ECE 275B Homework # 1 Solutions – Winter 2018

1. (a) Because x_i are assumed to be independent realizations of a *continuous* random variable, it is almost surely (a.s.)[1] the case that

$$x'_1 < x'_2 < \cdots < x'_n$$

Thus, the preimage of $\mathbf{x}'$ contains the $n!$ possible arrangements of the a.s. distinct values of the (ordered) components of $\mathbf{x}'$. As there is no reason to prefer any one arrangement over another, the classical probability assumption of equally likely events[2] yields

$$p_\theta(\mathbf{x} \,|\, T(\mathbf{x}) = \mathbf{x}') = \frac{1}{n!}\, \chi\left\{\mathbf{x} \in T^{-1}(\mathbf{x}')\right\} \quad \text{a.s.}$$

which is independent of the unknown parameter vector θ.[3] Thus the vector of order statistics $\mathbf{x}' = T(\mathbf{x})$ is sufficient as a consequence of the classical (Fisherian) definition of sufficiency.

(b) Because of the assumption of independence,

$$p_\theta(\mathbf{x}) = p_\theta(x_1) \cdots p_\theta(x_n) = p_\theta(x'_1) \cdots p_\theta(x'_n) = g(\theta, T(\mathbf{x}))\,.$$

Thus the vector of order statistics $\mathbf{x}' = T(\mathbf{x})$ is sufficient as a consequence of the Neyman–Fisher Factorization Theorem.

2. The solution closely follows the development done in lecture. Invoking the *Axiom of Choice,* for each coset[4] $[x]$ of the equivalence class (partition) defined in the problem statement we can choose a unique representative point $\xi \in [x]$ to serve as an index for the coset,

$$A_\xi \triangleq [x] \quad \text{iff} \quad [x] = [\xi]\,.$$

We then define a statistic $S(\cdot)$ which takes its values in the set of index variables by

$$\xi = S(x) \quad \text{iff} \quad x \in A_\xi = [\xi]\,.$$

Note this statistic induces the original partition,

$$x \overset{S}{\sim} x' \quad \text{iff} \quad S(x) = S(x') = \xi \quad \text{iff} \quad x, x' \in A_\xi = [\xi] \quad x, x' \sim \xi\,.$$

[1] Recall that "almost surely" means "with probability equal to one."

[2] Sometimes called the "Principle of *Insufficient Reason*" because there is *no* "reason" to consider any one outcome as more probable than another.

[3] As usual, $\chi\{A\}$ denotes the characteristic (or indicator) function of the event A.

[4] By definition $y \in [x]$ iff $y \sim x$.

Sufficiency of *S*. By *definition* of the equivalence class (partition) defined in the problem statement, $x \sim \xi = S(x)$ implies that

$$\frac{p_\theta(x)}{p_\theta(S(x))} = g(x, S(x))$$

or

$$p_\theta(x) = p_\theta(S(x)) \cdot g(x, S(x)) \triangleq f(\theta, S(x)) \cdot h(x)$$

so that S (and hence the partition) is sufficient from the Neyman–Fisher Factorization Theorem.

Necessity of *S*. Let $T(x)$ be *any* sufficient statistic for $p_\theta(x)$. Then $T(x) = T(x')$ implies that[5]

$$\frac{p_\theta(x)}{p_\theta(x')} = \frac{p_\theta(x, T(x))}{p_\theta(x', T(x'))} = \frac{p_\theta(x|T(x))\, p_\theta(T(x))}{p_\theta(x'|T(x'))\, p_\theta(T(x'))} = \frac{p_\theta(x|T(x))}{p_\theta(x'|T(x'))} = \frac{p(x|T(x))}{p(x'|T(x'))} \triangleq c(x, x')$$

where $c(x, x')$ is independent of θ. By the definition of the equivalence class induced by the statistic S defined above, this in turn implies that $S(x) = S(x')$. Thus the partition induced by S is coarser than that induced by any sufficient statistic T, which is true iff S is a function of any such T. Thus (by definition of necessity), S is a necessary statistic for $p_\theta(x)$.

3. If you had difficulty with this problem please come to my office hour.

4. On its domain of positive support a k–parameter exponential family distribution has the form

$$\ln p_\theta(y) = Q(\theta)^T T(y) - b(\theta) + a(y) = \sum_{j=1}^{k} Q_j(\theta) T_j(y) - b(\theta) + a(y)\,. \tag{1}$$

If the natural statistics[6] plus the constant function,

$$\{1, T_1(y), \cdots, T_k(y)\}\ ,$$

are linearly dependent, then at least one of the natural statistics (say, without loss of generality, $T_k(y)$) can be written as an affine combination of the others,

$$T_k(y) = \alpha_0 + \alpha_1\, T_1(y) + \cdots + \alpha_{k-1}\, T_{k-1}(y) = \alpha_0 + \sum_{j=1}^{k-1} \alpha_j\, T_j(y)\,. \tag{2}$$

[5]Recall that $\{x\} = \{x\} \cap \{T(x)\} = \{x\} \cap \{x' : T(x') = T(x)\}$ induces a "cut" $p(x) = p(x, T(x)) = p(x|T(x))\, p(T(x))$.

[6]Why is $T(y)$ *always* a vector of sufficient statistics?

Thus

$$\begin{aligned}\sum_{j=1}^{k} Q_j(\theta)T_j(y) - b(\theta) + a(y) &= \sum_{j=1}^{k-1} \left[Q_j(\theta) + \alpha_j\, Q_k(\theta)\right] T_j(y) - \left[b(\theta) - \alpha_0\, Q_k(\theta)\right] + a(y) \\ &= \sum_{j=1}^{k-1} \widetilde{Q}_j(\theta)T_j(y) - \tilde{b}(\theta) + a(y)\end{aligned}$$

where

$$\widetilde{Q}_j(\theta) \triangleq Q_j(\theta) + \alpha_j\, Q_k(\theta),\ j = 1, \cdots, k-1$$

and

$$\tilde{b}(\theta) \triangleq b(\theta) - \alpha_0\, Q_k(\theta)\,,$$

showing that we can reduce $p_\theta(y)$ to a $(k-1)$–parameter exponential family distribution.

Similarly, if the natural parameters plus the constant function,

$$\{1, Q_1(\theta), \cdots, Q_k(\theta)\}\,,$$

are a linearly dependent set of functions of θ, then at least one of them (say $Q_k(\theta)$) can be written as an affine combination of the others

$$Q_k(\theta) = \beta_0 + \sum_{j=1}^{k-1} \beta_j\, Q_j(\theta)$$

and again we can reduce $p_\theta(y)$ to a $(k-1)$-parameter exponential family distribution. (With an appropriate possible redefinition of the term $a(y)$.)

Important Comments

If the set $\{1, T_1(y), \cdots, T_k(y)\}$ is linearly independent, it forms a $(k+1)$-dimensional basis for the function space of log-probability functions of the regular statistical model $\mathcal{P} = \{p_\theta(y) |\, \theta \in \Theta\}$.[7] Also note that the ability to perform a reduction in rank due to linear dependence in $\{1, Q_1(\theta), \cdots, Q_k(\theta)\}$ *contradicts* the assumption that the natural parameter space, $\mathbb{Q}$, is "solid" (i.e., has nonempty interior, $\mathring{\mathbb{Q}} \neq \emptyset$). *Thus* if the k natural statistics plus constant function are independent *and* the natural parameter space is solid *then* the dimension $k+1$ cannot be further reduced and the rank, k, of the distribution is minimal.

[7] See Equation (1) and note that $a(y)$ is a parameter-independent function that can be move to the left-hand-side of the equation and absorbed into the loglikelihood.

The *standard assumptions* which are assumed to hold for a ***regular*** *k-parameter exponential family distribution* are as follows:

(a) The support of the density, $p_\theta(y)$, is independent of θ.

(b) The parameter space, $\Theta \subset \mathbb{R}^m$, has nonempty interior, $\mathring{\Theta} \neq \emptyset$.

(c) The k-dimensional natural statistics plus the constant function are linearly independent.

(d) The mapping, $Q(\cdot) : \Theta \to \mathbb{Q}$, from the parameter space, Θ, to the k-dimensional natural parameter space, $\mathbb{Q}$, is one-to-one.[8]

(e) The natural parameter space has nonempty interior, $\mathring{\mathbb{Q}} \neq \emptyset$.[9]

(f) The natural parameters, $Q(\theta)$, and $b(\theta)$ are twice continuously differentiable with respect to θ.[10]

Regular k-parameter distributions are of great utility because the natural (sufficient) statistics are *complete* (and thus minimal) and can therefore be used to construct uniformly minimum variance unbiased estimators (UMVUEs) via the Rao-Blackwell procedure. They also have very useful and convenient convexity properties.[11]

If the interior of the natural parameter space is not open (does not contain a k-dimensional rectangle) and yet the natural statistics plus the constant function are nonetheless linearly independent, this means that the natural parameters are either discrete points and/or constrained to live on a nonlinear hypersurface (manifold or "hypercurve"). In the latter case, the k-parameter distribution is called a *curved* exponential family distribution. The natural statistic is still sufficient, but generally it is not complete.[12]

[8] This is the requirement that $Q(\theta_1) = Q(\theta_2) \implies \theta_1 = \theta_2$, which ensures that the (nonnatural) parameter θ is identifiable if the natural parameter $Q(\theta)$ is identifiable.

[9] This assumption, that $\mathbb{Q}$ is "solid", together with condition (c), ensures that the (sufficient) natural statistics are complete, and hence minimal. I.e., conditions (c) and (e) ensure that the rank, k, is minimal and that the natural parameters are complete and identifiable.

[10] Note the tacit assumption that the log-partition function $b(\theta)$ exits. This means that the "raw" distribution is integrable and hence can be normalized to be a proper probability distribution. Also recall that the log-partition function only depends on θ though the natural parameter $Q(\theta)$, so we often abuse notation and write "$b(Q)$".

[11] Specifically, **1)** the negative log-likelihood function is convex in the natural parameter vector, Q, and **2)** the "total" natural parameter space, which is defined to be the set of all natural parameter values for which the log-partition function, $b(Q)$, exists, is a convex set. This means that the maximum likelihood estimate of the natural parameter vector is the only solution to the likelihood equation, greatly simplifying the problem of computing this estimate.

[12] However, if the convex hull of a curved or discrete parameter space has a nonempty interior the natural statistic is minimal, though not necessarily complete. The minimality is shown, e.g., in the textbook *Statistics and Econometric Models*, Vol. 1, 2005, C. Gourieroux and A. Monfort.

5. Note that

$$y - \langle y \rangle = y - \overline{y} + \overline{y} - \langle y \rangle$$

and therefore

$$\begin{aligned}(y - \langle y \rangle)(y - \langle y \rangle)^T &= (y - \overline{y})(y - \overline{y})^T + (y - \overline{y})(\overline{y} - \langle y \rangle)^T \\ &\quad + (\overline{y} - \langle y \rangle)(y - \overline{y})^T + (\overline{y} - \langle y \rangle)(\overline{y} - \langle y \rangle)^T .\end{aligned}$$

Now note that

$$\begin{aligned}\mathrm{E}\left\{(y - \overline{y})(\overline{y} - \langle y \rangle)^T\right\} &= \mathrm{E}\left\{\mathrm{E}\left\{(y - \overline{y})(\overline{y} - \langle y \rangle)^T \mid x\right\}\right\} \\ &= \mathrm{E}\left\{(\overline{y} - \overline{y})(\overline{y} - \langle y \rangle)^T\right\} = 0 .\end{aligned}$$

Therefore,

$$\mathrm{Cov}\{y\} = \mathrm{E}\left\{(y - \overline{y})(y - \overline{y})^T\right\} + \mathrm{Cov}\{\overline{y}\} \geq \mathrm{Cov}\{\overline{y}\} = \mathrm{Cov}\{\mathrm{E}\{y|\, x\}\} . \qquad (3)$$

Note that the left–hand–side equation of can also be written as

$$\mathrm{Cov}\{y\} = \mathrm{E}\{\mathrm{Cov}\{y|\, x\}\} + \mathrm{Cov}\{\mathrm{E}\{y|\, x\}\} .$$

The left–hand–side equation of Inequality (3) can also be interpreted as a "matrix pythagorean theorem".[13]

6. This problem is essentially done in Example 5.8 of Kay (once we recognize that an unbiased estimate is provided by twice an unbiased estimate of the mean). The only part of the homework problem which is unsolved in Example 5.8 is proving that the sufficient statistic

$$T(\mathcal{Y}^m) = \max_{1 \leq i \leq m} y_i$$

is complete. As shown in Kay, sufficiency easily follows from

$$p_\theta(\mathcal{Y}^m) = \left(\frac{1}{\theta^m}\,\chi\left\{\max_{1 \leq i \leq m} y_i \leq \theta\right\}\right) \cdot \left(\chi\left\{\min_{1 \leq i \leq m} y_i \geq 0\right\}\right) = g(T(\mathcal{Y}^m), \theta) \cdot h(\mathcal{Y}^m)$$

for any $\theta > 0$, and the Neyman–Fisher Factorization Theorem.[14] Completeness follows from noting (as per the development on page 115 of Kay) that any measurable function of T, say $W(t)$, has an expectation given by

$$\mathrm{E}_\theta\{W(T)\} = \frac{m}{\theta^m} \int_0^\theta W(t) t^{m-1} dt .$$

[13] This is most easily seen by taking $\langle y \rangle = 0$. Further note that then taking the trace of both sides of the left–hand–equation of Inequality (3) yields the "regular" pythagorean theorem.

[14] Here, $\chi\{\cdot\}$ denotes the so–called *characteristic* (or *indicator*) *function*.

Let $W(t) = W^+(t) - W^-(t)$, where $W^+(t) \geq 0$ and $W^-(t) \geq 0$ are *nonnegative* functions of t.[15] Then $\mathrm{E}_\theta\{W(T)\} = 0$ for all $\theta > 0$ if and only if

$$\int_0^\theta W^+(t)t^{m-1}dt = \int_0^\theta W^-(t)t^{m-1}dt \geq 0$$

for all $\theta > 0$, which, in turn, is true if and only if,[16]

$$\int_{\theta_1}^{\theta_2} W^+(t)t^{m-1}dt = \int_{\theta_1}^{\theta_2} W^-(t)t^{m-1}dt \geq 0$$

for every θ_1 and θ_2 such that $\theta_2 \geq \theta_1 > 0$. Because the integrands are positive and the equality must hold for every $\theta_2 \geq \theta_1 \geq 0$, it must therefore be the case that $W^+(t) = W^-(t)$ for almost all $t \geq 0$.[17] Therefore $W(t) = 0$ for almost all t showing that T is complete.[18]

7. Kay 5.13. Note that like the previous problem the density is not regular (in particular the area of positive support again depends on the unknown parameter θ) so that we cannot compute a Cramér–Rao lower bound. Note that we can write the sample data pdf as

$$p_\theta(\mathcal{X}^N) = \left(\mathrm{e}^{N\theta}\,\chi\left\{\min_{1\leq n\leq N} x[n] \geq \theta\right\}\right)\cdot\left(\mathrm{e}^{-\sum_{n=1}^N x[n]}\right) = g(T(\mathcal{X}^N),\theta)\cdot h(\mathcal{X}^N)\,.$$

Therefore, from the Neyman–Fisher Factorization Theorem a sufficient statistic is determined to be

$$T(\mathcal{X}^N) = \min_{1\leq n\leq N} x[n]\,.$$

[15] This can be done for any real function $W(t)$.

[16] $\int_0^{\theta_2} W^\pm(t)t^{m-1}dt = \int_0^{\theta_1} W^\pm(t)t^{m-1}dt + \int_{\theta_1}^{\theta_2} W^\pm(t)t^{m-1}dt$. and $\int_0^{\theta_1} W^+(t)t^{m-1}dt = \int_0^{\theta_1} W^-(t)t^{m-1}dt$.

[17] Let $\theta_1 = t \geq 0$ and $\theta_2 = t + \epsilon$, then in the limit of small $\epsilon > 0$ the equality becomes $\epsilon W^+(t)t^{m-1} = \epsilon W^-(t)t^{m-1} \implies W^+(t) = W^-(t)$.

[18] A more rigorously proof goes as follows: From the last inequality involving integrals, it follows that

$$\int_A W^+(t)t^{m-1}dt = \int_A W^-(t)t^{m-1}dt \geq 0$$

for every Borel measurable set A (recall that the Borel σ-algebra is the smallest σ-algebra containing the intervals in $\mathbb{R}$). In particular, if we take $A = \{t|W(t) > 0\}$ (for which the right-hand integral must take the value zero) it follows that $W^+ = 0$ a.e. Similarly, taking $A = \{t|W(t) < 0\}$ it follows that $W^- = 0$ a.e.

We now proceed to find the distribution function of T,

$$\begin{aligned} P_\theta(T \le t) &= 1 - P_\theta(T > t) = 1 - P_\theta(\min_{1\le n\le N} x[n] > t) \\ &= 1 - P_\theta(x[1] > t, \cdots, x[N] > t) \\ &= 1 - \left(\int_{\max\{t,\theta\}}^{\infty} \mathrm{e}^{\theta - x} dx \right)^N \\ &= 1 - \mathrm{e}^{N(\theta - \max\{t,\theta\})} . \end{aligned}$$

Differentiating the distribution function with respect to t we obtain the pdf,

$$p_\theta(t) = \begin{cases} N\mathrm{e}^{N(\theta - t)} & t \ge \theta \\ 0 & t < \theta \end{cases} .$$

The expected value of T can now be computed,

$$\mathrm{E}_\theta \{T\} = \theta + \frac{1}{N} .$$

An unbiased estimator is then obviously given by

$$\widehat{\theta} = T - \frac{1}{N} = \min_{1\le n\le N} x[n] - \frac{1}{N} .$$

It can be shown that T is a complete sufficient statistic.[19] Therefore, from the RBLS Theorem, we have found the UMVUE of the unknown parameter θ. Note that from the pdf of T we can compute the (uniformly optimal, parameter dependent) error variance if we so desire.

8. Moon 10.5.8. This is a generalization of the previous problem. The general class of such non–regular exponential families (i.e., exponential family–like, but with parameter–dependent support) is discussed in the text by Ferguson cited in Footnote 19.

   (a) For σ known, μ unknown we have

   $$p_\theta(\mathcal{X}^n) = \left(\mathrm{e}^{n\frac{\mu}{\sigma}} \chi \left\{ \min_{1\le k\le n} x_k \ge \mu \right\} \right) \cdot \left(\frac{\mathrm{e}^{-\frac{1}{\sigma}\sum_{k=1}^n x_k}}{\sigma^n} \right) = g(T(\mathcal{X}^n), \theta) \cdot h(\mathcal{X}^n) .$$

   From the N–F Factorization Theorem, $T(\mathcal{X}^n) = \min_{1\le k\le n} x_k$ is a sufficient statistic for μ. It is also complete.

[19] See, e.g., *Mathematical Statistics: A Decision Theoretic Approach,* T.S. Ferguson, Academic Press, 1967, Exercise #4, page 137.

(b) For σ unknown and μ known we have

$$p_\theta(\mathcal{X}^n) = \left(\frac{\mathrm{e}^{n\frac{\mu}{\sigma} - \frac{1}{\sigma}\sum_{k=1}^n x_k}}{\sigma^n} \right) \cdot \left(\chi\left\{ \min_{1\le k\le n} x_k \ge \mu \right\} \right) = g(T(\mathcal{X}^n), \theta) \cdot h(\mathcal{X}^n)\,.$$

Thus $T(\mathcal{X}^n) = \sum_{k=1}^n x_k$ is a sufficient statistic for σ.

(c) For both σ and μ unknown we have

$$p_\theta(\mathcal{X}^n) = \left(\frac{\mathrm{e}^{n\frac{\mu}{\sigma} - \frac{1}{\sigma}\sum_{k=1}^n x_k}}{\sigma^n} \chi\left\{ \min_{1\le k\le n} x_k \ge \mu \right\} \right) \cdot 1 = g(T(\mathcal{X}^n), \theta) \cdot h(\mathcal{X}^n)\,.$$

Therefore $T(\mathcal{X}^n) = (\sum_{k=1}^n x_k,\ \min_{1\le k\le n} x_k)^T$ is a sufficient statistic for $\theta = (\sigma, \mu)^T$.

9. Kay 5.15. If you have trouble with the first two parts of this problem, please come see me at my office hour. Note that the Gaussian, Rayleigh, and Exponential distributions are regular exponential families and that T is a complete (and hence minimal) sufficient statistic. Note that below we can find an UMVUE for each case, *but only if we choose an appropriate parameterization*. This is because of the strong constraint that the estimator be uniformly unbiased.

(a) Gaussian. Here $\theta = \mu$, $T(\mathbf{x}) = \sum_{n=1}^N x[n]$ and $E_\theta\{T(\mathbf{x})\} = N\mu$. Therefore the UMVUE for μ is given by

$$\widehat{\mu} = \frac{1}{N} T(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N x[n]\,.$$

(b) Rayleigh. Here $\theta = \sigma^2$, $T(\mathbf{x}) = \sum_{n=1}^N x^2[n]$ and $E_\theta\{T(\mathbf{x})\} = 2\sigma^2 N$. Therefore

$$\widehat{\sigma^2} = \frac{1}{2N} T(\mathbf{x}) = \frac{1}{2N} \sum_{n=1}^N x^2[n]$$

is the UMVUE for σ^2.

(c) Exponential. Here the appropriate parameterization is a little trickier. Now we take $\theta = \frac{1}{\lambda}$. We also have $T(\mathbf{x}) = \sum_{n=1}^N x[n]$ and $E_\theta\{T(\mathbf{x})\} = \frac{N}{\lambda}$. Thus the UMVUE is given by

$$\widehat{\theta} = \widehat{\left(\frac{1}{\lambda}\right)} = \frac{1}{N} T(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N x[n]\,.$$

10. Under the multivariate gaussian assumption we have that

$$p_{\boldsymbol{\theta}}(\mathbf{y}) = c \exp\left\{ -\frac{1}{2} \|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|^2_{\mathbf{R}^{-1}} \right\}$$

where the normalizing constant c is independent of $\boldsymbol{\theta}$ and the full column–rank matrix $\mathbf{A}$ is $m \times k$.

(a) With $T(\mathbf{y}) \triangleq \mathbf{A}^T\mathbf{R}^{-1}\mathbf{y}$ we have

$$\|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|^2_{\mathbf{R}^{-1}} = \|\mathbf{y}\|^2_{\mathbf{R}^{-1}} - 2\boldsymbol{\theta}^T T(\mathbf{y}) + \|\mathbf{A}\boldsymbol{\theta}\|^2_{\mathbf{R}^{-1}}.$$

Thus, as a consequence of the NFFT, T is sufficient and $p_{\boldsymbol{\theta}}(\mathbf{y})$ is seen to be an exponential family distribution. Because of the full column rank assumption on $\mathbf{A}$, the rows of $\mathbf{A}^T$ are linearly independent which means that the components of T are linearly independent functions of $\mathbf{y}$. Because the parameter vector is unconstrained, the parameter space has nonempty interior so that T is complete and therefore minimal.[20] Because of the assumption that $\mathbf{A}$ has full column rank, it must be the case that $k \leq m$. Thus the k–dimensional minimum sufficient statistic realization value $t = T(\mathbf{y})$ is no larger than the dimension m of the raw data. If it is the case that $k < m$ then data compression has occurred, which is especially nice in the case when $k \ll m$.

(b) With T a complete, minimum sufficient statistic, the RBLS Theorem tells us that the UMVUE (if it exists[21]) must be a function of T. Noting that

$$\mathrm{E}_{\boldsymbol{\theta}}\{T\} = \mathbf{A}^T\mathbf{R}^{-1}\mathbf{A}\boldsymbol{\theta},$$

with $\mathbf{A}$ full column–rank, it is evident that the UMVUE is given by

$$\widehat{\boldsymbol{\theta}}(\mathbf{y}) = \left(\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A}\right)^{-1} T(\mathbf{y}) = \left(\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{R}^{-1}\mathbf{y}.$$

11. The solution to this problem involves a very simple application of the result from the previous section. Define $T_k \triangleq T(\mathbf{y_k})$ and

$$\boldsymbol{\Sigma}_k \triangleq \mathrm{E}\left\{\mathbf{v}_k\mathbf{v}_k^T\right\} = \mathrm{diag}(\sigma_1, \cdots \sigma_k) = \mathrm{diag}(\boldsymbol{\Sigma}_{k-1} \quad \sigma_k)\,.$$

Then

$$T_k = \mathbf{A}_k^T\boldsymbol{\Sigma}_k^{-1}\mathbf{y_k} = \mathbf{A}_{k-1}^T\boldsymbol{\Sigma}_{k-1}^{-1}\mathbf{y_{k-1}} + \frac{\mathbf{r}_k^T}{\sigma_k}\,y[k] = T_{k-1} + \mathbf{b}_k\,y[k]$$

where $\mathbf{b}_k \triangleq \frac{\mathbf{r}_k^T}{\sigma_k}$. Note that both T_k and $\mathbf{b}_k$ are n–dimensional for all times k.

12. Kay 7.7. In particular, we will show that $\widehat{\theta}_{\mathrm{ML}}$ is the solution to the equation

$$\mathrm{E}_\theta\{B(\mathbf{X})\} = B(\mathbf{x})$$

where $\mathbf{x} = \mathbf{X}(\omega)$ denotes the realization vector of the N iid samples. Recall that we have defined $f'(\theta)$ as

$$f'(\theta) = \frac{\partial}{\partial\theta} f(\theta)\,.$$

[20] This is a fundamental property of regular, full–rank, nonempty parameter–set exponential family distributions.

[21] Remember that the set of uniformly unbiased estimators might be empty for an unfelicitous choice of parameterization.

Because the samples are iid, we have that

$$p_\theta(\mathbf{x}) = \exp\left\{A(\theta)B(\mathbf{x}) + C(\mathbf{x}) + ND(\theta)\right\}$$

where

$$B(\mathbf{x}) = \sum_{k=1}^{N} B(x_i) \tag{4}$$

is a complete sufficient statistic for θ and

$$C(\mathbf{x}) = \sum_{k=1}^{N} C(x_i)\,.$$

The MLE is found as a solution of the likelihood equation,

$$S_\theta(\mathbf{x}) = 0\,, \tag{5}$$

where $S_\theta(\mathbf{x})$ is the score function,

$$S_\theta(\mathbf{x}) = \frac{\partial}{\partial\theta} \ln p_\theta(\mathbf{x}) = A'(\theta)B(\mathbf{x}) + ND'(\theta)\,. \tag{6}$$

I.e., the MLE is a solution to

$$A'(\theta)B(\mathbf{x}) + ND'(\theta) = 0\,. \tag{7}$$

It is evident, then, that the MLE can be equivalently found as a solution to the equation

$$\boxed{B(\mathbf{x}) = -N\,\frac{D'(\theta)}{A'(\theta)}} \tag{8}$$

Now because we are working with a regular exponential family, we can interchange the order of differentiation and integration in expressions involving $p_\theta(\mathbf{x})$. In particular, differentiating the expression

$$\int p_\theta(\mathbf{x}) d\mathbf{x} = 1$$

with respect to θ yields the fact that the score has zero mean uniformly in θ

$$\mathrm{E}_\theta\left\{S_\theta(\mathbf{X})\right\} = 0 \tag{9}$$

an expression derived and discussed in class last quarter. The likelihood equation (5) and the zero-mean condition (9) are the key equations needed to solve this homework problem.

From (6), equation (9) is seen to be equivalent to

$$\mathrm{E}_\theta \left\{ A'(\theta) B(\mathbf{X}) + N D'(\theta) \right\} = 0$$

which can be rearranged as

$$\boxed{\mathrm{E}_\theta \left\{ B(\mathbf{X}) \right\} = -N \, \frac{D'(\theta)}{A'(\theta)}} \tag{10}$$

If we take θ itself to be the natural parameter

$$\theta = A$$

then this becomes

$$\mathrm{E}_A \left\{ B(\mathbf{X}) \right\} = -N \, D'(A), \tag{11}$$

where D is here taken to be a function of the natural parameter A and $D'(A)$ denotes the derivative of D with respect to A.

Finally, comparison of equations (8) and (10) shows that the MLE can be found as a solution to the equation

$$\boxed{\mathrm{E}_\theta \left\{ B(\mathbf{X}) \right\} = B(\mathbf{x})}$$

as claimed.

(a) Unit variance, unknown mean Gaussian case. We have,

$$p(x;\mu) = \phi(x;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x-\mu)^2 \right\} = \exp\left\{ \mu x - \frac{1}{2}x^2 - \frac{1}{2}(\mu^2 + \ln 2\pi) \right\} .$$

Thus we have,

$$A(\mu) = \mu, \quad B(x) = x, \quad C(x) = -\frac{x^2}{2}, \quad D(\mu) = -\frac{1}{2}(\mu^2 + \ln 2\pi),$$

and therefore,

$$A'(\mu) = 1, \quad D'(\mu) = -\mu, \quad B(\mathbf{x}) = \sum_{n=1}^{N} B(x[n]) = \sum_{n=1}^{N} x[n] \, .$$

Substitution into equation (7) above and rearranging yields the answer,

$$\widehat{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x[n] \, .$$

(b) Exponential Distribution. Here we have,

$$p(x;\lambda) = \lambda \exp\left\{ -\lambda x \right\} = \exp\left\{ -\lambda x + \ln \lambda \right\} .$$

Thus,

$$A(\lambda) = -\lambda, \quad B(x) = x, \quad C(x) = 0, \quad D(\lambda) = \ln \lambda,$$

so that

$$A'(\lambda) = -1, \quad D'(\lambda) = \frac{1}{\lambda}, \quad B(\mathbf{x}) = \sum_{n=1}^{N} B(x[n]) = \sum_{n=1}^{N} x[n]\,.$$

Substituting into equation (7) above and rearranging yields

$$\widehat{\lambda}_{\mathrm{ML}} = \left(\frac{1}{N} \sum_{n=1}^{N} x[n] \right)^{-1}.$$