VE414 Lecture 16

Jing Liu

UM-SJTU Joint Institute

November 2, 2019

- According to the IEEE Computer Society Journal, the 10 algorithms below:
  - Metropolis Algorithm for Monte Carlo
  - Simplex Method for Linear Programming
  - Krylov Subspace Iteration Methods
  - The Decompositional Approach to Matrix Computations
  - The Fortran Optimizing Compiler
  - QR Algorithm for Computing Eigenvalues
  - Quicksort Algorithm for Sorting
  - Fast Fourier Transform
  - Integer Relation Detection
  - Fast Multipole Method

  had the greatest influence on the development and practice of science and engineering in the 20th century. The list is in chronological order.
- Metropolis-Hastings is a modern generalisation of Metropolis algorithm.

- Recall of the main reason that rejection sampling is not feasible for a large p, where $\mathcal{D} \in \mathbb{R}^p$ is the support of the target distribution $f_{\mathbf{Y}}$, is that it is often very difficult to come up with suitable proposal distribution $g_{\mathbf{Y}}$.
- The breakthrough parallel to Gibbs is to think **locally**! That is, instead of coming up with a proposal $g_{\mathbf{Y}}$ works for every where in $\mathcal{D}$ for every iteration, we consider conditional proposal distributions,

$$g_{\cdot|\mathbf{Y}^{(t-1)}}$$

  that changes from iteration to iteration depending on where we are, $\mathbf{Y}^{(t-1)}$.
- The price to pay in return of this convenience is having only a Markov Chain

$$\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \cdots, \mathbf{y}^{(t)}, \cdots, \mathbf{y}^{(n)}\}$$

  instead of a sample of independent realisations of $\mathbf{Y}$ from $f_{\mathbf{Y}}$.

Q: Can you see why Metropolis-Hastings could ever be better than Gibbs?

Algorithm 1: Metropolis-Hastings

Input : function $f_{\mathbf{Y}}$, and $g_{\mathbf{Y}|\mathbf{Y}^*}$, initial value $\mathbf{y}^{(0)}$, size n

Output : sample array $[\mathbf{y}^{(t)}]_{n \times p}$

Function `MH`($f_{\mathbf{Y}}$, $\mathbf{y}^{(0)}$, n):

  for $t \leftarrow 1$ **to** n **do**

    $\mathbf{z} \sim g_{\mathbf{Y}|\mathbf{Y}^*=\mathbf{y}^{(t-1)}}$; `/* draw from the proposal */`

    $\alpha \leftarrow \min\left\{1, \frac{f_{\mathbf{Y}}(\mathbf{z}) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y}^{(t-1)} \mid \mathbf{z}\right)}{f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{z} \mid \mathbf{y}^{(t-1)}\right)}\right\}$

    $v \sim \mathrm{Uniform}(0, 1)$; `/* draw uniform */`

    if $v \leq \alpha$ **then**

      $\mathbf{y}^{(t)} \leftarrow \mathbf{z}$; `/* accept the new */`

    else

      $\mathbf{y}^{(t)} \leftarrow \mathbf{y}^{(t-1)}$; `/* reject the new */`

    end if

  end for

  return $\left[\mathbf{y}^{(t)}\right]_{n \times p}$; `/* samples */`

end

Q: What does the quantity α in the Metropolis-Hastings algorithm represent?

$$\alpha = \min\left\{1, \frac{f_{\mathbf{Y}}\left(\mathbf{z}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y}^{(t-1)} \mid \mathbf{z}\right)}{f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{z} \mid \mathbf{y}^{(t-1)}\right)}\right\}$$

- It represents the probability of accepting

$$\mathbf{Y}^{(t)} = \mathbf{z}$$

given we have drawn $\mathbf{z}$ from the conditional proposal and $\mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}$.

- Notice we will add another copy of $\mathbf{y}^{(t-1)}$ if we reject the new value $\mathbf{z}$.
- Secondly note it does not depend on the normalisation constant, i.e.

$$\begin{aligned}
\frac{f_{\mathbf{Y}}\left(\mathbf{y}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y}^{(t-1)} \mid \mathbf{y}\right)}{f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y} \mid \mathbf{y}^{(t-1)}\right)} &= \frac{\frac{1}{A} q_{\mathbf{Y}}\left(\mathbf{y}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y}^{(t-1)} \mid \mathbf{y}\right)}{\frac{1}{A} q_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y} \mid \mathbf{y}^{(t-1)}\right)} \\
&= \frac{q_{\mathbf{Y}}\left(\mathbf{y}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y}^{(t-1)} \mid \mathbf{y}\right)}{q_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y} \mid \mathbf{y}^{(t-1)}\right)}
\end{aligned}$$

where $A = \int_{\mathcal{D}} q_{\mathbf{Y}} \, d\mathbf{y}$ and $A f_{\mathbf{Y}} = q_{\mathbf{Y}}$ is the normalisation constant.

Q: Intuitively, why do you think Metropolis-Hastings provides samples from $f_{\mathbf{Y}}$?

- If we can sample from $f_{\mathbf{Y}}$, then there is no need to use conditional proposals, we will simply sample from $f_{\mathbf{Y}}$, and Metropolis-Hastings will accept all draws

$$\alpha = \min\left\{1, \frac{f_{\mathbf{Y}}\left(\mathbf{z}\right) \cdot f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right)}{f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right) \cdot f_{\mathbf{Y}}\left(\mathbf{z}\right)}\right\} = 1$$

Q: What would we have if we sample uniformly and accept $\mathbf{z}$ if

$$v \le \min\left\{1, \frac{f_{\mathbf{Y}}\left(\mathbf{z}\right)}{f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right)}\right\} \qquad \text{where} \quad v \sim \mathrm{Uniform}\left(0, 1\right)$$

- We would end up with a Markov Chain that "travels" on the support of $f_{\mathbf{Y}}$ in such a way the amount of "time" spend in each "location" in the support

$$\mathbf{y} \in \mathcal{D}$$

is directly proportional to the "height" of the density function as $n \to \infty$.

- Now, of course, we do not expect a uniform conditional proposal to be any good, so let us relax it by considering any conditional proposal such that

$$g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y} \mid \mathbf{y}^*\right) = g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y}^* \mid \mathbf{y}\right)$$

to simplify the situation before Metropolis-Hastings algorithm in general.

- When the conditional proposal distribution satisfies

$$g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y} \mid \mathbf{y}^*\right) = g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y}^* \mid \mathbf{y}\right)$$

then the distribution $g_{\mathbf{Y}|\mathbf{Y}^*}$ is known as symmetric, and

$$v \leq \min\left\{1, \frac{f_{\mathbf{Y}}\left(\mathbf{z}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y}^{(t-1)} \mid \mathbf{z}\right)}{f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{z} \mid \mathbf{y}^{(t-1)}\right)}\right\} = \min\left\{1, \frac{f_{\mathbf{Y}}\left(\mathbf{z}\right)}{f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right)}\right\}$$

which reduces Metropolis-Hastings to the original Metropolis algorithm.

- Similarly, it can be show that it has the joint $f_{\mathbf{Y}}$ as the invariant distribution.

Q: Intuitively, why do we have the following acceptance probability

$$\alpha = \min\left\{1, \frac{f_{\mathbf{Y}}\left(\mathbf{z}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{y}^{(t-1)} \mid \mathbf{z}\right)}{f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{z} \mid \mathbf{y}^{(t-1)}\right)}\right\}$$

when we sample from a non-symmetric conditional proposal?

- The goal is still to have a Markov Chain that "travels" on the support of $f_{\mathbf{Y}}$ in such a way the amount of "time" spend at each "location" in the support

$$\mathbf{y} \in \mathcal{D}$$

  is directly proportional to the "height" of the density function as $n \to \infty$.
- Intuitively, using asymmetric conditional proposal allows the Markov Chain to explore more region in $\mathcal{D}$, doing so can help prevent the Markov Chain being trapped in some region of $\mathcal{D}$, thus is particularly useful when the joint has many local peaks.
- The additional term in α makes sure the invariant distribution is the joint.

- Notice the Metropolis-Hastings acceptance probability

$$\alpha = \Pr\left(\text{accept} \mid \mathbf{Z} = \mathbf{z}, \mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}\right) = \alpha\left(\mathbf{z}, \mathbf{y}^{(t-1)}\right)$$

is a function of $\mathbf{z}$ and $\mathbf{y}^{(t-1)}$ given a new value $\mathbf{Z} = \mathbf{z}$ and $\mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}$.

- The probability of accepting any new value given only $\mathbf{Y}^{(t-1)}$ is given by

$$\Pr\left(\text{accept} \mid \mathbf{Y}^{(t-1)}\right) = \int_{\mathcal{D}} \alpha\left(\mathbf{z}, \mathbf{y}^{(t-1)}\right) \cdot g_{\mathbf{Y}|\mathbf{Y}^*}\left(\mathbf{z} \mid \mathbf{y}^{(t-1)}\right) d\mathbf{z}$$

- The transition kernel of the Metropolis-Hastings algorithm is

$$\begin{aligned}\kappa\left(\mathbf{y}^{(t-1)}, \mathbf{y}^{(t)}\right) = \alpha \cdot g_{\mathbf{Y}|\mathbf{Y}^*}&\left(\mathbf{y}^{(t)} \mid \mathbf{y}^{(t-1)}\right) \\ &+ \left(1 - \Pr\left(\text{accept} \mid \mathbf{Y}^{(t-1)}\right)\right) \cdot \delta_{\mathbf{y}^{(t-1)}}\left(\mathbf{y}^{(t)}\right)\end{aligned}$$

where $\delta_{\mathbf{y}^{(t-1)}}\left(\mathbf{y}^{(t)}\right)$ denotes Dirac-delta function .

Theorem

The Metropolis-Hastings kernel satisfies the detailed balance equation

$$\kappa\left(\mathbf{y}^{(t-1)}, \mathbf{y}^{(t)}\right) f_{\mathbf{Y}}\left(\mathbf{y}^{(t-1)}\right) = \kappa\left(\mathbf{y}^{(t)}, \mathbf{y}^{(t-1)}\right) f_{\mathbf{Y}}\left(\mathbf{y}^{(t)}\right)$$

so $f_{\mathbf{Y}}(\mathbf{y})$ is the invariant distribution of the Markov Chain. If the joint being positive $f_{\mathbf{Y}}(\mathbf{y}), f_{\mathbf{Y}}(\mathbf{y}^*) > 0$ guarantees the conditional proposal is also positive

$$g_{\mathbf{Y}|\mathbf{Y}^*}(\mathbf{y} \mid \mathbf{y}^*) > 0$$

for all $\mathbf{y}, \mathbf{y}^* \in \mathcal{D}$ of the joint distribution, then the sequence

$$\{f_{\mathbf{Y}^{(1)}}, f_{\mathbf{Y}^{(2)}}, \ldots\}$$

corresponding to the Metropolis-Hastings converges to $f_{\mathbf{Y}}$ for every $\mathbf{y}_0 \in \mathcal{D}$, and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} h\left(\mathbf{Y}^{(t)}\right) \to \mathbb{E}\left[h\left(\mathbf{Y}\right)\right]$$

- Recall, Gibbs will fail in practice when the joint density function $f_{\mathbf{Y}}$ that we need to sample from does not satisfy the positivity condition.

Q: Can you think of an example that Metropolis-Hastings will fail?

$$Y \sim \text{Uniform}\left(\mathcal{I}_1 \cup \mathcal{I}_2\right)$$

where $\mathcal{I}_1 = [0, 1]$ and $\mathcal{I}_2 = [2, 3]$, and

$$\text{Uniform}\left(y^{(t-1)} - \delta, y^{(t-1)} + \delta\right)$$

- If $\delta \leq 1$, then the Markov Chain

$$\{Y^{(1)}, Y^{(2)}, \ldots\}$$

generated by Metropolis-Hastings will not converge, that is, the distributions

$$\{f_{Y^{(1)}}, f_{Y^{(2)}}, \ldots\}$$

will not converge to the joint density f_Y.

Q: When should we use Metropolis-Hastings/Gibbs?

- In its original form, Gibbs sampling is the simplest of the MCMC algorithms, and it is our first choice for large conditionally conjugate models,

$$f_{\lambda_1|\{X_1,\dots X_n,\lambda_2,K\}} \sim \mathrm{Gamma}\left(\alpha_1 + \sum_{i=1}^{k} x_i, \beta_1 + k\right)$$

$$f_{\lambda_2|\{X_1,\dots X_n,\lambda_1,K\}} \sim \mathrm{Gamma}\left(\alpha_2 + \sum_{i=k+1}^{n} x_i, \beta_2 + n - k\right)$$

$$f_{K|\{X_1,\dots X_n,\lambda_1,\lambda_2\}} \propto \lambda_1^{\sum_{i=1}^{k} x_i} \lambda_2^{\sum_{i=k+1}^{n} x_i} \exp\left((\lambda_2 - \lambda_1)\cdot k\right)$$

  where we can easily sample each conditional posterior distribution.

- Metropolis-Hastings is used for models that are not conditionally conjugate.
- It also used in complicated models where the joint is unlikely to be unimodal.
- For a big complicated models, various combinations of the two are used.

Q: Can you see Gibbs is a special kind of Metropolis-Hastings?