Markov Chain Monte Carlo

Shinichi Nakajima

Technische Universität Berlin

14. Jan. 2016

Bayesian Learning (easy cases)

Gaussian

$$\text{Posterior:} \quad p(\boldsymbol{w}|\mathcal{D}) \propto p(\boldsymbol{w}, \mathcal{D}) \propto \underbrace{\exp\left(-\frac{\boldsymbol{\mu}^\top \boldsymbol{C}^{-1}\boldsymbol{\mu}}{2}\right)}_{\text{prior}} \underbrace{\prod_{i=1}^{n} \exp\left(-\frac{(\boldsymbol{\mu}-\boldsymbol{x}^{(i)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{x}^{(i)})}{2}\right)}_{\text{likelihood}}$$

Complete the square! $\quad \propto \exp\left(-\frac{(\boldsymbol{\mu}-\boldsymbol{m}_\mu)^\top \boldsymbol{S}_\mu^{-1}(\boldsymbol{\mu}-\boldsymbol{m}_\mu)}{2}\right)$

$$\therefore \quad p(\boldsymbol{w}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{m}_\mu, \boldsymbol{S}_\mu)$$

Multinomial

$$\text{Posterior:} \quad p(\boldsymbol{w}|\mathcal{D}) \propto p(\boldsymbol{w}, \mathcal{D}) \propto \underbrace{\prod_{k=1}^{K} \theta_k^{\phi_k - 1}}_{\text{prior}} \underbrace{\prod_{k=1}^{K} \theta_k^{x_k}}_{\text{likelihood}}$$

Add exponents! $\quad \propto \prod_{k=1}^{K} \theta_k^{x_k + \phi_k - 1}$

$$\therefore \quad p(\boldsymbol{w}|\mathcal{D}) = \mathrm{Dir}\left(\boldsymbol{\theta}; \{x_k + \phi_k\}_{k=1}^{K}\right)$$

These two patterns cover most of the cases, including approximate learning!

Bayesian learning is computationally hard, but Bayesian do easy work. All what you need are square completion, addition, and Wikipedia (to consult on moments)!

If Wikipedia does not know, ...

moments (expectation of some function $f(\boldsymbol{w})$) are approximated by

$$\int f(\boldsymbol{w}) \cdot p(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w} \approx \frac{1}{J} \sum_{j=1}^{J} f(\boldsymbol{w}^{(j)}), \quad \text{where} \quad \boldsymbol{w}^{(j)} \sim p(\boldsymbol{w}|\mathcal{D}).$$

MCMC Sampling (Metropolis-Hastings, Gibbs Sampling)

or

$$\int f(\boldsymbol{w}) \cdot p(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w} \approx \int f(\boldsymbol{w}) \cdot q(\boldsymbol{w}) d\boldsymbol{w} \quad \text{where} \quad q(\boldsymbol{w}) \approx p(\boldsymbol{w}|\mathcal{D}).$$

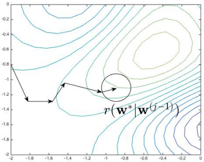
Deterministic Approximation (Laplace Approximation, Variational Bayes, Expectation Propagation)

$q(\boldsymbol{w})$ must be in a known form.

Sampling from *unnormalized* distribution

We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w}, \mathcal{D})}_{\text{known!}}.$$



Metropolis-Hastings (MH) (a general method):

In the j-th iteration,

1. Draw a sample $\boldsymbol{w}^* \sim r(\boldsymbol{w}|\boldsymbol{w}^{(j-1)})$, e.g., $r(\boldsymbol{w}|\boldsymbol{w}^{(j-1)}) \propto \exp\left(-\frac{\|\boldsymbol{w}-\boldsymbol{w}^{(j-1)}\|^2}{2\gamma^2}\right)$.
2. Set the j-th sample to

$$\boldsymbol{w}^{(j)} = \begin{cases} \boldsymbol{w}^* & \text{with probability} \quad T, \\ \boldsymbol{w}^{(j-1)} & \text{with probability} \quad 1 - T, \end{cases}$$

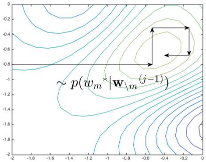
with acceptance probability (which satisfies detailed balance property appendix)

$$T = \min\left(1, \frac{p\left(\boldsymbol{w}^*, \mathcal{D}\right) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^*\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) r\left(\boldsymbol{w}^*|\boldsymbol{w}^{(j-1)}\right)}\right).$$

Sampling from *unnormalized* distribution

We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w}, \mathcal{D})}_{\text{known!}}.$$

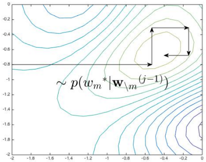


If $p\left(w_m|\boldsymbol{w}_{\setminus m}, \mathcal{D}\right)$ is in a known form (sampler available),

Sampling from *unnormalized* distribution

We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w},\mathcal{D})}_{\text{known!}}.$$



If $p\left(w_m|\boldsymbol{w}_{\backslash m},\mathcal{D}\right)$ is in a known form (sampler available),

Gibbs Sampling (GS) (a special case of MH) is efficient:

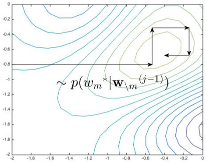
In the j-th iteration,

1. Draw a sample *element* $w_m^* \sim p\left(w_m|\boldsymbol{w}_{\backslash m}^{(j-1)},\mathcal{D}\right)$, and set $\boldsymbol{w}_{\backslash m}^* = \boldsymbol{w}_{\backslash m}^{(j-1)}$.
2. Set $\boldsymbol{w}_m^{(j)} = \boldsymbol{w}^*$ (with probability 1).

The reason why the acceptance probability is one is

$$T = \min\left(1, \frac{p(\boldsymbol{w}^*,\mathcal{D})r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^*\right)}{p\left(\boldsymbol{w}^{(j-1)},\mathcal{D}\right)r\left(\boldsymbol{w}^*|\boldsymbol{w}^{(j-1)}\right)}\right)$$

Sampling from *unnormalized* distribution



We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w}, \mathcal{D})}_{\text{known!}}.$$

If $p\left(w_m|\boldsymbol{w}_{\backslash m}, \mathcal{D}\right)$ is in a known form (sampler available),

Gibbs Sampling (GS) (a special case of MH) is efficient:

In the j-th iteration,

1. Draw a sample *element* $w_m^* \sim p\left(w_m|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right)$, and set $\boldsymbol{w}_{\backslash m}^* = \boldsymbol{w}_{\backslash m}^{(j-1)}$.
2. Set $\boldsymbol{w}_m^{(j)} = \boldsymbol{w}^*$ (with probability 1).

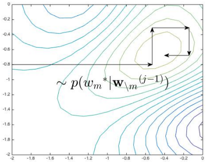
The reason why the acceptance probability is one is

$$T = \min\left(1, \frac{p(\boldsymbol{w}^*, \mathcal{D}) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^*\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) r\left(\boldsymbol{w}^*|\boldsymbol{w}^{(j-1)}\right)}\right) = \min\left(1, \frac{p(\boldsymbol{w}^*, \mathcal{D}) p\left(w_m^{(j-1)}|\boldsymbol{w}_{\backslash m}^*, \mathcal{D}\right) p\left(\boldsymbol{w}_{\backslash m}^{(j-1)}|\boldsymbol{w}_{\backslash m}^*, \mathcal{D}\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) p\left(w_m^*|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right) p\left(\boldsymbol{w}_{\backslash m}^*|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right)}\right).$$

Sampling from *unnormalized* distribution

We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w},\mathcal{D})}_{\text{known!}}.$$



If $p\left(w_m|\boldsymbol{w}_{\backslash m},\mathcal{D}\right)$ is in a known form (sampler available),

Gibbs Sampling (GS) (a special case of MH) is efficient:

In the j-th iteration,

1. Draw a sample *element* $w_m^* \sim p\left(w_m|\boldsymbol{w}_{\backslash m}^{(j-1)},\mathcal{D}\right)$, and set $\boldsymbol{w}_{\backslash m}^* = \boldsymbol{w}_{\backslash m}^{(j-1)}$.
2. Set $\boldsymbol{w}_m^{(j)} = \boldsymbol{w}^*$ (with probability 1).

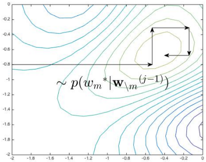
The reason why the acceptance probability is one is

$$T = \min\left(1, \frac{p(\boldsymbol{w}^*,\mathcal{D})r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^*\right)}{p\left(\boldsymbol{w}^{(j-1)},\mathcal{D}\right)r\left(\boldsymbol{w}^*|\boldsymbol{w}^{(j-1)}\right)}\right) = \min\left(1, \frac{p(\boldsymbol{w}^*,\mathcal{D})p\left(w_m^{(j-1)}|\boldsymbol{w}_{\backslash m}^*,\mathcal{D}\right)p\left(\boldsymbol{w}_{\backslash m}^{(j-1)}|\boldsymbol{w}_{\backslash m}^*,\mathcal{D}\right)}{p\left(\boldsymbol{w}^{(j-1)},\mathcal{D}\right)p\left(w_m^*|\boldsymbol{w}_{\backslash m}^{(j-1)},\mathcal{D}\right)p\left(\boldsymbol{w}_{\backslash m}^*|\boldsymbol{w}_{\backslash m}^{(j-1)},\mathcal{D}\right)}\right).$$

Sampling from *unnormalized* distribution

We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w}, \mathcal{D})}_{\text{known!}}.$$



If $p\left(w_m|\boldsymbol{w}_{\backslash m}, \mathcal{D}\right)$ is in a known form (sampler available),

Gibbs Sampling (GS) (a special case of MH) is efficient:

In the j-th iteration,

1. Draw a sample *element* $w_m^* \sim p\left(w_m|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right)$, and set $\boldsymbol{w}_{\backslash m}^* = \boldsymbol{w}_{\backslash m}^{(j-1)}$.
2. Set $\boldsymbol{w}_m^{(j)} = \boldsymbol{w}^*$ (with probability 1).

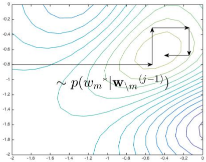
The reason why the acceptance probability is one is

$$T = \min\left(1, \frac{p(\boldsymbol{w}^*, \mathcal{D}) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^*\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) r\left(\boldsymbol{w}^*|\boldsymbol{w}^{(j-1)}\right)}\right) = \min\left(1, \frac{p(\boldsymbol{w}^*, \mathcal{D}) p\left(w_m^{(j-1)}|\boldsymbol{w}_{\backslash m}^*, \mathcal{D}\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) p\left(w_m^*|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right)}\right).$$

Sampling from *unnormalized* distribution

We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w}, \mathcal{D})}_{\text{known!}}.$$



If $p\left(w_m|\boldsymbol{w}_{\backslash m}, \mathcal{D}\right)$ is in a known form (sampler available),

Gibbs Sampling (GS) (a special case of MH) is efficient:

In the j-th iteration,

1. Draw a sample *element* $w_m^* \sim p\left(w_m|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right)$, and set $\boldsymbol{w}_{\backslash m}^* = \boldsymbol{w}_{\backslash m}^{(j-1)}$.
2. Set $\boldsymbol{w}_m^{(j)} = \boldsymbol{w}^*$ (with probability 1).

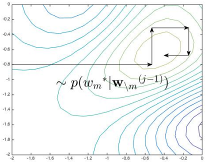
The reason why the acceptance probability is one is

$$T = \min\left(1, \frac{p(\boldsymbol{w}^*, \mathcal{D}) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^*\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) r\left(\boldsymbol{w}^*|\boldsymbol{w}^{(j-1)}\right)}\right) = \min\left(1, \frac{p(\boldsymbol{w}^*, \mathcal{D}) p\left(w_m^{(j-1)}|\boldsymbol{w}_{\backslash m}^*, \mathcal{D}\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) p\left(w_m^*|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right)}\right).$$

Sampling from *unnormalized* distribution

We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w}, \mathcal{D})}_{\text{known!}}.$$



If $p\left(w_m|\boldsymbol{w}_{\setminus m}, \mathcal{D}\right)$ is in a known form (sampler available),

Gibbs Sampling (GS) (a special case of MH) is efficient:

In the j-th iteration,

1. Draw a sample *element* $w_m^* \sim p\left(w_m|\boldsymbol{w}_{\setminus m}^{(j-1)}, \mathcal{D}\right)$, and set $\boldsymbol{w}_{\setminus m}^* = \boldsymbol{w}_{\setminus m}^{(j-1)}$.
2. Set $\boldsymbol{w}_m^{(j)} = \boldsymbol{w}^*$ (with probability 1).

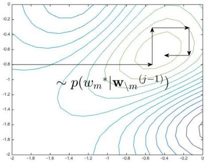
The reason why the acceptance probability is one is

$$T = \min\left(1, \frac{p(\boldsymbol{w}^*, \mathcal{D}) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^*\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) r\left(\boldsymbol{w}^*|\boldsymbol{w}^{(j-1)}\right)}\right) = \min\left(1, \frac{p\left(w_m^*|\boldsymbol{w}_{\setminus m}^*, \mathcal{D}\right) p\left(\boldsymbol{w}_{\setminus m}^*, \mathcal{D}\right) p\left(w_m^{(j-1)}|\boldsymbol{w}_{\setminus m}^*, \mathcal{D}\right)}{p\left(w_m^{(j-1)}|\boldsymbol{w}_{\setminus m}^{(j-1)}, \mathcal{D}\right) p\left(\boldsymbol{w}_{\setminus m}^{(j-1)}, \mathcal{D}\right) p\left(w_m^*|\boldsymbol{w}_{\setminus m}^{(j-1)}, \mathcal{D}\right)}\right).$$

Sampling from *unnormalized* distribution

We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w}, \mathcal{D})}_{\text{known!}}.$$



If $p\left(w_m|\boldsymbol{w}_{\backslash m}, \mathcal{D}\right)$ is in a known form (sampler available),

Gibbs Sampling (GS) (a special case of MH) is efficient:

In the j-th iteration,

1. Draw a sample *element* $w_m^* \sim p\left(w_m|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right)$, and set $\boldsymbol{w}_{\backslash m}^* = \boldsymbol{w}_{\backslash m}^{(j-1)}$.
2. Set $\boldsymbol{w}_m^{(j)} = \boldsymbol{w}^*$ (with probability 1).

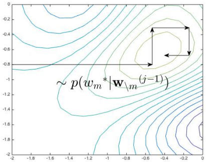
The reason why the acceptance probability is one is

$$T = \min\left(1, \frac{p(\boldsymbol{w}^*, \mathcal{D}) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^*\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) r\left(\boldsymbol{w}^*|\boldsymbol{w}^{(j-1)}\right)}\right) = \min\left(1, \frac{p\left(w_m^*|\boldsymbol{w}_{\backslash m}^*, \mathcal{D}\right) p\left(\boldsymbol{w}_{\backslash m}^*, \mathcal{D}\right) p\left(w_m^{(j-1)}|\boldsymbol{w}_{\backslash m}^*, \mathcal{D}\right)}{p\left(w_m^{(j-1)}|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right) p\left(\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right) p\left(w_m^*|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right)}\right).$$

Sampling from *unnormalized* distribution

We need samples such that

$$\boldsymbol{w} \sim \underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{unknown}} \propto \underbrace{p(\boldsymbol{w}, \mathcal{D})}_{\text{known!}}.$$



If $p\left(w_m|\boldsymbol{w}_{\backslash m}, \mathcal{D}\right)$ is in a known form (sampler available),

Gibbs Sampling (GS) (a special case of MH) is efficient:

In the j-th iteration,

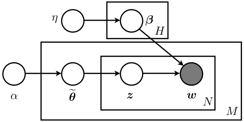
1. Draw a sample *element* $w_m^* \sim p\left(w_m|\boldsymbol{w}_{\backslash m}^{(j-1)}, \mathcal{D}\right)$, and set $\boldsymbol{w}^*_{\backslash m} = \boldsymbol{w}_{\backslash m}^{(j-1)}$.
2. Set $\boldsymbol{w}_m^{(j)} = \boldsymbol{w}^*$ (with probability 1).

The reason why the acceptance probability is one is

$$T = \min\left(1, \frac{p(\boldsymbol{w}^*, \mathcal{D}) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^*\right)}{p\left(\boldsymbol{w}^{(j-1)}, \mathcal{D}\right) r\left(\boldsymbol{w}^*|\boldsymbol{w}^{(j-1)}\right)}\right) = \min\left(1, \frac{p\left(w_m^*|\boldsymbol{w}^*_{\backslash m}, \mathcal{D}\right) p\left(\boldsymbol{w}^*_{\backslash m}, \mathcal{D}\right) p\left(w_m^{(j-1)}|\boldsymbol{w}^*_{\backslash m}, \mathcal{D}\right)}{p\left(w_m^{(j-1)}|\boldsymbol{w}^*_{\backslash m}, \mathcal{D}\right) p\left(\boldsymbol{w}^*_{\backslash m}, \mathcal{D}\right) p\left(w_m^*|\boldsymbol{w}^*_{\backslash m}, \mathcal{D}\right)}\right) = 1.$$

Latent Dirichlet Allocation

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



Likelihood and Prior

$$p(\boldsymbol{w}^{(n,m)}|\boldsymbol{\Theta}, \boldsymbol{B}) = \prod_{l=1}^{L} ((\boldsymbol{B}\boldsymbol{\Theta}^{\top})_{l,m})^{w_l^{(n,m)}},$$

$$p(\boldsymbol{\Theta}) \propto \prod_{m=1}^{M} \prod_{h=1}^{H} (\Theta_{m,h})^{\alpha-1},$$

$$p(\boldsymbol{B}) \propto \prod_{h=1}^{H} \prod_{l=1}^{L} (B_{l,h})^{\eta-1}.$$

$\boldsymbol{\Theta} \in [0,1]^{M\times H}$:Document parameter

$\boldsymbol{B} \in [0,1]^{L\times H}$:Topic parameter

M :# of documents

L :vocabulary size

H :# of topics ($\leq \min(M, L)$)

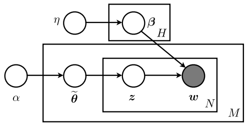
$N^{(m)}$:# of words in m-th document

α, β :Hyperparameters

Word distribution is a mixture of multinomial *topic* distribution.

Latent Dirichlet Allocation

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



Likelihood and Prior

$$p(\boldsymbol{w}^{(n,m)}|\boldsymbol{\Theta}, \boldsymbol{B}) = \prod_{l=1}^{L} \left(\sum_{h=1}^{H} B_{l,h} \Theta_{m,h} \right)^{w_l^{(n,m)}},$$

$$p(\boldsymbol{\Theta}) \propto \prod_{m=1}^{M} \prod_{h=1}^{H} (\Theta_{m,h})^{\alpha-1},$$

$$p(\boldsymbol{B}) \propto \prod_{h=1}^{H} \prod_{l=1}^{L} (B_{l,h})^{\eta-1}.$$

$\boldsymbol{\Theta} \in [0,1]^{M\times H}$:Document parameter

$\boldsymbol{B} \in [0,1]^{L\times H}$:Topic parameter

M :# of documents

L :vocabulary size

H :# of topics ($\leq \min(M, L)$)

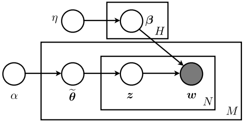
$N^{(m)}$:# of words in m-th document

α, β :Hyperparameters

Sum in the probability (mixture) is intractable.

Latent Dirichlet Allocation

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



Complete Likelihood and Prior

$$p(\boldsymbol{w}^{(n,m)}, \boldsymbol{z}^{(n,m)} | \boldsymbol{\Theta}, \boldsymbol{B}) = \prod_{l=1}^{L} \left(\prod_{h=1}^{H} \left(B_{l,h} \Theta_{m,h} \right)^{z_h^{(n,m)}} \right)^{w_l^{(n,m)}},$$

$$p(\boldsymbol{\Theta}) \propto \prod_{m=1}^{M} \prod_{h=1}^{H} (\Theta_{m,h})^{\alpha-1},$$

$$p(\boldsymbol{B}) \propto \prod_{h=1}^{H} \prod_{l=1}^{L} (B_{l,h})^{\eta-1}.$$

$\boldsymbol{\Theta} \in [0,1]^{M \times H}$:Document parameter

$\boldsymbol{B} \in [0,1]^{L \times H}$:Topic parameter

$\boldsymbol{z}^{(n,m)} \in [0,1]^{H}$:Topic assignment for each word

M :# of documents

L :vocabulary size

H :# of topics ($\leq \min(M, L)$)

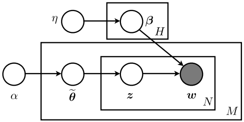
$N^{(m)}$:# of words in m-th document

α, β :Hyperparameters

Latent variable $\boldsymbol{z}^{(n,m)}$ changes the sum to the product,

Latent Dirichlet Allocation

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



Complete Likelihood and Prior

$$p(\boldsymbol{w}^{(n,m)}, \boldsymbol{z}^{(n,m)} | \boldsymbol{\Theta}, \boldsymbol{B}) = \prod_{h=1}^{H} (\Theta_{m,h})^{z_h^{(n,m)}} \quad \prod_{l=1}^{L} (B_{l,h})^{w_l^{(n,m)} z_h^{(n,m)}},$$

$$p(\boldsymbol{\Theta}) \propto \prod_{m=1}^{M} \prod_{h=1}^{H} (\Theta_{m,h})^{\alpha-1},$$

$$p(\boldsymbol{B}) \propto \prod_{h=1}^{H} \prod_{l=1}^{L} (B_{l,h})^{\eta-1}.$$

$\boldsymbol{\Theta} \in [0,1]^{M\times H}$:Document parameter

$\boldsymbol{B} \in [0,1]^{L\times H}$:Topic parameter

$\boldsymbol{z}^{(n,m)} \in [0,1]^{H}$:Topic assignment for each word

M :# of documents

L :vocabulary size

H :# of topics ($\leq \min(M, L)$)

$N^{(m)}$:# of words in m-th document

α, β :Hyperparameters

Latent variable $\boldsymbol{z}^{(n,m)}$ changes the sum to the product, and makes likelihood separable.

Joint is NOT in known form, but ...

Posterior: $p\left(\{z^{(n,m)}\}, \Theta, B | \{w^{(n,m)}\}\right) \propto p(\{z^{(n,m)}\}, \Theta, B, \{w^{(n,m)}\})$, where

$$p(\{z^{(n,m)}\}, \Theta, B, \{w^{(n,m)}\}) = p(\{w^{(n,m)}\}, \{z^{(n,m)}\} | \Theta, B) p(\Theta) p(B)$$
$$\propto \prod_{m=1}^{M} \prod_{n=1}^{N^{(m)}} \prod_{h=1}^{H} (\Theta_{m,h})^{z_h^{(n,m)} + (\alpha-1)/N^{(m)}} \prod_{l=1}^{L} (B_{l,h})^{w_l^{(n,m)} z_h^{(n,m)} + (\eta-1)/(MN^{(m)})}.$$

Joint distribution (on $\{z^{(n,m)}\}, \Theta, B$) is not in a known form, but *conditionals*

$$p(\{z^{(n,m)}\} | \Theta, B, \{w^{(n,m)}\}) \propto \prod_{m=1}^{M} \prod_{n=1}^{N^{(m)}} \prod_{h=1}^{H} \left(\Theta_{m,h} \prod_{l=1}^{L} (B_{l,h})^{w_l^{(n,m)}}\right)^{z_h^{(n,m)}}$$ Multinomial

$$p(\Theta | \{z^{(n,m)}\}, \{w^{(n,m)}\}) \propto \prod_{m=1}^{M} \prod_{n=1}^{N^{(m)}} \prod_{h=1}^{H} (\Theta_{m,h})^{z_h^{(n,m)} + (\alpha-1)/N^{(m)}}$$ Dirichlet

$$p(B | \{z^{(n,m)}\}, \{w^{(n,m)}\}) \propto \prod_{m=1}^{M} \prod_{n=1}^{N^{(m)}} \prod_{h=1}^{H} \prod_{l=1}^{L} (B_{l,h})^{w_l^{(n,m)} z_h^{(n,m)} + (\eta-1)/(MN^{(m)})}$$ Dirichlet

are in known forms. → Gibbs sampling!

Gibbs Sampling

(Naive) Gibbs Sampling:

In j-th step,

1. For each (n, m) independently, draw $\boldsymbol{z}^{(j,n,m)} \sim p(\boldsymbol{z}^{(n,m)}|\boldsymbol{\Theta}^{(j-1)}, \boldsymbol{B}^{(j-1)}, \{\boldsymbol{w}^{(n,m)}\})$,
2. For each m independently, draw $\widetilde{\boldsymbol{\theta}}_m^{(j)} \sim p(\widetilde{\boldsymbol{\theta}}_m|\{\boldsymbol{z}^{(j,n,m)}\}, \{\boldsymbol{w}^{(n,m)}\})$,
3. For each h independently, draw $\boldsymbol{\beta}_h^{(j)} \sim p(\boldsymbol{\beta}_h|\{\boldsymbol{z}^{(j,n,m)}\}, \{\boldsymbol{w}^{(n,m)}\})$.

$$p(\{\boldsymbol{z}^{(n,m)}\}|\boldsymbol{\Theta}, \boldsymbol{B}, \{\boldsymbol{w}^{(n,m)}\}) \propto \prod_{m=1}^{M} \prod_{n=1}^{N^{(m)}} \prod_{h=1}^{H} \left(\Theta_{m,h} \prod_{l=1}^{L} (B_{l,h})^{w_l^{(n,m)}}\right)^{z_h^{(n,m)}}$$ Multinomial

$$p(\boldsymbol{\Theta}|\{\boldsymbol{z}^{(n,m)}\}, \{\boldsymbol{w}^{(n,m)}\}) \propto \prod_{m=1}^{M} \prod_{n=1}^{N^{(m)}} \prod_{h=1}^{H} (\Theta_{m,h})^{z_h^{(n,m)}+(\alpha-1)/N^{(m)}}$$ Dirichlet

$$p(\boldsymbol{B}|\{\boldsymbol{z}^{(n,m)}\}, \{\boldsymbol{w}^{(n,m)}\}) \propto \prod_{m=1}^{M} \prod_{n=1}^{N^{(m)}} \prod_{h=1}^{H} \prod_{l=1}^{L} (B_{l,h})^{w_l^{(n,m)} z_h^{(n,m)}+(\eta-1)/(MN^{(m)})}$$ Dirichlet

Conditionally independent! ($\rightarrow$ easily parallelized) But we can do better.

Gibbs Sampling

Since joint is in (independent) Dirichlet forms of $\boldsymbol{\Theta}$ and $\boldsymbol{B}$, given $\{z^{(n,m)}\}$, we can marginalize:

$$p(\{z^{(n,m)}\},\{\boldsymbol{w}^{(n,m)}\}) = \int p(\{z^{(n,m)}\},\boldsymbol{\Theta},\boldsymbol{B},\{\boldsymbol{w}^{(n,m)}\})d\boldsymbol{\Theta}d\boldsymbol{B}$$

$$\propto \prod_{h=1}^{H}\left\{\left(\prod_{m=1}^{M}\Gamma\left(\alpha+\sum_{n=1}^{N^{(m)}} z_h^{(n,m)}\right)\right)\left(\frac{\prod_{l=1}^{L}\Gamma\left(\eta+\sum_{m=1}^{M}\sum_{n=1}^{N^{(m)}} w_l^{(n,m)} z_h^{(n,m)}\right)}{\Gamma\left(L\eta+\sum_{m=1}^{M}\sum_{n=1}^{N^{(m)}} z_h^{(n,m)}\right)}\right)\right\}.$$

Now $\{z^{(n,m)}\}$ are mutually dependent, and

$$p(z^{(n,m)}|\{z^{(n',m')}\}_{(n',m')\neq(n,m)},\{\boldsymbol{w}^{(n,m)}\}) = \frac{p(\{z^{(n,m)}\},\{\boldsymbol{w}^{(n,m)}\})}{p(\{z^{(n',m')}\}_{n'\neq n,m'\neq m},\{\boldsymbol{w}^{(n,m)}\})}$$

$$\propto \prod_{h=1}^{H}\left\{\frac{\left(\alpha+\sum_{n'\neq n} z_h^{(n',m)}\right)\left(\eta+\sum_{(n',m')\neq(n,m)} w_{\acute{l}(\boldsymbol{w}^{(n,m)})}^{(n',m')} z_h^{(n',m')}\right)}{\left(L\eta+\sum_{(n',m')\neq(n,m)} z_h^{(n',m')}\right)}\right\}^{z_h^{(n,m)}}.$$

For derivation, see Wikipedia for Dirichlet moments, and use $\Gamma(\tau+1)=\tau\Gamma(\tau)$.

Gibbs Sampling

Collapsed Gibbs Sampling:

In j-th step,

1. For a pair (n, m) chosen,
   draw a sample $z^{(j,n,m)} \sim p\left(z^{(n,m)} | \{z^{(j-1,n',m')}\}_{(n',m')\neq(n,m)}, \{w^{(n,m)}\}\right)$,
   and set $z^{(j,n',m')} = z^{(j-1,n',m')}$ for $(n', m') \neq (n, m)$.

Usually, (n, m) is chosen sequentially, and pick a sample every after all $\{z^{(n,m)}\}$ are updated.

$$p(z^{(n,m)} | \{z^{(n',m')}\}_{(n',m')\neq(n,m)}, \{w^{(n,m)}\}) = \frac{p(\{z^{(n,m)}\}, \{w^{(n,m)}\})}{p(\{z^{(n',m')}\}_{n'\neq n, m'\neq m}, \{w^{(n,m)}\})}$$

$$\propto \prod_{h=1}^{H} \left\{ \frac{\left(\alpha + \sum_{n'\neq n} z_h^{(n',m)}\right)\left(\eta + \sum_{(n',m')\neq(n,m)} w_{l(w^{(n,m)})}^{(n',m')} z_h^{(n',m')}\right)}{\left(L\eta + \sum_{(n',m')\neq(n,m)} z_h^{(n',m')}\right)} \right\}^{z_h^{(n,m)}}.$$

Gibbs Sampling

Collapsed Gibbs Sampling:

In j-th step,

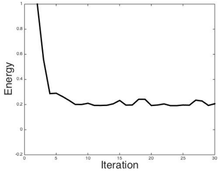
1. For a pair (n, m) chosen,
draw a sample $z^{(j,n,m)} \sim p\left(z^{(n,m)} | \{z^{(j-1,n',m')}\}_{(n',m') \neq (n,m)}, \{w^{(n,m)}\}\right)$,
and set $z^{(j,n',m')} = z^{(j-1,n',m')}$ for $(n', m') \neq (n, m)$.

Usually, (n, m) is chosen sequentially, and pick a sample every after all $\{z^{(n,m)}\}$ are updated.

* The energy in the right figure is a collapsed likelihood

$$p(\{z^{(n,m)}\}, \{w^{(n,m)}\}) = \int p(\{z^{(n,m)}\}, \boldsymbol{\Theta}, \boldsymbol{B}, \{w^{(n,m)}\}) d\boldsymbol{\Theta} d\boldsymbol{B},$$

which is proportional to the posterior $p(\{\widehat{z}^{(n,m)}\}|\{w^{(n,m)}\})$.



Gibbs Sampling

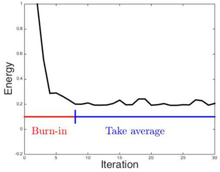
Collapsed Gibbs Sampling:

In j-th step,

1. For a pair (n, m) chosen,
draw a sample $z^{(j,n,m)} \sim p\left(z^{(n,m)} | \{z^{(j-1,n',m')}\}_{(n',m') \neq (n,m)}, \{\boldsymbol{w}^{(n,m)}\}\right)$,
and set $z^{(j,n',m')} = z^{(j-1,n',m')}$ for $(n', m') \neq (n, m)$.

Usually, (n, m) is chosen sequentially, and pick a sample every after all $\{z^{(n,m)}\}$ are updated.

With the estimator $\widehat{z}^{(n,m)} = \frac{1}{J - J_{\text{burn-in}}} \sum_{j=J_{\text{burn-in}}+1}^{J} z^{(j,n,m)}$,



Gibbs Sampling

Collapsed Gibbs Sampling:

In j-th step,

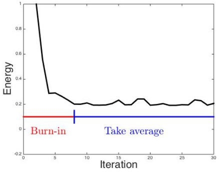
1. For a pair (n, m) chosen,
   draw a sample $z^{(j,n,m)} \sim p\left(z^{(n,m)} | \{z^{(j-1,n',m')}\}_{(n',m') \neq (n,m)}, \{w^{(n,m)}\}\right)$,
   and set $z^{(j,n',m')} = z^{(j-1,n',m')}$ for $(n', m') \neq (n, m)$.

Usually, (n, m) is chosen sequentially, and pick a sample every after all $\{z^{(n,m)}\}$ are updated.

With the estimator $\widehat{z}^{(n,m)} = \frac{1}{J - J_{\text{burn-in}}} \sum_{j=J_{\text{burn-in}}+1}^{J} z^{(j,n,m)}$,

$$p(\widetilde{\theta}_m | \{\widehat{z}^{(n,m)}\}, \{w^{(n,m)}\}) \propto \prod_{h=1}^{H} (\Theta_{m,h})^{\alpha - 1 + \sum_{n=1}^{N^{(m)}} \widehat{z}_h^{(n,m)}}$$

$$p(\beta_h | \{\widehat{z}^{(n,m)}\}, \{w^{(n,m)}\}) \propto \prod_{l=1}^{L} (B_{l,h})^{\eta - 1 + \sum_{m=1}^{M} \sum_{n=1}^{N^{(m)}} w_l^{(n,m)} \widehat{z}_h^{(n,m)}}$$



Gibbs Sampling

Collapsed Gibbs Sampling:

In j-th step,

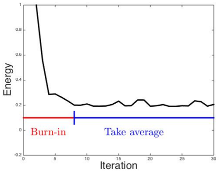
1. For a pair (n, m) chosen,
   draw a sample $z^{(j,n,m)} \sim p\left(z^{(n,m)} | \{z^{(j-1,n',m')}\}_{(n',m') \neq (n,m)}, \{w^{(n,m)}\}\right)$,
   and set $z^{(j,n',m')} = z^{(j-1,n',m')}$ for $(n', m') \neq (n, m)$.

Usually, (n, m) is chosen sequentially, and pick a sample every after all $\{z^{(n,m)}\}$ are updated.

With the estimator $\widehat{z}^{(n,m)} = \frac{1}{J - J_{\text{burn-in}}} \sum_{j=J_{\text{burn-in}}+1}^{J} z^{(j,n,m)}$,

$$p(\widetilde{\theta}_m | \{\widehat{z}^{(n,m)}\}, \{w^{(n,m)}\}) = \mathrm{Dir}\left(\{\alpha + \textstyle\sum_{n=1}^{N^{(m)}} \widehat{z}_h^{(n,m)}\}_{h=1}^{H}\right)$$

$$p(\beta_h | \{\widehat{z}^{(n,m)}\}, \{w^{(n,m)}\}) = \mathrm{Dir}\left(\{\eta + \textstyle\sum_{m=1}^{M} \sum_{n=1}^{N^{(m)}} w_l^{(n,m)} \widehat{z}_h^{(n,m)}\}_{l=1}^{L}\right)$$



More efficient! But lost *independence*... ($\{z^{(n,m)}\}$ are *mutually dependent*).

Gibbs Sampling

Collapsed Gibbs Sampling:

In j-th step,

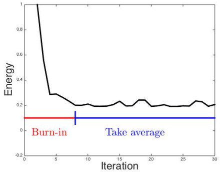
1. For a pair (n,m) chosen, estimate
$\widehat{\boldsymbol{\Theta}}_{\setminus(n,m)} = \langle \boldsymbol{\Theta} \rangle_{p(\boldsymbol{\Theta}|\{\boldsymbol{z}^{(j-1,n',m')}\}_{\setminus(n,m)},\{\boldsymbol{w}^{(n,m)}\})}$, $\quad \widehat{\boldsymbol{B}}_{\setminus(n,m)} = \langle \boldsymbol{B} \rangle_{p(\boldsymbol{B}|\{\boldsymbol{z}^{(j-1,n',m')}\}_{\setminus(n,m)},\{\boldsymbol{w}^{(n,m)}\})}$.
2. Draw a sample $\boldsymbol{z}^{(j,n,m)} \sim p\left(\boldsymbol{z}^{(n,m)} | \widehat{\boldsymbol{\Theta}}_{\setminus(n,m)}, \widehat{\boldsymbol{B}}_{\setminus(n,m)}, \{\boldsymbol{w}^{(n,m)}\}\right)$,
and set $\boldsymbol{z}^{(j,n',m')} = \boldsymbol{z}^{(j-1,n',m')}$ for $(n',m') \neq (n,m)$.

Usually, (n,m) is chosen sequentially, and pick a sample every after all $\{\boldsymbol{z}^{(n,m)}\}$ are updated.

With the estimator $\widehat{\boldsymbol{z}}^{(n,m)} = \frac{1}{J - J_{\text{burn-in}}} \sum_{j=J_{\text{burn-in}}+1}^{J} \boldsymbol{z}^{(j,n,m)}$,

$$p(\widetilde{\boldsymbol{\theta}}_m | \{\widehat{\boldsymbol{z}}^{(n,m)}\}, \{\boldsymbol{w}^{(n,m)}\}) = \mathrm{Dir}\left(\{\alpha + \textstyle\sum_{n=1}^{N^{(m)}} \widehat{z}_h^{(n,m)}\}_{h=1}^{H}\right)$$

$$p(\boldsymbol{\beta}_h | \{\widehat{\boldsymbol{z}}^{(n,m)}\}, \{\boldsymbol{w}^{(n,m)}\}) = \mathrm{Dir}\left(\{\eta + \textstyle\sum_{m=1}^{M} \sum_{n=1}^{N^{(m)}} w_l^{(n,m)} \widehat{z}_h^{(n,m)}\}_{l=1}^{L}\right)$$



Amounts to global parameter update after each element sampling.

Gibbs Sampling

Approximate Collapsed Gibbs Sampling

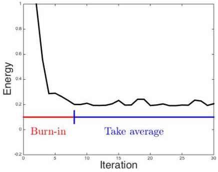
In j-th step,

1. For chosen pairs (n, m), independently draw $z^{(j,n,m)} \sim p(z^{(n,m)} | \widehat{\Theta}^{(j-1)}, \widehat{B}^{(j-1)}, \{w^{(n,m)}\})$,
2. Estimate $\widehat{\Theta}^{(j)} = \langle \Theta \rangle_{p(\Theta | \{z^{(j,n,m)}\}, \{w^{(n,m)}\})}$,
3. Estimate $\widehat{B}^{(j)} = \langle B \rangle_{p(B | \{z^{(j,n,m)}\}, \{w^{(n,m)}\})}$.

With the estimator $\widehat{z}^{(n,m)} = \frac{1}{J - J_{\text{burn-in}}} \sum_{j=J_{\text{burn-in}}+1}^{J} z^{(j,n,m)}$,

$$p(\widetilde{\theta}_m | \{\widehat{z}^{(n,m)}\}, \{w^{(n,m)}\}) = \mathrm{Dir}\left(\{\alpha + \sum_{n=1}^{N^{(m)}} \widehat{z}_h^{(n,m)}\}_{h=1}^{H}\right)$$

$$p(\beta_h | \{\widehat{z}^{(n,m)}\}, \{w^{(n,m)}\}) = \mathrm{Dir}\left(\{\eta + \sum_{m=1}^{M} \sum_{n=1}^{N^{(m)}} w_l^{(n,m)} \widehat{z}_h^{(n,m)}\}_{l=1}^{L}\right)$$



This should work.

Variational Bayesian Learning

Variational Bayesian Approximation:

In j-th step,

1. For chosen pairs (n,m), estimate $\widehat{z}^{(j,n,m)} = \langle z^{(n,m)} \rangle_{p(z^{(n,m)}|\widehat{\Theta}^{(j-1)},\widehat{B}^{(j-1)},\{w^{(n,m)}\})}$,
2. Estimate $\widehat{\Theta}^{(j)} = \langle \Theta \rangle_{p(\Theta|\{z^{(j,n,m)}\},\{w^{(n,m)}\})}$,
3. Estimate $\widehat{B}^{(j)} = \langle B \rangle_{p(B|\{z^{(j,n,m)}\},\{w^{(n,m)}\})}$.

Variational Bayes is similar.

Variational Bayesian Learning

Variational Bayesian Approximation:

In j-th step,

1. For chosen pairs (n, m), estimate $\widehat{z}^{(j,n,m)} = \langle z^{(n,m)} \rangle_{p(z^{(n,m)} | \widehat{\Theta}^{(j-1)}, \widehat{B}^{(j-1)}, \{w^{(n,m)}\})}$,
2. Estimate $\widehat{\Theta}^{(j)} = \langle \Theta \rangle_{p(\Theta | \{z^{(j,n,m)}\}, \{w^{(n,m)}\})}$,
3. Estimate $\widehat{B}^{(j)} = \langle B \rangle_{p(B | \{z^{(j,n,m)}\}, \{w^{(n,m)}\})}$.

For each k (a part of unknowns) in turn,

Gibbs sampling draws a sample from conditional $w_k \sim p\left(w_k | \boldsymbol{w}_{\backslash k}, \mathcal{D}\right)$.

- Slow (each iteration gives one sample from the distribution).
- Accurate (correlation between $\{w_k\}$ is taken into account).

Variational Bayes estimates the mean of conditional $\widehat{w}_k = \langle w_k \rangle_{p(w_k | \widehat{\boldsymbol{w}}_{\backslash k}, \mathcal{D})}$.

- Fast (each iteration estimates whole distribution).
- Inaccurate (correlation between $\{w_k\}$ is neglected).

Variational Bayes is similar.

Properties of MCMC methods

- Converges to the Bayesian posterior.
- (Generally) slower than deterministic methods.
- To get independent samples, we have to subsample from the MCMC sequence.
- Efficient methods are being developed (Hamiltonian Monte Carlo, distributed computation, stochastic gradient).

Appendix: Detailed Balance Property

Transition probability is (all probability below is conditional on $\mathcal{D}$)

$$\begin{aligned} p\left(\boldsymbol{w}^{(j)}|\boldsymbol{w}^{(j-1)}\right) &= T\left(\boldsymbol{w}^{(j)}|\boldsymbol{w}^{(j-1)}\right) \cdot r\left(\boldsymbol{w}^{(j)}|\boldsymbol{w}^{(j-1)}\right) \\ &= \min\left(r\left(\boldsymbol{w}^{(j)}|\boldsymbol{w}^{(j-1)}\right), \frac{p\left(\boldsymbol{w}^{(j)}\right)r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^{(j)}\right)}{p\left(\boldsymbol{w}^{(j-1)}\right)}\right). \end{aligned}$$

Therefore,

$$\begin{aligned} p\left(\boldsymbol{w}^{(j)}|\boldsymbol{w}^{(j-1)}\right) p\left(\boldsymbol{w}^{(j-1)}\right) &= \min\left(p\left(\boldsymbol{w}^{(j-1)}\right) r\left(\boldsymbol{w}^{(j)}|\boldsymbol{w}^{(j-1)}\right), p\left(\boldsymbol{w}^{(j)}\right) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^{(j)}\right)\right) \\ &= \min\left(p\left(\boldsymbol{w}^{(j)}\right) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^{(j)}\right), p\left(\boldsymbol{w}^{(j-1)}\right) r\left(\boldsymbol{w}^{(j)}|\boldsymbol{w}^{(j-1)}\right)\right) \\ &= T\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^{(j)}\right) r\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^{(j)}\right) p\left(\boldsymbol{w}^{(j)}\right) \\ &= p\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^{(j)}\right) p\left(\boldsymbol{w}^{(j)}\right) \qquad \text{Detailed Balance Property} \end{aligned}$$

$$\therefore \int p\left(\boldsymbol{w}^{(j)}|\boldsymbol{w}^{(j-1)}\right) p\left(\boldsymbol{w}^{(j-1)}\right) d\boldsymbol{w}^{(j-1)} = \int p\left(\boldsymbol{w}^{(j-1)}|\boldsymbol{w}^{(j)}\right) p\left(\boldsymbol{w}^{(j)}\right) d\boldsymbol{w}^{(j-1)} = p\left(\boldsymbol{w}^{(j)}\right)$$

Detailed balance $\Rightarrow$ $p\left(\boldsymbol{w}^{(j)}\right)$ is stationary of Markov process. back