T.61.5140 Machine Learning: Advanced Probablistic Methods
Hollmén, Raiko (Spring 2008)
Problem session, 4th and 11th of April, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

These are programming excercises that are discussed on 4th of April and demonstrated on 11th of April. The Matlab files are on the course webpage.

Let us use a notation $N(x \mid \mu, \sigma^2)$ to denote a Gaussian distribution of x with mean μ and variance σ^2.

1. Study the unknown distribution $p_u(x)$ by drawing samples from it. You can get values of $p_u(x)$ by calling the Matlab function unknown_ p.m with the argument x which may be a scalar or a vector. Use rejection sampling (page 528) with a proposal distribution $q(x) = N(0,2)$ (Gaussian with mean 0 and variance 2) and scaling k=1.5 (You get values of $kq(x)$ as 1.5*gaussian(x,0,2)). (a) Plot a figure that shows the accepted samples with a black dot ('.k') and rejected samples with a cyan dot ('.c'). Use both 100 and 1000 samples. (b) Estimate the expected values $E(x)$ and $E(\tanh(x))$ over distribution $p_u(x)$ by using the accepted samples.

Solution:

```
% Problem 1(a)
figure(1); clf;
xx=-5:0.01:5;
k=1.5;
yy=k*gaussian(xx,0,2);
for i=1:2,
  nsamples = 100*(i==1) + 1000*(i==2);
  x = randn(1,nsamples)*sqrt(2);
  y = k*rand(1,nsamples).*gaussian(x,0,2);
  accepted = find(y<unknown_p(x));
  subplot(1,2,i);
  plot(xx,[unknown_p(xx);yy]);
  hold on;
  plot(x,y,'.c');
  plot(x(accepted),y(accepted),'.k');
```

```
end;
```

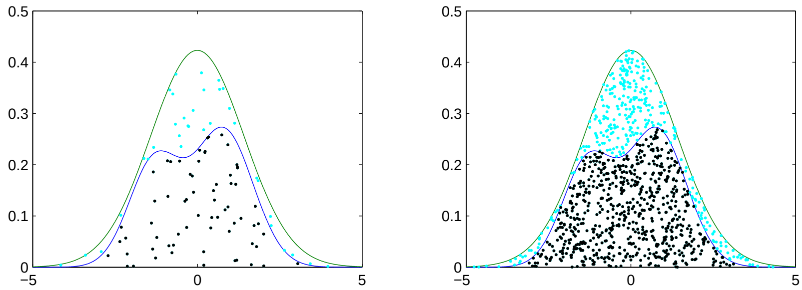

Figure 1: Problem 1. Left: 100 samples, Right: 1000 samples.

Solution of 1(b) is given with 2(b) and 2(c).

2. Study the same distribution $p_u(x)$ by using importance sampling (page 532). (a) Plot the importance weights $p_u(x)/q(x)$. (b) Estimate $E(x)$ and $E(\tanh(x))$ over distribution $p_u(x)$ by weighting all samples. (c) Was it more accurate than rejection sampling?

Solution:

```
% Problem 2(a)
figure(2); clf;
importance = unknown_p(xx)./gaussian(xx,0,2);
plot(xx,[importance; unknown_p(xx); gaussian(xx,0,2)]);
legend('importance p_u/q','unknown p_u','proposal q');
```

2(b). For estimating $E(x)$ and $E(\tanh(x))$, and to compare the accuracy, both sampling methods were tested 1000 times. By finding the mean and the standard deviation of these repetitions for each method, we can estimate the accuracy. The means were close to $E(x) = -0.04$ and $E(\tanh(x)) = -0.007$ for both methods, but the standard deviations were 15%–20% larger for rejection sampling than importance sampling.

```
% accuracy comparison 1(b),2(b),2(c)
% repeat the estimation a thousand times
for i=1:1000,
```

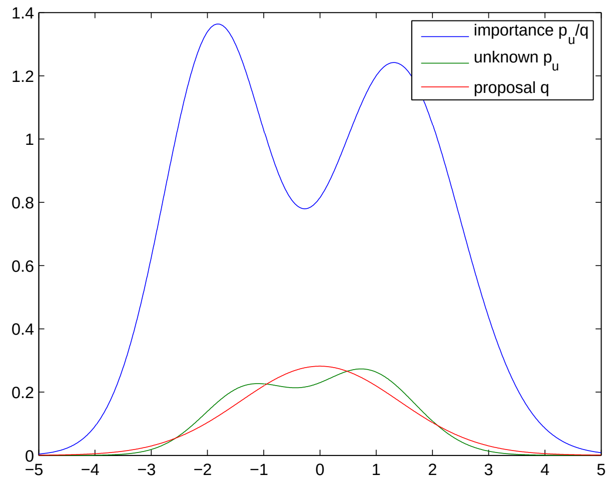

Figure 2: Problem 2(a). Importance weight $p_u(x)/q(x)$ as a function of x.

```
  nsamples = 100;
  x = randn(1,nsamples)*sqrt(2);
  y = k*rand(1,nsamples).*gaussian(x,0,2);
  accepted = find(y<unknown_p(x));
  importance = unknown_p(x)./gaussian(x,0,2);
  Ex1(i) = mean(x(accepted));
  Etanhx1(i) = mean(tanh(x(accepted)));
  Ex2(i) = sum(x.*importance)/sum(importance);
  Etanhx2(i) = sum(tanh(x).*importance)/sum(importance);
end;
mean(Ex1)
mean(Ex2)
std(Ex1)
std(Ex2)
mean(Etanhx1)
mean(Etanhx2)
std(Etanhx1)
std(Etanhx2)
```

3. Let us study a model with two variables: $x_1 \sim N(x_1 \mid 0, 1.0)$ and $x_2 \sim N(x_1 \mid x_1, 0.1)$ where the two Gaussians are independent. (a) Solve

$p(x_1 \mid x_2)$. (b) Initialize $x_1^{(0)} = x_2^{(0)} = 0$ and plot 50 samples from the model by Gibbs sampling (page 542) with a line connecting consequtive samples.

Solution:

(a) Let us first prove a lemma, the product of two independent Gaussians (ignoring normalization constants). Let us use the logarithmic scale:

$$\ln\left[N(x \mid \mu_1, \sigma_1^2) N(x \mid \mu_2, \sigma_2^2)\right] \tag{1}$$

$$= \frac{-(x-\mu_1)^2}{2\sigma_1^2} + \frac{-(x-\mu_2)^2}{2\sigma_2^2} + \text{const} \tag{2}$$

$$= -\frac{1}{2\sigma_1^2}x^2 + \frac{\mu_1}{\sigma_1^2}x - \frac{1}{2\sigma_2^2}x^2 + \frac{\mu_2}{\sigma_2^2}x + \text{const} \tag{3}$$

$$= -\frac{1}{2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)x^2 + \left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right)x + \text{const} \tag{4}$$

$$= -\frac{1}{2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)\left(x^2 - 2\frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}x\right) + \text{const} \tag{5}$$

$$= -\frac{1}{2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)\left(x - \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\right)^2 + \text{const} \tag{6}$$

$$= \ln N\left(x \mid \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}, \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}\right). \tag{7}$$

Note that the ignored constant changes between lines (5) and (6). We showed that the product of two Gaussians is another Gaussian whose mean is a weighted average of the two means, and the variance is smaller than either of the two variances. This result might be useful elsewhere, too.

Then let us study the problem at hand, starting with Bayes theorem:

$$
\begin{aligned}
p(x_1 \mid x_2) &= \frac{p(x_2 \mid x_1)p(x_1)}{p(x_2)} && (8)\\
&= \text{const}\exp\frac{-(x_1 - x_2)^2}{2\cdot 0.1}\exp\frac{-x_1^2}{2\cdot 1.0} && (9)\\
&= N(x_1 \mid x_2, 0.1)N(x_1 \mid 0, 1.0)\text{const} && (10)\\
&= N\left(x_1 \mid \frac{\frac{x_2}{0.1}+\frac{0}{1.0}}{\frac{1}{0.1}+\frac{1}{1.0}}, \left(\frac{1}{0.1}+\frac{1}{1.0}\right)^{-1}\right) && (11)\\
&= N\left(x_1 \mid \frac{10}{11}x_2, \frac{1}{11}\right). && (12)
\end{aligned}
$$

The problem is now solved, but let us still discuss the term $p(x_2 \mid x_1)$. When we think about it the normal way, x_1 is a constant and p gives the distribution over x_2. But $p(x_2 \mid x_1)$ is actually a function over two variables, x_1 and x_2. If we assume x_2 is known, $p(x_2 \mid x_1)$ is the likelihood function of x_1. Likelihood functions are not distributions (they don't have to normalize to 1 - consider for instance the case where x_2 is independent of x_1 and the likelihood function is a positive constant $p(x_2 \mid x_1) = p(x_2) > 0$ with respect to x_1). But in this case, the likelihood function is $\text{const}\cdot\exp\frac{-(x_1-x_2)^2}{2\cdot 0.1}$ which has a Gaussian form so we can write $p(x_2 \mid x_1) = N(x_1 \mid x_2, 0.1)\cdot\text{const}$.

```
% Problem 3(b)
figure(3); clf; hold on;
x1(1) = 0;
x2(1) = 0;
for iter=1:50,
  x1(iter+1) = 10/11*x2(iter)+randn(1)*sqrt(1/11);
  plot([x1(iter);x1(iter+1)],[x2(iter);x2(iter)],'b');
  x2(iter+1) = x1(iter+1)+randn(1)*sqrt(0.1);
  plot([x1(iter+1);x1(iter+1)],[x2(iter);x2(iter+1)],'b');
  plot([x1(iter);x1(iter+1)],[x2(iter);x2(iter+1)],'r');
end;
```

4. Let us study a model with three variables, x_1, x_2, and x_3. The model is such that x_1 and x_2 are drawn from the same distribution $p_u(x)$ as before,

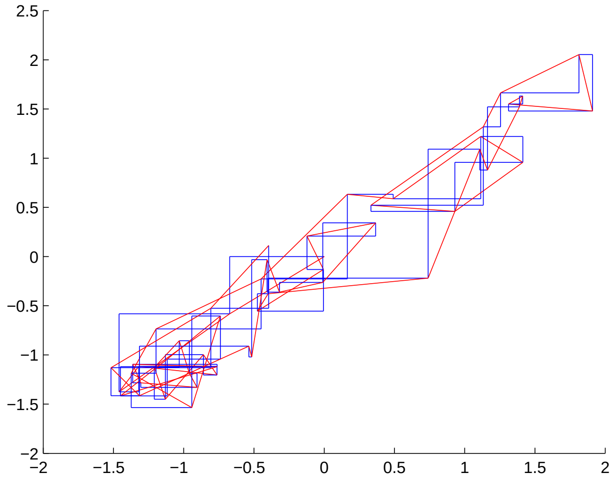

Figure 3: Problem 3(b). Blue lines show the alternate updates of x_1 and x_2 while the red lines show the two updates combined.

and $x_3 \sim N(x_3 \mid x_1 + x_2, 1.0)$. Draw samples from $p(x_1, x_2 \mid x_3 = 1.0)$ by using Metropolis algorithm (page 538). Use a proposal distribution $x_1^* \sim N(x_1^* \mid x_1^{(\tau)}, 1.0)$ and $x_2^* \sim N(x_2^* \mid x_2^{(\tau)}, 1.0)$ (simultaneously, not alternately as in Gibbs sampling). Plot 50 samples with a line connecting consequtive samples '-' and rejected proposals with a cross 'x'.

```
% Problem 4
figure(4); clf;
x3 = 1.0;
for i=1:8,
  % variance of the proposed jump:
  propvar = 0.01*(mod(i,4)==1)+0.1*(mod(i,4)==2) ...
            +1.0*(mod(i,4)==3)+10.0*(mod(i,4)==0);
  nsamples = 50*(i<=4)+500*(i>=5);
  subplot(2,4,i); axis([-4 4 -4 4]); hold on;
  title(sprintf('nsamples = %d, prop var = %3.2f',nsamples,propvar))
  x1(1) = 0;
  x2(1) = 0;
  for iter = 1:nsamples,
```

```
    prob_old = unknown_p(x1(iter))*unknown_p(x2(iter)) ...
               *gaussian(x3,x1(iter)+x2(iter),1.0);
    x1prop = x1(iter)+randn(1)*sqrt(propvar);
    x2prop = x2(iter)+randn(1)*sqrt(propvar);
    prob_prop = unknown_p(x1prop)*unknown_p(x2prop) ...
                *gaussian(x3,x1prop+x2prop,1.0);
    if prob_prop/prob_old > rand(1), % accept proposed jump?
      x1(iter+1)=x1prop;
      x2(iter+1)=x2prop;
      plot([x1(iter);x1(iter+1)],[x2(iter);x2(iter+1)],'b-');
    else,
      x1(iter+1)=x1(iter);
      x2(iter+1)=x2(iter);
      plot(x1prop,x2prop,'rx');
    end;
  end;
end;
```

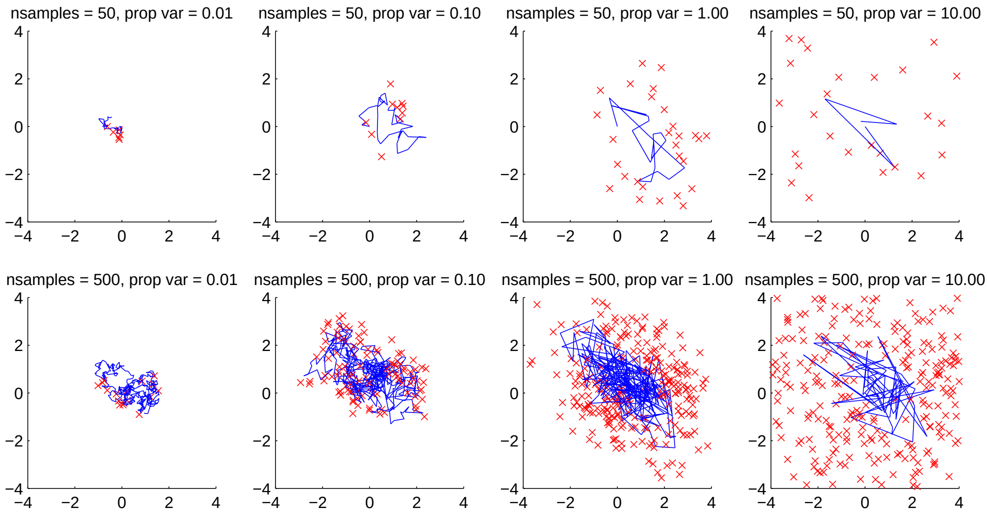

Figure 4: Problem 4. The top row has 50 samples while the bottom row has 500 samples. The variance of the proposal distribution is 0.01 (left), 0.1 (second from left), 1.0 (third from left), and 10.0 (right). Note how the samples are close to each other in the left because the variance is too small and most steps are rejected in the right because the variance is too large.