AI Research – Criteo AI Lab

Out of Academia… and a Deep Dive on a PAC-Bayes Wasserstein

Liva Ralaivola, VP Research, Criteo AI Lab
Prof@AMU, on secondment
@CriteoAILab, @LivaRalaivola

Journée Fondements Mathématiques de l'IA,
Sorbonne Université
Oct. 2, 2023

CRITEO

Outline

Outline

Fact sheet

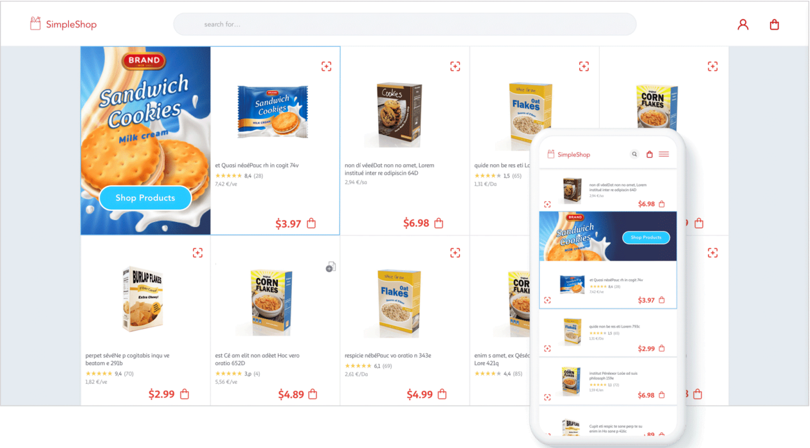
Capsule bio

2019. Full-time Criteo, head of Research of Innovation
2018. Part-time researcher at Criteo x Prof. at Aix-Marseille Université
2011. Prof. at Aix-Marseille Université
2004. Assist Prof. at Aix-Marseille Université
2004. Postdoc UC Irvine
2003. (20 years ago!!) PhD in this very place (almost, cf. Capitaine Scott)

Miscellaneous info

Criteo AI Lab. 20+ permanent researchers, 10+ PhD students, 120 (ML) engineers
Publications. Publications at NeurIPS, ICML, ICLR, AISTATS...
Partnerships. Universities, INRIA (cf. FAIRPLAY joint team)
Inner beat. Bi-annual evaluation, quarterly company-level synchro (OKR)

Criteo: from Retargeting to Retail Media Advertising



The History of Machine Learning at Criteo in a Glimpse

Criteo founded

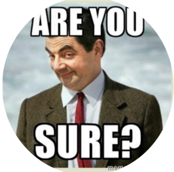
2005

Started in ads

2008

ML powered ads

2012



Sales > clicks

2014

CRITEO AI Lab

AI Lab creation

~30 researchers (10 PhD)

~60 engineers

2018

Data valuation through Machine Learning

Expand the science and technology of scalable AI
for the Open Internet data to be rightly valued,
secured and transparent commodities

$\Rightarrow$ **ML Science:** game theory, reinforcement learning, deep learning for structured data, generative AI, privacy-preserving ML

Data valuation through Machine Learning

FAIRPLAY: Criteo x INRIA joint team, with ENSAE
Coopetitive AI: fairness, privacy, incentivization

«L'objectif derrière le travail de l'équipe-projet est ainsi d'améliorer les systèmes automatiques de places de marché, mais également d'être en mesure de connaître le degré de discrimination de certains algorithmes, le tout en restant compatible avec les notions de protection de vie privée.»[1]

- **DU-Shapley: A Shapley Value Proxy for Efficient Dataset Valuation.** F. Garrido-Lucero, B. Heymann, M. Vono, P. Loiseau, V. Perchet, Arxiv, 2023
- **Collaborative Ad Transparency: Promises and Limitations.** E. Gkiouzepi , A. Andreou , O. Goga , P. Loiseau, Symposium on Security and Privacy, 2023
- **An algorithmic solution to the Blotto game using multi-marginal couplings.** V. Perchet , P. Rigollet , T. Le Gouic Economics and Computation, 2022
- ...

[1] https://www.inria.fr/fr/comment-eviter-discrimination-donnees-utilisateurs-publicite-fairplay-criteo

Scaling Machine Learning

600 TB
Data created per day

50K servers
3K Hadoop nodes
6M queries per second
<100 ms to answer a query

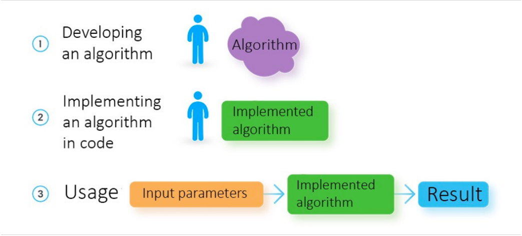
162Bn
DIN A4 pages

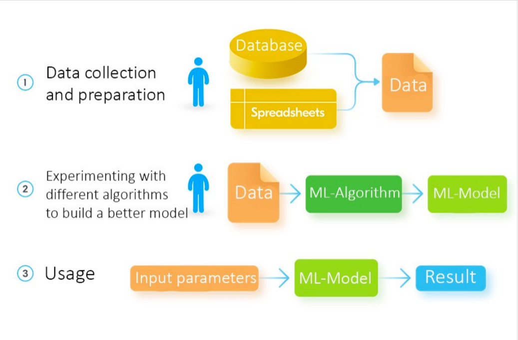
81M Kg
DIN A4 pages

13.5K
Elephants

Scaling Machine Learning

AI, a new way of programming: everything old is new again[2]





Problems: distributed learning, robust learning, privacy-preserving learning, deployment, platform, infra, verification, complexity...

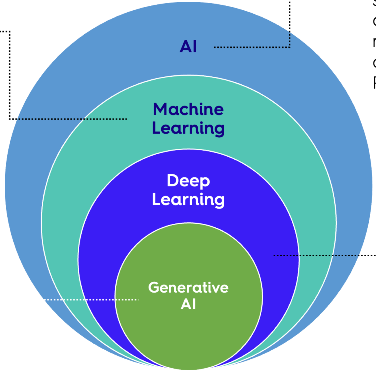
[2] https://towardsdatascience.com/machine-learning-vs-traditional-programming-c066e39b5b17

Educating on ML... beyond Master and PhD students

Targetting C-levels, commercial teams, HR teams...

Machine Learning is a subfield of AI that deals with systems/models that improve as they are exposed to more data (eg. cats vs dogs images classification

Generative AI is a subfield of Deep Learning where the systems / models can produce real-like signals (images, texts, sounds)



AI is a branch of computer science and mathematics that deals with systems that can reason, learn and act autonomously (eg. AlphaGo, Poker)

Deep Learning is a subfield of Machine Learning where the ML systems / models are « Deep » Artificial Neural Networks (eg. scene labeling)

Educating on ML... beyond Master and PhD students

and also technical teams...

Bootcamps. 10-day internal education on ML to software developers
Voyagers/Khali. 2-week to 2-quarter internal mobility, e.g.

- Horizontal Personalized Federated Learning for Criteo Keyword Model
- Improved Generalized Linear Value Function Approximation in Episodic Reinforcement Learning
- Game theory for data-sharing mechanisms

Reading and coding groups. Group reading of a book or implementation of a tutorial (experimental)
Hackathons. Annual 3-day Hackathon

Outline

Objective of the work

Joint work with R. Ohana (Flatiron, NY), K. Nadjahi (MIT, Boston), A. Rakotomamonjy (Criteo) *@ICML2023*

Provide a theoretical analysis for Adaptive Sliced Wasserstein Distances...

- SWD are distances between measures that are cheap to compute
- that primarily rely on a Uniform sampling of **slices**
- and that extend to non-uniform sampling of **slices**

using an alignment between Adaptive SWD and the PAC-Bayes Theory

- PAC Bayes bounds primarily characterize the generalization ability of the stochastic **Gibbs predictor** [Alquier, 2021]
- Adaptive Sliced Wasserstein Distances **do compute** the error of a **Gibbs predictor**

From Wasserstein Distance...

Definition (Wasserstein distance)

Let $p \in [1, \infty)$. The p-Wasserstein distance between μ, ν two measures on Ω is given by

$$\mathrm{W}_p^p(\mu, \nu) \doteq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathsf{X} \times \mathsf{X}} \|x - y\|_p^p d\pi(x, y) \,, \tag{1}$$

where $\Pi(\mu, \nu) \subset \mathcal{P}(\mathsf{X} \times \mathsf{X})$ denotes the set of probability measures on $\mathsf{X} \times \mathsf{X}$, whose marginals with respect to the first and second variables are μ and ν respectively.

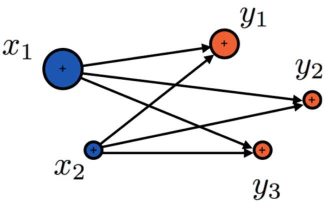
to Empirical Wasserstein Distances

Definition

- Given two probability distributions μ, ν on Ω with metric $\|\cdot\|_q$
- For $(\mathbf{x}_i)_{i=1}^n \sim \mu$, $(\mathbf{y}_i)_{i=1}^n \sim \nu$, let $\hat{\mu}_n^{\mathbf{a}} = \sum_{i=1}^n a_i \delta_{\mathbf{x}_i}$ and $\hat{\nu}_n^{\mathbf{b}} = \sum_{i=1}^n b_i \delta_{\mathbf{y}_i}$, with $\mathbf{a}$ and $\mathbf{b}$ distributions, the p-Wasserstein distance is

$$W_p^p(\hat{\mu}_n^{\mathbf{a}}, \hat{\nu}_n^{\mathbf{b}}) \doteq \min_{\Gamma \in \mathbf{P}} \left\{ \langle \Gamma, \mathbf{C}_q \rangle_F = \sum_{i,j} \gamma_{i,j} \|\mathbf{x}_i - \mathbf{y}_j\|_p^p \right\}$$

where $\mathbf{P} \doteq \{\Gamma \in (\mathbb{R}^+)^{n \times n} |\, \Gamma \mathbf{1}_{n_t} = \mathbf{a}, \Gamma^T \mathbf{1}_{n_s} = \mathbf{b}\}$



to (Empirical) Sliced Wasserstein Distance...

Definition (Sliced Wasserstein Distance)

- sample some random directions $\mathbf{u} \in \mathbb{S}^{d-1}$ uniformly
- project data on each random direction
- compute all 1D Wasserstein distances (cheap) and average them

$$\mathrm{SWD}_p^p(\hat{\mu}_n, \hat{\nu}_n) \doteq \frac{1}{k}\sum_{j=1}^{k} W_p^p\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{x}_i{}^\top \mathbf{u}_j}, \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{y}_i{}^\top \mathbf{u}_j}\right)$$

$\mathrm{SWD}_p^p(\hat{\mu}_n, \hat{\nu}_n)$ is an estimator of

$$\mathrm{SWD}_p^p(\hat{\mu}_n, \hat{\nu}_n; \rho) \doteq \int_{\mathbb{S}^{d-1}} W_p^p\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{x}_i{}^\top \mathbf{u}}, \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{y}_i{}^\top \mathbf{u}}\right)\rho(\mathbf{u})d\mathbf{u}$$

when ρ is the uniform distribution on $\mathbb{S}^{d-1}$

to (Empirical) Sliced Wasserstein Distance...

Definition (Sliced Wasserstein Distance)

- sample some random directions $\mathbf{u} \in \mathbb{S}^{d-1}$ uniformly
- project data on each random direction
- compute all 1D Wasserstein distances (cheap) and average them

$$\mathrm{SWD}_p^p(\hat{\mu}_n, \hat{\nu}_n) \doteq \frac{1}{k}\sum_{j=1}^{k} W_p^p\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{x}_i^\top \mathbf{u}_j}, \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{y}_i^\top \mathbf{u}_j}\right)$$

$\mathrm{SWD}_p^p(\hat{\mu}_n, \hat{\nu}_n)$ is an estimator of

$$\mathrm{SWD}_p^p(\hat{\mu}_n, \hat{\nu}_n; \rho) \doteq \mathbf{E}_{\mathbf{u}\sim\rho} W_p^p\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{x}_i^\top \mathbf{u}}, \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{y}_i^\top \mathbf{u}}\right) = \mathbf{E}_{\mathbf{u}\sim\rho} W_p^p\left(\mathbf{u}_\sharp^* \hat{\mu}_n, \mathbf{u}_\sharp^* \hat{\nu}_n\right)$$

when ρ is the uniform distribution on $\mathbb{S}^{d-1}$ ($\mathbf{u}_\sharp^* \hat{\mu}_n$ is the push-forward of $\hat{\mu}_n$)

to Adaptive Sliced Wasserstein Distances

Looking for ρ that makes the most discriminative Sliced Wasserstein Distance

- *Max-SW* [Deshpande et al., 2019] learns **a single** slice that maximizes the distance:

$$\mathrm{maxSW}(\hat{\mu}_n, \hat{\nu}_n) \doteq \max_{\delta_\theta : \theta \in \mathbb{S}^{d-1}} \mathrm{SW}_p^p(\hat{\mu}_n, \hat{\nu}_n; \delta_\theta) \tag{2}$$

- *Distributional SW* [Nguyen et al., 2021] learns **the whole** distribution of slices that maximizes the distance:

$$\mathrm{DSW}(\hat{\mu}_n, \hat{\nu}_n) \doteq \sup_{\mathbf{E}_{\theta,\theta' \sim \rho}[|\theta^\top \theta'|] \leq C,\ \rho \in \mathcal{P}(\mathbb{S}^{d-1}),} \mathrm{SW}_p^p(\hat{\mu}_n, \hat{\nu}_n; \rho) \tag{3}$$

What is missing, what do we provide

- What is the generalization power of the learned ρ? Is ρ discriminative on 'unseen' data?
- We introduce PAC-Bayesian results to answer the above questions

PAC Bayes Theory: bounds on the risk of the Gibbs predictor

Definition (Risks)

The *empirical ℓ-risk* $\hat{r}_\ell(\omega, S_n)$ (ℓ being a loss function) and *true ℓ-risk* $r_\ell(\omega)$ with training data $S_n = \{z_1, \ldots, z_n\}$ and parameters $\omega \in \Omega$ are

$$\hat{r}_\ell(\omega, S_n) \doteq \frac{1}{n}\sum_{i=1}^{n} \ell(\omega, z_i), \qquad r_\ell(\omega) \doteq \mathbf{E}_{z\sim\xi}[\ell(\omega, z)]$$

Theorem ([Catoni, 2003])

Let $\rho_0 \in \mathcal{P}(\Omega)$ a prior distribution. Assume bounded loss $0 \le \ell \le C$. For all $\lambda > 0$, for any $\delta \in (0,1)$, with probability $> 1-\delta$ (over dataset S_n): $\forall \rho \in \mathcal{P}(\Omega)$,

$$\mathbf{E}_{\omega\sim\rho}[r_\ell(\omega)] \le \mathbf{E}_{\omega\sim\rho}[\hat{r}_\ell(\omega, S_n)] + \frac{\lambda C^2}{8n} + \frac{1}{\lambda}\left\{KL(\rho||\rho_0) + \log\frac{1}{\delta}\right\}, \tag{4}$$

with $KL(\rho||\rho_0)$ the Kullback-Leibler divergence between ρ and ρ_0.

PAC Bayes Theory: bounds on the risk of the Gibbs predictor

Remarks

- The Gibbs predictor is a stochastic predictor which, upon a call:
  1. samples a predictor ω according to ρ
  2. outputs a prediction according to ω
- PAC-Bayes bounds focus on aggregated risks ($\mathbf{E}_{\omega\sim\rho}[r_\ell(\omega)]$ and $\mathbf{E}_{\omega\sim\rho}[\hat{r}_\ell(\omega, S_n)]$) instead of the risk the aggregated predictor $\omega_\rho \doteq \mathbf{E}_{\omega\sim\rho}\omega$
- They provide tight bounds on the risk of the Gibbs predictor
- Multiple works have turned PAC Bayes bounds into learning algorithms, even for ω_ρ

The key observation to our work

$\mathrm{SWD}_p^p(\hat{\mu}_n, \hat{\nu}_n; \rho) = \mathbf{E}_{\mathbf{u}\sim\rho} W_p^p\left(\mathbf{u}_\sharp^* \hat{\mu}_n, \mathbf{u}_\sharp^* \hat{\nu}_n\right)$ **is** an aggregated risk, if the loss considered is W_p^p

Main results

Theorem (PAC Bayes Sliced Wasserstein)

With some conditions on the distributions μ *and* ν *captured by* $\varphi_{\mu,\nu,p}$ *and* $\psi_{\mu,\nu,p}(n) : \mathbb{N}^* \to \mathbb{R}_+$*, the following holds.*
Let $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$*.* $\forall \delta \in (0,1)$*, with prob. at least* $1-\delta$*:* $\forall \rho \in \mathcal{P}(\mathbb{S}^{d-1})$*,*

$$\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \frac{\lambda}{n}\varphi_{\mu,\nu,p} - \frac{1}{\lambda}\left\{KL(\rho||\rho_0) + \log\left(\frac{1}{\delta}\right)\right\} - \psi_{\mu,\nu,p}(n) \leq \mathrm{SW}_p^p(\mu, \nu; \rho)$$

Notes

- **Interpretation:** Generalization guarantees on the learned distribution ρ over population distribution μ, ν given training set μ_n, ν_n
- **Actionable feature:** maximize the l.h.s over ρ to maximize generalization over μ, ν
- Many levels of samplings, expectations, crux of the proof is to identify the right concentration phenomenon

Main results

Theorem (PAC Bayes Sliced Wasserstein)

With some conditions on the distributions μ *and* ν *captured by* $\varphi_{\mu,\nu,p}$ *and* $\psi_{\mu,\nu,p}(n) : \mathbb{N}^* \to \mathbb{R}_+$*, the following holds.*
Let $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$*.* $\forall \delta \in (0,1)$*, with prob. at least* $1-\delta$*:* $\forall \rho \in \mathcal{P}(\mathbb{S}^{d-1})$*,*

$$\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \frac{\lambda}{n}\varphi_{\mu,\nu,p} - \frac{1}{\lambda}\left\{KL(\rho||\rho_0) + \log\left(\frac{1}{\delta}\right)\right\} - \psi_{\mu,\nu,p}(n) \leq \mathrm{SW}_p^p(\mu, \nu; \rho)$$

Specific cases

- Bounded support measures with diam. Δ: $\varphi_{\mu,\nu,p} = \frac{\Delta^{2p}}{2}$, $\psi_{\mu,\nu,p}(n) \propto p\Delta^p n^{-1/2}$
- Sub-Gaussian measures of var. σ^2 and τ^2: $\varphi_{\mu,\nu,1} = \sigma^2 + \tau^2$, $\psi_{\mu,\nu,p}(n) \propto \frac{\log n}{\sqrt{n}}$.
- Bernstein moment condition (BMC). μ (σ^2, b)-BMC and ν (τ^2, c)-BMC, $\sigma_\star^2 \doteq \max(\sigma^2, \tau^2)$, $b_\star \doteq \max(b, c)$: $\varphi_{\mu,\nu,1}(\lambda, n) = 2\sigma_\star^2(n - 2b_\star\lambda)^{-1}$, $\psi_{\mu,\nu,p}(n) \propto \frac{\log n}{\sqrt{n}}$.

Derived Algorithm: Optimization of the Bound

In spirit

- Given a training dataset $\{(x_i, y_i)\}_{i=1}^{n}$ and a prior $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$, find $\rho^\star(\mu_n, \nu_n)$ such that

$$\rho^\star(\mu_n, \nu_n) = \arg\sup_{\rho \in \mathcal{F}} \mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \frac{\mathrm{KL}(\rho||\rho_0)}{\lambda}$$

The algorithmic way

- **Input:** dataset $\{(x_i, y_i)\}_{i=1}^{n}$, parameter λ, prior ρ_0, initialization $\rho^{(0)}$, nb. iterations T, LR η
- **for** $t \leftarrow 1$ to T
  - $\mathcal{L}(\rho^{(t-1)}) \leftarrow \mathrm{SW}_p^p(\mu_n, \nu_n; \rho^{(t-1)}) - \mathrm{KL}(\rho^{(t-1)}||\rho_0)/\lambda$
  - $\rho^{(t)} \leftarrow \rho^{(t-1)} + \eta \nabla_\rho \mathcal{L}(\rho^{(t-1)})$
- **Output:** $\rho^{(T)}$

Excerpt of numerical simulations

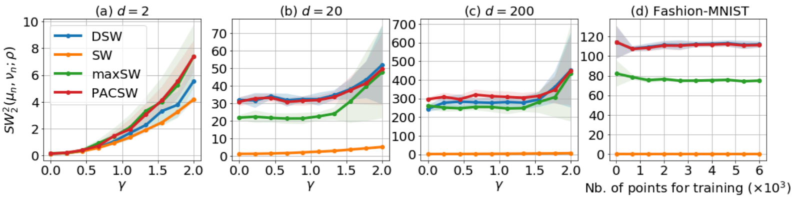


Figure 2. $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$ with (a-c) $\mu = \mathcal{N}(\mathbf{0}, \Sigma_d)$, $\nu = \mathcal{N}(\gamma\mathbf{1}, \Sigma_d)$, $n = 1000$, against γ, (d) classes 4 and 5 of Fashion-MNIST, against n. ρ is learned on the train set, and we report values on the test set.

Observations

- PACSW and DSW are always amongst the distances that generalize better
- Computing PACSW is, as of now, computationally demanding (KL estimation)

Conclusion and Outlooks

Conclusion

- First PAC Bayesian generalization bound on Adaptive Slice Wasserstein Distances
- Compelling numerical results
- DSW is a competitor, with less guarantees but more efficiency

Outlooks

- Further improve the computational efficiency of our algorithm
- Extend the usage to generative modelling (cf. paper)
- Connection between SWD and (Sparse) Principal Component Analysis
- Nothing to do with the above: a bandit approach to Sliced Wasserstein distances

Outline

General Conclusion

Real-world problems. Provide inspiration for academic research
Innovation and Transfer. A work on its own
Education. Key, at all levels, in all departments, in the entire society
Collaborations. AI is perfect place for cross-collaborations
Great experience!

Thanks

Outline

References I

Alquier, P. (2021).
User-friendly introduction to PAC-Bayes bounds.

Catoni, O. (2003).
A PAC-Bayesian approach to adaptive classification.
preprint LPMA 840.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. (2019).
Max-sliced wasserstein distance and its use for gans.
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10648--10656.

Nguyen, K., Ho, N., Pham, T., and Bui, H. (2021).
Distributional sliced-wasserstein and applications to generative modeling.
In International Conference on Learning Representations.

Ohana, R., Nadjahi, K., Rakotomamonjy, A., and Ralaivola, L. (2023).
Shedding a pac-bayesian light on adaptive sliced-wasserstein distances.