Machine Learning and Pattern Recognition
Tutorial 7

Instructor: Iain Murray

1. **Bring your questions:** As this is the last tutorial, make sure you ask about anything that's unclear from the course so far, including previous tutorial questions. Check the worked answers for past tutorials before asking about them.

2. **KL-divergence:** Sketch an elliptical probability density contour (enclosing the region of highest probability) for a two-dimensional, strongly-correlated Gaussian distribution p. Add contours to your sketch for two different q distributions, both with diagonal covariance: one minimizing $\mathrm{KL}(p||q)$ and the other minimizing $\mathrm{KL}(q||p)$. Label your contours.

   If keen: derive the diagonal covariances Λ of each q distribution for a p distribution with covariance Σ. For neatness, assume all the Gaussians are zero mean.

3. **KL-divergence and Monte Carlo:**

   Importance sampling uses a proposal distribution q to estimate an expectation under 'true distribution' p. Would it be a good idea to match these distributions using a KL-divergence? If so, which KL should be preferred and why?

   The Metropolis algorithm often uses Gaussian proposals centered on the current state: $q(\mathbf{x}'; \mathbf{x}) = \mathcal{N}(\mathbf{x}'; \mathbf{x}, \Sigma)$. Would it be a good idea to fit Σ by matching q (with mean set to $\mathbf{x}$) to the target distribution by minimizing a KL? If so, which?

4. **Optional:** Differentiating Gaussian integrals and Monte Carlo:

   In this course, variational approximations are always Gaussian $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$. Variational cost functions include expectations under this distribution of the form:

$$I = E_q\left[f(\theta)\right] = \int f(\theta)\, \mathcal{N}(\theta; \mu, \Sigma)\, \mathrm{d}\theta, \qquad (1)$$

   where f is the log-likelihood, or its contribution from a single data point. We can usually evaluate gradients $\mathbf{g}(\theta) = \nabla f(\theta)$, as used for maximum likelihood training.

   An empirical average over samples $\theta \sim q$ can approximate the above expectation. In 1-dimension, or for independent dimensions, a sample $\theta \sim \mathcal{N}(\mu, \sigma^2)$ can be generated by $\nu \sim \mathcal{N}(0, 1)$ and $\theta = \mu + \sigma\nu$. The full covariance generalization is $\theta = \mu + L\nu$, where $\nu \sim \mathcal{N}(0, I)$, L is the Cholesky decomposition of the covariance, the lower-triangular matrix such that $\Sigma = LL^\top$. The Cholesky decomposition is more useful than the covariance itself for most Gaussian computations, so we'll fit L as our variational parameter instead of Σ. (We'd also log-transform the positive diagonal elements of this matrix to create a fully-unconstrained optimization problem.)

   We need unbiased estimates of derivatives to optimize the variational parameters μ and L by stochastic gradient descent. *Approach 1:* Write the integral as an average under random choices ν:

$$I = E_\nu\left[f(\mu + L\nu)\right], \qquad \frac{\partial I}{\partial \mu} = E_\nu\left[\mathbf{g}(\mu + L\nu)\right], \qquad \frac{\partial I}{\partial L} = E_\nu\left[\mathbf{g}(\mu + L\nu)\nu^\top\right].$$

   Approach 2: We didn't have to do the change of variables $\theta \to \nu$. We could instead have differentiated inside the integral in (1):

$$\begin{aligned}\frac{\partial I}{\partial \mu} &= \int f(\theta) \frac{\partial}{\partial \mu} \mathcal{N}(\theta; \mu, \Sigma)\, \mathrm{d}\theta = \int f(\theta) \Sigma^{-1}(\theta - \mu) \mathcal{N}(\theta; \mu, \Sigma)\, \mathrm{d}\theta \\ &= E_{q(\theta)}[f(\theta)\Sigma^{-1}(\theta - \mu)], \quad \text{(and so on for } \frac{\partial I}{\partial L}\text{; messy but quite doable).}\end{aligned}$$

   Under either approach we can get a straight-forward Monte Carlo estimate of the derivatives.

   The questions: What are some pros and cons of these two approaches? You might want to consider a special case, such as $f(\theta) = \theta$. What are some pros and cons of stochastic variational inference compared to the Laplace approximation?