Problem 1 is courtesy of Prasad.

1. Taking the expectation of (11.2)

$$\mathrm{E}[\hat{f}] = \frac{1}{L}\sum_{l=1}^{L}\mathrm{E}[f(\mathbf{z}^{(l)})]$$
$$= \mathrm{E}[f]$$

since each $\mathbf{z}^{(l)}$ is chosen independently from $p(\mathbf{z})$.

Equation (11.3) can be proved using the facts:

a) If $Y = aX$ then $\mathrm{Var}(Y) = a^2\mathrm{Var}(X)$

b) If X_i are independent and $Y = \sum_i X_i$ then $\mathrm{Var}(Y) = \sum_i \mathrm{Var}(X_i)$

Taking the variance of (11.2)

$$\mathrm{Var}[\hat{f}] = \frac{1}{L^2}\sum_{l=1}^{L}\mathrm{Var}[f(\mathbf{z}^{(l)})]$$
$$= \frac{1}{L^2}L\mathrm{Var}[f]$$
$$= \frac{1}{L}\mathrm{E}\left[(f - \mathrm{E}[f])^2\right]$$

Problem 2 is courtesy of Ernesto.

2. (a) The blue curve in figure 11.2 is called the cumulative distribution function (CDF) of the random variable y.

   (b) In order to sample from the red distribution, we must sample numbers from the vertical axis uniformly between 0 and 1. We then draw a horizontal line and see where this horizontal line intersects the blue curve (the CDF of y), and take the y-value (the horizontal value in the figure) of that point of intersection as our sample point.

   More of the samples will end up near the peaks of the red curve because where the red curve has peaks, the blue curve grows rapidly, which means it takes up more vertical space per unit of horizontal space. Since the vertical axis values are being sampled uniformly, there is a higher chance that the horizontal values will end up in the places where the blue curve grows rapidly.

   (c) To show that 11.5 is true, we must show that the PDF of y is given by $p(z)|\frac{dz}{dy}|$. Let $F(z)$ be the CDF of z and $Q(y)$ be the CDF of y, such that $F(z_0) = \mathbf{P}(z \leq z_0)$ and $Q(y_0) = \mathbf{P}(y \leq y_0)$. Remembering that $y = f(z)$, we have

$$\begin{aligned} Q(y_0) &= \mathbf{P}(y \leq y_0) \\ &= \mathbf{P}(f(z) \leq y_0) \\ &= \mathbf{P}(z \leq f^{-1}(y_0)) \\ &= F(z_0), \end{aligned}$$

   where we have used $z = f^{-1}(y)$. Since it is the PDF of y that we are interested in, we must take the derivative of Q in order to get it. Thus

$$\begin{aligned} Q'(y_0) &= \frac{d}{dy} F(z_0) \\ &= F'(z_0) \frac{dz}{dy} \\ &= p(z) \frac{dz}{dy}. \end{aligned}$$

   If we call the PDF of y $p(y)$ (only because the author does it), and if f is strictly increasing (so is f^{-1}), we get the desired result

$$p(y) = p(z) \left| \frac{dz}{dy} \right|.$$

Problem 3 is courtesy of Abhishek.

3) PRML 11.2) Suppose $z \sim Unif(0,1)$ and $y = h^{-1}(z)$. Show that $y \sim p(y)$.
ans

$$\begin{aligned} P(y \leq y_0) &= P(h(y) \leq h(y_0)) \\ &= P(z \leq h(y_0)) \\ &= h(y_0) \end{aligned}$$

So y has cdf $h(.)$, hence $y \sim p(y)$.

Problem 4 is courtesy of Prasad.

4. Let us begin by deriving equation 11.45. Consider the acceptance probability for each step of the Metropolis-Hastings algorithm

$$A(\mathbf{z}', \mathbf{z}) = \min\left(1, \frac{p(\mathbf{z}')q(\mathbf{z}|\mathbf{z}')}{p(\mathbf{z})q(\mathbf{z}'|\mathbf{z})}\right)$$

Multiplying both sides by $p(\mathbf{z})q(\mathbf{z}'|\mathbf{z})$, which amounts to multiplying either the first or the second argument of the min function whichever is smaller. This deduces to

$$p(\mathbf{z})q(\mathbf{z}'|\mathbf{z})A(\mathbf{z}', \mathbf{z}) = \min(p(\mathbf{z}')q(\mathbf{z}|\mathbf{z}'), p(\mathbf{z})q(\mathbf{z}'|\mathbf{z}))$$

By the symmetry of $\mathbf{z}$ and $\mathbf{z}'$, we can interchange the two and write equation 11.45

$$\begin{aligned} p(\mathbf{z})q(\mathbf{z}'|\mathbf{z})A(\mathbf{z}', \mathbf{z}) &= \min(p(\mathbf{z}')q(\mathbf{z}|\mathbf{z}'), p(\mathbf{z})q(\mathbf{z}'|\mathbf{z})) \\ &= \min(p(\mathbf{z})q(\mathbf{z}'|\mathbf{z}), p(\mathbf{z}')q(\mathbf{z}|\mathbf{z}')) \\ &= p(\mathbf{z}')q(\mathbf{z}|\mathbf{z}')A(\mathbf{z}, \mathbf{z}') \end{aligned}$$

which implies that the Markov chain formed by the samples of the MH algorithm has $p(\mathbf{z})$ as its stationary distribution and the transition probabilities of the Markov chain are given by $T(\mathbf{z}', \mathbf{z}) = q(\mathbf{z}|\mathbf{z}')A(\mathbf{z}, \mathbf{z}')$. In order to make it

Problem 5 is courtesy of Abhishek.

5) Show that the MCMC sampling is ergodic.

ans Let the intial sample be from some probability distribution $r_0(x)$ and the Markov Chain has stationary distribution $\pi(x)$. Let us show that the distribution at time n can be expressed as:

$$p_n(x) = (1-(1-\nu)^n)\pi(x) + (1-\nu)^n r_n(x) \tag{5.1}$$

where $\nu \in (0,1)$ and $r_n(x)$ is a probability distribution.

To prove (5.1), we use induction on n.

When $n = 0$, (5.1) says

$$p_0(x) = r_0(x)$$

which is true by assumption. Suppose (5.1) holds for $n = 1, 2, \ldots, N$. Want to show that it holds for $n = N+1$. Using the definition of transition function as in 11.38, we can get $p_{N+1}(x)$ from $p_N(x)$ as follows:

$$p_{N+1}(x) = \sum_{x'} p_N(x')T(x',x) \tag{5.2}$$

Substitute the expression of $p_N(x)$ from (5.1) into (5.2) to get

$$p_{N+1}(x) = \sum_{x'} \left[(1-(1-\nu)^N)\pi(x') + (1-\nu)^N r_N(x')\right] T(x',x) \tag{5.3}$$

$$= \left(1-(1-\nu)^N\right) \sum_{x'} \pi(x')T(x',x) + (1-\nu)^N \sum_{x'} r_N(x')T(x',x) \tag{5.4}$$

$$= \left(1-(1-\nu)^N\right) \pi(x) + (1-\nu)^N \sum_{x'} r_N(x')T(x',x) \tag{5.5}$$

$$= \left(1-(1-\nu)^{N+1}\right) \pi(x) + \left(1-(1-\nu)^N\right) \pi(x) - \left(1-(1-\nu)^{N+1}\right) \pi(x)$$
$$+(1-\nu)^N \sum_{x'} r_N(x')T(x',x) \tag{5.6}$$

$$= \left(1-(1-\nu)^{N+1}\right) \pi(x) - \nu(1-\nu)^N \pi(x) + (1-\nu)^N \sum_{x'} r_N(x')T(x',x) \tag{5.7}$$

$$= \left(1-(1-\nu)^{N+1}\right) \pi(x) + (1-\nu)^{N+1} \left[\frac{1}{1-\nu} \sum_{x'} r_N(x')T(x',x) - \frac{\nu}{1-\nu}\pi(x)\right] \tag{5.8}$$

$$= \left(1-(1-\nu)^{N+1}\right) \pi(x) + (1-\nu)^{N+1} r_{N+1}(x) \tag{5.9}$$

To get (5.5) from (5.4), I used the fact that π is a stationary distribution for the MC. In (5.9), $r_{N+1}(x)$ has the expression

$$r_{N+1}(x) = \frac{1}{1-\nu} \sum_{x'} r_N(x')T(x',x) - \frac{\nu}{1-\nu}\pi(x) \tag{5.10}$$

We need to show that this defines a probability distribution for suitable choice of ν. $\sum_{x'} r_N(x')T(x',x)$ defines a probability distribution over x (that is the distribution of x_{N+1} if x_N has distribution $r_N(x)$), so does $\pi(x)$. Hence $r_{N+1}(x)$ is a weighted sum of two probability distributions, the weights adding up to 1, so

$$\sum_x r_{N+1}(x) = 1$$

Remains to show that $r_{N+1}(x) \geq 0$ for all x. That is equivalent to showing

$$\sum_{x'} r_N(x')T(x',x) \geq \nu\pi(x) \tag{5.10}$$

For that assume $T(x',x) > 0$ for all x, x'; and

$$m \doteq \inf_{x,x'} T(x',x) > 0$$

Then

$$\sum_{x'} r_N(x')T(x',x) \geq m \sum_{x'} r_N(x') = m$$

So if $\nu \leq m$, then

$$\nu\pi(x) \leq \nu \leq m \leq \sum_{x'} r_N(x')T(x',x)$$

which proves (5.10). Hence by induction, we have proved (5.1). Then from (5.1)

$$\begin{aligned} |\pi(x) - p_n(x)| &= |(1-\nu)^n\pi(x) - (1-\nu)^n r_n(x)| \\ &= (1-\nu)^n |\pi(x) - r_n(x)| \\ &\leq (1-\nu)^n \end{aligned} \tag{5.11}$$

Since we choose $0 < \nu < 1$, (5.11) converges to 0 as $n \to \infty$, which means $p_n(x) \to \pi(x)$. The inequality in (5.11) and the bound m on ν suggest that closer the transition probabilities are to 1, faster will be the rate of convergence.