CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Reversible Jump MCMC

Lecturer: Michael I. Jordan *Scribes: Shawn Chang and Barry Chen*

1 Approximation of Marginal-likelihood (continued)

Continuing our discussion on marginal likelihood approximation methods and the GMM example, we treat the GMM as a graphical model. After moralization of the graph, we may proceed to approximate $p(k|y)$ for each k. However, here we run into difficulty using only within-model approximation method, since different k will change the number of parameters we have and therefore the distribution for z. Thus, we need to consider between-model approximation methods such as Reversible Jump MCMC. But first let's recall the simpler Metropolis-Hastings method.

1.1 Metropolis-Hastings Method

Let x be a "state" and $\pi(x)$ a distribution we'd like to sample from. Set up a Markov Chain $p(x, x')$ that has $\pi(x)$ as its equilibrium distribution. A sufficient condition is the detailed balance, which states that $\pi(x)p(x, x') = \pi(x')p(x', x)$.

The Metropolis-Hastings method satisfies the detailed balance condition, and works as follows: when at x, propose a move to x' with probability $q(x, x')$, and then accept the proposal with probability $\alpha = \min\{1, \frac{\pi(x')q(x',x)}{\pi(x)q(x,x')}\}$, for all x and x'.

1.2 Reversible Jump MCMC

Reversible Jump (RJ) MCMC builds upon the Metropolis-Hastings method. Let's illustrate the main idea with a simple example. In Figure 1, when $k = 0$ we have the 1-D uniform distribution as shown by the elongated bar; when $k = 1$, we have a 2-D uniform distribution on the triangle. Let's see how we are able to "jump" between the two models.

Consider three possible moves:

- a) for $k = 0$, $x \to U(x - \epsilon, x + \epsilon)$.
- b) for $k = 1$, swap x_1 and x_2.
- c) Jump between $k = 0$ and $k = 1$ as:
  - for $k = 0$, choose u from $U(0, 1)$ and propose $(x_1, x_2) = (x, u)$
  - for $k = 1$, set $x = x_1$.

Note in move (c), we have a bijection $(x, u) \leftrightarrow (x1, x2)$.

At each step, a move is chosen as follows:

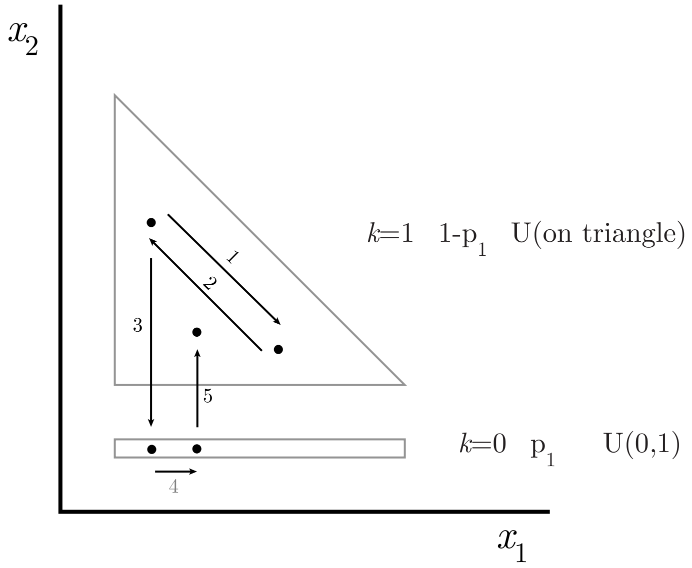

Figure 1: Example of Reversible Jump MCMC.

- for $k = 0$, choose (a) with probability $1 - r_1$ and (c) with probability r_1.
- for $k = 1$, choose (b) with probability $1 - r_2$ and (c) with probability r_2.

For example, as shown by the light arrows in Figure 1, we may jump within the triangle along arrow 1 and 2, from triangle to bar along arrow 3, within the bar along arrow 4, and back to the triangle along arrow 5.

In general, a proposal is accepted with probability

$$\alpha = \min\{1, \frac{(y|x')p(x')}{p(y|x)p(x)} \frac{g(k,u'|x')}{g(k,u|x)} \left|\frac{\partial(x',u')}{\partial(x,u)}\right|\} \tag{1}$$

where $g(k,u|x)$ is the joint distribution of k and u while at x, and $\left|\frac{\partial(x',u')}{\partial(x,u)}\right|$ is the determinant of Jacobian matrix. Note that (1) satisfies the detailed balance condition.

In our example,

$$\begin{aligned} \alpha &= \min\{1, \frac{p_2 I(x_2 < x_1)}{p_1} \frac{r_2}{r_2 \cdot 1} |1|\} && (2) \\ &= \min\{1, \frac{2p_2r_2}{p_1r_1}\} I(u < x) && (3) \end{aligned}$$

where $I()$ is the indicator function.

In general, for a given problem, we would like to first try exact inference; if intractable, we can try Gibb's sampling; if still difficult, we turn to the RJ MCMC methods. Back to our Gaussian mixture model example, one may use Reversible Jump MCMC to sample k while use the faster Gibb's sampling for other tractable nodes with conjugate prior distributions.

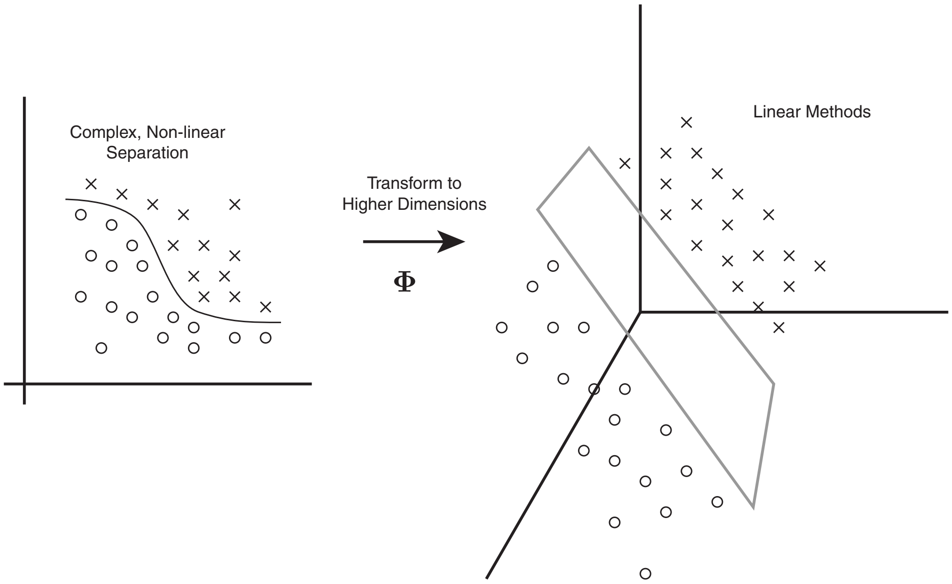

Figure 2: Illustration of kernel trick that transforms difficult problem into higher-dimensional but simpler problem.

2 Mercer Kernels

For certain tasks, such as some difficult classification problems, we may use kernel tricks to transform the problem into higher dimensional space where the the problem may be solved more easily. For example, in Figure 2, a classification boundary in the original 2-D space (left graph) may be of a complex nonlinear shape. However, once transformed into a 3-D space (right graph) through Φ, a simple hyper-plane may be adequate for classification.

Some simple examples of kernel functions include:

- Gaussian kernel: $k(x, x') = e^{-\frac{1}{2}\|x-x'\|^2} = \Phi(x)^T\Phi(x')$, where Φ is of infinite dimension.
- Polynomial kernel: $k(x, x') = (1 + x^T x')^n = \Phi(x)^T\Phi(x')$.

We will start our discussion on kernel methods by studying Support Vector Machines (SVM). As a classifier, there are two important ideas behind SVM:

- Optimal margin classifier.
- Kernel tricks.

The idea of optimal margin classifier is illustrated in Figure 3, where we would like to find a classification boundary that maximizes the minimum distance to each data class.

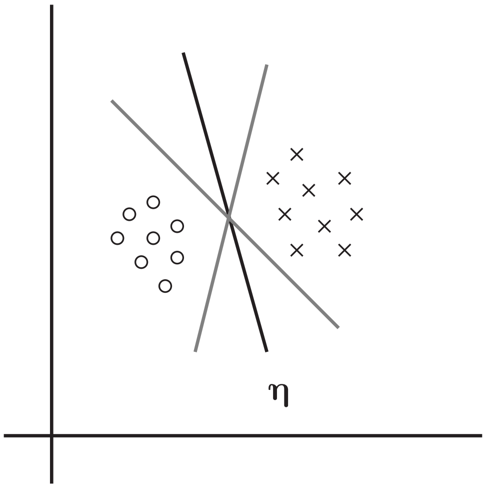

Figure 3: Illustration of the idea of Optimal Marginal Classifier in Support Vector Machine.