

Markov Chain Monte Carlo and variational inference

Chapter 0

Randal Douc and Sholom Schechtman

Télécom SudParis, Institut Polytechnique de Paris
randal.douc@telecom-sudparis.eu



Outline

- ① Introduction
- ② Variational Inference
- ③ MCMC Methods

Outline

- 1 Introduction
- 2 Variational Inference
- 3 MCMC Methods

Objective

Approximate π sur $(X, \mathcal{X})$ by drawing random variables $X_1, \dots, X_n$ “approximately” distributed according to π .

- In Bayesian inference :
 - Observations $Y_{1:m}$ to be fit with a family of densities $p_\beta(y_{1:m})$ with a priori distribution $\lambda(d\beta) = \lambda(\beta)\gamma(d\beta)$.
 - Posterior distribution is the target distribution :

$$\pi(\beta) = \frac{\lambda(\beta)p_\beta(Y_{1:m})}{\int \lambda(d\beta')p_{\beta'}(Y_{1:m})}, \triangleright \text{law of } \beta \text{ knowing } Y_1, \dots, Y_n .$$

- In what follows, $\pi(x) = C\pi_u(x)$, where C is unknown.

Objective

Approximate π sur $(X, \mathcal{X})$ by drawing random variables $X_1, \dots, X_n$ “approximately” distributed according to π .

- In Bayesian inference :
 - Observations $Y_{1:m}$ to be fit with a family of densities $p_\beta(y_{1:m})$ with a priori distribution $\lambda(d\beta) = \lambda(\beta)\gamma(d\beta)$.
 - Posterior distribution is the target distribution :

$$\pi(\beta) = \frac{\lambda(\beta)p_\beta(Y_{1:m})}{\int \lambda(d\beta')p_{\beta'}(Y_{1:m})}, \triangleright \text{law of } \beta \text{ knowing } Y_1, \dots, Y_n.$$

- In what follows, $\pi(x) = C\pi_u(x)$, where C is unknown.

Objective

Approximate π sur $(X, \mathcal{X})$ by drawing random variables $X_1, \dots, X_n$ “approximately” distributed according to π .

- In Bayesian inference :
 - Observations $Y_{1:m}$ to be fit with a family of densities $p_\beta(y_{1:m})$ with a priori distribution $\lambda(d\beta) = \lambda(\beta)\gamma(d\beta)$.
 - Posterior distribution is the **target distribution** :

$$\pi(\beta) = \frac{\lambda(\beta)p_\beta(Y_{1:m})}{\int \lambda(d\beta')p_{\beta'}(Y_{1:m})}, \blacktriangleright \text{law of } \beta \text{ knowing } Y_1, \dots, Y_n .$$

- In what follows, $\pi(x) = C\pi_u(x)$, where C is unknown.

Objective

Approximate π sur $(X, \mathcal{X})$ by drawing random variables $X_1, \dots, X_n$ “approximately” distributed according to π .

- In Bayesian inference :
 - Observations $Y_{1:m}$ to be fit with a family of densities $p_\beta(y_{1:m})$ with a priori distribution $\lambda(d\beta) = \lambda(\beta)\gamma(d\beta)$.
 - Posterior distribution is the **target distribution** :

$$\pi(\beta) = \frac{\lambda(\beta)p_\beta(Y_{1:m})}{\int \lambda(d\beta')p_{\beta'}(Y_{1:m})}, \blacktriangleright \text{law of } \beta \text{ knowing } Y_1, \dots, Y_n .$$

- In what follows, $\pi(x) = C\pi_u(x)$, where C is unknown.

Outline

- 1 Introduction
- 2 Variational Inference
- 3 MCMC Methods

Principle of VI (Variational Inference)

- Approximate π par $(p_\theta)_{\theta \in \Theta}$ (this family is often diverse and each distribution can be drawn).
- Choose the best candidate using a **divergence** (ex: Kullback-Leibler, α -divergence).
- **Pros** : Quick, efficient for large dimension spaces.
- **Cons** : Statistical properties not well known and expressivity of the family of distributions.

Principle of VI (Variational Inference)

- Approximate π par $(p_\theta)_{\theta \in \Theta}$ (this family is often diverse and each distribution can be drawn).
 - Choose the best candidate using a **divergence** (ex: Kullback-Leibler, α -divergence).
-
- **Pros :** Quick, efficient for large dimension spaces.
 - **Cons :** Statistical properties not well known and expressivity of the family of distributions.

Principle of VI (Variational Inference)

- Approximate π par $(p_\theta)_{\theta \in \Theta}$ (this family is often diverse and each distribution can be drawn).
- Choose the best candidate using a **divergence** (ex: Kullback-Leibler, α -divergence).
- **Pros :** Quick, efficient for large dimension spaces.
- **Cons :** Statistical properties not well known and expressivity of the family of distributions.

Principle of VI (Variational Inference)

- Approximate π par $(p_\theta)_{\theta \in \Theta}$ (this family is often diverse and each distribution can be drawn).
- Choose the best candidate using a **divergence** (ex: Kullback-Leibler, α -divergence).
- **Pros :** Quick, efficient for large dimension spaces.
- **Cons :** Statistical properties not well known and expressivity of the family of distributions.

α -divergence

Definition

Define $D_\alpha(\mathbb{P} \parallel \mathbb{Q}) = \int_{\mathcal{X}} f_\alpha \left(\frac{q(x)}{p(x)} \right) p(x) \lambda(dx)$, where

- ❶ $f_0(u) = -\log(u)$,
- ❷ $f_1(u) = u \log(u)$,
- ❸ $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}(u^\alpha - 1)$ pour $\alpha \notin \{0, 1\}$.

Lemma

$D_\alpha(\mathbb{P} \parallel \mathbb{Q}) \geq 0$, with equality if $\mathbb{P} = \mathbb{Q}$.

In VI, the Optimisation problem consists in solving

$$\arg \min_{\theta \in \Theta} D_\alpha(\pi \parallel p_\theta) = \arg \min_{\theta \in \Theta} \int_{\mathcal{X}} f_\alpha \left(\frac{p_\theta(x)}{\pi(x)} \right) \pi(x) \lambda(dx)$$

α -divergence

Definition

Define $D_\alpha(\mathbb{P} \parallel \mathbb{Q}) = \int_{\mathcal{X}} f_\alpha \left(\frac{q(x)}{p(x)} \right) p(x) \lambda(dx)$, where

- ❶ $f_0(u) = -\log(u)$,
- ❷ $f_1(u) = u \log(u)$,
- ❸ $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}(u^\alpha - 1)$ pour $\alpha \notin \{0, 1\}$.

Lemma

$D_\alpha(\mathbb{P} \parallel \mathbb{Q}) \geq 0$, with equality if $\mathbb{P} = \mathbb{Q}$.

In VI, the Optimisation problem consists in solving

$$\arg \min_{\theta \in \Theta} D_\alpha(\pi \parallel p_\theta) = \arg \min_{\theta \in \Theta} \int_{\mathcal{X}} f_\alpha \left(\frac{p_\theta(x)}{\pi(x)} \right) \pi(x) \lambda(dx)$$

α -divergence

Definition

Define $D_\alpha(\mathbb{P} \parallel \mathbb{Q}) = \int_{\mathcal{X}} f_\alpha \left(\frac{q(x)}{p(x)} \right) p(x) \lambda(dx)$, where

- ❶ $f_0(u) = -\log(u)$,
- ❷ $f_1(u) = u \log(u)$,
- ❸ $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}(u^\alpha - 1)$ pour $\alpha \notin \{0, 1\}$.

Lemma

$D_\alpha(\mathbb{P} \parallel \mathbb{Q}) \geq 0$, with equality if $\mathbb{P} = \mathbb{Q}$.

In VI, the Optimisation problem consists in solving

$$\arg \min_{\theta \in \Theta} D_\alpha(\pi \parallel p_\theta) = \arg \min_{\theta \in \Theta} \int_{\mathcal{X}} f_\alpha \left(\frac{p_\theta(x)}{\pi(x)} \right) \pi(x) \lambda(dx)$$

THE ELBO (Evidence Lower Bound)

An important case

We consider $\alpha = 1$. Write $\pi(x) = \tilde{\pi}(x)/C$ where

- $\tilde{\pi}(x)$ is explicit
- $C = \int_{\mathcal{X}} \tilde{\pi}(x) \lambda(dx)$ is an "unexplicit" constant

Then,

$$\begin{aligned} D_1(\pi \| p_\theta) &= \int_{\mathcal{X}} \log \left(\frac{p_\theta(x)}{\pi(x)} \right) p_\theta(x) \lambda(dx) \\ &= - \underbrace{\int_{\mathcal{X}} \log \left(\frac{\tilde{\pi}(x)}{p_\theta(x)} \right) p_\theta(x) \lambda(dx)}_{ELBO} + \log C \geq 0 \end{aligned}$$

key facts

- Maximising ELBO $\Leftrightarrow$ Minimising $D_1(\pi \| p_\theta)$.
- ELBO : $\int_{\mathcal{X}} \log \left(\frac{\tilde{\pi}(x)}{p_\theta(x)} \right) p_\theta(x) \lambda(dx) \leq \log C = \log \int_{\mathcal{X}} \tilde{\pi}(x) \lambda(dx)$.

THE ELBO (Evidence Lower Bound)

An important case

We consider $\alpha = 1$. Write $\pi(x) = \tilde{\pi}(x)/C$ where

- $\tilde{\pi}(x)$ is explicit
- $C = \int_{\mathcal{X}} \tilde{\pi}(x) \lambda(dx)$ is an "unexplicit" constant

Then,

$$\begin{aligned} D_1(\pi \| p_\theta) &= \int_{\mathcal{X}} \log \left(\frac{p_\theta(x)}{\pi(x)} \right) p_\theta(x) \lambda(dx) \\ &= - \underbrace{\int_{\mathcal{X}} \log \left(\frac{\tilde{\pi}(x)}{p_\theta(x)} \right) p_\theta(x) \lambda(dx)}_{ELBO} + \log C \geq 0 \end{aligned}$$

key facts

- Maximising ELBO $\Leftrightarrow$ Minimising $D_1(\pi \| p_\theta)$.
- ELBO : $\int_{\mathcal{X}} \log \left(\frac{\tilde{\pi}(x)}{p_\theta(x)} \right) p_\theta(x) \lambda(dx) \leq \log C = \log \int_{\mathcal{X}} \tilde{\pi}(x) \lambda(dx)$.

THE ELBO (Evidence Lower Bound)

An important case

We consider $\alpha = 1$. Write $\pi(x) = \tilde{\pi}(x)/C$ where

- $\tilde{\pi}(x)$ is explicit
- $C = \int_{\mathcal{X}} \tilde{\pi}(x) \lambda(dx)$ is an "unexplicit" constant

Then,

$$\begin{aligned} D_1(\pi \| p_\theta) &= \int_{\mathcal{X}} \log \left(\frac{p_\theta(x)}{\pi(x)} \right) p_\theta(x) \lambda(dx) \\ &= - \underbrace{\int_{\mathcal{X}} \log \left(\frac{\tilde{\pi}(x)}{p_\theta(x)} \right) p_\theta(x) \lambda(dx)}_{ELBO} + \log C \geq 0 \end{aligned}$$

key facts

- Maximising ELBO $\Leftrightarrow$ Minimising $D_1(\pi \| p_\theta)$.
- ELBO : $\int_{\mathcal{X}} \log \left(\frac{\tilde{\pi}(x)}{p_\theta(x)} \right) p_\theta(x) \lambda(dx) \leq \log C = \log \int_{\mathcal{X}} \tilde{\pi}(x) \lambda(dx)$.

Outline

- ① Introduction
- ② Variational Inference
- ③ MCMC Methods**

MCMC and VI

Principle of MCMC (Markov chain by Monte Carlo)

- Construct a Markov chain $(X_n)_{n \in \mathbb{N}}$ with π as stationary distribution.
- Convergence : the law of X_n converges to π when $n \rightarrow \infty$.
- Algorithms : Metropolis-Hastings, Gibbs sampling, etc.

Comparison VI vs MCMC

Critère	VI	MCMC
Precision	Biased approximation	Asymptotically exact
Speed	Quick	Slow
Dimension	high dimension	degrades with high dimensions

MCMC and VI

Principle of MCMC (Markov chain by Monte Carlo)

- Construct a Markov chain $(X_n)_{n \in \mathbb{N}}$ with π as stationary distribution.
- Convergence : the law of X_n converges to π when $n \rightarrow \infty$.
- Algorithms : Metropolis-Hastings, Gibbs sampling, etc.

Comparison VI vs MCMC

Critère	VI	MCMC
Precision	Biased approximation	Asymptotically exact
Speed	Quick	Slow
Dimension	high dimension	degrades with high dimensions