Homework 2
(due Friday, Feb 10, 2006)

February 1, 2006

1. The Fisher information matrix arises from multiparameter densities, where the (i, j) entry of the matrix is given by

$$g_{ij}(\theta) = \int p(\mathbf{x}|\theta) \frac{\partial}{\partial \theta^i} \log p(\mathbf{x}|\theta) \frac{\partial}{\partial \theta^j} \log p(\mathbf{x}|\theta) d\mathbf{x} \tag{1}$$

Intuitively one can think of the Fisher information as *a measure of the amount of information present in the data about a parameter θ.*

The Fisher information matrix also satisfies the properties of a metric on a Riemannian manifold. Don't worry too much about what exactly a Riemannian manifold is at this point. In this manifold, $p \in M$ is a probability density with its local coordinates defined by the model parameters. For example, a bivariate Gaussian density can be represented as a single point on *4*-dimensional manifold with coordinates $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)^T$, where as usual these represent the mean and standard deviation of the density. It can be shown that many of the other common distance measures on probability densities (e.g. Kullback-Leibler, Jensen-Shannon, etc.) can be written in terms of the Fisher-Rao metric given that the densities are close. For example, the Kullback-Leibler distance between two parametric densities θ and $\theta + \delta\theta$ is proportional to the Fisher-Rao metric g by

$$D\left(p(x|\theta + \delta\theta)||p(x|\theta)\right) \approx \frac{1}{2} \delta\theta^T g \delta\theta \tag{2}$$

In other words, the Fisher-Rao metric is equal to, within a constant, a quadratic form with the Hessian being the second derivative of the Kullback-Leibler distance. Thus given two parametric densities, we can formulate a path length between them as

$$s = \int_0^1 \sum_{i=1}^{M} \sum_{j=1}^{M} g_{ij} \dot{\theta}^i \dot{\theta}^j dt \tag{3}$$

where M is the cardinality of the set $\{\theta^i\}$ and $\dot{\theta}^i = \frac{d\theta^i}{dt}$ is the parameter time derivative. Technically, (3) is the square of the geodesic distance, but has the same minimizer as $\int_0^1 \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{M} g_{ij} \dot{\theta}^i \dot{\theta}^j} dt$. The functional (3) is minimized using standard calculus of variations techniques leading to the following Euler-Lagrange equations

$$\frac{\delta s}{\delta \theta^k} = -2 \sum_{i=1}^{M} g_{ki} \ddot{\theta}^i + \sum_{i=1}^{M} \sum_{j=1}^{M} \left\{ \frac{\partial g_{ij}}{\partial \theta^k} - \frac{\partial g_{ik}}{\partial \theta^j} - \frac{\partial g_{kj}}{\partial \theta^i} \right\} \dot{\theta}^i \dot{\theta}^j = 0. \tag{4}$$

This is a highly non-linear system of partial differential equations (PDEs) and not analytically solvable except in special cases. One can use gradient descent to find a local solution to the system with update equations

$$\theta^k_{\tau+1}(t) = \theta^k_\tau(t) - \alpha_\tau \frac{\delta s}{\delta\theta^k_\tau(t)}, \forall t \tag{5}$$

where τ represents the iteration step and α the step size.

In (5), the path parameter t has been discretized. Consequently, you have to use discrete approximations to the derivatives $\dot{\theta}^k = \theta^k(t+1) - \theta^k(t)$ and $\ddot{\theta}^k = \theta^k(t+1) - 2\theta^k(t) + \theta^k(t-1)$.

- Let $p(x|\theta) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x-\mu^1)^2\} + \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x-\mu^2)^2\}$. Assume that $(\mu^1(0), \mu^2(0)) = (-1,-1)$ and $(\mu^1(1), \mu^2(1)) = (1,1)$. Divide the path interval t into ten time steps. Initialize the geodesic by a straight line from $(-1,-1)$ to (1,1). In order to compute g_{ij} and its derivatives, you'll have to numerically perform the integration in $g_{ij}(\theta) = \int p(\mathbf{x}|\theta)\frac{\partial}{\partial\theta^i}\log p(\mathbf{x}|\theta)\frac{\partial}{\partial\theta^j}\log p(\mathbf{x}|\theta)d\mathbf{x}$ and in the integrals of $\frac{\partial g_{ij}}{\partial\theta^k}$. Since x is one-dimensional, assume an integration interval of [-10,10]. You will have to carefully take care of underflow errors. Run a gradient descent algorithm until you get reasonable convergence of the entire path. At each step τ, you should choose a step size parameter α_τ such that $s(\tau+1) \leq s(\tau)$. Show the resulting geodesic.
- Prove that the Fisher geodesic for a simple Gaussian $p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$ is a straight line.