WASSERSTEIN PAC-BAYES LEARNING: ON THE INTRICATIONS BETWEEN GENERALISATION AND OPTIMISATION

Maxime Haddouche

INRIA Lille, MODAL Project-Team

Tuesday 12th September, 2023

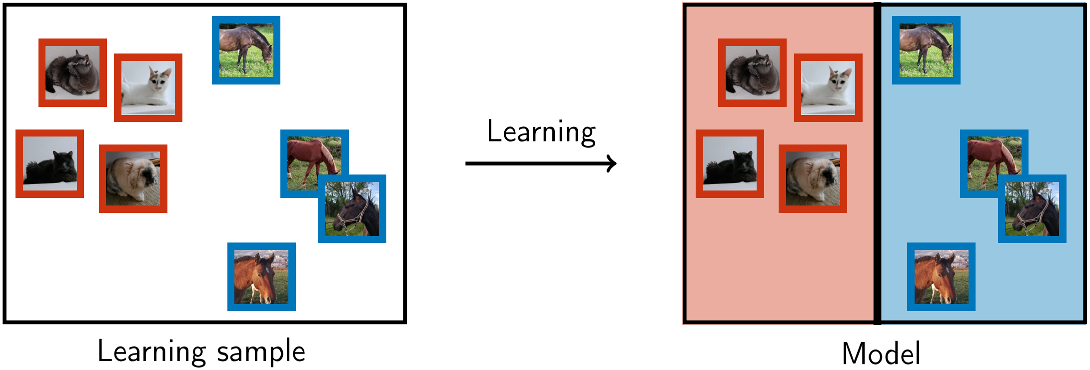
1. Introduction

2. Wasserstein PAC-Bayes to intricate generalisation and optimisation

3. Towards practical performances

Figures extracted from Paul Viallard's slides.

Example of supervised classification task: Predict if an image contains a cat or a horse

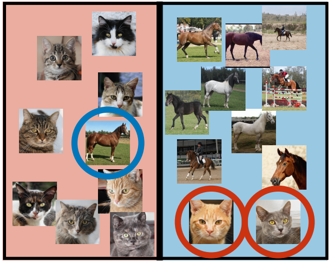


GENERALIZATION BOUNDS IN BATCH LEARNING

How many errors on the learning sample?
0 error!

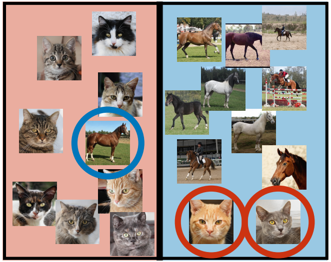
GENERALIZATION BOUNDS IN BATCH LEARNING

How many errors on the learning sample?
0 error!



How many errors on new examples?
3 errors...

How many errors on the learning sample?
0 error!



How many errors on new examples?
3 errors. . .

Can we have guarantees on the number of errors on new examples?

Generalization Bounds

$$\textit{true risk}(\text{pred}) \leq \textit{empirical risk}(\text{pred}) + \textit{complexity}(\text{pred, number of examples})$$

WHAT IS PAC-BAYES LEARNING?

- A branch of learning theory providing generalisation bounds
- Emerged in the late 90s with the works of Shawe-Taylor *et al.* (1997) and McAllester (1998, 1999).
- Recently proposed non-vacuous generalisation bounds valid during neural nets (NNs) training phase (no test set) (Dziugaite *et al.*, 2017)

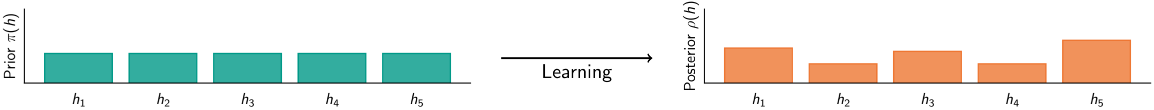
For more details see the recent surveys of:

1. Alquier (2021): `https://arxiv.org/abs/2110.11216`
2. Guedj (2019): `https://arxiv.org/abs/1901.05353`

BASIC SETTING

Setting:

- Model/predictor $h \in \mathcal{H}$, Data space $\mathcal{Z}$
- Loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0, 1]$
- m-sized learning sample $\mathcal{S} \in \mathcal{Z}^m$, $\mathcal{S} := \{\mathbf{z}_i\}_{i=1}^m \sim \mu^m$
- True risk $\mathrm{R}_\mu(h) = \mathbb{E}_{\mathbf{z}\sim\mu}\, \ell(h, \mathbf{z})$ and empirical risk $\mathrm{R}_\mu(h) = \frac{1}{m}\sum_{i=1}^m \ell(h, \mathbf{z}_i)$
- Space of distributions over $\mathcal{H}$: $\mathcal{M}(\mathcal{H})$
- PAC-Bayes: learning a posterior Q $\in \mathcal{M}(\mathcal{H})$ from a prior P $\in \mathcal{M}(\mathcal{H})$



McAllester's bound (Shawe-Taylor *et al.*, 1997; McAllester, 1998; Maurer, 2004)
For any prior P on $\mathcal{H}$, for any $\delta \in (0,1]$, we have with probability at least $1-\delta$ over $\mathcal{S} \sim \mu^m$ for all Q $\in \mathcal{M}(\mathcal{H})$

$$\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathrm{R}_{\mu}(h)\right] \leq \mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathrm{R}_{\mathcal{S}}(h)\right] + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathrm{Q}\|\mathrm{P}) + \ln\frac{2\sqrt{m}}{\delta}\right]}$$

where $\mathrm{KL}(\mathrm{Q}\|\mathrm{P}) = \mathbb{E}_{h\sim \mathrm{Q}} \ln\left(\frac{d\mathrm{Q}}{d\mathrm{P}}(h)\right)$

PAC-BAYESIAN BOUND IN BATCH LEARNING

McAllester's bound (Shawe-Taylor *et al.*, 1997; McAllester, 1998; Maurer, 2004)
For any prior P on $\mathcal{H}$, for any $\delta \in (0, 1]$, we have with probability at least $1 - \delta$ over $\mathcal{S} \sim \mu^m$ for all $Q \in \mathcal{M}(\mathcal{H})$

$$\mathop{\mathbb{E}}_{h\sim Q}\left[R_\mu(h)\right] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[R_\mathcal{S}(h)\right] + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right]}$$

where $\mathrm{KL}(Q\|P) = \mathbb{E}_{h\sim Q} \ln\left(\frac{dQ}{dP}(h)\right)$

- **No explicit dependency in the dimension of the problem** (potentially hidden in the KL term): potential tight bounds in practice (Dziugaite *et al.*, 2017, 2018; Pérez-Ortiz *et al.*, 2021).
- **Right-hand side is fully empirical**

A SIMPLE ROUTE OF PROOF

Step 1: A key ingredient: change of measure inequality

For any function f, any $Q \ll P$:

$$\underset{h\sim Q}{\mathbb{E}}[f(h)] - \ln\left(\underset{h\sim P}{\mathbb{E}}[\exp\circ f(h)]\right) \leq \mathrm{KL}(Q, P).$$

A SIMPLE ROUTE OF PROOF

Step 1: A key ingredient: change of measure inequality

For any function f, any $\mathrm{Q} \ll \mathrm{P}$:

$$\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[f(h)] - \ln\left(\mathop{\mathbb{E}}_{h\sim\mathrm{P}}[\exp\circ f(h)]\right) \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}).$$

Step 2: Markov's inequality

With probability at least $1-\delta$:

$$\begin{aligned}\mathop{\mathbb{E}}_{h\sim\mathrm{P}}[\exp\circ f(h)] &\leq \frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{P}}[\exp\circ f(h)]\right], \\ &= \frac{1}{\delta}\mathop{\mathbb{E}}_{h\sim\mathrm{P}}\left[\mathop{\mathbb{E}}_{\mathcal{S}}[\exp\circ f(h)]\right]. \qquad \text{(P data-free + Fubini)}\end{aligned}$$

A SIMPLE ROUTE OF PROOF (2)

Step 3: Choosing the right f.

Take $f((h) = m\,\mathrm{kl}\left(\mathrm{R}_{\mu}(h), \mathrm{R}_{\mathcal{S}}(h)\right)$ (kl= KL of Bernoullis).
Then Maurer (2004): for any h, loss in $[0,1]$:

$$\mathop{\mathbb{E}}_{\mathcal{S}}[\exp \circ f(h)] \leq 2\sqrt{m}$$

To conclude: $\mathrm{kl}(p,q) \geq 2(p-q)^2$.

TOWARDS PRACTICAL ALGORITHMS

High-probability PAC-Bayes bound = Generalisation-driven learning algorithm.

Catoni's PAC-Bayes algorithm (Alquier *et al.*, 2016, Theorem 4.1 subgaussian case): for $\lambda > 0$,

$$\mathrm{Q}^* := \mathrm{argmin}_{\mathrm{Q}} \mathop{\mathbb{E}}_{h\sim \mathrm{Q}} \left[\mathrm{R}_{\mathcal{S}}(h)\right] + \frac{\mathrm{KL}(\mathrm{Q}\|\mathrm{P})}{\lambda}$$

which leads to the explicit formulation of the **Gibbs posterior** $\mathrm{Q}^* := \mathrm{P}_{-\lambda \mathrm{R}_{\mathcal{S}}}$:

$$\frac{d\mathrm{Q}^*}{d\mathrm{P}}(h) = \frac{\exp\left(-\lambda \mathrm{R}_{\mathcal{S}}(h)\right)}{\mathbb{E}_{h\sim \mathrm{P}}\left[\exp\left(-\lambda \mathrm{R}_{\mathcal{S}}(h)\right)\right]}.$$

STRENGTHS OF PAC-BAYES

- Various PAC-Bayes algorithms can be derived and successfully applied to Stochastic NNs (Pérez-Ortiz *et al.*, 2021).
- PAC-Bayes is flexible enough to encompass various learning situations (bandits, reinforcement/online/meta/lifelong learning)
- PAC-Bayes holds for heavy-tailed losses (not only bounded/subgaussians) (Chugg *et al.*, 2023; Haddouche *et al.*, 2023a).

STRENGTHS OF PAC-BAYES

- Various PAC-Bayes algorithms can be derived and successfully applied to Stochastic NNs (Pérez-Ortiz *et al.*, 2021).
- PAC-Bayes is flexible enough to encompass various learning situations (bandits, reinforcement/online/meta/lifelong learning)
- PAC-Bayes holds for heavy-tailed losses (not only bounded/subgaussians) (Chugg *et al.*, 2023; Haddouche *et al.*, 2023a).

A major issue

Use of KL= impossible to consider Dirac measures (deterministic predictors)

Amit *et al.* (2022): replace KL divergence by Integral Probability Metrics. In particular: 1-Wasserstein is an IPM

Wasserstein distance

Given distance $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ and a Polish space $(\mathcal{A}, d)$, for any probability measures Q and P on $\mathcal{A}$, the Wasserstein distance is defined by

$$\mathrm{W}_1(\mathrm{Q}, \mathrm{P}) := \inf_{\gamma \in \Gamma(\mathrm{Q}, \mathrm{P})} \left\{ \mathop{\mathbb{E}}_{(a,b)\sim\gamma} d(a, b) \right\},$$

where $\Gamma(\mathrm{Q}, \mathrm{P})$ is the set of joint probability measures $\gamma \in \mathcal{M}(\mathcal{A}^2)$ such that the marginals are Q and P.

Amit *et al.* (2022): replace KL divergence by Integral Probability Metrics. In particular: 1-Wasserstein is an IPM

Wasserstein distance

Given distance $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ and a Polish space $(\mathcal{A}, d)$, for any probability measures Q and P on $\mathcal{A}$, the Wasserstein distance is defined by

$$\mathrm{W}_1(\mathrm{Q}, \mathrm{P}) := \inf_{\gamma \in \Gamma(\mathrm{Q}, \mathrm{P})} \left\{ \mathop{\mathbb{E}}_{(a,b) \sim \gamma} d(a, b) \right\},$$

where $\Gamma(\mathrm{Q}, \mathrm{P})$ is the set of joint probability measures $\gamma \in \mathcal{M}(\mathcal{A}^2)$ such that the marginals are Q and P.

Such a distance allows considering Dirac distributions, W_1 reduces to d in this case.

Kantorovich-Rubinstein duality

For any 1-Lipschitz function f:

$$\mathrm{W}_1(\mathrm{Q}, \mathrm{P}) \geq \mathop{\mathbb{E}}_{h \sim \mathrm{Q}}[f(h)] - \mathop{\mathbb{E}}_{h \sim \mathrm{P}}[f(h)]$$

REPLACING THE CHANGE OF MEASURE INEQUALITY

Kantorovich-Rubinstein duality

For any 1-Lipschitz function f:

$$W_1(Q, P) \geq \mathop{\mathbb{E}}_{h \sim Q}[f(h)] - \mathop{\mathbb{E}}_{h \sim P}[f(h)]$$

- This duality acts as a surrogate of change of measure for 1-Lipschitz functions
- Using it, Amit *et al.* (2022) recovered a McAllester-typed bound for finite classes of predictors.

Kantorovich-Rubinstein duality

For any 1-Lipschitz function f:

$$W_1(Q, P) \geq \mathop{\mathbb{E}}_{h \sim Q}[f(h)] - \mathop{\mathbb{E}}_{h \sim P}[f(h)]$$

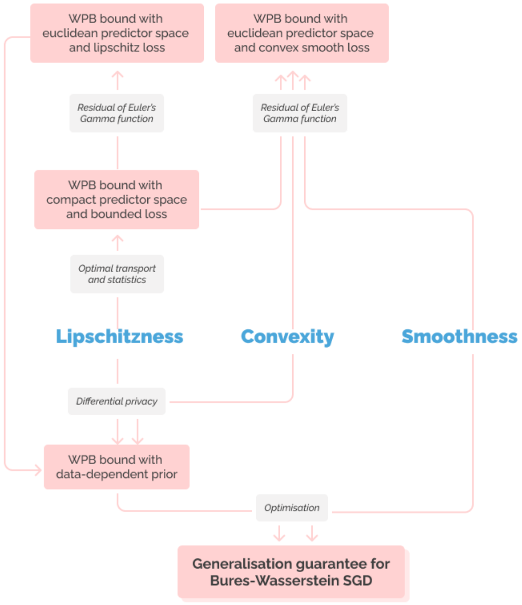
- This duality acts as a surrogate of change of measure for 1-Lipschitz functions
- Using it, Amit *et al.* (2022) recovered a McAllester-typed bound for finite classes of predictors.

1. Can we obtain high probability Wasserstein PAC-Bayes bounds (WPB) for infinite classes of predictors?
2. Are the geometric properties of the Waserstein useful in learning theory?
3. Can we obtain new generalisation-driven learning algorithms based on W_1?

PRESENTATION OF THE RESULTS

1. We obtain WPB bounds for infinite classes of predictors with a classical convergence rate $\mathcal{O}(1/\sqrt{m})$ at the cost of the curse of dimensionality. (Haddouche *et al.*, 2023b)
   $\mapsto$ Asymptotic yet interpretable guarantees
2. We show that it is possible to exploit the geometric convergence guarantees of the *Bures-Wasserstein SGD* to explain its generalisation ability (Haddouche *et al.*, 2023b)
3. We derive efficient learning algorithms from a WPB bound not implying the dimension at the cost of no explicit convergence rate. (Viallard *et al.*, 2023)

A LINK BETWEEN GENERALISATION AND OPTIMISATION



BEYOND KANTOROVICH-RUBINSTEIN DUALITY

Finite $\mathcal{H}$: Kantorovich-Rubinstein duality enough to obtain a sample-sized dependent lipschitz constant on f appearing (in the PB proof)

BEYOND KANTOROVICH-RUBINSTEIN DUALITY

Finite $\mathcal{H}$: Kantorovich-Rubinstein duality enough to obtain a sample-sized dependent lipschitz constant on f appearing (in the PB proof)
Such a property is not retrievable for infinite $\mathcal{H}$, need to find another tool

Finite $\mathcal{H}$: Kantorovich-Rubinstein duality enough to obtain a sample-sized dependent lipschitz constant on f appearing (in the PB proof)
Such a property is not retrievable for infinite $\mathcal{H}$, need to find another tool

Villani *et al.* (2009, Theorem 5.10)

Let $(\mathcal{X}, Q)$ and $(\mathcal{Y}, P)$ be two Polish probability spaces and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a nonnegative lower semicontinuous cost function:

$$\min_{\pi \in \Pi(Q,P)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) = \sup_{\substack{(\psi,\phi)\in L^1(Q)\times L^1(P) \\ \phi-\psi\leq c}} \left[\mathop{\mathbb{E}}_{Y\sim P}[\phi(Y)] - \mathop{\mathbb{E}}_{X\sim Q}[\phi(X)] \right],$$

where $L_1(P)$ refers to the set of all functions integrable with respect to P and the condition $\phi - \psi \leq c$ means that for all $x, y \in \mathcal{X} \times \mathcal{Y}, \phi(y) - \psi(x) \leq c(x,y)$.

A WPB BOUND FOR COMPACT PREDICTOR SPACE

Villani *et al.* (2009, Theorem 5.10) with $c_\varepsilon(x,y) = ||x-y|| + \varepsilon \to W_\varepsilon = W_1 + \varepsilon$
This + covering number tricks and PB route of proof gives a bound on the *generalisation gap* $\Delta_S(Q) = \mathbb{E}_{h\sim Q}[R_\mu(h) - R_S(h)]$:

A WPB BOUND FOR COMPACT PREDICTOR SPACE

Villani *et al.* (2009, Theorem 5.10) with $c_\varepsilon(x, y) = ||x - y|| + \varepsilon \to \mathrm{W}_\varepsilon = \mathrm{W}_1 + \varepsilon$
This + covering number tricks and PB route of proof gives a bound on the *generalisation gap* $\Delta_S(\mathrm{Q}) = \mathbb{E}_{h\sim \mathrm{Q}}[\mathrm{R}_\mu(h) - \mathrm{R}_S(h)]$:

Theorem

For any $\delta > 0$, assume that $\ell \in [0, 1]$ is K-Lipschitz wrt to h and that $\mathcal{H}$ is a compact of $\mathbb{R}^d$ bounded in norm by R. Let $\mathrm{P} \in \mathcal{P}_1(\mathcal{H})$ a (data-free) prior distribution. Then, with probability $1 - \delta$, for any posterior distribution $\mathrm{Q} \in \mathcal{P}_1(\mathcal{H})$:

$$|\Delta_S(\mathrm{Q})| \leq \sqrt{2K(2K+1)\frac{2d \log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(W_1(\mathrm{Q}, \mathrm{P}) + \varepsilon_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}},$$

with $\varepsilon_m = \mathcal{O}\left(1 + \sqrt{d\log(Rm)/m}\right)$.

- From now, $\mathcal{H} = \mathbb{R}^d$.
- $C_{\alpha,\beta,M} := \left\{\mathcal{N}(m,\Sigma) \in \mathrm{BW}(\mathbb{R}^d) \mid ||m|| \leq M,\ \alpha\mathrm{Id} \preceq \Sigma \preceq \beta\mathrm{Id}\right\}.$

- From now, $\mathcal{H} = \mathbb{R}^d$.
- $C_{\alpha,\beta,M} := \left\{\mathcal{N}(m, \Sigma) \in \mathrm{BW}(\mathbb{R}^d) \mid ||m|| \leq M,\ \alpha\mathrm{Id} \preceq \Sigma \preceq \beta\mathrm{Id}\right\}.$

Two sets of assumptions

- **(A1)** ℓ is uniformly K-Lipschitz over $\mathcal{H}$: for all $z, h \rightarrow \ell(h, z)$ is K-lipschitz, and $\sup_{z\in\mathcal{Z}} ||\ell(0, z)|| = D < +\infty$.
- **(A2)** For any $z \in \mathcal{Z}$, $\ell(., z)$ is continuously differentiable over $\mathcal{H}$, $\ell(., z)$ is also a convex L- smooth (*i.e*, its gradient is L-Lipschitz) and $\sup_{z\in\mathcal{Z}} ||\nabla_h \ell(0, z)|| = D < +\infty$.

Boundedness assumption is no longer required!

Theorem

Assume that $d \geq 3$, $\mathcal{H} = \mathbb{R}^d$ and that the (unbounded) loss satisfies **(A1)**. For any $\delta > 0, 0 \leq \alpha \leq \beta, M \geq 0$, let P $\in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, with probability $1-\delta$, for any posterior distribution Q $\in C_{\alpha,\beta,M}$, the following bound holds.

Asymptotic regime $(d \log(d) < \log(m))$

$$|\Delta_S(\mathrm{Q})| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1 + W_1(\mathrm{Q},\mathrm{P})\right) + \left(1 + K^2\log(m)\right)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

In all these formulas, $\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d), \log(m))$.

WPB BOUNDS FOR GAUSSIAN DISTRIBUTIONS (2)

Under **(A2)**, a similar bound can be reached (see Haddouche *et al.*, 2023b)

WPB BOUNDS FOR GAUSSIAN DISTRIBUTIONS (2)

Under **(A2)**, a similar bound can be reached (see Haddouche *et al.*, 2023b)

Tradeoff

Trading lipschitzness for smoothness has a cost: no constant K attenuating the impact of the dimension anymore.

TAKE-HOME MESSAGES

1. Bounds for low-data regime ($d \leq m$) and transitory regime ($m > d, \; d \log(d) \geq \log(m)$) are also available in the paper $\rightarrow$ worse dependencies in the dimension.
2. The Lipschitz constant attenuates the impact of the dimension.
3. PAC-Bayes with KL: statistical assumptions (*e.g.* boundedness). WPB involves geometric ones.

WPB WITH DATA-DEPENDENT PRIORS

Limitation

PAC-Bayes prior is arbitrary. Is it possible to replace the prior by the distribution we target?

Limitation

PAC-Bayes prior is arbitrary. Is it possible to replace the prior by the distribution we target?

Yes if the target is differentially private. Dziugaite *et al.* (2018) exploited that, when $\ell \in [0, 1]$, the Gibbs posterior is differentially private.

Limitation

PAC-Bayes prior is arbitrary. Is it possible to replace the prior by the distribution we target?

Yes if the target is differentially private. Dziugaite *et al.* (2018) exploited that, when $\ell \in [0,1]$, the Gibbs posterior is differentially private.

For lipschitz unbounded losses, it is possible to obtain a similar asymptotic bound than the Gaussian one by replacing the Gaussian prior P with the Gibbs posterior $\mathrm{Q}^* = \mathrm{P}_{-\frac{\lambda}{2K}}$

THE BURES-WASSERSTEIN SGD

A variational inference algorithm

Goal: find $\hat{\mathrm{Q}}$ the best Gaussian approximation of $\mathrm{Q}^* := \mathrm{P}_{-\frac{\lambda}{2K}\mathrm{R}_{\mathcal{S}}}$.

Algorithm 1: Bures-Wasserstein SGD.

Parameters : Strong convexity parameter $\alpha > 0$, radius $M > 0$; step size $\eta > 0$, initial mean m_0, initial covariance Σ_0

Set up $\hat{Q}_0 = \mathcal{N}(m_0, \Sigma_0)$.
for $k = 0..N-1$ **do**
  Draw a sample $X_k \sim \hat{Q}_k$.
  Set $m_k^+ = m_k - \eta \nabla V_S(X_k)$.
  Set $M_k = I - \eta(\nabla V^2(X_k) - \Sigma_k^{-1})$.
  Set $\Sigma_k^+ = M_k \Sigma_k M_k$.
  Set $m_{k+1} = \mathcal{P}_M(m_k^+)$, $\Sigma_{k+1} = \mathrm{clip}^{1/\alpha}\, \Sigma_k^+$.
  Set $\hat{Q}_{k+1} = \mathcal{N}(m_{k+1}, \Sigma_{k+1})$
end
Return $(\hat{Q}_k)_{k=1\ldots N}$.

THE BURES-WASSERSTEIN SGD (2)

Theorem

Assume having a smooth convex loss with a log-strongly convex prior. Under technical assumptions on $\eta, \hat{\mathrm{Q}}_0$, Bures-Wasserstein SGD satisfies for all $k \in \mathbb{N}$,

$$\mathbb{E} W_2^2\left(\hat{Q}_k, \hat{\mathrm{Q}}\right) \leq \exp(-\alpha k \eta) W_2^2\left(\hat{Q}_0, \hat{\mathrm{Q}}\right) + \frac{36 d \eta}{\alpha^2}.$$

In particular, $\mathbb{E} W_2^2\left(\hat{Q}_k, \hat{\mathrm{Q}}\right) \leq \varepsilon^2$ with suitable η, k.

BURES-WASSERSTEIN SGD GENERALISES!

Main assumptions (see Haddouche *et al.* (2023b) for technical ones

(A3): $\mathcal{H} = \mathbb{R}^d$ ℓ is twice differentiable, L-smooth, convex and uniformly K-Lipschitz over $\mathcal{H}$.
$\mathrm{P} = \mathcal{N}(0, \Sigma)$ with $\Sigma = \mathrm{diag}(\gamma), 1 \geq \gamma > 0$. Also $\lambda \leq 2K$ in the definition of Q^*.

Theorem (informal)

Assume **(A3)**, $d \geq 3$. Let $\beta_m = \mathcal{O}(1/\sqrt{m})$ and fix any $\beta_m < \delta < 1$. Bures-Wasserstein SGD, with adapted initialisation and parameters η, N satisfies, with probability $1 - 2\delta$:
Asymptotic regime $(d \log(d) < \log(m))$

$$|\Delta_S(\hat{Q}_N)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1 + W_1(\hat{\mathrm{Q}}, \mathrm{Q}^*)\right) + (1 + K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right),$$

where $\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d), \log(m))$.

CONCLUSION

Take-home messages

- Geometric optimisation guarantees are useful to explain generalisation
- Gaussian approximations are costful (if not well-suited) for generalisation.
- A good Lipschitz constant can compensate the impact of dimensionality

CONCLUSION

Take-home messages

- Geometric optimisation guarantees are useful to explain generalisation
- Gaussian approximations are costful (if not well-suited) for generalisation.
- A good Lipschitz constant can compensate the impact of dimensionality

What is next?

- Our WPB bounds suffers from the explicit impact of the dimension. Can we avoid it, as in classical PAC-Bayes?
- Can we relax the Lipschitzness assumption? It was crucial for differential privacy, but might be replaced elsewhere (e.g. by smoothness).
- 2-Wasserstein distance catches more efficiently the geometry of the predictor space, could we avoid the use of the Kantorovich-Rubinstein duality to directly exploit this distance instead of using W_1 as intermediary?

TOWARDS PRACTICAL PERFORMANCES

Previous results are meaningful asymptotically because of the impact of dimension. **Can we remove this constraint?**

TOWARDS PRACTICAL PERFORMANCES

Previous results are meaningful asymptotically because of the impact of dimension. **Can we remove this constraint?**

Yes! At the cost of no explicit convergence rate.

TOWARDS PRACTICAL PERFORMANCES

Previous results are meaningful asymptotically because of the impact of dimension. **Can we remove this constraint?**

Yes! At the cost of no explicit convergence rate.

Various advantages

- No explicit dimension term
- Allows easily heavy-tailed losses
- Allows easily non-iid data

Idea: split $\mathcal{S}$ into L parts $\mathcal{S}_1, ..., \mathcal{S}_L$ and exploit supermartingale techniques.

Assumptions:

- ℓ is non-negative and K-Lipschitz
- for any $1 \leq i \leq L, \mathcal{S}$, $\mathbb{E}_{h\sim \mathrm{P}_i(.,\mathcal{S}), z\sim\mu}\left[\ell(h,z)^2\right] \leq 1$
- Prior $\mathrm{P}_{i,\mathcal{S}}$ depend on $\mathcal{S}/\mathcal{S}_i$.

Theorem

For any $\delta \in (0,1]$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, the following holds for the distributions $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(\mathcal{S}, .)$ and for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathrm{R}_\mu(h) - \hat{\mathrm{R}}_{\mathcal{S}}(h)\right] \leq \sum_{i=1}^{L} \frac{2|\mathcal{S}_i|K}{m} \mathrm{W}(\mathrm{Q}, \mathrm{P}_{i,\mathcal{S}}) + \sum_{i=1}^{L} \sqrt{\frac{|\mathcal{S}_i| \ln \frac{L}{\delta}}{m^2}},$$

where $\mathrm{P}_{i,\mathcal{S}}$ *does not* depend on $\mathcal{S}_i$.

Remark

The previous bound if vacuous if $K = m$ (online setting)

ONLINE COUNTERPART FOR NON IID DATA

Remark

The previous bound if vacuous if $K = m$ (online setting)

Solution

The same set of technique allows a refined bound for online learning (see Viallard *et al.*, 2023, Theorems 3&4)

Why is it great?

- Zero assumption about the data distribution
- Still valid for heavy tailed losses
- Consider a sequence of priors/posteriors $\rightarrow$ more flexible.

NEW OPTIMISATION GOALS

Batch

$$\mathrm{argmin}_{h_{\mathbf{w}}\in\mathcal{H}} \left\{ \hat{\mathrm{R}}_{\mathcal{S}}(h_{\mathbf{w}}) + \varepsilon \left[\sum_{i=1}^{K} \frac{|\mathcal{S}_i|}{m} \|\mathbf{w} - \mathbf{w}_i\|_2 \right] \right\}.$$

Online

$$\forall i \geq 1, \quad h_i \in \mathrm{argmin}_{h_{\mathbf{w}}\in\mathcal{H}} \ell(h_{\mathbf{w}}, \mathbf{z}_i) + \|\mathbf{w} - \mathbf{w}_{i-1}\| \\ s.t. \ \ \|\mathbf{w} - \mathbf{w}_{i-1}\| \leq 1.$$

EXPERIMENTS

Classification problem on MNIST solved with linear models and fully connected neural networks.

(a) Linear model – batch learning

Dataset	Alg. 1 ($\frac{1}{m}$)		Alg. 1 ($\frac{1}{\sqrt{m}}$)		ERM	
	$\mathfrak{R}_S(h)$	$\mathfrak{R}_\mu(h)$	$\mathfrak{R}_S(h)$	$\mathfrak{R}_\mu(h)$	$\mathfrak{R}_S(h)$	$\mathfrak{R}_\mu(h)$
ADULT	.165	**.166**	.165	.167	.166	.167
FASHIONMNIST	.128	.151	.126	**.148**	.139	.153
LETTER	.285	.297	.287	**.296**	.287	.297
MNIST	.200	.216	.066	.092	.065	**.091**
MUSHROOMS	.001	**.001**	.001	**.001**	.001	**.001**
NURSERY	.766	**.773**	.760	**.773**	.794	.807
PENDIGITS	.049	**.059**	.050	.061	.052	.064
PHISHING	.063	**.067**	.065	.069	.064	**.067**
SATIMAGE	.144	**.200**	.138	.201	.148	.209
SEGMENTATION	.057	**.216**	.164	.386	.087	.232
SENSORLESS	.129	**.129**	.131	.131	.134	.136
TICTACTOE	.388	.299	.013	**.021**	.228	.238
YEAST	.527	.497	.524	.504	.470	**.427**

(b) Linear model – online learning

Alg. 2		OGD	
$\mathfrak{C}_S$	$\mathfrak{C}_\mu$	$\mathfrak{C}_S$	$\mathfrak{C}_\mu$
.230	**.236**	.248	.248
.223	**.282**	.540	.548
.919	.935	.916	**.926**
.284	**.310**	.378	.397
.218	.222	.082	**.087**
.794	.807	.789	**.805**
.342	**.484**	.589	.600
.226	.242	.226	**.220**
.669	.938	.635	**.888**
.749	**.803**	.738	.893
.906	.910	.825	**.830**
.443	.468	.390	**.303**
.699	.713	.667	**.708**

(c) NN model – batch learning

Dataset	Alg. 1 ($\frac{1}{m}$)		Alg. 1 ($\frac{1}{\sqrt{m}}$)		ERM	
	$\mathfrak{R}_S(h)$	$\mathfrak{R}_\mu(h)$	$\mathfrak{R}_S(h)$	$\mathfrak{R}_\mu(h)$	$\mathfrak{R}_S(h)$	$\mathfrak{R}_\mu(h)$
ADULT	.164	.164	.166	.165	.165	**.163**
FASHIONMNIST	.159	.163	.156	**.160**	.163	.167
LETTER	.259	.272	.250	**.260**	.258	.270
MNIST	.112	.120	.084	**.094**	.119	.127
MUSHROOMS	.000	**.000**	.000	**.000**	.000	**.000**
NURSERY	.706	**.719**	.706	**.719**	.706	**.719**
PENDIGITS	.009	.023	.021	.032	.009	**.022**
PHISHING	.042	**.050**	.039	.054	.046	.055
SATIMAGE	.132	.184	.149	**.172**	.141	.189
SEGMENTATION	.145	**.250**	.189	.373	.174	.389
SENSORLESS	.076	.079	.077	.079	.075	**.078**
TICTACTOE	.392	.301	.000	.038	.000	**.023**
YEAST	.679	.666	.487	**.478**	.644	.682

(d) NN model – online learning

Alg. 2		OGD	
$\mathfrak{C}_S$	$\mathfrak{C}_\mu$	$\mathfrak{C}_S$	$\mathfrak{C}_\mu$
.241	.254	.248	**.248**
.096	**.327**	.397	.446
.829	**.945**	.958	.963
.092	**.265**	.470	.521
.082	**.122**	.202	.217
.800	**.805**	.793	.806
.323	**.537**	.871	.879
.164	**.222**	.331	.318
.401	**.763**	.626	.857
.619	**.857**	.739	.913
.899	.910	.622	**.633**
.388	**.309**	.397	**.309**
.662	**.720**	.702	**.720**

Thank you for your attention!

Questions?

REFERENCES

John Shawe-Taylor and Robert Williamson. A PAC Analysis of a Bayesian Estimator. *COLT*. (1997).

David McAllester. Some PAC-Bayesian Theorems. *COLT*. (1998).

David McAllester. Some PAC-Bayesian Theorems. *Machine Learning*. (1999).

Andreas Maurer. A Note on the PAC Bayesian Theorem. *CoRR*. cs.LG/0411099. (2004).

Cédric Villani *et al.* Optimal transport: old and new. Vol. 338. *Springer*. (2009).

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*. (2016).

Gintare Karolina Dziugaite and Daniel Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *UAI*. (2017).

REFERENCES

Gintare Karolina Dziugaite and Daniel Roy. Data-dependent PAC-Bayes priors via differential privacy. *NeurIPS*. (2018).

Benjamin Guedj. A Primer on PAC-Bayesian Learning. *CoRR*. abs/1901.05353. (2019).

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *CoRR*. abs/2110.11216. (2021).

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter Risk Certificates for Neural Networks. *Journal of Machine Learning Research*. (2021).

Ron Amit, Baruch Epstein, Shay Moran, and Ron Meir. Integral Probability Metrics PAC-Bayes Bounds. *Conference on Neural Information Processing Systems (NeurIPS)*. (2022).

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A unified recipe for deriving (time-uniform) PAC-Bayes bounds. *CoRR*. abs/2302.03421. (2023).

REFERENCES

Maxime Haddouche and Benjamin Guedj. PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales. *Transactions on Machine Learning Research*. (2023).

Maxime Haddouche and Benjamin Guedj. Wasserstein PAC-Bayes Learning: A Bridge Between Generalisation and Optimisation. *CoRR*. abs/2304.07048. (2023).

Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. *arXiv preprint arXiv:2306.04375*. (2023).