$Cov(Y, Y+Z) = E[Y(Y+Z)] - E[Y]E[Y+Z]$

$= E[Y^2] + E[YZ] - E[Y](E[Y] + E[Z])$

$= \underbrace{E[Y^2] - E[Y]^2}_{Cov(Y,Y)} + \underbrace{E[YZ] - E[Y]E[Z]}_{Cov(Y,Z)}$

Jointly Gaussian random variables/vectors

If $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ are jointly Gaussian random vectors then $X|Y = y$ is also a Gaussian random vector with

- (Conditional) mean vector $\mu_{X|y}$:

$$\begin{aligned}\mu_{X|y} &= E[X|Y=y] = E[X] + \mathrm{Cov}(X,Y)(\mathrm{Cov}(Y))^{-1}(y - E[Y]) \\ &= \mu_X + \Sigma_{XY}\Sigma_Y^{-1}(y - \mu_Y).\end{aligned}$$

- (Conditional) covariance matrix $\Sigma_{X|y}$:

$$\begin{aligned}\Sigma_{X|y} &= \mathrm{Cov}(X - \mu_{X|y}|Y=y) \\ &= \mathrm{Cov}(X) - \mathrm{Cov}(X,Y)(\mathrm{Cov}(Y))^{-1}\mathrm{Cov}(Y,X) \\ &= \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}.\end{aligned}$$

Note: $\mu_{X|y}$ depends on the value of y but $\Sigma_{X|y}$ does not.

Boston University
Department of Electrical and Computer Engineering

ENG EC 503 (Ishwar) Learning from Data

Assignment 2 Solution

● Spring 2017 Prakash Ishwar

Issued: Tue 31 Jan 2017 **Due:** 5pm Wed 8 Jan 2017 in box outside PHO440
Required reading: Your notes from lectures and slides on website.

Problem 2.1 *(MMSE, MMAE, MAP, and ML decision rules)* Let $X = Y + Z$ where Y and Z are independent Gaussians with zero means and variances σ_Y^2 and σ_Z^2 respectively which are *known*. *Derive* the MMSE, MMAE, MAP, and ML decision rules for estimating Y for a given test point $X = x$. The expressions will be functions of x, σ_Y^2 and σ_Z^2. For the MMSE estimate, also *derive* an expression for the Bayes risk $E[(Y - h_{MMSE}(X))^2]$.

Solution: Since Y and Z are independent Gaussians and X is their sum, all three are jointly Gaussian. Thus, $Y|X = x \sim \mathcal{N}(E[Y|X=x], \mathrm{Cov}(Y|X=x))$ where,

Conditional mean

$$E[Y|X=x] = E[Y] + \mathrm{Cov}(Y,X)\mathrm{Cov}^{-1}(X)(x - E[X]) = 0 + \mathrm{Cov}(Y, Y+Z)\mathrm{Cov}(Y+Z, Y+Z)x = \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_Z^2}x$$

$\mathrm{Cov}(Y,Y) + \mathrm{Cov}(Y,Z)$; $\mathrm{Cov}(Y,Y) = \sigma_Y^2$; $\mathrm{Cov}(Y,Z) = 0$: independent ⇒ uncorrelated

since $\mathrm{Cov}(Y,Z) = 0$ and

Conditional Covariance

$$\mathrm{Cov}(Y|X=x) = \mathrm{Cov}(Y) - \mathrm{Cov}(Y,X)\mathrm{Cov}^{-1}(X)\mathrm{Cov}(X,Y) = \frac{\sigma_Y^2\sigma_Z^2}{\sigma_Y^2 + \sigma_Z^2}.$$

Thus,

$$Y|X = x \sim \mathcal{N}\left(\frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_Z^2}x, \frac{\sigma_Y^2\sigma_Z^2}{\sigma_Y^2 + \sigma_Z^2}\right).$$

The posterior pdf of Y conditioned on $X = x$ is a scalar Gaussian which is symmetric about the posterior mean $E[Y|X=x]$ and also reaches its maximum value there.

Thus the posterior mean, median, and mode all coincide and the MMSE, MMAE, and MAP decision rules are all identical and given by:

$$h_{MMSE}(x) = h_{MMAE}(x) = h_{MAP}(x) = E[Y|X=x] = \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_Z^2}x.$$

The associated MSE (Bayes risk for squared loss) is given by

doesn't depend on X

$$E[(Y - h_{MMSE}(X))^2] = E[E[(Y - E[Y|X])^2|X]] = E[\mathrm{Cov}(Y|X)] = E\left[\frac{\sigma_Y^2\sigma_Z^2}{\sigma_Y^2 + \sigma_Z^2}\right] = \frac{\sigma_Y^2\sigma_Z^2}{\sigma_Y^2 + \sigma_Z^2}.$$

To derive the ML decision rule, first note that likelihood $p(x|y) = \mathcal{N}(y, \sigma_Z^2)(x)$, i.e., a Gaussian with mean at y and variance σ_Z^2. The Gaussian attains its unique maximum value at the location of its mean which will be at x only if $y = x$. Thus $h_{ML}(x) = x$. Although not asked, the MSE of the ML rule is given by $E[(X - Y)^2] = E[Z^2] = \sigma_Z^2$.

Because $X = Y + Z$

If X & Y are Gaussians:

$$\begin{pmatrix}X\\Y\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \begin{pmatrix}\Sigma_{XX} & \Sigma_{XY}\\\Sigma_{YX} & \Sigma_{YY}\end{pmatrix}\right)$$

known that: $X|Y \sim \mathcal{N}\left(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y), \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\right)$

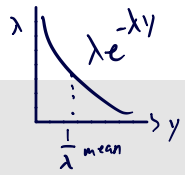
So, conditional covariance matrix $\mathrm{Var}(X|Y) = \Sigma_X - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$ is constant and doesn't depend on Y. Therefore, $E[\mathrm{Var}(X|Y)] = \mathrm{Var}(X|Y)$

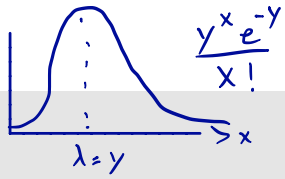
Problem 2.2 *(MMSE decision rule for light intensity)* The intensity Y of a light source is exponentially distributed with a *known* mean $1/\lambda$. Given $Y = y$, let $X_1, \ldots, X_n \sim$ IID Poisson(y) be n IID photon-count measurements of the light source.

(a) *Derive* the MMSE estimate of Y given $X_1 = x_1, \ldots, X_n = x_n$, as a function of $\lambda, n, x_1, \ldots, x_n$.

(b) *Determine* the limit to which the MMSE estimate converges to as $\lambda \to 0$. *Compare* with the ML estimate of Y given $X_1 = x_1, \ldots, X_n = x_n$.

Useful result: $\forall a \in (0, \infty), \int_0^\infty t^k e^{-at} dt = \frac{k!}{a^{k+1}}$.





$\exp(\lambda) = \lambda e^{-\lambda y}$, $y \geq 0$

light source intensity

Photon Count given $Y = y$

$\Rightarrow \lambda = y$, Poisson(y)

Solution:

(a) We have

$$p(y) = \lambda e^{-\lambda y} 1(y \geq 0), \quad p(x_j|y) = \frac{y^{x_j}}{x_j!} e^{-y}, \ j = 1, \ldots, n.$$

Thus,

$$\begin{aligned} p(y|x_1, \ldots, x_n) &\propto p(y) p(x_1, \ldots, x_n|y) \quad \text{(Proportional)} \\ &= p(y) \prod_{j=1}^n p(x_j|y) \\ &= \lambda e^{-\lambda y} 1(y \geq 0) \prod_{j=1}^n \frac{y^{x_j}}{x_j!} e^{-y} \\ &\propto e^{-\lambda y} y^{(\sum_{j=1}^n x_j)} e^{-ny} 1(y \geq 0) \\ &= e^{-(\lambda+n)y} y^{(\sum_{j=1}^n x_j)} 1(y \geq 0). \end{aligned}$$

Want MMSE estimate $Y|X$

$h_{MMSE} = E[Y|X = x] = \int_0^\infty y \, p(y|x) \, dy$ (Posterior)

$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$

$p(x|y) = \text{Poisson}(y) = \frac{y^x e^{-y}}{x!}$, $y \geq 0$

$p(y) = \lambda e^{-\lambda y}$

Therefore,

$$\begin{aligned} h_{MMSE}(x_1, \ldots, x_n) &= E[Y|X_1 = x_1, \ldots, X_n = x_n] \\ &= \frac{\int_0^\infty y e^{-y(\lambda+n)} y^{(\sum_{j=1}^n x_j)} dy}{\int_0^\infty e^{-y(\lambda+n)} y^{(\sum_{j=1}^n x_j)} dy} \\ &= \frac{\int_0^\infty e^{-y(\lambda+n)} y^{(1+\sum_{j=1}^n x_j)} dy}{\int_0^\infty e^{-y(\lambda+n)} y^{(\sum_{j=1}^n x_j)} dy} \\ &\overset{\text{hint}}{=} \frac{(1 + \sum_{j=1}^n x_j)!}{(\lambda+n)^{(1+1+\sum_{j=1}^n x_j)}} \cdot \frac{(\lambda+n)^{(1+\sum_{j=1}^n x_j)}}{(\sum_{j=1}^n x_j)!} \\ &= \frac{(1 + \sum_{j=1}^n x_j)}{(\lambda+n)}. \end{aligned}$$

$p(x) = \sum_{j=1}^n p(x|y_j) p(y_j)$

$p(y|x = x_1, \cdots x_n) = \frac{p(x_1, \cdots x_n|y) p(y)}{p(x_1, \cdots x_n)}$

$= \left(\frac{y^{\sum x_j} e^{-y}}{\sum x_j!} \right) \left(\lambda e^{-\lambda y} \right)$ $(y \geq 0)$

(b) For convenience, define $s_n := \sum_{j=1}^n x_j$. Then as $\lambda \to 0$, $h_{MMSE}(x_1, \ldots, x_n) = \frac{1+s_n}{\lambda+n} \to \frac{1+s_n}{n}$. The ML estimate of Y is given by

$$\begin{aligned} h_{ML}(x_1, \ldots, x_n) &= \arg\max_y p(x_1, \ldots, x_n|y) \\ &= \arg\max_y e^{-ny} y^{s_n} \quad \text{drop denominator scaling term} \\ &= \arg\max_y (-ny + s_n \ln y). \end{aligned}$$

If we define $g(y) := (-ny + s_n \ln y)$ then $g'(y) = -n + s_n/y$ and $g''(y) = -s_n/y^2 \leq 0$. Thus $g(y)$ attains a unique global maximum when $g'(y) = 0$. By solving this we get

$$h_{ML}(x_1, \ldots, x_n) = \frac{s_n}{n} = \frac{\sum_{j=1}^{n} x_j}{n}$$

which differs from the MMSE estimate by an additive term $1/n$ which disappears when $n \to \infty$. Note that as λ goes to zero, the prior on Y becomes essentially uniform over the entire positive real axis and its effect on the optimum estimate diminishes. Although the MMSE estimate is not the same as the MAP estimate, the diminished effect of the prior is similar in spirit to the result that MAP reduces to ML for a uniform prior.

Problem 2.3 *(Soft Thresholding decision rule)* Let $X = Y + Z$ where $Y \perp\!\!\!\perp Z$ (i.e., Y and Z are independent), $Z \sim \mathcal{N}(0, \sigma^2)$, $\sigma > 0$, and $Y \sim p(y) = 0.5\lambda \exp(-\lambda|y|)$, $\lambda > 0$ (Note: $p(y)$ is the prior pdf of Y). *Derive* $h_{\text{MAP}}(x)$, the MAP estimate of Y based on $X = x$.

Maximum Aposteriori Probability (MAP) Decision Rule

- $h_{\text{MAP}}(\mathbf{x})$ = posterior mode = $\arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x})$
- Alternative forms:

$$\begin{aligned} h_{\text{MAP}}(\mathbf{x}) &= \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}) \\ &= \arg\max_{y \in \mathcal{Y}} p(\mathbf{x}, y) \\ &= \arg\max_{y \in \mathcal{Y}} p(\mathbf{x}|y)p(y) \\ &= \arg\min_{y \in \mathcal{Y}} -\ln p(\mathbf{x}|y) - \ln p(y) \end{aligned}$$

Solution: *Solution-1:*

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ \delta(x - \mu) \end{cases}$$

$$p(x, y) = \frac{\lambda}{2} e^{-\lambda|y|} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-y)^2}$$

$$\Rightarrow h_{\text{MAP}}(x) = \arg\min_y \left[-\ln p(x, y)\right]$$

$$= \arg\min_y \underbrace{\left[(x - y)^2 + 2\lambda\sigma^2|y|\right]}_{\psi(y)}$$

$\psi(y)$ is a continuous function that is differentiable everywhere except at $y = 0$ where it has non-matching left and right derivatives. We have:

$$\psi'(y) = \frac{d}{dy}\psi(y) = \begin{cases} 2\left[(y - x) + \lambda\sigma^2 \text{sign}(y)\right] & \ldots y \neq 0 \\ \text{not differentiable} & \ldots y = 0. \end{cases}$$

$$= \begin{cases} 2\left(y - x - \lambda\sigma^2\right) & \ldots y < 0 \\ 2\left(y - x + \lambda\sigma^2\right) & \ldots 0 < y \\ \text{not differentiable} & \ldots y = 0 \end{cases}$$

Case-1: $|x| \leq \lambda\sigma^2$

$$\psi'(y) = \begin{cases} < 0 & y < 0 \\ > 0 & 0 < y \\ \text{not differentiable} & y = 0. \end{cases}$$

Thus, ψ is strictly decreasing for $y < 0$ and strictly increasing for $y > 0$. Therefore, ψ is minimized at $y = 0$.
Conclusion:

$$h_{\text{MAP}}(x) = 0 \text{ if } |x| \leq \lambda\sigma^2.$$

Case-2: $x > \lambda\sigma^2$

$$\psi'(y) = \begin{cases} < 0 & y < 0 \\ \text{not differentiable} & y = 0 \\ < 0 & 0 < y < x - \lambda\sigma^2 \\ \geq 0 & x - \lambda\sigma^2 \leq y. \end{cases}$$

Thus, ψ is decreasing for $y < x - \lambda\sigma^2$ and increasing for $x - \lambda\sigma^2 \leq y$. Therefore, ψ is minimized at $y = x - \lambda\sigma^2$.

Conclusion:

$$h_{\text{MAP}}(x) = x - \lambda\sigma^2 \text{ if } x > \lambda\sigma^2.$$

Case-3: $x < -\lambda\sigma^2$

$$\psi'(y) = \begin{cases} < 0 & y \leq x + \lambda\sigma^2 \\ > 0 & x + \lambda\sigma^2 < y < 0 \\ \text{not differentiable} & y = 0 \\ > 0 & x - \lambda\sigma^2 < y. \end{cases}$$

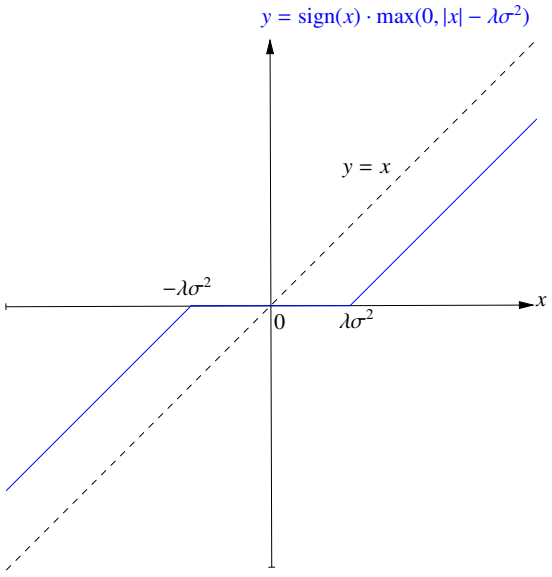
Thus, ψ is decreasing for $y < x + \lambda\sigma^2$ and increasing for $x + \lambda\sigma^2 \leq y$. Therefore, ψ is minimized at $y = x + \lambda\sigma^2$.

Conclusion:

$$h_{\text{MAP}}(x) = x + \lambda\sigma^2 \text{ if } x < -\lambda\sigma^2.$$

Combining all cases:

$$\begin{aligned} h_{\text{MAP}}(x) &= \begin{cases} 0 & \text{if } |x| \leq \lambda\sigma^2 \\ x - \lambda\sigma^2 & \text{if } \lambda\sigma^2 < x \\ x + \lambda\sigma^2 & \text{if } x < -\lambda\sigma^2 \end{cases} \\ &= \begin{cases} 0 & \text{if } |x| \leq \lambda\sigma^2 \\ x - \text{sign}(x)\lambda\sigma^2 & \text{if } |x| > \lambda\sigma^2 \end{cases} \\ &= \text{sign}(x) \cdot \max(0, |x| - \lambda\sigma^2). \end{aligned}$$



Solution: *Solution-2:* Consider the following broad class of priors $p(y) \propto e^{-b(y)}$ where $b(y)$ is symmetric, nondecreasing for $y \geq 0$, continuous, continuously differentiable everywhere *except perhaps at* $y = 0$, and $b'(0+) > 0$ (these properties are satisfied for the Laplacian prior with $b(y) = \lambda|y|$). We will show that the MAP estimator for such a prior is (i) antisymmetric, (ii) a *shrinkage function*, i.e., for all x, $|h_{MAP}(x)| \leq |x|$, and (iii) exhibits a *threshold effect*, i.e., there exists a threshold $\tau > 0$ such that $h_{MAP}(x) = 0$ for all $|x| < \tau$.

The MAP estimator minimizes the negative log posterior:

$$\begin{aligned} h_{MAP}(x) &= \arg\min_y[-\ln p(y|x)] \\ &= \arg\min_y[-\ln p(x|y) - \ln p(y)] \\ &= \arg\min_y \left[\frac{(x-y)^2}{2} + \sigma^2 b(y)\right] \\ &= \arg\min_y \psi(y) \end{aligned}$$

where

$$\psi(y) := \left[\frac{(x-y)^2}{2} + \sigma^2 b(y)\right].$$

Since $b(y)$ is symmetric, it follows that $h_{MAP}(x)$ is an antisymmetric function of x. Since $b(y)$ is symmetric and nondecreasing in y for $y \geq 0$, it follows that if $x \geq 0$ then $\psi(y)$ is nonincreasing for all $y < 0$ and nondecreasing for all $y > x$ and if $x \leq 0$ then $\psi(y)$ is nonincreasing for all $y < x$ and nondecreasing for all $y > 0$. Thus if $\psi(y)$ is minimized at y_0 then $|y_0| \leq |x|$. Hence, the MAP estimate is *shrunk towards zero*, i.e., the MAP estimator is a shrinkage function. For all $y \neq 0$,

$$\psi'(y) = y - x + \sigma^2 b'(y)$$

If $\psi(y)$ is minimized at $y_0 \neq 0$ then y_0 must be a root of $\psi'(y)$, i.e., $\psi'(y_0) = 0$. Such a nonzero y_0 must satisfy the (possibly) nonlinear equation

$$x = y_0 + \sigma^2 b'(y_0). \tag{1}$$

Since y and $b'(y)$ have the same sign, it follows that y_0 and x also have the same sign. Since y and $b'(y)$ are continuous functions of y for all $y \neq 0$ and $b'(0+) > 0$, it follows that there must exist a nonempty neighborhood $(-\tau, \tau)$ of $x = 0$ for which (1) has no solution. In this case, $\psi(y)$ is minimized at $y = 0$. Thus for all $|x| < \tau$, the only possible minimizer is $h_{MAP}(x) = 0$ and the shrinkage function exhibits a *threshold effect* with threshold which is at least equal to τ.

For the Laplacian prior, $b(y) = \lambda|y|$, which is not differentiable at zero, $b'(0+) = \lambda$, and for all nonzero y, $b'(y) = \lambda\text{sign}(y)$. If we define the *soft-threshold* $\tau_0 := \lambda\sigma^2$ then

$$\begin{aligned} x &= y_0 + \tau_0\text{sign}(y_0) \\ \Rightarrow |x|\text{sign}(x) &= y_0 + \tau_0\text{sign}(y_0) \\ \Rightarrow y_0 &= (|x| - \tau_0)\text{sign}(y_0) \end{aligned}$$

since $\text{sign}(y_0) = \text{sign}(x)$. This shows that if $|x| < \tau_0$ then $\psi'(y_0) = 0$ has no solution and the minimizer of $\psi(y)$ is $y = 0$. If, however, $|x| \geq \tau_0$, then $y_0 = (|x| - \tau_0)\text{sign}(x)$ is the root of $\psi'(y_0) = 0$ and also the minimizer of $\psi(y)$.

Thus for the Laplacian prior,

$$h_{MAP}(x) = \text{sign}(x)\max(0, |x| - \tau_0)$$

which is called the soft-thresholding rule.

Problem 2.4 *(ML learning of Categorical model parameters)* Let $x_1, \ldots, x_n$ be n IID training samples from a categorical distribution: $p(x|\theta) = \theta_x$ where $\theta = (\theta_1, \ldots, \theta_m)^\top$ is an *unknown* pmf over the first m positive integers. *Derive* the expression for $\hat{\theta}_{ML}(x_1, \ldots, x_n)$.
Useful result: For any two pmfs $p(x)$ and $q(x)$ over the same set, the scalar quantity

$$D(p\|q) := \sum_x p(x) \ln(p(x)/q(x)) \geq 0,$$

i.e., it is always nonnegative, and is equal to zero if, and only if, $p(x) = q(x)$ for all x, i.e., if the pmfs are identical. The nonnegative scalar quantity $D(p\|q)$ is called the Kullback-Liebler (KL) divergence or relative entropy of the pair of pmfs (p, q). This is an *asymmetric* measure of distance between pmfs and plays a fundamental role in Probability, Statistics, Machine Learning, and Information Theory. A related nonnegative scalar quantity is $\sum_x p(x) \ln(1/p(x))$ which is called the entropy of the pmf p.

Solution: For $i = 1, \ldots, m$, let $n_i := \sum_{j=1}^n 1(x_j = i)$ denote the number of training samples that are equal to i and $\hat{p}_i := n_i/n$ the corresponding fraction of training samples. Then noting that $n_i = n\hat{p}_i$ we have

$$\begin{aligned}
\hat{\theta}_{ML}(x_1, \ldots, x_n) &= \arg \max_{\text{pmfs } \theta} p(x_1, \ldots, x_n|\theta) \\
&= \arg \max_\theta \prod_{j=1}^n p(x_j|\theta) \\
&= \arg \max_\theta \prod_{j=1}^n \theta_{x_j} = \arg \max_\theta \prod_{i=1}^m (\theta_i)^{n_i} \\
&= \arg \max_\theta \sum_{i=1}^m n_i \ln \theta_i = \arg \max_\theta \sum_{i=1}^m \hat{p}_i \ln \theta_i \\
&= \arg \min_\theta \sum_{i=1}^m -\hat{p}_i \ln \theta_i = \arg \min_\theta \sum_{i=1}^m \hat{p}_i \ln \left(\frac{1}{\theta_i}\right) \\
&= \arg \min_\theta \left[\sum_{i=1}^m \hat{p}_i \ln \left(\frac{\hat{p}_i}{\theta_i}\right) + \underbrace{\sum_{i=1}^m \hat{p}_i \ln \left(\frac{1}{\hat{p}_i}\right)}_{\text{Entropy of } \hat{p}\text{; not a function of } \theta} \right] \\
&= \arg \min_\theta \sum_{i=1}^m \hat{p}_i \ln \left(\frac{\hat{p}_i}{\theta_i}\right) \\
&= \arg \min_\theta D(\hat{p}\|\theta) \\
&= \hat{p}
\end{aligned}$$

the empirical pmf of the categories. Thus, for $i = 1, \ldots, m$,

$$\hat{\theta}_{ML,i} = \hat{p}_i = \frac{n_i}{n} = \frac{\#i}{n} = \frac{1}{n} \sum_{j=1}^n 1(x_j = i).$$

Problem 2.5 *(ML learning of scalar Gaussian model parameters)* Let $\hat{\mu}_{ML}$ and $\widehat{\sigma^2}_{ML}$ denote the ML estimate of the mean and variance parameters of a scalar Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ based on observing n

IID training samples $x_1, \ldots, x_n$.

(a) *Derive* the expressions for $\hat{\mu}_{ML}$ and $\widehat{\sigma^2}_{ML}$ in terms of the training samples $x_1, \ldots, x_n$ and n. *Tip:* using an empirical random variable can reduce clutter.

(b) *Compute* the expected values $E[\hat{\mu}_{ML}]$ and $E[\widehat{\sigma^2}_{ML}]$ of the ML estimates with respect to the distribution on the training set. Based on your expressions, are the ML estimates of the mean and variance unbiased? If not, propose a slight modification to make them unbiased.

Solution:

(a) Let $\theta = (\mu, \sigma^2)$ and $\mathcal{D} := \{x_1, \ldots, x_n\}$. Then $p(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. For convenience, let us define $X_{\text{emp}} \sim \text{Uniform}\{x_1, \ldots, x_n\}$. Then,

$$E[X_{\text{emp}}] = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad \text{Var}(X_{\text{emp}}) = \frac{1}{n}\sum_{j=1}^{n}(x_j - E[X_{\text{emp}}])^2.$$

With these preliminaries,

$$\begin{aligned}
\hat{\theta}_{ML}(x_1, \ldots, x_n) &= \arg\max_{\theta} p(\mathcal{D}|\theta) \\
&= \arg\min_{\theta} \sum_{j=1}^{n} -\frac{1}{n} \ln p(x_j|\theta) \\
&= \arg\min_{\theta} E\left[-\ln p(X_{\text{emp}}|\theta)\right] \\
&= \arg\min_{\theta} E\left[\frac{(X_{\text{emp}} - \mu)^2}{2\sigma^2} + \frac{1}{2}\ln\sigma^2\right] \\
&= \arg\min_{\mu,\sigma^2} \left[\frac{1}{\sigma^2}\left(E[X_{\text{emp}}] - \mu\right)^2 + \frac{\text{Var}(X_{\text{emp}})}{\sigma^2} + \ln\sigma^2\right].
\end{aligned}$$

For any choice of σ^2 the expression inside the square brackets above is clearly minimized if

$$\mu = \hat{\mu}_{ML} = E[X_{\text{emp}}] = \frac{1}{n}\sum_{j=1}^{n} x_j.$$

With this choice for μ (which does not depend on the value of σ^2), we proceed to determine the optimum value for σ^2 next:

$$\begin{aligned}
\widehat{\sigma^2}_{ML} &= \arg\min_{\sigma^2}\left[\frac{\text{Var}(X_{\text{emp}})}{\sigma^2} + \ln\sigma^2\right], \quad \ldots \quad \text{setting } \sigma^2 = t \cdot \text{Var}(X_{\text{emp}}), t \geq 0, \\
&= \text{Var}(X_{\text{emp}}) \cdot \arg\min_{t\geq 0}\left[\frac{1}{t} + \ln t\right].
\end{aligned}$$

The function $g(t) = t^{-1} + \ln t$ has a unique global minimizer at $t = 1$ where its value equals $g(1) = 1$ since, $g'(t) = -t^{-2} + t^{-1} = 0 \Rightarrow t = 1$ and at $t = 1$, $g''(t) = 2(1)^{-3} - (1)^{-2} = 1 > 0$. Thus we conclude that

$$\hat{\mu}_{ML} = E[X_{\text{emp}}] = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad \widehat{\sigma^2}_{ML} = \text{Var}(X_{\text{emp}}) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{\mu}_{ML})^2.$$

(b) The ML estimate of the mean is unbiased since,

$$E[\hat{\mu}_{ML}] = E\left[\frac{1}{n}X_j\right] = \frac{1}{n}\sum_j E[X_j] = \mu.$$

The computation of $E[\widehat{\sigma^2}_{ML}]$ is a bit more involved because $\hat{\mu}_{ML}$ which appears inside it is correlated to each X_j. We have

$$E[\widehat{\sigma^2}_{ML}] = E\left[\frac{1}{n}\sum_{j=1}^{n}(X_j - \hat{\mu}_{ML})^2\right] = \frac{1}{n}\sum_{j=1}^{n} E[(X_j - \hat{\mu}_{ML})^2] = E[(X_1 - \hat{\mu}_{ML})^2],$$

where the last equality follows from the observation that all the n pairs $(X_1, \hat{\mu}_{ML}), \ldots, (X_n, \hat{\mu}_{ML})$ have the same joint distribution since $X_1, \ldots, X_n$ are IID and $\hat{\mu}_{ML}$ is invariant to any permutation of the n samples $X_1, \ldots, X_n$. Now,

$$\begin{aligned}
E[(X_1 - \hat{\mu}_{ML})^2] &= \frac{1}{n^2}E\left((n-1)X_1 - \sum_{j=2}^{n} X_j\right)^2 = \frac{1}{n^2}E\left[E\left((n-1)X_1 - \sum_{j=2}^{n} X_j\right)^2 \middle| X_1\right] \\
&= \frac{1}{n^2}E[((n-1)X_1 - (n-1)\mu)^2 + (n-1)\sigma^2] = \frac{1}{n^2}\left[(n-1)^2\sigma^2 + (n-1)\sigma^2\right] = \\
&= \frac{n-1}{n}\sigma^2 \\
&\neq \sigma^2.
\end{aligned}$$

Thus the ML estimate of the variance is not unbiased. It cannot be made unbiased when $n = 1$ (unless $\sigma = 0$) since then the estimate of variance is exactly zero. For $n > 1$, it can be made unbiased by scaling it by the factor $n/(n-1)$. Thus an unbiased estimate of the variance for $n > 1$ is given by

$$\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \hat{\mu}_{ML})^2.$$

Problem 2.6 *(ML learning of the support of a uniform distribution)*
$X_1, \ldots, X_n \sim$ IID Uniform($[0, \theta]$) where $\theta > 0$ is a deterministic unknown parameter.

(a) Derive the ML estimate of θ given $X_1 = x_1, \ldots, X_n = x_n$ as a function of $x_1, \ldots, x_n$.

(b) Show that the ML estimate is biased. Then explain how to modify it to make it unbiased.

Solution: $X_1, \ldots, X_n \sim$ IID pdf $p(x) = \frac{1}{\theta}1(x \leq \theta) \Rightarrow$ CDF $X_j : F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{\theta} & 0 \leq x \leq \theta \\ 1 & \text{otherwise.} \end{cases}$

(a)

$$\begin{aligned}\widehat{\theta}_{ML}(x_1,\ldots,x_n) &= \arg\max_{\theta>0}\prod_{j=1}^{n} p(x_j)\\ &= \arg\max_{\theta>0}\prod_{j=1}^{n}\left(\frac{1}{\theta}1(x_j\le\theta)\right)\end{aligned}$$

Now,

$$\prod_{j=1}^{n}\left(\frac{1}{\theta}1(x_j\le\theta)\right) = \begin{cases}0 & \theta < \max(x_1,\ldots,x_n)\\ \frac{1}{\theta^n} & \text{otherwise.}\end{cases}$$

Thus,

$$\begin{aligned}\widehat{\theta}_{ML}(x_1,\ldots,x_n) &= \arg\max_{\theta\ge\max(x_1,\ldots,x_n)}\frac{1}{\theta^n}\\ &= \max(x_1,\ldots,x_n)\end{aligned}$$

(b) Let $Z_n = \max(X_1,\ldots,X_n)$. It's CDF is given by:

$$\begin{aligned}F_{Z_n}(z) &= P(\max(X_1,\ldots,X_n)\le z)\\ &= P(X_1\le z,\ldots,X_n\le z)\\ &\overset{\text{I.I.D.}}{=} \prod_{j=1}^{n} P(X_j\le z)\\ &= \prod_{j=1}^{n} F(z)\\ &= (F(z))^n\,.\end{aligned}$$

Taking the derivative of the CDF, we get the pdf $f_{Z_n}(z) = n(F(z))^{n-1}p(z)$. The expected value of the ML estimate is then given by

$$\begin{aligned}E[\widehat{\theta}_{ML}(X_1,\ldots,X_n)] &= \int_{-\infty}^{\infty} z f_{Z_n}(z)dz\\ &= \int_{-\infty}^{\infty} z n(F(z))^{n-1}p(z)dz\\ &= \int_0^{\theta} z n\left(\frac{z}{\theta}\right)^{n-1}\frac{1}{\theta}dz\\ &= \frac{n}{\theta^n}\int_0^{\theta} z^n dz\\ &= \frac{n}{\theta^n}\frac{\theta^{n+1}}{(n+1)}\\ &= \frac{n}{n+1}\theta\\ &\neq \theta.\end{aligned}$$

Thus the expected value of the ML estimate of θ does not equal θ and is therefore biased. It can be made unbiased by scaling the ML esitmate by the factor $(n+1)/n$.