Approximate Inference via Sampling (2) MCMC algos: MH and Gibbs Sampling

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Some MCMC Algorithms

Metropolis-Hastings (MH) Sampling (1960)

- Suppose we wish to generate samples from a target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$
- Assume a suitable proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, e.g., $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2\boldsymbol{I})$
- In each step, draw $\boldsymbol{z}^*$ from $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$ and accept $\boldsymbol{z}^*$ with probability

Favors acceptance of $\boldsymbol{z}^*$ if it is more probable than $\boldsymbol{z}^{(\tau)}$ (under $p(\boldsymbol{z})$)

Favor acceptance of $\boldsymbol{z}^*$ if our proposal allows reverting to the older state $\boldsymbol{z}^{(\tau)}$ from $\boldsymbol{z}^*$

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)\, q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})\, q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right)$$

Favor acceptance of $\boldsymbol{z}^*$ if it had very low chance of being generated by the proposal but it does have high probability $\tilde{p}(\boldsymbol{z}^*)$ under the target

- Transition function of this Markov Chain: $T(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)}) = A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})$
- Exercise: Show that $T(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})$ satisfies the detailed balance property

$$p(\boldsymbol{z})T(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}) = p(\boldsymbol{z}^{(\tau)})T(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$$

The MH Sampling Algorithm

- Initialize $\boldsymbol{z}^{(1)}$ randomly
- For $\ell = 1,2, \dots, L$
  - Sample $\boldsymbol{z}^* \sim q(\boldsymbol{z}^*|\boldsymbol{z}^{(\ell)})$ and $u \sim \text{Unif}(0,1)$
  - Compute acceptance probability

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\ell)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)q(\boldsymbol{z}^{(\ell)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\ell)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\ell)})}\right)$$

  - If $A\left(\boldsymbol{z}^*, \boldsymbol{z}^{(\ell)}\right) > u$

$$\boldsymbol{z}^{(\ell+1)} = \boldsymbol{z}^*$$

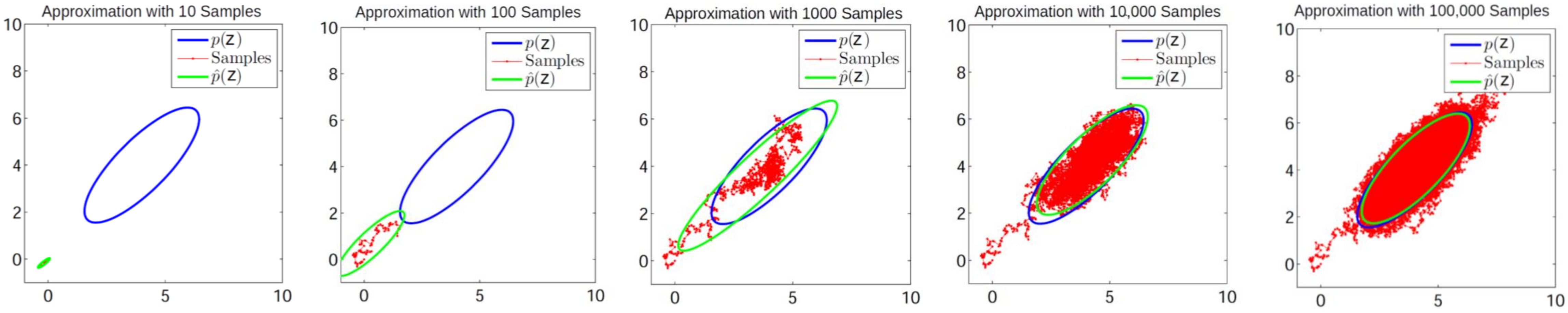
Meaning accepting $\boldsymbol{z}^*$ with probability $A\left(\boldsymbol{z}^*, \boldsymbol{z}^{(\ell)}\right)$

  - Else

$$\boldsymbol{z}^{(\ell+1)} = \boldsymbol{z}^{(\ell)}$$

MH Sampling in Action: A Toy Example..

- Target distribution $p(\mathbf{z}) = \mathcal{N}\left(\begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right)$
- Proposal distribution $q(\mathbf{z}^{(t)}|\mathbf{z}^{(t-1)}) = \mathcal{N}\left(\mathbf{z}^{(t-1)}, \begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$

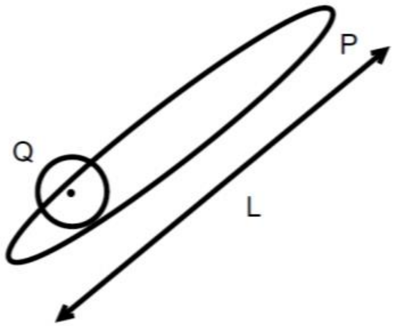


MH Sampling: Some Comments

- If prop. distrib. is symmetric, we get Metropolis Sampling algo (Metropolis, 1953) with

$$A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^*)}{\tilde{p}(\mathbf{z}^{(\tau)})}\right)$$

- Some limitations of MH sampling
  - Can sometimes have very slow convergence (also known as slow "mixing")



$Q(\mathbf{z}|\mathbf{z}^{(\tau)}) = \mathcal{N}(\mathbf{z}|\mathbf{z}^{(\tau)}, \sigma^2 \mathbf{I})$

σ large $\Rightarrow$ many rejections

σ small $\Rightarrow$ slow diffusion

$\sim \left(\frac{L}{\sigma}\right)^2$ iterations required for convergence

  - Computing acceptance probability can be expensive*, e.g., if $\boldsymbol{p(z)} = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ is some target posterior then $\tilde{p}(\boldsymbol{z})$ would require computing likelihood on all the data points (expensive)

*Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget (Korattikara et al, 2014), Firefly Monte Carlo: Exact MCMC with Subsets of Data {(Maclaurin and Adams, 2015)

Gibbs Sampling (Geman & Geman, 1984)

- Goal: Sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = [z_1, z_2, \dots, z_M]$
- Suppose we can't sample from $p(\boldsymbol{z})$ but can sample from each conditional $p(z_i|\boldsymbol{z}_{-i})$
  - In Bayesian models, can be done easily if we have a locally conjugate model
- For Gibbs sampling, the proposal is the conditional distribution $p(z_i|\boldsymbol{z}_{-i})$
- Gibbs sampling samples from these conditionals in a cyclic order

Hence no need to compute it

- Gibbs sampling is equivalent to MH sampling with acceptance prob. = 1

$$A(\boldsymbol{z}^*, \boldsymbol{z}) = \frac{p(\boldsymbol{z}^*)q(\boldsymbol{z}|\boldsymbol{z}^*)}{p(\boldsymbol{z})q(\boldsymbol{z}^*|\boldsymbol{z})} = \frac{p(z_i^*|\boldsymbol{z}^*_{-i})p(\boldsymbol{z}^*_{-i})p(z_i|\boldsymbol{z}^*_{-i})}{p(z_i|\boldsymbol{z}_{-i})p(\boldsymbol{z}_{-i})p(z_i^*|\boldsymbol{z}_{-i})} = 1$$

where we use the fact that $\boldsymbol{z}^*_{-i} = \boldsymbol{z}_{-i}$

Since only one component is changed at a time

Gibbs Sampling: Sketch of the Algorithm

- M: Total number of variables, T: number of Gibbs sampling iterations

1. Initialize $\{z_i : i = 1, \ldots, M\}$

Assuming $\boldsymbol{z} = [z_1, z_2, \ldots, z_M]$

2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

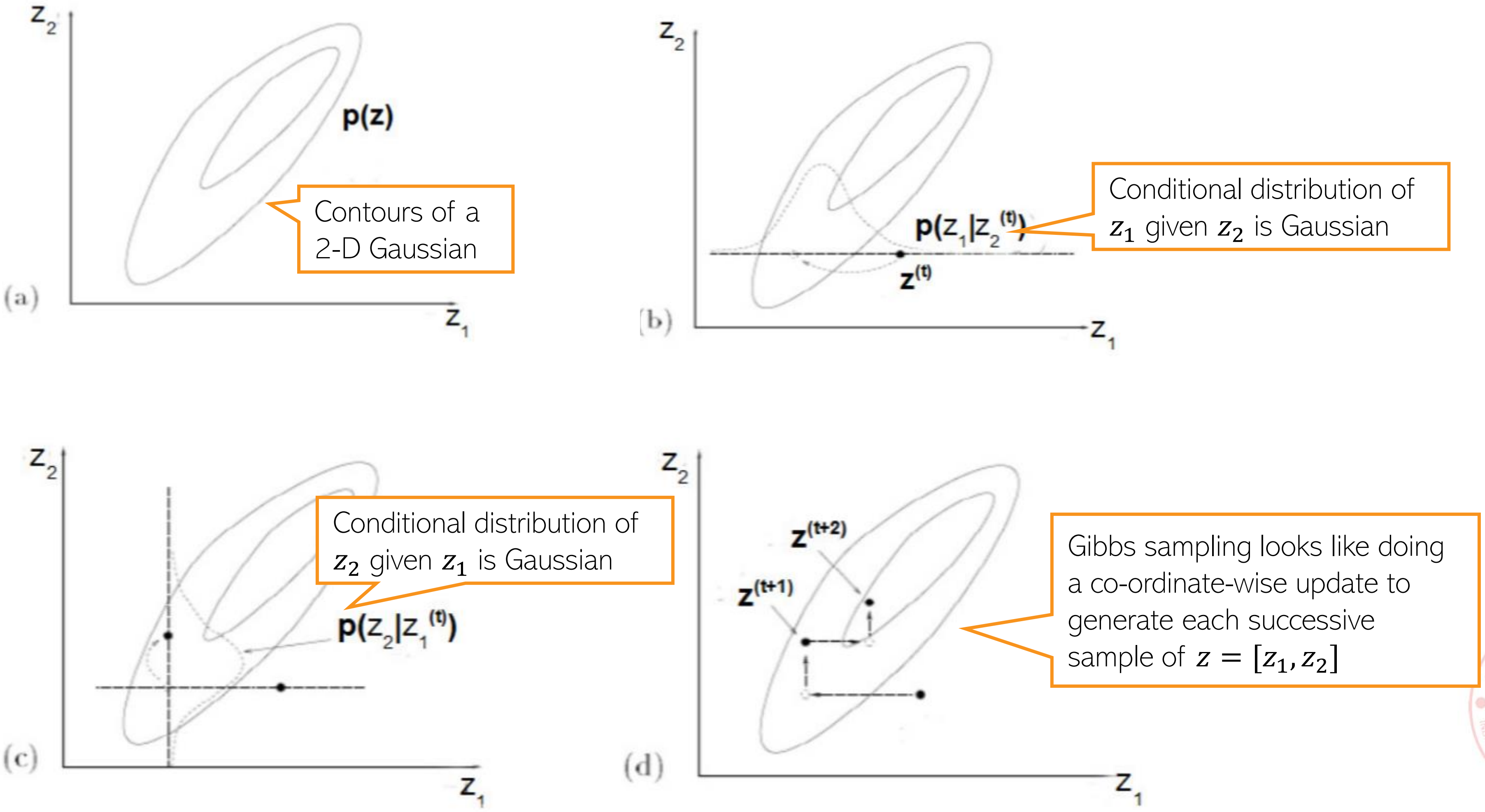
CP of each component of $\boldsymbol{z}$ uses the most recent values (from this or the previous iteration) of all the other components

Each iteration will give us one sample $\boldsymbol{z}^{(\tau)}$ of $\boldsymbol{z} = [z_1, z_2, \ldots, z_M]$

- Note: Order of updating the variables usually doesn't matter (but see "Scan Order in Gibbs Sampling: Models in Which it Matters and Bounds on How Much" from NIPS 2016)

Gibbs Sampling: A Simple Example

- Can sample from a 2-D Gaussian using 1-D Gaussians



Gibbs Sampling: Some Comments

- One of the most popular MCMC algorithms
- Very easy to derive and implement for locally conjugate models
- Many variations exist, e.g.,
  - Blocked Gibbs: sample more than one component jointly (sometimes possible)
  - Rao-Blackwellized Gibbs: Can collapse (i.e., integrate out) the unneeded components while sampling. Also called "collapsed" Gibbs sampling
  - MH within Gibbs: If CPs are not easy to sample distributions
- Instead of sampling from CPs, an alternative is to use the mode of the CPs
  - Called the "Iterative Conditional Mode" (ICM) algorithm
  - ICM doesn't give the posterior though – it's more like ALT-OPT to get (approx) MAP estimate

Coming Up Next

- Using posterior's gradient info in sampling algorithms
- Online MCMC algorithms
- Recent advances in MCMC
- Some other practical issues (convergence etc)