1 Probability and independence

a) Decomposition

We aim to validate

$$(X \perp Y, W \mid Z) \implies (X \perp Y \mid Z) \tag{1.1}$$

Proof. We suppose the statement $(X \perp Y,\ W \mid Z)$ is true. It follows from the definition of the conditional independence that $p(x, y, w \mid z) = p(x \mid z)p(y, w \mid z)$ for all $x \in \Omega_x$, $(y, w) \in \Omega_y \times \Omega_w$ and $z \in \Omega_z$. We then consider the marginalize $p(x, y, w \mid z)$:

$$\begin{aligned}
p(x, y|z) &= \sum_{w\ \in\ \Omega_w} p(x, y, w \mid z) \\
&= \sum_{w\ \in\ \Omega_w} p(x \mid z)p(y, w \mid z) \\
&= p(x \mid z) \sum_{w\ \in\ \Omega_w} p(y, w \mid z) \\
&= p(x \mid z)p(y \mid z)
\end{aligned}$$

from which we conclude that $(X \perp Y \mid Z)$ $\square$. By symmetry of the argument, we can show that $(X \perp W \mid Z)$ is true as well.

b)

We aim to validate

$$(X \perp Y \mid Z) \text{ and } (X, Y \perp W \mid Z) \implies (X \perp W \mid Z) \tag{1.2}$$

Proof. Suppose $(X, Y \perp W \mid Z)$ and $(X \perp Y \mid Z)$ are true. We know from the symmetry and decomposition properties of the conditional independence that $(X, Y \perp W \mid Z) \implies (W \perp X, Y \mid Z) \implies (X \perp W \mid Z)$. Therefore $(X \perp W \mid Z)$ is true $\square$.

c)

We aim to validate

$$(X \perp Y, W \mid Z) \text{ and } (Y \perp W \mid Z) \implies (X, W \perp Y \mid Z) \tag{1.3}$$

Proof. Suppose $(X \perp Y, W \mid Z)$ is true. Then it follows from the definition of conditional independence that

$$p(x, y, w \mid z) = p(x \mid z)p(y, w \mid z)$$

Then assume $(Y \perp W \mid Z)$ is true. The second factor can be factorized

$$p(x, y, w \mid z) = p(x \mid z)p(y \mid z)p(w \mid z)$$

From the decomposition property, we know $(X \perp W \mid Z)$ is true. Thus

$$p(x, y, w \mid z) = p(x, w \mid z)p(y \mid z)$$

From which we conclude $(X, W \perp Y \mid Z)$ is true $\square$.

d)

We aim to validate

$$(X \perp Y \mid Z) \text{ and } (X \perp Y \mid W) \implies (X \perp Y \mid Z, W) \tag{1.4}$$

Counter example. We consider the following R.V.

1. X: Person A arrive late for diner;
2. Y: Person B arrive late for diner;
3. W: They come from the same city;
4. Z: They work in the same city.

For this situation, we see that X and Y are conditionally independent when given either W or Z. If we know they are from the same city, then they might work in different cities and take different route home. Thus knowing person A was late doesn't inform us on the probability of person B to arrive late.

A similar argument can be made for $(X \perp Y \mid Z)$.

Thus the LHS of the proposition is true, yet the RHS is clearly false in our case. Assuming we were given that W and Z are true, then we are given the geolocalisation of person A and B. If we were given that person A would be late for diner, then we'd be able to make a good guess that person B would be late as well (they would both be impacted by the same traffic jam or whatnot). Thus the proposition is false.

2 Bayesian inferance and MAP

Let $\mathbf{X}_1, \dots, \mathbf{X}_n \mid \boldsymbol{\pi} \overset{\text{iid}}{\sim} \text{Multinomial}(1, \boldsymbol{\pi})$ on k element. The values are sampled from a set of cardinality 2, that is $x_j^{(i)} \in \{0, 1\}$. Each R.V. has only one non-zero entry for a given trial, that is $\sum_{j=1}^k x_j^{(i)} = 1$.

We assume a Dirichlet prior $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha})$ with a PDF

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \pi_j^{\alpha_j - 1}$$

a)

Since the data is IID, they are mutually independent by definition. Being given the parameters of their Multinomial distribution (or a subset for that matter) does not change the independence of the $\mathbf{X}$'s. Thus,

$$(\mathbf{X}_i \perp \mathbf{X}_j \mid \boldsymbol{\pi}) \, \forall \, (i, j) \, \in \{1, \dots, k\} \times \{1, \dots, k\}$$

Of course, none of the vector can be mutually nor conditionally independent to $\boldsymbol{\pi}$ since it contains information about the distribution of the one hot vectors $\mathbf{X}_i$. In this case $\boldsymbol{\pi}$ are the probabilities of one of the k entry to be equal to one. Even giving one of these away is enough to impact the posterior distribution of the conditional $p(x_i \mid x_\ell, \pi_j)$ for example.

b)

The posterior distribution $p(\boldsymbol{\pi} \mid x_1, \dots, x_n)$ is computed via the Bayes rule

$$p(\boldsymbol{\pi} \mid \mathbf{x}_{1:n}) = \frac{p(\mathbf{x}_{1:n} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi})}{p(\mathbf{x}_{1:n})}$$

where $p(\boldsymbol{\pi}) = p(\boldsymbol{\pi} \mid \boldsymbol{\alpha})$ is the prior for $\boldsymbol{\pi}$ defined above. For the sake of determining the posterior distribution, we can postpone the derivation of the marginal likelihood. The likelihood $p(\mathbf{x}_{1:n} \mid \boldsymbol{\pi})$ is the probability mass function corresponding to n trials of a $k-$sided die throw. We define the vector $\boldsymbol{\chi} \equiv \sum_{i=1}^n \mathbf{x}_j$ with the property

$$\sum_{j=1}^k \chi_j = n$$

It becomes clear that the likelihood follows the Multinomial$(n, \boldsymbol{\pi})$ distribution. The PMF is given by

$$p(\mathbf{x}_{1:n} \mid \boldsymbol{\pi}) = \binom{n}{\chi_1, \dots, \chi_k} \prod_{j=1}^k \pi_j^{\chi_j} \propto \prod_{i=1}^n \prod_{j=1}^k \pi_j^{x_j^{(i)}}$$

Where it is agreed that $\chi_j = \sum_{i=1}^n x_j^{(i)}$. Therefore, the posterior must be

$$p(\boldsymbol{\pi} \mid \mathbf{x}_{1:n}) \propto \prod_{i=1}^n \prod_{j=1}^k \pi_j^{x_j^{(i)}} \prod_{\ell=1}^k \pi_\ell^{\alpha_\ell - 1}$$

We use the fact that we can swap around product operator for real numbers.

$$p(\boldsymbol{\pi} \mid \mathbf{x}_{1:n}) \propto \prod_{i=1}^n \prod_{j=1}^k \prod_{\ell=1}^k \pi_\ell^{\alpha_\ell - 1} \pi_j^{x_j^{(i)}} = \prod_{j=1}^k \pi_j^{\sum_{i=1}^n x_j^{(i)} + \alpha_j - 1}$$

We can readily see that the resulting distribution will be a Dirichlet with updated α_ℓ's.

The posterior distribution is a Dirichlet distribution with parameters $\alpha_j' = \alpha_j + \sum_{i=1}^n x_j^{(i)}$.

c) Marginal Likelihood

The marginal likelihood $p(\mathbf{x}_{1:n})$ is a normalizing constant defined as the integral of the numerator over all instantiation of $\boldsymbol{\pi}$

$$p(\mathbf{x}_{1:n}) = \int_{\boldsymbol{\Delta}_k} p(\mathbf{x}_{1:n} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) d^{(k)}\boldsymbol{\pi}$$

where $\boldsymbol{\Delta}_k$ is the probability simplex. In term of the quantities defined above, this is

$$p(\mathbf{x}_{1:n}) = \int_{\boldsymbol{\Delta}_k} d^{(k)}\boldsymbol{\pi} \binom{n}{\chi_1, \dots, \chi_k} \prod_{j=1}^{k} \pi_j^{\chi_j} \left(\frac{\Gamma(\sum_{\ell=1}^{k} \alpha_\ell)}{\prod_{\ell=1}^{k} \Gamma(\alpha_\ell)} \prod_{\ell=1}^{k} \pi_\ell^{\alpha_\ell - 1} \right)$$

The π_j's are independent variables since the simplex $\boldsymbol{\Delta}_k$ is crucially defined as an affine plane in an Euclidian space which is supported by a set of orthonormal vectors. Thus, our task is to evaluate k identical integrals of the form

$$\int_0^1 d\pi_j \pi_j^{\sum_{i=1}^{n} x_j^{(i)} + \alpha_j - 1} = \left(\sum_{i=1}^{n} x_j^{(i)} + \alpha_j \right)^{-1}, \quad \{\alpha_j > 0\}$$

Thus

$$p(\mathbf{x}_{1:n}) = \binom{n}{\chi_1, \dots, \chi_k} \frac{\Gamma(\sum_{\ell=1}^{k} \alpha_\ell)}{\prod_{\ell=1}^{k} \Gamma(\alpha_\ell)} \prod_{j=1}^{k} \left(\sum_{i=1}^{n} x_j^{(i)} + \alpha_j \right)^{-1}$$

d) $\hat{\boldsymbol{\pi}}_{\mathbf{MAP}}$

The maximum *a posteriori* of the Multinomial distribution is

$$\hat{\boldsymbol{\pi}}_{\text{MAP}} \equiv \underset{\boldsymbol{\pi} \,\in\, \boldsymbol{\Delta}_k}{\text{argmax}}\; p(\boldsymbol{\pi} \mid \mathbf{x}_{1:n})$$

Where the probability simplex is defined as

$$\boldsymbol{\Delta}_k = \left\{ \boldsymbol{\pi} \in \mathbb{R}^k \;\middle|\; \pi_j \in [0,1] \text{ and } \sum_{j=1}^{k} \pi_j = 1 \right\}$$

To satisfy the constraint $g(\boldsymbol{\pi}) = 1 - \sum_{j=1}^{k} \pi_j$, we use the Lagrange multiplier λ s.t. the optimisation of the log posterior becomes

$$\hat{\boldsymbol{\pi}}_{\text{MAP}} = \underset{(\boldsymbol{\pi}, \lambda) \,\in\, \mathbb{R}^{k+1}}{\text{argmax}} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} x_j^{(i)} + \alpha_j - 1 \right) \log \pi_j + \lambda g(\boldsymbol{\pi})$$

Here we ignore the normalizing constants which become additive constants in the log posterior optimization problem. The solution is found where

$$\begin{aligned} \nabla_{\boldsymbol{\pi}} \log p(\boldsymbol{\pi} \mid \mathbf{x}_{1:n}) + \lambda g(\boldsymbol{\pi}) &= 0 \\ g(\boldsymbol{\pi}) &= 0 \end{aligned}$$

The first condition yields

$$[\nabla_{\boldsymbol{\pi}} \log p(\boldsymbol{\pi} \mid \mathbf{x}_{1:n}) + \lambda g(\boldsymbol{\pi})]_\ell \Big|_{\substack{\pi_\ell = \pi_\ell^* \\ \lambda = \lambda^*}} = 0 \implies \frac{\sum_{i=1}^{n} x_\ell^{(i)} + \alpha_\ell - 1}{\pi_\ell^*} = \lambda^*$$

Replacing this result in the second condition, we get

$$1 - \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} x_j^{(i)} + \alpha_j - 1}{\lambda^*} = 0 \implies \lambda^* = n + \sum_{j=1}^{k} \alpha_j - 1$$

Where we swapped the sum over the $x_j^{(i)}$ and used the fact that $\mathbf{x}_j$ are one hot vectors. Thus

$$\boxed{(\hat{\boldsymbol{\pi}}_{\text{MAP}})_j = \pi_j^* = \frac{\sum_{i=1}^{n} x_j^{(i)} + \alpha_j - 1}{n + \alpha_j - 1} \;\in\; [0,1]}$$

3 Properties of estimators

a) Poisson

Let n trials $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$ where $\lambda = \mathbb{E}_x[x]$. The pmf of the Poisson is

$$p(x \mid \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \; \forall \; x \in \mathbb{N}$$

Such that the pmf of n trials should be

$$p(x_{1:n} \mid \lambda) \propto \prod_{j=1}^{n} p(x_j \mid \lambda)$$

I MLE

Using the log likelihood, we define the MLE estimation of λ as

$$\hat{\lambda}_{\text{MLE}} = \underset{\lambda \,\in\, \mathbb{R}_{>0}}{\text{argmax}} \sum_{j=1}^{n} (x_j \log \lambda - \lambda)$$

Which is found where

$$\nabla_\lambda \log p(x_{1:n} \mid \lambda)\Big|_{\lambda=\lambda^*} = 0$$

That is

$$\nabla_\lambda \log p(x_{1:n} \mid \lambda) = \lambda^{-1} \sum_{j=1}^{n} (x_j - 1)$$

Thus

$$\boxed{\hat{\lambda}_{\text{MLE}} = \frac{1}{n} \sum_{j=1}^{n} x_j}$$

II Bias

The bias is defined as

$$\text{Bias}(\lambda, \hat{\lambda}_{\text{MLE}}) \equiv \mathbb{E}_x[\hat{\lambda}_{\text{MLE}}] - \lambda$$

The expectation value of the MLE estimator is

$$\begin{aligned}
\mathbb{E}_x[\hat{\lambda}_{\text{MLE}}] &= \mathbb{E}_x\left[\frac{1}{n}\sum_{j=1}^{n} x'_j\right] \\
&= \frac{1}{n}\sum_{j=1}^{k} \mathbb{E}_x[x_j] \\
&= \lambda
\end{aligned}$$

Therefore the **MLE estimator of a Poisson distribution is an unbiased estimator**.

III Variance

The variance of the estimator is

$$\text{Var}(\hat{\lambda}_{\text{MLE}}) \equiv \mathbb{E}_X[\hat{\lambda}^2_{\text{MLE}}] - \mathbb{E}^2_x[\hat{\lambda}_{\text{MLE}}]$$

We need to evaluate the first term. To do this, we first use the Multinomial theorem to expand $\hat{\lambda}^2_{\text{MLE}}$:

$$\hat{\lambda}^2_{\text{MLE}} = \frac{1}{n^2} \sum_{k_1+\dots+k_n=2} \binom{2}{k_1, \dots, k_n} x_1^{k_1} \dots x_n^{k_n}$$

Then we use both the linearity of the expectation operator and the fact that the R.V. $X_1, \dots, X_n$ are independent to factorize the expectation of a "cross" product

$$\mathbb{E}_x[X_i X_j] = \mathbb{E}_x[X_i]\,\mathbb{E}_x[X_j], \quad \forall \;\; i, j \;\; \{\text{iid}\}$$

to get

$$\mathbb{E}_x[\hat{\lambda}^2_{\text{MLE}}] = \frac{1}{n^2} \sum_{k_1+\cdots+k_n=2} \binom{2}{k_{1:n}} \mathbb{E}_x[x_1^{k_1}] \dots \mathbb{E}_x[x_n^{k_n}]$$

The sum can be separated into quadratic and linear term, s.t.

$$\mathbb{E}_x[\hat{\lambda}^2_{\text{MLE}}] = \frac{1}{n^2}\left(n\,\mathbb{E}_x[x^2] + 2\binom{n}{2}\lambda^2 \right)$$

We used the fact that $\mathbb{E}_x[1] = 1$ and $\mathbb{E}_x[x_j] = \lambda$, $\forall j$. To estimate the quadratic term, we can use a magic trick by adding zero inside the operator argument. Using its linear property

$$\mathbb{E}_x[x^2] = \mathbb{E}_x[x(x-1) + x] = \mathbb{E}_x[x(x-1)] + \lambda$$

It turns out that

$$\begin{aligned}
\mathbb{E}_x[x(x-1)] &= \sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!}e^{-\lambda} \\
&= e^{-\lambda}\sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty} \frac{\lambda^{x+2}}{x!} \\
&= \lambda^2
\end{aligned}$$

By noticing the sum is the Taylor series of e^{λ}. In the end, we get

$$\mathbb{E}_x[\hat{\lambda}^2_{\text{MLE}}] = \frac{1}{n^2}\left(n\lambda + n^2\lambda^2\right)$$

Where we expanded the binom coefficient $2\binom{n}{2} = n(n-1)$. The variance is thus

$$\boxed{\text{Var}(\hat{\lambda}_{\text{MLE}}) = \frac{\lambda}{n}}$$

IV Consistency

As $n \to \infty$, the estimator give an unbiased estimate of λ with a variance that goes to 0. Thus, the **estimator is consistent**.

b) Bernoulli

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli(p)}$ and let $n > 10$. We consider the estimator

$$\hat{p} \equiv \frac{1}{10}\sum_{i=1}^{10} X_i$$

I Bias

We first note that the expected value of a Bernoulli is

$$\mathbb{E}_x[x] = p$$

Since $x \in \{0, 1\}$ and p is the probability that $X = 1$. Therefore,

$$\text{Bias}(p, \hat{p}) = \frac{1}{10}\sum_{j=1}^{10}\mathbb{E}_x[x_j] - p = 0$$

$\hat{p}$ is an unbiased estimator.

II Variance

The variance is

$$\begin{aligned}
\mathrm{Var}(\hat{p}) &= \mathbb{E}_x[\hat{p}^2] - \mathbb{E}_x^2[\hat{p}] \\
&= \frac{1}{100}\big(10\,\mathbb{E}_x[x^2] + 90\,\mathbb{E}_x^2[x]\big) - p^2 \\
&= \frac{p}{10} + p^2\left(\frac{90}{100} - 1\right) \\
&= \frac{1}{10}(p - p^2)
\end{aligned}$$

III Consistency

This estimator is not consistent since the variance is constant as $n \to \infty$.

c) Uniform

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathrm{Uniform}(0, \theta)$. The pdf of this distribution is

$$p(x_i \mid \theta) = \begin{cases} \theta^{-1}, & x \in [0, \theta] \\ 0, & \text{otherwise} \end{cases} = \frac{1}{\theta}\mathbf{1}_{\{0 \le x_i \le \theta\}}$$

for $\theta \in \mathbb{R}_{>0}$. We used the indicator function

$$\mathbf{1}_A(x) \equiv \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

I MLE

Given n samples, we want an estimator of the maximum possible value of $\mathbf{X}$. The MLE is

$$\hat{\theta}_{\text{MLE}} = \underset{\theta\,\in\,\mathbb{R}_{>0}}{\operatorname{argmax}} \frac{1}{\theta^n}\prod_{i=1}^n \mathbf{1}_{\{0 \le x_i \le \theta\}}$$

Where we used the fact that the data is iid. We can see that the product only depends on the boundary cases of the dataset, that is

$$\hat{\theta}_{\text{MLE}} = \underset{\theta\,\in\,\mathbb{R}_{>0}}{\operatorname{argmax}} \frac{1}{\theta^n}\mathbf{1}_{\{0 \le \min \mathbf{X}\}}\mathbf{1}_{\{\max \mathbf{X} \le \theta\}}$$

One can see that θ should be as low as possible to maximize θ^{-n}, yet not too low s.t. it make the second indicator function 0. The obvious choice is therefore

$$\boxed{\hat{\theta}_{\text{MLE}} = \max \mathbf{X}}$$

We show that $T(\theta) = \hat{\theta}_{\text{MLE}}$ is a sufficient statistic. To show this, we user the Fisher-Neyman theorem which garantees that the statistic is sufficient if the probability density can be factorized as $p(\mathbf{X}) = h(\mathbf{X})g(\theta, T(\theta))$. First, we use the fact that the data is iid:

$$p(\mathbf{X}) = \prod_{i=1}^n p(x_i)$$

Then replacing by the Uniform distribution

$$p(\mathbf{X}) = \prod_{i=1}^n \frac{1}{\theta}\mathbf{1}_{\{0 \le x_i \le \theta\}}$$

For the probability to be non-zero, only the boundary cases are important. That is

$$p(\mathbf{X}) = \frac{1}{\theta^n}\mathbf{1}_{\{0 \le \min \mathbf{X}\}}\mathbf{1}_{\{\max \mathbf{X} \le \theta\}} = h(\mathbf{X})\frac{1}{\theta^n}\mathbf{1}_{\{T(\theta) \le \theta\}}$$

We identify $h(\mathbf{X}) = \mathbf{1}_{\{0 \le \min \mathbf{X}\}}$ and the rest with the function $g(\theta, T(\theta))$. Thus we have shown $T(\theta)$ is a sufficient statistic by the Fisher-Neyman theorem.

II Bias

The bias of this estimator is

$$\text{Bias}(\theta, \hat{\theta}_{\text{MLE}}) = \mathbb{E}_x[\hat{\theta}_{\text{MLE}}] - \theta = \frac{1}{\theta}\int_0^\theta \max \mathbf{X} dx - \theta$$

Integrating $\max \mathbf{X}$ over all possible values, we get (in absolute value)

$$\boxed{\left\|\text{Bias}(\theta, \hat{\theta}_{\text{MLE}})\right\|_2 = \frac{\theta}{2}}$$

III Variance

The variance is

$$\text{Var}(\hat{\theta}_{\text{MLE}}) = \mathbb{E}_x[\hat{\theta}^2_{\text{MLE}}] - \mathbb{E}^2_x[\hat{\theta}_{\text{MLE}}]$$

Computing the integral for all the possible values for $(\max \mathbf{X})^2$, we get

$$\boxed{\text{Var}(\hat{\theta}_{\text{MLE}}) = \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}}$$

IV Consistency

This estimator is not consistent.

d) Gaussian

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2),\ \ \sigma, \mu \in \mathbb{R}$. We define the mean as $\bar{X} = \frac{1}{n}\sum_{i=1}^n x_i$.

I MLE

The likelihood, using the fact that the data is iid, is

$$p(\mathbf{X} \mid \mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n}\prod_{i=1}^n \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

The MLE for μ and σ^2 can be derived from the log likelihood

$$\hat{\theta}_{\text{MLE}} = (\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}}) = \underset{(\mu,\sigma^2)\,\in\,\mathbb{R}^2}{\text{argmax}} \ -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^n \frac{(x_i-\mu)^2}{\sigma^2}$$

The solution is found where

$$\nabla_\theta \log p(\mathbf{X} \mid \theta) = 0$$

That is

$$\partial_\mu \implies \sum_{i=1}^n (x_i - \hat{\mu}) = 0$$

From which we find

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{X}$$

Also, we have

$$\partial_{\sigma^2} \implies -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2}\sum_{i=1}^n \frac{(x_i-\bar{X})^2}{\hat{\sigma}^4} = 0$$

Thus

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{X})^2$$

II Bias

The bias for the normal variance MLE $\hat{\sigma}^2_{\text{MLE}}$ is

$$\text{Bias}(\sigma^2, \hat{\sigma}^2_{\text{MLE}}) = \mathbb{E}_x[\hat{\sigma}^2_{\text{MLE}}] - \sigma^2$$

Where

$$\mathbb{E}_x[\hat{\sigma}^2_{\text{MLE}}] = \frac{1}{n\sigma\sqrt{2\pi}} \sum_{i=1}^{n} \int_{-\infty}^{\infty} (x_i - \bar{X})^2 \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx$$

We used the linearity of the integral to swap the sum operator. We define some useful properties of the gaussian integral which we do not derive. They correspond to the non-central moments of the gaussian.

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) &= \varphi(x) \\
\int_{-\infty}^{\infty} \varphi\left(\frac{x-\mu}{\sigma}\right) dx &= \sigma \\
\int_{-\infty}^{\infty} x\varphi\left(\frac{x-\mu}{\sigma}\right) dx &= \mu\sigma \\
\int_{-\infty}^{\infty} x^2\varphi\left(\frac{x-\mu}{\sigma^2}\right) dx &= \sigma\mu^2 + \sigma^3
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}_x[\hat{\sigma}^2_{\text{MLE}}] &= \frac{1}{n\sigma} \sum_{i=1}^{n} \int_{-\infty}^{\infty} (x_i^2 - 2x_i\bar{X} + \bar{X}^2)\varphi\left(\frac{x-\mu}{\sigma}\right) dx \\
&= \frac{1}{n\sigma} \sum_{i=1}^{n} \left(\sigma\mu^2 + \sigma^3 - 2\bar{X}\mu\sigma + \bar{X}\sigma\right) \\
&= \mu^2 + \sigma^2 - 2\bar{X}\mu + \bar{X}
\end{aligned}$$

Thus,

$$\boxed{\text{Bias}(\sigma, \hat{\sigma}^2_{\text{MLE}}) = (\mu - \bar{X})^2}$$

III Variance

The $\hat{\sigma}^2_{\text{MLE}}$ variance is

$$\text{Var}(\hat{\sigma}^2_{\text{MLE}}) = \mathbb{E}_x[\hat{\sigma}^4_{\text{MLE}}] - \mathbb{E}^2_x[\hat{\sigma}^2_{\text{MLE}}]$$

Using the gaussian integral properties, we can solve this system using only the operators $\mathbb{E}_x[x] = \mu$ and $\mathbb{E}_x[x^2] = \mu^2 + \sigma^2$.