# Basics of Bayesian inferences and Bayesian deep learning

-Summary-

Reference : https://www.stat.cmu.edu/~larry/=sml/Bayes.pdf

# Review

- Difference between frequentist and Bayesian approaches for statistical inference:

**Frequentist versus Bayesian Methods**

- In frequentist inference, probabilities are interpreted as long run frequencies. The goal is to create procedures with long run frequency guarantees.
- In Bayesian inference, probabilities are interpreted as subjective degrees of belief. The goal is to state and analyze your beliefs.

- Bayesian inference procedure:
  1. Choose prior distribution for $p(\theta)$ : express our beliefs about a parameter $\theta$ before we see any data.
  2. Choose statistical model $p(x|\theta)$ : reflects our beliefs about $x$ given $\theta$.
  3. After observing data $\mathcal{D}_n = \{X_1, \dots X_n\}$, we update our beliefs and calculate the posterior distributions $p(\theta|\mathcal{D}_n)$.

# Review

- Our goal is to find posterior distribution $p(\theta|\mathcal{D}_n)$ and estimate $E[g(\theta)|\mathcal{D}_n]$ by $\sum_{i=1}^{n} g\left(\theta^{(s)}\right)$
  - **When $\boldsymbol{p(\theta|\mathcal{D}_n)}$ can be computed explicitly** -> Use Monte Carlo (MC) method:

**<MC algorithm>** :

for $j = 1$ to $s$:

1. Sample $\theta^{(j)} \sim p(\theta|\mathcal{D}_n)$

**This algorithm gives independent Samples $\{\boldsymbol{\theta}^{(1)}, \ldots \boldsymbol{\theta}^{(s)}\}$**

For example : $x_i|\mu,\tau \sim N(\mu,\frac{1}{\tau})$, $\mu|\tau, x \sim N\left(\mu_0,\frac{1}{\tau}\right)$, $\tau \sim \text{Gamma}(\alpha,\beta)$:

$$p(\mu|x_1,\ldots x_n) \propto \left(2\beta + (\mu-\mu_0)^2 + \sum_{i=1}^{n}(x_i-\mu)^2\right)^{-\alpha-\frac{(n+1)}{2}}$$

$$p(\tau|x_1,\ldots x_n) = dGamma\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{n}{2(n+1)}(\bar{x}-\mu_0)^2\right)$$

# Review

- **When the posterior distribution $\boldsymbol{p(\theta|\mathcal{D}_n)}$ is intractable** (or cannot given by implicit form):

  (Let's denote $\theta = (\theta_1, ..\theta_k)$, $\theta_{-i} = (\theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$ )

  If the full conditional distribution $p(\theta_i|\theta_{-i}, \mathcal{D}_n)$ is **known** -> Use Gibbs sampling:

**<Gibbs sampling algorithm>** :

Given the current state $\theta^{(s)} = \left(\theta_1^{(s)}, \ldots \theta_k^{(s)}\right)$

( # $\theta^{(0)}$ can be chosen by appropriate estimator of $\theta$)

1. Sample $\theta_1^{(s+1)} \sim p\left(\theta \middle| \theta_{-1}^{(s)}, x_1, \ldots x_n\right)$

...

K. Sample $\theta_k^{(s+1)} \sim p\left(\theta \middle| \theta_{-k}^{(s)}, x_1, \ldots x_n\right)$

K+1. Set $\theta^{(s+1)} = \left(\theta_1^{(s+1)}, \ldots \theta_k^{(s+1)}\right)$

**This algorithm gives dependent Samples $\{\boldsymbol{\theta}^{(1)}, \ldots \boldsymbol{\theta}^{(s)}\}$**

# Review

- **When the posterior distribution $\boldsymbol{p(\theta|\mathcal{D}_n)}$ is intractable** (or cannot given by implicit form):

  If the full conditional distribution $p(\theta_i|\theta_{-i}, \mathcal{D}_n)$ is **unknown**: -> Use Metropolis Algorithm

**<Metropolis algorithm>** :

Given the current state $\theta^{(s)} = \left(\theta_1^{(s)}, \ldots \theta_k^{(s)}\right)$

1. Sample $\theta^* \sim J(\theta|\theta^{(s)})$
2. Compute the acceptance ratio:

$$r = \frac{p(\theta^*|x_1, \ldots, x_n)}{p\left(\theta^{(s)}\middle|x_1, \ldots, x_n\right)} = \frac{p(x_1, \ldots x_n|\theta^*)p(\theta^*)}{p\left(x_1, \ldots x_n\middle|\theta^{(s)}\right)p(\theta^{(s)})}$$

3. Set $\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability} \quad \min(r, 1) \\ \theta^{(s)} & \text{with probability } 1 - \min(r, 1) \end{cases}$

**This algorithm gives dependent Samples $\{\boldsymbol{\theta}^{(1)}, \ldots \boldsymbol{\theta}^{(s)}\}$**

Note: $J(\theta|\theta^{(s)})$ is a symmetric proposal distribution

(ex: $J\left(\theta\middle|\theta^{(s)}\right) = \text{uniform}\left(\theta^{(s)} - \delta, \theta^{(s)} + \delta\right)$ or $N(\theta^{(s)}, \delta^2)$ , where $\delta$ is given)

# Review

- Can we **generalize** the Gibbs sampling and Metropolis algorithm? -> Metropolis-Hasting

**<Metropolis-Hasting algorithm>** :

Given the current state $\theta^{(s)} = \left(\theta_1^{(s)}, \ldots \theta_k^{(s)}\right)$:

for $i = 1$ to $k$ :

1. Update $\theta_i$:

   a. Sample $\theta_i^* \sim J_i\left(\theta_i \middle| \theta^{(s)}\right)$

   b. Compute the acceptance ratio :

$$r = \frac{p\left(\theta_i^*, \theta_{-i}^{(s)}\right)}{p\left(\theta_i^{(s)}, \theta_{-i}^{(s)}\right)} \times \boxed{\frac{J_i\left(\theta_i^{(s)} \middle| \theta_i^*, \theta_{-i}^{(s)}\right)}{J_i\left(\theta_i^* \middle| \theta_i^{(s)}, \theta_{-i}^{(s)}\right)}}$$

   c. Set $\theta_i^{(s+1)} = \begin{cases} \theta_i^* & \text{with probability} \quad \min(r, 1) \\ \theta_i^{(s)} & \text{with probability } 1 - \min(r, 1) \end{cases}$

**Note:**

**1. This terms becomes 1 when $J_i$ is symmetric.**

**(the algorithm regresses to Metropolis algorithm)**

**2. Acceptance ratio $r = 1$ when $J_i$ is full conditional distribution of $\boldsymbol{\theta_i}$.**

**(the algorithm regresses to Gibbs Sampling)**

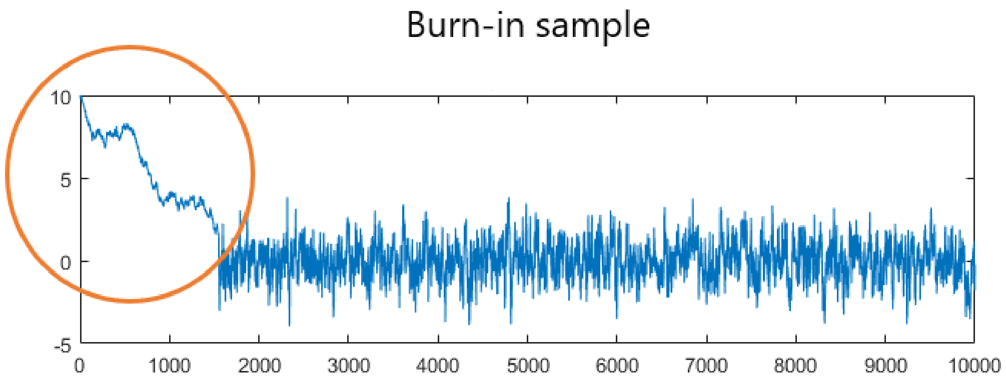
- Here, $J_i(\theta_i | \theta^{(s)})$ does not need to be symmetric (i.e : $J_i(\theta_a | \theta_b) \neq J_i(\theta_b | \theta_a)$)

# Review

- Recall that the Gibbs sampling / Metropolis (-Hasting) algorithms give ‘**dependent**’ samples.
  - In other words, the samples will be autocorrelated within a Markov chain, and we want independent samples. (How to obtain nearly independent samples among them??)

  - Q: Given $s$ samples by MCMC, how many independent samples can be induced (or considered) from these samples? -> Check **Effective Sample Size** (ESS)

  - Q: Under the MCMC process, Are all the samples important as samples? -> No (Use **burn-in**, for example: drop 1500 samples at the initial MCMC process)



# Review

- Recall that the Gibbs sampling / Metropolis (-Hasting) algorithms give '**dependent**' samples.

  - How to select samples to make nearly independent sample set? -> Use **thinning:**
    1. Check autocorrelation of MCMC chain
    2. Pick lag number so that ACF is reasonably low. (here, lag = $40$).
    3. Among attained $s$ samples by MCMC, pick every $40^{th}$ samples to make a nearly independent sample set.