MDL exercises, sixth handout:
(due April 6th, 14:00)

Recall from the previous week that, 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:

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

1. Let $\{p_{\mu,\sigma^2} \mid (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$ be the i.i.d.-normal family with mean μ and variance σ^2,

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

extended to n outcomes by taking product distributions. We turn this into a 1-dimensional family with members p_μ by setting σ^2 to 1, so that the density of x according to p_μ is given by $p_\mu(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(x-\mu)^2}$. This is called the normal *location* family.

   a) Show that in this case, the KL divergence between p_μ and $p_{\mu'}$ reduces to the squared Euclidean distance between the parameters.
   b) Show that (1) holds with $\theta = \mu$.
   c) Give a formula for $I(\mu)$ as a function of μ.

2. Consider the i.i.d.-normal family of the previous example; now we turn this into a 1-dimensional family by setting $\mu = 0$. We then get the so-called i.i.d. normal *scale* family $\{p_\sigma \mid \sigma > 0\}$ with mean 0 and variance σ^2, and the density of x according to p_σ is given by

$$p_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{x^2}{2\sigma^2}},$$

again extended to n outcomes by taking product distributions.

   a) Give a formula for the KL divergence between the distributions indexed by σ and σ'.
   b) Show that (1) holds with $\theta = \sigma$.
   c) Give a formula for $I(\sigma)$ as a function of σ.

3. Jeffreys' prior.

   a) Compute Jeffreys' prior for the normal distribution with mean μ in $[-K, K]$ and standard deviation σ in $[a, b]$ for a, b and K greater than zero. (The ranges are necessary to make the normalization factor converge.)
   b) Show that for the Bernoulli model we have $I(\theta) = 1/(\theta(1-\theta))$. For bonus points, you may also show that $\int_{\theta=0}^{1} \sqrt{\det I(\theta)} d\theta = \pi$. (You will need this in subsequent questions.)
   c) Show that Jeffreys' prior for the Bernoulli model is invariant if we reparameterize from $\theta = P(X = 1)$ to $\eta = -\ln(P(X = 1))$ in the domain $(0, \infty)$. That is, show that (take a deep breath), for any non-empty interval $\Theta = [a, b]$, Jeffreys' prior probability on Θ, calculated with respect to the first parameterization, is equal to Jeffreys' prior probability on the corresponding interval in the second parameterization, where Jeffreys' prior is calculated with respect to the second parameterization.

4. Recall that for well-behaved k-parameter statistical models the asymptotic worst case minimal regret is $\frac{k}{2}\log\frac{n}{2\pi}+\log\int_{\theta\in\Theta}\sqrt{\det I(\theta)}d\theta+o(1)$. Here, Θ is the parameter domain and a function is $o(1)$ if it converges to zero. For the Bernoulli model this evaluates to:

$$\frac{1}{2}\log\frac{n}{2\pi}+\log\pi+o(1) \tag{2}$$

In a previous lecture we described a two-part "index" code for the Bernoulli model. In the first part, we encoded the number of ones n_1 in the sequence with a uniform code, using $\log(n+1)$ bits. Then we gave the index in the list of all sequences of length n with that number of ones, using $\log\binom{n}{n_1}$ bits. Using the (by now) well-known facts that, (a), for the Bernoulli model, $-\log p_{\hat{\theta}}(x^n)=nH(\hat{\theta})$ and (b), that $\log\binom{n}{n_1}$ can be approximated in terms of $H(\hat{\theta})$, compute the asymptotic regret for this code and compare it to what we found in (2). For what sequences do you get a regret that is significantly different from minimax optimal? (Hint: for (b), use (4.36) in the book on page 129; this is a more precise approximation than what was done in earlier exercises)

5. Let $\mathcal{M}$ be the set of all Normal distributions with variance >0 and arbitrary mean (so $\mathcal{M}$ is a two-parameter model). Define the parameter vector $\theta=\binom{\mu}{\sigma}$ and let $\theta'=\binom{\mu'}{\sigma'}$. Find (a) an explicit expression for $D(P_\theta\|P_{\theta'})$ and (b) find an approximation of (a) by doing a second-order Taylor expansion of the logarithm appearing in your answer for (a). Now (c) find an explicit expression for the Fisher-information approximation $\frac{1}{2}(\theta-\theta')^T I(\theta)(\theta-\theta')$ (where superscript T is used to denote vector transpose). How do (b) and (c) relate?