MDL exercises, fifth handout

(due March 30th) (note: see also back side of paper; there are 4 exercises!)

1. Let $\{p_\theta \mid \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$ be a smoothly parameterized i.i.d. 1-dimensional model (see page 65 in the book) and let $I(\theta)$ denote the Fisher information at θ. You may assume that, in the exercises below, the order of taking expectations and differentiating can be interchanged, i.e. the expected value of a derivative is the derivative of the expected value.

   a) Show that, for θ, θ' in the interior of Θ, the KL divergence (relative entropy) satisfies

   $$D(\theta \| \theta') = \frac{1}{2} I(\theta)(\theta - \theta')^2 + O\left((\theta - \theta')^3\right). \tag{1}$$

   b) For a variety of models in their standard parameterizations, including the Poisson, geometric, normal and Bernoulli families, the following facts hold: (1) $I(\theta)$ is a continuous function of θ; (2) for every parameter θ and every sequence $x^n = x_1, \ldots, x^n$ such that both θ and the ML estimator $\hat{\theta}$ fall in the interior of Θ, we have:

   $$\frac{1}{n}\left(-\log \frac{p_\theta(x^n)}{p_{\hat{\theta}}(x^n)}\right) = D(\hat{\theta} \| \theta). \tag{2}$$

   Now suppose that we restrict the model to a subset Θ' of the interior of Θ where Θ' is some finite interval of length A. We discretize Θ' to a finite set $\ddot{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$ of m parameter values at distance $A/\sqrt{n}$, where $m = \sqrt{n} + 1$.

   Now consider the two-part code that works as follows: the data x^n are encoded in two stages: we first code the $\theta \in \ddot{\Theta}$ that maximizes the probability of the data. Here we use a uniform code on $\ddot{\Theta}$. We then code the data using the Shannon-Fano code based on the θ we encoded in the first stage.

   Assume that we get data such that, for all large n, $\hat{\theta} \in \Theta'$. Show, using (1) and (2) that the number of bits $L(x^n)$ we need to encode the data in this way satisfies

   $$-\log p_{\hat{\theta}}(x^n) < L(x^n) \leq -\log p_{\hat{\theta}}(x^n) + \frac{1}{2}\log n + C$$

   for some constant C independent of n.

2. Consider the Bernoulli model. Compute the probability that the first two outcomes are different on the basis of four different universal models/codes:

   - The Bayesian model with uniform prior
   - The Bayesian model with Jeffreys' prior (Hint: use that for this universal model the following variation of Laplace's rule of succession holds: $\bar{P}(X_{n+1} = 1 \mid X^n = x^n) = (n_1 + (1/2))/(n+1)$, where n_1 is the number of 1s in X^n).
   - The NML model for sample size 2
   - The NML model for sample size 3

3. Recall that the NML code is defined such that it has a constant regret of $\log \sum_{x^n} P(x^n \mid \hat{\theta}(x^n))$. With n_0 and n_1 defined as usual, show that in the case of the Bernoulli model this is equal to:

   $$\log \sum_{x^n \in \mathcal{X}^n} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0} \tag{3}$$

4. Suppose that we model data with a uniform distribution on the real numbers between 0 and $\theta > 0$.

   a) Given outcomes $x_1, \ldots, x_n$, what is the maximum likelihood value for θ? (yes, you had this question before, but it serves as a warm-up for the following question!)

   b) Explain why a formula like (1) *cannot* be proven for the uniform distributions on $[0, \theta]$. In what way then is the model of uniform distributions crucially different from the Bernoulli and the normal family?

   c) Show that (2) *does* hold for the uniform model.