Learning with BOT - Bregman and Optimal Transport divergences

Andrew Chee, Sébastien Loustau

Submitted to Bernoulli

Learning with BOT - Bregman and Optimal Transport divergences

ANDREW CHEE[1], and SÉBASTIEN LOUSTAU[2]

[1]*Cornell University, Operations Research and Information Engineering,Ithaca, New-York*
E-mail: ac2766@cornell.edu
[2]*Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques et de leurs Applications de Pau, France*
E-mail: sebastien.loustau@univ-pau.fr

The introduction of the Kullback-Leibler divergence in PAC-Bayesian theory can be traced back to the work of [1]. It allows to design learning procedure with generalization errors based on an optimal trade-off between accuracy on the training set, and complexity. This complexity is penalized thanks to the Kullback-Leibler divergence from a prior distribution, modeling a domain knowledge over the set of candidates or weak learners. In the context of high dimensional statistics, it gives rise to sparsity oracle inequalities or more recently sparsity regret bounds, where the complexity is measured thanks to ℓ_0 or ℓ_1-norms. In this paper, we propose to extend the PAC-Bayesian theory to get more generic regret bounds for sequential weighted averages, where (1) the measure of complexity is based on any ad-hoc criterion and (2) the prior distribution could be very simple. These results arise by introducing a new measure of divergences from the prior in terms of Bregman divergence or Optimal Transport.

Keywords: PAC-Bayesian theory; Convex duality; Bregman divergence; Optimal transport; Sequential learning; Regret bounds

1. Introduction

1.1. PAC-Bayesian theory

PAC-Bayesian machine learning theory can be traced back to the work of Shawe-Taylor and Williamson (see [2, 3]) and Mac-Allester ([4, 1]). The goal of PAC-Bayesian theory is to get distribution free generalization bounds, ie that holds a priori, but where some prior or arbitrary domain knowledge is available on the set of candidate models. As in Bayesian statistics, it leads to a posteriori estimates of the generalization performances based on informative priors. Historically, it was proposed as an alternative to the structural risk minimization problem based on the Vapnik-Chervonenkis theory since Bayesian algorithms minimize a risk expression involving a likelihood or goodness of fit term based on the training data, and a prior probability, leading to a trade-off between empirical accuracy and complexity in terms of divergence from the prior. Formally, in the discrete case, for a set of finite weak learners $\mathcal{G} = \{g_1, \dots, g_p\}$ and a prior distribution $\pi = (\pi_k)_{k=1}^p$ over this set $\mathcal{G}$, the first PAC-Bayesian bound appears in [4]. Given a loss function $\ell(g, \cdot)$ and a distribution S over a sample $z_1, \dots, z_m$, [4] shows the existence of a decision $\hat{g} \in \mathcal{G}$ such that with probability $1 - \delta$, the generalization error $R(\cdot) := \mathbb{E}_{z\sim S}\ell(\cdot, z)$ of $\hat{g}$ is bounded as follows:

$$R(\hat{g}) \leq \min_{g\in\{g_1,\dots,g_p\}} \left(\frac{1}{m}\sum_{t=1}^{m} \ell(g, z_t) + \sqrt{\frac{\log\frac{1}{\pi(g)} + \log\frac{1}{\delta}}{2m}} \right). \tag{1}$$

It leads to the model selection of a particular candidate $\hat{g} \in \mathcal{G}$ that trades off the goodness of fit with the minimum description length $-\log \pi(\hat{g})$ (see [5]). The result holds for any prior π but is interesting if there exists a prior π giving high probabilities on rules $g \in \mathcal{G}$ that fit well to the training problem. In term of domain knowledge, this is equivalent to suppose that particular candidates $g \in \mathcal{G}$ with low complexity fit well to the learning problem. This model selection principle is outperformed in [1] where (1) is extended to uncountable set $\mathcal{G}$. In this case, a stochastic algorithm is prefered and leads to the existence of a distribution $\hat{\rho} \in \mathcal{P}(\mathcal{G})$ the set of probability distributions over $\mathcal{G}$ such that with probability $1-\delta$, the expected generalization error is bounded as follows:

$$\mathbb{E}_{g\sim\hat{\rho}} R(g) \leq \inf_{\rho\in\mathcal{P}(\mathcal{G})} \left(\mathbb{E}_{g\sim\rho} \frac{1}{m} \sum_{t=1}^{m} \ell(g, z_t) + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\delta} + \log m + 2}{\lambda} \right), \tag{2}$$

where $\mathcal{K}(\rho,\pi)$ is the Kullback-Leibler divergence from prior π to ρ. Noting that since $\mathcal{K}(\delta_g,\pi) = -\log\pi(g)$ for any π and g, (2) appears as a powerfull generalization of (1). Minimizing the RHS in (2) corresponds to compute the convex conjugate of $\mathcal{K}(\cdot,\pi)$ and leads to a stochastic algorithm based on the following exponential weighted average or Gibbs posterior:

$$\hat{\rho}(g) := \frac{1}{Z_\lambda} \exp\left(-\frac{\lambda}{m} \sum_{t=1}^{m} \ell(g, z_t) \right) \pi(g), \tag{3}$$

where $\lambda > 0$ is a temperature parameter. This distribution reaches a trade-off between accuracy over the sample and Kullback-Leibler divergence from the prior, where the introduction of the Kullback-Leibler divergence is based on a Chernoff bound in [1, Lemma 4].

Based on these theoretical foundations, PAC-Bayesian theory has been widely used in high dimensional statistics, where $g = \sum_{i=1}^{p} \theta_i f_i \in \mathcal{G}$ often represents the linear span of a set of dictionary functions $\{f_1, \ldots, f_p\}$, where p is potentially huge. In this context, a sparsity scenario means that a small subset of the dictionary provides a nearly complete description of the underlying phenomenon. [6] proves generalization performances under this sparsity scenario (see also [7, 8, 9]). Then, instead of using a model selection technique leading to sparse estimators such as the lasso (see [10, 11]), [6] proposes a stochastic mirror averaging that reach a PAC-Bayesian inequality as in (2). Using a sparsity prior, one gets sparsity oracle inequalities. These techniques provide theoretical trade-off between goodness of fit and complexity in terms of ℓ_0 or ℓ_1-norm of the solution. Whereas these complexity measures have been widely used in the literature (see [12] for an application in deep learning), we expect in this paper more generic penalties to use any ad-hoc criterion for the set of candidates learning machines.

1.2. Main contributions

In this contribution we provide PAC-Bayesian bounds such as (2) replacing the Kullback-Leibler divergence and its usual convex conjugate (3) with Bregman and optimal transport divergences. We first extend the materials presented above to a more general divergence then the Kullback-Leibler divergence, namely Bregman divergences. In this context, we show that the two main ingredients, namely the convex duality gathering with the cancellation argument originated in [13] can be still applied to our context. It leads to a new family of stochastic algorithms that generalize the previous exponential weighted averages with equivalent theoretical guarantees. We also propose to introduce optimal transport as a promising alternative to Bregman (or Kullback) divergences. Indeed, by proving the main assumption originated in [14] (see Assumption $\mathbb{A}(\Pi, \delta_\lambda)$ below), we lead to a new kind of procedure where the exponential averages are replaced by optimal transport

optimization that satisfies a new PAC-Bayesian bound. With these theoretical results, we expect numerous possible regret bounds and new theoretical trade-off between goodness of fit and different complexity measures related with energy constraints (see [12] for a survey of some recent advances in the environmental impact and electirc consumption of learning machines).

In the sequel, we adopt the online learning scenario where a sequence of deterministic values $\{z_t, t = 1, \ldots, T\}$ is observed, where $z_t \in \mathcal{Z}$ could be a couple (x_t, y_t) in supervised learning, or an input data in unsupervised learning. Based on a loss function $\ell(g, z)$ that measures the loss of decision $g \in \mathcal{G}$ at observation $z \in \mathcal{Z}$, the goal of the forecaster is to build a sequence of distributions $\{\hat{\rho}_t, t = 1, \ldots, T\}$ with small expected cumulative loss. We are hence looking at regret bounds of the following type:

$$\sum_{t=1}^{T} \mathbb{E}_{g\sim\hat{\rho}_t} \ell(z_t, g) \leq \inf_{g\in\mathcal{G}} \left(\sum_{t=1}^{T} \ell(z_t, g) + \text{pen}(g) \right) + \delta_T, \tag{4}$$

where $\text{pen}(g)$ extends the sparsity paradigm where $\text{pen}(g) = \|g\|_{0,1}$ thanks to the introduction of Bregman divergences in (2), and $\delta_T > 0$ is a residual term. To build this sequence of mixtures, we act sequentially as follows. At each round $t \geq 1$, we have timely access to a temporal prior $\hat{\rho}_t$ derived from the previous observations. More precisely, given z_t, we are looking at a randomized decision $g \sim \hat{\rho}_{t+1}$ where the posterior distribution $\hat{\rho}_{t+1} \in \mathcal{P}(\mathcal{G})$ is the solution of the following minimization:

$$\hat{\rho}_{t+1} := \arg \min_{\rho\in\mathcal{P}(\mathcal{G})} \left\{ \mathbb{E}_{g\sim\rho} h_t(g) + \frac{1}{\lambda} \mathcal{D}(\rho, \hat{\rho}_t) \right\}, \tag{5}$$

where $h_t(g)$ is related with the loss of decision $g \in \mathcal{G}$ at z_t and $\mathcal{D}(\rho, \hat{\rho}_t)$ is a suitable divergence from the temporal prior to the candidate posterior ρ. Parameter $\lambda > 0$ governs the trade-off between the two terms. Depending on the choice of $\lambda > 0$ and the penalty above, our procedure reaches automatically the desire trade-off between fitting the data and penalty in terms of domain knowledge. It is important to note that minimization (5) is equivalent to compute the Legendre–Fenchel transformation of $\mathcal{D}(\cdot, \hat{\rho}_t)$ at each step $t \geq 1$, and leads to different kind of posterior distribution. In this paper, we investigate two divergences:

- Bregman divergences $\mathcal{D}(\rho, \pi) = B_\Phi(\rho, \pi)$ (see Section 2.1), where Φ governs the explicit form of the penalty as well as the sequence of distributions $\{\hat{\rho}_t, t = 1 \ldots, T\}$,
- Optimal transport $\mathcal{D}(\rho, \pi) = \mathcal{W}_\alpha(\rho, \pi)$ (see Section 2.2), where the introduction of a cost function $C(\cdot, \cdot) : \mathcal{G} \times \mathcal{G}$ lead to more flexibility.

Recall that in the PAC-Bayesian literature cited above, the proposed sequential procedure based on minimization (5) uses the classical Kullback-Leibler divergence where $\mathcal{D}(\rho, \pi) = \mathcal{K}(\rho, \pi)$. It leads to the vanilla Gibbs posterior (3) since in this case the unique solution of (5) can be written explicitly as follows:

$$\hat{\rho}_{t+1}(dg) = \frac{1}{Z_t} \exp(-h_t(g)) d\hat{\rho}_t(g) = \cdots = \frac{1}{Z} \exp\left(-\sum_{u=1}^{t} h_t(g) \right) d\pi(g),$$

where $\hat{\rho}_1 = \pi$ is the prior distribution.

Finally, the main results of the paper occur under an assumption on the same flavour as [7]. It can be generalized in our context as follows:

Assumption $\mathbb{A}(\Pi, \delta_\lambda)$ $\forall \pi \in \mathcal{P}(\mathcal{G}), \exists \Pi(\pi) \in \mathcal{P}(\mathcal{G}) : \forall z \in \mathcal{Z}$:

$$\mathbb{E}_{g'\sim\Pi(\pi)} \ell(g', z) \leq \frac{1}{\lambda} \mathbb{E}_{g'\sim\Pi(\pi)} \min_{\rho\in\mathcal{P}(\mathcal{G})} \left\{ \mathbb{E}_{g\sim\rho} [\lambda(\ell(g, z)) + \delta_\lambda(z, g, g')] + \mathcal{D}(\rho, \pi) \right\}.$$

Below we show particular cases when $\mathbb{A}(\Pi, \delta_\lambda)$ holds for both Bregman divergences and Optimal Transport leading to regret bounds for weighted averages built sequentially following Algorithm 1 below.

Algorithm 1 General Algorithm

init. $\lambda > 0$, $t = 1$, $\pi \in \mathcal{P}(\mathcal{G})$ a prior distribution. $\mathcal{D}$ a divergence such that $\mathbb{A}(\Pi, \delta_\lambda)$ holds.
Let $\tilde{\rho}_1 = \Pi \circ \pi$.
repeat Predict z_t according to $\hat{g}_t \sim \tilde{\rho}_t$. Observe z_t and compute

$$h_t(g) = \ell(z_t, g) + \delta_\lambda(z_t, g, \hat{g}_t) \quad \text{for all } g \in \mathcal{G}.$$

return $\tilde{\rho}_{t+1} = \Pi \circ \hat{\rho}_{t+1}$ where Π is defined in $\mathbb{A}(\Pi, \delta_\lambda)$ and $\hat{\rho}_{t+1}$ is the solution of the general optimisation problem:

$$\min_{\rho \in \mathcal{P}(\mathcal{G})} \left\{ \mathbb{E}_{g \sim \rho} \left(\ell(g, z_t) + \delta_\lambda(z_t, g, \hat{g}_t) \right) + \mathcal{D}(\rho, \hat{\rho}_t) \right\}$$

$t = t + 1$
until $t = T + 1$

2. PAC-Bayesian Inequalities

In this section we propose to state the main results of the paper. In the sequel, we consider a measurable space $(\mathcal{G}, \Omega)$ endowed with a σ-finite measure ν, where $\mathcal{G}$ is a set of decisions. We denote by $\mathcal{P}(\mathcal{G}) := \{\rho : \mathcal{G} \to \mathbb{R}_+ : \int \rho(g)\nu(dg) = 1\}$ the set of probability measure on $\mathcal{G}$. The main objective of PAC-Bayesian theory is to propose data-dependent posterior distribution $\mathcal{P}(\mathcal{G})$ with some theoretical guarantees. In Theorem 2.2, we control the expected cumulative loss of Algorithm 2, where Bregman divergences are proposed as regularizers. It generalizes the classical setting based on the usual Kullback-Leibler divergence. In Theorem 2.4, we go one step further and propose to introduce optimal transport as a promising alternative to measure the divergences to the prior. It leads to a control of the expected cumulative loss of Algorithm 3 where the introduction of a generic cost function allows to launch more generic regularizers and regret bounds in Corollary 2.5 and Corollary 2.6.

2.1. Bregman divergences

Given a strictly convex function $\Phi : \mathcal{P}(\mathcal{G}) \mapsto \mathbb{R}$, twice-continuously Fréchet-differentiable on $\mathrm{relint}_p \mathcal{P}(\mathcal{G})$, the relative interior of $\mathcal{P}(\mathcal{G})$ with respect to $L^p(\nu)$, we define the functional Bregman divergence between two probabilities on $\mathcal{P}(\mathcal{G})$ as follows:

$$\mathcal{B}_\Phi(\rho, \mu) := \Phi(\rho) - \Phi(\mu) - \langle \nabla\Phi(\mu), \rho - \mu \rangle, \tag{6}$$

where $\nabla\Phi(\mu)$ stands for the Fréchet derivative of Φ at point μ. Equation (6) generalizes the standard vector Bregman divergence $B_\Phi(u, v) = \Phi(u) - \Phi(v) - \nabla\Phi(v)^T(u - v)$ for $u, v \in \mathbb{R}^p$ where $\nabla\Phi(v)$ is the standard gradient of Φ at point v. Morever, as a particular case, we also introduce the class of pointwise Bregman divergences introduced by Csiszár in [15] as:

$$\mathcal{B}_s(\rho, \mu) = \int_{\mathcal{G}} s(\rho(g)) - s(\mu(g)) - s'(\mu(g))(\rho(g) - \mu(g))\nu(dg), \tag{7}$$

where $s : \mathbb{R}_+ \to \mathbb{R}$ is constrained to be differentiable and strictly convex and the limit $\lim_{x\to 0} s(x)$ and $\lim_{x\to 0} s'(x)$ must exist. By definition, B_s is non-negative and satisfies main properties of classical Bregman divergence such as convexity, linearity, and generalized Pythagorean theorem (see for instance [16] for a proof). An important fact with this pointwise Bregman divergence is the existence of an equivalent functional $\Phi : \mathcal{P}(\mathcal{G}) \mapsto \mathbb{R}$ and a functional Bregman divergence for any pointwise Bregman divergence if the measure ν is finite. However, the reverse is not true. Different function s in pointwise Bregman divergences (7) leads to different Bregman divergences. For instance, let $s(x) = x \log x$. Then $B_s(\rho, \mu) = \mathcal{K}(\rho, \mu)$ where $\mathcal{K}(\cdot, \cdot)$ is the Kullback-Leibler divergence. Moreover, if $s(x) = x^2$, then $B_s(\rho, \mu) = \|\rho - \mu\|_{2,\nu}$ whereas if $s(x) = -\log x$ we find the Itakura-Saito distance.

The following lemma is useful to state the main theoretical result.

Lemma 2.1. *Let $\pi \in \mathcal{P}(\mathcal{G})$ where $\mathcal{G}$ is finite. For some function $h : \mathcal{G} \to \mathbb{R}^+$, let us consider the minimization problem:*

$$\min_{\rho\in\mathcal{P}(\mathcal{G})} \left\{ \mathbb{E}_{g\sim\rho}[h(g)] + \mathcal{B}(\rho, \pi) \right\}, \tag{8}$$

where $\mathcal{B} \in \{\mathcal{B}_\Phi, \mathcal{B}_s\}$ is a Bregman divergence according to definition (6) *or* (7)*.*

- *If $\mathcal{B} = \mathcal{B}_\Phi$ for $\Phi : \mathcal{P}(\mathcal{G}) \to \mathbb{R}$ a strictly convex and continuously differentiable function, and $h(\cdot)$ is convex and continuously differentiable, the minimization problem* (8) *admits an unique solution $\hat{\rho}_{h,\pi}$ such that:*

$$\hat{\rho}_{h,\pi}(g) = \max\left(0, ((\nabla_g \Phi)^{-1}\left(\nabla_g \Phi(\pi(g)) - h(g) + c_{h,\pi}\right))\right), \forall g \in \mathcal{G},$$

  where $c_{h,\pi} > 0$ is uniquely defined such that $\sum_{g\in\mathcal{G}} \hat{\rho}_{h,\pi}(g) = 1$.
- *More generally if a minimizer $\hat{\rho}_{h,\pi}$ of* (8) *exists, we have:*

$$\hat{\rho}_{h,\pi}(g) = \max\left(0, ((\nabla_g \Phi)^{-1}\left(\nabla_g \Phi(\pi(g)) - h(g) + c_{h,\pi}\right))\right), \forall g \in \mathcal{G},$$

  for some constant $c_{h,\pi} > 0$ such that $\sum_{g\in\mathcal{G}} \hat{\rho}_{h,\pi}(g) = 1$.
- *If $\mathcal{B} = \mathcal{B}_s$ for some differentiable and strictly convex $s : \mathbb{R}_+ \to \mathbb{R}$ such that the limit $\lim_{x\to 0} s(x)$ as well as $\lim_{x\to 0} s'(x)$ must exist and $\lim_{x\to 0} s'(x) = -\infty$, the minimization problem* (8) *admits an unique solution*

$$\hat{\rho}_{h,\pi}(g) = (s')^{-1}(s'(\pi(g)) - h(g) + c_{h,\pi}),$$

  where $c_{h,\pi} > 0$ is uniquely defined such that $\sum_{g\in\mathcal{G}} \hat{\rho}_{h,\pi}(g) = 1$.

Moreover, for any sequence $(h_t)_{t=1}^T$ and prior distribution $\hat{\rho}_1 \in \mathcal{P}(\mathcal{G})$, under the previous assumptions, we have:

$$\hat{\rho}_{T+1} = \hat{\rho}_{\sum_{t=1}^T h_t, \hat{\rho}_1}, \tag{9}$$

where $\hat{\rho}_{T+1}$ is the solution of (8) *with $h = h_T$ and $\pi = \hat{\rho}_T$.*

The proof is postponed to Section 3.

Remark 1. *This lemma is useful to prove the PAC-Bayesian inequality stated in Theorem 2.2 for Algorithm 2. It generalizes the convex duality formula stated originally in [17] which corresponds in Lemma 2.1 to the particular case $\mathcal{B} = \mathcal{B}_s$ for $s(x) = x \log x$. In this case, we deal with the Gibbs measure $\hat{\rho}_{h,\pi}(g) = \exp\{\log \pi(g) + 1 - h(g) + c_{h,\pi} - 1\} = \frac{1}{Z} \exp\{-h(g)\}\pi(g)$.*

Remark 2. *The last statement* (9) *is useful in the proof of Theorem 2.2 to extend the cancellation argument originally stated in [13] in the classical Kullback-Leibler case.*

Remark 3. *Lemma 2.1 is restricted to a finite set of weak learners $\mathcal{G}$ for simplicity. Recent extensions of the KKT conditions to the infinite dimensional case (see [18]) can be used in order to consider uncountable set $\mathcal{G}$ but it is out of the scope of the present paper.*

Previous Lemma is useful to control the expected regret of Algorithm 2 as follows.

Theorem 2.2. *Let $\lambda > 0$ and (Π, δ_λ) defined below such that $\mathbb{A}(\Pi, \delta_\lambda)$ holds. Consider a sequence of distribution $(\tilde{\rho}_t)_{t=1}^T$ based on Algorithm 2, where $\mathcal{B} \in \{\mathcal{B}_\Phi, \mathcal{B}_s\}$ and $(h_t)_{t=1}^T$ satisfy assumption of Lemma 2.1. Then for any deterministic sequence $\{z_t, t = 1, \dots T\}$, for any prior $\pi \in \mathcal{P}(\mathcal{G})$, we have:*

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \tilde{\rho}_t} \ell(\hat{g}_t, z_t) \leq \min_{\rho \in \mathcal{P}(\mathcal{G})} \left\{ \mathbb{E}_{g \sim \rho} \sum_{t=1}^{T} \bar{\ell}(g, z_t) + \frac{\mathcal{B}(\rho, \pi)}{\lambda} \right\},$$

where $\lambda \bar{\ell}(g, z) := \lambda \ell(g, z) + \mathbb{E}_{(\hat{g}_1, \dots, \hat{g}_t)} \delta_\lambda(z_t, g, \hat{g}_t)$.

Sketch of proof. Since $\mathbb{A}(\Pi, \delta_\lambda)$ holds, we have for any $\pi \in \mathcal{P}(\mathcal{G})$, for any $z \in \mathcal{Z}$:

$$\mathbb{E}_{g' \sim \Pi(\pi)} \ell(g', z) \leq \frac{1}{\lambda} \mathbb{E}_{g' \sim \Pi(\pi)} \min_{\rho \in \mathcal{P}(\mathcal{G})} \left\{ \mathbb{E}_{g \sim \rho} \lambda \left(\ell(g, z) + \delta_\lambda(z, g, g') \right) + \mathcal{B}(\rho, \pi) \right\}. \tag{10}$$

We first apply (10) for $t = 1, \dots, T$ $z = z_t$, $\pi = \hat{\rho}_t := \hat{\rho}_{h_{t-1}, \hat{\rho}_{t-1}}$ in minimization (8) for $h_t(g) = \lambda(\ell(z_t, g)) + \delta_\lambda(z_t, g, \hat{g}_t)$, with $h_0 \equiv 0$ and $\hat{\rho}_0$ correspond to the prior π. Then, summing accross iterations yields:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{(\hat{g}_1, \dots, \hat{g}_t)} \ell(z_t, g') &\leq \frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{g' \sim \tilde{\rho}_t} \left\{ \mathbb{E}_{g \sim \hat{\rho}_{t+1}} h_t(g) + B_\Phi(\hat{\rho}_{t+1}, \hat{\rho}_t) \right\} \\
&= \frac{1}{\lambda} \mathbb{E}_{(\hat{g}_1, \dots, \hat{g}_T)} \sum_{t=1}^{T} \left\{ \mathbb{E}_{g \sim \hat{\rho}_{t+1}} h_t(g) + \Phi(\hat{\rho}_{t+1}) - \Phi(\hat{\rho}_t) - \nabla \Phi(\hat{\rho}_t)(\hat{\rho}_{t+1} - \hat{\rho}_t) \right\} \\
&= \frac{1}{\lambda} \mathbb{E}_{(\hat{g}_1, \dots, \hat{g}_T)} \sum_{t=1}^{T} \left\{ \mathbb{E}_{g \sim \hat{\rho}_{t+1}} h_t(g) + \Phi(\hat{\rho}_{T+1}) - \Phi(\pi) - \sum_{t=1}^{T} \nabla \Phi(\hat{\rho}_t)(\hat{\rho}_{t+1} - \hat{\rho}_t) \right\}.
\end{aligned}$$

Moreover, notice that by Lemma 3.1 (see Section 3), we have under the assumptions of Lemma 2.1:

$$\sum_{t=1}^{T} \nabla \Phi(\hat{\rho}_t)(\hat{\rho}_{t+1} - \hat{\rho}_t) = \nabla \Phi(\pi)(\hat{\rho}_{T+1} - \pi) - \mathbb{E}_{g \sim \hat{\rho}_{T+1}} \sum_{t=1}^{T-1} h_t(g) + \sum_{t=1}^{T-1} \mathbb{E}_{g \sim \hat{\rho}_{t+1}} h_t(g).$$

Then, gathering with the previous computations, we arrive at:

$$\sum_{t=1}^{T} \mathbb{E}_{g' \sim \tilde{\rho}_t} \ell(z_t, g') \leq \frac{1}{\lambda} \mathbb{E}_{(\hat{g}_1, \dots, \hat{g}_T)} \mathbb{E}_{g \sim \hat{\rho}_{T+1}} \sum_{t=1}^{T} h_t(g) + \Phi(\hat{\rho}_{T+1}) - \Phi(\pi) - \nabla \Phi(\pi)(\hat{\rho}_{T+1} - \pi)$$

$$= \frac{1}{\lambda}\mathbb{E}_{(\hat{g}_1,\ldots,\hat{g}_T)}\mathbb{E}_{g\sim\hat{\rho}_{T+1}}\sum_{t=1}^{T} h_t(g) + B_\Phi(\hat{\rho}_{T+1},\pi)$$

$$= \frac{1}{\lambda}\mathbb{E}_{(\hat{g}_1,\ldots,\hat{g}_T)}\min_{\rho\in\mathcal{P}(\mathcal{G})}\left\{\mathbb{E}_{g\sim\rho}\sum_{t=1}^{T} h_t(g) + B_\Phi(\rho,\pi)\right\}$$

where we use the last statement of Lemma 2.1 for the last equality.

Remark 4. *Theorem 2.2 controls the expected cumulative loss of Algorithm 2 in the online learning scenario presented in Section 1. Similar results can be stated in the i.i.d. case to get a control of the expected risk of a slightly modified version of Algorithm 2 by using a mirror averaging (see [6] or [7]).*

Remark 5. *Theorem 2.2 generalizes the standard PAC-Bayesian bounds with Kullback-Leibler divergences to Bregman divergences. Indeed, in Theorem 2.2, if Φ involves a Kullback-Leibler divergence, we get a Gibbs measure in Algorithm 2 and find the existing bounds stated in [7].*

Remark 6. *The introduction of alternatives to the Kullback-Leibler divergence in PAC-Bayesian theory has been studied in the literature. [19] studies Φ-divergence in a probabilistic context, whereas [20] uses Renyi divergence. Moreover [21] propose new posteriors in a Bayesians setting for variational inference for maximum likelihood estimators. Recently [22] also states regret bounds for unbounded losses thanks to the introduction of a Φ-divergence.*

Remark 7. *Assumption $\mathbb{A}(\Pi,\delta_\lambda)$ is necessary to conduct the proof. It can be traced back to [14]. It is well-known that this assumption is satisfied for any loss function in the Kullback-Leibler case for a particular quadratic function δ_λ (see [7]). In the sequel, we prove this assumption in the Optimal Transport case to lead to a new kind of procedure where exponential averages are replaced by optimal transport optimization.*

Remark 8. *Theorem 2.2 is the main ingredient to derive the same kind of result for the optimal transport divergence, as well as regret bound in Section 2.2.*

Algorithm 2 Bregman Algorithm

---

init. $\lambda > 0$, $t = 1$, $\pi \in \mathcal{P}(\mathcal{G})$ a prior distribution. $\mathcal{B}_\Phi$ a Bregman divergence. Suppose $\mathbb{A}(\Pi,\delta_\lambda)$ holds.
Let $\tilde{\rho}_1 = \Pi \circ \pi$.
repeat Predict z_t with $\hat{g}_t \sim \tilde{\rho}_t$. Observe z_t and compute

$$h_t(g) = \ell(g, z_t) + \delta_\lambda(z_t, g, \hat{g}_t) \quad \text{for all } g \in \mathcal{G}.$$

return $\tilde{\rho}_{t+1} = \Pi \circ \hat{\rho}_{t+1}$ where

$$\hat{\rho}_{t+1}(g) = \max\left(0, (\nabla_g\Phi)^{-1}\left(\nabla_g\Phi(\hat{\rho}_t(g)) - h_t(g) + c_{h_t,\hat{\rho}_t}\right)\right) \text{ for all } g \in \mathcal{G}.$$

$t = t + 1$
until $t = T + 1$

2.2. Optimal Transport

Given two probability measure $\rho, \pi \in \mathcal{P}(\mathcal{G})$, the Kantorovitch formulation of optimal transport between ρ and π is given by:

$$\mathcal{W}_C(\rho, \pi) := \min_{\Lambda \in \Delta(\rho,\pi)} \int_{\mathcal{G}\times\mathcal{G}} C(g, g')d\Lambda(g, g'), \tag{11}$$

where $\Delta(\rho, \pi) = \{\Lambda \in \mathcal{P}(\mathcal{G} \times \mathcal{G}) : \Lambda_1 = \rho \text{ and } \Lambda_2 = \pi\}$ and Λ_1 (resp. Λ_2) stands for the first (resp. second) marginal of Λ. An optimizer of (11) is called a transportation plan and quantifies how mass is moved from π to ρ whereas the cost function $C(\cdot, \cdot) : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ measure the cost of mooving a unit mass from g' to g. For an mathematical introduction of optimal transport, we refer to the monograph [23].

Optimal transport has recently inspired the machine learning community to get a variety of divergences between probability distributions (see the monograph of [24]). It rises to several applications where comparisons of complex and high dimensional objects are needed : time series (see for instance [25, 26]), images (see [27, 28]) or neuro-images ([29, 30]). Moreover in these machine learning applications, some form of regularization has been proposed to avoid the curse of dimensionality as well as to build scalable versions of the original optimization problem (11) (see [31]). For that purpose, entropy regularization makes the optimal transport problem more tractable and eliminates a number of analytic and computational difficulties. That is, we can consider a perturbation of the minimization problem (11) given by

$$\mathcal{W}_\alpha(\rho, \pi) := \min_{\Lambda \in \Delta(\rho,\pi)} \left\{ \int_{\mathcal{G}\times\mathcal{G}} C(g, g')d\Lambda(g, g') + \alpha(H(\rho) + H(\pi) - H(\Lambda)) \right\}, \tag{12}$$

for some $\alpha > 0$, where H is the Shannon entropy. Note that the quantity $H(\rho) + H(\pi) - H(\Lambda) \geq 0$. Moreover, since $\mathcal{W}_\alpha(\cdot, \pi)$ is strictly convex for any distribution $\pi \in \mathcal{P}(\mathcal{G})$, we use divergence (12) in the rest of the paper.

Lemma 2.3. *Assume $\mathcal{G}$ is finite. Let $\pi \in \mathcal{P}(\mathcal{G})$, $\alpha > 0$ and $C : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ a cost function and consider the minimization problem:*

$$\min_{\rho\in\mathcal{P}} \left\{ \mathbb{E}_{g\sim\rho}[h(g)] + \mathcal{W}_\alpha(\rho, \pi) \right\}. \tag{13}$$

Then we have:

$$\min_{\rho\in\mathcal{P}} \left\{ \mathbb{E}_{g\sim\rho}[h(g)] + \mathcal{W}_\alpha(\rho, \pi) \right\} = \alpha\mathcal{H}(\pi) - \langle \pi, v \rangle,$$

where $v : \mathcal{G} \to \mathbb{R}_+$ is the KKT multiplier and satisfies:

$$v(g') = -\alpha \log \sum_{g\in\mathcal{G}} M_\alpha(g, g') \exp\left(\frac{h(g)}{\alpha}\right), \forall g' \in \mathcal{G},$$

for M_α the inverse of the kernel matrix $K_\alpha = \left(\exp\left(-\frac{C(g,g')}{\alpha}\right)\right)_{g,g'\in\mathcal{G}}$.

Then, for any $\lambda > 0$, $\mathbb{A}(\Pi, \delta_\lambda)$ holds for:

$$\begin{cases} \delta_\lambda(z, g, g') = \frac{\lambda^2}{2\alpha}(\ell(z, g) - \ell(z, g'))^2 + \alpha \log A(\pi) \\ \Pi(\pi)(g) = A(\pi)\mathbb{E}_{g'\sim\pi} \exp\left(-\frac{C(g,g')}{\alpha}\right), \forall g \in \mathcal{G}, \end{cases} \tag{14}$$

where $A(\pi)$ is the normalizing constant.

Remark 9. *The first statement is useful to prove* (14)*. It uses the KKT conditions involved in the minimization problem* (13)*. As before, the restriction to a finite set $\mathcal{G}$ only appears since we use the classical finite dimensional Lagrange theory. Extensions to infinite and uncountable set of candidates $\mathcal{G}$ is possible using recent extensions of the KKT conditions to the infinite dimensional case (see [18]).*

Remark 10. *The second statement ensures the existence of a functional $\delta_\lambda(z,\cdot,\cdot)$ and an operator $\Pi(\cdot)$ such that assumption $\mathbb{A}(\Pi,\delta_\lambda)$ holds. As far as we know, this is the first time this assumption is used with non-trivial transformation $\Pi(\cdot)$ and Optimal Transport divergences instead of classical Kullback-Leibler divergence and duality. Operator Π depends on the cost function $C(\cdot,\cdot)$ chosen in the optimal transport divergence and acts as a regularizer based on this cost.*

This lemma is useful to prove the PAC-Bayesian bound for Algorithm 3.

Theorem 2.4. *Assume $\mathcal{G}$ is finite. Let $\lambda > 0$. Consider a sequence of distribution $(\tilde{\rho}_t)_{t=1}^T$ based on Algorithm 3. Then for any deterministic sequence $\{z_1,\ldots,z_T\}$, for any prior $\pi \in \mathcal{P}(\mathcal{G})$, we have:*

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t\sim\tilde{\rho}_t} \ell(\hat{g}_t, z_t) \leq \min_{\rho\in\mathcal{P}(\mathcal{G})} \left\{ \mathbb{E}_{g\sim\rho} \sum_{t=1}^{T} \bar{\ell}(g, z_t) + \frac{\mathcal{W}_\alpha(\rho,\pi)}{\lambda} \right\} + \Delta_{T,\lambda}\left(\mathcal{B}_{\Phi_\alpha}, \mathcal{W}_\alpha\right),$$

where $\bar{\ell}(g,z) := \ell(g,z) + \frac{\lambda}{2\alpha}\mathbb{E}_{(\hat{g}_1,\ldots,\hat{g}_t)}\left(\ell(g,z) - \ell(\hat{g}_t,z)\right)^2 + \alpha\log A(\pi)$ and the extra-term $\Delta_T\left(\mathcal{B}_{\Phi_\alpha},\mathcal{W}_\alpha\right)$ is defined in Section 3.

Remark 11. *The proof uses Theorem 2.2 with a particular function $\Phi_\alpha : \rho \mapsto \mathcal{W}_\alpha(\rho,\nu)$ for some $\nu \in \mathcal{P}(\mathcal{G})$ and allows us to extend the previous PAC-Bayesian bound to entropy regularized optimal transportation (see Section 3 for a detailled proof).*

Remark 12. *Theorem 2.4 allows us to derive regret bounds for Algorithm 3. It gives more flexibility into the regularization thanks to the introduction of a particular cost function C in* (12)*.*

Algorithm 3 Optimal transport PAC-Bayesian Algorithm

init. $\lambda, \alpha > 0$, $t = 1$, $\pi \in \mathcal{P}(\mathcal{G})$ a prior distribution. C a cost function in $\mathcal{W}_\alpha$.
Let $\tilde{\rho}_1 = \Pi \circ \pi$.
repeat Observe z_t and predict $\hat{g}_t \sim \tilde{\rho}_t$. Compute:

$$h_t(g) = \ell(g, z_t) + \frac{\lambda}{2\alpha}(\ell(g,z_t) - \ell(\hat{g}_t, z_t))^2 + \alpha\log A(\hat{\rho}_t), \quad \text{for all } g \in \mathcal{G},$$

where $A(\hat{\rho}_t)$ is defined in Lemma 2.3.
return $\tilde{\rho}_{t+1} = \Pi \circ \hat{\rho}_{t+1}$ where Π is defined in Lemma 2.3 and $\hat{\rho}_{t+1}$ is the solution of the minimization problem in Lemma 2.3 with $h = h_t$.
$t = t + 1$
until $t = T$

2.3. Generic regret bounds

We are now on time to apply Theorem 2.4 in the following toy generic example. Let us consider a finite set of learning machines $\mathcal{G} = \{g_1,\ldots,,g_p\}$, where $p \geq 1$ is the number of learners. Suppose we have at

hand a prior knowledge on these learning machines, in terms of generalization power, as well as another generic criterion to minimize (such as for instance energy consumption or carbon footprint, see [12] and the references therein). We denote by $\{(\mathrm{Err}_1, \mathrm{Crit}_1), \ldots, (\mathrm{Err}_p, \mathrm{Crit}_p)\}$ these sequences of criteria, where g_i has generalization error estimate Err_i and score Crit_i for the generic criterion. We consider in the sequel two different scenarii corresponding to Corollary 2.5 and Corollary 2.6. In the first scenario, we suppose the simplest multiple-criteria decision-making problem, where for any $\eta > 0$, there exists an unique optimal decision $g_\eta^\star = g_{i_\eta^\star} \in \mathcal{G}$ such that:

$$i_\eta^\star = \arg\min_{g \in \mathcal{G}} \left(\mathrm{Err}_i + \eta \mathrm{Crit}_i\right),$$

where $\eta > 0$ governs the trade-off between both criteria. Large value of $\eta > 0$ corresponds for instance to a small energy budget whereas $\eta = 0$ corresponds to a classical machine learning problem. With this in mind, consider a new task based on a deterministic sample $\{z_1, \ldots, z_T\}$ and a loss function $\ell(g, z)$. We want to draw a distribution, or mixture, on the finite set $\mathcal{G}$ able to trades off dynamically the loss over the new set of observations, and the generic criterion. For that purpose, we start from $g_\eta^\star$, the best compromise at hand and adapts sequentially the mixture thanks on Algorithm 3 as follows. At each round $t \geq 1$, we choose a new distribution $\tilde{\rho}_{t+1} \in \mathcal{P}(\mathcal{G})$ that minimizes the expected loss on new observation z_t and the transport from $\tilde{\rho}_t$, where the optimal transport is based on a cost function depending on the sequence Crit, such as for instance $C(g_i, g_j) := \mathrm{Crit}_{\mathrm{i}} - \mathrm{Crit}_j$. We hence have the following result based on a direct application of Theorem 2.4:

Corollary 2.5. *Let $\pi = \delta_{g_\eta^\star}$ the Dirac measure on $g_\eta^\star$ and consider Algorithm 3 with $C(g_i, g_j) := C(\mathrm{Crit}_{\mathrm{i}}, \mathrm{Crit}_j)$ for any $i, j = 1, \ldots p$ in the optimal transport divergence* (11)*. Then we have, for any deterministic sequence $\{z_1, \ldots, z_T\}$:*

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \tilde{\rho}_t} \ell(\hat{g}_t, z_t) \leq \min_{g \in \mathcal{G}} \left\{ \sum_{t=1}^{T} \bar{\ell}(g, z_t) + \frac{C(g, g_{i^\star})}{\lambda} \right\} + \Delta_T,$$

where $\bar{\ell}$ and $\Delta_T > 0$ are defined in Theorem 2.4 and $\lambda > 0$.

The proof is straightforward using Theorem 2.4 with Dirac prior $\delta_{g_\eta^\star}$.

However, in practice, the existence and uniqueness of $g_\eta^\star$ is rarely satisfied due to the unstability of the problem. Indeed, the generalization power, as well as the generic criterion, could be considered as stochastic values depending on unknown parameters such as the variability of the problem, the considered hardware, or some other practices. As a result, it could be more realistic to consider a prior π_b based on a weighted average of candidate learners $\{g_1, \ldots, g_p\}$. Moreover, if we have at hand a budget $b > 0$, we can choose π_b as a solution of the following optimization:

$$\begin{cases} \min_\rho \mathbb{E}_{g \sim \rho} \widehat{\mathrm{Err}}(g) \\ \quad \text{s.t. } \mathbb{E}_{g \sim \rho} \widehat{\mathrm{Crit}}(g) \leq b, \end{cases}$$

where $\widehat{\mathrm{Err}}(g)$ and $\widehat{\mathrm{Crit}}(g)$ are respectively estimates of the true but untractable generalization power and, for instance, electric consumption of learner $g \in \mathcal{G}$. In this case, strarting from π_b for some $b > 0$, we have the following result:

Corollary 2.6. *Let $\pi = \pi_b$ for some $b > 0$ in Algorithm 3 and $C(g_i, g_j) := C(\mathrm{Crit}_i, \mathrm{Crit}_j)$ for any $i, j = 1, \dots p$ in the optimal transport divergence* (11)*. Then we have for the sequence of distribution $\{\tilde{\rho}_t, t = 1 \dots, T\}$ based on Algorithm 3:*

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \tilde{\rho}_t} \ell(\hat{g}_t, z_t) \leq \inf_{i=1,\dots,p} \left\{ \sum_{t=1}^{T} \bar{\ell}(g, z_t) + \frac{\mathbb{E}_{g' \sim \pi_b} C(g_i, g')}{\lambda} \right\} + \Delta_T,$$

where $\bar{\ell}$ and $\Delta_T > 0$ are defined in Theorem 2.4.

The proof is also straightforward using Theorem 2.4 with prior π_b.

3. Proofs

3.1. Proof of Lemma 2.1

We start with the proof of Lemma 2.1. Since $\mathcal{G}$ is finite, the solution can be described directly by the Karush-Kuhn-Tucker (KKT) theory (see [32]). We first formally write the optimization problem as:

$$\begin{cases} \min_\rho \left\{ F_{h,\pi}(\rho) := \mathbb{E}_{g \sim \rho}[h(g)] + B_\Phi(\rho, \pi) \right\} \\ \quad \text{s.t. } \rho(g) \geq 0, \forall g \in \mathcal{G} \\ \qquad \sum_{g \in \mathcal{G}} \rho(g) - 1 = 0 \end{cases}$$

For the first statement, since Φ and h are convex and continuously differentiable, $F_{h,\pi}(\cdot)$ is convex and continuously differentiable and by the KKT conditions involved in the strong duality theorem, we first have for $\hat{\rho} := \hat{\rho}_{h,\pi}$ minimizer of $F_{h,\pi}$:

$$h(g) + \nabla_g \Phi(\hat{\rho}(g)) - \nabla_g \Phi(\pi(g)) = \mu(g) - \lambda, \forall g \in \mathcal{G},$$

where $\mu : \mathcal{G} \to \mathbb{R}_+$ and $\lambda \geq 0$ are KKT multipliers.

Moreover, the complementary slackness condition (see [32], Theorem 5.21) implies that if $\hat{\rho}(g) \neq 0$, then $\mu(g) = 0$. That is, if $\hat{\rho}(g) \neq 0$, then

$$\nabla_g \Phi(\hat{\rho}(g)) = \nabla_g \Phi(\pi(g)) - h(g) - \lambda \tag{15}$$

while if $\hat{\rho}(g) = 0$, then

$$\nabla_g \Phi(0) = \nabla_g \Phi(\hat{\rho}(g)) = \nabla_g \Phi(\pi(g)) - h(g) - \lambda + \mu(g) \geq \nabla_g \Phi(\pi(g)) - h(g) - \lambda.$$

Combining the above conditions, there is some constant $c_{h,\pi}$ such that for every $g \in \mathcal{G}$:

$$\nabla_g \Phi(\hat{\rho}(g)) = \max\{a_g, \nabla_g \Phi(\pi(g)) - h(g) + c_{h,\pi}\},$$

where $a_g = \inf_{r \in (0,1)} \nabla_g \Phi(r)$. In particular, for every $g \in \mathcal{G}$,

$$\hat{\rho}(g) = \max\left\{0, (\nabla_g \Phi)^{-1} \left(\nabla_g \Phi(\pi(g)) - h(g) + c_{h,\pi}\right)\right\}$$

where $c_{h,\pi}$ satisfies the following constraint:

$$\sum_{g\in\mathcal{G}} \hat{\rho}(g) = \sum_{g\in\mathcal{G}} \max\{0, \min\{1, (\nabla_g \Phi)^{-1}\left(\nabla_g \Phi(\pi(g)) - h(g) + c_{h,\pi}\right)\}\} = 1.$$

Moreover, by strict convexity of Φ, it is straightforward to see that $c_{h,\pi}$ is unique.

For the second statement, we suppose the existence of a minimizer. Then, by KKT conditions, (15) holds. Moreover, the uniqueness of the minimizer follows for the dual convexity formula. Indeed, consider another distribution ρ' and define the family of convex combinations $\rho_\epsilon = (1-\epsilon)\hat{\rho} + \epsilon\rho'$ for $\epsilon \in [0,1]$.

Then, consider the first variation of $\rho \mapsto B_\Phi(\rho, \pi)$:

$$\begin{aligned}
\frac{d}{d\epsilon}\left(\mathbb{E}_{\rho_\epsilon} h + B_\Phi(\rho_\epsilon, \pi)\right)\Big|_{\epsilon=0} &= \frac{d}{d\epsilon}\Phi(\rho_\epsilon) - \frac{d}{d\epsilon}(\nabla\Phi(\pi) - h)(\rho_\epsilon - \pi)\Big|_{\epsilon=0} \\
&= \nabla\Phi(\rho_\epsilon)(\rho' - \hat{\rho}) - (\nabla\Phi(\pi) - h)(\rho' - \hat{\rho})\big|_{\epsilon=0} \\
&= \nabla\Phi(\hat{\rho})(\rho' - \hat{\rho}) - (\nabla\Phi(\pi) - h)(\rho' - \hat{\rho}) \\
&= (\nabla\Phi(\pi) - h + c\mathbb{1})(\rho' - \hat{\rho}) - (\nabla\Phi(\pi) - h)(\rho' - \hat{\rho}) \\
&= 0.
\end{aligned}$$

Then, consider the second variation

$$\begin{aligned}
\frac{d^2}{d\epsilon^2}\left(\mathbb{E}_{\rho_\epsilon} h + B_\Phi(\rho_\epsilon, \pi)\right)\Big|_{\epsilon=0} &= \nabla^2\Phi(\rho_\epsilon)(\rho' - \hat{\rho})^2\Big|_{\epsilon=0} \\
&= \nabla^2\Phi(\hat{\rho})(\rho' - \hat{\rho})^2 \\
&> 0.
\end{aligned}$$

This shows that $\hat{\rho}$ is indeed a local minimizer. Moreover, the global convexity of the Bregman divergence in the first argument yields to the uniqueness of the global minimizer.

For the third statement, we provide a pointwise Bregman divergence as in (7), where the measure ν is assumed to be finite. Moreover, suppose that $\lim_{x\to 0} s'(x) \to -\infty$.

As noted in the above remark, this is a special case of the general Bregman divergence. Naturally, $B_s = B_\Phi$, where $\Phi(p) = \int s(p(x))\nu(dx)$ and $\nabla\Phi(p)(q) = \int s'(p(x))q(x)\,\nu(dx)$. Recall, as in the discrete case, that under the assumptions, the monotonicity of s' implies that it is invertible on $\mathbb{R}$. It follows that the inverse problem in (15) has solution $\hat{\rho}(g) = (s')^{-1}(s'(\pi(g)) - h(g) + c)$. Indeed, the fact that there is a unique c such that $\hat{\rho}$ is a distribution follows from the strict convexity of s and

$$\frac{d}{dc}\int (s')^{-1}(s'(\pi(g)) - h(g) + c)\,\nu(dg) = \int \frac{1}{s''(\hat{\rho}(g))}\,\nu(dg) > 0.$$

Then, we use the intermediate value theorem with

$$\int (s')^{-1}(s'(\pi(g)) - h(g) + M)\,\nu(dg) \geq \int (s')^{-1}(s'(\pi(g))\,\nu(dg) = 1$$

and

$$\int (s')^{-1}(s'(\pi(g)) - h(g) + m)\,\nu(dg) \leq \int (s')^{-1}(s'(\pi(g))\,\nu(dg) = 1,$$

where $M = \sup_g h(g)$ and $m = \inf_g h(g)$.

For the last statement, considering for concision a pointwise Bregman divergence as above, we have coarselly:

$$\begin{aligned}\hat{\rho}_{T+1}(g) &:= (s')^{-1}(s(\hat{\rho}_T(g)) - h_T(g) + c_T)\\ &= (s')^{-1}(s(\hat{\rho}_{T-1}(g)) - h_{T-1}(g) + c_{T-1} - h_T(g) + c_T)\\ &\vdots\\ &= (s')^{-1}(s(\hat{\rho}_1(g)) - \sum_{t=1}^{T} h_t(g) + c).\end{aligned}$$

3.2. Proof of Theorem 2.2

The proof of Lemma 2.2 is based on the following lemma.

Lemma 3.1. *Let $(\hat{\rho}_t)_{t=1}^{T+1}$ the sequence of distribution defined in the proof of Theorem 2.2. Then under the assumptions of Lemma 2.1:*

$$\sum_{t=1}^{T} \nabla\Phi(\hat{\rho}_t)(\hat{\rho}_{t+1} - \hat{\rho}_t) = \nabla\Phi(\pi)(\hat{\rho}_{T+1} - \pi) - \mathbb{E}_{g\sim\hat{\rho}_{T+1}} \sum_{t=1}^{T-1} h_t(g) + \sum_{t=1}^{T-1} \mathbb{E}_{g\sim\hat{\rho}_{t+1}} h_t(g).$$

Proof. From Lemma 2.1, since $\mathcal{G}$ is finite and using KKT conditions, (15) holds. Then by definition of the sequence $(\hat{\rho}_t)_{t=1}^{T+1}$, we have for any $t = 1, \ldots, T$, and any $\rho, \rho' \in \mathcal{P}(\mathcal{G})$:

$$\begin{aligned}\nabla\Phi(\hat{\rho}_t)(\rho - \rho') &= \left(\nabla\Phi(\hat{\rho}_{t-1}) - h_{t-1} + c_{h_t,\hat{\rho}_{t-1}}\right)(\rho - \rho')\\ &= \nabla\Phi(\hat{\rho}_{t-1})(\rho - \rho') - \sum_{g\in\mathcal{G}} h_{t-1}(g)\left(\rho(g) - \rho'(g)\right),\end{aligned}$$

where $\hat{\rho}_0 = \hat{\rho}_1$. Then, applying extensively this equality and telescoping terms, we hence have:

$$\begin{aligned}&\sum_{t=1}^{T} \nabla\Phi(\hat{\rho}_t)(\hat{\rho}_{t+1} - \hat{\rho}_t)\\ &= \left\{\sum_{t=1}^{T-1} \nabla\Phi(\hat{\rho}_t)(\hat{\rho}_{t+1} - \hat{\rho}_t)\right\} + \nabla\Phi(\hat{\rho}_{T-1})(\hat{\rho}_{T+1} - \hat{\rho}_T) - \sum_{g\in\mathcal{G}} h_{T-1}(g)\left(\hat{\rho}_{T+1}(g) - \hat{\rho}_T(g)\right)\\ &= \left\{\sum_{t=1}^{T-2} \nabla\Phi(\hat{\rho}_t)(\hat{\rho}_{t+1} - \hat{\rho}_t)\right\} + \nabla\Phi(\hat{\rho}_{T-1})(\hat{\rho}_{T+1} - \hat{\rho}_{T-1}) - \sum_{g\in\mathcal{G}} h_{T-1}(g)\left(\hat{\rho}_{T+1}(g) - \hat{\rho}_T(g)\right)\\ &= \left\{\sum_{t=1}^{T-2} \nabla\Phi(\hat{\rho}_t)(\hat{\rho}_{t+1} - \hat{\rho}_t)\right\} + \nabla\Phi(\hat{\rho}_{T-2})(\hat{\rho}_{T+1} - \hat{\rho}_{T-1}) - \sum_{t=T-2}^{T-1}\sum_{g\in\mathcal{G}} h_t(g)\left(\hat{\rho}_{T+1}(g) - \hat{\rho}_{t+1}(g)\right)\end{aligned}$$

$$\vdots$$

$$= \nabla\Phi(\pi)(\hat{\rho}_{T+1} - \pi) - \sum_{t=1}^{T-1} \sum_{g \in \mathcal{G}} h_t(g) \left(\hat{\rho}_{T+1}(g) - \hat{\rho}_{t+1}(g)\right)$$

$$= \nabla\Phi(\pi)(\hat{\rho}_{T+1} - \pi) - \mathbb{E}_{g \sim \hat{\rho}_{T+1}} \sum_{t=1}^{T-1} h_t(g) + \sum_{t=1}^{T-1} \mathbb{E}_{g \sim \hat{\rho}_{t+1}} h_t(g).$$

3.3. Proof of Lemma 2.3

We write the minimization problem of Lemma 2.3 with the embedded optimal transport problem to consider a single minimization problem over $\Lambda \in \Delta(\pi)$, the space of couplings whose right marginal is π.

$$\begin{cases} \min \sum_{g,g' \in \mathcal{G}} \Lambda(g,g')(C(g,g') + h(g)) + \alpha \left(H(\pi) + \sum_{g,g' \in \mathcal{G}} \Lambda(g,g')(\log \Lambda(g,g') - \log \sum_{g'' \in \mathcal{G}} \Lambda(g,g'')) \right) \\ \quad \text{s.t.} \sum_{g \in \mathcal{G}} \Lambda(g,g') = \pi(g') \quad \forall g' \in \mathcal{G} \\ \qquad \Lambda(g,g') \geq 0, \quad \forall g, g' \in \mathcal{G}. \end{cases} \tag{16}$$

Then, by the Karush-Kuhn-Tucker (KKT) conditions, there exist $u \in \mathbb{R}_+^{\mathcal{G} \times \mathcal{G}}$, $v \in \mathbb{R}_+^{\mathcal{G}}$ such that for every $g, g' \in \mathcal{G}$,

$$C(g,g') + h(g) - u(g,g') + v(g') + \alpha \left(\log \Lambda(g,g') - \log \sum_{\bar{g} \in \mathcal{G}} \Lambda(g,\bar{g}) \right) = 0. \tag{17}$$

Moreover, $\Lambda(g,g') \cdot u(g,g') = 0$ for all $g, g' \in \mathcal{G}$, so either $u(g,g')$ or $\Lambda(g,g')$ is zero. Thus, when multiplying the constraint by $\Lambda(g,g')$, we can assume in general that $u(g,g') = 0$. Multiplying by $\Lambda(g,g')$ and summing in both g and g' yields:

$$\begin{aligned} & F(h,\pi) - \alpha H(\pi) \\ &= \sum_{g,g'} \Lambda(g,g') \left(C(g,g') + h(g)\right) + \alpha \left(\sum_{g,g'} \Lambda(g,g') \log \Lambda(g,g') - \sum_{g} \sum_{g'} \Lambda(g,g') \log \sum_{g'} \Lambda(g,g') \right) \\ &= - \sum_{g,g' \in \mathcal{G}} \Lambda(g,g') v(g') \\ &= -\langle \pi, v \rangle. \end{aligned}$$

Thus, the optimal value satisfies

$$F(h,\pi) = \alpha H(\pi) - \langle \pi, v \rangle.$$

We use the other constraints to solve for v explicitly. Denote $\rho(g) = \sum_{\bar{g} \in \mathcal{G}} \Lambda(g,\bar{g})$. Exponentiating the KKT conditions yields:

$$\Lambda(g,g') = \rho(g)\exp\left(-\frac{C(g,g')+h(g)-u(g,g')+v(g')}{\alpha}\right).$$

Summing in g yields:

$$\pi(g') = \sum_{g\in\mathcal{G}} \rho(g)\exp\left(-\frac{C(g,g')+h(g)-u(g,g')+v(g')}{\alpha}\right). \tag{18}$$

On the other hand, we see that $\Lambda(g,g')=0$ if and only if $\rho(g)=0$. Consequently, the equation holds in general with $u(g,g')=0$. For $\rho(g)\neq 0$, summing in g' yields

$$1 = \sum_{g'\in\mathcal{G}} \exp\left(-\frac{C(g,g')+h(g)+v(g')}{\alpha}\right).$$

Let $\hat{v}_{\pm\alpha}(g') = e^{\pm v(g')/\alpha}$ and $\hat{h}_{\pm\alpha}(g) = e^{\pm h(g)/\alpha}$. Let $\hat{C}_{-\alpha}$ be the matrix with entries $\exp(-C(g,g')/\alpha)$ and under the assumption that it is invertible, let B_α be its inverse. Then, we rewrite the equation above in matrix form

$$\hat{h}_{+\alpha} = \hat{C}_{-\alpha}\hat{v}_{-\alpha} \quad \Longrightarrow \quad \hat{v}_{-\alpha} = B_\alpha \hat{h}_{+\alpha}.$$

That is,

$$v(g') = -\alpha\log\left(\sum_{g\in\mathcal{G}} B_\alpha(g,g')e^{h(g)/\alpha}\right).$$

Finally, to prove the second statement, let $\lambda>0$. Let $\hat{\rho}$ be the minimizer of (13) and let $\Pi:\mathcal{P}(\mathcal{G})\to\mathcal{P}(\mathcal{G})$ such that for any $g\in\mathcal{G}$:

$$\Pi(\pi)(g) := A(\pi)^{-1}\mathbb{E}_{g'\sim\pi}e^{-C(g,g')/\alpha},$$

where $A(\pi)$ is the normalizing constant defined as:

$$A(\pi) = \sum_{g\in\mathcal{G}} \mathbb{E}_{g'\sim\pi}e^{-C(g,g')/\alpha}.$$

By equation (18), we have:

$$\pi(g')\hat{v}_{+\alpha}(g') = \sum_{g\in\mathcal{G}} \hat{\rho}(g)\exp\left(-\frac{C(g,g')+h(g)}{\alpha}\right).$$

Recall that

$$\begin{aligned} F(h,\pi) &= \alpha H(\pi) - \langle \pi, v\rangle \\ &= -\alpha\left\langle \pi, \log\pi + \frac{v}{\alpha}\right\rangle \\ &= \left\langle \pi, -\alpha\log(\pi e^{v/\alpha})\right\rangle \end{aligned}$$

$$= \langle \pi, -\alpha \log(\pi \hat{v}_{+\alpha}) \rangle$$
$$= -\alpha \mathbb{E}_{g' \sim \pi} \log \left(\sum_{g \in \mathcal{G}} \hat{\rho}(g) \exp\left(-\frac{C(g,g') + h(g)}{\alpha} \right) \right)$$

Then by Jensen's inequality, to show $\mathbb{A}(\Pi, \delta_\lambda)$, it suffices to show:

$$\mathbb{E}_{g' \sim \Pi(\pi)} \mathbb{E}_{g'' \sim \pi} \sum_{g \in \mathcal{G}} \hat{\rho}(g) \exp\left(-\frac{C(g,g'') + h(z,g,g')}{\alpha} \right) \leq 1$$

Compute:

$$\begin{aligned}
&\mathbb{E}_{g' \sim \Pi(\pi)} \mathbb{E}_{g'' \sim \pi} \sum_{g \in \mathcal{G}} \hat{\rho}(g) \exp\left(-\frac{C(g,g'') + h(z,g,g')}{\alpha} \right) \\
&= \mathbb{E}_{g' \sim \Pi(\pi)} \sum_{g \in \mathcal{G}} \hat{\rho}(g) \mathbb{E}_{g'' \sim \pi} \exp\left(-\frac{C(g,g'')}{\alpha} \right) \exp\left(-\frac{h(z,g,g')}{\alpha} \right) \\
&\leq \sum_{g \in \mathcal{G}} \mathbb{E}_{g'' \sim \pi} \exp\left(-\frac{C(g,g'')}{\alpha} \right) \mathbb{E}_{g' \sim \Pi(\pi)} \exp\left(-\frac{h(z,g,g')}{\alpha} \right) \\
&= A(\pi)\, \mathbb{E}_{g,g' \sim \Pi(\pi)} \exp\left(-\frac{h(z,g,g')}{\alpha} \right) \\
&= A(\pi)\, \mathbb{E}_{g \sim \hat{\rho}, g' \sim \Pi(\pi)} \exp\left(-\frac{\lambda(\ell(g,z) - \ell(g',z)) + \delta_\lambda(z,g,g')}{\alpha} \right) \\
&= \mathbb{E}_{g,g' \sim \Pi(\pi)} \exp\left(-\frac{\lambda}{\alpha}(\ell(g,z) - \ell(g',z)) - \frac{1}{2}\left(\frac{\lambda}{\alpha}(\ell(g,z) - \ell(g',z)) \right)^2 \right) \\
&\leq 1.
\end{aligned}$$

3.4. Proof of Theorem 2.4

Let $\alpha > 0$. First note that from Lemma 2.3, we have for any $\pi \in \mathcal{P}(\mathcal{G})$, for any $z \in \mathcal{Z}$:

$$\mathbb{E}_{g' \sim \Pi(\pi)} \ell(g', z) \leq \frac{1}{\lambda} \mathbb{E}_{g' \sim \Pi(\pi)} \min_{\rho \in \mathcal{P}(\mathcal{G})} \left\{ \mathbb{E}_{g \sim \rho} \left(\lambda \ell(g,z) + \delta_\lambda(z,g,g') \right) + \mathcal{W}_\alpha(\rho, \pi) \right\}, \tag{19}$$

where Π and δ_λ are defined in Lemma 2.3. Then applying (19) for $t = 1, \ldots, T$, with $\pi = \hat{\rho}_t$ minimizer of (13) in Lemma 2.3 for $h_{t-1} = \lambda \ell(z_{t-1}, g) + \delta_\lambda(z_{t-1}, g, \hat{g}_{t-1})$ and $h_0 \equiv 0$ and summing accross iterations yields:

$$\sum_{t=1}^{T} \mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \ell(g', z_t) \leq \frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \min_{\rho \in \mathcal{P}(\mathcal{G})} \{ \mathbb{E}_{g \sim \rho} h_t(g) + \mathcal{W}_\alpha(\rho, \hat{\rho}_t) \}. \tag{20}$$

The idea of the proof is to decompose the RHS in (20) by introducing a particular Bregman divergence, and to use Theorem 2.2 as a main ingredient. For that purpose, let $\pi^\star \in \mathcal{P}(\mathcal{G})$ and consider the function $\Phi_\alpha : \rho \mapsto \mathcal{W}_\alpha(\rho, \pi^\star)$ based on the regularized optimal transport (12). First, note that by the entropy regularization, Φ_α is strictly convex and differentiable is its arguments for any $\alpha > 0$. Then we can decompose the RHS in (20) as follows:

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \ell(z_t, g') &\leq \frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \min_{\rho \in \mathcal{P}(\mathcal{G})} \{\mathbb{E}_{g \sim \rho} h_t(g) + \mathcal{W}_\alpha(\rho, \rho_t)\} \\ &= \frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \min_{\rho \in \mathcal{P}(\mathcal{G})} \{\mathbb{E}_{g \sim \rho} h_t(g) + \mathcal{B}_{\Phi_\alpha}(\rho, \rho_t) + \epsilon_t(\rho)\},\end{aligned}$$

where $\epsilon_t(\rho)$ is defined as :

$$\epsilon_t(\rho) = \mathcal{W}_\alpha(\rho, \rho_t) - \mathcal{B}_{\Phi_\alpha}(\rho, \rho_t).$$

By definition of the sequence $\{\hat{\rho}_t, t = 1, \ldots, T\}$, we can now decompose the RHS as follows:

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \ell(z_t, g') &= \frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{g' \sim \Pi(\rho_t)} \min_{\rho \in \mathcal{P}(\mathcal{G})} \{\mathbb{E}_{g \sim \rho} h_t(g) + \mathcal{B}_{\Phi_\alpha}(\rho, \hat{\rho}_t) + \epsilon_t(\rho)\} \\ &\leq \frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \min_{\rho \in \mathcal{P}(\mathcal{G})} \{\mathbb{E}_{g \sim \rho} h_t(g) + \mathcal{B}_{\Phi_\alpha}(\rho, \hat{\rho}_t)\} + \frac{1}{\lambda} \sum_{t=1}^{T} \epsilon_t(\hat{\rho}_{t+1}) \\ &:= \Sigma_T\left(\mathcal{B}_{\Phi_\alpha}\right) + \frac{1}{\lambda} \sum_{t=1}^{T} \epsilon_t(\hat{\rho}_{t+1}).\end{aligned}$$

We start with the control of $\Sigma_T\left(\mathcal{B}_{\Phi_\alpha}\right)$. To use Theorem 2.2, we introduce the sequence $\{\hat{\nu}_t,\, t = 1, \ldots T\}$ where $\hat{\nu}_{t+1}$ is the solution of the minimization (8) with couple $(h, \pi) = (h_t, \hat{\nu}_t)$ and where $\hat{\nu}_1 = \hat{\rho}_1$ corresponds to the prior π. Then we can write:

$$\begin{aligned}\Sigma_T\left(\mathcal{B}_{\Phi_\alpha}\right) &= \frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \min_{\rho \in \mathcal{P}(\mathcal{G})} \{\mathbb{E}_{g \sim \rho} h_t(g) + \mathcal{B}_{\Phi_\alpha}(\rho, \hat{\rho}_t)\} \\ &\leq \frac{1}{\lambda} \sum_{t=1}^{T} \left[\mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \mathbb{E}_{g \sim \hat{\nu}_{t+1}} h_t(g) + \mathcal{B}_{\Phi_\alpha}(\hat{\nu}_{t+1}, \hat{\nu}_t)\right] + \frac{1}{\lambda} \sum_{t=1}^{T} [\mathcal{B}_{\Phi_\alpha}(\hat{\nu}_{t+1}, \hat{\rho}_t) - \mathcal{B}_{\Phi_\alpha}(\hat{\nu}_{t+1}, \hat{\nu}_t)] \\ &:= \frac{1}{\lambda} \sum_{t=1}^{T} \left[\mathbb{E}_{g' \sim \Pi(\hat{\rho}_t)} \mathbb{E}_{g \sim \hat{\nu}_{t+1}} h_t(g) + \mathcal{B}_{\Phi_\alpha}(\hat{\nu}_{t+1}, \hat{\nu}_t)\right] + \frac{1}{\lambda} \sum_{t=1}^{T} \delta_t,\end{aligned}$$

where δ_t is defined by:

$$\delta_t = \mathcal{B}_{\Phi_\alpha}(\hat{\nu}_{t+1}, \hat{\rho}_t) - \mathcal{B}_{\Phi_\alpha}(\hat{\nu}_{t+1}, \hat{\nu}_t).$$

Then, following the same lines as in the proof of Theorem 2.2, Lemma 3.1 allows us to write:

$$\begin{aligned}\Sigma_T\left(\mathcal{B}_{\Phi_\alpha}\right) &\leq \frac{1}{\lambda}\min_{\rho\in\mathcal{P}(\mathcal{G})}\left\{\mathbb{E}_{g\sim\rho}\sum_{t=1}^{T}\mathbb{E}_{g'\sim\Pi(\hat{\rho}_t)}h_t(g)+\mathcal{B}_{\Phi_\alpha}(\rho,\pi)\right\}+\frac{1}{\lambda}\sum_{t=1}^{T}\delta_t\\ &\leq \frac{1}{\lambda}\min_{\rho\in\mathcal{P}(\mathcal{G})}\left\{\mathbb{E}_{g\sim\rho}\sum_{t=1}^{T}\bar{\ell}(z_t,g)+\mathcal{W}_\alpha(\rho,\pi)-\epsilon_1(\rho)\right\}+\frac{1}{\lambda}\sum_{t=1}^{T}\delta_t,\end{aligned}$$

where $\epsilon_1(\rho)$ is defined above. Hence we get the result by choosing $\pi^\star$ in Lemma 3.2 in order to control the residual term:

$$\begin{aligned}\Delta_T\left(\mathcal{B}_{\Phi_\alpha},\mathcal{W}_\alpha\right) &:= \frac{1}{\lambda}\sum_{t=1}^{T}\left(\epsilon_t(\hat{\rho}_{t+1})+\delta_t\right)\\ &= \frac{1}{\lambda}\sum_{t=1}^{T}\left(\mathcal{W}_\alpha(\hat{\rho}_{t+1},\hat{\rho}_t)-\mathcal{B}_{\Phi_\alpha}(\hat{\rho}_{t+1},\hat{\rho}_t)+\mathcal{B}_{\Phi_\alpha}(\hat{\nu}_{t+1},\hat{\rho}_t)-\mathcal{B}_{\Phi_\alpha}(\hat{\nu}_{t+1},\hat{\nu}_t)\right).\end{aligned}$$

Lemma 3.2. *Let $\alpha>0$ and $\pi^\star\in\mathcal{P}(\mathcal{G})$. Let $\Phi_\alpha:\rho\mapsto\mathcal{W}_\alpha(\rho,\pi^\star)$. Then we have:*

$$\mathcal{B}_{\Phi_\alpha}(\rho,\pi)=\mathcal{W}_\alpha(\rho,\pi^\star)-\mathcal{W}_\alpha(\pi,\pi^\star)+\langle v,\rho-\pi\rangle,$$

where $v:=v(\rho,\pi^\star)\in\mathbb{R}_+^{\mathcal{G}}$ is the KKT multiplier based on the optimization $\mathcal{W}_\alpha(\rho,\pi^\star)$ defined in (12)*.*

Proof First note that by the entropy regularization, Φ_α is strictly convex and differentiable is its arguments for any $\alpha>0$. Then $\mathcal{B}_{\Phi_\alpha}$ is well defined in (6). Moreover since $\mathcal{G}$ is finite, Φ_α can be described directly by the Karush-Kuhn-Tucker (KKT) multipliers. We first formally write the optimization problem as:

$$\begin{cases}\min_\Lambda \sum_{g,g'\in\mathcal{G}}\Lambda(g,g')C(g,g')+\alpha\left(H(\pi^\star)+\sum_{g,g'\in\mathcal{G}}\Lambda(g,g')\left[\log\Lambda(g,g')-\log\rho(g)\right]\right)\\ \quad\text{s.t. }\Lambda(g,g')\geq 0,\forall g,g'\in\mathcal{G}\\ \qquad\sum_{g'\in\mathcal{G}}\Lambda(g,g')-\rho(g)=0,\forall g\in\mathcal{G}\\ \qquad\sum_{g\in\mathcal{G}}\Lambda(g,g')-\pi^\star(g')=0,\forall g'\in\mathcal{G}\end{cases}\tag{21}$$

By the KKT conditions and the strong duality theorem, there exist $u\in\mathbb{R}_+^{\mathcal{G}\times\mathcal{G}}$, $v,v'\in\mathbb{R}_+^{\mathcal{G}}$ such that for every $g,g'\in\mathcal{G}$:

$$C(g,g')-u(g,g')+v(g)+v'(g')+\alpha\left(\log\Lambda^\star(g,g')-\log\rho(g)\right)=0,$$

where $\Lambda^\star$ is the solution of (21). Then, gathering with the same computations as in the proof of Lemma 2.3, we have:

$$\mathcal{W}_\alpha(\rho,\pi^\star)=\alpha H(\pi^\star)-\langle\rho,v\rangle-\langle\pi^\star,v'\rangle$$

By definition of the Bregman divergence (6) and the choice of Φ_α, we hence have:

$$\mathcal{B}_{\Phi_\alpha}(\rho,\pi)=\mathcal{W}_\alpha(\rho,\pi^\star)-\mathcal{W}_\alpha(\pi,\pi^\star)+\langle v,\rho-\pi\rangle.$$

References

[1] David A McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

[2] J. Shawe-Taylor and R.C. Williamson. A pac-analysis of a bayesian estimator. In *Conference on Learning Theory*, pages 2–9, 1997.

[3] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Functional bregman divergence and bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 44(5), 1998.

[4] David A McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

[5] Peter D. Grünwald. *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[6] Arnak S Dalalyan, Alexandre B Tsybakov, et al. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012.

[7] Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

[8] Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14(Mar):729–769, 2013.

[9] Le Li, Benjamin Guedj, Sébastien Loustau, et al. A quasi-bayesian perspective to online clustering. *Electronic journal of statistics*, 12(2):3071–3113, 2018.

[10] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[12] Andrew Chee and Sebastien Loustau. Sparsity regret bounds for xnor-nets++. *https://hal.archives-ouvertes.fr/hal-03262679*, 2021.

[13] Andrew R Barron. Are bayes rules consistent in information? In *Open Problems in Communication and Computation*, pages 85–91. Springer, 1987.

[14] Anatoli Juditsky, Philippe Rigollet, Alexandre B Tsybakov, et al. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

[15] Imre Csiszár. Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68(1-2):161–186, 1995.

[16] Béla A Frigyik, Santosh Srivastava, and Maya R Gupta. Functional bregman divergence and bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54(11):5130–5139, 2008.

[17] O Catoni. Statistical learning theory and stochastic optimization. In *Ecole d'été de Probabilités de Saint-Flour XXXI—2001*. Lecture Notes in Math.1851. Springer, Berlin, 2004.

[18] Jihong Park, Sumudu Samarakoon, Mehdi Bennis, and Mérouane Debbah. Wireless network intelligence at the edge. *Proceedings of the IEEE*, 107(11):2204–2239, 2019.

[19] Pierre Alquier and Benjamin Guedj. Simpler pac-bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

[20] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

[21] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*, 2019.

[22] Pierre Alquier. Non-exponentially weighted aggregation: regret bounds for unbounded loss functions. *arXiv preprint arXiv:2009.03017*, 2020.

[23] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[24] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[25] Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Apr 2007.

[26] Michael Muskulus and Sjoerd Verduyn-Lunel. Wasserstein distances in the analysis of time series and dynamical systems. *Physica D: Nonlinear Phenomena*, 240(1):45–58, 2011.

[27] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739, 2017.

[28] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

[29] Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. *CoRR*, abs/1503.08596, 2015.

[30] Hicham Janati, Thomas Bazeille, Thirion Bertrand, Cuturi Marco, and Gramfort Alexandre. Multi-subject meg/eeg source imaging with sparse multi-task regression. *NeuroImage*, 220, 2020.

[31] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[32] Cristianini, Nello, and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. repr. *Introduction to Support Vector Machines and other Kernel-Based Learning Methods*, 22, 01 2001.