Lectures on Variational Inference

1) Approximate Bayesian Inference in Machine Learning

Pierre Alquier

RIKEN AIP
Center for Advanced Intelligence Project

Heilbronn Institute – University of Bristol – Nov. 27, 2019

Statistical learning problem

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$.

Statistical learning problem

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$.
2. We have a loss function

$$\ell : \Theta \times \mathcal{X} \to \mathbb{R}_+.$$

Statistical learning problem

1. We observe $X_1, \ldots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$.
2. We have a loss function

$$\ell : \Theta \times \mathcal{X} \to \mathbb{R}_+.$$

3. Objective : base on the sample, learn $\theta_0 \in \Theta$ which minimizes the risk

$$R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)].$$

Example 1 : supervised classification

1. We observe $X_1 = (Z_1, Y_1), \dots, X_n = (Z_n, Y_n)$ i.i.d from P^0 unknown in $\mathbb{R}^d \times \{0, 1\}$.

Example 1 : supervised classification

1. We observe $X_1 = (Z_1, Y_1), \dots, X_n = (Z_n, Y_n)$ i.i.d from P^0 unknown in $\mathbb{R}^d \times \{0,1\}$.
2. Consider a set of predictors $(f_\theta, \theta \in \Theta)$ with $f_\theta : \mathbb{R}^d \to \{0,1\}$ and

$$\ell(\theta, (z,y)) = \mathbf{1}_{y \neq f_\theta(z)} = \begin{cases} 0 \text{ if } y = f_\theta(z), \\ 1 \text{ if } y \neq f_\theta(z). \end{cases}$$

Example 1 : supervised classification

1. We observe $X_1 = (Z_1, Y_1), \ldots, X_n = (Z_n, Y_n)$ i.i.d from P^0 unknown in $\mathbb{R}^d \times \{0, 1\}$.
2. Consider a set of predictors $(f_\theta, \theta \in \Theta)$ with $f_\theta : \mathbb{R}^d \to \{0, 1\}$ and

$$\ell(\theta, (z, y)) = \mathbf{1}_{y \neq f_\theta(z)} = \begin{cases} 0 \text{ if } y = f_\theta(z), \\ 1 \text{ if } y \neq f_\theta(z). \end{cases}$$

3. Objective : learn $\theta_0 \in \Theta$ which minimizes the classification error

$$R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)] = \mathbb{P}_{(Z,Y) \sim P^0}[Y \neq f_\theta(Z)].$$

Example 2 : parametric estimation

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$, with p.d.f p^0.

Example 2 : parametric estimation

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$, with p.d.f p^0.
2. Consider a parametric family of probability distributions : $(P_\theta, \theta \in \Theta)$ with p.d.f p_θ and

$$\ell(x, \theta) = -\log p_\theta(x) = \log \left(\frac{1}{p_\theta(x)} \right).$$

Example 2 : parametric estimation

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$, with p.d.f p^0.
2. Consider a parametric family of probability distributions : $(P_\theta, \theta \in \Theta)$ with p.d.f p_θ and

$$\ell(x, \theta) = -\log p_\theta(x) = \log\left(\frac{1}{p_\theta(x)}\right).$$

3. Objective : learn $\theta_0 \in \Theta$ which minimizes

$$R(\theta) = \mathbb{E}_{X \sim P^0}\left[\log\left(\frac{1}{p_\theta(X)}\right)\right]$$

Example 2 : parametric estimation

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$, with p.d.f p^0.
2. Consider a parametric family of probability distributions : $(P_\theta, \theta \in \Theta)$ with p.d.f p_θ and

$$\ell(x,\theta) = -\log p_\theta(x) = \log\left(\frac{1}{p_\theta(x)}\right).$$

3. Objective : learn $\theta_0 \in \Theta$ which minimizes

$$R(\theta) = \mathbb{E}_{X\sim P^0}\left[\log\left(\frac{p^0(X)}{p_\theta(X)}\right)\right] - \mathbb{E}_{X\sim P^0}\left[\log\left(p^0(X)\right)\right]$$

Example 2 : parametric estimation

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$, with p.d.f p^0.
2. Consider a parametric family of probability distributions : $(P_\theta, \theta \in \Theta)$ with p.d.f p_θ and

$$\ell(x, \theta) = -\log p_\theta(x) = \log\left(\frac{1}{p_\theta(x)}\right).$$

3. Objective : learn $\theta_0 \in \Theta$ which minimizes the Kullback divergence

$$R(\theta) = \mathcal{K}(P_0, P_\theta) - \text{ constant.}$$

Example 2 : MLE and Bayesian inference

$$R(\theta) = -\mathbb{E}_{X \sim P^0} \left[\log \left(p_\theta(X) \right) \right].$$

Example 2 : MLE and Bayesian inference

$$R(\theta) = -\mathbb{E}_{X\sim P^0}\left[\log\left(p_\theta(X)\right)\right].$$

Estimator of $R(\theta)$: $R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p_\theta(X_i)\right) = -\frac{\log(L_n(\theta))}{n}$.

Example 2 : MLE and Bayesian inference

$$R(\theta) = -\mathbb{E}_{X\sim P^0}\left[\log\left(p_\theta(X)\right)\right].$$

Estimator of $R(\theta)$: $R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p_\theta(X_i)\right) = -\frac{\log(L_n(\theta))}{n}.$

MLE	Bayesian inference

Example 2 : MLE and Bayesian inference

$$R(\theta) = -\mathbb{E}_{X\sim P^0}\left[\log\left(p_\theta(X)\right)\right].$$

$$\text{Estimator of } R(\theta) : R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p_\theta(X_i)\right) = -\frac{\log(L_n(\theta))}{n}.$$

MLE	Bayesian inference
$\hat{\theta} = \arg\min_{\theta\in\Theta} R_n(\theta)$	

Example 2 : MLE and Bayesian inference

$$R(\theta) = -\mathbb{E}_{X\sim P^0}\left[\log\left(p_\theta(X)\right)\right].$$

Estimator of $R(\theta)$: $R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p_\theta(X_i)\right) = -\frac{\log(L_n(\theta))}{n}$.

MLE	Bayesian inference
$\hat{\theta} = \arg\min_{\theta\in\Theta} R_n(\theta)$	prior $\pi(\theta)$ $\pi(\theta\|X_1,\dots,X_n) = \frac{L_n(\theta)\pi(\theta)}{\int L_n(\theta)\pi(\theta)\mathrm{d}\theta}$

Example 2 : MLE and Bayesian inference

$$R(\theta) = -\mathbb{E}_{X\sim P^0}\left[\log\left(p_\theta(X)\right)\right].$$

Estimator of $R(\theta)$: $R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p_\theta(X_i)\right) = -\frac{\log(L_n(\theta))}{n}$.

MLE	Bayesian inference
$\hat{\theta} = \arg\min_{\theta\in\Theta} R_n(\theta)$	prior $\pi(\theta)$ $\pi(\theta\|X_1,\dots,X_n) = \frac{L_n(\theta)\pi(\theta)}{\int L_n(\theta)\pi(\theta)\mathrm{d}\theta}$

The posterior

$$\pi(\theta|X_1,\dots,X_n) \propto \exp[-nR_n(\theta)]\pi(\theta).$$

General solution : Gibbs posterior

For the general machine learning problem :

1. $X_1, \dots, X_n$ i.i.d from P^0,
2. $R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)]$.

General solution : Gibbs posterior

For the general machine learning problem :

1. $X_1, \dots, X_n$ i.i.d from P^0,
2. $R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)]$.

Define

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i).$$

General solution : Gibbs posterior

For the general machine learning problem :

1. $X_1, \ldots, X_n$ i.i.d from P^0,
2. $R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)]$.

Define

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i).$$

Gibbs posterior, EWA, ...

$$\pi_{n,\alpha}(\theta) \propto \exp[-\alpha n R_n(\theta)] \pi(\theta).$$

Back to Example 2 : tempered posteriors

Note : in the case of parametric estimation,

$$\pi_{n,\alpha}(\theta) \propto \exp[-\alpha n R_n(\theta)]\pi(\theta) = \exp\left[\alpha \sum_{i=1}^{n} \log p_\theta(X_i)\right] \pi(\theta).$$

Back to Example 2 : tempered posteriors

Note : in the case of parametric estimation,

$$\pi_{n,\alpha}(\theta) \propto \exp[-\alpha n R_n(\theta)]\pi(\theta) = \exp\left[\alpha \sum_{i=1}^{n} \log p_\theta(X_i)\right] \pi(\theta).$$

The tempered posterior

$$\pi_{n,\alpha}(\theta) \propto [L_n(\theta)]^\alpha \pi(\mathrm{d}\theta).$$

Back to Example 2 : tempered posteriors

Note : in the case of parametric estimation,

$$\pi_{n,\alpha}(\theta) \propto \exp[-\alpha n R_n(\theta)]\pi(\theta) = \exp\left[\alpha \sum_{i=1}^{n} \log p_\theta(X_i)\right] \pi(\theta).$$

The tempered posterior

$$\pi_{n,\alpha}(\theta) \propto [L_n(\theta)]^\alpha \pi(\mathrm{d}\theta).$$

Tempered posteriors are actually very useful for statistical inference.

Back to Example 2 : tempered posteriors

- easier to sample from.

R.M. Neal. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*.

G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

Back to Example 2 : tempered posteriors

- easier to sample from.

R.M. Neal. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*.

G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

- more robust to model misspecification.

P. Grünwald and T. Van Ommen (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*.

Back to Example 2 : tempered posteriors

- easier to sample from.

R.M. Neal. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*.

G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

- more robust to model misspecification.

P. Grünwald and T. Van Ommen (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*.

- theoretical analysis simpler.

A. Bhattacharya, D. Pati & Y. Yang (2019). Bayesian fractional posteriors. *The Annals of Statistics*.

Notation – summary

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$.

Notation – summary

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$.
2. Loss function ℓ.

Notation – summary

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$.
2. Loss function ℓ.
3. Minimize $R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)]$.

Notation – summary

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$.
2. Loss function ℓ.
3. Minimize $R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)]$.
4. Empirical risk

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i).$$

Notation – summary

1. We observe $X_1, \dots, X_n$ i.i.d from P^0 unknown in $\mathcal{X}$.
2. Loss function ℓ.
3. Minimize $R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)]$.
4. Empirical risk

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i).$$

5. Temperature $\alpha > 0$, Gibbs or tempered posterior

$$\pi_{n,\alpha}(\theta) \propto \exp[-\alpha n R_n(\theta)] \pi(\theta).$$

Computational problem

Apart from a few classical examples, $\pi_{n,\alpha}$ is intractable.

Computational problem

Apart from a few classical examples, $\pi_{n,\alpha}$ is intractable.
Popular methods to compute / sample from the (tempered) posterior :

- Monte-Carlo methods : MCMC (Gibbs Sampler, Metropolis-Hastings), SMC, Langevin Monte-Carlo, ABC etc.

Computational problem

Apart from a few classical examples, $\pi_{n,\alpha}$ is intractable. Popular methods to compute / sample from the (tempered) posterior :

- Monte-Carlo methods : MCMC (Gibbs Sampler, Metropolis-Hastings), SMC, Langevin Monte-Carlo, ABC etc.
- optimization methods : variational approximations or variational inference (VI) and expectation-propagation (EP).

Roadmap

I) Lecture 1 : approximate Bayesian inference in ML.

1. Introduction.
2. Definition of VI.
3. Examples of VI.

Roadmap

I) Lecture 1 : approximate Bayesian inference in ML.

1. Introduction.
2. Definition of VI.
3. Examples of VI.

II) Lecture 2 : statistical analysis of VI.

With theorems and all !

Roadmap

I) Lecture 1 : approximate Bayesian inference in ML.

1. Introduction.
2. Definition of VI.
3. Examples of VI.

II) Lecture 2 : statistical analysis of VI.

With theorems and all !

III) Seminar (Thursday next week) : online VI.

For data streams or large-scale learning.

Lecture 1

1. Introduction : computational issues in Bayesian learning
   - Bayesian learning
   - Computational issues
   - Roadmap
2. Variational Approximations : Definition
   - Definition of VI
   - The ELBO
   - Strategies for ELBO maximization
3. Examples of Variational Approximations in Machine Learning
   - Recommender systems and matrix completion
   - Deep learning

Variational Approximations : Definition

Reminder on Kullback divergence

Definition – Kullback divergence

Let P and Q be two probability distributions with p.d.f p and q respectively. Then :

$$\mathcal{K}(P,Q) = \mathbb{E}_{U\sim P}\left[\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(U)\right)\right] = \int \log\left(\frac{p(u)}{q(u)}\right) p(u)\mathrm{d}u.$$

Reminder on Kullback divergence

Definition – Kullback divergence

$$\mathcal{K}(P,Q) = \mathbb{E}_{U\sim P}\left[\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(U)\right)\right] = \int \log\left(\frac{p(u)}{q(u)}\right) p(u)\mathrm{d}u.$$

Theorem

$\mathcal{K}(P,Q) \geq 0$ and $\mathcal{K}(P,Q) = 0 \Leftrightarrow P = Q$.

Reminder on Kullback divergence

Definition – Kullback divergence

$$\mathcal{K}(P,Q) = \mathbb{E}_{U\sim P}\left[\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(U)\right)\right] = \int \log\left(\frac{p(u)}{q(u)}\right) p(u)\mathrm{d}u.$$

Theorem

$\mathcal{K}(P,Q) \geq 0$ and $\mathcal{K}(P,Q) = 0 \Leftrightarrow P = Q$.

Proof :

$$\mathcal{K}(P,Q) = -\int \log\left(\frac{q(u)}{p(u)}\right) p(u)\mathrm{d}u$$
$$\geq -\log\left(\int \left[\frac{q(u)}{p(u)}\right] p(u)\mathrm{d}u\right) \geq \log(1) = 0.$$

Definition of VI

1. Chose a tractable family $\mathcal{F}$ of probability distributions on the parameter θ,

Definition of VI

1. Chose a tractable family $\mathcal{F}$ of probability distributions on the parameter θ,
2. Define
$$\tilde{\pi}_{n,\alpha} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}).$$

Definition of VI

1. Chose a tractable family $\mathcal{F}$ of probability distributions on the parameter θ,
2. Define
$$\tilde{\pi}_{n,\alpha} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}).$$

Examples of $\mathcal{F}$:

- parametric approximation
$$\mathcal{F} = \left\{ \mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+ \right\}.$$

Definition of VI

1. Chose a tractable family $\mathcal{F}$ of probability distributions on the parameter θ,
2. Define

$$\tilde{\pi}_{n,\alpha} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}).$$

Examples of $\mathcal{F}$:

- parametric approximation

$$\mathcal{F} = \left\{ \mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+ \right\}.$$

- mean-field approximation, $\theta = (\theta_1, \theta_2) \in \Theta = \Theta_1 \times \Theta_2$,

$$\mathcal{F} : \{ \rho : \rho(\mathrm{d}\theta) = \rho_1(\mathrm{d}\theta_1) \otimes \rho_2(\mathrm{d}\theta_2) \}.$$

The ELBO

$$
\begin{aligned}
0 &\leq \mathcal{K}(\rho, \pi_{n,\alpha}) \\
&= \mathbb{E}_{\theta\sim\rho}\left[\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi_{n,\alpha}}(\theta)\right)\right]
\end{aligned}
$$

The ELBO

$$
\begin{aligned}
0 &\leq \mathcal{K}(\rho, \pi_{n,\alpha}) \\
&= \mathbb{E}_{\theta\sim\rho}\left[\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi_{n,\alpha}}(\theta)\right)\right] \\
&= \mathbb{E}_{\theta\sim\rho}\left[\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\frac{\mathbb{E}_{\theta\sim\pi}\left[\exp(-n\alpha R_n(\theta))\right]}{\exp(-n\alpha R_n(\theta))}\right)\right]
\end{aligned}
$$

The ELBO

$$
\begin{aligned}
0 & \leq \mathcal{K}(\rho, \pi_{n,\alpha}) \\
& = \mathbb{E}_{\theta\sim\rho}\left[\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi_{n,\alpha}}(\theta)\right)\right] \\
& = \mathbb{E}_{\theta\sim\rho}\left[\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\frac{\mathbb{E}_{\theta\sim\pi}\left[\exp(-n\alpha R_n(\theta))\right]}{\exp(-n\alpha R_n(\theta))}\right)\right] \\
& = n\alpha\mathbb{E}_{\theta\sim\rho}\left[R_n(\theta)\right] + \mathcal{K}(\rho, \pi) + \log \mathbb{E}_{\theta\sim\pi}\left[\exp(-n\alpha R_n(\theta))\right].
\end{aligned}
$$

The ELBO

$$
\begin{aligned}
0 &\leq \mathcal{K}(\rho, \pi_{n,\alpha}) \\
&= \mathbb{E}_{\theta\sim\rho}\left[\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi_{n,\alpha}}(\theta)\right)\right] \\
&= \mathbb{E}_{\theta\sim\rho}\left[\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\frac{\mathbb{E}_{\theta\sim\pi}\left[\exp(-n\alpha R_n(\theta))\right]}{\exp(-n\alpha R_n(\theta))}\right)\right] \\
&= n\alpha\mathbb{E}_{\theta\sim\rho}\left[R_n(\theta)\right] + \mathcal{K}(\rho,\pi) + \log\mathbb{E}_{\theta\sim\pi}\left[\exp(-n\alpha R_n(\theta))\right].
\end{aligned}
$$

That is,

$$
\begin{aligned}
\text{Evidence} &= \log\mathbb{E}_{\theta\sim\pi}\left[\exp(-n\alpha R_n(\theta))\right] \\
&\geq -n\alpha\mathbb{E}_{\theta\sim\rho}\left[R_n(\theta)\right] - \mathcal{K}(\rho,\pi) \\
&= \mathrm{ELBO}(\rho) \text{ (Evidence Lower Bound).}
\end{aligned}
$$

Alternative definition of VI with ELBO

We end up with two definitions of VI.

Alternative definition of VI with ELBO

We end up with two definitions of VI.

1. best approximation of the posterior

$$\tilde{\pi}_{n,\alpha} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}),$$

Alternative definition of VI with ELBO

We end up with two definitions of VI.

1. best approximation of the posterior

$$\tilde{\pi}_{n,\alpha} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}),$$

2. maximization of the evidence lower bound

$$\begin{aligned} \tilde{\pi}_{n,\alpha} &= \arg\max_{\rho \in \mathcal{F}} \mathrm{ELBO}(\rho) \\ &= \arg\max_{\rho \in \mathcal{F}} \left\{ -n\alpha \mathbb{E}_{\theta \sim \rho} \left[R_n(\theta) \right] - \mathcal{K}(\rho, \pi) \right\}. \end{aligned}$$

Donsker and Varadhan's variational inequality

Remark : from the above inequality

$$0 \leq \mathcal{K}(\rho, \pi_{n,\alpha}) = -\mathrm{ELBO}(\rho) + \text{evidence},$$

Donsker and Varadhan's variational inequality

Remark : from the above inequality

$$0 \leq \mathcal{K}(\rho, \pi_{n,\alpha}) = -\mathrm{ELBO}(\rho) + \text{evidence},$$

it is clear that without the constraint $\rho \in \mathcal{F}$, the ELBO is maximized by $\rho = \pi_{n,\alpha}$.

Donsker and Varadhan's variational inequality

Remark : from the above inequality

$$0 \leq \mathcal{K}(\rho, \pi_{n,\alpha}) = -\mathrm{ELBO}(\rho) + \text{evidence},$$

it is clear that without the constraint $\rho \in \mathcal{F}$, the ELBO is maximized by $\rho = \pi_{n,\alpha}$.

Theorem : Donsker and Varadhan's variational inequality

$$\mathbb{E}_{\theta\sim\pi_{n,\alpha}}[R_n(\theta)] + \frac{\mathcal{K}(\pi_{n,\alpha}, \pi)}{n\alpha} = \inf_{\rho} \left\{ \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathcal{K}(\rho, \pi)}{n\alpha} \right\}.$$

How to maximize the ELBO ?

How to maximize the ELBO ?

Parametric variational inference : $\mathcal{F} = \{q_\lambda, \lambda \in \Lambda\}$.

How to maximize the ELBO ?

Parametric variational inference : $\mathcal{F} = \{q_\lambda, \lambda \in \Lambda\}$.

Gradient algorithm

$$\lambda_{t+1} = \lambda_t + \eta \nabla \mathrm{ELBO}(q_{\lambda_t}).$$

How to maximize the ELBO ?

Parametric variational inference : $\mathcal{F} = \{q_\lambda, \lambda \in \Lambda\}$.

Gradient algorithm

$$\lambda_{t+1} = \lambda_t + \eta \nabla \mathrm{ELBO}(q_{\lambda_t}).$$

Usually, the gradient is not available in closed-form but often it is possible to build an unbiased estimate of it : $\hat{\nabla}\mathrm{ELBO}(q_{\lambda_t})$.

How to maximize the ELBO ?

Parametric variational inference : $\mathcal{F} = \{q_\lambda, \lambda \in \Lambda\}$.

Gradient algorithm

$$\lambda_{t+1} = \lambda_t + \eta \nabla \mathrm{ELBO}(q_{\lambda_t}).$$

Usually, the gradient is not available in closed-form but often it is possible to build an unbiased estimate of it : $\hat{\nabla}\mathrm{ELBO}(q_{\lambda_t})$.

Stochastic gradient algorithm

$$\lambda_{t+1} = \lambda_t + \eta \hat{\nabla} \mathrm{ELBO}(q_{\lambda_t}).$$

Stochastic gradient of the ELBO

Ex : $q_\lambda = \mathcal{N}(\mu, \sigma^2 I)$, $\lambda = (\mu, \sigma) \in \mathbb{R}^d \times \mathbb{R}_+^*$, $\pi = \mathcal{N}(0, I)$.

Stochastic gradient of the ELBO

Ex : $q_\lambda = \mathcal{N}(\mu, \sigma^2 I)$, $\lambda = (\mu, \sigma) \in \mathbb{R}^d \times \mathbb{R}_+^*$, $\pi = \mathcal{N}(0, I)$.

$$\begin{aligned}
&\mathrm{ELBO}(q_\lambda) \\
&= -\mathbb{E}_{\theta \sim q_\lambda}[n\alpha R_n(\theta)] - \mathcal{K}(q_\lambda, \pi) \\
&= -\mathbb{E}_{\theta \sim \mathcal{N}(0,I)}[n\alpha R_n(\mu + \sigma\theta)] - \frac{\|\mu\|^2 + k(\sigma^2 - \log(\sigma^2) - 1)}{2}
\end{aligned}$$

and so, under for a smooth R_n,

$$\nabla \mathrm{ELBO}(q_\lambda) = -\mathbb{E}_{\theta \sim \mathcal{N}(0,I)}[n\alpha \nabla_\lambda R_n(\mu + \sigma\theta)] - \begin{pmatrix} \mu \\ \frac{k}{2}\left(\sigma - \frac{2}{\sigma}\right) \end{pmatrix},$$

Stochastic gradient of the ELBO

Ex : $q_\lambda = \mathcal{N}(\mu, \sigma^2 I)$, $\lambda = (\mu, \sigma) \in \mathbb{R}^d \times \mathbb{R}_+^*$, $\pi = \mathcal{N}(0, I)$.

$$\begin{aligned} &\mathrm{ELBO}(q_\lambda) \\ &= -\mathbb{E}_{\theta \sim q_\lambda}[n\alpha R_n(\theta)] - \mathcal{K}(q_\lambda, \pi) \\ &= -\mathbb{E}_{\theta \sim \mathcal{N}(0,I)}[n\alpha R_n(\mu + \sigma\theta)] - \frac{\|\mu\|^2 + k(\sigma^2 - \log(\sigma^2) - 1)}{2} \end{aligned}$$

and so, under for a smooth R_n,

$$\nabla \mathrm{ELBO}(q_\lambda) = -\mathbb{E}_{\theta \sim \mathcal{N}(0,I)}[n\alpha \nabla_\lambda R_n(\mu + \sigma\theta)] - \begin{pmatrix} \mu \\ \frac{k}{2}\left(\sigma - \frac{2}{\sigma}\right) \end{pmatrix},$$

$$\hat{\nabla} \mathrm{ELBO}(q_\lambda) = -\begin{pmatrix} \frac{n\alpha}{m} \sum_{j=1}^m \nabla R_n(\mu + \sigma\theta_j) + \mu \\ \frac{n\alpha}{m} \sum_{j=1}^m \theta_j \nabla R_n(\mu + \sigma\theta_j) + \frac{k}{2}\left(\sigma - \frac{2}{\sigma}\right) \end{pmatrix}.$$

Mean-field approximation

Mean-field variational approximation :

$$\mathcal{F} : \{\rho : \rho(\mathrm{d}\theta) = \rho_1(\mathrm{d}\theta_1) \otimes \rho_2(\mathrm{d}\theta_2)\} .$$

Alternate optimization

$$\rho_1^{t+1} = \arg\max_{\rho_1} \mathrm{ELBO}(\rho_1 \times \rho_2^t)$$
$$\rho_2^{t+1} = \arg\max_{\rho_2} \mathrm{ELBO}(\rho_1^{t+1} \times \rho_2)$$

Mean-field approximation : explicit formula

$$\rho_1^{t+1} = \arg\max_{\rho_1} \mathrm{ELBO}(\rho_1 \otimes \rho_2^t)$$

Mean-field approximation : explicit formula

$$\rho_1^{t+1} = \arg\max_{\rho_1} \mathrm{ELBO}(\rho_1 \otimes \rho_2^t)$$

Assume $\pi = \pi_1 \times \pi_2$.

$$\max_{\rho_1} \left\{ \alpha n \mathbb{E}_{\theta_1 \sim \rho_1} \mathbb{E}_{\theta_2 \sim \rho_2^t} [R_n(\theta_1, \theta_2)] + \mathcal{K}(\rho_1, \pi_1) + \mathcal{K}(\rho_2^t, \pi_2) \right\}.$$

Mean-field approximation : explicit formula

$$\rho_1^{t+1} = \arg\max_{\rho_1} \mathrm{ELBO}(\rho_1 \otimes \rho_2^t)$$

Assume $\pi = \pi_1 \times \pi_2$.

$$\max_{\rho_1} \left\{ \alpha n \mathbb{E}_{\theta_1 \sim \rho_1} \mathbb{E}_{\theta_2 \sim \rho_2^t} [R_n(\theta_1, \theta_2)] + \mathcal{K}(\rho_1, \pi_1) + \mathcal{K}(\rho_2^t, \pi_2) \right\} .$$

Use Donsker and Varadhan's variational formula.

$$\rho_1^{t+1}(\theta_1) \propto \exp\left[-n\alpha \mathbb{E}_{\theta_2 \sim \rho_2^t} [R_n(\theta_1, \theta_2) | \theta_1] \right] \pi_1(\theta_1).$$

Further reading

Recent survey on variational inference :

D. M. Blei, A. Kucukelbir & J. D. McAuliffe (2017). Variational inference : A review for statisticians. *Journal of the American Statistical Association*.

Examples

Example 1 : recommendation via matrix completion

				
Claire	4	?	3	...
Nial	?	4	?	...
Brendon	?	5	4	...
Andrew	?	4	?	...
Adrian	1	?	?	...
Damien	?	1	?	...
⋮	⋮	⋮	⋮	⋱

The Netflix challenge

Matrix completion : notations

The parameter θ is a matrix $M^0 \in \mathbb{R}^{m \times p}$, with $m, p \geq 1$.

Matrix completion : notations

The parameter θ is a matrix $M^0 \in \mathbb{R}^{m \times p}$, with $m, p \geq 1$.
Under P_M, the observations are random entries of this matrix with possible noise :

$$Y_i = M^0_{i_k, j_k} + \varepsilon_k$$

where the (i_k, j_k) are i.i.d $\mathcal{U}(\{1, \dots, m\} \times \{1, \dots, p\})$.

Matrix completion : notations

The parameter θ is a matrix $M^0 \in \mathbb{R}^{m \times p}$, with $m, p \geq 1$. Under P_M, the observations are random entries of this matrix with possible noise :

$$Y_i = M^0_{i_k, j_k} + \varepsilon_k$$

where the (i_k, j_k) are i.i.d $\mathcal{U}(\{1, \dots, m\} \times \{1, \dots, p\})$. Assume that the ε_k are i.i.d $\mathcal{N}(0, \sigma^2)$, σ^2 known.

Matrix completion : notations

The parameter θ is a matrix $M^0 \in \mathbb{R}^{m\times p}$, with $m,p \geq 1$. Under P_M, the observations are random entries of this matrix with possible noise :

$$Y_i = M^0_{i_k,j_k} + \varepsilon_k$$

where the (i_k, j_k) are i.i.d $\mathcal{U}(\{1,\dots,m\} \times \{1,\dots,p\})$. Assume that the ε_k are i.i.d $\mathcal{N}(0,\sigma^2)$, σ^2 known. We have

$$\mathcal{K}(P_M, P_N) = \frac{1}{mp} \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{(M_{i,j} - N_{i,j})^2}{2\sigma^2} = \frac{\|M - N\|_F^2}{2\sigma^2 mp}.$$

Matrix completion : notations

The parameter θ is a matrix $M^0 \in \mathbb{R}^{m\times p}$, with $m, p \geq 1$. Under P_M, the observations are random entries of this matrix with possible noise :

$$Y_i = M^0_{i_k, j_k} + \varepsilon_k$$

where the (i_k, j_k) are i.i.d $\mathcal{U}(\{1,\dots,m\} \times \{1,\dots,p\})$. Assume that the ε_k are i.i.d $\mathcal{N}(0, \sigma^2)$, σ^2 known. We have

$$\mathcal{K}(P_M, P_N) = \frac{1}{mp} \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{(M_{i,j} - N_{i,j})^2}{2\sigma^2} = \frac{\|M - N\|_F^2}{2\sigma^2 mp}.$$

Usual assumption : M^0 is low-rank.

Prior specification - main idea

Define :

$$\underbrace{M}_{p\times m} = \underbrace{U}_{p\times k}\underbrace{V^T}_{k\times m}.$$

Prior specification - main idea

Define :

$$\underbrace{M}_{p\times m} = \underbrace{U}_{p\times k}\underbrace{V^T}_{k\times m}.$$

Let $U_{\cdot,\ell} \sim \mathcal{N}(0,\gamma I)$ denote the ℓ-th column of M, we have :

$$M = \sum_{\ell=1}^{k} U_{\cdot,\ell}(V_{\cdot,\ell})^T \quad \Rightarrow \quad \mathrm{rank}(M) \leq k.$$

Prior specification - adaptation

R. Salakhutdinov & A. Mnih (2008). Bayesian probabilistic matrix factorization using MCMC. *Proceedings of ICML*.

Prior specification - adaptation

R. Salakhutdinov & A. Mnih (2008). Bayesian probabilistic matrix factorization using MCMC. *Proceedings of ICML.*

$$M = \sum_{\ell=1}^{k} U_{\cdot,\ell}(V_{\cdot,\ell})^T$$

with k large - e.g. $k = \min(p, m)$.

Prior specification - adaptation

R. Salakhutdinov & A. Mnih (2008). Bayesian probabilistic matrix factorization using MCMC. *Proceedings of ICML*.

$$M = \sum_{\ell=1}^{k} U_{\cdot,\ell} (V_{\cdot,\ell})^T$$

with k large - e.g. $k = \min(p, m)$.

Definition of π :

- $U_{\cdot,\ell}, V_{\cdot,\ell} \sim \mathcal{N}(0, \gamma_\ell I)$,
- γ_ℓ is itself random, such that most of the $\gamma_\ell \simeq 0$

$$\frac{1}{\gamma_\ell} \sim \mathrm{Gamma}(a, b).$$

Variational approximation

Y. J. Lim & Y. W. Teh (2007). Variational Bayesian approach to movie rating prediction. *Proceedings of KDD cup and workshop*.

Mean-field approximation, $\mathcal{F}$ given by :

$$\rho(\mathrm{d}U, \mathrm{d}V, \mathrm{d}\gamma) = \bigotimes_{i=1}^{m} \rho_{U_i}(\mathrm{d}U_{i,\cdot}) \bigotimes_{j=1}^{p} \rho_{V_j}(\mathrm{d}V_{j,\cdot}) \bigotimes_{k=1}^{K} \rho_{\gamma_k}(\gamma_k).$$

Variational approximation

Y. J. Lim & Y. W. Teh (2007). Variational Bayesian approach to movie rating prediction. *Proceedings of KDD cup and workshop.*

Mean-field approximation, $\mathcal{F}$ given by :

$$\rho(\mathrm{d}U, \mathrm{d}V, \mathrm{d}\gamma) = \bigotimes_{i=1}^{m} \rho_{U_i}(\mathrm{d}U_{i,\cdot}) \bigotimes_{j=1}^{p} \rho_{V_j}(\mathrm{d}V_{j,\cdot}) \bigotimes_{k=1}^{K} \rho_{\gamma_k}(\gamma_k).$$

It can be shown that

1. ρ_{U_i} is $\mathcal{N}(\mathbf{m}_{i,\cdot}^T, \mathcal{V}_i)$,
2. ρ_{V_j} is $\mathcal{N}(\mathbf{n}_{j,\cdot}^T, \mathcal{W}_j)$,
3. ρ_{γ_k} is inverse-$\Gamma(a + (m_1 + m_2)/2, \beta_k)$,

for some $m \times K$ matrix $\mathbf{m}$ whose rows are denoted by $\mathbf{m}_{i,\cdot}$, some $p \times K$ matrix $\mathbf{n}$ and some vector $\beta = (\beta_1, \ldots, \beta_K)$.

The VB algorithm

The parameters are updated iteratively through the formulae

1. moments of U :

$$\mathbf{m}_{i,\cdot}^T := \frac{2\alpha}{n}\mathcal{V}_i \sum_{k:i_k=i} Y_{i_k,j_k}\mathbf{n}_{j_k,\cdot}^T$$

$$\mathcal{V}_i^{-1} := \frac{2\alpha}{n}\sum_{k:i_k=i}\left[\mathcal{W}_{j_k} + \mathbf{n}_{j_k,\cdot}\mathbf{n}_{j_k,\cdot}^T\right] + \left(a + \frac{m_1+m_2}{2}\right)\mathbf{diag}(\beta)^{-1}$$

2. moments of V :

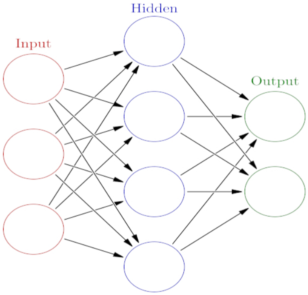
$$\mathbf{n}_{j,\cdot}^T := \frac{2\alpha}{n}\mathcal{W}_j \sum_{k:j_k=j} Y_{i_k,j_k}\mathbf{m}_{i_k,\cdot}^T$$

$$\mathcal{W}_j^{-1} := \frac{2\alpha}{n}\sum_{k:j_k=j}\left[\mathcal{V}_{i_k} + \mathbf{m}_{i_k,\cdot}\mathbf{m}_{i_k,\cdot}^T\right] + \left(a + \frac{m_1+m_2}{2}\right)\mathbf{diag}(\beta)^{-1}$$

3. moments of γ :

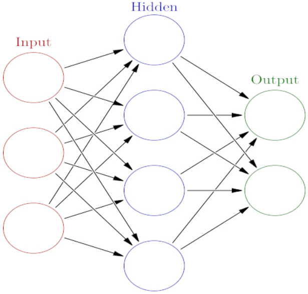
$$\beta_k := \frac{1}{2}\left[\sum_{i=1}^{m_1}\left(\mathbf{m}_{i,k}^2 + (\mathcal{V}_i)_{k,k}\right) + \sum_{j=1}^{m_2}\left(\mathbf{n}_{j,k}^2 + (\mathcal{V}_j)_{k,k}\right)\right].$$

Example 2 : deep learning



Source : Wikipedia.

Example 2 : deep learning



Source : Wikipedia.

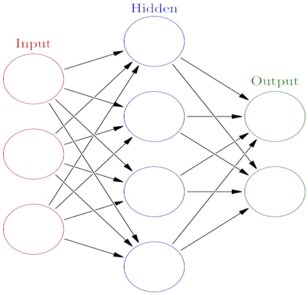
Neural network, recursive definition :

$$f_{0,\theta}(x) = x,$$

$$f_{\ell+1,\theta}^{(i)}(x) = \varphi\left(\sum_{j=1}^{s_\ell} \theta_{i,j}^{(\ell)} f_{\ell,\theta}^{(j)}(x)\right),$$

$$f_\theta(x) = \psi\left(\sum_{j=1}^{s_L} \theta_{i,j}^{(L)} f_{\ell,\theta}^{(j)}(x)\right).$$

Example 2 : deep learning



Source : Wikipedia.

Neural network, recursive definition :

$$f_{0,\theta}(x) = x,$$

$$f^{(i)}_{\ell+1,\theta}(x) = \varphi\left(\sum_{j=1}^{s_\ell} \theta^{(\ell)}_{i,j} f^{(j)}_{\ell,\theta}(x)\right),$$

$$f_{\theta}(x) = \psi\left(\sum_{j=1}^{s_L} \theta^{(L)}_{i,j} f^{(j)}_{\ell,\theta}(x)\right).$$

Prior π : independent

$$\theta^{(\ell)}_{i,j} \sim \mathcal{N}(0, \sigma^2_\ell).$$

Example 2 : deep learning

The posterior is *extremely* complicated.

Example 2 : deep learning

The posterior is *extremely* complicated.

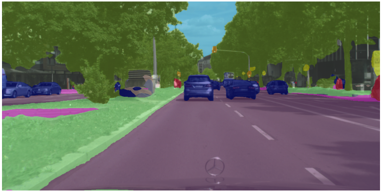
Using a mean-field variational approximation where all the $\theta_{i,j}^{(\ell)}$ are independent $\mathcal{N}(m_{i,j}^{(\ell)}, (\sigma_{i,j}^{(\ell)})^2)$ *a posteriori*, the authors of :

K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota & M. E. Khan (2019). Practical Deep Learning with Bayesian Principles. *NeurIPS*.

proposed a refined stochastic gradient algorithm and reached state-of-the-art performances on large datasets such as CIFAR-10 and ImageNet.

Example 2 : deep learning

K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota & M. E. Khan (2019). Practical Deep Learning with Bayesian Principles. *NeurIPS*.



Picture : Roman Bachmann.