Non-Exponentially Weighted Aggregation: Regret Bounds for Unbounded Loss Functions

Pierre Alquier [1]

Abstract

We tackle the problem of online optimization with a general, possibly unbounded, loss function. It is well known that when the loss is bounded, the exponentially weighted aggregation strategy (EWA) leads to a regret in $\sqrt{T}$ after T steps. In this paper, we study a generalized aggregation strategy, where the weights no longer depend exponentially on the losses. Our strategy is based on Follow The Regularized Leader (FTRL): we minimize the expected losses plus a regularizer, that is here a ϕ-divergence. When the regularizer is the Kullback-Leibler divergence, we obtain EWA as a special case. Using alternative divergences enables unbounded losses, at the cost of a worst regret bound in some cases.

1. Introduction

We focus in this paper on the online optimization problem as formalized for example in (Shalev-Shwartz, 2012): at each time step $t \in \mathbb{N}$, a learning machine has to make a decision $\theta_t \in \Theta$. Then, a loss function $\ell_t : \Theta \to \mathbb{R}_+$ is revealed and the machine suffers loss $\ell_t(\theta_t)$. Typical example include online linear regression, where $\ell_t(\theta) = \left(y_t - \theta^T x_t\right)^2$ for some $x_t \in \mathbb{R}^d$ and $y_t \in \mathbb{R}$, or online linear classification with $\ell_t(\theta) = \mathbf{1}_{\{y_t \neq \text{sign}(\theta^T x_t)\}}$ or $\ell_t(\theta) = \max(1 - \theta^T x_t, 0)$ for some $x_t \in \mathbb{R}^d$ and $y_t \in \{-1, +1\}$. The objective is to design a strategy for the machine that will ensure that the regret at time T,

$$\mathcal{R}_T := \sum_{t=1}^{T} \ell_t(\theta_t) - \inf_{\theta \in \Theta} \sum_{t=1}^{T} \ell_t(\theta), \tag{1}$$

satisfies $\mathcal{R}_T = o(T)$.

Various strategies were investigated under different assumptions. When the functions ℓ_t are convex, methods based on the sub-gradient of ℓ_t can be used. Such strategies lead to regret in $\sqrt{T}$ under the additional assumption that the ℓ_t are Lipschitz. The regret bounds and strategies are detailed in Chapter 2 in (Shalev-Shwartz, 2012). Another very popular strategy is the so-called exponentially weighted aggregation (EWA) that is based on the probability distribution:

$$\rho^t(\mathrm{d}\theta) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(\theta)\right) \pi(\mathrm{d}\theta)}{\int \exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(\vartheta)\right) \pi(\mathrm{d}\vartheta)} \tag{2}$$

for some prior distribution π on Θ and some learning rate $\eta > 0$. Drawing $\theta_t \sim \rho^t$ leads to an expected regret in $\sqrt{T}$, under the strong assumption that the losses ℓ_t are uniformly bounded, see (Gerchinovitz, 2011).

Is is actually well known that

$$\rho^t = \underset{\rho \in \mathcal{P}(\Theta)}{\text{argmin}} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta \sim \rho}[\ell_s(\theta)] + \frac{\mathrm{KL}(\rho||\pi)}{\eta} \right\} \tag{3}$$

where KL is the Kullback-Leibler divergence and $\mathcal{P}(\Theta)$ is the set of all probability distributions on Θ equipped with a suitable σ-algebra (rigorous notations will come in Subsection 1.2). In this paper, we will study a generalization of the EWA strategy given by

$$\rho^t = \underset{\rho \in \mathcal{P}(\Theta)}{\text{argmin}} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta \sim \rho}[\ell_s(\theta)] + \frac{D_\phi(\rho||\pi)}{\eta} \right\}, \tag{4}$$

where D_ϕ can be any ϕ-divergence (on the condition that this minimizer exists, which will be discussed). Such a strategy is known as "Follow The Regularized Leader" (FTRL) in the online optimization community, and has been studied extensively in the finite Θ case (Shalev-Shwartz, 2012; Hazan, 2016; Orabona, 2019). In Bayesian statistics, (4) was advocated recently in (Li & Turner, 2016; Knoblauch et al., 2019). Some generalization error bounds were proven by (Alquier & Guedj, 2018). However, (Alquier & Guedj, 2018) is written in the batch setting, and the error bounds thus require strong assumptions: for $\theta \in \Theta$, the $(\ell_t(\theta))_{t \in \mathbb{N}}$ must be independent and identically distributed random variables.

Let us call ρ^t the D_ϕ-posterior associated to the ϕ-divergence D_ϕ, the prior π, the learning rate η and the

[1]RIKEN AIP, Tokyo, Japan. Correspondence to: Pierre Alquier <pierrealain.alquier@riken.jp>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

sequence of losses $(\ell_s)_{s\in\mathbb{N}}$. In this paper, we study D_ϕ-posteriors in the online setting, which allows to get completely rid of the stochastic assumptions of (Alquier & Guedj, 2018). First, we prove a regret bound on the D_ϕ-posterior. Our proof follows the same scheme as the study of FTRL in (Shalev-Shwartz, 2012; Orabona, 2019), but in the general case (Θ is not assumed to be finite). Interestingly, when D_ϕ is the χ^2 divergence, our bound holds under very general assumptions – in particular, it does not require that the losses are bounded, Lipschitz, nor convex, but that might be at the cost of a larger regret. We also provide explicit forms for the D_ϕ-posterior. It turns out that it extends the idea of EWA beyond the exponential function, thus the title of the paper. Finally, it is known that EWA is not always feasible in practice. A way to overcome this issue is to use variational approximations of EWA. We thus propose an algorithm that can be seen as the generalization of online variational inference to ϕ-divergences, and provide a regret bound.

1.1. Related works

The case $D_\phi = \mathrm{KL}$, (3) has been studied under the name "multiplicative update", aggregating strategy, EWA (Vovk, 1990; Littlestone & Warmuth, 1994; Catoni, 2004) to name a few. Regret bounds in $\sqrt{T}$ can be found in (Stoltz, 2005; Cesa-Bianchi & Lugosi, 2006; Devaine et al., 2013) in the case where Θ is finite, we refer the reader to (Gerchinovitz, 2011) for the general case. Note that in (Shalev-Shwartz, 2012; Hazan, 2016; Orabona, 2019), EWA is studied as a special case of the FTRL strategy (Follow The Regularized Leader), in the case where Θ is finite. This point of view is the main inspiration of the proofs in this paper, even though we deal here with a general set Θ. Also, note that smaller regret in $\log T$ is feasible under a stronger assumption: exp-concavity (Hazan et al., 2007; Cesa-Bianchi & Lugosi, 2006; Audibert, 2009). (Reid et al., 2015; Mhammedi & Williamson, 2018) also studied small regrets and used for this a generalization of EWA beyond the KL divergence, but here again the study was restricted to a finite set Θ. Similar techniques were also considered by (Audibert & Bubeck, 2009; Zimmert & Seldin, 2019) in incomplete information problems (bandits).

Given a statistical model, that is, a family of densities p_θ with respect to a reference measure ν on some space $\mathcal{X}$, and i.i.d random variables X_1, X_2, ..., drawn from some probability distribution on $\mathcal{X}$, one can define the loss $\ell_t(\theta) = -\log p_\theta(X_t)$. In this case, for $\eta = 1$, ρ^t is actually the posterior distribution of θ given $X_1, \ldots, X_{t-1}$ used in Bayesian statistics. Thus, EWA is also sometimes refered to as "generalized Bayes". (Li & Turner, 2016) proposed (4) as one further generalization of Bayes, using Rényi divergences instead of KL. More recently, (Knoblauch et al., 2019) advocated for a use of taylored losses and divergences. Note that in the batch setting, a general theory allows to provide risk bounds for generalized Bayes (or EWA): PAC-Bayes bounds (Shawe-Taylor & Williamson, 1997; McAllester, 1999; Catoni, 2007; Alquier, 2008), see (Guedj, 2019) for a recent survey. PAC-Bayes bounds for generalized Bayes with the χ^2-divergence were proven in (Honorio & Jaakkola, 2014) and for the Rényi divergence in (Bégin et al., 2016). (Alquier & Guedj, 2018; Ohnishi & Honorio, 2021) showed that while these bounds are usually less tight than standard PAC-Bayes bounds, they allow to get rid of the boundedness assumption in these results. The corresponding optimal posteriors are derived in (Alquier & Guedj, 2018). Other techniques to get rid of boundedness are discussed in (Holland, 2019; Rivasplata et al., 2020) in the batch case.

The idea of variational approximations is to minimize (3) over a restricted set of probability distributions in order to get a feasible approximation of ρ^t, see (Blei et al., 2017; Alquier, 2020) for recent surveys. In the online setting, online variational approximations are studied by (Khan & Lin, 2017; Khan & Nielsen, 2018) and led to the first scaling of Bayesian principles to state-of-the-art neural networks (Osawa et al., 2019). In the i.i.d setting, a series of paper established the first theoretical results on variational inference, for many of them through a connection with PAC-Bayes bounds (Alquier et al., 2016; Sheth & Khardon, 2017; Dziugaite & Roy, 2018; Chérief-Abdellatif & Alquier, 2018; Chérief-Abdellatif, 2019; Wang & Blei, 2019b; Alquier & Ridgway, 2020; Yang et al., 2020; Wang & Blei, 2019a; Jaiswal et al., 2020; Zhang & Gao, 2020; Cherief-Abdellatif, 2020; Plummer et al., 2020; Banerjee et al., 2021; Frazier et al., 2021; Medina et al., 2021). Up to our knowledge, the only regret bound for online variational inference can be found in (Chérief-Abdellatif et al., 2019). The analysis of our generalized online variational approximation is based on this work.

1.2. Notations

Let us now provide accurate notations and a few basic assumptions that will be used throughout the paper. We assume that the set Θ is equipped with a σ-algebra $\mathcal{T}$. Let $\mathcal{P}(\Theta)$ denote the set of all probability distributions on $(\Theta, \mathcal{T})$. Let $\pi \in \mathcal{P}(\Theta)$ be a probability distribution called the *prior* and $(\ell_s)_{s\in\mathbb{N}}$ be a sequence of functions called losses, $\ell_s : \Theta \to \mathbb{R}_+$, assumed to be $\mathcal{T}$-measurable.

Let $\mathcal{M}(\Theta)$ be the set of all finite, signed measures on $(\Theta, \mathcal{T})$. Note that $\mathcal{P}(\Theta) \subsetneq \mathcal{M}(\Theta)$. A norm N on $\mathcal{M}(\Theta)$ is a function $N : \mathcal{M}(\Theta) \to [0, \infty]$ with i) $N(\nu) = 0 \Leftrightarrow \nu = 0$, ii) $N(\nu + \mu) \leq N(\nu) + N(\mu)$ and iii) for $\lambda \in \mathbb{R}$, $N(\lambda.\nu) = |\lambda| N(\nu)$. A norm N on $\mathcal{M}(\Theta)$ induces a metric on $\mathcal{P}(\Theta)$ given by $d_N(\mu, \nu) = N(\nu - \mu)$. For example, the total variation norm $N_{\mathrm{TV}}(\nu) = \sup_{A\in\mathcal{T}} |\nu(A)|$ leads to the

classical total variation distance on $\mathcal{P}(\Theta)$.

Given a sctrictly convex function $\phi : \mathbb{R}_+ \to \mathbb{R}\cup\{+\infty\}$ with $\phi(1) = 0$ and $\inf_{x\geq 0}\phi(x) > -\infty$, define the ϕ-divergence between ρ and $\pi \in \mathcal{P}(\Theta)$ by

$$D_\phi(\rho||\pi) = \mathbb{E}_{\theta\sim\pi}\left[\phi\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right)\right] \text{ if } \rho \ll \pi \qquad (5)$$

and $+\infty$ otherwise. By Jensen's inequality, $D_\phi(\rho||\pi) \geq 0$. Put $\mathcal{P}_{D_\phi,\pi}(\Theta) = \{\rho \in \mathcal{P}(\Theta) : D_\phi(\rho||\pi) < +\infty\}$.

A real-valued function f is said to be upper semicontinuous if for any α, $\{x : f(x) \geq \alpha\}$ is closed. For any real-valued function f, we will denote by f_+ the function defined by $f_+(x) = \max(f(x), 0)$. Given a function $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ that is not uniformly infinite, we will let f^* denote its convex conjugate, that is, for any $y \in \mathbb{R}^d$,

$$f^*(y) = \sup_{x\in\mathbb{R}^d}\left\{x^T y - f(x)\right\} \in \mathbb{R}\cup\{+\infty\}. \qquad (6)$$

1.3. Outline of the paper

We state our general regret bound in Section 2. In particular, we show that for some divergences, our result extends the results known for EWA to unbounded losses. We then provide an explicit form for the D_ϕ-posterior in Section 3. We study generalized online variational inference in Section 4. Section 5 contains the proofs of the results in Sections 2 and 3, the remaining proofs are in the Appendix.

2. A Regret Bound for D_ϕ-Posteriors

2.1. General result

Theorem 2.1. *Assume that there is a norm N on $\mathcal{M}(\Theta)$ and real numbers $\alpha, L > 0$ such that*

- *for any $\rho \in \mathcal{M}(\Theta)$, $N(\rho) \geq N_{\mathrm{TV}}(\rho)$,*
- *for any $t \in \mathbb{N}$, for any $(\rho,\rho') \in \mathcal{P}_{D_\phi,\pi}(\Theta)^2$,*

$$|\mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)] - \mathbb{E}_{\theta\sim\rho'}[\ell_t(\theta)]| \leq LN(\rho-\rho'), \qquad (7)$$

- *for any $\gamma \in [0,1]$, for any $(\rho,\rho') \in \mathcal{P}_{D_\phi,\pi}(\Theta)^2$,*

$$D_\phi(\gamma\rho+(1-\gamma)\rho'||\pi) \leq -2\alpha\gamma(1-\gamma)N(\rho-\rho')^2 + \gamma D_\phi(\rho||\pi) + (1-\gamma)D_\phi(\rho'||\pi). \qquad (8)$$

Assume that each ℓ_t is π-integrable. Then ρ^t in (4) exists, is unique, and

$$\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{D_\phi(\rho||\pi)}{\eta}\right\}. \qquad (9)$$

The assumptions have a simple interpretation: (7) states that each $\mathbb{E}_{\theta\sim\rho}[\ell_t]$ is L-Lipschitz in ρ with respect to the norm N, while (8) states that D_ϕ, as a function of its first argument, is α-strongly convex with respect to N.

Regarding the choice of η, $\eta \sim 1/\sqrt{T}$ seems natural (indeed, in the countable case studied below, it leads to regrets in $\sqrt{T} = o(T)$). However, this choice depends on the horizon T. The doubling trick can be used to avoid this dependence, see e.g. (Cesa-Bianchi & Lugosi, 2006).

When the losses ℓ_t are convex, Jensen's inequality gives $\mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \geq \ell_t[\mathbb{E}_{\theta\sim\rho^t}(\theta)]$. We can thus use the posterior mean $\mathbb{E}_{\theta\sim\rho^t}(\theta)$ instead of a randomized strategy.

Corollary 2.2. *Under the assumptions of Theorem 2.1, assuming moreover that each ℓ_t is convex, and writing $\hat{\theta}_t = \mathbb{E}_{\theta\sim\rho^t}(\theta)$, we have*

$$\sum_{t=1}^T \ell_t(\hat{\theta}_t) \leq \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{D_\phi(\rho||\pi)}{\eta}\right\}. \qquad (10)$$

We now apply Theorem 2.1 to classical divergences.

Example 2.1. *Consider $\phi(x) = x\log x$ so that $D_\phi(\rho||\pi) = \mathrm{KL}(\rho||\pi)$ the Kullback-Leibler divergence. Assuming that, for any $t \in \mathbb{N}$, $|\ell_t(\theta)| \leq L$ holds π-almost surely on θ, we have, for $(\rho,\rho') \in \mathcal{P}_{D_\phi,\pi}(\Theta)^2$,*

$$\int \ell_t(\theta)\rho(\mathrm{d}\theta) - \int \ell_t\rho'(\mathrm{d}\theta) \qquad (11)$$

$$= \int \ell_t(\theta)\left|\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta)\right|\pi(\mathrm{d}\theta) \qquad (12)$$

$$\leq L\int \left|\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta)\right|\pi(\mathrm{d}\theta) \qquad (13)$$

that is, (7) holds with the norm on $\mathcal{M}(\Theta)$:

$$N(\rho) = \int \left|\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right|\pi(\mathrm{d}\theta) = 2N_{\mathrm{TV}}(\rho). \qquad (14)$$

It is known that (8) holds with $\alpha = 1$, the calculations are detailed in the discrete case page 30 in (Shalev-Shwartz, 2012) and can be directly extended to the general case. So

$$\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)] + \eta L^2 T + \frac{\mathrm{KL}(\rho||\pi)}{\eta}\right\}. \qquad (15)$$

This is essentially the same result as Theorem 2.2 page 16 in (Cesa-Bianchi & Lugosi, 2006). Note however that a different proof technique is used there, that leads to better constants: the term in $\eta L^2 T$ is replaced by $\eta L^2 T/8$.

Before considering a new example, let us simply remind the definition of strong convexity: a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is said to be α-strongly convex with respect to a norm $\|\cdot\|$ when, for any $(u, v) \in (\mathbb{R}^d)^2$ and $\gamma \in [0, 1]$, $\varphi(\gamma u + (1-\gamma)v) \leq \gamma\varphi(u) + (1-\gamma)\varphi(v) - \alpha\gamma(1-\gamma)\|u-v\|^2/2$. It is known that when $\varphi : \mathbb{R} \to \mathbb{R}$ is twice differentiable and $\|\cdot\|$ is the Euclidean norm, this is equivalent to the condition: $\forall u, \varphi''(u) \geq \alpha$. Plugging $u = \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)$ and $v = \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta)$ in this definition and integrating with respect to π immediately yields the following.

Lemma 2.3. *Assume that $\phi : \mathbb{R} \to \mathbb{R}$ with $\phi(1) = 0$ is α-strongly convex, then the ϕ-divergence D_ϕ satisfies (8) for the $\mathcal{L}_2(\pi)$-norm*

$$N_2(\rho) := \sqrt{\int \left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right)^2 \pi(\mathrm{d}\theta)} \geq 2N_{\mathrm{TV}}(\rho) \qquad (16)$$

(extended by $+\infty$ when $\rho \ll \pi$ does not hold).

Example 2.2. *Now, $\phi(x) = x^2 - 1$, so $D_\phi(\rho||\pi) = \chi^2(\rho||\pi)$ the χ^2-divergence. As $x \mapsto x^2$ is 2-strongly convex, Lemma 2.3 gives (8) with $N = N_2$. Moreover,*

$$\begin{aligned}\left|\int \ell_t(\theta)\rho(\mathrm{d}\theta) - \int \ell_t\rho'(\mathrm{d}\theta)\right| &\leq \int \ell_t(\theta)\left|\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta)\right|\pi(\mathrm{d}\theta) \\ &\leq N_2(\rho-\rho')\left(\int \ell_t(\theta)^2\pi(\mathrm{d}\theta)\right)^{1/2}. \qquad (17)\end{aligned}$$

So, we obtain (7) under the only assumption that, for any $t \in \mathbb{R}$, $\int \ell_t(\theta)^2\pi(\mathrm{d}\theta) \leq L^2$.

As the application of Theorem 2.1 to the context of the previous example is new to our knowledge, we state it now as a separate corollary.

Corollary 2.4. *Define ρ^t as in (4) with $D_\phi = \chi^2$. Assume that for any $t \in \mathbb{R}$,*

$$\int \ell_t(\theta)^2\pi(\mathrm{d}\theta) \leq L^2 \qquad (18)$$

for some $L > 0$, then

$$\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)] + \frac{\eta L^2 T}{2} + \frac{\chi^2(\rho||\pi)}{\eta}\right\}. \qquad (19)$$

It is important to note that (18) allows choices of priors that are not possible with EWA. Consider for example classification with the exponential loss $\ell_t(\theta) = \exp(-y_t x_t^T\theta)$ or with the hinge loss $\ell_t(\theta) = \max(0, 1 - y_t x_t^T\theta)$, where $\Theta = \mathbb{R}^d$, $y_t \in \{-1, +1\}$ and $\|x_t\| \leq 1$. In this case, (18) will be satisfied with any Gaussian prior. However, we don't have $\ell_t(\theta) \leq L$ uniformly on $\mathbb{R}^d$: this prevents to use EWA with such a prior. Another example (quadratic loss) is provided in Appendix A.

Remark 2.1. *One of the anonymous Referees suggested an alternative proof for Corollary 2.4, in the finite Θ case: rewrite the χ^2 divergence as a weighted quadratic norm between ρ and π, and use the results on weighted ℓ_p norms in Section 5 of (Orabona et al., 2015). This would require some adaptation of the proof to constrain ρ to belong to the simplex, but it would be interesting to compare Corollary 2.4 to the results obtained in this way.*

2.2. Comparison of the bounds in the countable case

In this subsection, $\Theta = \{\theta_0, \theta_1, \dots\}$ is countable. Consider any prior π. In this case, we upper bound the infimum in (9) by its restriction to all Dirac masses. We obtain:

Corollary 2.5. *Under the conditions of Theorem 2.1, assuming in addition that $\Theta = \{\theta_0, \theta_1, \dots\}$ we have:*

$$\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{j\in\mathbb{N}}\left\{\sum_{t=1}^T \ell_t(\theta_j) + \frac{\eta L^2 T}{\alpha} + \frac{\pi(\theta_j)\phi\left(\frac{1}{\pi(\theta_j)}\right) + (1-\pi(\theta_j))\phi(0)}{\eta}\right\}. \qquad (20)$$

In any case, chosing $\eta = 1/\sqrt{T}$ will lead to a regret in $\sqrt{T}$:

$$\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{j\in\mathbb{N}}\left\{\sum_{t=1}^T \ell_t(\theta_j) + \left[\frac{L^2}{\alpha} \pi(\theta_j)\phi\left(\frac{1}{\pi(\theta_j)}\right) + (1-\pi(\theta_j))\phi(0)\right]\sqrt{T}\right\}. \qquad (21)$$

Regarding the dependence on π, let us now to compare the bounds for $D_\phi = \mathrm{KL}$ and $D_\phi = \chi^2$.

Example 2.3. *When $D_\phi = \mathrm{KL}$, the assumption in (7) implies that $0 \leq \ell_t(\theta_j) \leq L$ for any $t, j \in \mathbb{N}$. In the case $\ell_t(\theta) = |y_t - f_\theta(x_t)|$ this can be obtained by assuming that $|y_t| \leq L/2$ where L is known, so that the predictors will be designed or truncated by the user to stay in the interval $[-L/2, L/2]$. In this case, the bound in (20) becomes*

$$\sum_{t=1}^T \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{j\in\mathbb{N}}\left\{\sum_{t=1}^T \ell_t(\theta_j) + \eta L^2 T + \frac{\log\left(\frac{1}{\pi(\theta_j)}\right)}{\eta}\right\}. \qquad (22)$$

Example 2.4. *When $D_\phi = \chi^2$, the bound in (20) becomes*

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{j\in\mathbb{N}}\left\{\sum_{t=1}^{T}\ell_t(\theta_j) + \frac{\eta L^2 T}{2} + \frac{\frac{1}{\pi(\theta_j)}-1}{\eta}\right\} \quad (23)$$

which is much worse for large j's (for which we necessarily will have $\pi(\theta_j)$ small). On the other hand, the assumption in (7) only requires

$$0 \leq \sum_{j=0}^{\infty}\pi(\theta_j)\ell_t(\theta_j)^2 \leq L^2 \quad (24)$$

for any $t \in \mathbb{N}$. In the case $\ell_t(\theta) = |y_t - f_\theta(x_t)|$ this can be obtained by assuming that $|y_t| \leq c$ where c is unknown. Indeed the user might be tempted to use predictors with various magnitude: $|f_{\theta_j}(x)| \leq c_j$ where c_j grows with j. In order to ensure (7) we must take a prior π such that

$$L^2 := 2c^2 + 2\sum_{j=0}^{\infty}\pi(\theta_j)c_j^2 < +\infty. \quad (25)$$

Remark 2.2. *The take-home message of these examples is that the χ^2 divergence allows unbounded losses, but at the cost of a worst regret bound. One of the anonymous Referees asked whether it is possible to get the best of both worlds, that is, unbounded losses with the same regret bound of EWA. This is of course a very important question, we are not aware of existing answers. In an additional example in Appendix A, we show however that it is possible to mitigate the deterioration of the bound in the unbounded case.*

Remark 2.3. *Another anonymous Referee pointed out that (Kalnishkan & Vyugin, 2008) also derived regret bounds for unbounded losses. In their bound (8), there is a term in εT where $\varepsilon > 0$ is some tuning parameter, thus, when ϵ is constant, their bound is in T and not in $\sqrt{T}$. Chosing $\varepsilon = 1/\sqrt{T}$ in their bound leads to non-explicit regret bounds because of the term L_ε.*

3. Explicit D_ϕ-Posteriors: Non-Exponentially Weighted Aggregation

We now provide an explicit formula for the D_ϕ-posterior ρ^t.

Proposition 3.1. *Assume that ϕ is differentiable, strictly convex and define $\tilde{\phi}$ on $\mathbb{R}$ by $\tilde{\phi}(x) = \phi(x)$ if $x \geq 0$ and $\tilde{\phi}(x) = +\infty$ otherwise. Then*

$$\tilde{\phi}^* = \sup_{x\in\mathbb{R}}[xy - \tilde{\phi}(x)] = \sup_{x\geq 0}[xy - \phi(x)] \quad (26)$$

is differentiable and for any $y \in \mathbb{R}$,

$$\nabla\tilde{\phi}^*(y) = \operatorname*{argmax}_{x\geq 0}\{xy - \phi(x)\}. \quad (27)$$

Assume moreover that $\tilde{\phi}^*(\lambda - a) - \lambda \to \infty$ when $\lambda \to \infty$, for any $a \geq 0$. Then

$$\lambda_t = \operatorname*{argmin}_{\lambda\in\mathbb{R}}\left\{\int\tilde{\phi}^*\left(\lambda - \eta\sum_{s=1}^{t-1}\ell_s(\theta)\right)\pi(\mathrm{d}\theta) - \lambda\right\} \quad (28)$$

exists, and

$$\rho^t(\mathrm{d}\theta) = \nabla\tilde{\phi}^*\left(\lambda_t - \eta\sum_{s=1}^{t-1}\ell_s(\theta)\right)\pi(\mathrm{d}\theta) \quad (29)$$

minimizes (4).

In the finite Θ case, (29) was proven by (Reid et al., 2015). An anonymous Referee pointed out that it can also be recovered thanks to (Teboulle, 1992). The techniques used in these papers cannot be used in the general case, though. Instead, we use new tools from (Agrawal & Horel, 2020), that are introduced in the proof.

A similar formula in the context of bandits (with a finite number of arms) can also be found in (Audibert & Bubeck, 2009). The distribution ρ^t is also related to the generalized exponential family in (Grünwald & Dawid, 2004) and the generalized MaxEnt models of (Frongillo & Reid, 2014).

Example 3.1. *First, $\phi(x) = x\log(x)$ so $D_\phi = \mathrm{KL}$. In this case, $\tilde{\phi}^*(y) = \exp(y-1)$ so $\nabla\tilde{\phi}^*(y) = \tilde{\phi}^*(y) = \exp(y-1)$. This leads to*

$$\lambda_t = -\log\int\exp\left[-\eta\sum_{s=1}^{t-1}\ell_s(\theta) - 1\right]\pi(\mathrm{d}\theta), \quad (30)$$

and

$$\rho^t(\mathrm{d}\theta) = \frac{\exp\left[-\eta\sum_{s=1}^{t-1}\ell_s(\theta)\right]\pi(\mathrm{d}\theta)}{\int\exp\left[-\eta\sum_{s=1}^{t-1}\ell_s(\vartheta)\right]\pi(\mathrm{d}\vartheta)}. \quad (31)$$

Example 3.2. *Then $\phi(x) = x^2 - 1$, so $D_\phi = \chi^2$. In this case, $\tilde{\phi}^*(y) = (y^2/4)\mathbf{1}_{\{y\geq 0\}}$, so $\nabla\tilde{\phi}^*(y) = (y/2)_+$ and*

$$\rho^t(\mathrm{d}\theta) = \left[\frac{\lambda_t - \eta\sum_{s=1}^{t-1}\ell_s(\theta)}{2}\right]_+\pi(\mathrm{d}\theta). \quad (32)$$

In this case, λ_t is not available in closed form, but it exists and is the only constant that will make the above sum to 1.

Example 3.3. *More generally, consider $\phi(x) = x^p - 1$. In this case $\nabla\tilde{\phi}^*(y) = (y/p)_+^{1/(p-1)}$, which leads to*

$$\rho^t(\mathrm{d}\theta) = \left[\frac{\lambda_t - \eta\sum_{s=1}^{t-1}\ell_s(\theta)}{p}\right]_+^{\frac{1}{p-1}}\pi(\mathrm{d}\theta). \quad (33)$$

This is quite similar to the Polynomially Weighted Average forecaster studied in Corollary 2.1 page 12 in (Cesa-Bianchi & Lugosi, 2006) in the finite Θ case, even though the normalization procedure is different.

Remark 3.1. *When $\Theta = \{\theta_1, \dots, \theta_M\}$ is finite, these results are simply opbtained by minimizing*

$$F(\rho_1^t, \dots, \rho_M^t) = \sum_{j=1}^M \rho_j^t \sum_{s=1}^{t-1} \ell_s(\theta_j) + \frac{\pi_j \phi\left(\frac{\rho_j^t}{\pi_j}\right)}{\eta} \quad (34)$$

under the constraint that $\rho_1^t + \dots + \rho_M^t = 1$ and that for all j, $\rho_j \geq 0$ (for the sake of simplicity, we wrote $\pi_j := \pi(\theta_j)$ and $\rho_j^t := \rho^t(\theta_j)$). The Lagrange operator is given by

$$\mathcal{L}(\rho_1^t, \dots, \rho_M^t, \lambda, \nu_1, \dots, \nu_M) = \sum_{j=1}^M \rho_j^t \sum_{s=1}^{t-1} \ell_s(\theta_j) + \frac{\sum_{j=1}^M \pi_j \phi\left(\frac{\rho_j^t}{\pi_j}\right)}{\eta} + \lambda \frac{1 - \sum_{j=1}^M \rho_j^t}{\eta} + \sum_{j=1}^M \nu_j \rho_j^t \quad (35)$$

(the notation λ is carefully chosen: it indeed corresponds to (28)). Under the assumptions of Proposition 3.1, the method of Lagrange multipliers will lead to (27). We believe that this derivation gives some insights on (29). So, we provide it in full length in Appendix C.

4. Generalized Online Variational Inference

Apart from the special case of conjugacy, the probability distribution ρ^t in (2) is not tractable. Thus, ρ^t in (4) is not expected to be tractable either. It can of course be implemented via Monte-Carlo methods, but the cost of these methods is often prohibitive for the online setting. In (Chérief-Abdellatif et al., 2019), the authors proposed to use a variational approximation, that is, to minimize (3) on a set smaller than $\mathcal{P}(\Theta)$. We here propose to extend this idea to the minimization in (4).

4.1. The algorithm

Let $(q_\mu)_{\mu \in M}$ be a set of probability distributions in $\mathcal{P}(\Theta)$, where M is some closed convex set in $\mathbb{R}^d$. We could define the variational approximation of ρ^t in this family by:

$$\operatorname*{argmin}_{\mu \in M} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta \sim q_\mu}[\ell_s(\theta)] + \frac{D_\phi(q_\mu || \pi)}{\eta} \right\}, \quad (36)$$

but even this problem might be challenging. We thus replace it by the linearized version

$$\mu_t = \operatorname*{argmin}_{\mu \in M} \left\{ \sum_{s=1}^{t-1} \left\langle \mu, \nabla_{\mu=\mu_s} \mathbb{E}_{\theta \sim q_\mu}[\ell_s(\theta)] \right\rangle + \frac{D_\phi(q_\mu || \pi)}{\eta} \right\}. \quad (37)$$

Observe that when $\mu \mapsto \mathbb{E}_{\theta \sim q_{\mu_s}}[\ell_s(\theta)]$ is convex, (37) can be seen as a convex relaxation of (36) as $\mathbb{E}_{\theta \sim q_\mu}[\ell_s(\theta)] \leq \mathbb{E}_{\theta \sim q_{\mu_s}}[\ell_s(\theta)] + \left\langle \mu, \nabla_{\mu=\mu_s} \mathbb{E}_{\theta \sim q_\mu}[\ell_s(\theta)] \right\rangle$.

Proposition 4.1. *Let $F(\mu) = D_\phi(q_\mu || \pi)$. Assume that F is a differentiable and strictly convex function on $\mathbb{R}^d$, then F^* is differentiable with*

$$\nabla F^*(\lambda) = \operatorname*{argmax}_{\mu \in M} \left[\langle \mu, \lambda \rangle - F(\mu) \right]. \quad (38)$$

Then the solution of (37) exists, is unique and given by

$$\mu_t = \nabla F^* \left(-\eta \sum_{s=1}^{t-1} \nabla_{\mu=\mu_s} \mathbb{E}_{\theta \sim q_\mu}[\ell_s(\theta)] \right). \quad (39)$$

Note the "Mirror Descent" structure of this strategy: we can simply initialize $\lambda_0 = 0$, and update at each step:

$$\begin{cases} \lambda_t = \lambda_{t-1} - \eta \nabla_{\mu=\mu_{t-1}} \mathbb{E}_{\theta \sim q_\mu}[\ell_{t-1}(\theta)], \\ \mu_t = \nabla F^*(\lambda_t) \end{cases} \quad (40)$$

(on mirror descent, see (Nemirovski & Yudin, 1983), and (Shalev-Shwartz, 2012) for an analysis in the online setting). That is, we have a simple update rule for the "dual parameters" λ_t, and then we compute $\mu_t = \nabla F^*(\lambda_t)$. An anonymous Referee also pointed out a similarity with "dual averaging" (Xiao, 2010).

4.2. Regret bound

Theorem 4.2. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. If each $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[\ell_s(\theta)]$ is convex and L-Lipschitz with respect to $\|\cdot\|$, if $\mu \mapsto D_\phi(q_\mu || \pi)$ is α-strongly convex with respect to $\|\cdot\|$,*

$$\sum_{t=1}^T \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\mu \in M} \left\{ \sum_{t=1}^T \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{D_\phi(q_\mu || \pi)}{\eta} \right\}. \quad (41)$$

Let us consider for example a location scale family $(q_\mu)_{\mu \in \mathbb{M}}$ with $\mu = (m, C)$, $m \in \mathbb{R}^k$ and C is a $k \times k$ matrix. That is, when $\vartheta \sim q_{(0, I_k)}$, then $m + C\vartheta \sim q_{(m,C)}$. It is proven in (Domke, 2019), under minimal assumptions on $q_{(0,I_k)}$, that if $\theta \mapsto \ell_t(\theta)$ is convex, then so is $(m, C) \mapsto \mathbb{E}_{\theta \sim q_{(m,C)}}[\ell_t(\theta)]$. In (Chérief-Abdellatif et al., 2019), it is proven that if $\theta \mapsto \ell_t(\theta)$ is L-Lipschitz, then $(m, C) \mapsto \mathbb{E}_{\theta \sim q_{(m,C)}}[\ell_t(\theta)]$ is $2L$-Lipschitz.

Example 4.1. *Consider Gaussian distributions. Using the above parametrization $\mu = (m, C)$ with $q_\mu = q_{(m,C)} = \mathcal{N}(m, C^T C)$, $C \in UT(d)$ the set of full-rank upper triangular $d \times d$ real matrices, and chosing as a prior $\pi = q_{(\bar{m}, \bar{C})}$ we have*

$$\mathrm{KL}(q_{(m,C)}, q_{(\bar{m}, \bar{C})}) = \frac{(m - \bar{m})^T (\bar{C}^T \bar{C})^{-1} (m - \bar{m})}{2} + \frac{\mathrm{tr}[(\bar{C}^T \bar{C})^{-1} (C^T C)] + \log\left(\frac{\det(\bar{C}^T \bar{C})}{\det(C^T C)}\right) - d}{2} \quad (42)$$

which is known to be strongly convex on $\mathbb{R}^d \times \mathcal{M}_C$ where $\mathcal{M}_C$ is any closed bounded subset of $UT(d)$. Formulas for the updates are derived in (Chérief-Abdellatif et al., 2019).

Other parametrizations can also be used in practice. For exponential families, (Khan & Nielsen, 2018) proposed a parametrization based on the expectation of the sufficient statistics. It enjoys very nice properties, and leads to excellent results in practice. However, Theorem 4.2 cannot be applied as the convexity assumption is generally not satisfied with this parametrization. The analysis of this algorithm in this case remains an important open question.

5. Proofs

We first remind a classical result in convex analysis, e.g page 95 in (Boyd & Vandenberghe, 2004) or (2.13) page 43 in (Shalev-Shwartz, 2012).

Lemma 5.1. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a function that is differentiable and strictly convex. Then, its convex conjugate f^* is differentiable and*

$$\nabla f^*(y) = \operatorname*{argmax}_{x \in \mathbb{R}^d} \left[x^T y - f(x)\right]. \tag{43}$$

Proof of Theorem 2.1: Let us start by proving the existence. For the sake of shortness, put

$$F(\rho) = \sum_{s=1}^{t-1} \int \ell_s(\theta)\rho(\mathrm{d}\theta) + \frac{D_\phi(\rho||\pi)}{\eta} \tag{44}$$

for any $\rho \in \mathcal{P}(\Theta)$ and $C = \inf_{\rho \in \mathcal{P}(\Theta)} F(\rho)$. For any $n \in \mathbb{N}$ there is a ρ_n^t such that $C \leq F(\rho_n^t) \leq C + 1/n$. Also, the ρ_n^t are absolutely continous with respect to π, otherwise, $D_\phi(\rho_n^t||\pi) = +\infty$. Then

$$\begin{aligned} C &\leq F\left(\frac{\rho_n^t + \rho_m^t}{2}\right) \\ &\leq \frac{F(\rho_n^t) + F(\rho_m^t)}{2} - \alpha N(\rho_n^t - \rho_m^t)^2/2 \\ &\leq C + 1/(2n) + 1/(2m) - \alpha N(\rho_n^t - \rho_m^t)^2/2 \end{aligned}$$

which leads to $N(\rho_n^t - \rho_m^t)^2 \leq 1/(\alpha n) + 1/(\alpha m)$, proving that ρ_n^t is a Cauchy sequence w.r.t the norm N. Thus, it is also a Cauchy sequence w.r.t the norm N_{TV} by the inequality $N(\rho_n^t - \rho_m^t) \geq N_{\mathrm{TV}}(\rho_n^t - \rho_m^t)$. From Proposition A.10 page 512 in (Ghosal & Van der Vaart, 2017), the set of probability distributions that are absolutely continuous with respect to π is complete for N_{TV}, so, there is a ρ_∞^t absolutely continuous with respect to π such that $N_{\mathrm{TV}}(\rho_n^t - \rho_\infty^t) \xrightarrow[n\to\infty]{} 0$. This can be rewritten as

$$\int \left|\frac{\mathrm{d}\rho_n^t}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho_\infty^t}{\mathrm{d}\pi}(\theta)\right| \pi(\mathrm{d}\theta) \xrightarrow[n\to\infty]{} 0. \tag{45}$$

This means that the nonnegative random variable $\frac{\mathrm{d}\rho_n^t}{\mathrm{d}\pi}$ converges to the random variable $\frac{\mathrm{d}\rho_\infty^t}{\mathrm{d}\pi}$ in $\mathcal{L}_1$, thus it converges in probability, and thus, there exists a subsequence $\frac{\mathrm{d}\rho_{n_k}^t}{\mathrm{d}\pi}$ that converges almost surely to $\frac{\mathrm{d}\rho_\infty^t}{\mathrm{d}\pi}$. Now, ϕ being lower-bounded, we can use Fatou lemma:

$$\begin{aligned} C &\leq F(\rho_\infty^t) \\ &= \int \left[\sum_{s=1}^{t-1} \ell_s(\theta)\frac{\mathrm{d}\rho_\infty^t}{\mathrm{d}\pi}(\theta) + \phi\left(\frac{\mathrm{d}\rho_\infty^t}{\mathrm{d}\pi}(\theta)\right)\right]\pi(\mathrm{d}\theta) \\ &= \int \liminf_k \left[\sum_{s=1}^{t-1} \ell_s(\theta)\frac{\mathrm{d}\rho_{n_k}^t}{\mathrm{d}\pi}(\theta) + \phi\left(\frac{\mathrm{d}\rho_{n_k}^t}{\mathrm{d}\pi}(\theta)\right)\right]\pi(\mathrm{d}\theta) \\ &\leq \liminf_k \int \left[\sum_{s=1}^{t-1} \ell_s(\theta)\frac{\mathrm{d}\rho_{n_k}^t}{\mathrm{d}\pi}(\theta) + \phi\left(\frac{\mathrm{d}\rho_{n_k}^t}{\mathrm{d}\pi}(\theta)\right)\right]\pi(\mathrm{d}\theta) \\ &= \liminf_k F(\rho_{n_k}^t) \leq \liminf_k \left(C + \frac{1}{n_k}\right) = C \end{aligned}$$

which proves that ρ_∞^t is indeed a minimizer of (4) (the previous series of inequalities follows the proof of the fact that ϕ-divergences are lower semi-continuous in Chapter 2 in (Keziou, 2003)). Let us now prove its uniqueness: assume that $\tilde{\rho}_\infty^t \neq \rho_\infty^t$ is another minimizer. Put $\bar{\rho}_\infty^t = (\tilde{\rho}_\infty^t + \rho_\infty^t)/2$, using (8) we have:

$$\begin{aligned} C \leq F(\bar{\rho}_\infty^t) &\leq \frac{F(\tilde{\rho}_\infty^t) + F(\rho_\infty^t)}{2} - \frac{\alpha}{2}N(\tilde{\rho}_\infty^t - \rho_\infty^t) \\ &= C - \alpha N(\tilde{\rho}_\infty^t - \rho_\infty^t)/2 < C, \end{aligned} \tag{46}$$

a contradiction. Thus, $\rho^t = \rho_\infty^t$ exists and is unique.

Let us now prove the regret bound. We follow the main steps of the analysis of the FTRL. We start by proving by induction on T that

$$\begin{aligned} &\sum_{s=1}^{T} \int \ell_s(\theta)\rho^{s+1}(\mathrm{d}\theta) \\ &\qquad \leq \inf_{\rho \in \mathcal{P}(\Theta)} \left[\sum_{s=1}^{T} \int \ell_s(\theta)\rho(\mathrm{d}\theta) + \frac{D_\phi(\rho||\pi)}{\eta}\right]. \end{aligned} \tag{47}$$

Indeed, for $T = 0$, the statement is simply $D_\phi(\rho||\pi)/\eta \geq 0$ that is true by definition of a divergence. Now, assuming that (47) is true at step T, we add $\int \ell_s(\theta)\rho^{T+1}(\mathrm{d}\theta)$ to each side of (47) to obtain

$$\begin{aligned} &\sum_{s=1}^{T+1} \int \ell_s(\theta)\rho^{s+1}(\mathrm{d}\theta) \leq \int \ell_{T+1}(\theta)\rho^{T+1}(\mathrm{d}\theta) \\ &\qquad + \min_{\rho \in \mathcal{P}(\Theta)} \left[\sum_{s=1}^{T} \int \ell_s(\theta)\rho(\mathrm{d}\theta) + \frac{D_\phi(\rho||\pi)}{\eta}\right]. \end{aligned} \tag{48}$$

Upper bounding the minimum in ρ by the value for $\rho = \rho^{T+1}$ we obtain

$$\sum_{s=1}^{T+1} \int \ell_s(\theta)\rho^{s+1}(\mathrm{d}\theta)$$
$$\leq \sum_{s=1}^{T+1} \int \ell_s(\theta)\rho^{T+1}(\mathrm{d}\theta) + \frac{D_\phi(\rho^{T+1}||\pi)}{\eta} \quad (49)$$
$$= \min_{\rho\in\mathcal{P}(\Theta)} \left[\sum_{s=1}^{T+1} \int \ell_s(\theta)\rho(\mathrm{d}\theta) + \frac{D_\phi(\rho||\pi)}{\eta}\right] \quad (50)$$

by the definition of ρ^{T+1}. This ends the proof of (47).

Now that (47) is proven, adding $\sum_{s=1}^{T} \int \ell_s(\theta)\rho^s(\mathrm{d}\theta)$ to each side and rearranging the terms leads to

$$\sum_{s=1}^{T} \int \ell_s(\theta)\rho^s(\mathrm{d}\theta) \leq \inf_{\rho\in\mathcal{P}(\Theta)} \left[\sum_{s=1}^{T} \int \ell_s(\theta)\rho(\mathrm{d}\theta) + \sum_{s=1}^{T} \left(\int \ell_s(\theta)\rho^s(\mathrm{d}\theta) - \int \ell_s(\theta)\rho^{s+1}(\mathrm{d}\theta)\right) + \frac{D_\phi(\rho||\pi)}{\eta}\right]. \quad (51)$$

The last step is thus to prove that, for any s,

$$\int \ell_s(\theta)\rho^s(\mathrm{d}\theta) - \int \ell_s(\theta)\rho^{s+1}(\mathrm{d}\theta) \leq \frac{\eta L^2}{\alpha}. \quad (52)$$

First, by (7),

$$\int \ell_s(\theta)\rho^s(\mathrm{d}\theta) - \int \ell_s(\theta)\rho^{s+1}(\mathrm{d}\theta) \leq LN(\rho^s - \rho^{s+1}). \quad (53)$$

Define $H_s(\rho) = \int \sum_{t=1}^{s-1} \ell_t(\theta)\rho(\mathrm{d}\theta) + D_\phi(\rho||\pi)/\eta$. Dividing (8) by η and adding $\gamma \int \sum_{t=1}^{s-1} \ell_t(\theta)\rho(\mathrm{d}\theta) + (1-\gamma)\int \sum_{t=1}^{s-1} \ell_s(\theta)\rho'(\mathrm{d}\theta)$ to each side, we obtain:

$$H_s(\gamma\rho + (1-\gamma)\rho') \leq \gamma H_s(\rho) + (1-\gamma)H_s(\rho') - \frac{2\alpha}{\eta}\gamma(1-\gamma)N(\rho-\rho')^2. \quad (54)$$

Now, put $h_s(u) = H_s(u\rho^s + (1-u)\rho^{s+1})$. Thanks to (54), we have $h_s(\gamma u + (1-\gamma)u') \leq \gamma h_s(u) + (1-\gamma)h_s(u') - \frac{\alpha}{2}\gamma(1-\gamma)(u-u')^2N(\rho^s-\rho^{s+1})^2$, that is: h_s is $\alpha N(\rho^s - \rho^{s+1})^2$-strongly convex. Moreover, $h_s(u)$ is minimized by $u = 0$, because by definition, $H_s(\rho)$ is minimized by $\rho = \rho^s$. Using the well-known property of strongly convex functions of a real variable, we obtain:

$$h_s(u) \geq h_s(0) + \frac{\alpha N(\rho^s - \rho^{s+1})^2}{2\eta}u^2 \quad (55)$$

and so, for $u = 1$,

$$H_s(\rho_s) \geq H_s(\rho^{s+1}) + \frac{\alpha N(\rho_s - \rho^{s+1})^2}{2\eta}. \quad (56)$$

We obtain in a similar way:

$$H_{s+1}(\rho^{s+1}) \geq H_{s+1}(\rho^s) + \frac{\alpha N(\rho_s - \rho^{s+1})^2}{2\eta}. \quad (57)$$

Summing (56) and (57) gives:

$$\int \ell_s(\theta)\rho^s(\mathrm{d}\theta) - \int \ell_s(\theta)\rho^{s+1}(\mathrm{d}\theta) \geq \frac{\alpha N(\rho^s - \rho^{s+1})^2}{\eta}. \quad (58)$$

Combining (58) with (53) gives:

$$N(\rho^s - \rho^{s+1}) \leq \sqrt{\frac{\eta}{\alpha}\left(\int \ell_s(\theta)\rho^s(\mathrm{d}\theta) - \int \ell_s(\theta)\rho^{s+1}(\mathrm{d}\theta)\right)} \quad (59)$$

which, using again (53), gives (52). □

Proof of Proposition 3.1: First, note that (26) is obvious from the definition of $\tilde{\phi}$. Then apply Lemma 5.1 to $f = \tilde{\phi}$ that is α-strongly convex. We obtain (27).

Let us now define $F_{\phi,\pi}(\rho) = D_\phi(\rho||\pi)$ and its convex conjugate, for $g : \Theta \to \mathbb{R}$ that is π-integrable,

$$F^*_{\phi,\pi}(g) = \sup_{\rho\in\mathcal{P}_{D,\phi}(\Theta)} \left[\int g(\theta)\rho(\mathrm{d}\theta) - D_\phi(\rho||\pi)\right]. \quad (60)$$

Then, by Proposition 4.3.2 in (Agrawal & Horel, 2020),

$$F^*_{\phi,\pi}(g) = \inf_{\lambda\in\mathbb{R}} \left\{\int \tilde{\phi}^*(g(\theta)+\lambda)\pi(\mathrm{d}\theta) - \lambda\right\}, \quad (61)$$

where the infimum is actually reached as soon as it is finite.

In our case, we apply this result to the nonpositive, π-integrable function

$$g_t(\theta) = -\eta \sum_{s=1}^{t-1} \ell_s(\theta). \quad (62)$$

Using Jensen's inequality, we have:

$$\int \tilde{\phi}^*(g_t(\theta)+\lambda)\pi(\mathrm{d}\theta) - \lambda \geq \tilde{\phi}^*\left(\int g_t(\theta)\pi(\mathrm{d}\theta) + \lambda\right) - \lambda. \quad (63)$$

This quantity is convex, ≥ 0 when $\lambda \leq 0$, and $\to \infty$ when $\lambda \to \infty$. So, its infimum is finite, and thus, according to (Agrawal & Horel, 2020), it is reached by some $\lambda = \lambda_t$.

Let us now define ρ^t as in (31): $\rho^t(\mathrm{d}\theta) = \nabla\tilde{\phi}^*(\lambda_t + g_t(\theta))\pi(\mathrm{d}\theta)$. A first step is to check that ρ^t is indeed a probability distribution. By differentiating (28) with respect to λ we obtain:

$$\frac{\partial}{\partial\lambda}\left[\int\tilde{\phi}^*(\lambda_t + g_t(\theta))\pi(\mathrm{d}\theta)\right]_{\lambda=\lambda_t} = 1. \tag{64}$$

Note that $\nabla\tilde{\phi}^*$ is the differential of a convex, differentiable function. Thus, it is a nondecreasing function, and it has no jumps. So, it is continuous, and so, we have

$$\int\rho^t(\mathrm{d}\theta) = \int\nabla\tilde{\phi}^*(\lambda + g(\theta))\pi(\mathrm{d}\theta) = 1. \tag{65}$$

Let us now remind the following formula, which can be found for example in (Boyd & Vandenberghe, 2004) page 95, for a convex and differentiable function f:

$$f^*(\nabla f(x)) = x^T\nabla f(x) - f(x). \tag{66}$$

Applying this formula to $f = \tilde{\phi}^*$ that is convex and differentiable, we obtain:

$$\tilde{\phi}^{**}(\nabla\tilde{\phi}^*(x)) = x^T\nabla\tilde{\phi}^*(x) - \tilde{\phi}^*(x). \tag{67}$$

Now, it is easy to check that the function $\tilde{\phi}$ is upper semi-continuous and convex. So, $\tilde{\phi}^{**} = \tilde{\phi}$ (e.g Exercice 3.39 page 121 in (Boyd & Vandenberghe, 2004)), and we obtain:

$$\tilde{\phi}(\nabla\tilde{\phi}^*(x)) = x^T\nabla\tilde{\phi}^*(x) - \tilde{\phi}^*(x). \tag{68}$$

So, we have:

$$\begin{aligned}\int\left[-\frac{g_t(\theta)}{\eta}\right]\rho^t(\mathrm{d}\theta) + \frac{D_\phi(\rho^t||\pi)}{\eta} \\ = \int\left[-\frac{g_t(\theta)}{\eta}\nabla\tilde{\phi}^*(\lambda_t + g_t(\theta))\right. \\ \left. + \frac{1}{\eta}\tilde{\phi}\left(\nabla\tilde{\phi}^*(\lambda_t + g_t(\theta))\right)\right]\pi(\mathrm{d}\theta)\end{aligned}$$

and applying (68) gives

$$\begin{aligned}\int\left[-\frac{g_t(\theta)}{\eta}\right]\rho^t(\mathrm{d}\theta) + \frac{D_\phi(\rho^t||\pi)}{\eta} \\ = \int\left[-\frac{g_t(\theta)}{\eta}\nabla\tilde{\phi}^*(\lambda_t + g(\theta))\right. \\ + \frac{(\lambda_t + g(\theta))}{\eta}\nabla\tilde{\phi}^*(\lambda_t + g(\theta)) \\ \left. - \tilde{\phi}^*(\lambda_t + g_t(\theta))\right]\pi(\mathrm{d}\theta) \\ = \lambda_t - \int\tilde{\phi}^*(\lambda_t + g_t(\theta))\pi(\mathrm{d}\theta) \\ = \min_{\rho\in\mathcal{P}_{D,\pi}(\Theta)}\left[-\int\frac{g_t(\theta)}{\eta}\rho^t(\mathrm{d}\theta) + \frac{D_\phi(\rho^t||\pi)}{\eta}\right]\end{aligned} \tag{69}$$

by (61). So ρ^t minimizes the desired criterion. □

The proof of Proposition 4.1 and Theorem 4.2 are provided in Appendix B.

Acknowledgements

The anonymous Referees suggested many clarifications and connections to existing works. I thank them deeply for this. I also would like to thank Emtiyaz Khan and all the members of the ABI team (RIKEN AIP), Dimitri Meunier (IIT Genoa), Badr-Eddine Chérief-Abdellatif (Univ. of Oxford), Jeremias Knoblauch and Lionel Riou-Durand (Univ. of Warwick) for useful discussions/comments that led to improvements of the paper.

References

Agrawal, R. and Horel, T. Optimal bounds between f-divergences and integral probability metrics. In *Proceedings of the 37th International Conference on Machine Learning, PMLR 119*, pp. 115–124, 2020.

Alquier, P. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17 (4):279–304, 2008.

Alquier, P. Approximate Bayesian inference. *Entropy*, 22: 1272, 2020.

Alquier, P. and Guedj, B. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Alquier, P. and Ridgway, J. Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, 3(48):1475–1497, 2020.

Alquier, P., Ridgway, J., and Chopin, N. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(239):1–41, 2016.

Audibert, J. Y. Fast learning rates in statistical inference through aggregation. *Annals of Statistics*, 37(4):1591–1646, 2009.

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pp. 1–122, 2009.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Banerjee, I., Rao, V. A., and Honnappa, H. PAC-Bayes bounds on Variational Tempered Posteriors for Markov Models. *Entropy* 23(3):313, 2021.

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*, pp. 435–444, 2016.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Catoni, O. *Statistical Learning Theory and Stochastic Optimization*. Saint-Flour Summer School on Probability Theory 2001, Lecture Notes in Mathematics. Springer, 2004.

Catoni, O. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. IMS Lecture Notes, Monograph Series, 56. 2007.

Cesa-Bianchi, N. and Lugosi, G. Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, 43(3): 247–264, 2001.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

Chérief-Abdellatif, B.-E. Convergence rates of variational inference in sparse deep learning. In *Proceedings of the 37th International Conference on Machine Learning, PMLR 119*, pp. 1831–1842, 2019.

Cherief-Abdellatif, B.-E. *Contributions to the theoretical study of variational inference and robustness*. PhD thesis, Institut Polytechnique de Paris, 2020.

Chérief-Abdellatif, B.-E. and Alquier, P. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12 (2):2995–3035, 2018.

Chérief-Abdellatif, B.-E., Alquier, P., and Khan, M. E. A generalization bound for online variational inference. *Proceedings of The Eleventh Asian Conference on Machine Learning, PMLR*, 101:662–677, 2019.

Devaine, M., Gaillard, P., Goude, Y., and Stoltz, G. Forecasting electricity consumption by aggregating specialized experts. *Machine Learning*, 90(2):231–260, 2013.

Domke, J. Provable smoothness guarantees for black-box variational inference. In *Proceedings of the 37th International Conference on Machine Learning, PMLR 119*, pp. 2587–2596, 2019.

Dziugaite, G. K. and Roy, D. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. In *International Conference on Machine Learning*, pp. 1377–1386, 2018.

Frazier, D. T., Loaiza-Maya, R., Martin, G. M., and Koo, B. Loss-based variational Bayes prediction. *arXiv preprint arXiv:2104.14054*, 2021.

Frongillo, R. and Reid, M. D. Convex foundations for generalized MaxEnt models. In *AIP Conference Proceedings*, volume 1636, pp. 11–16. American Institute of Physics, 2014.

Gerchinovitz, S. *Prédiction de suites individuelles et cadre statistique classique: étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation*. PhD thesis (in English), Paris 11, 2011.

Ghosal, S. and Van der Vaart, A. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Grünwald, P. D. and Dawid, A. P. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

Guedj, B. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, 2019.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3–4):157–325, 2016.

Hazan, E., Agarwal, A. and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Holland, M. PAC-Bayes under potentially heavy tails. In *Advances in Neural Information Processing Systems*, pp. 2715–2724, 2019.

Honorio, J. and Jaakkola, T. Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pp. 384–392, 2014.

Jaiswal, P., Honnappa, H., and Rao, V. A. Asymptotic consistency of loss-calibrated variational Bayes. *Stat*, 9 (1), 2020.

Kalnishkan, Y. and Vyugin, M. V. The weak aggregating algorithm and weak mixability. *Journal of Computer and System Sciences*, 74(8):1228–1244, 2008.

Keziou, A. *Utilisation des divergences entre mesures en statistique inferentielle*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2003.

Khan, M. E. and Lin, W. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *PMLR: Proceedings of ICML*, 54:878–887, 2017.

Khan, M. E. and Nielsen, D. Fast yet simple natural-gradient descent for variational inference in complex models. Invited paper at ISITA 2018, 2018.

Knoblauch, J., Jewson, J., and Damoulas, T. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*, 2019.

Li, Y. and Turner, R. E. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pp. 1073–1081, 2016.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

McAllester, D. A. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Medina, M. A., Olea, J. L. M., Rush, C., and Velez, A. On the robustness to misspecification of α-posteriors and their variational approximations. *arXiv preprint arXiv:2104.08324*, 2021.

Mhammedi, Z. and Williamson, R. C. Constant regret, generalized mixability, and mirror descent. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7430–7439, 2018.

Nemirovski, A. and Yudin, D. *Problem complexity and method efficiency in optimization*. Wiley-Inter-science Series in Discrete Mathematics, Wiley, XV, 1983.

Ohnishi, Y. and Honorio, J. Novel change of measure inequalities with applications to PAC-Bayesian bounds and Monte Carlo estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1711–1719. PMLR, 2021.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Orabona, F., Crammer, K., and Cesa-Bianchi, N. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

Osawa, K., Swaroop, S., Khan, M. E., Jain, A., Eschenhagen, R., Turner, R. E., and Yokota, R. Practical deep learning with Bayesian principles. In *Advances in neural information processing systems*, pp. 4287–4299, 2019.

Plummer, S., Pati, D., and Bhattacharya, A. Dynamics of coordinate ascent variational inference: A case study in 2d ising models. *Entropy*, 22(11):1263, 2020.

Reid, M. D., Frongillo, R. M., Williamson, R. C., and Mehta, N. Generalized mixability via entropic duality. In *Conference on Learning Theory*, pp. 1501–1522, 2015.

Rivasplata, O., Kuzborskij, I., Szepesvári, C., and Shawe-Taylor, J. PAC-Bayes analysis beyond the usual bounds. *arXiv preprint arXiv:2006.13057*, 2020.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Shawe-Taylor, J. and Williamson, R. C. A PAC analysis of a Bayesian estimator. In *Tenth annual conference on Computational learning theory*, volume 6, pp. 2–9, 1997.

Sheth, R. and Khardon, R. Excess risk bounds for the Bayes risk using variational inference in latent Gaussian models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5151–5161. Curran Associates, Inc., 2017.

Stoltz, G. *Incomplete information and internal regret in prediction of individual sequences*. PhD Thesis, Université Paris Sud-Paris XI, 2005.

Teboulle, M. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690, 1992.

Vovk, V. G. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, 1990.

Wang, Y. and Blei, D. Variational Bayes under model misspecification. In *Advances in Neural Information Processing Systems*, pp. 13357–13367, 2019a.

Wang, Y. and Blei, D. M. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019b.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

Yang, Y., Pati, D., and Bhattacharya, A. α-variational inference with statistical guarantees. *Annals of Statistics*, 48 (2):886–905, 2020.

Zhang, F. and Gao, C. Convergence rates of variational posterior distributions. *Annals of Statistics*, 48(4):2180–2207, 2020.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 467–475. PMLR, 2019.