Markov Chain Monte Carlo algorithms

Modèles probabilistes pour l'apprentissage

Théo Moins[1]

[1]Statify, Inria Grenoble Rhône-Alpes

November 21, 2022

Table of Contents

Reminder: Bayesian paradigm

Data: $\mathbf{x} := x_1, \cdots, x_n$

Reminder: Bayesian paradigm

Data: $\mathbf{x} := x_1, \cdots, x_n$

Model with parameters $\boldsymbol{\theta} := \theta_1, \cdots, \theta_D$: $p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta})$

Reminder: Bayesian paradigm

Data: $\mathbf{x} := x_1, \cdots, x_n$

Model with parameters $\boldsymbol{\theta} := \theta_1, \cdots, \theta_D$: $p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta})$

Prior uncertainty on $\boldsymbol{\theta}$: $p(\boldsymbol{\theta})$

Reminder: Bayesian paradigm

Data: $\mathbf{x} := x_1, \cdots, x_n$

Model with parameters $\boldsymbol{\theta} := \theta_1, \cdots, \theta_D$: $p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta})$

Prior uncertainty on $\boldsymbol{\theta}$: $p(\boldsymbol{\theta})$

$\Longrightarrow$ **Bayesian update:**

$$p(\boldsymbol{\theta} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Reminder: Bayesian paradigm

Data: $\mathbf{x} := x_1, \cdots, x_n$

Model with parameters $\boldsymbol{\theta} := \theta_1, \cdots, \theta_D$: $p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta})$

Prior uncertainty on $\boldsymbol{\theta}$: $p(\boldsymbol{\theta})$

$\implies$ **Bayesian update:**

$$p(\boldsymbol{\theta} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

What's next?

Reminder: Bayesian paradigm

Data: $\mathbf{x} := x_1, \cdots, x_n$

Model with parameters $\boldsymbol{\theta} := \theta_1, \cdots, \theta_D$: $p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta})$

Prior uncertainty on $\boldsymbol{\theta}$: $p(\boldsymbol{\theta})$

$\implies$ **Bayesian update:**

$$p(\boldsymbol{\theta} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

What's next? All computations reduce to posterior means of quantity of interest $f(\boldsymbol{\theta})$:

$$\mathbb{E}_{p(\cdot \mid \mathbf{x})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}$$

Reminder: MCMC

Bayesian inference on $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathbf{x}) \implies$ computation of $\mathbb{E}_{p(\cdot \mid \mathbf{x})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}$,

Reminder: MCMC

$$\text{Bayesian inference on } \boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathbf{x}) \implies \text{computation of } \mathbb{E}_{p(\cdot|\mathbf{x})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta},$$

$$\implies \text{computation of an integral}$$

(highly dimensional, non-explicit, etc.)

Reminder: MCMC

Bayesian inference on $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathbf{x}) \implies$ computation of $\mathbb{E}_{p(\cdot\mid\mathbf{x})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta}$,

$\implies$ computation of an integral

(highly dimensional, non-explicit, etc.)

MCMC (Markov Chain Monte Carlo):

Monte Carlo	Markov Chain
$\mathbb{E}[f(\boldsymbol{\theta})] \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)})$	$\boldsymbol{\theta}^{(i+1)} \mid \boldsymbol{\theta}^{(i)} \sim P(\boldsymbol{\theta}^{(i)}, \cdot)$

Reminder: MCMC

Bayesian inference on $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathbf{x}) \implies$ computation of $\mathbb{E}_{p(\cdot|\mathbf{x})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}$,

$\implies$ computation of an integral

(highly dimensional, non-explicit, etc.)

MCMC (Markov Chain Monte Carlo):

Monte Carlo	Markov Chain
$\mathbb{E}[f(\boldsymbol{\theta})] \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)})$	$\boldsymbol{\theta}^{(i+1)} \mid \boldsymbol{\theta}^{(i)} \sim P(\boldsymbol{\theta}^{(i)}, \cdot)$

With a correct choice of $P(\cdot, \cdot)$ one can often prove that

$$\frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)}) \to \mathbb{E}_{p(\cdot|\mathbf{x})}[f(\boldsymbol{\theta})], \quad \text{when } N \to +\infty.$$

Table of Contents

Metropolis–Hastings

Initialize $\boldsymbol{\theta}^{(0)}$;
for $i \leftarrow 1$ **to** N **do**
  Sample a candidate $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(i-1)})$;
  Compute an acceptation ratio $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)}) \in [0, 1]$;
  Accept the candidate $(\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*)$ with probability $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)})$;
  Otherwise reject $(\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)})$;
end

Algorithm 1: Metropolis-Hastings Algorithm

Illustration: `https://chi-feng.github.io/mcmc-demo/app.html`

Metropolis–Hastings

Acceptation ratio for the general framework:

$$\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)}) = \min\left(1, \frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})} \frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}\right)$$

Metropolis–Hastings

Acceptation ratio for the general framework:

$$\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)}) = \min\left(1, \frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})} \frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}\right)$$

- The ratio $\frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})}$ favours the candidates with the largest posterior density,

Metropolis–Hastings

Acceptation ratio for the general framework:

$$\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)}) = \min\left(1, \frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})}\frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}\right)$$

- The ratio $\frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})}$ favours the candidates with the largest posterior density,
- The ratio $\frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}$ can be seen as a correction term,

Metropolis–Hastings

Acceptation ratio for the general framework:

$$\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)}) = \min\left(1, \frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})} \frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}\right)$$

- The ratio $\frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})}$ favours the candidates with the largest posterior density,
- The ratio $\frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}$ can be seen as a correction term,
- $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)})$ is only an **acceptance probability**!

Metropolis–Hastings

Acceptation ratio for the general framework:

$$\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)}) = \min\left(1, \frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})} \frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}\right)$$

- The ratio $\frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})}$ favours the candidates with the largest posterior density,
- The ratio $\frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}$ can be seen as a correction term,
- $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)})$ is only an **acceptance probability**!

Special cases:

- **Metropolis**: $q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})$ (symmetrical proposal).

Metropolis–Hastings

Acceptation ratio for the general framework:

$$\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)}) = \min\left(1, \frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})} \frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}\right)$$

- The ratio $\frac{p(\boldsymbol{\theta}^*|\mathbf{x})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{x})}$ favours the candidates with the largest posterior density,
- The ratio $\frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}$ can be seen as a correction term,
- $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)})$ is only an **acceptance probability**!

Special cases:

- **Metropolis**: $q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})$ (symmetrical proposal).
- **Gibbs**: $q(\theta_k^*|\boldsymbol{\theta}^{(i)}) = p(\theta_k^*|\theta_1^{(i)}, \cdots, \theta_D^{(i)}, \mathbf{x})$ (full conditional, $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)}) = 1$)

Metropolis–Hastings

```
Initialize θ^(0);
for i ← 1 to N do
    Sample θ* ~ q(θ*|θ^(i−1));
    Sample u ~ U[0, 1];
    if u < α(θ* | θ^(i−1)) then
        θ^(i) = θ*                                  /* Accept */
    else
        θ^(i) = θ^(i−1)                             /* Reject */
    end
end
```

Algorithm 2: Metropolis–Hastings Algorithm

Performance of Random Walk Metropolis

To conclude:

1. **Proposal distribution** $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})$: favour large volumes ...

2. **Acceptance ratio** $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)})$: favour large densities.

$\implies$ The combination makes the selection toward the typical set.

Drawback: tuning is extremely hard in high dimension!

MCMC

To sum up:

MCMC algorithms

- Iterative sampling of the probability density,

MCMC

To sum up:

MCMC algorithms

- Iterative sampling of the probability density,
- Next sample depends on the precedent sample,

MCMC

To sum up:

MCMC algorithms

- Iterative sampling of the probability density,
- Next sample depends on the precedent sample,
- Based on an Acceptation-Rejection Rule,

MCMC

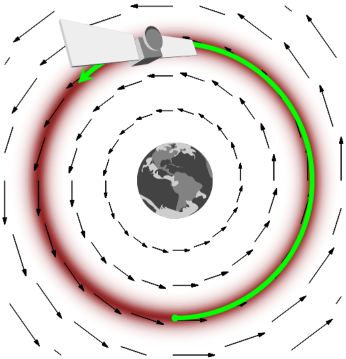
To sum up:

MCMC algorithms

- Iterative sampling of the probability density,
- Next sample depends on the precedent sample,
- Based on an Acceptation-Rejection Rule,
- Examples: Metropolis-Hastings (MH), Gibbs Sampler, Hamiltonian Monte Carlo (HMC) etc.

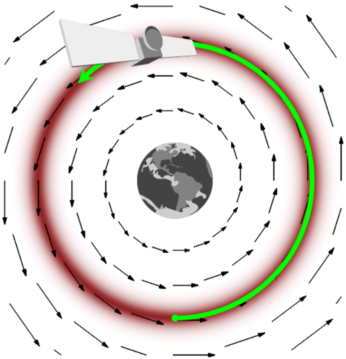
Table of Contents

Introduction to HMC



From Betancourt (2017)

Introduction to HMC

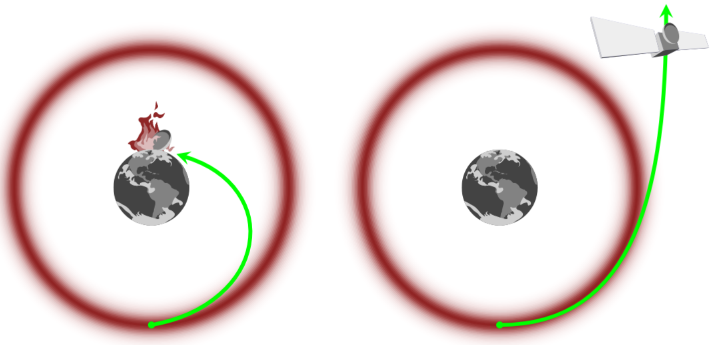


From Betancourt (2017)

- Probabilistic system $\rightarrow$ Physical system
- Mode of the distribution $\rightarrow$ Massive planet
- Gradient of the density $\rightarrow$ Gravitational field

Introduction to HMC

Conservative exploration:



From Betancourt (2017)

Introduction to HMC

For each dimension k, we add a momentum ξ_k to the position θ_k ($2D$ parameters).

Introduction to HMC

For each dimension k, we add a momentum ξ_k to the position θ_k ($2D$ parameters).

$$\implies p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathbf{x}) = p(\boldsymbol{\xi} \mid \boldsymbol{\theta}, \mathbf{x}) p(\boldsymbol{\theta} \mid \mathbf{x})$$

Introduction to HMC

For each dimension k, we add a momentum ξ_k to the position θ_k ($2D$ parameters).

$$\implies p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathbf{x}) = p(\boldsymbol{\xi} \mid \boldsymbol{\theta}, \mathbf{x}) p(\boldsymbol{\theta} \mid \mathbf{x})$$

Hamiltonian definition:

$$\begin{aligned} H(\boldsymbol{\theta}, \boldsymbol{\xi}) &= -\log p(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \mathbf{x}) \\ &= -\log p(\boldsymbol{\xi} \mid \boldsymbol{\theta}, \mathbf{x}) - \log p(\boldsymbol{\theta} \mid \mathbf{x}) \\ &= \qquad K \qquad + \qquad V, \qquad \text{with} \begin{cases} K : & \text{kinetic energy} \\ V : & \text{potential energy} \end{cases} \end{aligned}$$

Introduction to HMC

Hamilton's equation of motion:

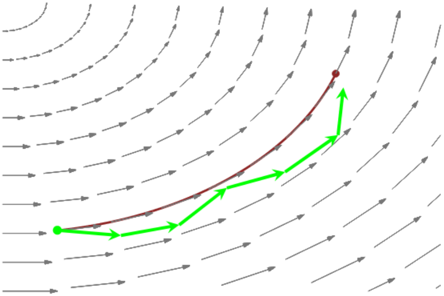
$$\frac{d\theta_k}{dt} = \frac{\partial H}{\partial \xi_k}, \qquad \frac{d\xi_k}{dt} = \frac{\partial H}{\partial \theta_k}. \tag{1}$$

Introduction to HMC

Hamilton's equation of motion:

$$\frac{d\theta_k}{dt} = \frac{\partial H}{\partial \xi_k}, \qquad \frac{d\xi_k}{dt} = \frac{\partial H}{\partial \theta_k}. \tag{1}$$

Solving (1) leads to obtain the position $\phi_t(\boldsymbol{\theta}, \boldsymbol{\xi}) \in \mathbb{R}^{2D}$ accross time t.



From Betancourt (2017)

Introduction to HMC

Initialize $\boldsymbol{\theta}^{(0)}$;
for $i \leftarrow 1$ **to** N **do**
    Sample a momentum $\boldsymbol{\xi}^{(i-1)} \sim q(\boldsymbol{\xi}^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)})$;
    Sample an amount of time $t \sim \mathcal{U}[0, T]$;
    Numerically solve the Hamilton's equation $(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t) \leftarrow \phi_t(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\xi}^{(i-1)})$;
    Accept the candidate $(\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_t)$ with probability $\alpha(\boldsymbol{\theta}_t, -\boldsymbol{\xi}_t \mid \boldsymbol{\theta}^{(i-1)}, \boldsymbol{\xi}^{(i-1)})$;
    Otherwise reject $(\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)})$;
end

Algorithm 3: HMC Algorithm

Introduction to HMC

Initialize $\boldsymbol{\theta}^{(0)}$;

for $i \leftarrow 1$ **to** N **do**

  Sample $\boldsymbol{\xi}^* \sim \mathcal{N}(0, M)$;

  Let $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^{(i-1)}$ and $\boldsymbol{\xi}_0 = \boldsymbol{\xi}^*$;

  for $l \leftarrow 1$ **to** L **do**

    $\boldsymbol{\xi}_{l-1/2} = \boldsymbol{\xi}_{l-1} - \epsilon \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}_0)/2$;

    $\boldsymbol{\theta}_l = \boldsymbol{\theta}_{l-1} + \epsilon \boldsymbol{\xi}_{l-1}$;

    $\boldsymbol{\xi}_l = \boldsymbol{\xi}_{l-1/2} - \epsilon \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}_{Leaps})/2$;

  end

  Sample $u \sim \mathcal{U}[0, 1]$;

  if $u < \alpha(\boldsymbol{\theta}_L, -\boldsymbol{\xi}_L \mid \boldsymbol{\theta}_0, \boldsymbol{\xi}_0)$ **then**

    $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_L$ `/* Accept */`

  else

    $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_0$ `/* Reject */`

  end

Introduction to HMC - Summary

Hamiltonian Monte Carlo

- Algorithm based on Hamiltonian Mechanics
- Time discretization simulated by the Leap Frog Algorithm
- Acceptation rule based on the discretization error in simulating Hamiltonian mechanics
- A lot of hyperparameters!

References:

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Table of Contents

Reminder: MCMC

Bayesian inference on $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathbf{x}) \implies$ computation of $\mathbb{E}_{p(\cdot \mid \mathbf{x})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}$.

Reminder: MCMC

Bayesian inference on $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathbf{x}) \implies$ computation of $\mathbb{E}_{p(\cdot|\mathbf{x})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}$.

MCMC (Markov Chain Monte Carlo):

Monte Carlo	Markov Chain
$\mathbb{E}[f(\boldsymbol{\theta})] \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)})$	$\boldsymbol{\theta}^{(i+1)} \mid \boldsymbol{\theta}^{(i)} \sim P(\boldsymbol{\theta}^{(i)}, \cdot)$

Reminder: MCMC

Bayesian inference on $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathbf{x}) \implies$ computation of $\mathbb{E}_{p(\cdot|\mathbf{x})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}$.

MCMC (Markov Chain Monte Carlo):

Monte Carlo	Markov Chain
$\mathbb{E}[f(\boldsymbol{\theta})] \approx \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)})$	$\boldsymbol{\theta}^{(i+1)} \mid \boldsymbol{\theta}^{(i)} \sim P(\boldsymbol{\theta}^{(i)}, \cdot)$

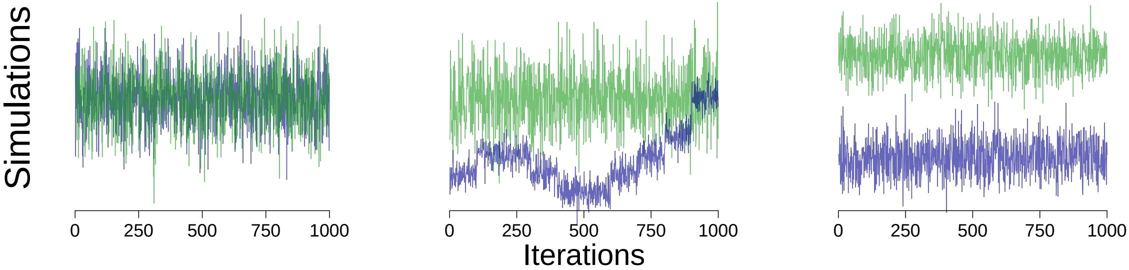
With a correct choice of $P(\cdot, \cdot)$ one can often prove that

$$\frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)}) \to \mathbb{E}_{p(\cdot|\mathbf{x})}[f(\boldsymbol{\theta})], \quad \text{when } N \to +\infty.$$

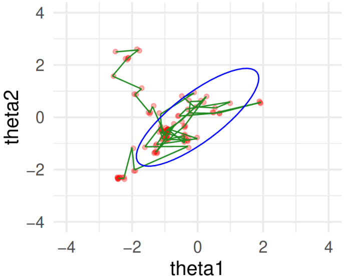
$\hookrightarrow$ How to choose N?

Has the chain(s) converged?

Two kind of convergence issues: mixing and stationarity (univariate case)

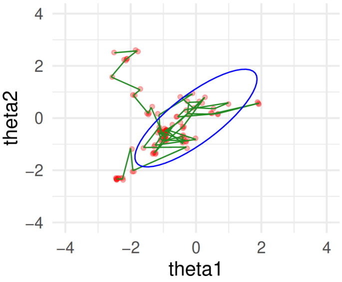


Has the chain(s) converged?



From Aki Vehtari : https://avehtari.github.io/BDA_course_Aalto/

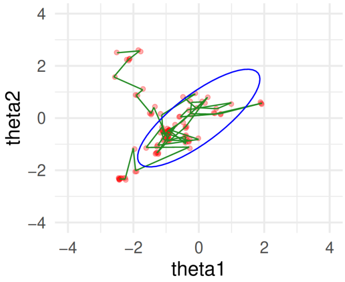
Has the chain(s) converged?



From Aki Vehtari : https://avehtari.github.io/BDA_course_Aalto/

Warm-up: discard the early iterations (usually the first half),

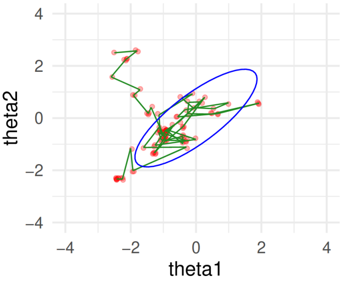
Has the chain(s) converged?



From Aki Vehtari : https://avehtari.github.io/BDA_course_Aalto/

Warm-up: discard the early iterations (usually the first half),
Thinning: keep every kth iterations from each sequence,

Has the chain(s) converged?



From Aki Vehtari : https://avehtari.github.io/BDA_course_Aalto/

Warm-up: discard the early iterations (usually the first half),
Thinning: keep every kth iterations from each sequence,
Multiple chains: reduce the influence of the starting point.

Autocorrelation

The dependency decreases the amount of effective information contained in each element of the chain.

Autocorrelation

The dependency decreases the amount of effective information contained in each element of the chain.

$\hookrightarrow$ Increase of the variance on the target parameters.

Autocorrelation

The dependency decreases the amount of effective information contained in each element of the chain.

$\hookrightarrow$ Increase of the variance on the target parameters.

Autocorrelation function at lag-t: correlation between elements of the sequence distant from t steps.

$$\mathrm{ACF}_t(\theta) = \frac{\mathrm{ACov}_t(\theta)}{\mathrm{Var}(\theta)} = \frac{\frac{1}{S-t}\sum_{s=1}^{S-t}(\theta^{(s)} - \overline{\theta})(\theta^{(s+t)} - \overline{\theta})}{\frac{1}{S-1}\sum_{s=1}^{S}(\theta^{(s)} - \overline{\theta})^2}$$

Autocorrelation

The dependency decreases the amount of effective information contained in each element of the chain.

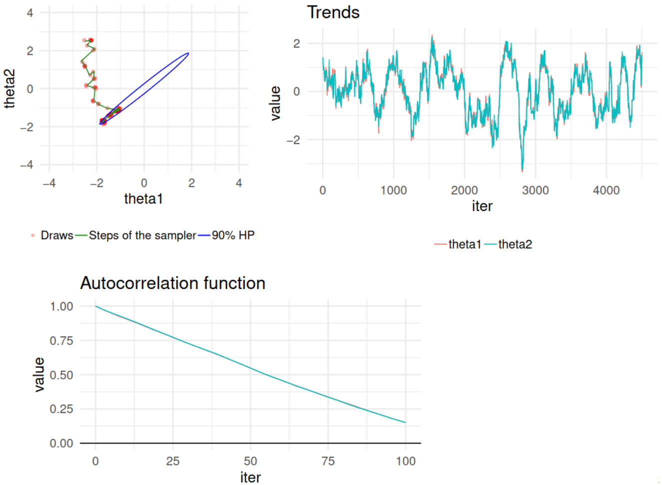
$\hookrightarrow$ Increase of the variance on the target parameters.

Autocorrelation function at lag-t: correlation between elements of the sequence distant from t steps.

$$\mathrm{ACF}_t(\theta) = \frac{\mathrm{ACov}_t(\theta)}{\mathrm{Var}(\theta)} = \frac{\frac{1}{S-t}\sum_{s=1}^{S-t}(\theta^{(s)} - \overline{\theta})(\theta^{(s+t)} - \overline{\theta})}{\frac{1}{S-1}\sum_{s=1}^{S}(\theta^{(s)} - \overline{\theta})^2}$$

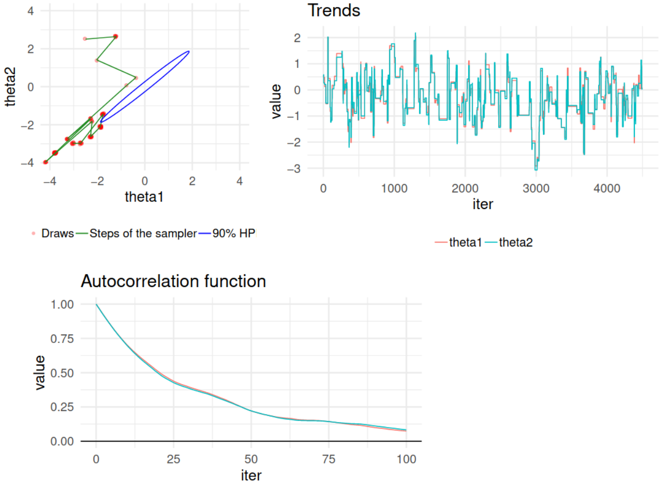
Ideal case: uncorrelated draws $\implies \mathrm{ACF}_t(\theta) = 0 \quad \forall t > 1$

Autocorrelation



From Aki Vehtari : https://avehtari.github.io/BDA_course_Aalto/

Autocorrelation



From Aki Vehtari : `https://avehtari.github.io/BDA_course_Aalto/`

Effective Sample Size (ESS)

In the independent case ($M = 1$ chain):

$$N.\mathrm{Var}(\overline{\theta}) = N.\mathbb{E}\left[(\overline{\theta} - \theta_0)^2\right] \xrightarrow[n \to +\infty]{} \mathrm{Var}(\theta \mid \boldsymbol{x})$$

Effective Sample Size (ESS)

In the independent case ($M = 1$ chain):

$$N.\mathrm{Var}(\overline{\theta}) = N.\mathbb{E}\left[(\overline{\theta} - \theta_0)^2\right] \xrightarrow[n \to +\infty]{} \mathrm{Var}(\theta \mid \boldsymbol{x})$$

Otherwise:

$$N.\mathrm{Var}(\overline{\theta}) \xrightarrow[n \to +\infty]{} \left(1 + 2\sum_{t=1}^{\infty} \mathrm{ACF}_t(\theta)\right) \mathrm{Var}(\theta \mid \boldsymbol{x}),$$

Effective Sample Size (ESS)

In the independent case ($M = 1$ chain):

$$N.\mathrm{Var}(\overline{\theta}) = N.\mathbb{E}\left[(\overline{\theta} - \theta_0)^2\right] \xrightarrow[n \to +\infty]{} \mathrm{Var}(\theta \mid \boldsymbol{x})$$

Otherwise:

$$N.\mathrm{Var}(\overline{\theta}) \xrightarrow[n \to +\infty]{} \left(1 + 2\sum_{t=1}^{\infty} \mathrm{ACF}_t(\theta)\right) \mathrm{Var}(\theta \mid \boldsymbol{x}),$$

ESS: Equivalent number of independent draws

$$\mathrm{ESS} = \frac{N}{1 + 2\sum_{t=1}^{N} \mathrm{ACF}_t(\theta)}$$

$\hookrightarrow$ Number of samples to obtain the same variance in the i.i.d case.

$\hat{R}$ (aka potential scale reduction factor)

Introduced by Gelman and Rubin (1992).
Consider m chains of size n, with $\theta^{(i,j)}$ denoting the ith draw from chain j.
Comparison of the **between-variance** B and the **within-variance** W of the chains:

$$\hat{R} = \sqrt{\frac{\hat{W} + \hat{B}}{\hat{W}}}$$

$\hat{R}$ (aka potential scale reduction factor)

Introduced by Gelman and Rubin (1992).
Consider m chains of size n, with $\theta^{(i,j)}$ denoting the ith draw from chain j.
Comparison of the **between-variance** B and the **within-variance** W of the chains:

$$\hat{R} = \sqrt{\frac{\hat{W} + \hat{B}}{\hat{W}}}$$

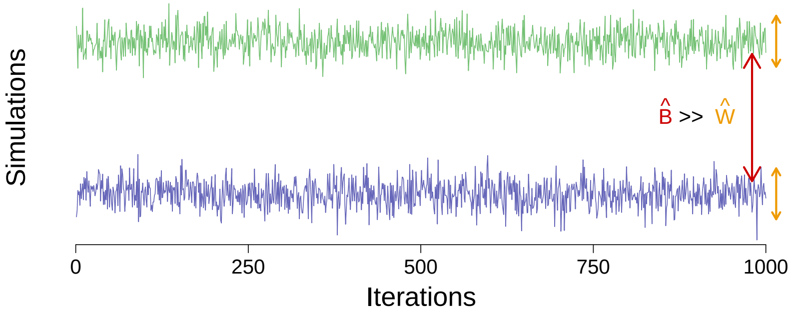
$$\text{Between var}: \hat{B} = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\theta}^{(\cdot,j)} - \bar{\theta}^{(\cdot,\cdot)})^2, \quad \text{where } \bar{\theta}^{(\cdot,j)} = \frac{1}{n}\sum_{i=1}^{n}\theta^{(i,j)}, \quad \bar{\theta}^{(\cdot,\cdot)} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}^{(\cdot,j)},$$

$$\text{Within var}: \hat{W} = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \quad \text{where } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta^{(i,j)} - \bar{\theta}^{(\cdot,j)})^2.$$

$\hat{R}$ (aka potential scale reduction factor)

Introduced by Gelman and Rubin (1992).
Consider m chains of size n, with $\theta^{(i,j)}$ denoting the ith draw from chain j.
Comparison of the **between-variance** B and the **within-variance** W of the chains:

$$\hat{R} = \sqrt{\frac{\hat{W} + \hat{B}}{\hat{W}}}$$



Limitations of $\hat{R}$

The main limitations are:

- It is not robust to certain types of non-convergence.
  $\hat{R}$ and potentially also rank-$\hat{R}$

Limitations of $\hat{R}$

The main limitations are:

- It is not robust to certain types of non-convergence.
  $\hat{R}$ and potentially also rank-$\hat{R}$
- It suffers from a lack of interpretability.
  What is R associated to $\hat{R}$?

Limitations of $\hat{R}$

The main limitations are:

- It is not robust to certain types of non-convergence.
  $\hat{R}$ and potentially also rank-$\hat{R}$
- It suffers from a lack of interpretability.
  What is R associated to $\hat{R}$?
- It must be compared to an arbitrary chosen threshold.
  $\hat{R} \geq 1.1$? 1.01?

Limitations of $\hat{R}$

The main limitations are:

- It is not robust to certain types of non-convergence.
  $\hat{R}$ and potentially also rank-$\hat{R}$
- It suffers from a lack of interpretability.
  What is R associated to $\hat{R}$?
- It must be compared to an arbitrary chosen threshold.
  $\hat{R} \geq 1.1$? 1.01?
- It is associated with a univariate parameter.
  How to manage multiple parameters?

Limitations of $\hat{R}$

The main limitations are:

- It is not robust to certain types of non-convergence.
  $\hat{R}$ and potentially also rank-$\hat{R}$
- It suffers from a lack of interpretability.
  What is R associated to $\hat{R}$?
- It must be compared to an arbitrary chosen threshold.
  $\hat{R} \geq 1.1$? 1.01?
- It is associated with a univariate parameter.
  How to manage multiple parameters?

BDA book recommends to use $\hat{R}$ and ESS in the following way:

$$\hat{R} \in [1, 1.01] \implies \text{“Chains are mixing well”}.$$
$$ESS > 400 \implies \text{“Enough data for estimation”}.$$

Example 1: Bayesian logistic regression

$$\boldsymbol{\beta} \sim \mathcal{N}(0, 0.35^2.\boldsymbol{I}_4), \quad y_j \sim \text{Bernoulli}\left(\frac{1}{1 + e^{-\boldsymbol{x}_j^\top \boldsymbol{\beta}}}\right).$$

Examples

Example 1: Bayesian logistic regression

$$\boldsymbol{\beta} \sim \mathcal{N}(0, 0.35^2.\boldsymbol{I}_4), \quad y_j \sim \text{Bernoulli}\left(\frac{1}{1+e^{-\boldsymbol{x}_j^\top \boldsymbol{\beta}}}\right).$$

Example 2: Hierarchical model (8 Schools)

$\hookrightarrow$ Test the effectiveness of coaching courses.

y_j: coaching effect for the school j

$$\mu \sim \mathcal{N}(0,5), \quad \tau \sim \mathcal{N}(0,10),$$
$$\theta_j \sim \mathcal{N}(\mu,\tau), \quad y_j \sim \mathcal{N}(\theta_j, \sigma_j^2)$$

Model "in between" the separate model and the joint model.

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

From BDA book