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*.

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 $\mathcal{F}$:

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

We have the equivalent definition :

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

with

$$\mathrm{ELBO}(\rho) = -\alpha \int \frac{1}{n} \sum_{i=1}^{n} \log p_{\theta}(X_i) \rho(\mathrm{d}\theta) + KL(\rho, \pi).$$

Outline of the talk

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

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

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 questions :

Can we define a sequential update for variational approximations ? What about the theoretical guarantees ?

Outline of the talk

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 questions :

Can we define a sequential update for variational approximations ? What about the theoretical guarantees ?

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

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

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

Consistency at rate r_n

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

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

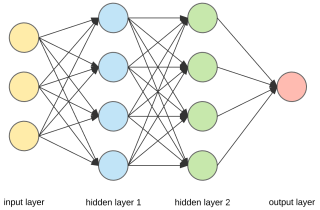
Nonparametric regression

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

Nonparametric regression & Deep Neural Networks

Nonparametric regression

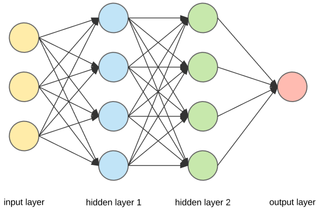
- $X_i \sim \mathcal{U}([-1,1]^d)$,
- $Y_i = f_0(X_i) + \zeta_i$,
- $\zeta_i \sim \mathcal{N}(0, \sigma^2)$.



Nonparametric regression & Deep Neural Networks

Nonparametric regression

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



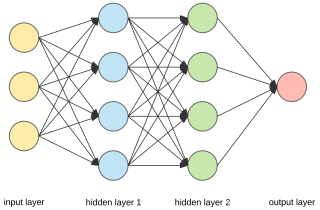
Deep neural networks

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

Nonparametric regression & Deep Neural Networks

Nonparametric regression

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



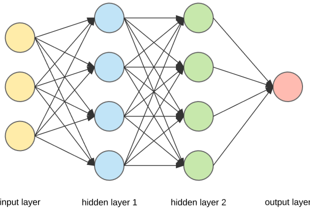
Deep neural networks

- 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.

1. Basics of variational inference
   - Tempered posteriors
   - Variational approximations
   - Challenges in VI theory
2. Consistency of variational inference
   - Posterior consistency
   - Theoretical results
   - Example
3. Online variational inference algorithms
   - **Bayes & online learning**
   - Online variational inference
   - Simulations

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 $\theta_1 \rightarrow \theta_2$,
   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

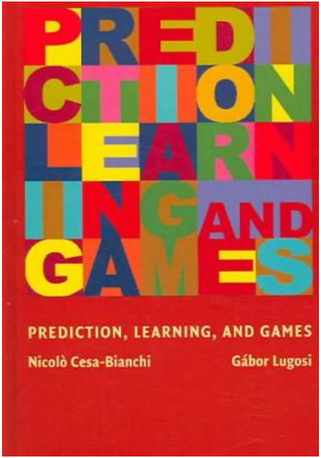
1. 1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $\ell(x_1; \theta_1)$
2. 1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $\ell(x_2; \theta_2)$
3. 1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,
   3. incur loss $\ell(x_3; \theta_3)$
4. ...

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

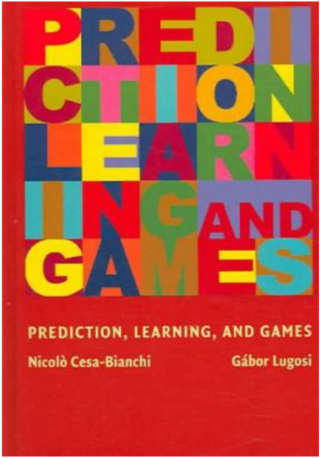
$$\sum_{t=1}^{T} \ell(x_t; \theta_t)$$

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 :

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

Online gradient algorithm (OGA)

Online gradient algorithm (OGA)

- Learning rate α.

Online gradient algorithm (OGA)

- Learning rate α.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.

Online gradient algorithm (OGA)

- Learning rate α.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize θ_1.

Online gradient algorithm (OGA)

- Learning rate α.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize θ_1.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.

Online gradient algorithm (OGA)

- Learning rate α.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize θ_1.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- θ_{t+1} is the solution of :

$$\min_\theta \left\{ \theta^T \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) + \frac{\|\theta - \theta_1\|^2}{2\alpha} \right\}$$

Online gradient algorithm (OGA)

- Learning rate α.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize θ_1.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- θ_{t+1} is the solution of :

$$\min_\theta \left\{ \sum_{s=1}^{t} \ell_s(\theta\ ) \right\}$$

Online gradient algorithm (OGA)

- Learning rate α.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize θ_1.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- θ_{t+1} is the solution of :

$$\min_\theta \left\{ \sum_{s=1}^{t} \ell_s(\theta) + \frac{\|\theta - \theta_1\|^2}{2\alpha} \right\}$$

Online gradient algorithm (OGA)

- Learning rate α.
- Loss $\ell_t(\theta) := \ell(x_t;\theta)$.
- Initialize θ_1.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- θ_{t+1} is the solution of :

$$\min_\theta \left\{ \theta^T \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) + \frac{\|\theta - \theta_1\|^2}{2\alpha} \right\}$$

Online gradient algorithm (OGA)

- Learning rate α.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize θ_1.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- θ_{t+1} is the solution of :

$$\min_\theta \left\{ \theta^T \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) + \frac{\|\theta - \theta_1\|^2}{2\alpha} \right\}$$

Online gradient algorithm (OGA)

- Learning rate α.
- Loss $\ell_t(\theta) := \ell(x_t;\theta)$.
- Initialize θ_1.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- θ_{t+1} is the solution of :

$$\min_\theta \left\{ \theta^T \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) + \frac{\|\theta - \theta_1\|^2}{2\alpha} \right\}$$

and

$$\min_\theta \left\{ \theta^T \nabla_\theta \ell_t(\theta_t) + \frac{\|\theta - \theta_t\|^2}{2\alpha} \right\}.$$

Bayesian learning and VI

Bayesian learning and VI

- Bayesian inference / EWA :

$$\pi_{t+1,\alpha}(\mathrm{d}\theta) \propto \exp\left(-\alpha \sum_{s=1}^{t} \ell_s(x_s) \right) \pi(\mathrm{d}\theta).$$

Bayesian learning and VI

- Bayesian inference / EWA :

$$\pi_{t+1,\alpha}(\mathrm{d}\theta) \propto \exp\left(-\alpha \sum_{s=1}^{t} \ell_s(x_s)\right)\pi(\mathrm{d}\theta).$$

- Not tractable so resort to VI :

$$\begin{aligned}\tilde{\pi}_{t+1,\alpha} &= \arg\min_{q\in\mathcal{F}} KL(q, \pi_{t+1,\alpha}) \\ &= \arg\min_{q\in\mathcal{F}} \left\{ \mathbb{E}_{\theta\sim q}\left[\sum_{s=1}^{t} \ell_s(\theta)\right] + \frac{KL(q,\pi)}{\alpha} \right\}.\end{aligned}$$

Bayesian learning and VI

- Bayesian inference / EWA :

$$\pi_{t+1,\alpha}(\mathrm{d}\theta) \propto \exp\left(-\alpha\sum_{s=1}^{t}\ell_s(x_s)\right)\pi(\mathrm{d}\theta).$$

- Not tractable so resort to VI :

$$\begin{aligned}\tilde{\pi}_{t+1,\alpha} &= \arg\min_{q\in\mathcal{F}} KL(q,\pi_{t+1,\alpha}) \\ &= \arg\min_{q\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim q}\left[\sum_{s=1}^{t}\ell_s(\theta)\right] + \frac{KL(q,\pi)}{\alpha}\right\}.\end{aligned}$$

- Online formula for EWA :

$$\pi_{t+1,\alpha}(\mathrm{d}\theta) \propto \exp\left(-\alpha\ell_t(x_t)\right)\pi_{t,\alpha}(\mathrm{d}\theta).$$

Bayesian learning and VI

- Bayesian inference / EWA :

$$\pi_{t+1,\alpha}(\mathrm{d}\theta) \propto \exp\left(-\alpha\sum_{s=1}^{t}\ell_s(x_s)\right)\pi(\mathrm{d}\theta).$$

- Not tractable so resort to VI :

$$\begin{aligned}\tilde{\pi}_{t+1,\alpha} &= \arg\min_{q\in\mathcal{F}} KL(q,\pi_{t+1,\alpha}) \\ &= \arg\min_{q\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim q}\left[\sum_{s=1}^{t}\ell_s(\theta)\right]+\frac{KL(q,\pi)}{\alpha}\right\}.\end{aligned}$$

- Online formula for EWA :

$$\pi_{t+1,\alpha}(\mathrm{d}\theta) \propto \exp\left(-\alpha\ell_t(x_t)\right)\pi_{t,\alpha}(\mathrm{d}\theta).$$

- Equivalent online formulation for VI ?

A regret bound for EWA

A regret bound for EWA

Theorem

If the loss is bounded by B :

$$\sum_{t=1}^{T} \mathbb{E}_{\theta \sim \pi_{t,\alpha}}[\ell_t(\theta)] \leq \inf_{q} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta \sim q}[\ell_t(\theta)] + \frac{\alpha B^2 T}{8} + \frac{KL(q,\pi)}{\alpha} \right\}.$$

A regret bound for EWA

Theorem

If the loss is bounded by B :

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_{t,\alpha}}[\ell_t(\theta)] \leq \inf_{q} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q}[\ell_t(\theta)] + \frac{\alpha B^2 T}{8} + \frac{KL(q,\pi)}{\alpha} \right\}.$$

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

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_{t,\alpha}}[\ell_t(\theta)] \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + \mathcal{O}\left(\sqrt{dT\log(T)}\right)$$

A regret bound for EWA

Theorem

If the loss is bounded by B :

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_{t,\alpha}}[\ell_t(\theta)] \leq \inf_q \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q}[\ell_t(\theta)] + \frac{\alpha B^2 T}{8} + \frac{KL(q,\pi)}{\alpha} \right\}.$$

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

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_{t,\alpha}}[\ell_t(\theta)] \leq \inf_\theta \sum_{t=1}^{T} \ell_t(\theta) + \mathcal{O}\left(\sqrt{dT\log(T)}\right)$$

Equivalent regret bounds for VI ?

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 $\pi_{t,\alpha}$ in EWA...

SVA and SVB strategies

- SVA (Sequential Variational Approximation) :

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

SVA and SVB strategies

- SVA (Sequential Variational Approximation) :

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

SVA and SVB strategies

- SVA (Sequential Variational Approximation) :

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

- SVB (Streaming Variational Bayes) :

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

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 : $q_\mu = q_{m,\sigma} = \mathcal{N}\left(m, \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{pmatrix}\right)$.

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 :

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

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 :

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

where $\odot$ means "componentwise multiplication" and $h(x) = \sqrt{1+x^2} - x$ is also applied componentwise. We also have a similar formula for SVA.

A regret bound for SVA

Theorem (C.A., Alquier & Khan)

Assume that the expected loss is L-Lipschitz and convex.

A regret bound for SVA

Theorem (C.A., Alquier & Khan)

Assume that the expected loss is L-Lipschitz and convex. (this is for example the case as soon as the loss is convex in θ and L-Lipschitz, and μ is a location-scale parameter).

A regret bound for SVA

Theorem (C.A., Alquier & Khan)

Assume that the expected loss is L-Lipschitz and convex. Assume that $\mu \mapsto KL(q_\mu, \pi)$ is γ-strongly convex. Then SVA satisfies :

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

A regret bound for SVA

Theorem (C.A., Alquier & Khan)

Assume that the expected loss is L-Lipschitz and convex. Assume that $\mu \mapsto KL(q_\mu, \pi)$ is γ-strongly convex. Then SVA satisfies :

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

Application to Gaussian approximation leads to :

$$\sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + (1 + o(1)) \frac{2L}{\gamma} \sqrt{dT \log(T)}.$$

For SVB : some results in the Gaussian case.

Test on the Forest Cover Type dataset

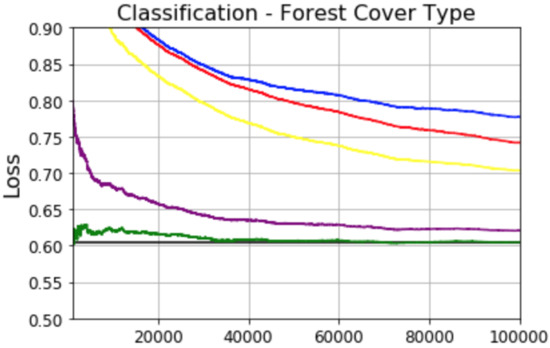


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Boston Housing dataset

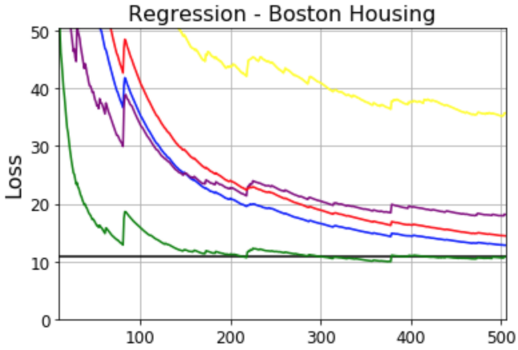


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Final remarks (1)

Using online-to-batch conversion, we can have algorithms for variational inference with provable statistical properties.

$$\sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + (1 + o(1)) \frac{2L}{\gamma} \sqrt{dT \log(T)}.$$

Final remarks (1)

Using online-to-batch conversion, we can have algorithms for variational inference with provable statistical properties.

$$\sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + (1 + o(1)) \frac{2L}{\gamma} \sqrt{dT \log(T)}.$$

Assuming that $x_1, \ldots, x_T$ are actually i.i.d from Q with density q, define $\hat{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \theta_t$ for the loss $\ell_t(\theta) := -\log p_\theta(x_t)$,

Final remarks (1)

Using online-to-batch conversion, we can have algorithms for variational inference with provable statistical properties.

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + (1+o(1))\frac{2L}{\gamma}\sqrt{dT\log(T)}.$$

Assuming that $x_1,\dots,x_T$ are actually i.i.d from Q with density q, define $\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T}\theta_t$ for the loss $\ell_t(\theta) := -\log p_\theta(x_t)$,

$$\mathbb{E}\left[KL\left(Q, P_{\hat{\theta}_T}\right)\right] \leq \inf_{\theta\in\Theta} KL\left(Q, P_\theta\right) + (1+o(1))\frac{2L}{\gamma}\sqrt{\frac{d\log(T)}{T}}.$$

Final remarks (2)

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

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

Final remarks (2)

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

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

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

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 !