Lecture 2. Computational methods Markov Chain Monte Carlo, Laplace approximation

Why computational methods?

Recall that in our target formula for posterior $p(\theta \mid x) = \frac{p(\theta)p(x \mid \theta)}{\int_{\mathbb{R}} p(\theta)p(X \mid \theta) d\theta}$

where θ are our parameters the **integral** below can get **really nasty**!

BUT: this integral is just a constant! Rewrite $p(\theta|x) = \frac{1}{Z}p(x,\theta)$, where Z is just a normalising constant, although possibly varying over a large range.

What to do?

Monte Carlo integration.

Assume we want to compute
$$E f(\theta | X) = \frac{\int f(\theta)p(\theta)p(X | \theta)d\theta}{\int p(\theta)p(X | \theta)d\theta}$$

where f is some function of parameters θ given the data X.

Monte Carlo integration.

Assume we want to compute
$$E f(\theta | X) = \frac{\int f(\theta)p(\theta)p(X | \theta)d\theta}{\int p(\theta)p(X | \theta)d\theta}$$

where f is some function of parameters θ given the data X.

Monte Carlo integration evaluates this integral by drawing independent samples $\{\theta_t, t = 1, ..., n\}$ from posterior distribution $p(\theta | X)$

Monte Carlo integration.

Assume we want to compute
$$E f(\theta | X) = \frac{\int f(\theta)p(\theta)p(X | \theta)d\theta}{\int p(\theta)p(X | \theta)d\theta}$$

where f is some function of parameters θ given the data X.

Monte Carlo integration evaluates this integral by drawing **independent samples** $\{\theta_t, t=1,...,n\}$ from posterior distribution $p(\theta \mid X)$ and then approximating $E f(\theta \mid X) \approx \frac{1}{n} \sum_{t=1}^{n} f(\theta_t)$

Monte Carlo integration.

Assume we want to compute
$$E f(\theta | X) = \frac{\int f(\theta)p(\theta)p(X | \theta)d\theta}{\int p(\theta)p(X | \theta)d\theta}$$

where f is some function of parameters θ given the data X.

Monte Carlo integration evaluates this integral by drawing **independent samples** $\{\theta_t, t=1,...,n\}$ from posterior distribution $p(\theta \mid X)$ and then approximating $E f(\theta \mid X) \approx \frac{1}{n} \sum_{t=1}^{n} f(\theta_t)$

(law of large numbers)

However:

1. $p(\theta | X)$ can be non-standard, and hence sampling independently from it would not be feasible.

However:

- 1. $p(\theta | X)$ can be non-standard, and hence sampling independently from it would not be feasible.
- 2. Good news: $\{\theta_t\}$ does not necessarily need to be independent. One of the ways of tackling the above problem is to

However:

- 1. $p(\theta | X)$ can be non-standard, and hence sampling independently from it would not be feasible.
- 2. Good news: $\{\theta_t\}$ does not necessarily need to be independent. One of the ways of tackling the above problem is to do it through a Markov chain having $p(\theta|X)$ as its stationary distribution.

This is called Markov chain Monte Carlo.

Markov chain. Suppose we generate a sequence of random variables $\{\theta_0, \theta_1, \dots\}$.

Markov chain. Suppose we generate a sequence of random variables $\{\theta_0, \theta_1, \dots\}$.

Each time $t \ge 0$ the next state θ_{t+1} is sampled from a distribution $P(\theta_{t+1} | \theta_t)$, which depends **only on** the current state of the chain θ_t and does not depend on its history $\{\theta_0, \dots \theta_{t-1}\}$.

Markov chain. Suppose we generate a sequence of random variables $\{\theta_0, \theta_1, \dots\}$.

Each time $t \ge 0$ the next state θ_{t+1} is sampled from a distribution $P(\theta_{t+1} | \theta_t)$, which depends **only on** the current state of the chain θ_t and does not depend on its history $\{\theta_0, \dots \theta_{t-1}\}$.

Subject to certain conditions the chain will gradually "forget" its initial state θ_0 and the distribution $P(\theta_t | \theta_0)$ will not depend on t or θ_0 and converge to a unique stationary distribution

Markov chain. Suppose we generate a sequence of random variables $\{\theta_0, \theta_1, \dots\}$.

Each time $t \ge 0$ the next state θ_{t+1} is sampled from a distribution $P(\theta_{t+1} | \theta_t)$, which depends **only on** the current state of the chain θ_t and does not depend on its history $\{\theta_0, \dots \theta_{t-1}\}$.

Subject to certain conditions the chain will gradually "forget" its initial state θ_0 and the distribution $P(\theta_t | \theta_0)$ will not depend on t or θ_0 and converge to a unique stationary distribution

Hence, after sufficiently long <u>burn-in</u> of m iterations points of $\{\theta_t, t = m + 1, ..., n\}$ will be samples from the stationary distribution and the desired integral can be re-written as

Markov chain. Suppose we generate a sequence of random variables $\{\theta_0, \theta_1, \dots\}$.

Each time $t \ge 0$ the next state θ_{t+1} is sampled from a distribution $P(\theta_{t+1} | \theta_t)$, which depends **only on** the current state of the chain θ_t and does not depend on its history $\{\theta_0, \dots \theta_{t-1}\}$.

Subject to certain conditions the chain will gradually "forget" its initial state θ_0 and the distribution $P(\theta_t | \theta_0)$ will not depend on t or θ_0 and converge to a unique stationary distribution

Hence, after **sufficiently long burn-in** of m iterations points of $\{\theta_t, t = m + 1, ..., n\}$ will be samples from the stationary distribution and the desired integral can be re-written as

$$E f(\theta | X) \approx \frac{1}{n-m} \sum_{t=m+1}^{n} f(\theta_t)$$

Markov chain. Suppose we generate a sequence of random variables $\{\theta_0, \theta_1, \dots\}$.

Each time $t \ge 0$ the next state θ_{t+1} is sampled from a distribution $P(\theta_{t+1} | \theta_t)$, which depends **only on** the current state of the chain θ_t and does not depend on its history $\{\theta_0, \dots \theta_{t-1}\}$.

Subject to certain conditions the chain will gradually "forget" its initial state θ_0 and the distribution $P(\theta_t | \theta_0)$ will not depend on t or θ_0 and converge to a unique stationary distribution

Hence, after **sufficiently long burn-in** of m iterations points of $\{\theta_t, t = m + 1, ..., n\}$ will be samples from the stationary distribution and the desired integral can be re-written as

$$E f(\theta \mid X) \approx \frac{1}{n-m} \sum_{t=m+1}^{n} f(\theta_t)$$

Important: We can construct an MCMC algorithm which will have $p(\theta \mid X)$ as the stationary distribution!

At each time t the next state θ_{t+1} is chosen by first sampling a candidate Y from a **proposal** distribution $q(. | \theta_t)$ which **depends only on the current state** θ_t (or not even that)

At each time t the next state θ_{t+1} is chosen by first sampling a candidate Y from a **proposal** distribution $q(. | \theta_t)$ which **depends only on the current state** θ_t (or not even that)

Candidate Y is then accepted to be the next state of the chain with probability $\alpha(\theta_t, Y)$, where $\alpha(\theta, Y) = \min\left(1, \frac{p(Y)p(X\,|\,Y)q(\theta\,|\,Y)}{p(\theta)p(X\,|\,\theta)q(Y\,|\,\theta)}\right)$.

At each time t the next state θ_{t+1} is chosen by first sampling a candidate Y from a **proposal** distribution $q(\cdot \mid \theta_t)$ which **depends only on the current state** θ_t (or not even that)

Candidate Y is then accepted to be the next state of the chain with probability $\alpha(\theta_t, Y)$, where $\alpha(\theta, Y) = \min\left(1, \frac{p(Y)p(X|Y)q(\theta|Y)}{p(\theta)p(X|\theta)q(Y|\theta)}\right)$.

Now denote $\pi(\theta) = p(\theta | X)$

At each time t the next state θ_{t+1} is chosen by first sampling a candidate Y from a **proposal** distribution $q(\cdot \mid \theta_t)$ which **depends only on the current state** θ_t (or not even that)

Candidate Y is then accepted to be the next state of the chain with probability $\alpha(\theta_t, Y)$,

where
$$\alpha(\theta, Y) = \min\left(1, \frac{p(Y)p(X|Y)q(\theta|Y)}{p(\theta)p(X|\theta)q(Y|\theta)}\right)$$
.

Now denote
$$\pi(\theta) = p(\theta | X) = \frac{p(\theta)p(X | \theta)}{\int p(\theta)p(X | \theta)d\theta}$$

$$P(\theta_{t+1} | \theta_t) = q(\theta_{t+1} | \theta_t) \alpha(\theta_t, \theta_{t+1}) + I(\theta_{t+1} = \theta_t) [1 - \int q(Y | \theta_t) \alpha(\theta_t, Y) dY]$$
(1)

At each time t the next state θ_{t+1} is chosen by first sampling a candidate Y from a **proposal** distribution $q(\cdot \mid \theta_t)$ which **depends only on the current state** θ_t (or not even that)

Candidate Y is then accepted to be the next state of the chain with probability $\alpha(\theta_t, Y)$,

where
$$\alpha(\theta, Y) = \min\left(1, \frac{p(Y)p(X|Y)q(\theta|Y)}{p(\theta)p(X|\theta)q(Y|\theta)}\right)$$
.

Now denote
$$\pi(\theta) = p(\theta | X) = \frac{p(\theta)p(X | \theta)}{\int p(\theta)p(X | \theta)d\theta}$$

$$P(\theta_{t+1} | \theta_t) = q(\theta_{t+1} | \theta_t) \alpha(\theta_t, \theta_{t+1}) + I(\theta_{t+1} = \theta_t) [1 - \int q(Y | \theta_t) \alpha(\theta_t, Y) dY]$$
(1)

acceptance of candidate $Y = \theta_{t+1}$

At each time t the next state θ_{t+1} is chosen by first sampling a candidate Y from a **proposal** distribution $q(\cdot \mid \theta_t)$ which **depends only on the current state** θ_t (or not even that)

Candidate Y is then accepted to be the next state of the chain with probability $\alpha(\theta_t, Y)$,

where
$$\alpha(\theta, Y) = \min\left(1, \frac{p(Y)p(X|Y)q(\theta|Y)}{p(\theta)p(X|\theta)q(Y|\theta)}\right)$$
.

Now denote
$$\pi(\theta) = p(\theta|X) = \frac{p(\theta)p(X|\theta)}{\int p(\theta)p(X|\theta)d\theta}$$

$$P(\theta_{t+1} | \theta_t) = q(\theta_{t+1} | \theta_t) \alpha(\theta_t, \theta_{t+1}) + I(\theta_{t+1} = \theta_t) [1 - \int q(Y | \theta_t) \alpha(\theta_t, Y) dY]$$
(1)

acceptance of candidate $Y = \theta_{t+1}$

rejection of all possible candidates Y

Recall
$$\alpha(\theta, Y) = \min\left(1, \frac{\pi(Y)q(\theta \mid Y)}{\pi(\theta)q(Y \mid \theta)}\right)$$
, and hence

$$\pi(\theta_t)q(\theta_{t+1}|\theta_t)\alpha(\theta_t,\theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t|\theta_{t+1})\alpha(\theta_{t+1},\theta_t)$$
(2)

Recall
$$\alpha(\theta, Y) = \min\left(1, \frac{\pi(Y)q(\theta \mid Y)}{\pi(\theta)q(Y \mid \theta)}\right)$$
, and hence

$$\pi(\theta_t)q(\theta_{t+1}|\theta_t)\alpha(\theta_t,\theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t|\theta_{t+1})\alpha(\theta_{t+1},\theta_t)$$
(2)

Hint: one of the α s in the equality above is equal to 1. Moreover, multiply (1) by $\pi(\theta_t)$

Recall
$$\alpha(\theta, Y) = \min\left(1, \frac{\pi(Y)q(\theta \mid Y)}{\pi(\theta)q(Y \mid \theta)}\right)$$
, and hence

$$\pi(\theta_t)q(\theta_{t+1}|\theta_t)\alpha(\theta_t,\theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t|\theta_{t+1})\alpha(\theta_{t+1},\theta_t)$$
(2)

Hint: one of the α s in the equality above is equal to 1. Moreover, multiply (1) by $\pi(\theta_t)$

$$\pi(\theta_t) P(\theta_{t+1} | \theta_t) = \pi(\theta_t) q(\theta_{t+1} | \theta_t) \alpha(\theta_t, \theta_{t+1}) + \pi(\theta_t) I(\theta_{t+1} = \theta_t) [1 - \int q(Y | \theta_t) \alpha(\theta_t, Y) dY]$$
(3)

Recall
$$\alpha(\theta, Y) = \min\left(1, \frac{\pi(Y)q(\theta \mid Y)}{\pi(\theta)q(Y \mid \theta)}\right)$$
, and hence

$$\pi(\theta_t)q(\theta_{t+1}|\theta_t)\alpha(\theta_t,\theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t|\theta_{t+1})\alpha(\theta_{t+1},\theta_t)$$
(2)

Hint: one of the α s in the equality above is equal to 1. Moreover, multiply (1) by $\pi(\theta_t)$

$$\pi(\theta_t) P(\theta_{t+1} | \theta_t) = \pi(\theta_t) q(\theta_{t+1} | \theta_t) \alpha(\theta_t, \theta_{t+1}) + \pi(\theta_t) I(\theta_{t+1} = \theta_t) [1 - \int q(Y | \theta_t) \alpha(\theta_t, Y) dY]$$
(3)

$$\pi(\theta_{t+1})P(\theta_t \mid \theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t \mid \theta_{t+1})\alpha(\theta_{t+1}, \theta_t) + \pi(\theta_{t+1})I(\theta_{t+1} = \theta_t)[1 - \int q(Y \mid \theta_{t+1})\alpha(\theta_{t+1}, Y)dY]$$
(4)

Recall
$$\alpha(\theta, Y) = \min\left(1, \frac{\pi(Y)q(\theta \mid Y)}{\pi(\theta)q(Y \mid \theta)}\right)$$
, and hence

$$\pi(\theta_t)q(\theta_{t+1}|\theta_t)\alpha(\theta_t,\theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t|\theta_{t+1})\alpha(\theta_{t+1},\theta_t)$$
(2)

Hint: one of the α s in the equality above is equal to 1. Moreover, multiply (1) by $\pi(\theta_t)$

$$\pi(\theta_t)P(\theta_{t+1}|\theta_t) = \pi(\theta_t)q(\theta_{t+1}|\theta_t)\alpha(\theta_t,\theta_{t+1}) + \pi(\theta_t)I(\theta_{t+1} = \theta_t)[1 - \int q(Y|\theta_t)\alpha(\theta_t,Y)dY]$$
(3)

$$\pi(\theta_{t+1})P(\theta_t | \theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t | \theta_{t+1})\alpha(\theta_{t+1}, \theta_t) + \pi(\theta_{t+1})I(\theta_{t+1} = \theta_t)[1 - \int q(Y | \theta_{t+1})\alpha(\theta_{t+1}, Y)dY]$$
(4)

The first terms on the right-hand side of (3) and (4) are equal by (2), and the second ones by equality $\theta_t = \theta_{t+1}$, therefore

Recall
$$\alpha(\theta, Y) = \min\left(1, \frac{\pi(Y)q(\theta \mid Y)}{\pi(\theta)q(Y \mid \theta)}\right)$$
, and hence

$$\pi(\theta_t)q(\theta_{t+1} \mid \theta_t)\alpha(\theta_t, \theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t \mid \theta_{t+1})\alpha(\theta_{t+1}, \theta_t)$$
(2)

Hint: one of the α s in the equality above is equal to 1. Moreover, multiply (1) by $\pi(\theta_t)$

$$\pi(\theta_t)P(\theta_{t+1}|\theta_t) = \pi(\theta_t)q(\theta_{t+1}|\theta_t)\alpha(\theta_t,\theta_{t+1}) + \pi(\theta_t)I(\theta_{t+1} = \theta_t)[1 - \int q(Y|\theta_t)\alpha(\theta_t,Y)dY]$$
(3)

$$\pi(\theta_{t+1})P(\theta_t | \theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t | \theta_{t+1})\alpha(\theta_{t+1}, \theta_t) + \pi(\theta_{t+1})I(\theta_{t+1} = \theta_t)[1 - \int q(Y | \theta_{t+1})\alpha(\theta_{t+1}, Y)dY]$$
(4)

The first terms on the left-hand side of (3) and (4) are equal by (2), and the second ones by equality $\theta_t = \theta_{t+1}$, therefore

 $\pi(\theta_t)P(\theta_{t+1} \mid \theta_t) = \pi(\theta_{t+1})P(\theta_t \mid \theta_{t+1})$. Let us integrate both sides with respect to θ_t

Recall
$$\alpha(\theta, Y) = \min\left(1, \frac{\pi(Y)q(\theta \mid Y)}{\pi(\theta)q(Y \mid \theta)}\right)$$
, and hence

$$\pi(\theta_t)q(\theta_{t+1}|\theta_t)\alpha(\theta_t,\theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t|\theta_{t+1})\alpha(\theta_{t+1},\theta_t)$$
(2)

Hint: one of the α s in the equality above is equal to 1. Moreover, multiply (1) by $\pi(\theta_t)$

$$\pi(\theta_t) P(\theta_{t+1} | \theta_t) = \pi(\theta_t) q(\theta_{t+1} | \theta_t) \alpha(\theta_t, \theta_{t+1}) + \pi(\theta_t) I(\theta_{t+1} = \theta_t) [1 - \int q(Y | \theta_t) \alpha(\theta_t, Y) dY]$$
(3)

$$\pi(\theta_{t+1})P(\theta_t | \theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t | \theta_{t+1})\alpha(\theta_{t+1}, \theta_t) + \pi(\theta_{t+1})I(\theta_{t+1} = \theta_t)[1 - \int q(Y | \theta_{t+1})\alpha(\theta_{t+1}, Y)dY]$$
(4)

The first terms on the left-hand side of (3) and (4) are equal by (2), and the second ones by equality $\theta_t = \theta_{t+1}$, therefore $\pi(\theta_t)P(\theta_{t+1}\,|\,\theta_t) = \pi(\theta_{t+1})P(\theta_t\,|\,\theta_{t+1})$. Let us integrate both sides with respect to θ_t

$$\int \pi(\theta_t) P(\theta_{t+1} \mid \theta_t) d\theta_t = \pi(\theta_{t+1}) \quad \text{Meaning: if } \theta_t \text{ is from the distribution } \pi(.), \text{ then } \theta_{t+1} \text{ will be also.}$$

Recall
$$\alpha(\theta, Y) = \min\left(1, \frac{\pi(Y)q(\theta \mid Y)}{\pi(\theta)q(Y \mid \theta)}\right)$$
, and hence

$$\pi(\theta_t)q(\theta_{t+1} \mid \theta_t)\alpha(\theta_t, \theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t \mid \theta_{t+1})\alpha(\theta_{t+1}, \theta_t)$$
(2)

Hint: one of the α s in the equality above is equal to 1. Moreover, multiply (1) by $\pi(\theta_t)$

$$\pi(\theta_t) P(\theta_{t+1} | \theta_t) = \pi(\theta_t) q(\theta_{t+1} | \theta_t) \alpha(\theta_t, \theta_{t+1}) + \pi(\theta_t) I(\theta_{t+1} = \theta_t) [1 - \int q(Y | \theta_t) \alpha(\theta_t, Y) dY]$$
(3)

$$\pi(\theta_{t+1})P(\theta_t | \theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t | \theta_{t+1})\alpha(\theta_{t+1}, \theta_t) + \pi(\theta_{t+1})I(\theta_{t+1} = \theta_t)[1 - \int q(Y | \theta_{t+1})\alpha(\theta_{t+1}, Y)dY]$$
(4)

The first terms on the left-hand side of (3) and (4) are equal by (2), and the second ones by equality $\theta_t = \theta_{t+1}$, therefore

$$\pi(\theta_t)P(\theta_{t+1} \mid \theta_t) = \pi(\theta_{t+1})P(\theta_t \mid \theta_{t+1})$$
. Let us integrate both sides with respect to θ_t

$$\pi(\theta_t)P(\theta_{t+1}\,|\,\theta_t)d\theta_t=\pi(\theta_{t+1})$$
 Meaning: if θ_t is from the distribution $\pi(\,.\,)$, then θ_{t+1} will be also.

Hence, once sample from stationary has been obtained, all subsequent samples are going to be from it. This means MCMC has converged. The period before convergence is called burn-in

Metropolis-Hastings: how it works in practice

- 1. Start at current position *X*.
- 2. Propose moving to a **new position** Y using proposal q(Y|X)
- 3. Accept/Reject the new position based on the position's adherence to the data and prior distributions using $\alpha(X,Y)$
 - If you accept: Move to the new position Y. Return to Step 1.
 - Else: Do not move to new position, stay at X. Return to Step 1.
- 4. After a large number of iterations, return all accepted positions.

The natural question: what should be the proposal distribution $q(Y | \theta)$?

The natural question: what should be the proposal distribution $q(Y | \theta)$?

1. The rate of convergence to the stationary distribution depends on it! And hence the **compute time**.

The natural question: what should be the proposal distribution $q(Y | \theta)$?

- 1. The rate of convergence to the stationary distribution depends on it! And hence the **compute time**.
- 2. Even if the chain converged it may **mix** slowly (move around the states). And hence one needs to **run it for longer** to obtain **reliable estimates**.

The natural question: what should be the proposal distribution $q(Y | \theta)$?

- 1. The rate of convergence to the stationary distribution depends on it! And hence the **compute time**.
- 2. Even if the chain converged it may **mix** slowly (move around the states). And hence one needs to **run it for longer** to obtain **reliable estimates**.
- 3. Proposal has to **explore the space efficiently**, sometimes it requires to perform experimentation and craftsmanship to construct a good one.

Jupyter notebook 2

Typical proposal distributions

Most typical one: random walk, $q(Y|\theta) = q(|Y - \theta|)$.

Typical proposal distributions

Most typical one: random walk, $q(Y|\theta) = q(|Y - \theta|)$.

Example: $Y \sim N(\theta_t, s)$, where N is a normal distribution and s is the custom standard deviation

Most typical one: random walk, $q(Y|\theta) = q(|Y - \theta|)$.

Example: $Y \sim N(\theta_t, s)$, where N is a normal distribution and s is the custom standard deviation

Important property: acceptance rate - how frequently the proposal gets accepted. Ideally should be 0.2-0.4

Most typical one: random walk, $q(Y|\theta) = q(|Y - \theta|)$.

Example: $Y \sim N(\theta_t, s)$, where N is a normal distribution and s is the custom standard deviation

Important property: acceptance rate - how frequently the proposal gets accepted. Ideally should be 0.2-0.4

This can be tuned during the burn-in period. In general:

Most typical one: random walk, $q(Y|\theta) = q(|Y - \theta|)$.

Example: $Y \sim N(\theta_t, s)$, where N is a normal distribution and s is the custom standard deviation

Important property: acceptance rate - how frequently the proposal gets accepted. Ideally should be 0.2-0.4

This can be tuned during the burn-in period. In general:

1. Acceptance **too high** -> chain mixes slowly. Acceptance **too low** -> chain stops moving.

Most typical one: random walk, $q(Y|\theta) = q(|Y - \theta|)$.

Example: $Y \sim N(\theta_t, s)$, where N is a normal distribution and s is the custom standard deviation

Important property: acceptance rate - how frequently the proposal gets accepted. Ideally should be 0.2-0.4

This can be tuned during the burn-in period. In general:

- 1. Acceptance **too high** -> chain mixes slowly. Acceptance **too low** -> chain stops moving.
- 2. The larger the variance of the proposal is the lower the acceptance rate is.

Most typical one: random walk, $q(Y|\theta) = q(|Y - \theta|)$.

Example: $Y \sim N(\theta_t, s)$, where N is a normal distribution and s is the custom standard deviation

Important property: acceptance rate - how frequently the proposal gets accepted. Ideally should be 0.2-0.4

This can be tuned during the burn-in period. In general:

- 1. Acceptance **too high** -> chain mixes slowly. Acceptance **too low** -> chain stops moving.
- 2. The larger the variance of the proposal is the lower the acceptance rate is.
- 3. This can be used during burn-in to reach the desired acceptance rate.

Instead of updating θ en bloc it is often more convenient and computationally efficient to divide θ into components $\{\theta_1...\theta_h\}$ and update them one by one.

Instead of updating θ en bloc it is often more convenient and computationally efficient to divide θ into components $\{\theta_1...\theta_h\}$ and update them one by one.

This means that instead of $q(Y|\theta)$ we will have $q(Y_i|\theta_{-i},\theta_i)$, where $\theta_{-i} = \{\theta_1...\theta_{i-1},\theta_{i+1}...\theta_h\}$.

Instead of updating θ *en bloc* it is often more convenient and computationally efficient to divide θ into components $\{\theta_1...\theta_h\}$ and update them one by one.

This means that instead of $q(Y|\theta)$ we will have $q(Y_i|\theta_{-i},\theta_i)$, where $\theta_{-i} = \{\theta_1...\theta_{i-1},\theta_{i+1}...\theta_h\}$.

Acceptance probability will then be $\alpha(\theta_{-i}, \theta_i, Y_i) = \min\left(1, \frac{\pi(Y_i \mid \theta_{-i})q(\theta_i \mid Y_i, \theta_{-i})}{\pi(\theta_i \mid \theta_{-i})q(Y_i \mid \theta_i, \theta_{-i})}\right)$

Instead of updating θ en bloc it is often more convenient and computationally efficient to divide θ into components $\{\theta_1...\theta_h\}$ and update them one by one.

This means that instead of $q(Y|\theta)$ we will have $q(Y_i|\theta_{-i},\theta_i)$, where $\theta_{-i} = \{\theta_1...\theta_{i-1},\theta_{i+1}...\theta_h\}$.

Acceptance probability will then be $\alpha(\theta_{-i}, \theta_i, Y_i) = \min \left(1, \frac{\pi(Y_i \mid \theta_{-i}) q(\theta_i \mid Y_i, \theta_{-i})}{\pi(\theta_i \mid \theta_{-i}) q(Y_i \mid \theta_i, \theta_{-i})} \right)$

Gibbs sampler: $q(Y_i | \theta_i, \theta_{-i}) = \pi(Y_i | \theta_{-i})$. Acceptance probability in this case is always equals to 1!

Instead of updating θ en bloc it is often more convenient and computationally efficient to divide θ into components $\{\theta_1...\theta_h\}$ and update them one by one.

This means that instead of $q(Y|\theta)$ we will have $q(Y_i|\theta_{-i},\theta_i)$, where $\theta_{-i} = \{\theta_1...\theta_{i-1},\theta_{i+1}...\theta_h\}$.

Acceptance probability will then be $\alpha(\theta_{-i}, \theta_i, Y_i) = \min \left(1, \frac{\pi(Y_i \mid \theta_{-i}) q(\theta_i \mid Y_i, \theta_{-i})}{\pi(\theta_i \mid \theta_{-i}) q(Y_i \mid \theta_i, \theta_{-i})} \right)$

Gibbs sampler: $q(Y_i | \theta_i, \theta_{-i}) = \pi(Y_i | \theta_{-i})$. Acceptance probability in this case is always equals to 1!

Gibbs sampling uses the property of tractability of all *conditional* posterior distributions to get samples from the unknown *full* posterior distribution of all model variables.

Gibbs sampling scheme

Assume we have data $X \sim p(X | \theta_1, \theta_2)$

- 1. Randomly initialize $\theta_1^{(0)}$ and sample $\theta_2^{(0)} \sim p(\theta_2 \mid X, \theta_1^{(0)})$
- 2. For step t = 1, ..., T
 - (a) Sample $\theta_1^{(t)} \sim p(\theta_1 | X, \theta_2^{t-1})$
 - (b) Sample $\theta_2^{(t)} \sim p(\theta_2 \mid X, \theta_1^t)$

The idea: find parameters μ and Σ such that $p(\theta \mid X) \approx N(\mu, \Sigma)$

The idea: find parameters μ and Σ such that $p(\theta | X) \approx N(\mu, \Sigma)$

Ingredients: Taylor series expansion and Maximum A Posteriori solution (MAP)

The idea: find parameters μ and Σ such that $p(\theta | X) \approx N(\mu, \Sigma)$

Ingredients: Taylor series expansion and Maximum A Posteriori solution (MAP)

$$p(\theta \mid X) = \frac{p(\theta, X)}{p(X)} = \frac{e^{\ln p(\theta, X)}}{\int e^{\ln p(\theta, x)} d\theta}, \text{ concentrate on } \ln p(\theta, X) \text{ as a function of } \theta$$

The idea: find parameters μ and Σ such that $p(\theta | X) \approx N(\mu, \Sigma)$

Ingredients: Taylor series expansion and Maximum A Posteriori solution (MAP)

$$p(\theta \mid X) = \frac{p(\theta, X)}{p(X)} = \frac{e^{\ln p(\theta, X)}}{\int e^{\ln p(\theta, x)} d\theta}, \text{ concentrate on } \ln p(\theta, X) \text{ as a function of } \theta$$

Taylor series up to the 2nd term: $f(\theta) \approx f(\theta_0) + (\theta - \theta_0)^T \nabla f(\theta_0) + \frac{1}{2} (\theta - \theta_0) \nabla^2 f(\theta_0) (\theta - \theta_0)^T \nabla f(\theta_0)$

The idea: find parameters μ and Σ such that $p(\theta | X) \approx N(\mu, \Sigma)$

Ingredients: Taylor series and Maximum A Posteriori solution (MAP)

$$p(\theta \mid X) = \frac{p(\theta, X)}{p(X)} = \frac{e^{\ln p(\theta, X)}}{\int e^{\ln p(\theta, x)} d\theta}, \text{ concentrate on } \ln p(\theta, X) \text{ as a function of } \theta$$

Taylor series up to the 2nd term: $f(\theta) \approx f(\theta_0) + (\theta - \theta_0)^T \nabla f(\theta_0) + \frac{1}{2} (\theta - \theta_0) \nabla^2 f(\theta_0) (\theta - \theta_0)^T \nabla f(\theta_0)$

Even a crocodile is shorter than this expression!

The idea: find parameters μ and Σ such that $p(\theta | X) \approx N(\mu, \Sigma)$

Ingredients: Taylor series and Maximum A Posteriori solution (MAP)

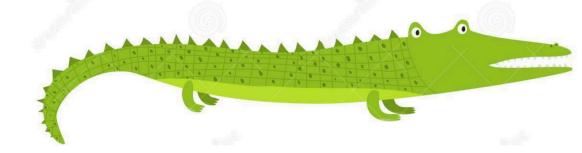
$$p(\theta \mid X) = \frac{p(\theta, X)}{p(X)} = \frac{e^{\ln p(\theta, X)}}{\int e^{\ln p(\theta, x)} d\theta}, \text{ concentrate on } \ln p(\theta, X) \text{ as a function of } \theta$$

Taylor series up to the 2nd term: $f(\theta) \approx f(\theta_0) + (\theta - \theta_0)^T \nabla f(\theta_0) + \frac{1}{2} (\theta - \theta_0) \nabla^2 f(\theta_0) (\theta - \theta_0)^T \nabla f(\theta_0)$

Even a crocodile is shorter than this expression!

Hence finding a good point (MAP):

$$\theta_0 = \theta_{MAP} = \arg \max_{\theta} p(\theta | X)$$



The idea: find parameters μ and Σ such that $p(\theta | X) \approx N(\mu, \Sigma)$

Ingredients: Taylor series and Maximum A Posteriori solution (MAP)

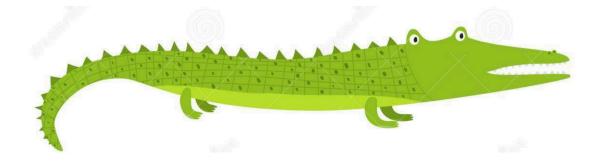
$$p(\theta \mid X) = \frac{p(\theta, X)}{p(X)} = \frac{e^{\ln p(\theta, X)}}{\int e^{\ln p(\theta, x)} d\theta}, \text{ concentrate on } \ln p(\theta, X) \text{ as a function of } \theta$$

Taylor series up to the 2nd term: $f(\theta) \approx f(\theta_0) + (\theta - \theta_0)^T \nabla f(\theta_0) + \frac{1}{2} (\theta - \theta_0) \nabla^2 f(\theta_0) (\theta - \theta_0)^T \nabla f(\theta_0)$

Even a crocodile is shorter than this expression!

Hence let us find a good point (MAP):

$$\theta_0 = \theta_{MAP} = \arg\max_{\theta} p(\theta | X) = \arg\max_{\theta} \frac{p(X, \theta)}{p(X)}$$



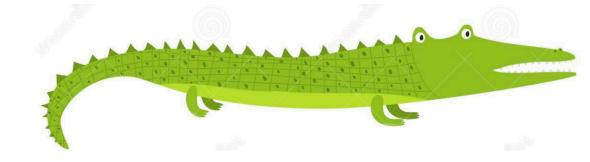
The idea: find parameters μ and Σ such that $p(\theta | X) \approx N(\mu, \Sigma)$

Ingredients: Taylor series and Maximum A Posteriori solution (MAP)

$$p(\theta \mid X) = \frac{p(\theta, X)}{p(X)} = \frac{e^{\ln p(\theta, X)}}{\int e^{\ln p(\theta, x)} d\theta}, \text{ concentrate on } \ln p(\theta, X) \text{ as a function of } \theta$$

Taylor series up to the 2nd term: $f(\theta) \approx f(\theta_0) + (\theta - \theta_0)^T \nabla f(\theta_0) + \frac{1}{2} (\theta - \theta_0) \nabla^2 f(\theta_0) (\theta - \theta_0)^T \nabla f(\theta_0)$

Even a crocodile is shorter than this expression!

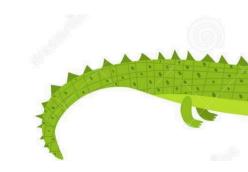


Hence let us find a good point (MAP):

$$\theta_0 = \theta_{MAP} = \arg\max_{\theta} p(\theta | X) = \arg\max_{\theta} \frac{p(X, \theta)}{p(X)} = \arg\max_{\theta} \ln p(X, \theta)$$

Note, that θ_{MAP} corresponds to **local maximum of the posterior**

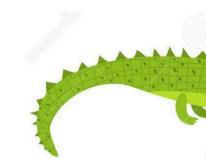
Hence $\nabla f(\theta_{MAP}) = 0$ and the second term of the "crocodile" conveniently gets zeroed down:



Note, that θ_{MAP} corresponds to local maximum of the posterior

Hence $\nabla f(\theta_{MAP}) = 0$ and the second term of the "crocodile" conveniently gets zeroed down:

$$f(\theta) \approx f(\theta_{MAP}) + (\theta - \theta_{MAP})^T \nabla f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T$$



Note, that θ_{MAP} corresponds to local maximum of the posterior

Hence $\nabla f(\theta_{MAP}) = 0$ and the second term of the "crocodile" conveniently gets zeroed down:



$$f(\theta) \approx f(\theta_{MAP}) + (\theta - \theta_{MAP})^T \nabla f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T = f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T$$

Note, that θ_{MAP} corresponds to local maximum of the posterior

Hence $\nabla f(\theta_{MAP}) = 0$ and the second term of the "crocodile" conveniently gets zeroed down:



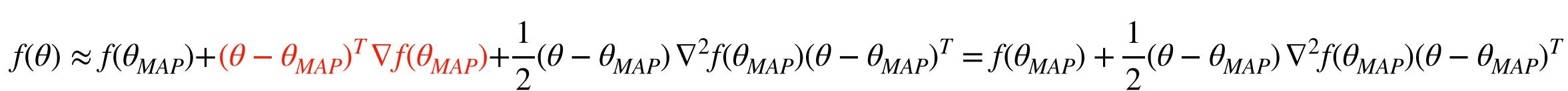
$$f(\theta) \approx f(\theta_{MAP}) + (\theta - \theta_{MAP})^T \nabla f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T = f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T$$

Now posterior, substitute $f(\theta)$ by $\ln p(X, \theta)$:

$$p(\theta | X) = \frac{e^{\ln p(X,\theta)}}{\int e^{\ln p(X,\theta)} d\theta} \approx$$

Note, that θ_{MAP} corresponds to local maximum of the posterior

Hence $\nabla f(\theta_{MAP}) = 0$ and the second term of the "crocodile" conveniently gets zeroed down:



Now posterior, substitute $f(\theta)$ by $\ln p(X, \theta)$:

$$p(\theta \mid X) = \frac{e^{\ln p(X,\theta)}}{\int e^{\ln p(X,\theta)} d\theta} \approx \frac{p(X,\theta_{MAP})e^{\frac{1}{2}(\theta - \theta_{MAP})^T \nabla^2 \ln p(X,\theta_{MAP})(\theta - \theta_{MAP})}}{\int p(X,\theta_{MAP})e^{\frac{1}{2}(\theta - \theta_{MAP})^T \nabla^2 \ln p(X,\theta_{MAP})(\theta - \theta_{MAP})} d\theta}$$



Note, that θ_{MAP} corresponds to local maximum of the posterior

Hence $\nabla f(\theta_{MAP}) = 0$ and the second term of the "crocodile" conveniently gets zeroed down:

$$f(\theta) \approx f(\theta_{MAP}) + (\theta - \theta_{MAP})^T \nabla f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T = f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T$$

Now posterior, substitute $f(\theta)$ by $\ln p(X, \theta)$:

$$p(\theta \mid X) = \frac{e^{\ln p(X,\theta)}}{\int e^{\ln p(X,\theta)} d\theta} \approx \frac{p(X,\theta_{MAP})e^{\frac{1}{2}(\theta - \theta_{MAP})^T \nabla^2 \ln p(X,\theta_{MAP})(\theta - \theta_{MAP})}}{\int p(X,\theta_{MAP})e^{\frac{1}{2}(\theta - \theta_{MAP})^T \nabla^2 \ln p(X,\theta_{MAP})(\theta - \theta_{MAP})} d\theta}$$

Looks like a Normal distribution!



Note, that θ_{MAP} corresponds to local maximum of the posterior

Hence $\nabla f(\theta_{MAP}) = 0$ and the second term of the "crocodile" conveniently gets zeroed down:

$$f(\theta) \approx f(\theta_{MAP}) + (\theta - \theta_{MAP})^T \nabla f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T = f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T$$

Now posterior, substitute $f(\theta)$ by $\ln p(X, \theta)$:

$$p(\theta \mid X) = \frac{e^{\ln p(X,\theta)}}{\int e^{\ln p(X,\theta)} d\theta} \approx \frac{p(X,\theta_{MAP})e^{\frac{1}{2}(\theta - \theta_{MAP})^T \nabla^2 \ln p(X,\theta_{MAP})(\theta - \theta_{MAP})}}{\int p(X,\theta_{MAP})e^{\frac{1}{2}(\theta - \theta_{MAP})^T \nabla^2 \ln p(X,\theta_{MAP})(\theta - \theta_{MAP})} d\theta}$$
 Looks

Looks like a Normal distribution!

(Pdf of the normal distribution $N(\mu, \Sigma)$ is $p(x, \mu, \Sigma) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$



Note, that θ_{MAP} corresponds to local maximum of the posterior

Hence $\nabla f(\theta_{MAP}) = 0$ and the second term of the "crocodile" conveniently gets zeroed down:

$$f(\theta) \approx f(\theta_{MAP}) + (\theta - \theta_{MAP})^T \nabla f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T = f(\theta_{MAP}) + \frac{1}{2} (\theta - \theta_{MAP}) \nabla^2 f(\theta_{MAP}) (\theta - \theta_{MAP})^T$$

Now posterior, substitute $f(\theta)$ by $\ln p(X, \theta)$:

$$p(\theta \mid X) = \frac{e^{\ln p(X,\theta)}}{\int e^{\ln p(X,\theta)} d\theta} \approx \frac{p(X,\theta_{MAP})e^{\frac{1}{2}(\theta - \theta_{MAP})^T \nabla^2 \ln p(X,\theta_{MAP})(\theta - \theta_{MAP})}}{\int p(X,\theta_{MAP})e^{\frac{1}{2}(\theta - \theta_{MAP})^T \nabla^2 \ln p(X,\theta_{MAP})(\theta - \theta_{MAP})} d\theta}$$

Looks like a Normal distribution!

(Pdf of the normal distribution $N(\mu, \Sigma)$ is $p(x, \mu, \Sigma) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$)

Hence
$$\theta | X \sim N(\theta_{MAP}, -(\nabla^2 \ln p(X, \theta_{MAP}))^{-1})$$

1. How to find MAP? Iterative procedure, gradient ascent.

In **pymc** function **find_map** which we already used in the first Jupyter notebook.

2. How to find Hessian $\nabla^2 \ln p(X, \theta)$?:

In pymc function find_hessian

However with the large number of parameters this also becomes too computationally challenging, hence one needs another method

Jupyter notebook 2 Laplace approximation