Lecture 4: Gibbs Sampling

Tianpei Xie

Oct. 8th., 2022

Contents

1 Basic Concept of Markov Chain Monte Carlo

- **Definition** ***A Markov Chain Monte Carlo (MCMC) method*** for the simulation of a distribution f is any method producing an ergodic Markov chain $(X_t)_t$ whose stationary distribution is f.

- The MCMC updates ***preserve the probability measure*** π when convergence is attained. That is, *when the Markov chain* ***converges***, the distribution of X_t is the same as the distribution of $X_{t+1}, X_{t+2}, \ldots$. Thus we have obtained a sequence of ***identically distributed (but dependent) samples***. When Markov chain converges (***mixing***), we can use samples the same way as we did in vanilla Monte Carlo to approximate the expectation. In particular, an MCMC estimator is

$$J_T = \widehat{\mathbb{E}}_{\boldsymbol{\pi}}\left[h(X)\right] = \frac{1}{T}\sum_{t=0}^{T} h(X_t). \tag{1}$$

  The ergodic theorem guarantees the (almost sure) convergence of the empirical average to $\mathbb{E}_{\boldsymbol{\pi}}\left[h(X)\right]$ where $\boldsymbol{\pi}$ is the stationary distribution. A sequence $(X_t)_t$ produced by a Markov chain Monte Carlo algorithm can thus be employed just as an i.i.d sample.

- The basic idea for Metropolis algorithm is to simulate π using the stationary distribution g from a Markov chain. Compare to analysis of Markov chain itself, which often starts from a known transition kernel, the Metropolis algorithm starts from a known stationary distribution g and is interested in how to prescribe an efficient transition kernel to reach the equilibrium.

- The ***Metropolis-Hastings Algorithm*** [Robert and Casella, 1999, Liu, 2001] is described as below:

  1. Given current configuration $\boldsymbol{X}_t$, draw $\boldsymbol{Y}$ from the proposal function $K(\boldsymbol{X}_t\,, \boldsymbol{Y})$.
  2. Compute the ***Hastings ratio*** (*acceptance function*):

$$r(\boldsymbol{X}_t\,, \boldsymbol{Y}) := \frac{\pi(\boldsymbol{Y})\, K(\boldsymbol{Y}, \boldsymbol{X}_t)}{\pi(\boldsymbol{X}_t)\, K(\boldsymbol{X}_t\,, \boldsymbol{Y})} \tag{2}$$

  3. (***Metropolis Rejection***) Generate a uniform random variable $U \in \mathcal{U}[0,1]$.
     Accept $\boldsymbol{X}_{t+1} = \boldsymbol{Y}$ if

$$U \leq \alpha(\boldsymbol{X}_t\,, \boldsymbol{Y}),$$

     where

$$\alpha(\boldsymbol{X}_t, \boldsymbol{Y}) = \min\left\{1,\ r(\boldsymbol{X}_t\,, \boldsymbol{Y})\right\}$$

  4. Otherwise, accept $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$.

- The *transition probability* of $(\boldsymbol{X}_t)_t$ from the Metropolis-Hastings algorithm is computed as

$$\begin{aligned} A(\boldsymbol{x}\,, \boldsymbol{y}) &= K(\boldsymbol{x}\,, \boldsymbol{y})\, \alpha(\boldsymbol{x}\,, \boldsymbol{y}) \\ &= K(\boldsymbol{x}\,, \boldsymbol{y})\, \min\left\{1,\ \frac{\pi(\boldsymbol{y})\, K(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})\, K(\boldsymbol{x}\,, \boldsymbol{y})}\right\} \end{aligned}$$

$$= \frac{\min\{\pi(\boldsymbol{x})\,K(\boldsymbol{x}\,,\boldsymbol{y})\,,\pi(\boldsymbol{y})\,K(\boldsymbol{y},\boldsymbol{x})\}}{\pi(\boldsymbol{x})} := \frac{\delta(\boldsymbol{x},\boldsymbol{y})}{\pi(\boldsymbol{x})}. \tag{3}$$

Note that due to the **rejection rule**, $A(\boldsymbol{x}\,,\boldsymbol{y}) \neq K(\boldsymbol{x}\,,\boldsymbol{y})$. We can see that since $\delta(\boldsymbol{x},\boldsymbol{y}) = \delta(\boldsymbol{y},\boldsymbol{x})$, the *detailed balance equation* holds

$$\pi(\boldsymbol{x})A(\boldsymbol{x},\boldsymbol{y}) = \pi(\boldsymbol{y})A(\boldsymbol{y},\boldsymbol{x}). \tag{4}$$

From the theorem on detailed balance equation, we see that π is guaranteed to be the stationary distribution of the Markov chain and the Markov chain $(\boldsymbol{X}_t)_t$ is *time-reversible*.

- In general, define a ***symmetric distribution*** g_σ so that $q(\boldsymbol{y}|\boldsymbol{x}) = g_\sigma(\|\boldsymbol{y}-\boldsymbol{x}\|_2) > 0$ for all $\|\boldsymbol{y}-\boldsymbol{x}\|_2 < \delta$. Here σ controls the "***range***" of the exploration. The random walk with transition kernel as g_σ will produce an ergodic Markov chain.

  The most common distributions g in this setup are the ***uniform distributions on spheres*** centered at the origin or standard distributions like the ***normal*** and the ***Student's t distributions***. *All these distributions usually need to be* ***scaled***.

- The ***Random-walk Metropolis-Hastings*** is described as below:

  1. Draw $\boldsymbol{\epsilon}_t$ from $g_\sigma(\boldsymbol{\epsilon})$ and set $\boldsymbol{Y} := \boldsymbol{X}_t + \boldsymbol{\epsilon}_t$;
  2. Compute the ratio:

  $$r(\boldsymbol{X}_t\,,\boldsymbol{Y}) := \frac{\pi(\boldsymbol{Y})}{\pi(\boldsymbol{X}_t)} \tag{5}$$

  3. Accept $\boldsymbol{X}_{t+1} = \boldsymbol{Y}$ with probability

  $$\alpha(\boldsymbol{X}_t,\boldsymbol{Y}) = \min\{1,\ r(\boldsymbol{X}_t\,,\boldsymbol{Y})\}$$

  4. Otherwise, accept $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$.

2 Gibbs Sampling

- Choosing a good proposal function $q(\boldsymbol{y}|\boldsymbol{x}) = K(\boldsymbol{x},\boldsymbol{y})$ is an art. When no prior information is available, people tend to choose it arbitrarily. In many applications, the *Random Walk Metropolis-Hastings* is commonly used as "unbiased proposals". If not subject to Metropolis rejection rule, this type of Markov chain would explore the entire space of $\mathcal{X}$, which is very inefficient.

- The basic update of ***Gibbs sampling*** involves two steps:

  1. Split the multi-dimensional sample $\boldsymbol{X} := (X_j, \boldsymbol{X}_{-j})$.
  2. Update X_j' by conditional distribution $\pi_j(x_j|\boldsymbol{x}_{-j})$; Then $\boldsymbol{X'} := (X_j', \boldsymbol{X}_{-j})$.

- As compared to Metropolis-Hastings algorithm, Gibbs sampling has a number of distinct features:

  – The ***acceptance rate*** of the Gibbs sampler is ***uniformly*** **equal to** 1. Therefore, every simulated value is accepted and the suggestions on the optimal acceptance rates for

Metropolis-Hastings do not apply in this setting. This also means that ***convergence assessment*** for this algorithm should be treated differently than for Metropolis-Hastings techniques.

- The use of the Gibbs sampler implies limitations on the choice of instrumental distributions and requires a **prior knowledge** of some analytical or ***probabilistic properties*** of π.
- The Gibbs sampler is, by construction, ***multidimensional***. Even though some components of the simulated vector may be artificial for the problem of interest, or unnecessary for the required inference, the construction is still at least two-dimensional.
- The Gibbs sampler does not apply to problems where *the number of parameters* ***varies***, because of the obvious lack of irreducibility of the resulting chain.

- Gibbs sampling is **preferred** in following situations:
  - For ***high dimensional*** joint distribution $\pi(\boldsymbol{x})$, ***the local factor*** $\pi(x_k|\boldsymbol{x}_{-k})$ can be represented in ***explicit function form*** and is ***easy to simulate***. For instance, in many *probabilistic graphical models* such as *Ising model*, *Gaussian graphical models*, *hierarchical models*, *non-parametric Bayesian models*, *Hidden Markov models* etc., the local factor can explicitly simulated. Similarly, in many complex system, people have ***more knowledge on the local dynamics*** as compared to the overall system.
  - The target domain $\mathcal{X}$ contains **certain structures** such as *clusters*, *low-dimensional subspaces/sub-manifolds* etc. A simple Metropolis-Hastings algorithm such as random walk is ***inefficient*** since it spends too much efforts on exploration of low density regions. In high dimensional spaces, this type of random walk is doomed to fail. Instead, Gibbs sampling use conditional distribution resulting from constraining the target distribution π on certain subspaces. As a result, no rejection is incurred at sampling.
- Finally, note that Gibbs sampling is heavily affected by the ***parameterization*** (or *decomposition*) of space of distributions (like *coordinate descent*).

2.1 Slice Sampling

- To understand the idea of Gibbs sampling, we revisit the ***fundamental theorem of simulation***. In order to simulate from distribution f, we generate uniform samples from the ***subgraph*** of f, $\mathfrak{L}(f) := \{(x, u) : 0 \leq u \leq f(x)\}$.
- We can choose to simulate this distribution using *random walk* as Markov chain with stationary distribution as $\mathcal{U}(\mathfrak{L}(f))$. A natural implementation of this random walk is to ***go one direction at a time*** along the x-axis and u-axis iteratively.
  1. Starting from a point $(x, u) \in \mathfrak{L}(f)$, move along u-axis, i.e. sampling according to the conditional distribution
  $$U|X = x \sim \mathcal{U}\{u : 0 \leq u \leq f(x)\}.$$
  This results in a point $(x, u') \in \mathfrak{L}(f)$.
  2. Then given $(x, u') \in \mathfrak{L}(f)$, move along x-axis, i.e. sampling according to the conditional

distribution

$$X|U = u' \sim \mathcal{U}\left\{x : u' \leq f(x)\right\}.$$

This results in a point $(x', u') \in \mathfrak{L}(f)$.

This set of proposals is the basis chosen for the original ***Slice Sampler***.

- This slice sampler ***do not need accept-rejection step*** in normal Metropolis-Hastings algorithm since it directly sample from the target distribution conditioned on each variable.
- Note that we can use the unnormalized density $f_1 = C\,f$ in the simulation without changing the distribution.
- The validity of slice sampling as an MCMC algorithm associated with f stems from the fact that both steps 1. and 2. successively *preserve the uniform distribution* on the subgraph of f:

$$\begin{aligned}
(x_t, u_{t+1}) &\sim f(x)\frac{\mathbb{1}\left\{u \in [0, f_1(x)]\right\}}{f_1(x)} && \propto \mathbb{1}\left\{u \in [0, f_1(x)]\right\} \\
(x_t, u_{t+1}, x_{t+1}) &\sim f(x_t)\frac{\mathbb{1}\left\{u_{t+1} \in [0, f_1(x_t)]\right\}}{f_1(x_t)}\frac{\mathbb{1}\left\{x_{t+1} \in A_{t+1}\right\}}{\mathrm{vol}(A_{t+1})} \\
\Rightarrow (u_{t+1}, x_{t+1}) &\sim \frac{1}{C}\mathbb{1}\left\{x_{t+1} \in A_{t+1}\right\}\int \frac{\mathbb{1}\left\{u_{t+1} \in [0, f_1(x)]\right\}}{\mathrm{vol}(A_{t+1})}dx && \propto \mathbb{1}\left\{x_{t+1} \in A_{t+1}\right\}
\end{aligned}$$

where $A_{t+1} := \{x : f(x) \geq u_{t+1}\}$.

- This can be extended to f with multiple components. See that

$$f(x) \propto \prod_i^k f_i(x)$$

where $f_i(x)$ is some positive functions, e.g. likelihoods. This decomposition can then be associated with k ***auxiliary variables*** ω_i, so that

$$f_i(x) = \int \mathbb{1}\left\{0 \leq \omega_i \leq f(x)\right\} d\omega_i.$$

Therefore, f is the marginal distribution of $(f, \omega_1, \ldots, \omega_k)$. This is called *de-marginalization* of f.

- The ***General Slice Sampler*** will be
  1. For $i = 1, \ldots, k$, sample $\omega_i^{(t+1)} \sim \mathcal{U}[0, f_i(x^{(t)})]$;
  2. sample $x^{(t+1)}$ from $\mathcal{U}(A_{t+1})$ where $A_{t+1} := \left\{x : f_i(x) \geq \omega_i^{(t+1)},\ i = 1, \ldots, k\right\}$
- **Lemma 2.1** *[Robert and Casella, 1999]*
  If h is bounded and supp(f_i) is bounded, the slice sampler above is ***uniformly ergodic****.*

2.2 Gibbs Sampling

- The general Gibbs Sampling is an extension of Slice Sampling, where each time sample one variable from the region $A_k^{(t+1)} := \{x_k : f(x_k, \boldsymbol{x}_{-k}) \geq u/f_{-k}(\boldsymbol{x}_{-k})\}$ where $f_{-k}(\boldsymbol{x}_{-k})$ is the

marginal distribution on the rest of variables $\boldsymbol{x}_{-k}$. In the limit, this is equivalent to sampling x_k from the conditional distribution $f_{k|-k}(x_k|\boldsymbol{x}_{-k})$.

- The most prominent feature of **Gibbs sampling** is that the underlying Markov chain is constructed by composing a sequence of ***conditional distributions*** *(local factors)* along *a set of directions (often along the coordinate axis).*
- Suppose $\boldsymbol{X} = [X_1, \ldots, X_d]$ and for given k, the rest of variables $\boldsymbol{X}_{-k} = [X_j, \forall\, j \neq k]$. Let $\pi(\cdot|\boldsymbol{x}_{-k})$ be the conditional distribution of target π on rest of variables $\boldsymbol{x}_{-k}$.
- The ***Random Scan Gibbs Sampling*** is described as below at $t+1$ iteration:

  1. Given $\boldsymbol{X}^{(t)} = [X_1^{(t)}, \ldots, X_d^{(t)}]$ from iteration t, ***randomly*** select a coordinate i from $\{1, \ldots, d\}$ with probability $(\alpha_1, \ldots, \alpha_d)$ (usually uniform $(1/d, \ldots, 1/d)$).
  2. Draw $X_i^{(t+1)}$ from conditional distribution $\pi(x_i|\boldsymbol{X}_{-i}^{(t)})$ and *leave the rest of variable* ***unchanged***. That is

$$\boldsymbol{X}_{-i}^{(t+1)} = \boldsymbol{X}_{-i}^{(t)}$$

- The ***Systematic Scan Gibbs Sampling*** is described as below at $t+1$ iteration:

  1. Given $\boldsymbol{X}^{(t)} = [X_1^{(t)}, \ldots, X_d^{(t)}]$ from iteration t, then for $i = 1, \ldots, d$:

     (a) draw $X_i^{(t+1)}$ from conditional distribution

$$\pi\left(x_i \middle| \boldsymbol{X}_{1:(i-1)}^{(t+1)}, \boldsymbol{X}_{(i+1):d}^{(t)}\right)$$

- It is easy to check that *every* single conditional update for both *Random Scan Gibbs Sampling* and *Systematic Scan Gibbs Sampling* **preserve the joint probability** π. Suppose $\boldsymbol{X}^{(t)} \sim \pi$, then $\boldsymbol{X}_{-i}^{(t)} \sim \pi(\boldsymbol{X}_{-i}^{(t)})$ its marginal distribution. Then

$$\pi(X_i^{(t+1)}|\boldsymbol{X}_{-i}^{(t)}) \times \pi(\boldsymbol{X}_{-i}^{(t)}) = \pi\left(X_i^{(t+1)}, \boldsymbol{X}_{-i}^{(t)}\right),$$

  which means that after one update the new configuration follows the same distribution π.

2.3 The Hammersley-Clifford Theorem

- A most surprising feature of the Gibbs sampler is that the *conditional distributions* contain ***sufficient information*** to produce a sample from the *joint distribution*. A similar case is the *coordinate ascent algorithm* in optimization problem. However, it is well known that this optimization method does not necessarily lead to the *global maximum*, but may end up in a saddlepoint. It is, therefore, somewhat remarkable that the full conditional distributions perfectly summarize the joint density, although the set of **marginal distributions obviously fails to do so**. The following result then shows that the joint density can be directly and constructively derived from the conditional densities.
- **Definition** Let $(X_1, X_2, \ldots, X_d) \sim \pi(x_1, x_2, \ldots, x_d)$, where π_i denotes the marginal distribution of X_i. If $\pi_i(x_i) > 0$ for every $i = 1, \ldots, d$, implies that $\pi(x_1, x_2, \ldots, x_d) > 0$, then π satisfies the ***positivity condition***.

- **Theorem 2.2** ***(Hammersley-Clifford)***
  Under the ***positivity condition****, the joint distribution* π *satisfies*

$$\pi(x_1, x_2, \ldots, x_d) = \prod_{j=1}^{d} \frac{\pi_{\ell_j}(x_{\ell_j} \,|\, x_{\ell_1}, x_{\ell_2}, \ldots, x_{\ell_{j-1}}, x'_{\ell_{j+1}}, \ldots, x'_{\ell_d})}{\pi_{\ell_j}(x'_{\ell_j} \,|\, x_{\ell_1}, x_{\ell_2}, \ldots, x_{\ell_{j-1}}, x'_{\ell_{j+1}}, \ldots, x'_{\ell_d})} \tag{6}$$

  for every permutation ℓ *on* $\{1, 2, \ldots, d\}$ *and every* $\boldsymbol{x}' \in \mathcal{X}$*.*

- In two-stage Gibbs sampling, the joint distribution π associated with conditionals $\pi_{X|Y}(x|y)$ and $\pi_{Y|X}(y|x)$ can be reconstructed as

$$\pi(x, y) = \frac{\pi_{Y|X}(y|x)}{\int \pi_{Y|X}(y|x)/\pi_{X|Y}(x|y)dy}$$

2.4 Partial Resampling

- From Hammersley-Clifford theorem, we see that ***any transition rule*** $A(x_j \to x'_j)$ that leaves the ***conditional distribution*** $\pi_j(x_j \,|\, \boldsymbol{x}_{-j})$ ***invariant*** will leave the *joint distribution* π *invariant*.

$$\begin{aligned}
&\int \pi(x_j, \boldsymbol{x}_{-j})\, A(x_j \to x'_j|\boldsymbol{x}_{-j})dx_j \\
&= \pi(\boldsymbol{x}_{-j}) \int \pi_j(x_j|\boldsymbol{x}_{-j})\, A(x_j \to x'_j|\boldsymbol{x}_{-j})dx_j \\
&= \pi(\boldsymbol{x}_{-j})\, \pi_j(x'_j|\boldsymbol{x}_{-j}) = \pi(x'_j, \boldsymbol{x}_{-j})
\end{aligned}$$

- This inspires the idea of ***partial resampling*** [Liu, 2001], which use *reparameterization* to improve the performance of sampling.

- Suppose we have a ***partition*** of sample space $\mathcal{X} = \bigcup_{\alpha \in A} \mathcal{X}_\alpha$, $\mathcal{X}_a \cap \mathcal{X}_b = \emptyset$, $a, b \in A$. Then the joint distribution π can be reconstructed as [Liu, 2001]

$$\pi(\boldsymbol{x}) = \int \nu_\alpha(\boldsymbol{x})\, \rho(\alpha)$$

  where $\nu_\alpha(\boldsymbol{x})$ is the conditional distribution of $\boldsymbol{X}$ on subspace $\mathcal{X}_\alpha$. Each $\mathcal{X}_\alpha$ is called a ***fiber***.

  It is seen that any transition rule $A(\boldsymbol{x} \to \boldsymbol{x}')$ that leaves $\nu_\alpha(\boldsymbol{x})$ on ***fiber*** $\mathcal{X}_\alpha$ ***invariant*** will *leave the joint distribution* π *invariant*.

- Choosing different fibers will allow Gibbs sampling updates in directions other than coordinate axis.

- The benefits of partial resampling include:
  - Similar to Metropolis move, a conditional move will reduce the *high dimensional* simulation problem into a ***low dimensional*** one;
  - It allows the sampler to follow the ***local dynamics*** of the target distribution;
  - It enables the ***non-axis moves*** that cannot be achieved by Gibbs sampler.

2.5 Probabilistic Structures of two-stage Gibbs Sampling Markov Chain

- For two-stage Gibbs sampling, not only $(X_t, Y_t)_t$ is a Markov chain, but also ***each subsequence*** $(X_t)_t$ and $(Y_t)_t$ ***is a Markov chain***. The transition probability for (X_t) is

$$K_X(x_t, x_{t+1}) = \int \pi_{X|Y}(x_{t+1} \,|\, y_{t+1})\, \pi_{Y|X}(y_{t+1} \,|\, x_t)\, dy_{t+1} \tag{7}$$

  which indeed depends on the past only through the last value of (X_t)·

- We have the following lemma

  Lemma 2.3 *[Robert and Casella, 1999]*
  Each of the sequences $(X_t)_t$ and $(Y_t)_t$ produced by the two-stage Gibbs sampling is a ***Markov chain*** *with corresponding stationary distributions*

$$\pi_X(x) = \int \pi(x, y)dy, \quad \pi_Y(y) = \int \pi(x, y)dx \tag{8}$$

  as the ***marginal distribution*** *of π.*

  Proof: The development of (7) shows that each chain is, individually, a Markov chain. The stationary distribution is computed via

$$\begin{aligned}
\pi_X(x_{t+1}) &= \int \pi_{X|Y}(x_{t+1}|y_{t+1})\pi_Y(y_{t+1})dy_{t+1} \\
&= \int \pi_{X|Y}(x_{t+1}|y_{t+1}) \int \pi_{Y|X}(y_{t+1}|x_t)\pi_X(x_t)dx_t dy_{t+1} \\
&= \int \left[\int \pi_{X|Y}(x_{t+1}|y_{t+1})\pi_{Y|X}(y_{t+1}|x_t)dy_{t+1} \right] \pi_X(x_t)dx_t \\
&= \int K_X(x_t, x_{t+1})\pi_X(x_t)dx_t.
\end{aligned}$$

  Thus $\pi_X(x)$ is the stationary distribution associated with (X_t) with kernel $K_X(x_t, x_{t+1})$. Under the positivity constraint, if π is positive, $\pi_{X|Y}(x|y)$ is positive on the (projected) support of π and every Borel set of $\mathcal{X}$ can be visited in a single iteration of Gibbs update, establishing the strong irreducibility. This development also applies to (Y_t). ∎

  This elementary reasoning shows, in addition, that if only the chain (X_t) is of interest and if the condition $\pi_{X|Y}(x|y) > 0$ holds for every pair (X', Y), ***irreducibility*** is satisfied. As shown further below, the "*dual*" chain (Y_t) can be used to establish some probabilistic properties of (X_t).

- For two-stage Gibbs sampling, if the sub-chain (X_t) is of interest, the sub-chain (Y_t) is called *dual* chain or ***instrumental chain***.

  Definition Two Markov chains (X_t) and (Y_t) are said to be ***conjugate*** to each other with the ***interleaving property*** (or ***interleaved***) if

  1. $X_{t+1} \perp\!\!\!\perp X_t \,|\, Y_t$
  2. $Y_t \perp\!\!\!\perp Y_{t-1} \,|\, X_t$
  3. (X_t, Y_{t-1}) and (X_t, Y_t) are ***identically distributed*** under ***stationarity***.

This implies the *interleaving link structure* $\ldots \to Y_{t-1} \to X_t \to Y_t \to X_{t+1} \to \ldots$

- **Lemma 2.4** *[Robert and Casella, 1999]*
*Each of the chains (X_t) and (Y_t) generated by a two-stage Gibbs sampling algorithm is **reversible**, and the chain (X_t, Y_t) satisfies the **interleaving property**.*

- The well-known concept of *Rao-Blackwellization* is based on the fact that a ***conditioning*** argument, using variables other than those directly of interest, can result in an improved procedure. We will see below that this phenomenon is more fundamental than a mere improvement of the variance of some estimators, as it provides a general technique to establish convergence properties for the chain *of interest* (X_t) based on the instrumental chain (Y_t), even when the latter is ***unrelated*** with the inferential problem. this use of the dual chain the ***Duality Principle*** when they used it in the setup of mixtures of distributions.

- **Theorem 2.5** *[Robert and Casella, 1999]*
*Consider a **Markov chain** (X_t) and a **sequence** (Y_t) of random variables generated from the conditional distributions*

$$X_t \,|\, y_t \sim \pi(x|\, y_t), \quad Y_{t+1} \,|\, x_t, y_t \sim f(y \,|\, x_t, y_t)$$

*If the chain (Y_t) is **ergodic** (geometrically or uniformly ergodic) and if $\pi^t_{y_0}$ denotes the distribution of (X_t) associated with the initial value y_0, the norm $\left\|\pi^t_{y_0} - \pi\right\|_{TV}$ goes to 0 when t goes to infinity (goes to 0 at a geometric or uniformly bounded rate).*

Note that this setting contains as a particular case *hidden Markov models*, where $Y_{t+1} \,|\, x_t, y_t \sim f(y \,|\, y_t)$ is not fully observed.

- **Theorem 2.6** *[Robert and Casella, 1999]*
*If (Y_t) is a **finite** state-space Markov chain, with state-space $\mathcal{Y}$, such that*

$$P(Y_{t+1} = k \,|\, y_t, x) > 0, \quad \forall k \in \mathcal{Y}, \forall x \in \mathcal{X},$$

*the sequence (X_t) derived from (Y_t) by the transition $\pi(x|\, y_t)$ is **uniformly ergodic**.*

Notice that this convergence result does not impose any constraint on the *transition* $\pi(x|\, y)$, which, for instance, is ***not necessarily everywhere positive***.

- **Corollary 2.7** *For two **interleaved** Markov chains, (X_t) and (Y_t), if (X_t) is **ergodic** (geometrically ergodic), then (Y_t) is **ergodic** (geometrically ergodic).*

- Finally, we turns to the rate of convergence:

Proposition 2.8 *[Robert and Casella, 1999]*
*If (Y_t) is geometrically convergent with **compact** state-space and with convergence rate ρ, there exists C_h such that*

$$\|\mathbb{E}\left[h(X_t)|y_0\right] - \mathbb{E}_\pi\left[h(X)\right]\| < C_h \rho^t$$

*for every function $h \in \mathcal{L}_1(\pi(\cdot|x))$ **uniformly** in $y_0 \in \mathcal{Y}$.*

2.6 Theoretical Justifications

- For multi-stage Gibbs sampling, we can extend the result from two-stage Gibbs sampling above. We first extend the distribution π by introducing new *auxiliary variables* $\boldsymbol{z}$.

Definition [Robert and Casella, 1999]
Given a probability density π, a density g that satisfies

$$\int_{\mathcal{Z}} g(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = \pi(\boldsymbol{x}) \tag{9}$$

is called a ***completion*** of π.

The density g is chosen so that the full conditionals of g are **easy to simulate from** and the Gibbs algorithm is implemented on g instead of π. For $p > 1$, write $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{z})$ and denote the conditional densities of $g(\boldsymbol{y}) = g(y_1, \ldots, y_p)$ by

$$Y_k \,|\, \boldsymbol{y}_{-k} \sim g_k(y_k \,|\, \boldsymbol{y}_{-k})$$

In principle, the Gibbs sampler does not require that the completion of π into g and of $\boldsymbol{x}$ in $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{z})$ should be related to the problem of interest. Indeed, there are settings where the vector $\boldsymbol{z}$ has no meaning from a statistical point of view and is only a useful device.

- Using the auxiliary variables, we can prove the following theorem.

  Theorem 2.9 *[Robert and Casella, 1999]*
  *For the systematic scan Gibbs sampler, if (Y_t) is **ergodic**, then the distribution g is a **stationary distribution** for the chain (Y_t) and π is the **limiting distribution** of the subchain (X_t).*

  Note that the kernel of (Y_t) is

$$K(\boldsymbol{y}, \boldsymbol{y}') = \prod_{j=1}^{p} g_j(y'_j | \boldsymbol{y}'_{1:(j-1)}, \boldsymbol{y}_{(j+1):p}).$$

- **Theorem 2.10** *For the systematic scan Gibbs sampler, if the density g satisfies the **positivity condition**, it is **irreducible**.*

- **Lemma 2.11** *[Robert and Casella, 1999]*
  *If the transition kernel associated with Gibbs sampling is **absolutely continuous** with respect to the dominating measure, the resulting chain is **Harris recurrent**.*

  Note that the condition on the Gibbs transition kernel yields an irreducible chain, and Harris recurrence was shown to follow from irreducibility.

- Finally, we have convergence guarantee:

  Theorem 2.12 *[Robert and Casella, 1999]*
  *If the transition kernel of the Gibbs chain (Y_t) is **absolutely continuous** with respect to the measure μ,*

  1. *If $h_1, h_2 \in \mathcal{L}_1(g)$ with $\mathbb{E}_g[h_2(Y)] := \int h_2(y) dg(y) \neq 0$, then*

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} h_1(Y_t)}{\sum_{t=1}^{T} h_2(Y_t)} = \frac{\mathbb{E}_g[h_1(Y)]}{\mathbb{E}_g[h_2(Y)]} = \frac{\int h_1(y) dg(y)}{\int h_2(y) dg(y)}, \quad a.e.\ g \tag{10}$$

  2. *If, in addition, (Y_t) is **aperiodic**, then, for **every** initial distribution μ,*

$$\lim_{t\to\infty} \left\| \int K^t(\boldsymbol{y}, \cdot) \mu(d\boldsymbol{y}) - g \right\|_{TV} = 0 \tag{11}$$

2.7 Gibbs Sampling as Metropolis-Hastings

- **Theorem 2.13** *The Gibbs sampling method is equivalent to the composition of d Metropolis-Hastings algorithms, with acceptance probabilities* ***uniformly*** *equal to* 1.

  Note that we can write the proposal function of j-th Metropolis-Hastings algorithms as

$$K_j(\boldsymbol{x}, \boldsymbol{x}') = \delta_{\boldsymbol{x}_{-j}}(\boldsymbol{x}'_{-j})\, \pi_j(x'_j|\boldsymbol{x}_{-j}), \quad j = 1, \ldots, d \tag{12}$$

  where $\delta_{\boldsymbol{x}_{-j}}(\boldsymbol{x}'_{-j}) = 1$ if $\boldsymbol{x}_{-j} = \boldsymbol{x}'_{-j}$, otherwise equal to 0. Therefore, the Hastings ratio is

$$\begin{aligned}
\frac{\pi(\boldsymbol{x}')\, K_j(\boldsymbol{x}', \boldsymbol{x})}{\pi(\boldsymbol{x})\, K_j(\boldsymbol{x}, \boldsymbol{x}')} &= \frac{\pi(\boldsymbol{x}')\delta_{\boldsymbol{x}'_{-j}}(\boldsymbol{x}_{-j})\, \pi_j(x_j|\boldsymbol{x}'_{-j})}{\pi(\boldsymbol{x})\delta_{\boldsymbol{x}_{-j}}(\boldsymbol{x}'_{-j})\, \pi_j(x'_j|\boldsymbol{x}_{-j})} \\
&= \frac{[\pi_j(x'_j|\boldsymbol{x}'_{-j})\pi(\boldsymbol{x}'_{-j})\delta_{\boldsymbol{x}'_{-j}}(\boldsymbol{x}_{-j})]\, \pi_j(x_j|\boldsymbol{x}'_{-j})}{[\pi_j(x_j|\boldsymbol{x}_{-j})\pi(\boldsymbol{x}_{-j})\delta_{\boldsymbol{x}_{-j}}(\boldsymbol{x}'_{-j})]\, \pi_j(x'_j|\boldsymbol{x}_{-j})} \\
&= \frac{\pi_j(x'_j|\boldsymbol{x}_{-j})\, \pi_j(x_j|\boldsymbol{x}'_{-j})}{\pi_j(x_j|\boldsymbol{x}'_{-j})\, \pi_j(x'_j|\boldsymbol{x}_{-j})} = 1
\end{aligned}$$

2.8 Metroplized Gibbs sampling

- When the state-space of interest is discrete, Liu suggested an completion strategy described above to improve the Gibbs sampling. In particular, we augment state $\boldsymbol{x}$ to be $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{z})$ and the distribution g replace π. We have the following lemma

  Lemma 2.14 *[Robert and Casella, 1999]*
  Consider two ***reversible*** *Markov chains on a* ***countable*** *state-space, with transition matrices* $\boldsymbol{K}_1$ *and* $\boldsymbol{K}_2$ *such that* $\boldsymbol{K}_1 \ll \boldsymbol{K}_2$*, which means that the* ***non-diagonal entries*** *of* $\boldsymbol{K}_2$ *is larger than those of* $\boldsymbol{K}_1$*. The chain associated with* $\boldsymbol{K}_2$ *dominates the chain associated with* $\boldsymbol{K}_1$ *in terms of variances.*

  Given a conditional distribution $g_i(y_i \,|\, \boldsymbol{y}^{(t)}_{-i})$ on a discrete space, the modification proposed by Liu (1995) is to use an additional Metropolis-Hastings step.

- The ***Metropolized Gibbs Sampling*** [Robert and Casella, 1999, Liu, 2001] is described as below:

  Given $\boldsymbol{Y}^{(t)}$, for $i = 1, \ldots, p$:

  1. (***Gibbs Update***) generate $Z_i \neq Y_i^{(t)}$ with probability

$$p_i := \frac{g_i(z_i \,|\, \boldsymbol{y}^{(t)}_{-i})}{1 - g_i(z_i \,|\, \boldsymbol{y}^{(t)}_{-i})}$$

  2. (***Metropolis Rejection***) Accept $Y_i^{(t+1)} = Z_i$ with probability

$$\alpha(\boldsymbol{X}_t, \boldsymbol{Y}) = \min\left\{1,\ \frac{1 - g_i(y_i^{(t)} \,|\, \boldsymbol{y}^{(t)}_{-i})}{1 - g_i(z_i \,|\, \boldsymbol{y}^{(t)}_{-i})}\right\}$$

The probability of moving from $Y_i^{(t)}$ to a different value is then necessarily higher than in the original Gibbs sampling algorithm. Following the Lemma above, we have:

- **Theorem 2.15** *[Robert and Casella, 1999]*
  The Metroplized Gibbs sampling is more efficient than the random scan Gibbs sampling in terms of variance.

References

Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.