MDL exercises 11 (due May 16th, 2017)

1. Let $\mathcal{M}_\phi$ be an exponential family of distributions on a single outcome, with carrier $r(x)$ equal to 1 for all x, i.e. a Maximum Entropy family. We extend $\mathcal{M}_\phi$ to n outcomes by taking product distributions, i.e. we model data as being i.i.d. according to a distribution in $\mathcal{M}_\phi$. Suppose that $P_{\hat{\beta}}$ is the maximum likelihood distribution within $\mathcal{M}_\phi$ for some data x^n.

   (a) Show that
   $$nH(P_{\hat{\beta}}) = -\ln P_{\hat{\beta}}(x^n). \tag{1}$$
   This means that the *expected* codelength (an average involving probabilities over *all* possible outcomes) according to the ML estimator is equal to the *actual codelength* (an average involving frequencies of the subset of *actually observed* outcomes).

   (b) (1) does not hold in general, but is in fact a very special property of Maximum Entropy families. Prove this by picking a model that is not a Maximum Entropy family and exhibiting a data set x^n for which (1) does not hold. For example, you may pick the model of 2 mixtures of Gaussians with mean 1 and varying means, that also featured in Exercise 2(f) of last week.

2. Consider the following family of distributions: the set of power law distributions, also known as the *Pareto family*: $P_\theta(n) = n^{-\theta}/\sum_{n=1}^{\infty} n^{-\theta}$ for $n \in \{1, 2, \ldots\}$ and $\theta > 1$. Prove that this is an exponential family. HINT: you can show that a family is an exponential family by rewriting it in the exponential form $\frac{1}{Z(\beta)} e^{\beta\phi(x)} r(x)$ for some function $\phi(x)$

3. Suppose $X_1, \ldots, X_n \in \{0, 1\}^n$ are outcomes of independent tosses of a fair coin.

   (a) Use Theorem 19.2 in the book (Sanov/Chernoff bound) to get a bound on $P(\sum_{i=1}^n X_i \leq 1)$ in terms of the KL divergence $D(\hat{\mu}\|\mu)$ where μ represents the Bernoulli distribution with mean μ. Also calculate $P(\sum_{i=1}^n X_i \leq 1)$ directly. How good is the bound?

   (b) Do the same for $P(\sum_{i=1}^n X_i \leq 0)$.

4. Let $\mathcal{P}$ consist of just two distributions: the Bernoulli distribution with mean $P(X = 1) = 0.49$ and the Bernoulli distribution with mean $P(X = 1) = 0.7$.

   (a) What is the maximum entropy distribution within the set $\mathcal{P}$?

   (b) What is the maximum entropy distribution within the convex closure of the set $\mathcal{P}$? (see also back side)

(c) If our goal is to code an outcome from some $P \in \mathcal{P}$ with small expected worst-case codelength, should we code according to the distribution under (a) or the distribution under (b)?