CS 194-10, Fall 2011
Assignment 0 Solutions

1. *In this question you will write a simple program in python that produces samples from various distributions, using* only *samples from the uniform distribution over the unit interval (that is, the only "source of randomness" you may use is calls to numpy.random.uniform()). The functions you will implement will enable sampling from four distributions: categorical (see also the multinomial for $n=1$), univariate Gaussian, multivariate Gaussian and general mixture distributions.*

   The solution is given in sampler_solution.py

2. *Prove that the sum of two independent Poisson variables is also a Poisson variable.*
   Let L and M be Poisson with rates λ_L and λ_M respectively, i.e.,

$$P(L=\ell)=e^{-\lambda_L}\frac{\lambda_L^\ell}{\ell!}\ ; \qquad P(M=m)=e^{-\lambda_M}\frac{\lambda_M^m}{m!}$$

   where ℓ and m range over the non-negative integers. Now let $N=L+M$; in general, the distribution of N is given by

$$P(N=n)=\sum_{\{\ell,m:\ell+m=n\}} P(L=\ell,M=m)\ .$$

   Here, L and M are independent and Poisson, and ℓ and m are non-negative integers, so we have

$$\begin{aligned} P(N=n) &= \sum_{\ell=0}^{n} P(L=\ell)P(M=n-\ell) \\ &= \sum_{\ell=0}^{n} e^{-\lambda_L}\frac{\lambda_L^\ell}{\ell!}e^{-\lambda_M}\frac{\lambda_M^{n-\ell}}{(n-\ell)!} \\ &= e^{-(\lambda_L+\lambda_M)}\sum_{\ell=0}^{n}\frac{\lambda_L^\ell}{\ell!}\frac{\lambda_M^{n-\ell}}{(n-\ell)!} \end{aligned}$$

   Now we notice that the summation looks suspiciously like a binomial expansion of $(\lambda_L+\lambda_M)^n$, except for a factor of $n!$ which we can easily supply:

$$\begin{aligned} P(N=n) &= e^{-(\lambda_L+\lambda_M)}\cdot\frac{1}{n!}\sum_{\ell=0}^{n}\frac{n!}{\ell!(n-\ell)!}\lambda_L^\ell\lambda_M^{n-\ell} \\ &= e^{-(\lambda_L+\lambda_M)}\cdot\frac{(\lambda_L+\lambda_M)^n}{n!} \end{aligned}$$

   which is Poisson with rate $\lambda_L+\lambda_M$.

3. *Let X_0 and X_1 be continuous random variables. Show that if*

$$P(X_0=x_0)=\alpha_0 e^{-\frac{1}{2}\left(\frac{(x_0-\mu_0)^2}{\sigma_0}\right)}$$
$$P(X_1=x_1|X_0=x_0)=\alpha_1 e^{-\frac{1}{2}\left(\frac{(x_1-x_0)^2}{\sigma}\right)}$$

   then there exists α, μ_1 and σ_1 such that

$$P(X_1=x_1)=\alpha e^{-\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1}\right)}\ .$$

Write explicitly the values of α, μ_1 and σ_1 that satisfy the above relations.

The trick here is to convert exponentials of arbitrary quadratic forms into exponentials of the standard Gaussian form. This is called *completing the square* in the exponent:

$$ax^2 + bx + c = a(x^2 + \tfrac{b}{a}x + \tfrac{c}{a}) = a[(x - \tfrac{-b}{2a})^2 - \tfrac{b^2}{4a^2} + \tfrac{c}{a}] \ .$$

Once we have a Gaussian for x_0 we can use the known value of the integral of a Gaussian, plus the fact that terms not involving x_0 can be moved outside the integral. Another suggestion: keep checking the dimensions of every term to make sure that erors have not crept in!

$$\begin{aligned}
P(X_1 = x_1) &= \int P(X_1 = x_1 \mid X_0 = x_0) P(X_0 = x_0) \mathrm{d}x_0 \\
&= \alpha_0\alpha_1 \int \exp\left(-\frac{1}{2}\frac{(x_0-\mu_0)^2}{\sigma_0^2} + \frac{(x_1-x_0)^2}{\sigma^2}\right) \mathrm{d}x_0 \\
&= \alpha_0\alpha_1 \int \exp\left(-\frac{1}{2}\frac{\sigma^2(x_0-\mu_0)^2 + \sigma_0^2(x_1-x_0)^2}{\sigma_0^2\sigma^2}\right) \mathrm{d}x_0 \\
&= \alpha_0\alpha_1 \int \exp\left(-\frac{1}{2}\frac{(\sigma^2+\sigma_0^2)x_0^2 - 2(\sigma^2\mu_0 + \sigma_0^2 x_1)x_0 + \sigma^2\mu_0^2 + \sigma_0^2 x_1^2}{\sigma_0^2\sigma^2}\right) \mathrm{d}x_0 \\
&= \alpha_0\alpha_1 \int \exp\left(-\frac{1}{2}\frac{x_0^2 - 2(\frac{\sigma^2\mu_0+\sigma_0^2 x_1}{(\sigma^2+\sigma_0^2)})x_0 + (\frac{\sigma^2\mu_0+\sigma_0^2 x_1}{(\sigma^2+\sigma_0^2)})^2}{\sigma_0^2\sigma^2/(\sigma^2+\sigma_0^2)}\right) \exp\left(-\frac{1}{2}\frac{\sigma^2\mu_0^2 + \sigma_0^2 x_1^2 - \frac{(\sigma^2\mu_0+\sigma_0^2 x_1)^2}{\sigma^2+\sigma_0^2}}{\sigma_0^2\sigma^2}\right) \mathrm{d}x_0
\end{aligned}$$

Now, we have an integral over an expression in Gaussian form. Suppose the normalization constant for this is α'; then the above expression reduces to

$$\begin{aligned}
P(X_1 = x_1) &= \frac{\alpha_0\alpha_1}{\alpha'} \exp\left(-\frac{1}{2}\frac{(\sigma^2\mu_0^2 + \sigma_0^2 x_1^2)(\sigma^2+\sigma_0^2) - (\sigma^2\mu_0 + \sigma_0^2 x_1)^2}{\sigma_0^2\sigma^2(\sigma^2+\sigma_0^2)}\right) \\
&= \frac{\alpha_0\alpha_1}{\alpha'} \exp\left(-\frac{1}{2}\frac{[\sigma_0^2(\sigma^2+\sigma_0^2) - \sigma_0^4]x_1^2 - 2\sigma^2\sigma_0^2\mu_0 x_1 + \sigma^2\mu_0^2(\sigma^2+\sigma_0^2) - \sigma^4\mu_0^2}{\sigma_0^2\sigma^2(\sigma^2+\sigma_0^2)}\right) \\
&= \frac{\alpha_0\alpha_1}{\alpha'} \exp\left(-\frac{1}{2}\frac{x_1^2 - 2\mu_0 x_1 + \mu_0^2}{\sigma^2+\sigma_0^2}\right) = \frac{\alpha_0\alpha_1}{\alpha'} \exp\left(-\frac{1}{2}\frac{(x_1-\mu_0)^2}{\sigma^2+\sigma_0^2}\right)
\end{aligned}$$

which is a Gaussian distribution with mean $\mu_1 = \mu_0$ and variance $\sigma_1^2 = \sigma^2 + \sigma_0^2$. Since $P(X_1 = x_1)$ is by construction a probability distribution, it is already normalized, so the normalization coefficient $\frac{\alpha_0\alpha_1}{\alpha'}$ is just the one required for a Gaussian of the given variance, i.e., $\alpha = 1/\sqrt{2\pi(\sigma^2+\sigma_0^2)}$.

4. *Find the eigenvalues and eigenvectors of the following matrix:*

$$\mathcal{A} = \begin{pmatrix} 13 & 5 \\ 2 & 4 \end{pmatrix} \ .$$

The eigenvectors $\mathbf{x} = (x_1, x_2)^T$ and eigenvalues λ satisfy $\mathcal{A}\mathbf{x} = \lambda\mathbf{x}$:

$$\begin{pmatrix} 13 & 5 \\ 2 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

or

$$\begin{aligned}
13x_1 + 5x_2 &= \lambda x_1 \\
2x_1 + 4x_2 &= \lambda x_2 \ .
\end{aligned}$$

You can solve this by straightforward algebra—for example, substituting $x_1 = \frac{\lambda-4}{2}x_2$ from the second equation into the first equation, dividing through by x_2, and rearranging, we get $\lambda^2 - 17\lambda + 42 = 0 = (\lambda-3)(\lambda-14)$, and hence the eigenvalues are 3 and 14. For $\lambda = 3$, we have $x_2 = -2x_1$, so the unit eigenvector is $(\sqrt{\frac{1}{5}}, -\sqrt{\frac{4}{5}})^T$; for $\lambda = 14$, we have $x_1 = 5x_2$ so the unit eigenvector is $(\sqrt{\frac{25}{26}}, -\sqrt{\frac{1}{26}})^T$.

A more systematic and general approach is to use the so-called *characteristic polynomial* equation:

$$\det(\mathcal{A} - \lambda\mathbf{I}) = 0 \ .$$

(That the eigenvalues satisfy this equation is explained on the Wikipedia page for eigenvalues and eigenvectors.) For a 2×2 matrix this reduces to $\lambda^2 - \mathrm{tr}(\mathcal{A})\lambda + \det\mathcal{A} = 0$; then we proceed as above.

5. *Provide one example for each of the following cases, where A and B are 2×2 matrices*:

   (a) $(\mathcal{A}+\mathbf{B})^2 \neq \mathcal{A}^2 + 2\mathcal{A}\mathbf{B} + \mathbf{B}^2$. The correct expansion of the LHS is $\mathcal{A}^2 + \mathcal{A}\mathbf{B} + \mathbf{B}\mathcal{A} + \mathbf{B}^2$, which is not identical to the RHS because matrix multiplication is not in general commutative. So we just need to find any pair of matrices for which commutativity fails, e.g.,

   $$\mathcal{A} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}; \quad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \ .$$

   (b) $AB = 0,\ A \neq 0,\ B \neq 0$. The pair of matrices in the preceding answer nearly works. We can zero out the one non-zero entry easily by flipping one of the signs:

   $$\mathcal{A} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}; \quad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix} \ .$$

6. *Let $\mathbf{u}$ denote a real vector normalized to unit length, i.e., $\mathbf{u}^T\mathbf{u} = 1$. Show that*

   $$\mathcal{A} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T$$

   is orthogonal, i.e., that $\mathcal{A}^T\mathcal{A} = \mathbf{I}$.
   This is a cute exercise in using the rules of matrix algebra. We have

   $$\begin{aligned}
   \mathcal{A}^T\mathcal{A} &= (\mathbf{I} - 2\mathbf{u}\mathbf{u}^T)^T(\mathbf{I} - 2\mathbf{u}\mathbf{u}^T) \qquad \text{(given)} \\
   &= (\mathbf{I} - 2\mathbf{u}\mathbf{u}^T)(\mathbf{I} - 2\mathbf{u}\mathbf{u}^T) \qquad \text{(using } \mathbf{I}^T{=}\mathbf{I},\ (\mathcal{A}\mathbf{B})^T{=}\mathbf{B}^T\mathcal{A}^T\text{)} \\
   &= \mathbf{I} - 4\mathbf{u}\mathbf{u}^T + 4(\mathbf{u}\mathbf{u}^T)(\mathbf{u}\mathbf{u}^T) \qquad \text{(expanding out)} \\
   &= \mathbf{I} - 4\mathbf{u}\mathbf{u}^T + 4\mathbf{u}(\mathbf{u}^T\mathbf{u})\mathbf{u}^T \qquad \text{(associativity)} \\
   &= \mathbf{I} - 4\mathbf{u}\mathbf{u}^T + 4\mathbf{u}\mathbf{u}^T \qquad \text{((unit length)} \\
   &= \mathbf{I} \ .
   \end{aligned}$$

7. *A function f is convex on a given set iff for $\lambda \in [0,1]$ and for all x, y in the set, the following holds,*

   $$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \, .$$

   Moreover, a univariate function $f(x)$ is convex on a set if its second derivative $f''(x)$ is non-negative everywhere in the set. Prove the following assertions:

   (a) $f(x) = x^3$ *is convex for* $x \geq 0$.
   The second derivative $f''(x) = 6x \geq 0$ for $x \geq 0$.

   (b) $f(x_1, x_2) = \max(x_1, x_2)$ *is convex on* $\mathbb{R}^2$.

(c) *If univariate f and g are convex on a set, $f+g$ is convex on the same set.*
Because f and g are univariate, we can use the derivative rule:

$$(f+g)''(x) = f''(x) + g''(x) \geq 0 \ .$$

The convexity of $f+g$ actually holds in general, not just for univariate functions:

$$\begin{aligned}(f+g)(\lambda x + (1-\lambda)y) &= f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \\ &\leq \lambda f(x) + (1-\lambda)f(y) + \lambda g(x) + (1-\lambda)g(y) \\ &= \lambda(f+g)(x) + (1-\lambda)(f+g)(y) \ .\end{aligned}$$

(d) *If univariate f and g are convex and nonnegative on a set, and have their minimum within the set at the same point, then fg is convex on the same set.*
Here we use the derivative rule. We have

$$\begin{aligned}(fg)''(x) &= (f'g + fg')'(x) \qquad \text{(derivative of product)} \\ &= (f''g + 2f'g' + fg'')(x) \qquad \text{(derivative of product).}\end{aligned}$$

Now, because f and g are convex and non-negative, we have immediately that $(f''g)(x)$ and $(fg'')(x)$ are non-negative on the set. Let the common minimum of f and g be x^*; for $x \leq x^*$, f' and g' are both non-positive, so $f'g'$ is non-negative, whereas for $x > x^*$, f' and g' are both non-negative, so $f'g'$ is also non-negative; hence, $(fg)''(x)$ is the sum of three non-negative terms. (Note that this proof extends to the case where $f' = g' = 0$ over an extended interval rather than just at a single point.)

8. *The* entropy *of a categorical distribution on K values is defined as*

$$H(p) = -\sum_{i=1}^{K} p_i \log p_i \ .$$

Using the method of Lagrange multipliers, find the categorical distribution that has the highest entropy.
We will maximize h with respect to each p_i, subject to the normalization constraint that $\sum_{i=1}^{K} p_i = 1$. To use the method of Lagrange multipliers, we first write the constraint as $\sum_{i=1}^{K} p_i - 1 = 0$ and then write the Lagrangian function as

$$\begin{aligned}\Lambda &= -\sum_{i=1}^{K} p_i \log(p_i) + \lambda\left(\sum_{i=1}^{K} p_i - 1\right) \\ \frac{\partial \Lambda}{\partial p_i} &= -\log p_i - 1 + \lambda = 0 \ .\end{aligned}$$

which we can rewrite as $\log p_i = \lambda - 1$. This equation says that all the p_is have the same value; to satisfy the normalization constraint, that value must be $1/K$. *In general*, to show that a stationary point is a maximum, we have to show that the Hessian is negative semidefinite; in this particular case, however, we observe that (1) the stationary point p^* is unique, and (2) $H(p^*) = \log K \geq 0$, whereas $H([1, 0, 0, \ldots, 0]) = 0$, so p^* must be a maximum.