## A Proofs

*Proof of Theorem 3.1.* It is enough to show that the EWA strategy leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t\sim\pi_t}[\hat{L}_t(\hat{g}_t)] \leq \inf_{\rho}\left\{ \mathbb{E}_{g\sim\rho}\left[\sum_{t=1}^{T}\hat{L}_t(g)\right] + \frac{\eta C^2 T}{8} + \frac{\mathcal{K}(\rho,\pi_1)}{\eta}\right\}. \tag{A.1}$$

Once this is done, we only have to use the assumption that the regret of the within-task algorithm on task $t$ is upper bounded by $\beta(g,m_t)$ to obtain that

$$\sum_{t=1}^{T}\hat{L}_t(g) = \sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\big(h^{g}_{t,i}\circ g(x_{t,i}),y_{t,i}\big) \leq \sum_{t=1}^{T}\left\{\beta(g,m_t) + \inf_{h\in\mathcal{H}}\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\big(h\circ g(x_{t,i}),y_{t,i}\big)\right\}$$

and we obtain the statement of the result.

It remains to prove (A.1). To this end, we follows the same guidelines as in the proof of Theorem 1 in (Audibert, 2006). First, note that

$$\pi_t(g) = \frac{\exp\left[-\eta\sum_{u=1}^{t-1}\hat{L}_u(g)\right]\pi_1(\mathrm{d}g)}{\int\exp\left[-\eta\sum_{u=1}^{t-1}\hat{L}_u(\gamma)\right]\pi_1(\mathrm{d}\gamma)} = \frac{\exp\left[-\eta\sum_{u=1}^{t-1}\hat{L}_u(g)\right]\pi_1(\mathrm{d}g)}{W_t} \tag{A.2}$$

where we introduce the notation $W_t$ for the sake of shortness. Put $E_t = \int \hat{L}_t(g)\pi_t(\mathrm{d}g) = \mathbb{E}_{\hat{g}_t\sim\pi_t}[\hat{L}_t(g)]$. Using Hoeffding's inequality on the bounded random variable $\hat{L}_t(g)\in[0,C]$ we have, for any $t$, that

$$\mathbb{E}_{\hat{g}_t\sim\pi_t}\left[\exp\left\{\eta(E_t-\hat{L}_t(g))\right\}\right] = \int\exp\left\{\eta(E_t-\hat{L}_t(g))\right\}\pi_t(\mathrm{d}g) \leq \exp\left\{\frac{C^2\eta^2}{8}\right\}$$

which can be rewritten as

$$\exp\left\{-\eta\mathbb{E}_{g_t\sim\pi_t}[\hat{L}_t(g_t)]\right\} \geq \exp\left(-\frac{C^2\eta^2}{8}\right)\mathbb{E}_{\hat{g}_t\sim\pi_t}\left\{\exp\left[-\eta\hat{L}_t(g_t)\right]\right\}. \tag{A.3}$$

Next, we note that

$$\begin{aligned}
\exp\left\{-\eta\sum_{t=1}^{T}\mathbb{E}_{\hat{g}_t\sim\pi_t}[\hat{L}_t(g_t)]\right\} &= \prod_{t=1}^{T}\exp\left\{-\eta\mathbb{E}_{g_t\sim\pi_t}[\hat{L}_t(g_t)]\right\}\\
&\geq \exp\left(-\frac{TC^2\eta^2}{8}\right)\prod_{t=1}^{T}\mathbb{E}_{\hat{g}_t\sim\pi_t}\left\{\exp\left[-\eta\hat{L}_t(g_t)\right]\right\}, \text{ using (A.3)}\\
&= \exp\left\{-\frac{TC^2\eta^2}{8}\right\}\prod_{t=1}^{T}\int\exp\left\{-\eta\hat{L}_t(g)\right\}\pi_t(\mathrm{d}g)\\
&= \exp\left\{-\frac{TC^2\eta^2}{8}\right\}\prod_{t=1}^{T}\int\frac{\exp\left\{-\eta\sum_{u=1}^{t}\hat{L}_u(g)\right\}}{W_t}\pi_1(\mathrm{d}g), \text{ using (A.2)}\\
&= \exp\left\{-\frac{TC^2\eta^2}{8}\right\}\prod_{T=1}^{T}\frac{W_{t+1}}{W_t} = \exp\left\{\frac{TC^2\eta^2}{8}\right\}W_{T+1}.
\end{aligned}$$

So

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}_{\hat{g}_t\sim\pi_t}[\hat{L}_t(g_t)] &\leq -\frac{\log W_{T+1}}{\eta} + \frac{TC^2\eta}{8}\\
&= -\frac{\log\int\exp\left[-\eta\sum_{t=1}^{T}\hat{L}_t(g)\right]\pi_1(\mathrm{d}g)}{\eta} + \frac{TC^2\eta}{8}
\end{aligned}$$

and finally we use (Catoni, 2004, Equation (5.2.1)) which states that

$$-\frac{\log \int \exp\left[-\eta \sum_{t=1}^{T} \hat{L}_t(g)\right] \pi_1(\mathrm{d}g)}{\eta} = \inf_{\rho} \left\{ \mathbb{E}_{g\sim\rho}\left[\sum_{t=1}^{T} \hat{L}_t(g)\right] + \frac{\mathcal{K}(\rho,\pi_1)}{\eta} \right\}.$$

□

*Proof of Theorem 4.3.* Let $D^*$ denote a minimizer to the optimization problem

$$\min_{D\in\mathcal{D}_K} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}).$$

We apply Theorem 3.1 and upper bound the infimum with respect to any $\rho$ by an infimum with respect to $\rho$ in the following parametric family

$$\rho_c(\mathrm{d}D) \propto \mathbf{1}\{\forall j = 1,\dots,K : \|D_{\cdot,j} - D^*_{\cdot,j}\| \leq c\}\pi_1(dD).$$

where $c$ is a positive parameter. Note that when $c$ is small, $\rho_c$ highly concentrates around $D^*$, but we will show this is at a price of an increase in $\mathcal{K}(\rho_c,\pi_1)$. The proof then proceeds in optimizing with respect to $c$.

We have that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\hat{g}_t\sim\pi_t}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i}\right]$$
$$\leq \inf_{c}\left\{\mathbb{E}_{D\sim\rho_c}\left[\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}) + \beta(m)\right] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho_c,\pi_1)}{\eta T}\right\}.$$

Now, we have

$$\mathcal{K}(\rho_c,\pi_1) = -\log \pi_1(\{\forall j = 1,\dots,K : \|D_{\cdot,j} - D^*_{\cdot,j}\| \leq c\}),$$

and

$$\pi_1(\{\forall j = 1,\dots,K : \|D_{\cdot,j} - D^*_{\cdot,j}\| \leq c\}) \geq \prod_{j=1}^{K}\left(\frac{\pi^{(d-1)/2}(c/2)^{d-1}}{\Gamma(\frac{d-1}{2}+1)} \Bigg/ \frac{2\pi^{(d+1)/2}}{\Gamma(\frac{d+1}{2})}\right) \geq \prod_{j=1}^{K}\left(\frac{c^{d-1}}{2^d\pi}\right)$$

where the first inequality follows by observing that, since $\pi_1$ is the uniform distribution on the unit $d$-sphere, the probability to be calculated is greater or equal to the ration between the volume of the $(d-1)$-ball with radius $c/2$ and the surface area of the unit $d$-sphere. So we get

$$\mathcal{K}(\rho_c,\pi_1) \leq Kd\log(1/c) + 3Kd.$$

Furthermore, using the notation

$$h_t^* := \arg\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t, D^*x_{t,i}\rangle, y_{t,i}\big),$$

we get

$$\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t, Dx_{t,i}\rangle, y_{t,i}\big) - \frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t^*, D^*x_{t,i}\rangle, y_{t,i}\big) \leq \frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t^*, Dx_{t,i}\rangle, y_{t,i}\big) - \frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t^*, D^*x_{t,i}\rangle, y_{t,i}\big).$$

Under the condition on the loss, we have

$$\Big|\ell(\langle h_t^*, Dx_{t,i}\rangle, y_{t,i}) - \ell\big(\langle h_t^*, D^*x_{t,i}\rangle, y_{t,i}\big)\Big| \leq \Phi\,\Big|\,\langle h_t^*, (D-D^*)x_{t,i}\rangle\,\Big|.$$

We obtain an upper-bound

$$\begin{aligned}
&\mathbb{E}_{D\sim\rho_c}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}) \\
&\qquad\leq \inf_{D\in\mathcal{D}_K}\left\{\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}) + \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\Phi\,|\,\langle h_t^*, (D-D^*)x_{t,i}\rangle\,|\right\}.
\end{aligned}$$

But then note that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\Phi\,|\,\langle h_t^*, (D-D^*)x_{t,i}\rangle\,| &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\Phi\sqrt{\langle h_t^*, (D-D^*)x_{t,i}\rangle^2} \\
&\leq \Phi\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\langle h_t^*, (D-D^*)x_{t,i}\rangle^2 \text{ (Jensen)}} \\
&= \Phi\sqrt{\frac{1}{T}\sum_{t=1}^{T}(h_t^*)^T(D-D^*)\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)(D-D^*)^T h_t^*} \\
&\leq \Phi\sqrt{\frac{1}{T}\sum_{t=1}^{T}\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)\|(D-D^*)^T h_t^*\|^2} \\
&\leq \Phi cB\sqrt{\frac{1}{T}\sum_{t=1}^{T}\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)}.
\end{aligned}$$

So Theorem 3.1 leads to

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{g_t\sim\pi_t}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i}\right] - \inf_{D\in\mathcal{D}_K}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}) \\
&\qquad\leq \inf_{c}\left\{c\Phi B\sqrt{\frac{1}{T}\sum_{t=1}^{T}\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)} + \frac{Kd}{\eta T}\log(1/c)\right\} + \frac{3Kd}{\eta T} + \beta(m) + \frac{\eta C^2}{8}.
\end{aligned}$$

The choices $c = \sqrt{\frac{1}{T}}$ and $\eta = \frac{2}{C}\sqrt{\frac{Kd}{T}}$ lead to the result. □

*Proof of Theorem 6.1.* The proof relies on an application of the well-known online-to-batch trick, discussed pedagogically in Section 5 page 186 in Shalev-Shwartz (2011). Still, it is very cumbersome, and it is easy to get confused. For these reasons, we think it is important to write it completely. We use the following notation for any random variable $V$, $\mathbb{E}_V$ is the expectation with respect to $V$. This is very important as the online-to-batch trick relies essentially on inverting the order of the random variables in the integration. We have:

$$\begin{aligned}
&\mathbb{E}[\ell(\hat{h}\circ\hat{g}(x), y)] \\
&= \mathbb{E}_{\mathcal{T}}\mathbb{E}_{\mathcal{I}}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{1,1},y_{1,1}),(x_{1,2},y_{1,2}),\dots,(x_{T,m},y_{T,m})}\mathbb{E}_P\mathbb{E}_{(x_1,y_1),\dots,(x_m,y_m)}\mathbb{E}_{(x,y)}[\ell(\hat{h}\circ\hat{g}(x), y)] \\
&= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{T,m},y_{T,m})}\mathbb{E}_P\mathbb{E}_{(x_1,y_1),\dots,(x_m,y_m)}\mathbb{E}_{(x,y)}[\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x), y)] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{T,m},y_{T,m})}\mathbb{E}_P\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{(x_1,y_1),\dots,(x_{i-1},y_{i-1})}\mathbb{E}_{(x,y)}[\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x), y)] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{T,m},y_{T,m})}\mathbb{E}_P\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{(x_1,y_1),\dots,(x_{i-1},y_{i-1})}\mathbb{E}_{(x_i,y_i)}[\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_i), y_i)] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{T,m},y_{T,m})}\mathbb{E}_P\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{(x_1,y_1),\dots,(x_m,y_m)}[\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_i), y_i)]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{T,m},y_{T,m})}\mathbb{E}_{P}\mathbb{E}_{(x_1,y_1),\dots,(x_m,y_m)}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_i),y_i)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{P_1,\dots,P_{t-1}}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{t-1,m},y_{t-1,m})}\mathbb{E}_{P}\mathbb{E}_{(x_1,y_1),\dots,(x_m,y_m)}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_i),y_i)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{P_1,\dots,P_{t-1}}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{t-1,m},y_{t-1,m})}\mathbb{E}_{P_t}\mathbb{E}_{(x_{t,1},y_{t,1}),\dots,(x_{t,m},y_{t,m})}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_{t,i}),y_{t,i})\right] \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{t,m},y_{t,m})}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_{t,i}),y_{t,i})\right] \\
&= \mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{t,m},y_{t,m})}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_{t,i}),y_{t,i})\right] \\
&\leq \mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{1,1},y_{1,1}),\dots,(x_{T,m},y_{T,m})}\inf_{\rho}\Bigg\{\mathbb{E}_{g\sim\rho}\Bigg[\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(h_t\circ g(x_{t,i}),y_{t,i}) \\
&\qquad + \frac{1}{T}\sum_{t=1}^{T}\beta(g,m)\Bigg] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho,\pi_1)}{\eta T}\Bigg\}, \text{ using Theorem 3.1,} \\
&\leq \inf_{\rho}\left\{\mathbb{E}_{g\sim\rho}\left[\mathbb{E}_{P\sim Q}\inf_{h_t\in\mathcal{H}}\mathbb{E}_{(x,y)\sim P}\ell(h_t\circ g(x),y)+\beta(g,m)\right]+\frac{\eta C^2}{8}+\frac{\mathcal{K}(\rho,\pi_1)}{\eta T}\right\}.
\end{aligned}
$$

□

# B Better Bounds for Dictionary Learning

We now state a refined version of the bounds for dictionary learning in Section 4. As pointed out in that section, while in general the bound

$$\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)\leq 1$$

is unimprovable, if the input vectors $x_{t,i}$ are i.i.d. random variables from uniform distribution on the unit sphere, then

$$\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T \xrightarrow[m\to\infty]{a.s.} \mathrm{Cov}(x_{t,i},x_{t,i})=\frac{1}{d}I$$

where $I$ is the identity matrix. Consequently,

$$\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)\xrightarrow[m\to\infty]{a.s.}\frac{1}{d}.$$

We can take advantage of this fact in order to improve the term $\beta(m)=\sup_{g\in\mathcal{G}}\beta(g,m)$, but only if we assume that we know in advance that $\lambda_{\max}\left(\sum_{i=1}^{m}x_{t,i}x_{t,i}^T/m\right)$ is not too large. This is the meaning of the following theorem.

**Theorem B.1.** *Assume that we know in advance that for all $t\in\{1,\dots,T\}$,*

$$\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)\leq\Lambda$$

*for some $\Lambda>0$. Assume the same assumptions as in Theorem 4.3, still with $\eta=\frac{2}{C}\sqrt{\frac{Kd}{T}}$. Use within tasks Algorithm 2 (online gradient) with a fixed gradient step $\zeta=B/(L\sqrt{2mK\Lambda})$. Then we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{g_t\sim\pi_t}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i}\right]-\inf_{g\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t,gx_{t,i}\rangle,y_{t,i}\big)$$
$$\leq\frac{C}{4}\sqrt{\frac{Kd}{T}}(\log(T)+7)+\frac{2BL\sqrt{2K\Lambda}}{\sqrt{m}}+\frac{B\Phi\sqrt{\Lambda}}{\sqrt{T}}.$$

In particular, note that when $\Lambda=1/d$ the bound becomes

$$\frac{C}{4}\sqrt{\frac{Kd}{T}}(\log(T)+7)+\frac{2BL\sqrt{2K}}{\sqrt{md}}+\frac{B\Phi}{\sqrt{dT}}.$$

*Proof.* We apply Theorem 4.3, so we only have to upper bound the term $\beta(g,m)$ for the online gradient algorithm with the prescribed step size. Note that in (Corollary 2.7 Shalev-Shwartz, 2011) we actually have the following regret bound for Algorithm 2 with fixed step size $\eta>0$:

$$\beta(g,m)=\frac{B^2}{2\eta m}+\frac{\eta}{m}\sum_{i=1}^{m}\|\nabla_{\theta=\theta_t}\ell(\langle\theta,gx_{t,i}\rangle,y_{t,i})\|^2.$$

By the $L$-Lipschitz assumption on $\ell$, $\|\nabla_{\theta=\theta_t}\ell(\langle\theta_t,gx_{t,i}\rangle,y_{t,i})\|^2\leq L^2\|gx_{t,i}\|^2$. So we have

$$\sum_{t=1}^{m}\|\nabla_{\theta=\theta_t}\ell(\langle\theta,gx_{t,i}\rangle,y_{t,i})\|^2\leq L^2\sum_{i=1}^{m}\|gx_{t,i}\|^2=L^2\sum_{i=1}^{m}\sum_{k=1}^{K}\langle g_{k,\cdot},x_{t,i}\rangle^2=L^2\sum_{i=1}^{m}\sum_{k=1}^{K}g_{k,\cdot}^Tx_{t,i}x_{t,i}^Tg_{k\cdot}$$
$$\leq mL^2\sum_{k=1}^{K}g_{k,\cdot}^T\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)g_{k\cdot}\leq mKL^2\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)\|g_{k\cdot}\|^2$$
$$\leq mKL^2\Lambda.$$

Consequently, $\beta(m)=\sup_g\beta(g,m)\leq B^2/(2\eta m)+\eta KL^2\Lambda$ and The choice $\eta\leq B/(L\sqrt{2mK\Lambda})$ leads to

$$\beta(m)=2BL\sqrt{2K\Lambda/m}.$$

□

## C Batch-Within-Online Lifelong Learning

In this last section of the appendix, we present an alternative approach for the batch-within-online setting discussed in Section 2. In this setting, the tasks are presented sequentially, but, for each task $t\in\{1,\ldots,T\}$ the dataset $\mathcal{S}_t$ is presented all at once and we assume it is obtained i.i.d. from a distribution $P_t$. Unlike to the reasoning in Section 6, where we assumed that the $P_t$ were i.i.d. from a distribution $Q$, here we make no assumptions on the generation process underlying the $P_t$'s, which may even be adversarial chosen.

Let us recap the setting. At each time $t\in\{1,\ldots,T\}$, a task is presented to the learner in the following manner:

1. nature choses $P_t$ , no assumption is made on this choice. This $P_t$ is not revealed to the forecaster.

2. nature draws the sample $\mathcal{S}_t=\big((x_{t,1},y_{t,1}),\ldots,(x_{t,m_t},y_{t,m_t})\big]$ i.i.d. from $P_t$, and this sample is revealed to the forecaster.

3. based on her/his current guess $\tilde{g}_t$ of $g$ and on the sample $\mathcal{S}_t$, the forecaster has to run her/his favourite learning algorithm $\hat{h}$ on $(\tilde{g}_t,\mathcal{S}_t)$ to get an estimate $\tilde{h}_t=\hat{h}(\tilde{g}_t,\mathcal{S}_t)$ based on an algorithm of his choice. Note that the forecaster observes $\tilde{r}_t:=r_t(\tilde{h}_t\circ\tilde{g}_t)$ where

$$r_t(f)=\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\big(f(x_{t,i}),y_{t,i}\big).$$

4. the forecaster incur the loss $R_t(\tilde{h}_t \circ \tilde{g}_t)$ where
$$R_t(f) = \mathbb{E}_{(x,y)\sim P_t}\left[\ell\big(f(x),y\big)\right].$$
Unfortunately, this quantity is not known to the forecaster.

At the end of time, we are interested in a strategy such that the compound regret
$$\mathcal{R} := \frac{1}{T}\sum_{t=1}^{T} R_t(\tilde{h}_t \circ \tilde{g}_t) - \inf_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} R_t(h_t \circ g)$$
is controled. The situation is similar to the setting discussed in the core of the paper: we will propose an EWA algorithm for transfer learning, EWA-TL, for which the regret will be controlled, on the condition that the learner chooses a suitable within task algorithm. In the online case, the within tasks algorithm was either EWA or OGA. In Subsection C.1 we discuss briefly the within task algorithm. In Subsection C.2 we present the EWA-TL algorithm and its theoretical analysis.

### C.1 Within-task Algorithms

We make an additional assumption, that is that the estimator $\hat{h}$ satisfies a bound in probability:
$$\mathbb{P}\Bigg[\forall g \in \mathcal{G}, |r(\hat{h}(g,\mathcal{S}_t)\circ g) - R_t(\hat{h}(g,\mathcal{S}_t)\circ g)| \leq \delta(g,m_t,\varepsilon) \text{ and } |R_t(\hat{h}(g,\mathcal{S}_t)\circ g) - \inf_{h\in\mathcal{H}} R_t(h\circ g)| \leq 2\delta(g,m_t,\varepsilon)\Bigg] \geq 1-\varepsilon. \quad \text{(C.1)}$$

In classification, when $\ell$ is the 0-1 loss function, and for any $g$, the family $\{h\circ g, h\in\mathcal{H}\}$ has a Vapnik-Chervonenkis dimension bounded by $V$, then the empirical risk minimizer (ERM)
$$\hat{h}(g,\mathcal{S}_t) = \arg\min_{h\in\mathcal{H}} r_t(h\circ g)$$
satisfies the above condition with
$$\delta(g,m_t,\varepsilon) = 2\sqrt{2\frac{V\log\left(\frac{2m_t \mathrm{e}}{V}\right) + \log\left(\frac{4}{\varepsilon}\right)}{m_t}},$$
see e.g. (Chapter 4, page 94 Vapnik, 1998). Similar rates can be obtained with PAC-Bayesian bounds (McAllester, 1998; Catoni, 2004), but we postpone the details to future work.

### C.2 EWA-TL

```
Algorithm 6 EWA-TL
Data A sequence of datasets
    $\mathcal{S}_t = \big((x_{t,1},y_{t,1}),\dots,(x_{t,m_t},y_{t,m_t})\big)$, $1 \leq t \leq T$, associated with different learning tasks; the datasets are revealed sequentially, but the points within each dataset $\mathcal{S}_t$ are revealed all at once.
Input A prior $\pi_1$, a learning parameter $\eta > 0$ and a learning algorithm $\hat{h}$ which satisfies (C.1).
Loop For $t = 1,\dots,T$
    i Draw $\hat{g}_t \sim \pi_t$.
    ii Run the within-task learning algorithm $\hat{t}$ on $\mathcal{S}_t$ to get $\tilde{h}_t = \hat{h}(\hat{g}_t,\mathcal{S}_t)$.
    iii Update
        $\pi_{t+1}(\mathrm{d}g) \propto \exp\Big\{-\eta\Big[r_t(\hat{h}(\mathcal{S}_t,g)\circ g) + \delta(g,m_t,\varepsilon/T)\Big]\Big\}\pi_{t-1}(\mathrm{d}g).$
```

We now provide a bound on the regret of EWA-TL.

**Theorem C.1.** *Under* (C.1)*, and assuming that there is a constant $C$ such that $0 \leq r_t(\hat{h}(\mathcal{S}_t, g) \circ g) + \delta(g, m_t, \varepsilon/T) \leq C$, with probability at least $1-\varepsilon$,*

$$\sum_{t=1}^{T} \mathbb{E}_{\tilde{g}_t \sim \pi_{t-1}} \left[ R_t(\tilde{h}_t \circ \tilde{g}_t)] \right] \leq \inf_{\rho} \left\{ \mathbb{E}_{g\sim\rho} \left[ \frac{1}{T} \sum_{t=1}^{T} \inf_{h\in\mathcal{H}} R_t(h \circ g) + \frac{4}{T} \sum_{t=1}^{T} \delta(g, m_t, \varepsilon/T) \right] + \frac{\eta T C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta} \right\}.$$

*Sketch of the proof.* First, follow the proof of Theorem 3.1 to get:

$$\sum_{t=1}^{T} \mathbb{E}_{\tilde{g}_t \sim \pi_{t-1}} \left[ r_t(\tilde{h}_t \circ \tilde{g}_t)] + \delta(\tilde{g}_t, m_t, \varepsilon/T) \right] \leq \inf_{\rho} \left\{ \sum_{t=1}^{T} \mathbb{E}_{g\sim\rho} \left[ r_t(\tilde{h}_t \circ g) + \delta(g, m_t, \varepsilon/T) \right] + \frac{\eta T C^2}{8} + \frac{\mathcal{K}(\rho, \pi)}{\eta} \right\}.$$

So, with probability at least $1-\varepsilon$,

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}_{\tilde{g}_t \sim \pi_{t-1}} \left[ R_t(\tilde{h}_t \circ \tilde{g}_t)] \right] &\leq \sum_{t=1}^{T} \mathbb{E}_{\tilde{g}_t \sim \pi_{t-1}} \left[ r_t(\tilde{h}_t \circ \tilde{g}_t)] + \delta(\tilde{g}_t, m_t, \varepsilon/T) \right] \\ &\leq \inf_{\rho} \left\{ \sum_{t=1}^{T} \mathbb{E}_{g\sim\rho} \left[ r_t(\tilde{h}_t \circ g) + \delta(g, m_t, \varepsilon/T) \right] + \frac{\eta T C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta} \right\} \\ &\leq \inf_{\rho} \left\{ \sum_{t=1}^{T} \mathbb{E}_{g\sim\rho} \left[ R_t(\hat{h}_t(g, \mathcal{S}_t) \circ g) + 2\delta(g, m_t, \varepsilon/T) \right] + \frac{\eta T C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta} \right\} \\ &\leq \inf_{\rho} \left\{ \mathbb{E}_{g\sim\rho} \left[ \sum_{t=1}^{T} \inf_{h\in\mathcal{H}} R_t(h \circ g) + 4 \sum_{t=1}^{T} \delta(g, m_t, \varepsilon/T) \right] + \frac{\eta T C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta} \right\}. \end{aligned}$$

□