ESE 531: Statistical Learning and Inference

Homework 3: Numerical Optimization and Evaluation of Estimators

1. Consider a random sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p(x;\theta)$. Prove whether or not a unique global optimum exists for the likelihood under the following population distributions:

   (a) Gaussian population with unknown mean: $p(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$, where $-\infty < x < \infty$ and $-\infty < \mu < \infty$.

   Solution. To show that a unique global optimium exists, we need to show that the negative log-likelihood is convex. First, we note that for an i.i.d. random sample, the log-likelihood is additive, that is:

   $$\ell(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ell(\theta; x_i)$$

   Since the sum of convex functions is convex, we only need to show that $f(\theta) \triangleq -\log p(x;\theta)$ is convex. For this case, we have that:

   $$\begin{aligned} f(\mu) &= -\log p(x;\mu) \\ &= -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\right) \\ &= -\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2\right) \\ &= \frac{\log(2\pi)}{2} + \frac{\log(\sigma^2)}{2} + \frac{(x-\mu)^2}{2\sigma^2} \end{aligned}$$

   To show that $f(\mu)$ is convex, we need to show that the second derivative is positive for all μ. It can be shown that the second derivative is given by:

   $$\frac{\partial^2 f}{\partial \mu^2} = \frac{1}{\sigma^2} \geq 0,$$

   since $\sigma^2 > 0$. Therefore, $f(\mu)$ is convex and a unique maximum exists for the likelihood.

   (b) Gaussian population with unknown precision (inverse variance): $p(x;\alpha) = \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{\alpha}{2}(x-\mu)^2\right)$, where $-\infty < x < \infty$ and $\alpha > 0$. Note, replacing $\alpha = \frac{1}{\sigma^2}$ recovers the standard parameterization for a Gaussian distribution.

   Solution. We can use the result from the previous part to deduce that showing the negative log-likelihood is convex for this population distribution is equivalent to showing that the following function is convex:

   $$f(\alpha) \triangleq \frac{\log(2\pi)}{2} - \frac{\log(\alpha)}{2} + \frac{\alpha(x-\mu)^2}{2}$$

   It can be shown that second derivative of $f(\alpha)$ is given by:

   $$\frac{\partial^2 f}{\partial \alpha^2} = \frac{1}{2\alpha^2} \geq 0$$

   since $\alpha > 0$. Thus, $f(\alpha)$ is convex and a unique maximum exists for the likelihood function.

   (c) Exponential population: $p(x;\lambda) = \lambda e^{-\lambda x}$, where $x > 0$ and $\lambda > 0$.

   Solution. Following the previous parts, in this part we need to show that:

   $$f(\lambda) \triangleq -\log(\lambda) + \lambda x$$

   is a convex function. The second derivative of $f(\lambda)$ can be shown to be:

   $$\frac{\partial^2 f}{\partial \alpha^2} = \frac{1}{\lambda^2} \geq 0$$

   since $\lambda > 0$. Thus, $f(\lambda)$ is convex and a unique maximum exists for the likelihood function.

(d) Gamma population: $p(x;\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} \exp\left(\frac{-x}{\theta}\right)$, where $x > 0$ and $\theta > 0$.

Solution. Following the previous parts, in this part we need to show that:

$$f(\theta) \triangleq -\log\left(\frac{x^{k-1}}{\Gamma(k)}\right) + k\log(\theta) + \frac{x}{\theta}$$

is a convex function. The first derivative of $f(\theta)$ can be shown to be:

$$\frac{\partial f}{\partial \theta} = \frac{k}{\theta} - \frac{x}{\theta^2}$$

Differentiating again, we have:

$$\frac{\partial^2 f}{\partial \theta^2} = \frac{-k}{\theta^2} + \frac{2x}{\theta^3} = \frac{1}{\theta^2}\left(\frac{2x}{\theta} - k\right)$$

The multiplier term $\frac{1}{\theta^2}$ is greater than or equal to 0, therefore the sign of $\frac{\partial^2 f}{\partial \theta^2}$ is the sign of the term $\frac{2x}{\theta} - k$. For a Gamma pdf, the parameter $k > 0$ and so the term is not always positive and depends on the values of x and k. So generally, a unique maximum does not exist for the log-likelihood $\ell(\theta; x_1, \ldots, x_n)$ under this parameterization of the population distribution.

2. Time series data can be modeled as autoregressive processes. For example, consider a random sample that following a first-order autoregressive process (i.e., an AR(1) process):

$$X_i = aX_{i-1} + \epsilon_t, \quad i = 1, \ldots, n$$

where $\epsilon_t \sim \mathcal{N}(0,1)$ and $X_0 = x_0$ is assumed to be fixed and known. Here, we say that X_i is conditionally independent of all other X_j given X_{i-1}.

(a) Write down the log-likelihood of a for observed data $x_1, \ldots, x_n$.

Solution. Using the conditional independence assumption, we can write:

$$p(x_{1:n}; a) = \prod_{i=1}^{n} p(x_i | x_{i-1}; a),$$

where the distribution $p(x_i|x_{i-1};a)$ is given by:

$$p(x_i|x_{i-1};a) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - ax_{i-1})^2}{2}\right)$$

Putting these two pieces together, we have:

$$\begin{aligned} p(x_{1:n}; a) &= \prod_{i=1}^{n} p(x_i|x_{i-1};a) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - ax_{i-1})^2}{2}\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - ax_{i-1})^2\right) \end{aligned}$$

Therefore, the log-likelihood is given by:

$$\ell(a; x_{1:n}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - ax_{i-1})^2$$

(b) Find the maximum likelihood estimator of a.

Solution. We can find the MLE of a by differentiating the log-likelihood $\ell(a; x_{1:n})$ w.r.t. a and setting it equal to 0:

$$\begin{aligned}
\frac{\partial \ell}{\partial a} &= \sum_{i=1}^{n} (x_i - a x_{i-1}) x_{i-1} \\
&\implies \sum_{i=1}^{n} (X_i - \widehat{a} X_{i-1}) X_{i-1} = 0 \\
&\implies \widehat{a} \sum_{i=1}^{n} X_{i-1}^2 = \sum_{i=1}^{n} X_i X_{i-1} \\
&\implies \widehat{a} = \frac{\sum_{i=1}^{n} X_i X_{i-1}}{\sum_{i=1}^{n} X_{i-1}^2}
\end{aligned}$$

(c) Suppose that X_k is a latent random variable; that is, assume that it cannot be observed and will be missing from the data record. Derive the EM algorithm updates for finding the maximum likelihood estimator.

Solution. Recall that the EM algorithm updates are obtained by taking the expected value of the latent variable given the most recent value of the unknown parameters (E-Step) and then solving for the MLE, where the latent variable is replaced by the expected value (M-Step). From the previous part, we already know the complete data log-likelihood of the observed data $x_{-k} = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$ and the latent variable x_k and know what the MLE is in the case X_k is observed (hence, the M-step is already solved in part b of this question). For the E-step, we need to find $\mathbb{E}[X_k | x_{-k}; \widehat{a}]$. Note that if $\widehat{a}$ is known and x_{k-1} is known, the distribution of X_k is a Gaussian distribution (this is due to the fact X_k is conditionally independent of X_{-k} given X_{k-1}):

$$p(x_k | x_{k-1}; \widehat{a}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_k - \widehat{a} x_{k-1})^2}{2} \right),$$

where the mean of the Gaussian is $\mu_k = \widehat{a} x_{k-1}$. To summarize, the EM algorithm for solving this MLE in the case that a data point x_k is missing is:

- Initialize $\widehat{a}^{(0)}$
- For $i = 1, \ldots,$ (until convergence):
  - E-step:
  $$\mu_k^{(i)} = \widehat{a} x_{k-1}$$
  - M-step:
  $$\widehat{a}^{(i)} = \frac{\sum_{i=1}^{n} x_i x_{i-1}}{\sum_{i=1}^{n} x_{i-1}^2},$$
  where we replace x_k with $\mu_k^{(i)}$, i.e., the result from the M-step.

3. Consider the probabilistic model:

$$X_i = A r^i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is white Gaussian noise with variance σ^2 and $r > 0$ is known.

(a) Find the CRLB for A.

Solution. The CRLB can be found by determining the Fisher information:

$$I(A) = -\mathbb{E}\left[\frac{\partial^2 \log p(x_{1:n}; A)}{\partial A^2} \right]$$

This can be done straightforwardly by first determining the joint density $p(x_{1:n}; A)$:

$$p(x_{1:n}; A) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - Ar^i\right)^2\right)$$

with corresponding logarithm given by:

$$\log p(x_{1:n}; A) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - Ar^i\right)^2$$

The first and second derivative of $\log p(x_{1:n}; A)$ can be found to be:

$$\frac{\partial \log p(x_{1:n}; A)}{\partial A} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i r^i - Ar^{2i})$$

$$\frac{\partial^2 \log p(x_{1:n}; A)}{\partial A^2} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} r^{2i}$$

Therefore, the Fisher information is given by:

$$I(A) = \frac{1}{\sigma^2}\sum_{i=1}^{n} r^{2i}$$

and the CRLB establishes that for any unbiased estimator $\tilde{A}$, we must have that:

$$\begin{aligned}\mathbb{V}[\tilde{A}] &\geq \frac{1}{I(A)} \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} r^{2i}}\end{aligned}$$

(b) Find the maximum likelihood estimator for A and derive its variance.

Solution. We can obtain the MLE $\widehat{A}$ by setting the first derivative of the log-likelihood (from the first part) equal to 0:

$$\begin{aligned}&\frac{\partial \log p(x_{1:n}; A)}{\partial A} = 0 \\ \implies &\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i r^i - Ar^{2i}) = 0 \\ \implies &\widehat{A} = \frac{\sum_{i=1}^{n} r^i X_i}{\sum_{i=1}^{n} r^{2i}}\end{aligned}$$

To obtain the variance we have:

$$\begin{aligned}\mathbb{V}[\widehat{A}] &= \mathbb{V}\left[\frac{\sum_{i=1}^{n} r^i X_i}{\sum_{i=1}^{n} r^{2i}}\right] \\ &= \frac{1}{\left(\sum_{i=1}^{n} r^{2i}\right)^2}\mathbb{V}\left[\sum_{i=1}^{n} r^i X_i\right] \\ &= \frac{1}{\left(\sum_{i=1}^{n} r^{2i}\right)^2}\sum_{i=1}^{n}\mathbb{V}\left[r^i X_i\right] \\ &= \frac{1}{\left(\sum_{i=1}^{n} r^{2i}\right)^2}\sum_{i=1}^{n} r^{2i}\mathbb{V}\left[X_i\right] \\ &= \frac{1}{\left(\sum_{i=1}^{n} r^{2i}\right)^2}\sum_{i=1}^{n} r^{2i}\sigma^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} r^{2i}}\end{aligned}$$

We can see that the variance of this estimator is equal to the inverse of the Fisher information. This implies that the MLE is an efficient estimator (in the finite sample sense, not just the asymptotic sense).

(c) What happens to the variance as $n \to \infty$ for various values of r?

Solution. When $|r| < 1$, the denominator is a (convergent) geometric series with radius r^2:

$$\begin{aligned}\lim_{n\to\infty} \sum_{i=1}^{n} r^{2i} &= \lim_{n\to\infty} \left(\sum_{i=0}^{n} (r^2)^i \right) - 1 \\ &= \frac{1}{1-r^2} - 1 \\ &= \frac{r^2}{1-r^2},\end{aligned}$$

where we had to subtract 1 since a geometric series starts with index $i = 0$. For this case, the variance of $\widehat{A}$ will be strictly greater than 0 even in the limit of infinite samples. If $|r| \geq 1$, the variance of $\widehat{A}$ will go to 0. Therefore, we have:

$$\lim_{n\to\infty} \mathbb{V}[\widehat{A}] = \begin{cases} \frac{(1-r^2)}{r^2}\sigma^2, & |r| < 1 \\ 0, & \text{otherwise} \end{cases}$$

4. Consider the probabilistic model:

$$X_i = r^i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is white Gaussian noise with variance σ^2. Find the CRLB for r.

Solution. This is a similar probabilistic model to the previous question, except with unknown coefficient r and $A = 1$. The log joint distribution is given by:

$$\log p(x_{1:n}; \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(x_i - r^i\right)^2$$

Taking first- and second-order derivative w.r.t. r, we have:

$$\begin{aligned}\frac{\partial \log p(x_{1:n}; r)}{\partial r} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (i r^{i-1} x_i - i r^{2i-1}) \\ \frac{\partial^2 \log p(x_{1:n}; r)}{\partial r^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} i(i-1) r^{i-2} x_i - i(2i-1) r^{2i-2}\end{aligned}$$

To find the Fisher information, we take the negative expected value of the second partial derivative of

the log joint distribution:

$$\begin{aligned}
I(r) &= -\mathbb{E}\left[\frac{\partial^2 \log p(x_{1:n}; r)}{\partial r^2}\right] \\
&= -\mathbb{E}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n} i(i-1)r^{i-2}X_i - i(2i-1)r^{2i-2}\right] \\
&= -\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} i(i-1)r^{i-2}\mathbb{E}[X_i] - i(2i-1)r^{2i-2}\right) \\
&= -\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} i(i-1)r^{i-2}r^{i} - i(2i-1)r^{2i-2}\right) \\
&= -\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} i(i-1)r^{2i-2} - i(2i-1)r^{2i-2}\right) \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n} i^2 r^{2i-2}
\end{aligned}$$

where we used the fact that $\mathbb{E}[X_i] = r^i$. Therefore, the CRLB for any unbiased estimator $\tilde{r}$ is established as:

$$\begin{aligned}
\mathbb{V}[\tilde{r}] &\geq \frac{1}{I(r)} \\
&= \frac{\sigma^2 r^2}{\sum_{i=1}^{n} i^2 r^{2i}}
\end{aligned}$$

5. Consider a discrete random sample from a Poisson population, i.e., $X_i \overset{i.i.d.}{\sim} \text{Poisson}(\alpha)$, where:

$$p(x; \alpha) = \frac{\alpha^x}{x!}e^{-\alpha}, \quad x = 0, 1, \ldots,$$

and $\alpha > 0$. Suppose that α is unknown and we would like to estimate it.

(a) Find the CRLB for α.

Solution. We proceed by finding the Fisher information. For this question, we will use the identity that:

$$-\mathbb{E}\left[\frac{\partial^2 \log p(x_{1:n}; \alpha)}{\partial \alpha^2}\right] = \sum_{i=1}^{n} -\mathbb{E}\left[\frac{\partial^2 \log p(x_i; \alpha)}{\partial \alpha^2}\right],$$

that is, for an independent random sample, the Fisher information of $X_1, \ldots, X_n$ is the sum of the individual Fisher information for each X_i. This makes the computation more simple, in the sense that we don't need to take products. For each individual component, we have:

$$\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \log p(x_i; \alpha)}{\partial \alpha^2}\right] &= -\mathbb{E}\left[\frac{\partial^2}{\partial \alpha^2}\left(X_i \log \alpha - \log(X_i!) - \alpha\right)\right] \\
&= -\mathbb{E}\left[-\frac{X_i}{\alpha^2}\right] \\
&= \frac{1}{\alpha^2}\mathbb{E}[X_i] \\
&= \frac{1}{\alpha},
\end{aligned}$$

where we have used the fact that $\mathbb{E}[X] = \alpha$ for a Poisson random variable. Therefore, we have that

$$I(\alpha) = \sum_{i=1}^{n} \frac{1}{\alpha} = \frac{n}{\alpha}$$

and the CRLB for any unbiased estimator $\tilde{\alpha}$ is given by:

$$\begin{aligned}\mathbb{V}[\tilde{\alpha}] &\geq \frac{1}{I(\alpha)} \\ &= \frac{\alpha}{n}\end{aligned}$$

(b) Derive the method of moments estimator of α and compute its variance. Is the estimator efficient?

Solution. The MOME estimator can easily be established by setting the expected value of X equal to the sample mean. Since we know for a Poisson random variable $\mathbb{E}[X] = \alpha$, we have the MOME esitmator is given by:

$$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The variance of this estimator can also easily be established by using the fact that the variance of a Poisson random variable is $\mathbb{V}[X] = \alpha$. Therefore, we have:

$$\begin{aligned}\mathbb{V}[\widehat{\alpha}] &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}[X_i] \\ &= \frac{\alpha}{n}\end{aligned}$$

Since the variance of the estimator achieves the CRLB (that is $\mathbb{V}[\widehat{\alpha}] = 1/I(\alpha)$), we say that the estimator is efficient.

(c) Derive the maximum likelihood estimator of α and compute its variance. Is the estimator efficient?

Solution. The log-likelihood can be established as:

$$\begin{aligned}\ell(\alpha; x_{1:n}) &= \sum_{i=1}^{n} \log p(x_i; \alpha) \\ &= \sum_{i=1}^{n} x_i \log \alpha - \log(x_i!) - \alpha \\ &= -n\alpha + \sum_{i=1}^{n} - \log(x_i!) + x_i \log \alpha\end{aligned}$$

Differentiating the log-likelihood w.r.t. α and setting equal to 0, we have:

$$\begin{aligned}\frac{\partial \ell(\alpha; x_{1:n})}{\partial \alpha} &= -n + \frac{1}{\alpha}\sum_{i=1}^{n} x_i \\ \implies \widehat{\alpha} &= \frac{1}{n}\sum_{i=1}^{n} X_i\end{aligned}$$

and thus the MLE is the sample mean, which is the same as the MOME. Since the estimators are the same, the MLE has the same variance and also is an efficient estimator.