Theoretical study of variational inference

Badr-Eddine Chérief-Abdellatif
CREST - ENSAE - Institut Polytechnique de Paris

RIKEN AIP Seminar
February 20, 2020

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by Q : $\frac{\mathrm{d}P_\theta}{\mathrm{d}Q} = p_\theta$. Prior π on Θ.

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by Q : $\frac{\mathrm{d}P_\theta}{\mathrm{d}Q} = p_\theta$. Prior π on Θ.

The likelihood

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by Q : $\frac{\mathrm{d}P_\theta}{\mathrm{d}Q} = p_\theta$. Prior π on Θ.

The likelihood

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$

The posterior

$$\pi_n(\mathrm{d}\theta) \propto L_n(\theta)\pi(\mathrm{d}\theta).$$

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by Q : $\frac{\mathrm{d}P_\theta}{\mathrm{d}Q} = p_\theta$. Prior π on Θ.

The likelihood

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$

The posterior

$$\pi_n(\mathrm{d}\theta) \propto L_n(\theta)\pi(\mathrm{d}\theta).$$

The tempered posterior - $0 < \alpha < 1$

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

Various reasons to use a tempered posterior

- Easier to sample from

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

Various reasons to use a tempered posterior

- Easier to sample from

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

- 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*.

Various reasons to use a tempered posterior

- Easier to sample from

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

- 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 easier

A. Bhattacharya, D. Pati & Y. Yang (2016). Bayesian fractional posteriors. *Preprint arxiv :1611.01125*.

Computation of the posterior

- Explicit form (conjugate models),

Computation of the posterior

- Explicit form (conjugate models),
- MCMC algorithms : Metropolis-Hastings, Gibbs sampler, Langevin Monte Carlo...

Computation of the posterior

- Explicit form (conjugate models),
- MCMC algorithms : Metropolis-Hastings, Gibbs sampler, Langevin Monte Carlo...

But...

- when the dimension is large, the convergence of MCMC can be extremely slow,

Computation of the posterior

- Explicit form (conjugate models),
- MCMC algorithms : Metropolis-Hastings, Gibbs sampler, Langevin Monte Carlo...

But...

- when the dimension is large, the convergence of MCMC can be extremely slow,
- when the model is complex or when the sample size is large, each evaluation of $\pi_{n,\alpha}(\theta)$ can be expensive.

Computation of the posterior

- Explicit form (conjugate models),
- MCMC algorithms : Metropolis-Hastings, Gibbs sampler, Langevin Monte Carlo...

But...

- when the dimension is large, the convergence of MCMC can be extremely slow,
- when the model is complex or when the sample size is large, each evaluation of $\pi_{n,\alpha}(\theta)$ can be expensive.

For these reasons, in the past 20 years, many methods targeting an approximation of $\pi_{n,\alpha}$ became popular : **ABC**, **EP algorithm**, variational inference...

Variational approximations : definitions

Idea of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

Variational approximations : definitions

Idea of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

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

Variational approximations : definitions

Idea of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

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

Examples :

- parametric approximation

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

Variational approximations : definitions

Idea of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

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

Examples :

- parametric approximation

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

- mean-field approximation, $\Theta = \Theta_1 \times \Theta_2$ and

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

Evidence Lower BOund (ELBO)

Note that :

$$\begin{aligned}\tilde{\pi}_{n,\alpha} &= \arg\min_{\rho\in\mathcal{F}} KL(\rho, \pi_{n,\alpha}) \\ &= \arg\min_{\rho\in\mathcal{F}} \underbrace{\left\{ -\alpha \int \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i)\rho(\mathrm{d}\theta) + KL(\rho, \pi) \right\}}_{-\mathrm{ELBO}(\rho)}.\end{aligned}$$

Evidence Lower BOund (ELBO)

Note that :

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

$$= \arg\min_{\rho \in \mathcal{F}} \underbrace{\left\{ -\alpha \int \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i) \rho(\mathrm{d}\theta) + KL(\rho, \pi) \right\}}_{-\mathrm{ELBO}(\rho)}.$$

So we have the equivalent definition :

$$\tilde{\pi}_{n,\alpha} := \arg\max_{\rho \in \mathcal{F}} \ \mathrm{ELBO}(\rho).$$

Outline of the talk

After this introduction :

Outline of the talk

After this introduction :

Section 2 will address the following question :

What are the conditions ensuring that $\tilde{\pi}_{n,\alpha}$ leads to good estimators ?

Outline of the talk

After this introduction :

Section 2 will address the following question :

What are the conditions ensuring that $\tilde{\pi}_{n,\alpha}$ leads to good estimators ?

We will study general conditions, an example (DNNs) and extensions.

Outline of the talk

After this introduction :

Section 2 will address the following question :

What are the conditions ensuring that $\tilde{\pi}_{n,\alpha}$ leads to good estimators ?

We will study general conditions, an example (DNNs) and extensions.

Section 3 will address the following question :

Are there efficient algorithms to (provably) compute $\tilde{\pi}_{n,\alpha}$?

Outline of the talk

After this introduction :

Section 2 will address the following question :

What are the conditions ensuring that $\tilde{\pi}_{n,\alpha}$ leads to good estimators ?

We will study general conditions, an example (DNNs) and extensions.

Section 3 will address the following question :

Are there efficient algorithms to (provably) compute $\tilde{\pi}_{n,\alpha}$?

We will see that fast algorithms from online optimization can be used to compute variational approximations.

Tools for the consistency of VB

The α-Rényi divergence for $\alpha \in (0,1)$

$$D_\alpha(P,R) = \frac{1}{\alpha - 1} \log \int (\mathrm{d}P)^\alpha (\mathrm{d}R)^{1-\alpha}.$$

Tools for the consistency of VB

The α-Rényi divergence for $\alpha \in (0,1)$

$$D_\alpha(P,R) = \frac{1}{\alpha-1}\log\int (\mathrm{d}P)^\alpha(\mathrm{d}R)^{1-\alpha}.$$

All the properties derived in :

T. Van Erven & P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 2014.

Among others, for $1/2 \leq \alpha$, link with Hellinger and Kullback :

$$\mathcal{H}^2(P,R) \leq D_\alpha(P,R) \xrightarrow[\alpha\nearrow 1]{} KL(P,R).$$

Notions of Concentration and Consistency

Concentration at rate r_n

$$\rho\left(\theta \in \Theta \;/\; D_\alpha(P_\theta, P_{\theta_0}) > M_n r_n\right) \xrightarrow[n\to+\infty]{} 0$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

Notions of Concentration and Consistency

Concentration at rate r_n

$$\rho\left(\theta \in \Theta \;/\; D_\alpha(P_\theta, P_{\theta_0}) > M_n r_n\right) \xrightarrow[n\to+\infty]{} 0$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

Consistency at rate r_n

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\rho(d\theta)\right] \leq r_n.$$

Notions of Concentration and Consistency

Concentration at rate r_n

$$\rho\Big(\theta \in \Theta \;/\; D_\alpha(P_\theta, P_{\theta_0}) > M_n r_n\Big) \xrightarrow[n \to +\infty]{} 0$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

Consistency at rate r_n

$$\mathbb{E}\Big[\int D_\alpha(P_\theta, P_{\theta_0})\rho(d\theta)\Big] \leq r_n.$$

Consistency implies concentration of the Bayesian distribution.

Technical condition for posterior concentration

Technical condition for posterior concentration

Prior mass condition for concentration of tempered posteriors

The rate (r_n) is such that

$$\pi[\mathcal{B}(r_n)] \geq e^{-nr_n}$$

where $\mathcal{B}(r) = \{\theta \in \Theta : KL(P_{\theta^0}, P_\theta) \leq r\}$.

Technical condition for posterior concentration

Prior mass condition for concentration of tempered posteriors

The rate (r_n) is such that

$$\pi[\mathcal{B}(r_n)] \geq e^{-nr_n}$$

where $\mathcal{B}(r) = \{\theta \in \Theta : KL(P_{\theta^0}, P_\theta) \leq r\}$.

Prior mass condition for concentration of Variational Bayes

The rate (r_n) is such that there exists $\rho_n \in \mathcal{F}$ such that

$$\int KL(P_{\theta^0}, P_\theta)\rho_n(\mathrm{d}\theta) \leq r_n, \text{ and } KL(\rho_n, \pi) \leq nr_n.$$

What do we know about $\pi_{n,\alpha}$?

Theorem, variant of (Bhattacharya, Pati & Yang)

Under the prior mass condition, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\pi_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

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

Extension of previous result to VB

Theorem (Alquier & Ridgway)

Under the extended prior mass condition, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

P. Alquier & J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 2019.

Misspecified case

Theorem (Alquier & Ridgway)

Under the extended prior mass condition, for any $\alpha \in (0, 1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Misspecified case

Theorem (Alquier & Ridgway)

Under the extended prior mass condition, for any $\alpha \in (0, 1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Assume now that $X_1, \ldots, X_n$ i.i.d $\sim Q \notin \{P_\theta, \theta \in \Theta\}$.

Misspecified case

Theorem (Alquier & Ridgway)

Under the extended prior mass condition, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Assume now that $X_1, \ldots, X_n$ i.i.d $\sim Q \notin \{P_\theta, \theta \in \Theta\}$.

Theorem (Alquier and Ridgway)

Under a similar condition, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, Q)\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{\alpha}{1-\alpha}\inf_\theta KL(Q, P_\theta) + \frac{1+\alpha}{1-\alpha} r_n.$$

Nonparametric regression & Deep Neural Networks

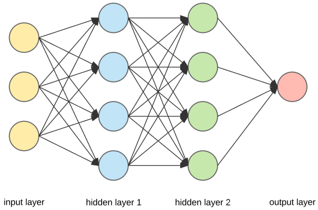
Nonparametric regression

- $X_i \sim \mathcal{U}([-1,1]^d)$,
- $Y_i = f_0(X_i) + \zeta_i$,
- $\zeta_i \sim \mathcal{N}(0, \sigma^2)$.

Nonparametric regression & Deep Neural Networks

Nonparametric regression

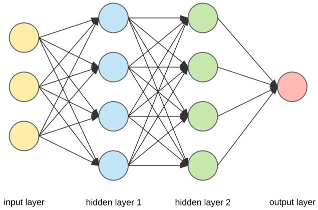
- $X_i \sim \mathcal{U}([-1,1]^d)$,
- $Y_i = f_0(X_i) + \zeta_i$,
- $\zeta_i \sim \mathcal{N}(0, \sigma^2)$.



Nonparametric regression & Deep Neural Networks

Nonparametric regression

- $X_i \sim \mathcal{U}([-1,1]^d)$,
- $Y_i = f_0(X_i) + \zeta_i$,
- $\zeta_i \sim \mathcal{N}(0, \sigma^2)$.



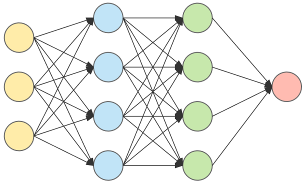
Deep neural networks

- Depth $L \geq 3$, width $D \geq d$, sparsity $S \leq T$.

Nonparametric regression & Deep Neural Networks

Nonparametric regression

- $X_i \sim \mathcal{U}([-1,1]^d)$,
- $Y_i = f_0(X_i) + \zeta_i$,
- $\zeta_i \sim \mathcal{N}(0, \sigma^2)$.



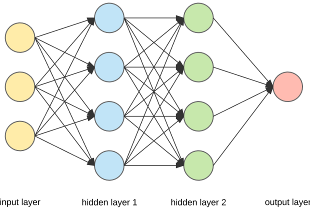
Deep neural networks

- Depth $L \geq 3$, width $D \geq d$, sparsity $S \leq T$.
- Parameter $\theta = \{(A_1, b_1), ..., (A_L, b_L)\}$.

Nonparametric regression & Deep Neural Networks

Nonparametric regression

- $X_i \sim \mathcal{U}([-1,1]^d)$,
- $Y_i = f_0(X_i) + \zeta_i$,
- $\zeta_i \sim \mathcal{N}(0, \sigma^2)$.



Deep neural networks

- Depth $L \geq 3$, width $D \geq d$, sparsity $S \leq T$.
- Parameter $\theta = \{(A_1, b_1), ..., (A_L, b_L)\}$.
- $f_\theta(x) = A_L \rho(A_{L-1} ... \rho(A_1 x + b_1) + ... + b_{L-1}) + b_L$.

ReLU Deep Neural Networks : convergence rates

Theorem (C.-A.)

Chose spike-and-slab prior and variational set on θ.

ReLU Deep Neural Networks : convergence rates

Theorem (C.-A.)

Chose spike-and-slab prior and variational set on θ. Then :

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \tilde{\pi}_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha} \inf_{\theta^*} \|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D},$$

with $\quad r_n^{S,L,D} \sim \frac{S \log(nL/S)}{n} \vee \frac{LS \log D}{n}$.

ReLU Deep Neural Networks : convergence rates

Theorem (C.-A.)

Chose spike-and-slab prior and variational set on θ. Then :

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \tilde{\pi}_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha} \inf_{\theta^*} \|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1+\frac{\sigma^2}{\alpha}\right) r_n^{S,L,D},$$

with $\quad r_n^{S,L,D} \sim \frac{S\log(nL/S)}{n} \vee \frac{LS\log D}{n}.$

If f_0 β-Hölder for suitable (S, L, D) :

ReLU Deep Neural Networks : convergence rates

Theorem (C.-A.)

Chose spike-and-slab prior and variational set on θ. Then :

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \tilde{\pi}_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha} \inf_{\theta^*} \|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1+\frac{\sigma^2}{\alpha}\right) r_n^{S,L,D},$$

with $\quad r_n^{S,L,D} \sim \frac{S\log(nL/S)}{n} \vee \frac{LS\log D}{n}$.

If f_0 β-Hölder for suitable (S, L, D) : $\quad \tilde{\mathcal{O}}\left(n^{-\frac{2\beta}{2\beta+d}}\right)$.

ReLU Deep Neural Networks : convergence rates

Theorem (C.-A.)

Chose spike-and-slab prior and variational set on θ. Then :

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \tilde{\pi}_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha}\inf_{\theta^*}\|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1+\frac{\sigma^2}{\alpha}\right) r_n^{S,L,D},$$

with $\quad r_n^{S,L,D} \sim \frac{S\log(nL/S)}{n} \vee \frac{LS\log D}{n}$.

If f_0 β-Hölder for suitable (S, L, D) : $\quad \tilde{\mathcal{O}}\left(n^{-\frac{2\beta}{2\beta+d}}\right)$.

C.-A.. Convergence Rates of Variational Inference in Sparse Deep Learning. *Preprint Arxiv*, 2019.

More extensions

1. more general models with latent variables :

Y. Yang, D. Pati & A. Bhattacharya. α-Variational Inference with Statistical Guarantees. *The Annals of Statistics*, 2019.

More extensions

1. more general models with latent variables :

   Y. Yang, D. Pati & A. Bhattacharya. α-Variational Inference with Statistical Guarantees. *The Annals of Statistics*, 2019.

2. case $\alpha = 1$, *i.e* approximation of the "usual" posterior :

   F. Zhang & C. Gao. Convergence Rates of Variational Posterior Distributions. *The Annals of Statistics*, 2019.

More extensions

1. more general models with latent variables :

Y. Yang, D. Pati & A. Bhattacharya. α-Variational Inference with Statistical Guarantees. *The Annals of Statistics*, 2019.

2. case $\alpha = 1$, *i.e* approximation of the "usual" posterior :

F. Zhang & C. Gao. Convergence Rates of Variational Posterior Distributions. *The Annals of Statistics*, 2019.

3. approximation based on another distance, for example :

$$\tilde{\pi}_{n,\alpha} := \arg\min_{\rho \in \mathcal{F}} \mathcal{W}(\rho, \pi_{n,\alpha}) \text{ (Wasserstein distance)},$$

J. Huggins, T. Campbell, M. Kasprzak & T. Broderick. Practical bounds on the error of Bayesian posterior approximations : a nonasymptotic approach. *Preprint arXiv*, 2018.

Online optimization

Online optimization

1. 1. initialize θ_1,

Online optimization

1.
   1. initialize θ_1,
   2. x_1 revealed,

Online optimization

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$

Online optimization

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2.
   1. update $\theta_1 \to \theta_2$,

Online optimization

1. 1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2. 1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,

Online optimization

1. 1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2. 1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $\ell(x_2; \theta_2)$

Online optimization

1. 1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2. 1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $\ell(x_2; \theta_2)$
3. 1. update $\theta_2 \to \theta_3$,

Online optimization

1. 1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2. 1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $\ell(x_2; \theta_2)$
3. 1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,

Online optimization

1. 1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2. 1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $\ell(x_2; \theta_2)$
3. 1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,
   3. incur loss $\ell(x_3; \theta_3)$
4. ...

Online optimization

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2.
   1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $\ell(x_2; \theta_2)$
3.
   1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,
   3. incur loss $\ell(x_3; \theta_3)$
4. ...

Objective :

Online optimization

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2.
   1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $\ell(x_2; \theta_2)$
3.
   1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,
   3. incur loss $\ell(x_3; \theta_3)$
4. ...

Objective : make sure that we learn to predict well **as fast as possible**.

Online optimization

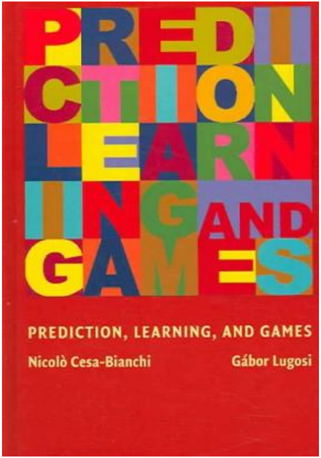
1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2.
   1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $\ell(x_2; \theta_2)$
3.
   1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,
   3. incur loss $\ell(x_3; \theta_3)$
4. ...

Objective : make sure that we learn to predict well **as fast as possible**. Keep

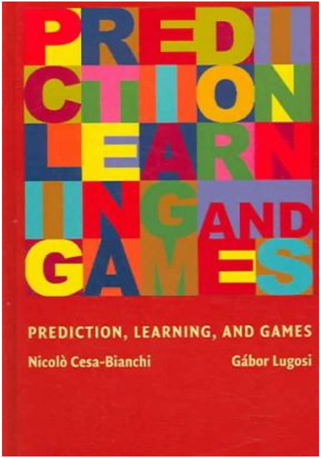
$$\sum_{t=1}^{T} \ell(x_t; \theta_t)$$

as small as possible for any T, **without stochastic assumptions on the data.**

Reference


PREDICTION LEARNING AND GAMES
PREDICTION, LEARNING, AND GAMES
Nicolò Cesa-Bianchi
Gábor Lugosi

Reference



The regret :

$$\mathrm{R}(T) = \sum_{t=1}^{T} \ell(x_t;\theta_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{T} \ell(x_t;\theta).$$

EWA strategy / multipicative update...

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.
- initialize $p_1 = \pi$ (the prior).

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.
- initialize $p_1 = \pi$ (the prior).
- loss $\ell(x;\theta) = -\log p_\theta(x)$.

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.
- initialize $p_1 = \pi$ (the prior).
- loss $\ell(x;\theta) = -\log p_\theta(x)$.

Algorithm 2 Exponentially Weighted Aggregation

1: **for** $t = 1, 2, \dots$ **do**
2: $\quad \theta_t = \mathbb{E}_{\theta \sim p_t}[\theta]$,
3: $\quad x_t$ revealed, update $p_{t+1}(\mathrm{d}\theta) = \frac{[p_\theta(x_t)]^\alpha p_t(\mathrm{d}\theta)}{\int [p_\vartheta(x_t)]^\alpha p_t(\mathrm{d}\vartheta)}$.
4: **end for**

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.
- initialize $p_1 = \pi$ (the prior).
- loss $\ell(x;\theta) = -\log p_\theta(x)$.

Algorithm 2 Exponentially Weighted Aggregation

1: **for** $t = 1, 2, \ldots$ **do**
2: $\theta_t = \mathbb{E}_{\theta \sim p_t}[\theta]$,
3: x_t revealed, update $p_{t+1}(\mathrm{d}\theta) = \frac{[p_\theta(x_t)]^\alpha p_t(\mathrm{d}\theta)}{\int [p_\vartheta(x_t)]^\alpha p_t(\mathrm{d}\vartheta)}$.
4: **end for**

- EWA algorithm has strong theoretical guarantees.

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.
- initialize $p_1 = \pi$ (the prior).
- loss $\ell(x;\theta) = -\log p_\theta(x)$.

Algorithm 2 Exponentially Weighted Aggregation

1: **for** $t = 1, 2, \dots$ **do**
2: $\quad \theta_t = \mathbb{E}_{\theta \sim p_t}[\theta]$,
3: $\quad x_t$ revealed, update $p_{t+1}(\mathrm{d}\theta) = \frac{[p_\theta(x_t)]^\alpha p_t(\mathrm{d}\theta)}{\int [p_\vartheta(x_t)]^\alpha p_t(\mathrm{d}\vartheta)}$.
4: **end for**

- EWA algorithm has strong theoretical guarantees.
- We recover the tempered posterior $p_t = \pi_{t,\alpha}$.

A regret bound for EWA

From now, $\theta \mapsto [-\log p_\theta(x_t)]$ is convex + bounded : $|\cdot| \leq C$.

A regret bound for EWA

From now, $\theta \mapsto [-\log p_\theta(x_t)]$ is convex + bounded : $|\cdot| \leq C$.

Theorem

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_p \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim p}[-\log p_\theta(x_t)] + \frac{\alpha C^2 T}{2} + \frac{KL(p,\pi)}{\alpha} \right\}.$$

A regret bound for EWA

From now, $\theta \mapsto [-\log p_\theta(x_t)]$ is convex + bounded : $|\cdot| \leq C$.

Theorem

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_p \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim p}[-\log p_\theta(x_t)] + \frac{\alpha C^2 T}{2} + \frac{KL(p,\pi)}{\alpha} \right\}.$$

Under similar assumptions than in the batch case, that is, the prior gives enough mass to relevant θ, and $\alpha \sim 1/\sqrt{T}$,

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_{\theta\in\Theta} \sum_{t=1}^{T}[-\log p_\theta(x_t)] + \mathrm{cst.}\sqrt{T}$$

$$\sum_{t=1}^{T} [-\log p_{\theta_t}(x_t)] \leq \inf_{\theta \in \Theta} \sum_{t=1}^{T} [-\log p_{\theta}(x_t)] + \text{cst.}\sqrt{T}$$

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_{\theta\in\Theta} \sum_{t=1}^{T}[-\log p_{\theta}(x_t)] + \mathrm{cst}.\sqrt{T}$$

Assuming that $x_1, \ldots, x_T$ are actually i.i.d from Q, with density q, define

$$\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T}\theta_t,$$

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_{\theta\in\Theta} \sum_{t=1}^{T}[-\log p_{\theta}(x_t)] + \mathrm{cst}.\sqrt{T}$$

Assuming that $x_1, \ldots, x_T$ are actually i.i.d from Q, with density q, define

$$\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T}\theta_t,$$

we have ("online-to-batch" conversion) :

$$\mathbb{E}\left[KL\left(Q, P_{\hat{\theta}_T}\right)\right] \leq \inf_{\theta\in\Theta} KL\left(Q, P_{\theta}\right) + \frac{\mathrm{cst}}{\sqrt{T}}.$$

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_{\theta\in\Theta} \sum_{t=1}^{T}[-\log p_{\theta}(x_t)] + \mathrm{cst}.\sqrt{T}$$

Assuming that $x_1, \ldots, x_T$ are actually i.i.d from Q, with density q, define

$$\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T}\theta_t,$$

we have ("online-to-batch" conversion) :

$$\mathbb{E}\left[KL\left(Q, P_{\hat{\theta}_T}\right)\right] \leq \inf_{\theta\in\Theta} KL\left(Q, P_{\theta}\right) + \frac{\mathrm{cst}}{\sqrt{T}}.$$

Drawback : θ_t is not tractable !

Variational approximations of EWA

B.-E. Chérief-Abdellatif, P. Alquier & M. E. Khan. A Generalization Bound for Online Variational Inference. *Proceedings of ACML*, 2019.

Variational approximations of EWA

B.-E. Chérief-Abdellatif, P. Alquier & M. E. Khan. A Generalization Bound for Online Variational Inference. *Proceedings of ACML*, 2019.

Parametric variational approximation :

$$\mathcal{F} = \{q_\mu, \mu \in M\}.$$

Objective : propose a way to update $\mu_t \to \mu_{t+1}$ so that q_{μ_t} leads to similar performances as p_t in EWA...

SVA and SVB strategies

Algorithm 3 SVA (Sequential Variational Approximation)

1: **for** $t = 1, 2, \ldots$ **do**
2: $\quad \theta_t = \mathbb{E}_{\theta \sim q_{\mu_t}}[\theta]$,
3: $\quad x_t$ revealed, update

$$\mu_{t+1} = \arg\min_{\mu \in M} \left\{ \sum_{i=1}^{t} \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_i)] + \frac{KL(q_\mu, \pi)}{\alpha} \right\}.$$

4: **end for**

SVA and SVB strategies

Algorithm 3 SVA (Sequential Variational Approximation)

1: **for** $t = 1, 2, \ldots$ **do**
2: $\quad \theta_t = \mathbb{E}_{\theta \sim q_{\mu_t}}[\theta]$,
3: $\quad x_t$ revealed, update

$$\mu_{t+1} = \arg\min_{\mu \in M} \left\{ \sum_{i=1}^{t} \mu^T \nabla_{\mu=\mu_i} \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_i)] + \frac{KL(q_\mu, \pi)}{\alpha} \right\}.$$

4: **end for**

SVA and SVB strategies

Algorithm 3 SVA (Sequential Variational Approximation)

1: **for** $t = 1, 2, \dots$ **do**
2: $\quad \theta_t = \mathbb{E}_{\theta \sim q_{\mu_t}}[\theta]$,
3: $\quad x_t$ revealed, update

$$\mu_{t+1} = \arg\min_{\mu \in M} \left\{ \mu^T \nabla_\mu \sum_{i=1}^{t} \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_i)] + \frac{KL(q_\mu, \pi)}{\alpha} \right\}.$$

4: **end for**

SVB (Streaming Variational Bayes) has update

$$\mu_{t+1} = \arg\min_{\mu \in M} \left\{ \mu^T \nabla_\mu \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_t)] + \frac{KL(q_\mu, q_{\mu_t})}{\alpha} \right\}.$$

NGVI strategy

NGVI (Natural Gradient Variational Inference) : fix some $\beta > 0$,

$$\mu_{t+1} = \arg\min_{\mu \in M} \left\{ \mu^T \nabla_\mu \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_t)] + \frac{KL(q_\mu, \pi)}{\alpha} + \frac{KL(q_\mu, q_{\mu_t})}{\beta} \right\}.$$

NGVI strategy

NGVI (Natural Gradient Variational Inference) : fix some $\beta > 0$,

$$\mu_{t+1} = \arg\min_{\mu \in M} \left\{ \mu^T \nabla_\mu \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_t)] + \frac{KL(q_\mu, \pi)}{\alpha} + \frac{KL(q_\mu, q_{\mu_t})}{\beta} \right\}.$$

M. E. Khan & W. Lin. Conjugate-computation variational inference : Converting variational inference in non-conjugate models to inferences in conjugate models. *AISTAT*, 2017.

An example : SVB with Gaussian approximations

As an example, assume that $\theta \in \mathbb{R}^d$, the prior is $\pi = \mathcal{N}(0, s^2 I)$ and that we use the variational approximation family : $q_\mu = q_{m,\sigma} = \mathcal{N}\left(m, \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{pmatrix}\right)$.

An example : SVB with Gaussian approximations

As an example, assume that $\theta \in \mathbb{R}^d$, the prior is $\pi = \mathcal{N}(0, s^2 I)$ and that we use the variational approximation family : $q_\mu = q_{m,\sigma} = \mathcal{N}\left(m, \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{pmatrix}\right)$.

In this case, the update in SVB is :

$$
\begin{aligned}
m_{t+1} &= m_t - \alpha\sigma_t^2 \odot \nabla_{m=m_t} \mathbb{E}_{\theta\sim q_{m,\sigma_t}}[-\log p_\theta(x_t)] \\
\sigma_{t+1} &= \sigma_t \odot h\left(\frac{\alpha\sigma_t \nabla_{\sigma=\sigma_t} \mathbb{E}_{\theta\sim q_{m_t,\sigma}}[-\log p_\theta(x_t)]}{2}\right)
\end{aligned}
$$

where $\odot$ means "componentwise multiplication" and $h(x) = \sqrt{1+x^2} - x$ is also applied componentwise.

An example : SVB with Gaussian approximations

As an example, assume that $\theta \in \mathbb{R}^d$, the prior is $\pi = \mathcal{N}(0, s^2 I)$ and that we use the variational approximation family : $q_\mu = q_{m,\sigma} = \mathcal{N}\left(m, \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{pmatrix} \right)$.

In this case, the update in SVB is :

$$
\begin{aligned}
m_{t+1} &= m_t - \alpha \sigma_t^2 \odot \nabla_{m=m_t} \mathbb{E}_{\theta \sim q_{m,\sigma_t}} [-\log p_\theta(x_t)] \\
\sigma_{t+1} &= \sigma_t \odot h\left(\frac{\alpha \sigma_t \nabla_{\sigma=\sigma_t} \mathbb{E}_{\theta \sim q_{m_t,\sigma}} [-\log p_\theta(x_t)]}{2} \right)
\end{aligned}
$$

where $\odot$ means "componentwise multiplication" and $h(x) = \sqrt{1+x^2} - x$ is also applied componentwise. We also have explicit formulas for SVA and NGVI (see the paper).

A regret bound for SVA

Theorem (C.A., Alquier & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex.

A regret bound for SVA

Theorem (C.A., Alquier & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex. (this is for example the case as soon as the log-likelihood is concave in θ and L-Lipschitz, and μ is a location-scale parameter).

A regret bound for SVA

Theorem (C.A., Alquier & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex. Assume that $\mu \mapsto KL(q_\mu, \pi)$ is γ-strongly convex. Then SVA satisfies :

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)]$$

$$\leq \inf_{\mu\in M}\left\{\mathbb{E}_{\theta\sim q_\mu}\left[\sum_{t=1}^{T}[-\log p_\theta(x_t)]\right] + \frac{\alpha L^2 T}{\gamma} + \frac{KL(q_\mu,\pi)}{\alpha}\right\}.$$

A regret bound for SVA

Theorem (C.A., Alquier & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex. Assume that $\mu \mapsto KL(q_\mu, \pi)$ is γ-strongly convex. Then SVA satisfies :

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)]$$

$$\leq \inf_{\mu\in M}\left\{\mathbb{E}_{\theta\sim q_\mu}\left[\sum_{t=1}^{T}[-\log p_\theta(x_t)]\right] + \frac{\alpha L^2 T}{\gamma} + \frac{KL(q_\mu,\pi)}{\alpha}\right\}.$$

For SVB : some results in the Gaussian case.

A regret bound for SVA

Theorem (C.A., Alquier & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex. Assume that $\mu \mapsto KL(q_\mu, \pi)$ is γ-strongly convex. Then SVA satisfies :

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)]$$
$$\leq \inf_{\mu\in M}\left\{\mathbb{E}_{\theta\sim q_\mu}\left[\sum_{t=1}^{T}[-\log p_\theta(x_t)]\right] + \frac{\alpha L^2 T}{\gamma} + \frac{KL(q_\mu,\pi)}{\alpha}\right\}.$$

For SVB : some results in the Gaussian case. For NGVI : we were not able to derive regret bounds until now.

Test on the Forest Cover Type dataset

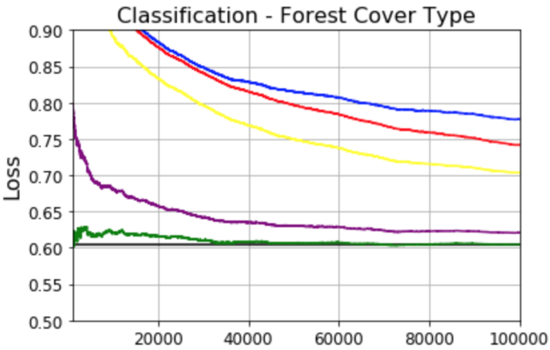


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Boston Housing dataset

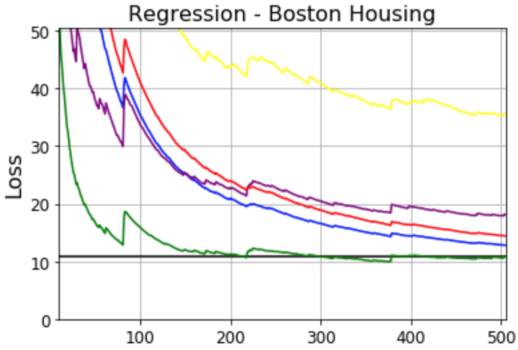


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Conclusions

1. Using online-to-batch conversion, we now have algorithms for variational inference with provable statistical properties after a finite number of steps.

Conclusions

1. Using online-to-batch conversion, we now have algorithms for variational inference with provable statistical properties after a finite number of steps.
2. SVA, SVB competitive with OGA (online gradient algorithm, "non-Bayesian").

Conclusions

1. Using online-to-batch conversion, we now have algorithms for variational inference with provable statistical properties after a finite number of steps.
2. SVA, SVB competitive with OGA (online gradient algorithm, "non-Bayesian").
3. NGVI is the best method on all datasets. Its theoretical analysis is thus an important open problem. Cannot be done with our current techniques (using natural parameters in exponential models lead to non-convex objectives).

Thank you !