APPENDIX

We remind that Appendix A contains an application of Theorem (2.1) to the continuous case. Appendix B contains the proofs of Proposition 4.1 and Theorem 4.2. Finally, Appendix C contains the derivation of ρ^t in the finite case thanks to Lagrange method of multipliers.

A. Comparison of the Bounds in the Continuous Case

As another example of application of Theorem (2.1), let us consider the case $\ell_t(\theta) = (\theta - y_t)^2$. We assume that $\sup_{t\in\mathbb{N}} |y_t| = C < +\infty$. We prove the following statements:

- for some choice of η and π, EWA (that is, (4) with $D_\phi = \mathrm{KL}$) leads to

  $$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{m\in[-C,C]} & \left\{ \sum_{t=1}^{T} (y_t - m)^2 \right. \\ & + 4C^2\sqrt{T\log(T)}(1+o(1)) \\ = \sum_{t=1}^{T} & (y_t - \bar{y}_T)^2 \\ & \left. + 4C^2\sqrt{T\log(T)}(1+o(1)) \right\}\end{aligned}$$

  where $\bar{y}_T = (1/T)\sum_{t=1}^{T} y_t$, but C has to be known by the user to reach this.

- for some choice of η and π, (4) with $D_\phi = \chi^2$ leads to

  $$\begin{aligned}& \sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \\ & \leq \inf_{m\in\mathbb{R}} \left\{ \sum_{t=1}^{T} (m - y_t)^2 + C' T^{\frac{2}{3}} (1+|m|)^5 \right\} \quad (70)\end{aligned}$$

  where C' is a constant that depends only on C, and none of these constants have to be known by the user.

There are various ways of using EWA in this context. The important point is that they all require the support of the prior π to be bounded (or to truncate its support at some point):

1. a first option is to use as a prior π the uniform distribution on $[-C, C]$. Of course, this is possible only if one knows C in advance! In this case, the losses are bounded by $4C^2$ and so the regret bound is given by

   $$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{\rho\in\mathcal{P}(\Theta)} & \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho}[(y_t - \theta)^2] \right. \\ & \left. + \frac{\eta C^2 T}{2} + \frac{\mathrm{KL}(\rho\|\pi)}{\eta} \right\}. \quad (71)\end{aligned}$$

   For $m \in [-C, C]$ and $\delta \in (0, 1)$, define $\rho_{m,\delta}$ as the uniform distribution on an inverval of length δC that contains m (one could think of $[m - \delta C/2, m + \delta C/2]$ but when $m = C$, this interval would not be included in $[-C, C]$...). We obtain:

   $$\begin{aligned}& \sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \\ & \leq \inf_{m\in[-C,C]} \inf_{\delta\in(0,1)} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho}[(\theta - y_t)^2] \right. \\ & \qquad \left. + 8\eta C^4 T + \frac{\log\left(\frac{2}{\delta}\right)}{\eta} \right\} \\ & \leq \inf_{m\in[-C,C]} \inf_{\delta\in(0,1)} \left\{ \sum_{t=1}^{T} \Big((y_t - m)^2 + C^2\delta^2 \right. \\ & \qquad \left. + 2C\delta|y_t - m| \Big) + 8\eta C^4 T + \frac{\log\left(\frac{2}{\delta}\right)}{\eta} \right\} \\ & \leq \inf_{m\in[-C,C]} \inf_{\delta\in(0,1)} \left\{ \sum_{t=1}^{T} (y_t - m)^2 + 5TC^2\delta \right. \\ & \qquad \left. + 8\eta C^4 T + \frac{\log\left(\frac{2}{\delta}\right)}{\eta} \right\} \\ & = \inf_{m\in[-C,C]} \left\{ \sum_{t=1}^{T} (y_t - m)^2 \right. \\ & \qquad \left. + 8\eta C^4 T + \frac{1 + \log\left(10TC^2\eta\right)}{\eta} \right\}\end{aligned}$$

   reached for $\delta = 1/(5\eta TC^2)$ (in $(0, 1)$ for T large enough). The choice $\eta = \sqrt{\log(T)/(4C^2\sqrt{T})}$ gives:

   $$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{m\in[-C,C]} & \left\{ \sum_{t=1}^{T} (y_t - m)^2 \right. \\ & \left. + 4C^2\sqrt{T\log(T)}(1+o(1)) \right\}. \quad (72)\end{aligned}$$

2. a second strategy is detailed for example in (Gerchinovitz, 2011), it consists in taking a heavy-tailed distribution on $\mathbb{R}$ for π, but to use as a predictor the projection of θ on the interval $[-C, C]$, that is, changing the loss in $|y_t - \mathrm{proj}_{[-C,C]}(\theta)|$. This would lead to a regret bound similar to (72), without improving the applicability of the result, in the sense that one has to know C to use the procedure.

3. a third approach was mentioned by an anonymous Referee, and can in principle be used when C is unknown. In this case, we take π as the uniform prior on $[-c, c]$ for some $c > 0$. The important point is that the loss $\ell_t(\theta)$ belongs to the interval $[0, (c+C)^2]$ whose upper bound is unknown. Based on techniques developped in (Cesa-Bianchi & Lugosi, 2001; Auer et al., 2002), Theorem 6 in (Cesa-Bianchi et al., 2007) upper bounds the regret of an adaptive version of EWA that can be used in the case where the losses belongs to an (unknown) bounded interval. This theorem is written in the finite Θ case, but it seems to be direct to extend the result to the general case. If one is "lucky", that is, if $c \geq C$, then one would recover a regret bound similar to (72). However, it might be that $C > c$. In this case, we only have the guarantee to perform as well as the best predictor in $[-c, c]$. If the best predictor m satisfies $|m| > c$ then this gives a linear regret.

Let us now use the strategy (4) with D being the χ^2 divergence, and with a prior π that is the student distribution $\mathcal{T}(k)$ with $k = 4$ degrees of freedom. First,

$$\int \ell_t(\theta)^2 \pi(\mathrm{d}\theta) = \int |y_t - \theta|^4 \pi(\mathrm{d}\theta)$$
$$\leq 8 \int |y_t|^4 \pi(\mathrm{d}\theta) + 8 \int |\theta|^4 \pi(\mathrm{d}\theta) \leq 8(C^4 + 24). \quad (73)$$

This gives the regret bound

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho}[(y_t - \theta)^2] \right.$$
$$\left. + \eta 8(C^4 + 24)T + \frac{\chi^2(\rho||\pi)}{\eta} \right\} \quad (74)$$

and here, let us consider, for $m \in \mathbb{R}$ and $\delta \in (0, 1)$, the uniform distribution $\rho_{m,\delta}$ on an interval of length δC that contains m. The regret bound becomes:

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)]$$
$$\leq \inf_{m\in\mathbb{R}} \inf_{\delta\in(0,1)} \left\{ \sum_{t=1}^{T} (m - y_t)^2 + \delta 5 C^2 T \right.$$
$$\left. + \eta 8(C^4 + 24)T + \frac{\chi^2(\rho_{m,\delta}||\pi)}{\eta} \right\}. \quad (75)$$

Note that the density of $\mathcal{T}(k)$ with respect to the Lebesgue measure is given by:

$$\frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \quad (76)$$

so we can derive the upper bound

$$\chi^2(\rho_{m,\delta}||\pi) \leq \frac{C''}{\delta\eta}(1 + |m|)^5 \quad (77)$$

for some $C'' > 0$ that depends only on C. This time, the choices $\delta = \eta = 1/T^{1/3}$ lead to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)]$$
$$\leq \inf_{m\in\mathbb{R}} \left\{ \sum_{t=1}^{T} (m - y_t)^2 + C' T^{\frac{2}{3}} (1 + |m|)^5 \right\} \quad (78)$$

where C' is a constant that depends only on C. The important point is that the strategy can be implemented without the knowledge of C nor C'. But also, this has an important cost, that is, the regret is now in $T^{2/3}$.

Remark A.1. *An anonymous Referee suggested that it is possible to first build predictors on nested intervals, and then to aggregate them via EWA to obtain adaptation to the unknown constant C. However, these predictors are not uniformly bounded, so we don't see how to apply the standard results on EWA to them.*

However, this suggestion leads to an improvement on (70) that will combine the ideas of EWA and non-exponentially weighted aggregation. The idea is to use EWA on nested intervals, and then to aggregate them using the χ^2 bound, which does not require boundedness.

More precisely, define ρ_k^t as the result of using EWA with a uniform prior on $[-k, k]$, for any $k \in \mathbb{N} \setminus \{0\}$. We have:

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho_k^t}[\ell_t(\theta)] \leq \inf_{m\in[-k,k]} \left\{ \sum_{t=1}^{T} (y_t - m)^2 \right.$$
$$\left. + 4k^2 \sqrt{T \log(T)} (1 + o(1)) \right\}.$$

It is then possible to adapt Corollary 2.4 to aggregate the various ρ_k^t, using a prior π on k. This leads to a posterior ρ^t on k such that

$$\sum_{t=1}^{T} \mathbb{E}_{k\sim\rho^t} \mathbb{E}_{\theta\sim\rho_k^t}[\ell_t(\theta)]$$
$$\leq \inf_{k\geq 1} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho_k^t}[\ell_t(\theta)] + \frac{\eta L^2 T}{2} + \frac{\frac{1}{\pi(k)-1}}{\eta} \right\}$$
$$\leq \inf_{k\geq 1} \inf_{m\in[-k,k]} \left\{ \sum_{t=1}^{T} (y_t - m)^2 \right.$$
$$\left. + 4k^2 \sqrt{T \log(T)} (1 + o(1)) + \frac{\eta L^2 T}{2} + \frac{\frac{1}{\pi(k)} - 1}{\eta} \right\}$$

where

$$L^2 = 2C^2 + 2\sum_{k=1}^{\infty} \pi(k)k^2.$$

The choice $\pi(k) \propto 1/k^4$ and $\eta \propto 1/\sqrt{T}$ leads to a bound in:

$$\sum_{t=1}^{T} \mathbb{E}_{k\sim\rho^t}\mathbb{E}_{\theta\sim\rho_k^t}[\ell_t(\theta)] \leq \inf_{m\in\mathbb{R}} \left\{ \sum_{t=1}^{T} (y_t - m)^2 \right.$$
$$\left. + (1 + m^4 + C^2)\sqrt{T\log(T)}(1 + o(1)) \right\}$$

which improves on (70).

B. Proofs of the Results in Section 4

Proof of Proposition 4.1: It is a direct application of Lemma 5.1 to $f = F$. □

Proof of Theorem 4.2: This proof follows step by step the classical analysis of FTRL, but we provide it for the sake of completeness. For short, let $\bar{L}_t(\mu) := \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$. First, by assumption, $\bar{L}_t$ is convex. By definition of the subgradient of a convex function,

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_{\theta\sim q_{\mu_t}}[\ell_t(\theta)] - \sum_{t=1}^{T}\mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)] \\
&\quad = \sum_{t=1}^{T}\bar{L}_t(\mu_t) - \sum_{t=1}^{t}\bar{L}_t(\mu) \\
&\quad \leq \sum_{t=1}^{T}\mu_t^T\nabla\bar{L}_t(\mu_t) - \sum_{t=1}^{T}\mu^T\nabla\bar{L}_t(\mu_t). \qquad (79)
\end{aligned}$$

Then, we prove by recursion on T that for any $\mu \in \mathbb{R}^d$,

$$\begin{aligned}
&\sum_{t=1}^{T}\mu_t^T\nabla\bar{L}_t(\mu_t) - \sum_{t=1}^{T}\mu^T\nabla\bar{L}_t(\mu_t) \\
&\quad \leq \sum_{t=1}^{T}\mu_t^T\nabla\bar{L}_t(\mu_t) - \sum_{t=1}^{T}\mu_{t+1}^T\nabla\bar{L}_t(\mu_t) \\
&\qquad\qquad + \frac{D_\phi(q_\mu||\pi)}{\eta} \qquad (80)
\end{aligned}$$

which is exactly equivalent to

$$\sum_{t=1}^{T}\mu_{t+1}^T\nabla\bar{L}_t(\mu_t) \leq \sum_{t=1}^{T}\mu^T\nabla\bar{L}_t(\mu_t) + \frac{D_\phi(q_\mu||\pi)}{\eta}. \quad (81)$$

Indeed, for $T = 0$, (81) just states that $D_\phi(q_\mu||\pi) \geq 0$ which is true by assumption. Assume that (81) holds for some integer $T-1$. We then have, for all $\mu \in \mathbb{R}^d$,

$$\begin{aligned}
&\sum_{t=1}^{T}\mu_{t+1}^T\nabla\bar{L}_t(\mu_t) \\
&\quad = \sum_{t=1}^{T-1}\mu_{t+1}^T\nabla\bar{L}_t(\mu_t) + \mu_{T+1}^T\nabla\bar{L}_T(\mu_T) \\
&\quad \leq \sum_{t=1}^{T-1}\mu^T\nabla\bar{L}_t(\mu_t) + \frac{D_\phi(q_\mu||\pi)}{\eta} + \mu_{T+1}^T\nabla\bar{L}_T(\mu_T)
\end{aligned}$$

as (81) holds for $T-1$. Apply this to $\mu = \mu_{T+1}$ to get

$$\begin{aligned}
&\sum_{t=1}^{T}\mu_{t+1}^T\nabla\bar{L}_t(\mu_t) \\
&\quad \leq \sum_{t=1}^{T}\mu_{T+1}^T\nabla\bar{L}_t(\mu_t) + \frac{D(q_{\mu_{T+1}}||\pi)}{\eta} \\
&\quad = \min_{m\in\mathbb{R}^d}\left[\sum_{t=1}^{T}m^T\nabla\bar{L}_t(\mu_t) + \frac{D(q_m||\pi)}{\eta}\right] \\
&\qquad \text{(by definition of } \mu_{T+1}\text{)}, \\
&\quad \leq \sum_{t=1}^{T}\mu^T\nabla\bar{L}_t(\mu_t) + \frac{D_\phi(q_\mu||\pi)}{\eta}
\end{aligned}$$

for all $\mu \in \mathbb{R}^d$. Thus, (81) holds for T. Thus, by recursion, (81) and (80) hold for all $T \in \mathbb{N}$.

The last step is to prove that for any $t \in \mathbb{N}$,

$$\mu_t^T\nabla\bar{L}_t(\mu_t) - \mu_{t+1}^T\nabla\bar{L}_t(\mu_t) \leq \frac{\eta L^2}{\alpha}. \qquad (82)$$

Indeed,

$$\begin{aligned}
\mu_t^T\nabla\bar{L}_t(\mu_t) &- \mu_{t+1}^T\nabla\bar{L}_t(\mu_t) \\
&= (\mu_t - \mu_{t+1})^T\nabla\bar{L}_t(\mu_t) \\
&\leq \|\mu_t - \mu_{t+1}\|\|\nabla\bar{L}_t(\mu_t)\|^* \\
&\leq L\|\mu_t - \mu_{t+1}\| \qquad (83)
\end{aligned}$$

as $\bar{L}_t$ is L Lipschitz w.r.t $\|\cdot\|$ (Lemma 2.6 page 27 in (Shalev-Shwartz, 2012) states that the conjugate norm of its gradient is bounded by L). Define

$$G_t(\mu) = \sum_{i=1}^{t-1}\mu^T\nabla\bar{L}_i(\mu_i) + \frac{D_\phi(q_\mu||\pi)}{\eta}.$$

We remind that by assumption, $\mu \mapsto D_\phi(q_\mu||\pi)/\eta$ is α/η-strongly convex with respect to $\|\cdot\|$. As the sum of a linear function and an α/η-strongly convex function, G_t is α/η-strongly convex. So, for any (μ, μ'),

$$G_t(\mu') - G_t(\mu) \geq (\mu' - \mu)^T\nabla G_t(\mu) + \frac{\alpha\|\mu' - \mu\|^2}{2\eta}.$$

As a special case, using the fact that μ_t is a minimizer of G_t, we have

$$G_t(\mu_{t+1}) - G_t(\mu_t) \geq \frac{\alpha \|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

In the same way,

$$G_{t+1}(\mu_t) - G_{t+1}(\mu_{t+1}) \geq \frac{\alpha \|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

Summing the two previous inequalities gives

$$\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \geq \frac{\alpha \|\mu_{t+1} - \mu_t\|^2}{\eta},$$

and so, combined with (83), this gives:

$$\|\mu_{t+1} - \mu_t\| \leq \sqrt{\frac{\eta}{\alpha} \left[\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t)\right]}.$$

Combining this inequality with (83) leads to (82).

Plugging (79), (80) and (82) together gives

$$\sum_{t=1}^T \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] - \sum_{t=1}^T \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)] \\ \leq \frac{\eta T L^2}{\alpha} + \frac{D_\phi(q_\mu \| \pi)}{\eta},$$

that is the statement of the theorem. □

C. Derivation of ρ^t in the Finite Case via Lagrange Method of Multipliers

Following Remark 3.1, we provide the derivation of ρ^t in the finite case, thanks to Lagrange method of multipliers. We remind that

$$\mathcal{L}(\rho_1^t, \ldots, \rho_M^t, \lambda, \nu_1, \ldots, \nu_M) = \sum_{j=1}^M \rho_j^t \sum_{s=1}^{t-1} \ell_s(\theta_j) \\ + \frac{\sum_{j=1}^M \pi_j \phi\left(\frac{\rho_j^t}{\pi_j}\right)}{\eta} + \lambda \frac{1 - \sum_{j=1}^M \rho_j^t}{\eta} + \sum_{j=1}^M \nu_j \rho_j^t. \quad (84)$$

So:

$$\frac{\partial}{\partial \rho_j^t} \mathcal{L}(\rho_1^t, \ldots, \rho_M^t, \lambda, \nu_1, \ldots, \nu_M) = \sum_{s=1}^{t-1} \ell_s(\theta_j) \\ + \frac{\phi'\left(\frac{\rho_j^t}{\pi_j}\right)}{\eta} + \frac{-\lambda}{\eta} + \nu_j. \quad (85)$$

Thus the first-order equation

$$\frac{\partial}{\partial \rho_j^t} \mathcal{L}(\rho_1^t, \ldots, \rho_M^t, \lambda, \nu_1, \ldots, \nu_M) = 0 \quad (86)$$

is equivalent to

$$\phi'\left(\frac{\rho_j^t}{\pi_j}\right) = \lambda - \eta \sum_{s=1}^{t-1} \ell_s(\theta_j) - \eta \nu_j. \quad (87)$$

Intuitively, the next step would be to apply the inverse of the function ϕ':

$$\frac{\rho_j^t}{\pi_j} = (\phi')^{-1} \left(\lambda - \eta \sum_{s=1}^{t-1} \ell_s(\theta_j) - \eta \nu_j\right). \quad (88)$$

Remind that the first order condition for ν_j is $\nu_j \geq 0$ and $\nu_j > 0 \Leftrightarrow \rho_j = 0$. So, we would obtain the simpler formula:

$$\rho_j^t = \pi_j \max\left\{0, (\phi')^{-1}\left(\lambda - \eta \sum_{s=1}^{t-1} \ell_s(\theta_j)\right)\right\}. \quad (89)$$

It turns out that, under the assumptions of Proposition 3.1, $(\phi')^{-1}$ indeed exists and $\nabla \tilde{\phi}^*(y) = \max\{0, (\phi')^{-1}(y)\}$. That is, (89) is equivalent to (29).