Unit 5: Gibbs Sampling

5.1 Various Gibbs Sampling Schemes

Let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, and consider a target probability distribution $\pi(x)$ with $x = (x_1, \ldots, x_d)$. For clarity, in this unit we will use the following notation. Given a vector x, we write $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$. Let $\pi_{i\,|\,-i}(\cdot \,|\, x_{-i})$ denote the conditional distribution of X_i given $X_{-i} = x_{-i}$, where X follows the distribution π, and similarly, denote the marginal distribution of X_i by π_i and the marginal distribution of X_{-i} by π_{-i}.

Define the transition kernel P_1 by

$$P_1(x, B_1 \times B_{-1}) = \mathbb{1}_{B_{-1}}(x_{-1}) \int_{B_1} \pi_{1\,|\,-1}(\mathrm{d}y_1 \,|\, x_{-1})$$

for any $B_1 \in \mathcal{B}(\mathcal{X}_1)$ and $B_{-1} \in \mathcal{B}(\mathcal{X}_2 \times \mathcal{X}_d)$. Define transition kernels $P_2, \ldots, P_d$ similarly. Note that equivalently, we can write

$$P_i(x, \mathrm{d}y) = \pi_{i\,|\,-i}(\mathrm{d}y_i \,|\, x_{-i}) \delta_{x_{-i}}(y_{-i}).$$

Hence, we will denote the density of $P_i(x, \cdot)$ by

$$p_i(x, y) = \pi_{i\,|\,-i}(y_i \,|\, x_{-i}) \mathbb{1}_{\{x_{-i} = y_{-i}\}},$$

(If $\pi_{i\,|\,-i}(y_i \,|\, x_{-i})$ is with respect to the measure μ on $\mathcal{X}_i$, then $p_i(x, \cdot)$ is the density with respect to the measure $\mu \times \delta_{x_{-i}}$ on $\mathcal{X}_i \times \mathcal{X}_{-i}$.)

The kernel P_i can only modify the i-th coordinate. In Unit 3, we have shown that actually P_i is a Metropolis–Hastings algorithm which always accepts the proposal, which implies that P_i is reversible w.r.t π. We can also directly check the detailed balance condition:

$$\pi(x) p_i(x, y) = \pi_{-i}(x_{-i}) \pi_{i\,|\,-i}(x_i \,|\, x_{-i}) \pi_{i\,|\,-i}(y_i \,|\, x_{-i}) \mathbb{1}_{\{x_{-i} = y_{-i}\}}.$$

This is symmetric in x, y, since $p_i(x, y) \neq 0$ only when $x_{-i} = y_{-i}$. So π is a stationary distribution of P_i (though not unique!), and we can write $\pi P_i = \pi$. In Gibbs sampling, we make use of all the d kernels, $P_1, P_2, \ldots, P_d$, so that it is possible to move between any two states. As we have discussed in Unit 3, there are many ways to combine these kernels, which lead to different "updating schemes" of Gibbs sampling.

Example 5.1 (deterministic sweep)**.** Consider the Gibbs sampler with transition kernel $P = P_1 P_2 \cdots P_d$. In each step of this Gibbs sampler, we update $x_1, \ldots, x_p$ sequentially. For example, suppose $d = 2$ and the current state is $x = (x_1, x_2)$. We first generate y_1 from the conditional distribution given x_2 and then generate y_2 from the conditional distribution given y_1. Since $\pi P_i = \pi$ for each i, we have $\pi P = \pi$. Note that this Gibbs sampler is usually not a reversible Markov chain (check that the detailed balance condition does not hold).

Example 5.2 (reversible sweep)**.** The transition kernel $P = P_1 P_2 \cdots P_{d-1} P_d P_{d-1} \cdots P_2 P_1$ also clearly has π as the stationary distribution. Further, P is reversible with respect to π.

Example 5.3 (random scan)**.** This was discussed in Unit 3. Random scan Gibbs sampler is the Markov chain with the kernel $P = d^{-1}(P_1 + \cdots + P_d)$. This is also reversible.

Example 5.4 (random permutation)**.** A more complicated reversible updating scheme can be constructed as follows. In each iteration, we generate an order τ uniformly from the symmetric group on $\{1, 2, \ldots, d\}$ and update the d coordinates according to τ. The resulting kernel can be writting as

$$P = \frac{1}{d!} \sum_{\tau} P_{\tau(1)} \cdots P_{\tau(d)}.$$

Exercise 5.1. Give an example where

$$\pi(x_1, x_2)\pi_{1\,|\,2}(y_1 \,|\, x_2)\pi_{2\,|\,1}(y_2 \,|\, y_1) \neq \pi(y_1, y_2)\pi_{1\,|\,2}(x_1 \,|\, y_2)\pi_{2\,|\,1}(x_2 \,|\, x_1).$$

This shows that the deterministic sweep Gibbs sampler is not reversible.

5.2 Gibbs Sampling for Multivariate Normal Targets

5.2.1 Convergence Rates

Let $\mathcal{X} = \mathbb{R}^n$ be our state space, and let π be the d-dimensional multivariate normal distribution $N(\mu, \Sigma)$. Assume Σ is invertible. Write $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ where $\mathcal{X}_i = \mathbb{R}^{n_i}$, and $n_1 + \cdots + n_d = n$. Partition the covariance matrix Σ accordingly:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1d} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{d1} & \Sigma_{d2} & \cdots & \Sigma_{dd}, \end{bmatrix}$$

where Σ_{ij} has dimension $n_i \times n_j$. Similarly, let $Q = \Sigma^{-1}$, and let Q_{ij} denote the (i, j)-th block of Q. Note that by the block matrix inversion formula,

$$Q_{11}^{-1} = \Sigma_{11} - \begin{bmatrix} \Sigma_{12} & \cdots & \Sigma_{1d} \end{bmatrix} \begin{bmatrix} \Sigma_{22} & \cdots & \Sigma_{2d} \\ \vdots & \ddots & \vdots \\ \Sigma_{d2} & \cdots & \Sigma_{dd} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{21} \\ \vdots \\ \Sigma_{d1} \end{bmatrix},$$

$$Q_{11}^{-1} \begin{bmatrix} Q_{12} & \cdots & Q_{1d} \end{bmatrix} = - \begin{bmatrix} \Sigma_{12} & \cdots & \Sigma_{1d} \end{bmatrix} \begin{bmatrix} \Sigma_{22} & \cdots & \Sigma_{2d} \\ \vdots & \ddots & \vdots \\ \Sigma_{d2} & \cdots & \Sigma_{dd} \end{bmatrix}^{-1}.$$

Let $X = (X_1, \ldots, X_d)$ denote the current sample, and consider the deterministic sweep Gibbs sampler which updates $X_1, \ldots, X_d$ sequentially. Denote the resulting new vector by

$Y = (Y_1, \ldots, Y_d)$. Since Y_1 is drawn from the conditional distribution given $X_2, \ldots, X_d$, we can write

$$Y_1 = \mu_1 - Q_{11}^{-1} \begin{bmatrix} Q_{12} & \cdots & Q_{1d} \end{bmatrix} \begin{bmatrix} X_2 - \mu_2 \\ \vdots \\ X_d - \mu_d \end{bmatrix} + Q_{11}^{-1/2} Z_1,$$

where Z_1 follows the standard normal distribution $N(0, I_{n_1})$ independently of other variables. Similarly, Y_2 is drawn from the conditional distribution given Y_1 and $X_3, \ldots, X_d$, which yields

$$Y_2 = \mu_2 - Q_{22}^{-1} \begin{bmatrix} Q_{21} & Q_{23} & \cdots & Q_{1d} \end{bmatrix} \begin{bmatrix} Y_1 - \mu_1 \\ X_3 - \mu_3 \\ \vdots \\ X_d - \mu_d \end{bmatrix} + Q_{22}^{-1/2} Z_2.$$

Repeating this calculation and doing some algebra, we obtain the following lemma [3].

Lemma 5.1. *Define the block diagonal matrix $D = \operatorname{diag}(Q_{11}, Q_{22}, \ldots, Q_{dd})$. Let $A = I - D^{-1}Q$, and denote the lower triangular and upper triangular parts of A by L and U, respectively (note that the diagonal blocks of A are zeros). Define $B = (I - L)^{-1}U$. Then we can write the deterministic sweep Gibbs update as*

$$Y \sim N(BX + (I - B)\mu,\, \Sigma - B\Sigma B^\top).$$

It was shown in [3] that the convergence rate of the deterministic sweep Gibbs sampler is determined by the spectral radius (maximum eigenvalue in absolute value) of the matrix B in Lemma 5.1; denote the spectral radius by $\rho(B)$. The precise statement is given in Theorem 5.1.

Theorem 5.1. *Let π denote the target distribution $N(\mu, \Sigma)$ and P denote the transition kernel of the deterministic sweep Gibbs sampler described above. We have*

$$\rho(B) = \inf \left\{ r \colon \lim_{t \to \infty} \frac{\int \{(P^t f)(x) - \pi(f)\}^2 \pi(x) \mathrm{d}x}{r^t} = 0,\ \forall f \in L^2(\pi) \right\}.$$

Proof. See [3]. □

Note that the smaller $\rho(B)$, the faster convergence the deterministic sweep Gibbs sampler achieves. Other updating schemes were also studied in [3]. For example, they showed that the convergence rate of the random scan Gibbs sampler is given by

$$\left(\frac{d - 1 + \lambda_1(A)}{d} \right)^d,$$

where $\lambda_1(A)$ denotes the largest eigenvalue of the matrix A in Lemma 5.1.

Exercise 5.2. Prove Lemma 5.1.

Exercise 5.3. Prove that all eigenvalues of A are real and $0 \leq \lambda_1(A) < 1$ (recall that we assume Σ is strictly positive definite so that Q exists).

5.2.2 Multilevel Random Effects Models

Suppose we have IJK grouped observations, denoted by $\{w_{ijk} \colon 1 \le i \le I,\, j \le 1 \le J,\, 1 \le k \le K\}$. Consider the following multilevel random effects model

$$\begin{aligned} w_{ijk} &= \mu + a_i + b_{ij} + \epsilon_{ijk}, \quad 1 \le i \le I,\, j \le 1 \le J,\, 1 \le k \le K, \\ \epsilon_{ijk} &\overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2), \quad 1 \le i \le I,\, j \le 1 \le J,\, 1 \le k \le K, \\ b_{ij} &\overset{\text{i.i.d.}}{\sim} N(0, \sigma_b^2), \quad 1 \le i \le I,\, 1 \le j \le J, \\ a_i &\overset{\text{i.i.d.}}{\sim} N(0, \sigma_a^2), \quad 1 \le i \le I, \\ p(\mu) &\propto 1, \end{aligned}$$

where $p(\mu) \propto 1$ means that we assign an improper prior distribution on μ (uniform over $\mathbb{R}$). Assume the hyperparameters $\sigma_a^2, \sigma_b^2, \sigma_e^2$ are given. It is easy to find the un-normalized posterior:

$$-2 \log p(\mu, a, b \,|\, w) = C + \sum_i \frac{a_i^2}{\sigma_a^2} + \sum_{i,j} \frac{b_{ij}^2}{\sigma_b^2} + \sum_{i,j,k} \frac{(w_{ijk} - \mu - a_i - b_{ij})^2}{\sigma_e^2},$$

where C is the unknown normalizing constant. So the joint posterior of μ, a, b is Gaussian, of which the precision matrix is very easy to write down (find it!), and then we can apply the theory of the last subsection to numerically find the convergence rate of the deterministic sweep or random scan Gibbs sampler targeting $p(\mu, a, b \,|\, w)$.

A more interesting observation is that different parameterizations of this model can lead to significantly different convergence rates of Gibbs sampling. For example, let's reparameterize the model as follows.

$$\begin{aligned} w_{ijk} &= \eta_{ij} + \epsilon_{ijk}, \quad 1 \le i \le I,\, j \le 1 \le J,\, 1 \le k \le K, \\ \epsilon_{ijk} &\overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2), \quad 1 \le i \le I,\, j \le 1 \le J,\, 1 \le k \le K, \\ \eta_{ij} &\overset{\text{i.i.d.}}{\sim} N(\gamma_i, \sigma_b^2), \quad 1 \le i \le I,\, 1 \le j \le J, \\ \gamma_i &\overset{\text{i.i.d.}}{\sim} N(\mu, \sigma_a^2), \quad 1 \le i \le I, \\ p(\mu) &\propto 1. \end{aligned}$$

This model is exactly the same as the previous one: all we have done is to transform the parameters by the mapping

$$\gamma_i = \mu + a_i, \quad \eta_{ij} = \gamma_i + b_{ij}.$$

The posterior can be written as

$$-2 \log p(\mu, \gamma, \eta \,|\, w) = C + \sum_i \frac{(\gamma_i - \mu)^2}{\sigma_a^2} + \sum_{i,j} \frac{(\eta_{ij} - \gamma_i)^2}{\sigma_b^2} + \sum_{i,j,k} \frac{(w_{ijk} - \eta_{ij})^2}{\sigma_e^2}.$$

Again, $p(\mu, \gamma, \eta \,|\, w)$ is Gaussian, and we can find the convergence rates of the corresponding deterministic sweep and random scan Gibbs samplers. Explicit expressions for the convergence rates of deterministic sweep schemes were obtained in [4].

Theorem 5.2. *Define* $\tilde{\sigma}_a^2 = \sigma_a^2/I$, $\tilde{\sigma}_b^2 = \sigma_b^2/IJ$, $\tilde{\sigma}_e^2 = \sigma_e^2/IJK$. *For the deterministic sweep Gibbs sampler targeting* $p(\mu, a, b \,|\, w)$, *the convergence rate is* $\max\{\tilde{\sigma}_a^2/(\tilde{\sigma}_a^2+\tilde{\sigma}_e^2), \tilde{\sigma}_b^2/(\tilde{\sigma}_b^2+\tilde{\sigma}_e^2)\}$. *For the deterministic sweep Gibbs sampler targeting* $p(\mu, \gamma, \eta \,|\, w)$, *the convergence rate is* $1 - \tilde{\sigma}_a^2\tilde{\sigma}_b^2/[(\tilde{\sigma}_a^2 + \tilde{\sigma}_b^2)(\tilde{\sigma}_b^2 + \tilde{\sigma}_e^2)]$.

Proof. See [4] □

Remark 5.1. This result implies that if $\tilde{\sigma}_a^2 \gg \tilde{\sigma}_b^2 \gg \tilde{\sigma}_e^2$, then the Gibbs sampling with parameterization (μ, γ, η) will be very efficient (convergence rate close to 0), while that with parameterization (μ, a, b) will be very inefficient (convergence rate close to 1). For an intuitive explanation, consider the extreme case $\sigma_e^2 = 0$, which implies that the value of $\mu + a_i + b_j$ for each (i, j) can be exactly determined from the data. If we run the Gibbs sampler with parameterization (μ, a, b), the chain will not be able to move at all, since the conditional posterior distribution of one parameter given the other two is degenerate. In contrast, consider the parameterization (μ, γ, η). The value of η is completely determined, but the chain can still be efficient in exploring the posterior distribution of (μ, γ). Of course one can also consider the parameterizations (μ, γ, b) and (μ, a, η), and the explicit expressions for the convergence rates were also derived in [4]. Any one of these four parameterizations can be the most efficient depending on the relations between $\tilde{\sigma}_a^2, \tilde{\sigma}_b^2, \tilde{\sigma}_e^2$. For more details and a more general theory, see [4].

5.3 Proximal Sampling

Proximal sampling is a class of sampling methods that became popular very recently. Consider a distribution π on $\mathbb{R}^d$ with density $\pi(x) \propto e^{-f(x)}$. Let's augment the state space and consider a joint distribution

$$\pi(x, y) \propto \exp\left\{-f(x) - \frac{1}{2\lambda}\|x - y\|_2^2\right\}, \tag{1}$$

where $\lambda > 0$ is a tuning parameter. As in Section 5.1, we use $\pi_{1\,|\,2}$ and $\pi_{2\,|\,1}$ to denote the two conditional distributions. Clearly,

$$\pi_{1\,|\,2}(x \,|\, y) = \frac{\exp\left\{-f(x) - \frac{1}{2\lambda}\|x - y\|_2^2\right\}}{\int \exp\left\{-f(z) - \frac{1}{2\lambda}\|z - y\|_2^2\right\} \mathrm{d}z},$$

and $\pi_{2\,|\,1}(y \,|\, x)$ is the density of $N(x, \lambda I_d)$. The proximal sampling algorithm is the Gibbs sampler that targets $\pi(x, y)$. Since $\pi(x, y)$ has $\pi(x)$ as the marginal distribution, we can collect samples $(X_t, Y_t)_{t\geq 0}$ from such a Gibbs sampler and then use $(X_t)_{t\geq 0}$ to approximate the distribution of interest, $\pi(x)$.

The main challenge is how to perform sampling from the conditional distribution $\pi_{1\,|\,2}(x \,|\, y)$. One possible method is to do rejection sampling. If we know $\pi(x)$ has light tails and we can find $m(x; y, \lambda) = \arg\min_x\{-f(x) - \frac{1}{2\lambda}\|x - y\|_2^2\}$, then a normal distribution with mean $m(x; y, \lambda)$ can be used as an efficient reference distribution [1]. Of course, one can also

consider more complicated schemes, including running another MCMC sampler targeting $\pi_{1\,|\,2}(x\,|\,y)$.

Naturally, one may wonder why such a scheme is useful, since apparently we are converting one sampling problem (i.e., $\pi(x)$) to a class of sampling problems (i.e., $\{\pi_{1\,|\,2}(x\,|\,y)\colon y \in \mathbb{R}^d\}$). To explain the motivation, we need the following definitions.

Definition 5.1. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and $\alpha, \beta > 0$. We say f is β-smooth if ∇f is β-Lipschitz; that is,

$$\|\nabla f(y) - \nabla f(x)\|_2 \le \beta \|y - x\|_2, \quad \forall x, y \in \mathbb{R}^d.$$

We say f is α-strongly convex if

$$f(y) \ge f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^d.$$

Remark 5.2. If f is twice differentiable, then f is β-smooth and α-strongly convex if

$$\alpha I_d \preceq \nabla^2 f(x) \preceq \beta I_d, \quad \forall x \in \mathbb{R}^d,$$

where the notation $A \preceq B$ means that $B - A$ is positive semi-definite. The ratio $\kappa = \beta/\alpha \ge 1$ is often called the condition number in the sampling literature. A smaller condition number implies that $\pi(x) \propto e^{-f(x)}$ is easier to sample from, since f is similar to a quadratic function (and thus π is similar to a normal distribution).

Now assume that f in (1) is twice differentiable, β-smooth and α-strongly convex. So it has condition number $\kappa(f) = \beta/\alpha$. Fix some $y \in \mathbb{R}^d$, $\lambda > 0$ and consider

$$g_y(x) = f(x) + \frac{1}{2\lambda}\|x - y\|_2^2.$$

Then g_y is again smooth and strongly convex, and its condition number is

$$\kappa(g_y) \le \frac{1 + \beta\lambda}{1 + \alpha\lambda} \le \frac{\beta}{\alpha} = \kappa(f).$$

If we know β, we can use $\lambda = 1/\beta$, which guarantees that $\kappa(g_y) \le 2$. Hence, the term $\|x - y\|_2^2/(2\lambda)$ regularizes the target density and makes it easier to sample from.

Moreover, suppose that f in (1) is twice differentiable and β-smooth, but f may not be strongly convex. If $1/\lambda \ge \beta$, then g_y is still strongly convex with condition number

$$\kappa(g_y) \le \frac{1 + \beta\lambda}{1 - \beta\lambda}.$$

For more details about the convergence properties of proximal sampling, see, e.g., [2, 1].

Exercise 5.4. Prove that if f is strongly convex, then it is strictly convex.

References

[1] Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

[2] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

[3] Gareth O Roberts and Sujit K Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(2):291–317, 1997.

[4] Giacomo Zanella and Gareth Roberts. Multilevel linear models, Gibbs samplers and multigrid decompositions (with discussion). *Bayesian Analysis*, 16(4):1309–1391, 2021.