^2 \\
&= \frac{a(a+1)\Gamma(a)}{b^2\Gamma(a)} - \left( \frac{a}{b} \right)^2 \\
&= \frac{a}{b^2},
\end{aligned}$$

where the second last step used the property  $\Gamma(x+1) = x\Gamma(x)$ .

(2.152)

$$\begin{aligned}
p(\mathbf{x}|\mu, \lambda) &= \prod_{n=1}^N \left( \frac{\lambda}{2\pi} \right)^{1/2} \exp \left\{ -\frac{\lambda}{2} (x_n - \mu)^2 \right\} \\
&= \left( \frac{\lambda}{2\pi} \right)^{N/2} \exp \left\{ -\frac{\lambda}{2} \sum_{n=1}^N (x_n - \mu)^2 \right\} \\
&\propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda\mu^2}{2} \right) \right]^N \exp \left\{ \lambda\mu \sum_{n=1}^N x_n - \frac{\lambda}{2} \sum_{n=1}^N x_n^2 \right\}.
\end{aligned}$$

(2.158)

$$\begin{aligned}
p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) d\tau \\
&= \int_0^\infty \frac{b^a e^{-b\tau} \tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\} d\tau \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \int_0^\infty \tau^{a-1/2} \exp\left\{-b\tau - \frac{\tau}{2}(x-\mu)^2\right\} d\tau.
\end{aligned}$$

By making the change of variable  $z = \tau[b + (x - \mu)^2/2]$ , we have

$$\begin{aligned}
p(x|\mu, a, b) &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \int_0^\infty z^{a-1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a+1/2} \exp(-z) \left|\frac{d\tau}{dz}\right| dz \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \int_0^\infty z^{a-1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a+1/2} \exp(-z) \left[b + \frac{(x-\mu)^2}{2}\right]^{-1} dz \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2} \int_0^\infty z^{a-1/2} \exp(-z) dz \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2} \Gamma(a + 1/2).
\end{aligned}$$

(2.162)

$$\begin{aligned}
\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) &= \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}(\eta|\nu/2, \nu/2) d\eta \\
&= \int_0^\infty \frac{|\eta\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \eta\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})\right\} \frac{1}{\Gamma(\nu/2)} \left(\frac{\nu}{2}\right)^{\nu/2} \eta^{\nu/2-1} \exp\left\{-\frac{\nu\eta}{2}\right\} d\eta \\
&= \frac{|\boldsymbol{\Lambda}|^{1/2} (\nu/2)^{\nu/2}}{(2\pi)^{D/2} \Gamma(\nu/2)} \int_0^\infty \exp\left\{-\frac{\eta}{2}(\Delta^2 + \nu)\right\} \eta^{D/2+\nu/2-1} d\eta,
\end{aligned}$$

where we have defined

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}),$$

and in the last step we used the property that  $|\eta\boldsymbol{\Lambda}| = \eta^D |\boldsymbol{\Lambda}|$ .

Now, we substitute  $z = \eta(\Delta^2 + \nu)/2$ , which gives

$$\begin{aligned}
\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) &= \frac{|\boldsymbol{\Lambda}|^{1/2} (\nu/2)^{\nu/2}}{(2\pi)^{D/2} \Gamma(\nu/2)} \int_0^\infty \exp(-z) \left(\frac{2z}{\Delta^2 + \nu}\right)^{D/2+\nu/2-1} \left|\frac{d\eta}{dz}\right| dz \\
&= \frac{|\boldsymbol{\Lambda}|^{1/2} (\nu/2)^{\nu/2}}{(2\pi)^{D/2} \Gamma(\nu/2)} \int_0^\infty \exp(-z) \left(\frac{2z}{\Delta^2 + \nu}\right)^{D/2+\nu/2-1} \frac{2}{\Delta^2 + \nu} dz \\
&= \frac{|\boldsymbol{\Lambda}|^{1/2} (\nu/2)^{\nu/2}}{(2\pi)^{D/2} \Gamma(\nu/2)} \left(\frac{2z}{\Delta^2 + \nu}\right)^{D/2+\nu/2} \int_0^\infty \exp(-z) z^{D/2+\nu/2-1} dz \\
&= \frac{|\boldsymbol{\Lambda}|^{1/2} (\nu/2)^{\nu/2}}{(2\pi)^{D/2} \Gamma(\nu/2)} \left(\frac{2z}{\Delta^2 + \nu}\right)^{D/2+\nu/2} \Gamma(D/2 + \nu/2) \\
&= \frac{\Gamma(D/2 + \nu/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2}.
\end{aligned}$$

(2.164)

$$\begin{aligned}
\mathbb{E}_{\mathbf{x} \sim \text{St}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)}[\mathbf{x}] &= \int_{\mathbf{x}} \mathbf{x} \text{St}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) d\mathbf{x} \\
&= \int_{\mathbf{x}} \mathbf{x} \left( \int_0^\infty \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1}) \text{Gam}(\eta | \nu/2, \nu/2) d\eta \right) d\mathbf{x} \\
&= \int_0^\infty \left( \int_{\mathbf{x}} \mathbf{x} \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1}) d\mathbf{x} \right) \text{Gam}(\eta | \nu/2, \nu/2) d\eta \\
&= \int_0^\infty \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1})}[\mathbf{x}] \text{Gam}(\eta | \nu/2, \nu/2) d\eta \\
&= \boldsymbol{\mu} \int_0^\infty \text{Gam}(\eta | \nu/2, \nu/2) d\eta \\
&= \boldsymbol{\mu},
\end{aligned}$$

where we used the property that the expected value of a Gaussian random variable is its mean, and the fact that the gamma distribution is normalized.

(2.165)

Firstly, notice that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x} \sim \text{St}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)}[\mathbf{x} \mathbf{x}^T] &= \int_{\mathbf{x}} \mathbf{x} \mathbf{x}^T \text{St}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) d\mathbf{x} \\
&= \int_{\mathbf{x}} \mathbf{x} \mathbf{x}^T \left( \int_0^\infty \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1}) \text{Gam}(\eta | \nu/2, \nu/2) d\eta \right) d\mathbf{x} \\
&= \int_0^\infty \left( \int_{\mathbf{x}} \mathbf{x} \mathbf{x}^T \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1}) d\mathbf{x} \right) \text{Gam}(\eta | \nu/2, \nu/2) d\eta \\
&= \int_0^\infty \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1})}[\mathbf{x} \mathbf{x}^T] \text{Gam}(\eta | \nu/2, \nu/2) d\eta \\
&= \int_0^\infty (\text{cov}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1})}[\mathbf{x}] + \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1})}[\mathbf{x}] \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1})}[\mathbf{x}^T]) \text{Gam}(\eta | \nu/2, \nu/2) d\eta \\
&= \int_0^\infty ((\eta \boldsymbol{\Lambda})^{-1} + \boldsymbol{\mu} \boldsymbol{\mu}^T) \text{Gam}(\eta | \nu/2, \nu/2) d\eta \\
&= \boldsymbol{\mu} \boldsymbol{\mu}^T \int_0^\infty \text{Gam}(\eta | \nu/2, \nu/2) d\eta + \boldsymbol{\Lambda}^{-1} \int_0^\infty \eta^{-1} \text{Gam}(\eta | \nu/2, \nu/2) d\eta \\
&= \boldsymbol{\mu} \boldsymbol{\mu}^T + \boldsymbol{\Lambda}^{-1} \int_0^\infty \eta^{-1} \text{Gam}(\eta | \nu/2, \nu/2) d\eta \\
&= \boldsymbol{\mu} \boldsymbol{\mu}^T + \frac{\boldsymbol{\Lambda}^{-1} (\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \int_0^\infty \eta^{\nu/2-2} \exp\left(-\frac{\nu\eta}{2}\right) d\eta.
\end{aligned}$$

Making the change of variable  $z = \nu\eta/2$ , we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{x} \sim \text{St}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)}[\mathbf{x}\mathbf{x}^T] &= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{\boldsymbol{\Lambda}^{-1}(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \int_0^\infty \left(\frac{2z}{\nu}\right)^{\nu/2-2} \exp(-z) \left|\frac{d\eta}{dz}\right| dz \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{\boldsymbol{\Lambda}^{-1}(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \int_0^\infty \left(\frac{2z}{\nu}\right)^{\nu/2-2} \exp(-z) \left(\frac{2}{\nu}\right) dz \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{\boldsymbol{\Lambda}^{-1}}{\Gamma(\nu/2)} \frac{\nu}{2} \int_0^\infty z^{\nu/2-2} \exp(-z) dz \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{\boldsymbol{\Lambda}^{-1}\nu}{2} \frac{\Gamma(\nu/2-1)}{\Gamma(\nu/2)} \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{\boldsymbol{\Lambda}^{-1}\nu}{2} \frac{\Gamma(\nu/2-1)}{(\nu/2-1)\Gamma(\nu/2-1)} \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{\nu}{\nu-2} \boldsymbol{\Lambda}^{-1},
\end{aligned}$$

where in the second last step we used the property  $\Gamma(x+1) = x\Gamma(x)$ .

Therefore,

$$\begin{aligned}
\text{cov}_{\mathbf{x} \sim \text{St}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)}[\mathbf{x}] &= \mathbb{E}_{\mathbf{x} \sim \text{St}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}_{\mathbf{x} \sim \text{St}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)}[\mathbf{x}]\mathbb{E}_{\mathbf{x} \sim \text{St}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)}[\mathbf{x}^T] \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{\nu}{\nu-2} \boldsymbol{\Lambda}^{-1} - \boldsymbol{\mu}\boldsymbol{\mu}^T \\
&= \frac{\nu}{\nu-2} \boldsymbol{\Lambda}^{-1}.
\end{aligned}$$

(2.228)

$$\begin{aligned}
\nabla_{\boldsymbol{\eta}} \ln p(\mathbf{X}|\boldsymbol{\eta}) &= \nabla_{\boldsymbol{\eta}} \left( \sum_{n=1}^N \ln h(\mathbf{x}_n) + N \ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) \right) \\
&= N \nabla_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}) + \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) \\
&= \mathbf{0}.
\end{aligned}$$

Rearranging the equation, we obtain

$$-\nabla_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}_{\text{ML}}) = \frac{1}{N} \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n).$$