# Contributions to the theoretical study of variational inference and robustness

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

CENTER FOR RESEARCH IN ECONOMICS AND STATISTICS

PhD Defense
June 23, 2020

1. Consistency of variational inference
   - Variational inference
   - Theoretical results
   - Examples
2. Online variational inference algorithms
   - Bayes & online learning
   - Online variational inference
   - Simulations
3. Robust MMD-based estimation
   - Robustness in statistics
   - MMD-based estimation
   - MMD-Bayes estimator

# Notations

Assume that we observe $X_1, \dots, X_n$ i.i.d from $P_0 = P_{\theta_0}$ in a model $\{P_\theta, \theta \in \Theta\}$ with likelihood $L_n(\theta)$. Prior $\pi$ on $\Theta$.

**The posterior**

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

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

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

# Notations

Assume that we observe $X_1, \dots, X_n$ i.i.d from $P_0 = P_{\theta_0}$ in a model $\{P_\theta, \theta \in \Theta\}$ with likelihood $L_n(\theta)$. Prior $\pi$ on $\Theta$.

**The posterior**

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

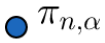
**The tempered posterior - $0 < \alpha < 1$**

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

**Computation of the posterior**

The classical MCMC algorithms may be slow when both the model dimension and the sample size are large. A more and more popular alternative : variational inference.

# Variational approximations : definition



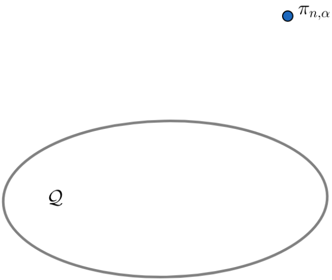

Idea of VB : chose a family $\mathcal{Q}$ of probability distributions on $\Theta$ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{Q}$ :

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

# Variational approximations : definition

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

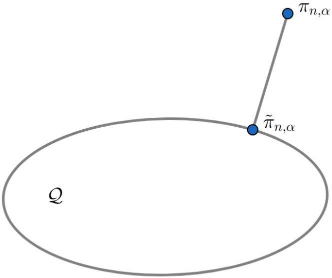
$$\tilde{\pi}_{n,\alpha} := \arg\min_{q \in \mathcal{Q}} KL(q, \pi_{n,\alpha}).$$



# Variational approximations : definition

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

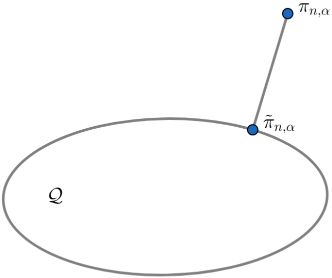
$$\tilde{\pi}_{n,\alpha} := \arg\min_{q \in \mathcal{Q}} KL(q, \pi_{n,\alpha}).$$



# Variational approximations : definition

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

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




Examples of sets $\mathcal{Q}$ :

- parametric ($\Theta \subset \mathbb{R}^d$) :

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

- mean-field ($\Theta = \Theta_1 \times \Theta_2$) :

$$q(\mathrm{d}\theta) = q_1(\mathrm{d}\theta_1) \times q_2(\mathrm{d}\theta_2).$$

# Tools for the consistency of VB

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

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

For $1/2 \leq \alpha$, link with Hellinger and Kullback :

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

# Tools for the consistency of VB

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

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

For $1/2 \leq \alpha$, link with Hellinger and Kullback :

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

**Consistency at rate $r_n$**

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq r_n \xrightarrow[n \to \infty]{} 0.$$

# Technical condition for posterior consistency

## Prior mass condition for consistency 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 consistency of Variational Bayes

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

$$\int KL(P_{\theta_0}, P_\theta) q_n(\mathrm{d}\theta) \leq r_n, \text{ and } KL(q_n, \pi) \leq nr_n.$$

# Consistency of the approximate posterior

**Theorem**

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

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

**Theorem**

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**

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**

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 $\sim P_0 \notin \{P_\theta, \theta \in \Theta\}$.

# Misspecified case

**Theorem**

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 $\sim P_0 \notin \{P_\theta, \theta \in \Theta\}$.

**Theorem**

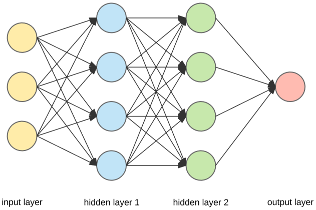
Under a similar condition, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_0)\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{\alpha}{1-\alpha} \inf_\theta KL(P_0, P_\theta) + \frac{1+\alpha}{1-\alpha} r_n.$$

# Nonparametric regression & Deep Neural Networks

## Nonparametric regression

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




## 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**

Chose spike-and-slab prior and variational set on $\theta$. 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

Chose spike-and-slab prior and variational set on $\theta$. 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 $f_0$ $\beta$-Hölder for suitable $(S, L, D)$ : $\quad \tilde{\mathcal{O}}\left(n^{-\frac{2\beta}{2\beta+d}}\right).$

# Related publications

- B.-E. C.-A., P. Alquier. Consistency of Variational Bayes Inference for Estimation and Model Selection in mixtures. *Electronic Journal of Statistics*, 2018.

- B.-E. C.-A. Consistency of ELBO Maximization for Model Selection. *Proceedings of AABI*, 2019.

- B.-E. C.-A. Convergence Rates of Variational Inference in Sparse Deep Learning. *Accepted at ICML*, 2020.

# Online learning

## Objective

Make sure that we learn to predict well **as fast as possible.**

# Online learning

## Objective

Make sure that we learn to predict well **as fast as possible.** Keep, **without stochastic assumptions on the data**, as small as possible for any $T$ :

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

## 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 learning

## Objective

Make sure that we learn to predict well **as fast as possible.** Keep, **without stochastic assumptions on the data**, as small as possible for any $T$ :

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

## The regret

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

What strategy can lead to a low regret ?

# Online gradient algorithm (OGA)

# Online gradient algorithm (OGA)

- Learning rate $\alpha$.

# Online gradient algorithm (OGA)

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

# Online gradient algorithm (OGA)

- Learning rate $\alpha$.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize $\theta_1$.

# Online gradient algorithm (OGA)

- Learning rate $\alpha$.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize $\theta_1$.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.

# Online gradient algorithm (OGA)

- Learning rate $\alpha$.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize $\theta_1$.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- $\theta_{t+1}$ 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 $\alpha$.
- Loss $\ell_t(\theta) := \ell(x_t;\theta)$.
- Initialize $\theta_1$.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- $\theta_{t+1}$ is the solution of :

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

# Online gradient algorithm (OGA)

- Learning rate $\alpha$.
- Loss $\ell_t(\theta) := \ell(x_t;\theta)$.
- Initialize $\theta_1$.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- $\theta_{t+1}$ 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 $\alpha$.
- Loss $\ell_t(\theta) := \ell(x_t;\theta)$.
- Initialize $\theta_1$.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- $\theta_{t+1}$ 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 $\alpha$.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize $\theta_1$.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- $\theta_{t+1}$ 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 $\alpha$.
- Loss $\ell_t(\theta) := \ell(x_t; \theta)$.
- Initialize $\theta_1$.
- Update $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \ell_t(\theta_t)$.
- $\theta_{t+1}$ 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 :

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

# Bayesian learning and VI

- Bayesian inference / EWA :

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

- Online formula for EWA :

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

# Bayesian learning and VI

- Bayesian inference / EWA :

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

- Online formula for EWA :

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

- Not tractable so resort to VI :

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

# Bayesian learning and VI

- Bayesian inference / EWA :

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

- Online formula for EWA :

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

- Not tractable so resort to VI :

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

- 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 $\theta$, 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 $\theta$, 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. C.-A., P. Alquier & M. E. Khan. A Generalization Bound for Online Variational Inference. *Proceedings of ACML*, 2019.

# Variational approximations of EWA

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

Parametric variational approximation :

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

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

# SVA and SVB strategies

- SVA (Sequential Variational Approximation) :

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

- SVB (Streaming Variational Bayes) :

# 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) :

# 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.

# A regret bound for SVA

**Theorem**

Assume that the expected loss $\mu \to \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$ is $L$-Lipschitz and convex.

# A regret bound for SVA

**Theorem**

Assume that the expected loss $\mu \to \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$ is $L$-Lipschitz and convex. (this is for example the case as soon as the loss $\ell_t(\theta)$ is convex in $\theta$ and $L$-Lipschitz, and $\mu$ is a location-scale parameter).

# A regret bound for SVA

**Theorem**

Assume that the expected loss $\mu \to \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$ is $L$-Lipschitz and convex. Assume that $\mu \mapsto KL(q_\mu, \pi)$ is $\gamma$-strongly convex. Then SVA satisfies :

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

# A regret bound for SVA

**Theorem**

Assume that the expected loss $\mu \to \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$ is $L$-Lipschitz and convex. Assume that $\mu \mapsto KL(q_\mu, \pi)$ is $\gamma$-strongly convex. Then SVA satisfies :

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

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

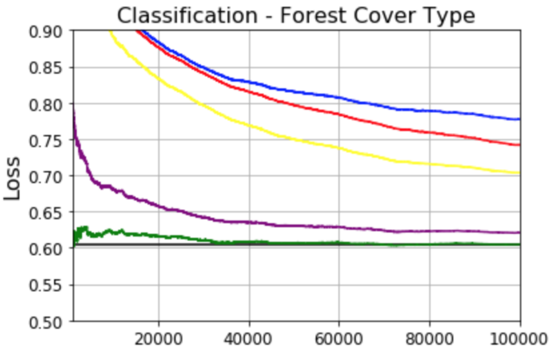



Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

# Test on the Boston Housing dataset

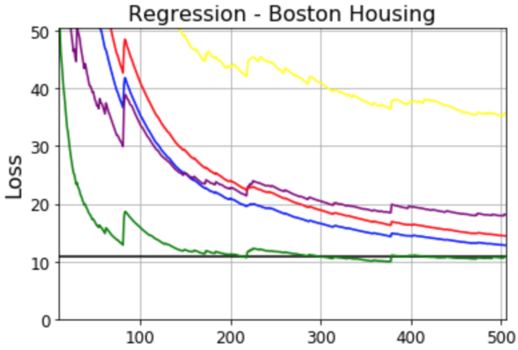



Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

# What is a “robust” estimator ?

A robust estimator $\hat{\theta}_n$ must be such that, for some distance $d$ on probability distributions,

# What is a "robust" estimator ?

A robust estimator $\hat{\theta}_n$ must be such that, for some distance $d$ on probability distributions,

1. when the model is well specified, that is, $P_0 = P_{\theta_0}$,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq r_n(\Theta) \xrightarrow[n\to\infty]{} 0.$$

# What is a "robust" estimator ?

A robust estimator $\hat{\theta}_n$ must be such that, for some distance $d$ on probability distributions,

1. when the model is well specified, that is, $P_0 = P_{\theta_0}$,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq r_n(\Theta) \xrightarrow[n\to\infty]{} 0.$$

2. in the misspecified case $P_0 = (1-\varepsilon)P_{\theta_0} + \varepsilon Q$, for any Q,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq c.\underbrace{d(P_0, P_{\theta_0})}_{\xrightarrow[\varepsilon\to 0]{} 0} + \underbrace{r_n(\Theta)}_{\xrightarrow[n\to\infty]{} 0}.$$

# What is a "robust" estimator ?

A robust estimator $\hat{\theta}_n$ must be such that, for some distance $d$ on probability distributions,

1. when the model is well specified, that is, $P_0 = P_{\theta_0}$,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq r_n(\Theta) \xrightarrow[n\to\infty]{} 0.$$

2. in the misspecified case $P_0 = (1-\varepsilon)P_{\theta_0} + \varepsilon Q$, for any Q,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq c.\underbrace{d(P_0, P_{\theta_0})}_{\xrightarrow[\varepsilon\to 0]{} 0} + \underbrace{r_n(\Theta)}_{\xrightarrow[n\to\infty]{} 0}.$$

Many popular estimators in statistics such as MLE do not satisfy these requirements in some settings.

# A typical example

Yatracos' skeleton estimate $\hat{\theta}_n^Y$ :

$$\mathbb{E}\left[d_{TV}(P_{\hat{\theta}_n^Y}, P_0)\right] \leq 3 d_{TV}(P_0, P_{\theta_0}) + C.\sqrt{\frac{\dim(\Theta)}{n}}$$

where

$$d_{TV}(P, Q) = \sup_E |P(E) - Q(E)|.$$

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*.

# A typical example

Yatracos' skeleton estimate $\hat{\theta}_n^Y$ :

$$\mathbb{E}\left[d_{TV}(P_{\hat{\theta}_n^Y}, P_0)\right] \leq 3 d_{TV}(P_0, P_{\theta_0}) + C.\sqrt{\frac{\dim(\Theta)}{n}}$$

where

$$d_{TV}(P, Q) = \sup_E |P(E) - Q(E)|.$$

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*.

But it cannot be computed in practice.

# A typical example

Yatracos' skeleton estimate $\hat{\theta}_n^Y$ :

$$\mathbb{E}\left[d_{TV}(P_{\hat{\theta}_n^Y}, P_0)\right] \leq 3 d_{TV}(P_0, P_{\theta_0}) + C . \sqrt{\frac{\dim(\Theta)}{n}}$$

where

$$d_{TV}(P,Q) = \sup_E |P(E) - Q(E)|.$$

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*.

But it cannot be computed in practice.

Additional requirement : an estimator must be tractable ! ! !

# Maximum Mean Discrepancy

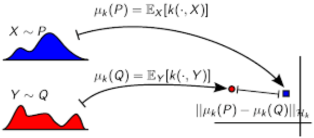
We consider a bounded p.d. kernel : $0 \leq k(x, y) \leq 1.$

# Maximum Mean Discrepancy

We consider a bounded p.d. kernel : $0 \leq k(x, y) \leq 1.$

**Kernel mean embedding**

$$\mu_k(P) = \mathbb{E}_{X \sim P}[k(\cdot, X)] \in \mathcal{H}_k.$$

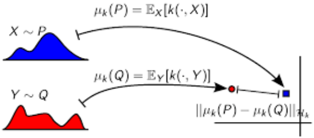


# Maximum Mean Discrepancy

We consider a bounded p.d. kernel : $0 \leq k(x, y) \leq 1$.

**Kernel mean embedding**

$\mu_k(P) = \mathbb{E}_{X \sim P}[k(\cdot, X)] \in \mathcal{H}_k$.



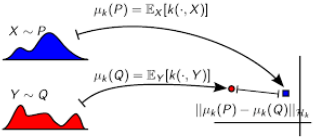

The kernel $k$ is characteristic (i.e. $\mu_k(\cdot)$ is injective).
Example : $k(x, y) = \exp(-\frac{\|x-y\|^2}{\gamma^2})$ is a characteristic kernel.

# Maximum Mean Discrepancy

We consider a bounded p.d. kernel : $0 \leq k(x,y) \leq 1$.

**Kernel mean embedding**

$\mu_k(P) = \mathbb{E}_{X\sim P}[k(\cdot,X)] \in \mathcal{H}_k$.




The kernel $k$ is characteristic (i.e. $\mu_k(\cdot)$ is injective).
Example : $k(x,y) = \exp(-\frac{\|x-y\|^2}{\gamma^2})$ is a characteristic kernel.

**Definition : the MMD distance**

$$\mathbb{D}_k(P,Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}.$$

# MMD-based estimator

$X_1, \ldots, X_n$ be i.i.d in $\mathcal{X}$ from a probability distribution $P_0$, model $\{P_\theta, \theta \in \Theta\}$, bounded p.d. kernel $0 \leq k(x,y) \leq 1$.

## Definition - MMD based estimator

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \mathbb{D}_k \left( P_\theta, \hat{P}_n \right) \text{ where } \hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

# MMD-based estimator

$X_1, \ldots, X_n$ be i.i.d in $\mathcal{X}$ from a probability distribution $P_0$, model $\{P_\theta, \theta \in \Theta\}$, bounded p.d. kernel $0 \leq k(x,y) \leq 1$.

## Definition - MMD based estimator

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, \hat{P}_n\right) \text{ where } \hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

## Theorem

$$\forall P_0, \quad \mathbb{E}\left[\mathbb{D}_k\left(P_{\hat{\theta}_n}, P_0\right)\right] \leq \underbrace{\inf_{\theta\in\Theta} \mathbb{D}_k(P_\theta, P_0)}_{\substack{\leq 2\varepsilon \text{ when} \\ P_\mathbf{0} = (1-\varepsilon)P_{\theta_\mathbf{0}} + \varepsilon Q}} + \frac{2}{\sqrt{n}}.$$

# How to compute $\hat{\theta}_n^{MMD}$ ?

We actually have (up to a constant)

$$\mathbb{D}_k^2(P_\theta, \hat{P}_n) = \mathbb{E}_{X,X' \sim P_\theta}[k(X, X')] - \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}_{X \sim P_\theta}[k(X_i, X)]$$

# How to compute $\hat{\theta}_n^{MMD}$ ?

We actually have (up to a constant)

$$\mathbb{D}_k^2(P_\theta, \hat{P}_n) = \mathbb{E}_{X,X' \sim P_\theta}[k(X, X')] - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{X \sim P_\theta}[k(X_i, X)]$$

and so

$$\nabla_\theta \mathbb{D}_k^2(P_\theta, \hat{P}_n) = 2\mathbb{E}_{X,X' \sim P_\theta}\left\{\left[k(X, X') - \frac{1}{n}\sum_{i=1}^{n} k(X_i, X)\right] \nabla_\theta[\log p_\theta(X)]\right\}$$

that can be approximated by sampling from $P_\theta$.

# Example : Gaussian mean estimation

Example : the model is given by $P_\theta = \mathcal{N}(\theta, \sigma^2)$ for $\theta \in \mathbb{R}$.

# Example : Gaussian mean estimation

Example : the model is given by $P_\theta = \mathcal{N}(\theta, \sigma^2)$ for $\theta \in \mathbb{R}$.

Using a Gaussian kernel $k(x, y) = \exp(-(x-y)^2/2)$, from the previous theorem and from the equality

$$\mathbb{D}_k^2\left(P_\theta, P_{\theta'}\right) = \sqrt{2}\left[1 - \exp\left(-\frac{(\theta-\theta')^2}{4\sigma^2}\right)\right]$$

we obtain

$$\mathbb{E}\left[(\hat{\theta}_n - \theta_0)^2\right] \leq 16\sigma^2\left(\varepsilon^2 + \frac{1}{n}\right).$$

(for $\varepsilon^2 + \frac{1}{n} \leq \frac{1}{4\sqrt{2}}$).

# Example : Gaussian mean estimation, simulations

Model : $\mathcal{N}(\theta, 1)$, and $X_1, \dots, X_n$ i.i.d $\mathcal{N}(\theta_0, 1)$, $n = 100$ and we repeat the experiment 200 times.

| | $\hat{\theta}_n^{MLE}$ | $\hat{\theta}_n^{MMD}$ |
|---|---|---|
| mean absolute error | 0.0722 | 0.0838 |

# Example : Gaussian mean estimation, simulations

Model : $\mathcal{N}(\theta, 1)$, and $X_1, \dots, X_n$ i.i.d $\mathcal{N}(\theta_0, 1)$, $n = 100$ and we repeat the experiment 200 times.

| | $\hat{\theta}_n^{MLE}$ | $\hat{\theta}_n^{MMD}$ |
|---|---|---|
| mean absolute error | 0.0722 | 0.0838 |

Now, $\varepsilon = 2\%$ of the observations drawn from a Cauchy.

| | | |
|---|---|---|
| mean absolute error | 0.2349 | 0.0953 |

# Example : Gaussian mean estimation, simulations

Model : $\mathcal{N}(\theta, 1)$, and $X_1, \ldots, X_n$ i.i.d $\mathcal{N}(\theta_0, 1)$, $n = 100$ and we repeat the experiment 200 times.

| | $\hat{\theta}_n^{MLE}$ | $\hat{\theta}_n^{MMD}$ |
|---|---|---|
| mean absolute error | 0.0722 | 0.0838 |

Now, $\varepsilon = 2\%$ of the observations drawn from a Cauchy.

| | | |
|---|---|---|
| mean absolute error | 0.2349 | 0.0953 |

Now, $\varepsilon = 1\%$ are replaced by 1000.

| | | |
|---|---|---|
| mean absolute error | 10.018 | 0.0903 |

1. Consistency of variational inference
   - Variational inference
   - Theoretical results
   - Examples
2. Online variational inference algorithms
   - Bayes & online learning
   - Online variational inference
   - Simulations
3. Robust MMD-based estimation
   - Robustness in statistics
   - MMD-based estimation
   - MMD-Bayes estimator

# Bayesian MMD-based estimation

Given a prior $\pi(\theta)$ we propose the following pseudo-posterior :

$$\pi_n^{\beta}(d\theta) \propto \exp\left(-\beta \mathbb{D}_k^2(P_\theta, \hat{P}_n)\right) \pi(d\theta).$$

**Theorem**

Let $\mathcal{B} = \{\theta \in \Theta / \mathbb{D}_k\left(P_{\theta_0}, P_\theta\right) \leq 1/\sqrt{n}\}$. Assume $(\pi, \beta)$ satisfies the prior mass condition : $\pi(\mathcal{B}) \geq e^{-\beta/\sqrt{n}}$. Then :

$$\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, P_0\right) \pi_n^{\beta}(\mathrm{d}\theta)\right] \leq 8 \inf_{\theta \in \Theta} \mathbb{D}_k^2\left(P_\theta, P_0\right) + \frac{16}{n}.$$

We also prove similar results for variational approximations, that can be computed by stochastic gradient descent :

$$q_\beta = \operatorname*{arg\,min}_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q}\left[\mathbb{D}_k^2\left(P_\theta, \hat{P}_n\right)\right] + \frac{\mathrm{KL}(q, \pi)}{\beta} \right\}.$$

# Related publications

B.-E. C.-A., P. Alquier. Finite sample properties of parametric MMD estimation : robustness to misspecification and dependence. *Preprint ArXiv*, 2019.

B.-E. C.-A., P. Alquier. MMD-Bayes : Robust Bayesian Estimation via Maximum Mean Discrepancy. *Proceedings of AABI*, 2020.

Thank you !