Minimisation stochastique de divergences et de fonctions

Rouen Juin 2024

Michel Broniatowski[1]

[1]LPSM, Sorbonne Université, Paris

19-21 Juin 2024

joint work with W Stummer, FAU, Erlangen

Contents

Generalities

Problem

Statistics: A class of estimators of

$$\Phi\left(\Omega, P^X\right) := \inf\left\{\Phi\left(\mathbf{Q}, P^X\right), \mathbf{Q} \in \Omega\right\}$$

and minimizers (assuming P^X not in Ω)

Example

$\Omega \subset \mathbb{S}^K$ is a set defined by quantile conditions, moment conditions or a tubular neighborhood of some distribution, etc. Ω with non void interior.

Problem

$\Psi : \mathbb{R}^K \to \mathbb{R}$ *continuous on "regular"* $\Omega \subset \mathbb{R}^K$

$$\Psi\left(\Omega\right) := \inf\left\{\Psi\left(\mathbf{Q}\right), \mathbf{Q} \in \Omega\right\}$$

and minimizers

Bare Simulation Minimizable divergences

$$\Omega \subset \mathbb{R}^K,\ int(\Omega) \text{ non void },\ Cl\left(\Omega\right) = Cl\left(Int\left(\Omega\right)\right)$$

$\Phi_{\mathbf{P}}\left(\mathbf{Q}\right)$ is minimizable is minimizable (upon $\mathbf{Q}$) on Ω iff there exist
A-a function $G : (0, \infty) \to \mathbb{R}$
B- a sequence $\left(\mathcal{X}_n, \mathcal{A}_n, \Pi_n\right)$ of probability spaces and on them a sequence $\mathbb{X}_n$ of $\mathbb{R}^K$ valued r.v.'s

$$G\left(-\lim_{n\to\infty}\frac{1}{n}\log\Pi_n\left(\mathbb{X}_n \in \Omega\right)\right) = \inf_{\mathbf{Q}\in\Omega}\Phi_{\mathbf{P}}\left(\mathbf{Q}\right)$$

Application: Inference in neural networks

Neural network: Trained Fashion MNIST 70000 images, 10 classes.
Two Parts: A convolution network and a dense network.
First part (convolution): 8544 weights
Second part (dense) 412778 weights $K = 412778$
Activation: Relu for all levels, Softmax (10 classes) for the last one
Training provides a set of weights a_0 leading to Prob success $P(a_0) = 0.91$ on test samples

$$\Omega_\eta := \left\{ a \in \mathbb{R}^K : P(a) \geq P(a_0) - \eta \right\}$$

$$\min_{a \in \Omega_\eta} \Phi(a) := \min_{a \in \Omega_\eta} \|a\|_1$$

and find sparse solutions among minimizers.

A remark: statistical divergences and corresponding distributions

W , $\quad E(W) = 1 \qquad E \exp tW$ finite on $V(0)$ (Cramer condition).

$$\varphi^W(x) := \sup_t tx - \log E \exp tW$$

Convex, $\varphi^W(1) = 0$, $\left(\varphi^W\right)'(1) = 0$ Hence divergence function.
W Normal (1,1), $\varphi^W(x) = (x-1)^2$ divergence function χ^2
W Inverse Gaussian (1,1) $\varphi^W(x) = \frac{(x-1)^2}{x}$ divergence function χ_m^2
W Poisson (1) $\varphi^W(x) = x \log x - x + 1$ divergence function KL
W Exp(1) $\varphi^W(x) = -\log x + x - 1$ divergence function Likelihood
etc

Divergence functions

Definition

(a) Let the "divergence-generator" be a convex function $\varphi :]a, b[\to [0, \infty]$ satisfying $\varphi(1) = 0. -\infty \leq a < 1 < b \leq \infty$. Moreover, we suppose that φ is strictly convex in a non-empty neighborhood $]t_-^{sc}, t_+^{sc}[\subseteq]a, b[$ of one $(t_-^{sc} < 1 < t_+^{sc})$. It is convenient to state

ConditionA generator φ should satisfy the representation

$$\varphi(t) = \sup_{z \in \mathbb{R}} \left(z \cdot t - \log \int_{\mathbb{R}} e^{zy} \, dF(y) \right), \qquad t \in \mathbb{R}, \tag{1}$$

for some probability distribution $F = F^W$ on the real line such that the function $z \mapsto MGF_F(z) := \int_{\mathbb{R}} e^{zy} \, dF(y)$ is finite on some open interval containing zero and $EW = 1$..

Examples

Example

A generic class Power generators

$$x \in \mathbb{R}_+^* \mapsto \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)}$$

for $\gamma \in \mathbb{R} \setminus \{0, 1\}$ and $\varphi_0(x) := -\log x + x - 1$, $\varphi_1(x) := x \log x - x + 1$ with $\varphi_\gamma(0) = \lim_{x \downarrow 0} \varphi_\gamma(x)$, $\varphi_\gamma(\infty) = \lim_{x \to \infty} \varphi_\gamma(x)$, for any $\gamma \in \mathbb{R}$. The Kullback-Leibler divergence (KL) : φ_1, the modified Kullback-Leibler (KL_m) : φ_0, the χ^2 divergence : φ_2, the modified χ^2 divergence (χ^2_m) : φ_{-1} and the Hellinger :$\varphi_{1/2}$. Possibly extended on $\mathbb{R}$, for the χ^2.

Cressie-Read (Power) divergences and natural exponential families

Weights W in the twisted family of some NEF (with $EW = 1$)

- For $\gamma < 0$ by stable distributions on $\mathbb{R}^+$ with characteristic exponent in $(0, 1)$. The resulting distributions define the Tweedie scale family (with basis these stable laws) Example in the NEF: Inverse Gaussian $(\gamma = -1/2)$
- For $\gamma = 0$ by the exponential distribution
- For $0 < \gamma < 1$ by Compound Gamma-Poisson distributions
- For $\gamma = 1$ by the Poisson distribution
- For $\gamma = 2$ by the normal distribution

Other values of γ do not yield NEF's , so no W. in that class

- Given φ identify W : Consider φ^* and since $\varphi(x) = \sup_t tx - \log E \exp tW$, $\varphi^*(t) = \log E \exp tW$; identify, if possible (Widder, Bernstein) or a catalogue

Extension of power type divergences (support)
Starting from a prototype divergence, modifying its properties (example : robustness, bounded φ');

- Starting from W and identifying φ

These approaches allow for the wide extension of the power class blended weight chi-square divergence of Lindsay (1994) for vectors

Jensen Shannon, Rényi type, etc Sanghvi's genetic difference measure (1953)

(squared) Vincze-Le Cam distance (cf. Vincze (1981), Le Cam (1986))

triangular discrimination (cf. Topsoe (2000)) etc
see Br- Stummer On a cornerstone of bare simulation optimization, Springer, GSI2023

Building consistent minimum divergence sequences, vector cases

Basic fact

Assume W satisfies Cramer condition and $\varphi(x) := \sup_t tx - \log E \exp tW$.

Theorem

(Cramer) For all $B \subset \mathbb{R}^K$ s.t. $cl(int(B)) = cl(B)$ with W_1 , .., W_n iid copies of W

$$-\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}^W \left(\frac{1}{n} \sum_{i=1}^{n} W_i \in B \right) = \inf_{b \in B} \varphi(b)$$

Question:From this to the minimisation of

$$\Phi_{\mathbf{P}}(\mathbf{Q}) := D_\varphi(\mathbf{Q}, \mathbf{P}) := \sum_{k=1}^{K} p_k \varphi\left(\frac{q_k}{p_k}\right)$$

over Ω and other dissimilarities between vectors in $\mathbb{R}^K$ or on $\mathbb{S}^K$ with $\mathbf{P} := (p_1, .., p_K) \in \mathbb{R}^K_{>0}$

Examples CASM divergences

For $\mathbf{P} := (p_1, .., p_K)$ a vector with positive entries in $\mathbb{R}^K$ and $\mathbf{Q} := (q_1, .., q_K)$ a vector in $\mathbb{R}^K$ define

$$D_\varphi(\mathbf{Q}, \mathbf{P}) := \sum_{k=1}^{K} p_k \varphi\left(\frac{q_k}{p_k}\right)$$

and for Ω a subset in $\mathbb{R}^K$ let

$$D_\varphi(\Omega, \mathbf{P}) := \inf\left\{D_\varphi(\mathbf{Q}, \mathbf{P}), \mathbf{Q} \in \Omega\right\}.$$

Let $M_{\mathbf{P}} := p_1 + .. + p_K$ denote the total mass of $\mathbf{P}$. Since

$$D_\varphi(\mathbf{Q}, \mathbf{P}) = D_{M_{\mathbf{P}}.\varphi}\left(\frac{\mathbf{P}}{M_{\mathbf{P}}}, \frac{\mathbf{Q}}{M_{\mathbf{P}}}\right)$$

we only consider minimization problems for $D_\varphi(\mathbf{Q}, \mathbf{P})$ as $\mathbf{Q} \in \Omega$ for vectors $\mathbf{P}$ with positive entries and total mass 1.

Bregman divergences

Scaled BREGMAN (Stummer, Vajda)

$$D_{\varphi,\mathbf{P}}\left(\mathbf{Q},\mathbf{Q}^{**}\right) := \sum_{k=1}^{K} p_k \varphi_k\left(\frac{q_k}{p_k}\right)$$

$$\varphi_k(t) := \varphi(t) - \varphi(t_k^{**}) - \varphi'(t_k^{**})\,(t - t_k^{**})$$

and

$$t_k^{**} := \frac{q_k^{**}}{p_k} \in \left(t_-^{sc}, t_+^{sc}\right).$$

$$D_{\varphi,\mathbf{P}}\left(\mathbf{Q},\mathbf{Q}^{**}\right) \geq 0,\ D_{\varphi,\mathbf{P}}\left(\mathbf{Q},\mathbf{Q}^{**}\right) = 0 \text{ iff } \ \mathbf{Q} = \mathbf{Q}^{**}$$

$$\Phi_{\mathbf{P},\mathbf{Q}^{**}}(\Omega) := \inf_{\mathbf{Q}\in\Omega} D_{\varphi,\mathbf{P}}\left(\mathbf{Q},\mathbf{Q}^{**}\right)$$

- $\mathbf{P}_{unif} = (1/K, .., 1/K)$: ordinary/classical Bregman divergence $D_{\varphi}^{BD}\left(\mathbf{Q},\mathbf{Q}^{**}\right) = \frac{1}{K} D_{\varphi,\mathbf{P}_{unif}}^{SBD}\left(\frac{\mathbf{Q}}{K}, \frac{\mathbf{Q}^{**}}{K}\right)$
- $D_{\varphi,\mathbf{P}}\left(\mathbf{Q},\mathbf{P}\right) = D_{\varphi}\left(\mathbf{Q},\mathbf{P}\right)$ (CASM)

Construction of the simulation scheme

$$\inf_{\mathbf{Q}\in\Omega} D_\varphi\left(\mathbf{Q},\mathbf{P}\right)$$

Remark: For Uniform $\mathbf{P} := (1/K, .., 1/K)$:equivalent to constrained maximization of entropies
Blocks of indexes induced by $\mathbf{P}$. Let n be such that for all $k = 1, .., K$ the product np_k is an integer.
$I_1^n := \{1, .., np_1\}$, $I_2^n := \{np_1 + 1, .., np_1 + np_2\}$be the second block, and so on until $I_K^n := \{n\left(p_1 + .. + p_{K-1}\right) + 1, .., n\}$.

Theorem

Consider n iid copies of W with distr Π *and define the vector* $\mathbf{V}_n$ *through*

$$\mathbf{V}_n := (V_{1,n}, .., V_{K,n}) \tag{2}$$

where

$$V_{k,n} := \frac{1}{n} \sum_{i \in I_k^n} W_i = \frac{p_k}{n_k} \sum_{i \in I_k^n} W_i .$$

It then holds

$$D_\varphi(\Omega, \mathbf{P}) = -\frac{1}{n} \log \Pi(\mathbf{V}_n \in \Omega) + O\left(\frac{\log n}{n}\right)$$

for all sets $\Omega \subset \mathbb{R}^K$ *satisfying the regularity condition; for convenience* Ω *bounded*

$$Int\Omega \neq \varnothing \text{ and } Cl(Int\Omega) = Cl\Omega. \tag{3}$$

CASM Estimators, first approach

Let $\delta > 0$

$$\Omega^* \text{ minimizers of } D_\varphi\left(\mathbf{Q},\mathbf{P}\right) \text{ over } \Omega$$

$$\left(\Omega^*\right)^\delta := \left\{\mathbf{Q} \in \mathbb{R}^K : d\left(\mathbf{Q},\Omega^*\right) < \delta\right\}$$

Then

$$\Pi\left(\mathbf{V}_n \in \left(\Omega^*\right)^\delta \middle| \mathbf{V}_n \in \Omega\right) \to 1$$

Algorithm Simulate $\mathbf{V}_n^1, .., \mathbf{V}_n^L$ iid . Order the $D_\varphi\left(\mathbf{V}_n^l, \mathbf{P}\right)$'s for $\mathbf{V}_n^l \in \Omega$ (HIT RATE!) and define

$$\mathbf{V}_{n,L}^* := \arg\min_{l=1,..,L}\left\{D_\varphi\left(\mathbf{V}_n^l, \mathbf{P}\right), \mathbf{V}_n^l \in \Omega\right\}$$

Then

$$\lim_{n\to\infty}\lim_{L\to\infty} d\left(\mathbf{V}_{n,L}^*,\Omega^*\right) = 0 \quad \text{pr}$$

$$\lim_{n\to\infty}\lim_{L\to\infty} D_\varphi\left(\mathbf{V}_{n,L}^*, \mathbf{P}\right) = D_\varphi\left(\Omega, \mathbf{P}\right) \quad \text{pr}$$

Bregman Simulation scheme

The instrumental sequence of rv's $\mathbf{V}_n$ is centered at some point $\mathbf{Q}^{**}$ with Large deviation rate $D_{\varphi,\mathbf{P}}^{SBD}(., \mathbf{Q}^{**})$

$$\mathbf{V} := \mathbf{V}_n := (V_1, \ldots, V_K)$$

$$\mathbf{V}_n := \Big(\frac{1}{n}\sum_{i\in I_1^{(n)}} V_{n,i}, \ldots, \frac{1}{n}\sum_{i\in I_K^{(n)} V_{n,i} i}\Big) \qquad ,$$

Now the components of each block have common distribution depending on the block whose sizes also here depend on $\mathbf{P}$.

Bregman vector case

For $k=1,..,K$, n_k i.i.d. random variables$V_{n,i}$, $i \in I_k^{(n)}$, with common distribution $[V_{n,i} \in \cdot\,] = U_k[\,\cdot\,]$ $(i \in I_k^{(n)})$ given by

$$dU_k(v) \;:=\; \frac{\exp\left(\tau_k \cdot v\right)}{E\exp(\tau_k \cdot W)}\, d\Pi^W(v),$$

and τ_k satisfies

$$\frac{d}{d\tau}\log E\exp(\tau_k \cdot W) = \frac{q_k^{**}}{p_k}$$

(Eescher transform, tilted (twisted) distribution).

$$E\mathbf{V}_n = \mathbf{Q}^{**}$$

Bregman vector case

Theorem

The instrumental sequence of rv's $\mathbf{V}_n$ *is centered at* $\mathbf{Q}^{**}$ *with Large deviation rate* $D_{\varphi,\mathbf{P}}^{SBD}(.,\mathbf{Q}^{**})$

$$\inf_{\mathbf{Q}\in\Omega} D_{\varphi,\mathbf{P}}^{SBD}(\mathbf{Q},\mathbf{Q}^{**}) = -\frac{1}{n}\log\Pi(\mathbf{V}_n\in\Omega)+O\left(\frac{\log n}{n}\right)$$

for regular Ω.

Bregman vector case

Example

equal sized ($\mathbf{P}$ uniform) and $\varphi_2(x) := (x-1)^2$ (ie $\Pi = N(1,1)$) for

$$D_{\varphi,\mathbf{P}}^{SBD}(\mathbf{Q},\mathbf{Q}^{**}) = C\left\|\mathbf{Q}-\mathbf{Q}^{**}\right\|^2$$

For $\varphi(x)$ and $\mathbf{P}$ uniform $D_{\varphi,\mathbf{P}}^{SBD}(\mathbf{Q},\mathbf{Q}^{**})$ is the Basu Harris Hjort Jones power divergence, other choices:Mahalanobis, etc

Numerical considerations

1-Direct estimator of $\inf_{\mathbf{Q}\in\Omega} D_\varphi(\mathbf{Q},\mathbf{P})$ for CASM: Substitute $\Pi(\mathbf{V}_n\in\Omega)$ by its empirical counterpart

$$\begin{aligned}\Pi(\mathbf{V}_n\in\Omega) &\simeq \frac{1}{L}\sum_{l=1}^{L} 1_\Omega\left(\mathbf{V}_n^l\right)\\ &\simeq \frac{1}{L}\sum_{l=1}^{L} 1_\Omega\left(\widetilde{\mathbf{V}}_n^l\right)\frac{\pi}{q}\left(\widetilde{\mathbf{V}}_n^l\right)\end{aligned}$$

(Importance sampling with $\widetilde{\mathbf{V}}_n^l$ iid under q); then $\frac{1}{n}\log\Pi(\mathbf{V}_n\in\Omega)$ $\simeq \frac{1}{n}\log\frac{1}{L}\sum_{l=1}^{L} 1_\Omega\left(\widetilde{\mathbf{V}}_n^l\right)\frac{\pi}{q}\left(\widetilde{\mathbf{V}}_n^l\right)$ and choose $\widetilde{\mathbf{V}}_n$ with expectation in Ω

2-For example when $\varphi=\varphi_\gamma$, all corrresponding W have infinitely divisible distribution. Therefore $\mathbf{V}_n$ can be simulated easily under Π or under its tilted (IS)

All calculation can be achieved as parallelized steps

Same for Bregman

Statistical context

Often $\mathbf{P}$ unknown and $\mathbf{P}_N(X_1, .., X_N) \to \mathbf{P}$ a.s. A model $\Omega \subset \mathbb{S}^K$ **An index of fit.**
(non, semi-parametric, ex moment conditions, tubular neighborhood of parametric model, etc)
Consider now the case when Ω is a subset in $\mathbb{S}^K$ with non void interior in the relative topology, with $Cl(Int\Omega) = Cl\Omega$.

$$D_\varphi(\Omega, \mathbf{P}_N) \qquad \text{is an index of fit}$$

The vector $\mathbf{V}_n$ is modified so that the sum of its components is 1 through

$$\overline{\mathbf{V}}_n := \left(\overline{V}_{1,n}, .., \overline{V}_{K,n}\right)$$

with

$$\overline{V}_{k,n} := \frac{\sum_{i \in I_k^n} W_i.}{\sum_{k=1}^K \sum_{i \in I_k^n} W_i.}$$

whenever $\sum_{i=1}^n W_i \neq 0$,($\overline{\mathbf{V}}_n = (\infty)^K$ otherwise) and the blocks I_k^n have lengths $n\mathbf{P}_N(k) \in \mathbb{N}_{>0}$.

Bregman , statistical context

Theorem

The following limit statement holds

Theorem

$$\lim_{n\to\infty} \frac{1}{n} \log \Pi \left(\overline{\mathbf{V}}_n \in \Omega \middle| \mathbf{P}_N \right) = - \inf_{m \neq 0} D_\varphi \left(m\Omega, \mathbf{P}_N \right) \tag{4}$$

Statistical context

Solving the minimization pb: get rid of the $\inf_{m \neq 0}$ For power divergences φ_γ: there exists F_γ **invertible, explicit** such that with $\Omega \subset \mathbb{S}^K$ for any $\mathbf{P}$

$$\inf_{m \neq 0} D_\varphi (m\Omega, \mathbf{P}) = F_\gamma(D_\varphi (\Omega, \mathbf{P}))$$

Theorem

$$D_{\varphi_\gamma} (\Omega, \mathbf{P}_N) = \lim_{n \to \infty} F_\gamma^{-1} \left(-\frac{1}{n} \log \Pi \left(\mathbf{V}_n \in \Omega | \, \mathbf{P}_N \right) \right)$$

$$D_{\varphi_\gamma} (\Omega, \mathbf{P}) = \lim_{N \to \infty} \lim_{n \to \infty} F_\gamma^{-1} \left(-\frac{1}{n} \log \Pi \left(\mathbf{V}_n \in \Omega | \, \mathbf{P}_N \right) \right)$$

Remark For Bregman divergences: no simple transformation from $\mathbb{R}^K$ to $\mathbb{S}^K$. Anyhow bounds for $D_\varphi (\Omega, \mathbf{P})$ (example Jensen Shannon, etc)

Estimators

Estimators: minimal value of $D_\varphi(\mathbf{Q},\mathbf{P})$ over $\Omega\subset\mathbb{R}^K$, Simulate $\mathbf{V}_n^1,..,\mathbf{V}_n^L$ iid

$$\widehat{D_\varphi(\Omega,\mathbf{P})}_{n,L} := -\frac{1}{n}\log\frac{1}{L}\sum_{l=1}^{L}1_\Omega(\mathbf{V}_n^l)$$

For minimal value of $D_{\varphi_\gamma}(\mathbf{Q},\mathbf{P})$ over $\Omega\subset\mathbb{S}^K$

$$\widehat{D_\varphi(\Omega,\mathbf{P})}_{n,L} := F_\gamma^{-1}\left(-\frac{1}{n}\log\frac{1}{L}\sum_{l=1}^{L}1_\Omega(\mathbf{V}_n^l)\right)$$

Better: IS simulate from a distribution centered at $\omega\in\Omega$ close to Ω^*, etc. Many possibilities; (B-Stummer IEEE 2023).

Intermediate remark

- **Starting from the divergence generator.** Simulate random vectors $\mathbf{V}_n$ on $\mathbb{R}^K$ or on $\mathbb{S}^K$ with prescribed LDP rates of the form $D_\varphi(., \mathbf{P})$ for prescribed $\mathbf{P}$.
- **Starting from a distribution** for W with $EW = 1$ and Cramer condition. Define $\mathbf{P}$. generate $\mathbf{V}_n$ (on $\mathbb{R}^K$ or on $\mathbb{S}^K$) with LDP rate D_φ (*unknown*)

In both cases $\mathbf{V}_n$ can be modified in order to be centered at any point in $\mathbb{R}^K$ or on $\mathbb{S}^K$. For example through $\mathbf{Q}^{**}$ when making use of scaled Bregman or modifying the LDP CASM rate accordingly

Adapting Varadhan's Lemma

A (any) sequence of rv's $\mathbf{Z}_n \subset \mathcal{X}$ with LDP rate J. A compact subset $\Omega \subset \mathcal{X}$ with regularity condition; a continuous function $\Phi : \mathcal{X} \to \mathbb{R}$. Let Φ^* the set of miimizers of Φ on Ω

$$\lim_{n\to\infty} \frac{1}{n} \log E \left[\exp n \left(J(\mathbf{Z}_n) - \Phi\left(\mathbf{Z}_n\right)\right) 1_\Omega \left(\mathbf{Z}_n\right)\right] = - \inf_{\mathbf{Q}\in\Omega} \Phi\left(\mathbf{Q}\right).$$

Simulate L iid copies of $\mathbf{Z}_n$ with LD rate J (possibly unknown). Rank $\Phi\left(\mathbf{Z}_n^l\right)$. Set

$$\mathbf{Z}_n^{L*} := \arg \min_{l=1,..,L} \left\{ \Phi\left(\mathbf{Z}_n^l\right) \ , \mathbf{Z}_n^l \in \Omega \right\}$$

Theorem

It holds

$$\lim_{n\to\infty} \lim_{L\to\infty} d\left(\mathbf{Z}_n^{L*}, \Phi^*\right) = 0 \text{ in probabilty}$$

$$\lim_{n\to\infty} \lim_{L\to\infty} \Phi\left(\mathbf{Z}_n^{L*}\right) = \inf_{\mathbf{Q}\in\Omega} \Phi\left(\mathbf{Q}\right) \quad \text{in probability}$$

Minimization of functionals, vector case

A general result: for regular compact $\Omega \subset \mathbb{R}^K$ and continuous Φ (also version for non compact Ω)

$$-\lim_{n\to\infty} \frac{1}{n} \log E_{\Pi} \left[\exp n \left(D_{\varphi} \left(\mathbf{V}_n, \mathbf{P} \right) - \Phi \left(\mathbf{V}_n \right) \right) 1_{\Omega} \left(\mathbf{V}_n \right) \right] = \inf_{\mathbf{Q}\in\Omega} \Phi \left(\mathbf{Q} \right) \tag{5}$$

Example

With $\Phi(\mathbf{Q}) = \Phi_{\mathbf{P}}(\mathbf{Q}) = \sum |q_k - p_k|$ and $W = Z_1 - Z_2 + 1$, and Ω regular (non compact)

Minimization of functionals, vector case

Importance Sampling

$$E_{\Pi}\left[\exp n\left(D_{\varphi}\left(\mathbf{V}_n,\mathbf{P}\right)-\Phi\left(\mathbf{V}_n\right)\right)1_{\Omega}\left(\mathbf{V}_n\right)\right]$$

$$\frac{1}{L}\sum_{l=1}^{L}\exp n\left(D_{\varphi}\left(\mathbf{T}_n^l,\mathbf{P}\right)-\Phi\left(\mathbf{T}_n^l\right)\right)\frac{d\Pi}{dR}\left(\mathbf{T}_n^l\right)1_{\Omega}\left(\mathbf{T}_n^l\right)$$

with $\mathbf{T}_n^l$'s iid R; if R centered on some ω in Ω, better hit rate; choice : R tilted from Π or centered and modified in variance, etc. Minimizers:

$$\arg\min\Phi\left(\mathbf{T}_n^l\right),\mathbf{T}_n^l\in\Omega$$

Minimization of functionals, vector case

Rates of convergence

$$\frac{1}{n}\log E_{\Pi}\left[\exp n\left(D_{\varphi}\left(\mathbf{V}_n,\mathbf{P}\right)-\Phi\left(\mathbf{V}_n\right)\right)1_{\Omega}\left(\mathbf{V}_n\right)\right]$$

$$f_{\mathbf{V}_n}(v)=\left[n^{K/2}\Psi(v)\exp -nD_{\varphi}\left(v,\mathbf{P}\right)\right]\left(1+O\left(\frac{1}{n}\right)\right)$$

(Jensen, etc) with $\Psi(v)$ defined through the 2^d order properties of φ.
Laplace method, etc give rate of cv $O\left(\frac{\log n}{n}\right)$
Plus rate in LLN as $L\to\infty$ for given n.

Minimization of functionals vector case

Example

With **P** uniform in (5) and some change of notation

$$\inf \|\mathbf{Q}\|_1$$

such that with $d << K$, with non degeneration assumptions on the entries $x_{i,k}$ so that $Int\Omega \neq \varnothing$ with

$$\Omega := \left\{ \mathbf{Q} \in \mathbb{R}^K \text{ such that } \sum_{i=1}^{d} \left(y_i - \sum_{k=1}^{K} x_{i,k}.q_k \right)^2 \leq \varepsilon \right\}$$

(Basis Pursuit Denoising Problem); see example Neural Network

Minimization of functionals, vector case

Estimators

Simulate L iid copies of $\mathbf{V}_n$ with LD rate $D_\varphi\left(.,\mathbf{P}\right)$. Set

$$\mathbf{V}_n^{L*} := \arg\min\left\{\Phi\left(\mathbf{V}_n^l\right), \mathbf{V}_n^l \in \Omega\right\} \text{ a proxy of } \Phi^*$$

$$\Phi\left(\mathbf{V}_n^{L*}\right) \text{ a proxy of } \Phi\left(\Omega\right)$$

as $L \longrightarrow \infty$ and $n \longrightarrow \infty$.
Choice of $\mathbf{V}_n$ (and therefore of $D_\varphi\left(.,\mathbf{P}\right)$). Take $D_\varphi\left(\mathbf{Q},\mathbf{P}\right) = D_{\varphi,\mathbf{R}}^{SBD}\left(\mathbf{Q},\mathbf{Q}^{**}\right)$ so that $E\mathbf{V}_n = \mathbf{Q}^{**} \in \Omega$, possibly with $\Phi\left(\mathbf{Q}^{**}\right) \approx \Phi\left(\Omega\right)$. (see example Neural Network)

Statistical context

.

Typically $\Omega \subset \mathbb{S}^K$,$\mathbf{P}$ unknown $(X_1 \ , .., X_N) \rightarrow P_N := \frac{1}{N} \sum \delta_{X_i}$

$$\Phi\left(\mathbf{Q}\right) := \Phi_{\mathbf{P}}\left(\mathbf{Q}\right).$$

Simulate L iid copies of $\overline{\mathbf{V}}_n$ for arbitrary R

$$\overline{V}_{n,k} := \frac{\sum_{i \in I_n^k} W_i}{\sum_{k=1}^{K} \sum_{i \in I_n^k} W_i}$$

$$\overline{\mathbf{V}}_n := \left(\overline{V}_{n,1}, .., \overline{V}_{n,K}\right)$$

The blocks of indices I_n^k are random $I_n^1 := \{1, .., nR(1)\}$, $I_n^2 := \{nR(1) + 1, .., n\left(R(1) + R(1)\right)\}$,etc

Statistical context

.

. If the $W_i^{\prime}s$ are iid such that φ_γ is the Legendre transform of $\log E \exp tW$ **LD rate** of $\mathbf{V}_n$ exists depending on R

$$\mathbf{Q} \to F_\gamma^{-1}\left(D_{\varphi_\gamma}(\mathbf{Q},R)\right).$$

More generally if W satisfies Cramer condition then $\overline{\mathbf{V}}_n$ obeys LDP with rate $\inf_{m\neq 0} D_\varphi\left(.,R\right)$,

Statistical context

However explicit rate not required

$$\overline{\mathbf{V}}_{n,N}^{L*} := \arg\min \left\{ \Phi_{P_N}\left(\overline{\mathbf{V}}_n^{l}\right), \overline{\mathbf{V}}_n^{l} \in \Omega \right\}$$

Then

$$\lim_{N\to\infty} \lim_{n\to\infty} d\left(\overline{\mathbf{V}}_{n,N}^{L*}, \Phi_{\mathbf{P}}^{*}\right) = 0 \quad \text{pr}$$

$$\lim_{N\to\infty} \lim_{n\to\infty} \Phi_{P_N}\left(\overline{\mathbf{V}}_n^{L*}\right) = \min\left\{\Phi_{\mathbf{P}}^{*}(\mathbf{Q}), \mathbf{Q} \in \Omega\right\} \text{ pr.}$$

Application: Inference in neural networks (with P Bertrand)

Neural network: Trained Fashion MNIST 70000 images, 10 classes.
Two Parts: A convolution network and a dense network.
First part (convolution): 8544 weights
Second part (dense) 412778 weights $K = 412778$
Activation: Relu for all levels, Softmax (10 classes) for the last one
Training provides a set of weights a_0 leading to Prob success $P(a_0) = 0.91$ on test sample (size 10000)

$$\Omega_\eta := \left\{ a \in \mathbb{R}^K : P(a) \geq P(a_0) - \eta \right\}$$

$$\min_{a \in \Omega_\eta} \Phi(a) := \min_{a \in \Omega_\eta} \|a\|_1$$

and find sparse solutions among minimizers.

Application: Inference in neural networks

Pruning: Minimizing $\|a\|_1$ and then pruning is known to increase the rate of null weights achieving Ω_η.
1- minimizing $\|a\|_1$ over Ω_η:Choose $W := Z_1 - Z_2 + 1$, $Z_{1,2}$ two gammas with same expectation, randomize on variance. Make a shift in order to achieve at step 1

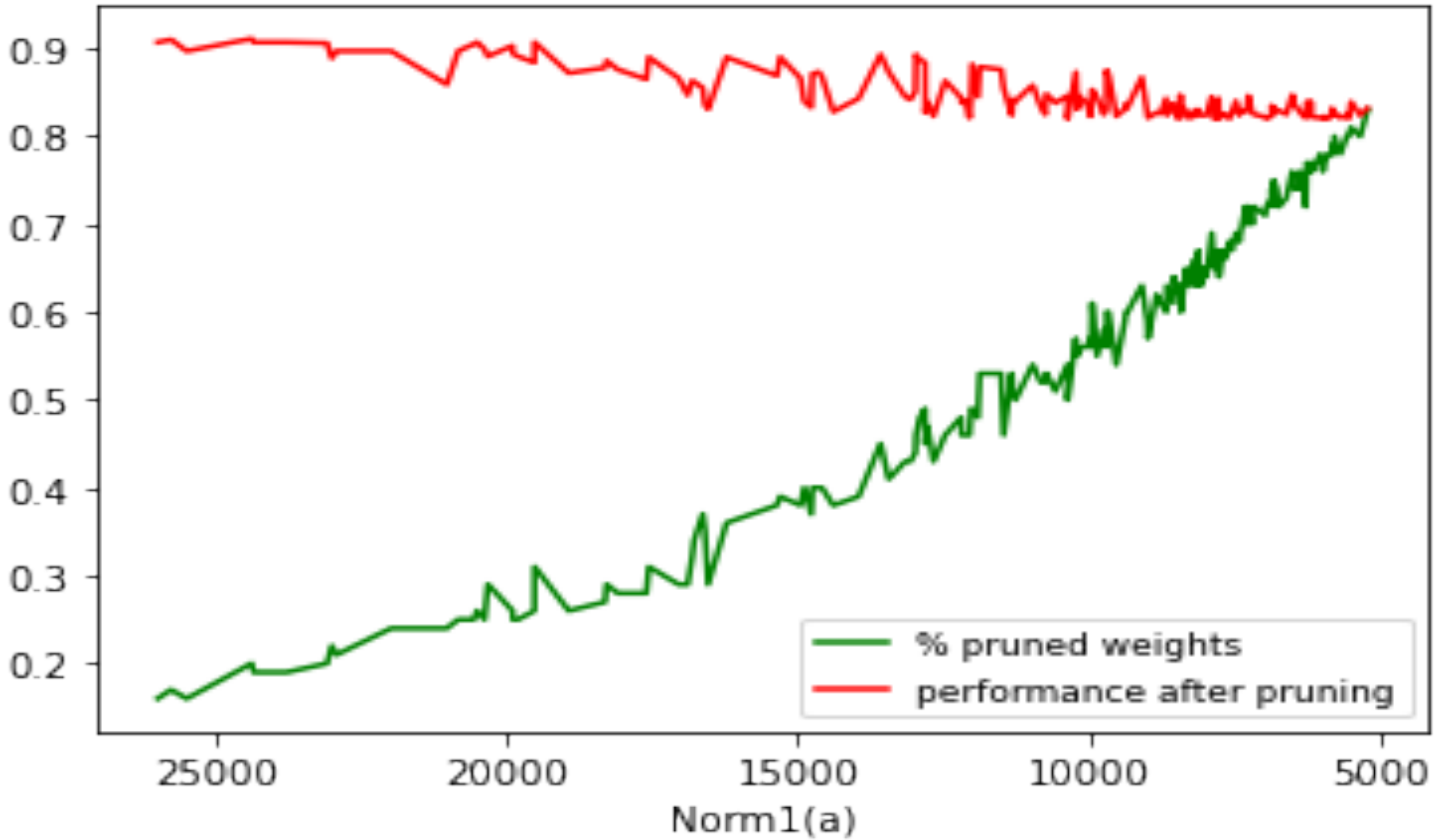
$$E\mathbf{V}_{n,1} = a_{0.}$$

Iteratively run the minimization on Ω_η, adapting $E\mathbf{V}_{n,i+1}$ to the value a_i with $\|a_i\|_1 < \|a_{i-1}\|_1$, get a_{i+1}, etc
2-Pruning: set all components of a_{i+1} less than 10^{-6} to 0. Denote α_{i+1} the resulting weights
2-Evaluate $P(\alpha_{i+1})$.

Application: Inference in neural networks

RESULTS

itération	$\|\|W\|\|$	Performance de $R(W)$	% removed	# minima
0-9	15221	86	39	83
10-19	9199	83	61	26
20-29	8286	82	63	15
30-39	7086	83	71	17
40-49	7328	82	69	10
50-59	6585	82	73	6
60-69	6937	82	71	5
70-79	7700	82	65	4
80-89	6834	82	72	4
90-99	7678	82	67	2



Remark Minimizing the $\|.\|_1$:sparse minimizers (under convex constraints) ; Donoho, Candès (2006), hence the success of pruning

References

Br-Stummer A bare simulation approach to some distance minimization problems 1-Foundations IEEE Transactions on Information Theory, vol. 69, no. 5, pp. 3062-3120, 2023

Br-Stummer A bare simulation approach to some distance minimization problems 2-Further foundations Arxiv, September 2023

Bertrand-Br-Stummer Neural network inference through bare simulation, Arxiv 2024

Br -A bootstrap procedure for the minimization of some divergences, Analytic Methods in Statistics, Proc. Math. Stat., 193, 1–22., Springer, Cham, 2017

Letac, Mora,Natural real exponential families with cubic variance functions Ann. Statist. 18 (1990), no. 1, 1–37.

Csiszar, Gamboa, Gassiat MEM pixel correlated solutions for generalized moment and interpolation problems.IEEE Trans. Inform. Theory45(1999), no.7, 2253–2270

Stummer , Vajda, On Bregman distances and divergences of probability measures IEEE Trans. Inform. Theory 58 (2012), no. 3, 1277–1288.

THANK YOU