A. Background and Results for Online Convex Optimization

Throughout the appendix we assume all subsets are convex and in $\mathbb{R}^d$ unless explicitly stated. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$, which we assume to be any norm on $\mathbb{R}^d$, and note that the dual norm of $\|\cdot\|_2$ is itself. For sequences of scalars $\sigma_1, \dots, \sigma_T \in \mathbb{R}$ we will use the notation $\sigma_{1:t}$ to refer to the sum of the first t of them. In the online learning setting, we will use the shorthand ∇_t to denote the subgradient of $\ell_t : \Theta \mapsto \mathbb{R}$ evaluated at action $\theta_t \in \Theta$. We will use $\mathrm{Conv}(S)$ to refer to the convex hull of a set of points S and $\mathrm{Proj}_S(\cdot)$ to be the projection to any convex subset $S \subset \mathbb{R}^d$.

A.1. Convex Functions

We first state the related definitions of *strong convexity* and *strong smoothness*:

Definition A.1. *An everywhere sub-differentiable function $f : S \mapsto \mathbb{R}$ is α-**strongly-convex** w.r.t. norm $\|\cdot\|$ if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2 \;\forall\, x, y \in S$$

Definition A.2. *An everywhere sub-differentiable function $f : S \mapsto \mathbb{R}$ is β-**strongly-smooth** w.r.t. norm $\|\cdot\|$ if*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y - x\|^2 \;\forall\, x, y \in S$$

We now turn to the *Bregman divergence* and a discussion of several useful properties (Bregman, 1967; Banerjee et al., 2005):

Definition A.3. *Let $f : S \mapsto \mathbb{R}$ be an everywhere sub-differentiable strictly convex function. Its **Bregman divergence** is defined as*

$$\mathcal{B}_f(x||y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle$$

The definition directly implies that $\mathcal{B}_f(\cdot||y)$ preserves the (strong or strict) convexity of f for any fixed $y \in S$. Strict convexity further implies $\mathcal{B}_f(x||y) \geq 0 \;\forall\, x, y \in S$, with equality iff $x = y$. Finally, if f is α-strongly-convex, or β-strongly-smooth, w.r.t. $\|\cdot\|$ then Definition A.1 implies $\mathcal{B}_f(x||y) \geq \frac{\alpha}{2}\|x - y\|^2$, or $\mathcal{B}_f(x||y) \leq \frac{\beta}{2}\|x - y\|^2$, respectively.

Claim A.1. *Let $f : S \mapsto \mathbb{R}$ be a strictly convex function on S, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ be a sequence satisfying $\alpha_{1:n} > 0$, and $x_1, \dots, x_n \in S$. Then*

$$\bar{x} = \frac{1}{\alpha_{1:n}}\sum_{i=1}^n \alpha_i x_i = \operatorname*{arg\,min}_{y \in S} \sum_{i=1}^n \alpha_i \mathcal{B}_f(x_i||y)$$

Proof. $\forall\, y \in S$ we have

$$\begin{aligned}
\sum_{i=1}^n \alpha_i \left(\mathcal{B}_f(x_i||y) - \mathcal{B}_f(x_i||\bar{x})\right) &= \sum_{i=1}^n \alpha_i \left(f(x_i) - f(y) - \langle \nabla f(y), x_i - y\rangle - f(x_i) + f(\bar{x}) + \langle \nabla f(\bar{x}), x_i - \bar{x}\rangle\right) \\
&= \left(f(\bar{x}) - f(y) + \langle \nabla f(y), y\rangle\right)\alpha_{1:n} + \sum_{i=1}^n \alpha_i \left(-\langle \nabla f(\bar{x}), \bar{x}\rangle + \langle \nabla f(\bar{x}) - \nabla f(y), x_i\rangle\right) \\
&= \left(f(\bar{x}) - f(y) - \langle \nabla f(y), \bar{x} - y\rangle\right)\alpha_{1:n} \\
&= \alpha_{1:n}\mathcal{B}_f(\bar{x}||y)
\end{aligned}$$

By Definition A.3 the last expression has a unique minimum at $y = \bar{x}$. □

A.2. Standard Online Algorithms

Here we provide a review of the online algorithms we use. Recall that in this setting our goal is minimizing regret:

Definition A.4. *The* **regret** *of an agent playing actions* $\{\theta_t \in \Theta\}_{t\in[T]}$ *on a sequence of loss functions* $\{\ell_t : \Theta \mapsto \mathbb{R}\}_{t\in[T]}$ *is*

$$\mathbf{R}_T = \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta)$$

Within-task our focus is on two closely related meta-algorithms, Follow-the-Regularized-Leader (FTRL) and (linearized lazy) Online Mirror Descent (OMD).

Definition A.5. *Given a strictly convex function* $R : \Theta \mapsto \mathbb{R}$*, starting point* $\phi \in \Theta$*, fixed learning rate* $\eta > 0$*, and a sequence of functions* $\{\ell_t : \Theta \mapsto \mathbb{R}\}_{t\geq 1}$*,* **Follow-the-Regularized Leader (**$\mathrm{FTRL}_{\phi,\eta}^{(R)}$**)** *plays*

$$\theta_t = \arg\min_{\theta\in\Theta} \mathcal{B}_R(\theta||\phi) + \eta \sum_{s<t} \ell_s(\theta)$$

Definition A.6. *Given a strictly convex function* $R : \Theta \mapsto \mathbb{R}$*, starting point* $\phi \in \Theta$*, fixed learning rate* $\eta > 0$*, and a sequence of functions* $\{\ell_t : \Theta \mapsto \mathbb{R}\}_{t\geq 1}$*,* **lazy linearized Online Mirror Descent (**$\mathrm{OMD}_{\phi,\eta}^{(R)}$**)** *plays*

$$\theta_t = \arg\min_{\theta\in\Theta} \mathcal{B}_R(\theta||\phi) + \eta \sum_{s<t} \langle \nabla_s, \theta\rangle$$

These formulations make the connection between the two algorithms – their equivalence in the linear case $\ell_s(\cdot) = \langle \nabla_s, \cdot\rangle$ – very explicit. There exists a more standard formulation of OMD that is used to highlight its generalization of OGD – the case of $R(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ – and the fact that the update is carried out in the dual space induced by R (Hazan, 2015, Section 5.3). However, we will only need the following regret bound satisfied by both (Shalev-Shwartz, 2011, Theorems 2.11 and 2.15)

Theorem A.1. *Let* $\{\ell_t : \Theta \mapsto \mathbb{R}\}_{t\in[T]}$ *be a sequence of convex functions that are* G_t*-Lipschitz w.r.t.* $\|\cdot\|$ *and let* $R : S \mapsto \mathbb{R}$ *be 1-strongly-convex. Then the regret of both* $\mathrm{FTRL}_{\eta,\phi}^{(R)}$ *and* $\mathrm{OMD}_{\eta,\phi}^{(R)}$ *is bounded by*

$$\mathbf{R}_T \leq \frac{\mathcal{B}_R(\theta^*||\phi)}{\eta} + \eta G^2 T$$

for all $\theta^* \in \Theta$ *and* $G^2 \geq \frac{1}{T}\sum_{t=1}^T G_t^2$*.*

We next review the online algorithms we use for the meta-update. The main requirement here is logarithmic regret guarantees for the case of strongly convex loss functions, which is satisfied by two well-known algorithms:

Definition A.7. *Given a sequence of strictly convex functions* $\{\ell_t : \Theta \mapsto \mathbb{R}\}_{t\geq 1}$*,* **Follow-the-Leader (FTL)** *plays arbitrary* $\theta_1 \in \Theta$ *and for* $t > 1$ *plays*

$$\theta_t = \arg\min_{\theta\in\Theta} \sum_{s<t} \ell_s(\theta)$$

Definition A.8. *Given a sequence of functions* $\{\ell_t : \Theta \mapsto \mathbb{R}\}_{t\geq 1}$ *that are* α_t*-strongly-convex w.r.t.* $\|\cdot\|_2$*,* **Adaptive OGD (AOGD)** *plays arbitrary* $\theta_1 \in \Theta$ *and for* $t > 1$ *plays*

$$\theta_{t+1} = \mathrm{Proj}_\Theta\left(\theta_t - \frac{1}{\alpha_{1:t}}\nabla f(\theta_t)\right)$$

Kakade & Shalev-Shwartz (2008, Theorem 2) and Bartlett et al. (2008, Theorem 2.1) provide for FTL and AOGD, respectively, the following regret bound:

Theorem A.2. *Let* $\{\ell_t : \Theta \mapsto \mathbb{R}\}_{t\in[T]}$ *be a sequence of convex functions that are* G_t*-Lipschitz and* α_t*-strongly-convex w.r.t.* $\|\cdot\|$*. Then the regret of both FTL and AOGD is bounded by*

$$\mathbf{R}_T \leq \frac{1}{2}\sum_{t=1}^T \frac{G_t^2}{\alpha_{1:t}}$$

One further useful fact about FTL and AOGD is that when run on a sequence of Bregman regularizers $\mathcal{B}_R(\theta_1||\cdot),\dots,\mathcal{B}_R(\theta_T||\cdot)$ they will play points in the convex hull $\operatorname{Conv}(\{\theta_t\}_{t\in[T]})$:

Claim A.2. *Let $R:\Theta\mapsto\mathbb{R}$ be 1-strongly-convex w.r.t. $\|\cdot\|$ and consider any $\theta_1,\dots,\theta_T\in\Theta^*$ for some convex subset $\Theta^*\subset\Theta$. Then for loss sequence $\alpha_1\mathcal{B}_R(\theta_1||\cdot),\dots,\alpha_T\mathcal{B}_R(\theta_T||\cdot)$ for any positive scalars $\alpha_1,\dots,\alpha_T\in\mathbb{R}_+$, if we assume $\phi_1\in\Theta^*$ then FTL will play $\phi_t\in\Theta^*\ \forall\ t$ and AOGD will as well if we further assume $R(\cdot)=\frac{1}{2}\|\cdot\|^2$.*

Proof. The proof for FTL follows directly from Claim A.1 and the fact that the weighted average of a set of points is in their convex hull. For AOGD we proceed by induction on t. The base case $t=1$ holds by the assumption $\phi_t\in\Theta^*$. In the inductive case, note that $\mathcal{B}_R(\theta_t||\phi_t)=\frac{1}{2}\|\theta_t-\phi_t\|_2^2$ so the gradient update is $\phi_{t+1}=\phi_t+\frac{\alpha_t}{\alpha_{1:t}}(\theta_t-\phi_t)$, which is on the line segment between ϕ_t and θ_t, so the proof is complete by the convexity of $\Theta^*\ni\phi_t,\theta_t$. $\square$

A.3. Online-to-Batch Conversion

Finally, as we are also interested in distributional meta-learning, we discuss some techniques for converting regret guarantees into generalization bounds, which are usually named *online-to-batch conversions*. We state some standard results below:

Proposition A.1. *If a sequence of bounded convex loss functions $\{\ell_t:\Theta\mapsto\mathbb{R}\}_{t\in[T]}$ drawn i.i.d. from some distribution $\mathcal{D}$ is given to an online algorithm with regret bound $\mathbf{R}_T$ that generates a sequence of actions $\{\theta_t\in\Theta\}_{t\in[T]}$ then for $\bar\theta=\frac{1}{T}\theta_{1:T}$ and any $\theta^*\in\Theta$ we have*

$$\mathop{\mathbb{E}}_{\mathcal{D}^T}\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\bar\theta)\le\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\theta^*)+\frac{\mathbf{R}_T}{T}$$

Proof. Applying Jensen's inequality and using the fact that θ_t only depends on $\ell_1,\dots,\ell_{t-1}$ we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{D}^T}\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\bar\theta)\le\frac{1}{T}\mathop{\mathbb{E}}_{\mathcal{D}^T}\sum_{t=1}^T\mathop{\mathbb{E}}_{\ell_t'\sim\mathcal{D}}\ell_t'(\theta_t)&=\frac{1}{T}\mathop{\mathbb{E}}_{\{\ell_t\}\sim\mathcal{D}^T}\left(\sum_{t=1}^T\mathop{\mathbb{E}}_{\ell_t'\sim\mathcal{D}}\ell_t'(\theta_t)-\ell_t(\theta_t)\right)+\frac{1}{T}\mathop{\mathbb{E}}_{\{\ell_t\}\sim\mathcal{D}^T}\left(\sum_{t=1}^T\ell_t(\theta_t)\right)\\&\le\frac{1}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{\{\ell_s\}_{s<t}\sim\mathcal{D}^{t-1}}\left(\mathop{\mathbb{E}}_{\ell_t'\sim\mathcal{D}}\ell_t'(\theta_t)-\mathop{\mathbb{E}}_{\ell_t\sim\mathcal{D}}\ell_t(\theta_t)\right)+\frac{\mathbf{R}_T}{T}+\frac{1}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\theta^*)\\&=\frac{\mathbf{R}_T}{T}+\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\theta^*)\end{aligned}$$

$\square$

Proposition A.2. *If a sequence of loss functions $\{\ell_t:\Theta\mapsto[0,1]\}_{t\in[T]}$ drawn i.i.d. from some distribution $\mathcal{D}$ is given to an online algorithm that generates a sequence of actions $\{\theta_t\in\Theta\}_{t\in[T]}$ then the following inequalities each hold w.p. $1-\delta$:*

$$\frac{1}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\theta_t)\le\frac{1}{T}\sum_{t=1}^T\ell_t(\theta_t)+\sqrt{\frac{2}{T}\log\frac{1}{\delta}}\qquad\textit{and}\qquad\frac{1}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\theta_t)\ge\frac{1}{T}\sum_{t=1}^T\ell_t(\theta_t)-\sqrt{\frac{2}{T}\log\frac{1}{\delta}}$$

Note that Cesa-Bianchi et al. (2004) only prove the first inequality; the second follows via the same argument but applying the symmetric version of the Azuma-Hoeffding inequality (Azuma, 1967).

Corollary A.1. *If a sequence of loss functions $\{\ell_t:\Theta\mapsto[0,1]\}_{t\in[T]}$ drawn i.i.d. from some distribution $\mathcal{D}$ is given to an online algorithm with regret bound $\mathbf{R}_T$ that generates a sequence of actions $\{\theta_t\in\Theta\}_{t\in[T]}$ then*

$$\mathop{\mathbb{E}}_{t\sim\mathcal{U}[T]}\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\theta_t)\le\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\theta^*)+\frac{\mathbf{R}_T}{T}+\sqrt{\frac{8}{T}\log\frac{1}{\delta}}\qquad\textit{w.p. }1-\delta$$

for any $\theta^*\in\Theta$.

Proof. By Proposition A.2 we have

$$\frac{1}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\theta_t)\le\frac{1}{T}\sum_{t=1}^T\ell_t(\theta^*)+\frac{\mathbf{R}_T}{T}+\sqrt{\frac{2}{T}\log\frac{1}{\delta}}\le\mathop{\mathbb{E}}_{\ell\sim\mathcal{D}}\ell(\theta^*)+\frac{\mathbf{R}_T}{T}+\sqrt{\frac{8}{T}\log\frac{1}{\delta}}$$

$\square$

B. Proofs of Theoretical Results

In this section we prove the main guarantees on task-averaged regret for our algorithms, as, lower bounds showing that the results are tight up to constant factors, and online-to-batch conversion guarantees for statistical LTL. We first define some necessary definitions, notations, and general assumptions.

Setting B.1. *Using the data given to Algorithm 2 define the following quantities:*

- *convenience coefficients* $\sigma_t = G_t\sqrt{m_t}$
- *the sequence of update parameters* $\{\hat\theta_t \in \Theta\}_{t\in[T]}$ *with average update parameter* $\hat\phi = \frac{1}{\sigma_{1:T}}\sum_{t=1}^T \sigma_t\hat\theta_t$
- *a sequence of reference parameters* $\{\theta'_t \in \Theta\}_{t\in[T]}$ *with average reference parameter* $\phi' = \frac{1}{\sigma_{1:T}}\sum_{t=1}^T \sigma_t\theta'_t$
- *a sequence* $\{\theta^*_t \in \Theta\}_{t\in[T]}$ *of optimal parameters in hindsight*
- *we will say we are in the "Exact" case if* $\hat\theta_t = \theta'_t = \theta^*_t\ \forall\, t$ *and the "Approx" case otherwise*
- $\kappa \geq 1, \Delta^*_t \geq 0$ *s.t.* $\sum_{t=1}^T \alpha_t\mathcal{B}_R(\theta^*_t||\phi_t) \leq \sum_{t=1}^T \alpha_t\Delta^*_t + \kappa\sum_{t=1}^T \alpha_t\mathcal{B}_R(\hat\theta_t||\phi_t)$ *for some nonnegative* α_t
- $\nu \geq 1, \Delta' \geq 0$ *s.t.* $\sum_{t=1}^T \sigma_t\mathcal{B}_R(\hat\theta_t||\hat\phi) \leq \Delta' + \nu\sum_{t=1}^T \sigma_t\mathcal{B}_R(\theta'_t||\phi')$
- $\Delta_{\max} \geq 0$ *s.t.* $\frac{1}{2}\|\theta'_t - \hat\theta_t\|^2 \leq \Delta_{\max}\ \forall\, t \in [T]$
- *average deviation* $\bar D^2 = \frac{1}{\sigma_{1:T}}\sum_{t=1}^T \sigma_t\mathcal{B}_R(\theta'_t||\phi')$ *of the reference parameters; assumed positive*
- *task diameter* $D^* = \max_{\theta,\phi\in\mathrm{Conv}(\{\theta'_t\}_{t\in[T]})}\sqrt{\mathcal{B}_R(\theta||\phi)}$*; assumed positive*
- *action diameter* $D^2 = \max\{D^{*2}, \max_{\theta\in\Theta}\mathcal{B}_R(\theta||\phi_1)\}$ *in the Exact case or* $\max_{\theta,\phi\in\Theta}\mathcal{B}_R(\theta||\phi)$ *in the Approx case*
- *universal constant* C' *s.t.* $\|\theta\| \leq C'\|\theta\|_2\ \forall\,\theta\in\Theta$ *and* ℓ_2*-diameter* $D' = \max_{\theta,\phi}\|\theta-\phi\|_2$ *of* Θ
- *upper bound* G' *on the Lipschitz constants of the functions* $\{\mathcal{B}_R(\hat\theta_t||\cdot)\}_{t\in[T]}$ *over* $\mathrm{Conv}(\{\hat\theta_t\}_{t=1}^T)$
- *we will say we are in the "Nice" case if* $\mathcal{B}_R(\theta||\cdot)$ *is 1-strongly-convex and* β*-strongly-smooth w.r.t.* $\|\cdot\|\ \forall\,\theta\in\Theta$
- *in the general case* META *is FTL; in the Nice case* META *may instead be AOGD re-initialized at* θ^*_1
- *convenience indicator* $\iota = 1_{\mathrm{META}=\mathrm{FTL}}$
- *effective meta-action space* $\hat\Theta = \mathrm{Conv}(\{\hat\theta_t\}_{t\in[T]})$ *if* META *is FTL or* Θ *if* META *is AOGD*
- $\mathrm{TASK}_{\eta,\phi} = \mathrm{FTRL}^{(R)}_{\eta,\phi}$ *or* $\mathrm{OMD}^{(R)}_{\eta,\phi}$

We make the following assumptions:

- *the loss functions* $\ell_{t,i}$ *are convex* $\forall\, t, i$
- *at time* $t = 1$ *the update algorithm* META *plays* $\phi_1 \in \Theta$ *satisfying* $\max_{\theta\in\Theta}\mathcal{B}_R(\theta||\phi_1) < \infty$
- *in the Approx case* R *is* β*-strongly-smooth for some* $\beta \geq 1$

B.1. Upper Bound

We first prove a technical result on the performance of FTL on a sequence of Bregman regularizers. We start by lower bounding the regret of FTL when the loss functions are quadratic.

Lemma B.1. *For any $\theta_1,\dots,\theta_T\in S$ and positive scalars $\alpha_1,\dots,\alpha_T\in\mathbb{R}_+$ define $\phi_t=\frac{1}{\alpha_{1:t}}\sum_{s=1}^t\alpha_t\theta_t$ and let ϕ_0 be any point in S. Then*

$$\sum_{t=1}^T\alpha_t\|\theta_t-\phi_{t-1}\|_2^2-\sum_{t=1}^T\alpha_t\|\theta_t-\phi_T\|_2^2\geq 0$$

Proof. We proceed by induction on T. The base case $T=1$ follows directly since $\phi_1=\theta_1$ and so the second term is zero. In the inductive case we have

$$\sum_{t=1}^{T-1}\alpha_t\|\theta_t-\phi_{t-1}\|_2^2-\sum_{t=1}^{T-1}\alpha_t\|\theta_t-\phi_{T-1}\|_2^2\geq 0$$

so it suffices to show

$$\phi_{T-1}=\arg\min_{\theta_T}\sum_{t=1}^T\alpha_t\|\theta_t-\phi_{t-1}\|_2^2-\sum_{t=1}^T\alpha_t\|\theta_t-\phi_T\|_2^2$$

in which case $\phi_T=\phi_{T-1}$ and both added terms are zero, preserving the inequality. The gradient and Hessian are

$$2\alpha_T(\theta_T-\phi_{T-1})+\frac{2\alpha_T}{\alpha_{1:T}}\sum_{t=1}^{T-1}\alpha_t(\theta_t-\phi_T)-2\alpha_T(\theta_T-\phi_T)\left(1-\frac{\alpha_T}{\alpha_{1:T}}\right)$$

$$2\alpha_T\left(1-\frac{\alpha_T\alpha_{1:T-1}}{\alpha_{1:T}^2}-1+\frac{2\alpha_T}{\alpha_{1:T}}-\frac{\alpha_T^2}{\alpha_{1:T}^2}\right)I=\frac{2\alpha_T^2}{\alpha_{1:T}}I\succeq 0$$

so the problem is strongly convex and thus has a unique global minimum. Setting the gradient to zero yields

$$0=\theta_T-\phi_{T-1}+\frac{1}{\alpha_{1:T}}\sum_{t=1}^{T-1}\alpha_t\theta_t-\frac{1}{\alpha_{1:T}}\sum_{t=1}^{T-1}\alpha_t\phi_T-\theta_T+\frac{\alpha_T}{\alpha_{1:T}}\theta_T+\phi_T-\frac{\alpha_T}{\alpha_{1:T}}\phi_T=\phi_T-\phi_{T-1}\implies\theta_T=\phi_{T-1}$$

□

We use this to show logarithmic regret of FTL when the loss functions are Bregman regularizers with changing first arguments. Note that such functions are in general only strictly convex, so the bounds from Theorem A.2 cannot be applied directly.

Lemma B.2. *Let $\mathcal{B}_R$ be a Bregman regularizer on S w.r.t. $\|\cdot\|$ and consider any $\theta_1,\dots,\theta_T\in S$. Then for loss sequence $\alpha_1\mathcal{B}_R(\theta_1||\cdot),\dots,\alpha_T\mathcal{B}_R(\theta_T||\cdot)$ for any positive scalars $\alpha_1,\dots,\alpha_T\in\mathbb{R}_+$ we have regret bound*

$$\mathbf{R}_T\leq\frac{G_R^2+1}{2}\sum_{t=1}^T\frac{\alpha_t}{\alpha_{1:t}}$$

where G_R is the Lipschitz constant of the Bregman regularizer $\mathcal{B}_R(\theta_t||\cdot)$ for any $t\in[T]$ on S w.r.t. the Euclidean norm.

Proof. Defining $\bar\phi=\frac{1}{\alpha_{1:T}}\sum_{t=1}^T\alpha_t\theta_t$, we apply Claim A.1 and Lemma B.1 to get

$$\begin{aligned}\mathbf{R}_T&=\sum_{t=1}^T\alpha_t\mathcal{B}_R(\theta_t||\phi_t)-\min_{\phi\in S}\sum_{t=1}^T\alpha_t\mathcal{B}_R(\theta_t||\phi)\\&\leq\sum_{t=1}^T\alpha_t\mathcal{B}_R(\theta_t||\phi_t)-\sum_{t=1}^T\alpha_t\mathcal{B}_R(\theta_t||\bar\phi)+\frac12\sum_{t=1}^T\alpha_t\|\theta_t-\phi_t\|_2^2-\frac12\sum_{t=1}^T\alpha_t\|\theta_t-\bar\phi\|_2^2\\&=\sum_{t=1}^T\alpha_t\mathcal{B}_R(\theta_t||\phi_t)+\frac{\alpha_t}{2}\|\theta_t-\phi_t\|_2^2-\min_{\phi\in S}\sum_{t=1}^T\alpha_t\mathcal{B}_R(\theta_t||\phi)+\frac{\alpha_t}{2}\|\theta_t-\phi\|_2^2\end{aligned}$$

Since Bregman regularizers are convex in the second argument, the above is the regret of playing FTL on a sequence of a_t-strongly-convex losses. Applying Kakade & Shalev-Shwartz (2008, Theorem 2) yields the result. □

The following result is our main theorem; Theorems 2.1 and 3.1 will follow as corollaries.

Theorem B.1. *In Setting B.1, Algorithm 2 has TAR bounded as*

$$\bar{\mathbf{R}} \leq \frac{1}{T}\left((2\kappa D+\varepsilon)\sigma_1 + \frac{\kappa C}{\rho D^*}\sum_{t=1}^T \frac{\sigma_t^2}{\sigma_{1:t}} + \kappa\left(\frac{\nu \bar{D}^2}{\rho D^*} + \gamma(\rho D^* + \mathcal{E}) + \varepsilon\right)\sigma_{1:T}\right)$$
$$+ \frac{1}{T}\left(\frac{\Delta_{1:T}^*}{\varepsilon} + \frac{\kappa\Delta'}{\rho D^*} + \sum_{k=0}^{\lfloor \log_\gamma \frac{\rho D^*+\mathcal{E}}{\varepsilon}\rfloor}\left(\frac{\kappa(\rho D^*+\mathcal{E})}{\gamma^k\varepsilon} + \gamma^k\varepsilon\right)\sigma_{t_k}\right)$$

for $C = \frac{G'^2}{2}$ in the Nice case or otherwise $C = \frac{C'D'(G'+1)}{2}$, $\rho = 1$ in the Exact case or $\rho = 2\sqrt{\beta}$ in the Approx case, and $\mathcal{E} = 2\sqrt{2\beta\Delta_{\max}}$.

Proof. We first use the β-strong-smoothness of R to provide a bound in the Approx setting of the distance from the initialization to the update parameter at each time $t \in [T]$:

$$\begin{aligned}
\mathcal{B}_R(\hat{\theta}_t||\phi_t) \leq \frac{\beta}{2}\|\hat{\theta}_t - \phi_t\|^2 &\leq \beta\left(\|\hat{\theta}_t - \theta_t'\|^2 + \|\theta_t' - \phi_t\|^2\right)\\
&\leq \beta\left(\|\hat{\theta}_t - \theta_t'\|^2 + \max_{s<t} 2\|\theta_t' - \theta_s'\|^2 + 2\|\theta_s' - \hat{\theta}_s\|^2\right)\\
&\leq 4\beta D^{*2} + 4\beta \max_t \|\theta_t' - \hat{\theta}_t\|^2\\
&\leq 4\beta D^{*2} + 8\beta\Delta_{\max}
\end{aligned}$$

Combining this bound with the Exact setting assumption yields $\mathcal{B}_R(\hat{\theta}_t||\phi_t) \leq \rho^2 D^{*2} + 8\beta\Delta_{\max} \leq \rho^2 D^{*2} + \mathcal{E}^2\ \forall\, t \in [T]$. We now turn to analyzing the regret by defining two "cheating" sequences: $\tilde{\phi}_t = \phi_t$ on all t except $t = 1$, when we set $\tilde{\phi}_1 = \theta_1^*$; similarly, $\tilde{D}_t = D_t$ on all t except $t = 1$ and any t s.t. $\mathcal{B}_R(\hat{\theta}_t||\phi_t) > D_t^2$, when we set $\tilde{D}_t = \rho D^* + \mathcal{E}$. In order to do this we add outside of the summation the corresponding regret of the true sequences whenever one of them is not the same as its "cheating" sequence. Note that by this definition all upper bounds of $\mathcal{B}_R(\hat{\theta}_t||\phi_t)$ also upper bound $\mathcal{B}_R(\hat{\theta}_t||\tilde{\phi}_t)$. Furthermore the times t s.t. $\mathcal{B}_R(\theta_t^*||\phi_t) > D_t^2$ corresponds exactly to the times that the violation count k is incremented in Algorithm 2 and thus this occurs at most $\log_\gamma \frac{\rho D^*+\mathcal{E}}{\varepsilon}$ times, as we multiply the diameter guess by γ each time it happens, which together with Lemma A.2 ensures that ϕ_t remains within $\max\{\gamma(\rho D^*+\mathcal{E}),\varepsilon\}$ of all the reference parameters θ_t'. We index these times by $k = 0,\dots,$ so that at each k the agent uses η_{t_k} set using $\gamma^k\varepsilon$.

$$\begin{aligned}
\bar{\mathbf{R}}\,T &= \sum_{t=1}^T \frac{\mathcal{B}_R(\theta_t^*||\phi_t)}{\eta_t} + \eta_t G_t^2 m_t\\
&\leq \frac{\Delta_{1:T}^*}{\varepsilon} + \sum_{t=1}^T\left(\frac{\kappa\mathcal{B}_R(\hat{\theta}_t||\phi_t)}{D_t} + D_t\right)\sigma_t \qquad \text{(substitute } \eta_t = \frac{D_t}{G_t\sqrt{m_t}} \text{ and } D_t \geq \varepsilon)\\
&\leq \left(\frac{\kappa\mathcal{B}_R(\hat{\theta}_1||\phi_1)}{D_1} + D_1\right)\sigma_1 + \frac{\Delta_{1:T}^*}{\varepsilon} \qquad \text{(substitute cheating sequence)}\\
&\qquad + \sum_{t=1}^T\left(\frac{\kappa\mathcal{B}_R(\hat{\theta}_t||\tilde{\phi}_t)}{\tilde{D}_t} + \tilde{D}_t\right)\sigma_t + \sum_{k=0}^{\lfloor \log_\gamma \frac{\rho D^*+\mathcal{E}}{\varepsilon}\rfloor}\left(\frac{\kappa\mathcal{B}_R(\hat{\theta}_{t_k}||\phi_{t_k})}{\gamma^k\varepsilon} + \gamma^k\varepsilon\right)\sigma_{t_k}\\
&\leq ((\kappa+1)D+\varepsilon)\sigma_1 + \frac{\Delta_{1:T}^*}{\varepsilon} + \kappa\sum_{t=1}^T\left(\frac{\mathcal{B}_R(\hat{\theta}_t||\tilde{\phi}_t)}{\tilde{D}_t} + \tilde{D}_t\right)\sigma_t + \sum_{k=0}^{\lfloor \log_\gamma \frac{\rho D^*+\mathcal{E}}{\varepsilon}\rfloor}\left(\frac{\kappa(\rho D^*+\mathcal{E})}{\gamma^k\varepsilon} + \gamma^k\varepsilon\right)\sigma_{t_k}
\end{aligned}$$

We now bound the third term. For any $t \in [T]$ define $B_t^2 = \mathcal{B}_R(\hat{\theta}_t||\tilde{\phi}_t)$ and $f_t(x) = \frac{B_t^2}{x} + x$. Its derivative $\partial_x f_t = 1 - \frac{B_t^2}{x^2}$ is nonnegative on $x \geq B_t$. Thus when $\tilde{D}_t \leq \rho D^* + \mathcal{E}$ we have $f(\tilde{D}_t) \leq f(\rho D^* + \mathcal{E})$, as by definition both are greater than B_t and so f_t is increasing on the interval between them. On the other hand, for $\tilde{D}_t \geq \rho D^* + \mathcal{E}$, either $\tilde{D}_t \leq \gamma(\rho D^* + \mathcal{E})$ by the tuning rule or, if we initialized $\varepsilon > \rho D^* + \mathcal{E}$, then $\tilde{D}_t = \varepsilon \; \forall \; t \in [T]$, so either way we have $f_t(\tilde{D}_t) \leq \frac{B_t^2}{\rho D^*} + \max\{\gamma(\rho D^* + \mathcal{E}), \varepsilon\} \; \forall \; t \in [T]$. Since $\gamma > 1$ this bounds $f(\tilde{D}_t)$ in the previous case $\tilde{D}_t \leq \rho D^* + \mathcal{E}$ as well, so we have

$$
\begin{aligned}
\bar{\mathbf{R}} T &\leq ((1+\kappa)D + \varepsilon)\sigma_1 + \frac{\Delta_{1:T}^*}{\varepsilon} + \kappa \sum_{t=1}^T \left(\frac{\mathcal{B}_R(\hat{\theta}_t||\tilde{\phi}_t)}{\tilde{D}_t} + \tilde{D}_t \right) \sigma_t + \sum_{k=0}^{\lfloor \log_\gamma \frac{\rho D^* + \mathcal{E}}{\varepsilon} \rfloor} \left(\frac{\kappa(\rho D^* + \mathcal{E})}{\gamma^k \varepsilon} + \gamma^k \varepsilon \right) \sigma_{t_k} \\
&\leq (2\kappa D + \varepsilon)\sigma_1 + \frac{\Delta_{1:T}^*}{\varepsilon} + \kappa \sum_{t=1}^T \left(\frac{\mathcal{B}_R(\hat{\theta}_t||\tilde{\phi}_t)}{\rho D^*} + \gamma(\rho D^* + \mathcal{E}) + \varepsilon \right) \sigma_t + \sum_{k=0}^{\lfloor \log_\gamma \frac{\rho D^* + \mathcal{E}}{\varepsilon} \rfloor} \left(\frac{\kappa(\rho D^* + \mathcal{E})}{\gamma^k \varepsilon} + \gamma^k \varepsilon \right) \sigma_{t_k} \\
&\leq (2\kappa D + \varepsilon)\sigma_1 + \frac{\Delta_{1:T}^*}{\varepsilon} + \frac{\kappa}{\rho D^*} \sum_{t=1}^T \left(\mathcal{B}_R(\hat{\theta}_t||\tilde{\phi}_t) - \mathcal{B}_R(\hat{\theta}_t||\hat{\phi}) \right) \\
&\qquad + \kappa \sum_{t=1}^T \left(\frac{\mathcal{B}_R(\hat{\theta}_t||\hat{\phi})}{\rho D^*} + \gamma(\rho D^* + \mathcal{E}) + \varepsilon \right) \sigma_t + \sum_{k=0}^{\lfloor \log_\gamma \frac{\rho D^* + \mathcal{E}}{\varepsilon} \rfloor} \left(\frac{\kappa(\rho D^* + \mathcal{E})}{\gamma^k \varepsilon} + \gamma^k \varepsilon \right) \sigma_{t_k} \\
&\leq (2\kappa D + \varepsilon)\sigma_1 + \frac{\Delta_{1:T}^*}{\varepsilon} + \frac{\kappa C}{\rho D^*} \sum_{t=1}^T \frac{\sigma_t^2}{\sigma_{1:t}} + \frac{\kappa \Delta'}{\rho D^*} \qquad \text{(Thm. A.2 and Lem. B.2)} \\
&\qquad + \kappa \sum_{t=1}^T \left(\frac{\nu \mathcal{B}_R(\theta_t'||\phi')}{\rho D^*} + \gamma(\rho D^* + \mathcal{E}) + \varepsilon \right) \sigma_t + \sum_{k=0}^{\lfloor \log_\gamma \frac{\kappa(\rho D^* + \mathcal{E})}{\varepsilon} \rfloor} \left(\frac{\kappa(\rho D^* + \mathcal{E})}{\gamma^k \varepsilon} + \gamma^k \varepsilon \right) \sigma_{t_k} \\
&= (2\kappa D + \varepsilon)\sigma_1 + \frac{\Delta_{1:T}^*}{\varepsilon} + \frac{\kappa C}{\rho D^*} \sum_{t=1}^T \frac{\sigma_t^2}{\sigma_{1:t}} + \frac{\kappa \Delta'}{\rho D^*} \\
&\qquad + \kappa \left(\frac{\nu \bar{D}^2}{\rho D^*} + \gamma(\rho D^* + \mathcal{E}) + \varepsilon \right) \sigma_{1:T} + \sum_{k=0}^{\lfloor \log_\gamma \frac{\rho D^* + \mathcal{E}}{\varepsilon} \rfloor} \left(\frac{\kappa(\rho D^* + \mathcal{E})}{\gamma^k \varepsilon} + \gamma^k \varepsilon \right) \sigma_{t_k}
\end{aligned}
$$

□

The following result corresponds to the general case of Theorem 2.1.

Corollary B.1. *In the Exact case of Setting B.1, if $G_t = G, m_t = m \; \forall \; t \in [T]$, the FAL variant of Algorithm 2 has TAR*

$$
\bar{\mathbf{R}} \leq \left(\frac{2D + 2\varepsilon + \frac{C}{D^*}(1 + \log T) + \frac{\gamma}{\gamma - 1}\left(\frac{D^{*2}}{\varepsilon} + D^* \right)}{T} + \frac{\bar{D}^2}{D^*} + \gamma D^* + \varepsilon \right) G\sqrt{m}
$$

If we assume known D, picking $\varepsilon = D\frac{1+\log T}{T}$ and $\gamma = \frac{1+\log T}{\log T}$ yields

$$
\bar{\mathbf{R}} \leq \left(\left(6D + \frac{C}{D^*} \right) \frac{1 + \log T}{T} + \frac{9}{2} D^* \right) G\sqrt{m}
$$

Proof. For $K = \lfloor \log_\gamma \frac{D^*}{\varepsilon} \rfloor$ we have

$$
\sum_{k=0}^K \left(\frac{D^{*2}}{\gamma^k \varepsilon} + \gamma^k \varepsilon \right) = \frac{(\gamma^{K+1} - 1)(D^{*2} + \gamma^K \varepsilon^2)}{\gamma^K(\gamma - 1)\varepsilon} \leq \frac{\gamma}{\gamma - 1} \left(\frac{D^{*2}}{\varepsilon} + D^* \right)
$$

The result follows by noting that in the exact case we have $\kappa = \nu = \rho = 1, \Delta_{1:T}^* = \Delta' = \Delta_{\max} = 0$, and substituting $\sum_{t=1}^T \frac{1}{t} \leq (1 + \log T)$. □

B.2. Lower Bound

The following lower bound, which extends Theorem 4.2 of Abernethy et al. (2008) to the multi-task setting, shows that the previous TAR guarantees are optimal up to a constant multiplicative factor. Note that while the result is stated in terms of the task divergence D^*, since $D^* \geq \bar{D}$ the same lower bound holds for the average task deviation as well.

Theorem B.2. *Suppose the action space is $\Theta \subset \mathbb{R}^d$ for $d \geq 3$ and for each task $t \in [T]$ an adversary must play a a sequence of m_t convex G_t-Lipschitz functions $\ell_{t,i} : \Theta \mapsto \mathbb{R}$ whose optimal actions in hindsight $\arg\min_{\theta\in\Theta} \sum_{i=1}^{m_t} \ell_{t,i}(\theta)$ are contained in some fixed ℓ_2-ball $\Theta^* \subset \Theta$ with center ϕ^* and diameter D^*. Then the adversary can force the agent to have task-averaged regret at least $\frac{D^*}{4T}\sum_{t=1}^T G_t\sqrt{m_t}$.*

Proof. Let $\{\theta_{t,i}\}_{i=1}^m$ be the sequence of actions of the agent on task t. Define $c(\theta) = \frac{G_t}{2}\max\{0, \|\theta - \phi^*\|_2 - D^*\}$, which is 0 on Θ^* and an upward-facing cone with vertex $\left(\phi^*, -\frac{G_t D^*}{2}\right)$ and slope $\frac{G_t}{2}$ on the complement. The strategy of the adversary at round i of task t will be to play $\ell_{t,i}(\theta) = \langle\nabla_{t,i}, \theta - \phi^*\rangle + c(\theta)$, where $\nabla_{t,i}$ satisfies $\|\nabla_{t,i}\|_2 = \frac{G_t}{2}$, $\langle\nabla_{t,i}, \theta_{t,i} - \phi^*\rangle = 0$, and $\langle\nabla_{t,i}, \nabla_{t,1:i-1}\rangle = 0$. Such a $\nabla_{t,i}$ always exists for $d \geq 3$. Note that these conditions imply that along any direction from ϕ^* the total loss $\sum_{i=1}^{m_t}\ell_{t,i}(\theta)$ is increasing outside Θ^* and so is minimized inside Θ^*, so we have

$$\min_{\theta\in\Theta}\sum_{i=1}^{m_t}\ell_{t,i}(\theta) = \min_{\theta\in\Theta^*}\sum_{i=1}^{m_t}\langle\nabla_{t,i}, \theta - \phi^*\rangle = \min_{\|\theta-\phi^*\|_2 \leq \frac{D^*}{2}}\langle\theta - \phi^*, \nabla_{t,1:m_t}\rangle = -\frac{D^*}{2}\|\nabla_{t,1:m_t}\|_2$$

Note that the condition $\langle\nabla_{t,i}, \theta_{t,i} - \phi^*\rangle = 0$ and the nonnegativity of $c(\theta)$ implies that the loss of the agent is at least 0, and so the agent's regret on task t satisfies $\mathbf{R}_{m_t} \geq \frac{D^*}{2}\|\nabla_{t,1:m_t}\|_2$. By the condition $\langle\nabla_{t,i}, \nabla_{t,1:i-1}\rangle = 0$ we have that

$$\|\nabla_{t,1:i}\|_2^2 = \|\nabla_{t,i} + \nabla_{t,1:i-1}\|_2^2 = \|\nabla_{t,i}\|_2^2 + \|\nabla_{t,1:i-1}\|_2^2 = \frac{G_t^2}{4} + \|\nabla_{t,i-1}\|_2^2$$

and so by induction on i with base case $\|\nabla_{t,1}\|_2 = \frac{G_t}{2}$ we have $\|\nabla_{t,1:m_t}\|_2 = \frac{G_t}{2}\sqrt{m_t} \implies \mathbf{R}_{m_t} \geq \frac{G_t D^*}{4}\sqrt{m_t}$. Substituting the regret on each task into $\bar{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^T \mathbf{R}_{m_t}$ completes the proof. $\square$

B.3. Task-Averaged Regret for Approximate Meta-Updates

For the Approx variants of FMRL we need a bound on the distance between the last or average iterate of FTRL/OMD and the best parameter in hindsight. This necessitates further assumptions on the loss functions besides convexity, as a task may otherwise have functions with very small losses, even far away from the optimal parameter, in which case the last iterate of FTRL/OMD will be far away if the initial point is far away from the optimum. Here we make use of the α-QG assumption on the average loss functions to obtain stability of the estimates w.r.t. the true loss.

Lemma B.3. *Let $\ell_1, \dots, \ell_m$ be a sequence of convex losses on Θ with $L(\theta) = \frac{1}{m}\sum_{i=1}^m \ell(\theta)$ being α-QG w.r.t. $\|\cdot\|$ and define $\hat{\theta} = \arg\min_{\theta\in\Theta} \mathcal{B}_R(\theta||\phi) + \eta m L(\theta)$ to be the last iterate of running $\operatorname{FTRL}_{\eta,\phi}^{(R)}$ for $\eta > 0, \phi \in \Theta$, and $R : \Theta \mapsto \mathbb{R}$ 1-strongly-convex w.r.t. $\|\cdot\|$. Then the closest minimum $\theta^* \in \Theta$ of L to $\hat{\theta}$ satisfies*

$$\frac{1}{2}\|\theta^* - \hat{\theta}\|^2 \leq \frac{\mathcal{B}_R(\theta^*||\phi) - \mathcal{B}_R(\hat{\theta}||\phi)}{\alpha\eta m}$$

Proof. We have by definition of θ' and $\hat{\theta}$ that

$$\mathcal{B}_R(\theta^*||\phi) + \eta m L(\theta^*) \geq \mathcal{B}_R(\hat{\theta}||\phi) + \eta m L(\hat{\theta})$$

On the other hand since L is α-QG we have that

$$L(\hat{\theta}) \geq L(\theta^*) + \frac{\alpha}{2}\|\theta^* - \hat{\theta}\|^2$$

Multiplying the second inequality by ηm and adding it to the first yields the result. □

Proposition B.1. *In Setting B.1 , if for each task $t \in [T]$ the losses $\ell_{t,1}, \dots, \ell_{t,m_t}$ satisfy the α-QG condition as in Lemma B.3 and $\varepsilon \geq \max_t \frac{4\beta G_t}{\alpha\sqrt{m_t}}$, then for $\hat{\theta}_t$ set according to the FLI-Online algorithm and $\theta_t^* = \theta_t' \;\forall\, t \in [T]$ we have*

$$\kappa = 4\beta, \qquad \Delta_t^* = 0 \;\forall\, t \in [T], \qquad \nu = 3\beta, \qquad \Delta' = \frac{6\beta D^2}{\alpha\varepsilon}\sum_{t=1}^T \frac{G_t\sigma_t}{\sqrt{m_t}}, \qquad \Delta_{\max} = \max_t \frac{D^2 G_t}{\alpha\varepsilon\sqrt{m_t}}$$

Proof. Applying the triangle inequality, Jensen's inequality, and Lemma B.3 yields the first two values:

$$\|\theta_t^* - \phi_t\|^2 \leq 2\|\theta_t^* - \hat{\theta}_t\|^2 + 2\|\hat{\theta}_t - \phi_t\|^2 \leq \frac{4\mathcal{B}_R(\theta_t^*||\phi_t)}{\alpha\eta_t m_t} + 4\mathcal{B}_R(\hat{\theta}_t||\phi_t) \leq \frac{2\beta}{\alpha\eta_t m_t}\|\theta_t^* - \phi_t\|^2 + 4\mathcal{B}_R(\hat{\theta}_t||\phi_t)$$

$$\implies \mathcal{B}_R(\theta_t^*||\phi_t) \leq \frac{\beta}{2}\|\theta_t^* - \phi_t\|^2 \leq \frac{2\beta\mathcal{B}_R(\hat{\theta}_t||\phi_t)}{1 - \frac{2\beta}{\alpha\eta_t m_t}} \leq 4\beta\mathcal{B}_R(\hat{\theta}_t||\phi_t)$$

Here in the last step we used the fact that $\varepsilon \geq \frac{4\beta G_t}{\alpha\sqrt{m_t}} \implies \eta_t \geq \frac{4\beta}{\alpha m_t} \;\forall\, t \in [T]$. For the next two values, noting that for FLI-Online, $\theta_t^* = \theta_t' \;\forall\, t \in [T]$ we have by the triangle inequality and Titu's lemma that

$$\|\phi' - \hat{\phi}\|^2 = \frac{1}{(\sigma_{1:T})^2}\left\|\sum_{t=1}^T \sigma_t\theta_t' - \sum_{t=1}^T \sigma_t\hat{\theta}_t\right\|^2 \leq \frac{1}{(\sigma_{1:T})^2}\left(\sum_{t=1}^T \sigma_t\|\theta_t' - \hat{\theta}_t\|\right)^2 \leq \frac{1}{\sigma_{1:T}}\sum_{t=1}^T \sigma_t\|\theta_t' - \hat{\theta}_t\|^2$$

Therefore since $\eta \geq \frac{\varepsilon}{\sigma_t}$ and $\mathcal{B}_R(\theta_t^*||\phi_t) \leq D^2$ we have that

$$\sum_{t=1}^T \sigma_t\mathcal{B}_R(\hat{\theta}_t||\hat{\phi}) \leq \frac{3\beta}{2}\sum_{t=1}^T \sigma_t(\|\hat{\theta}_t - \theta_t'\|^2 + \|\theta_t' - \phi'\|^2 + \|\phi' - \hat{\phi}\|^2) \leq 3\beta\sum_{t=1}^T \sigma_t\left(\frac{2\mathcal{B}_R(\theta_t^*||\phi_t)}{\alpha\eta_t m_t} + \mathcal{B}_R(\theta_t'||\phi')\right)$$

The last value follows directly by Lemma B.3, $\eta_t \geq \frac{\varepsilon}{\sigma_t}$, and the bound D^2 on the maximum Bregman divergence. □

The following upper bound yields Theorem 3.1:

Corollary B.2. *In the Approx. case of Setting B.1, if* $G_t = G, m_t = m \;\forall\, t \in [T], \gamma = \frac{1+\log T}{\log T}$*, and* $\varepsilon = \frac{4\beta G}{\alpha \sqrt[6]{m}} + D\frac{1+\log T}{T}$ *then the FLI-Online variant of Algorithm 2 has TAR*

$$\bar{\mathbf{R}} = \mathcal{O}\left(\frac{D}{D^*}\left(\frac{\log T}{T} + \frac{1}{\sqrt[6]{m}}\right) + D^*\right) G\sqrt{m}$$

Proof. Substitute Proposition B.1 into Theorem B.1 and simplify. □

Lemma B.4. *Let* $\ell_1, \dots, \ell_m : \Theta \mapsto [0,1]$ *be a sequence of convex losses on* Θ *drawn i.i.d. from some distribution* $\mathcal{D}$ *with risk* $\mathbb{E}_{\ell\sim\mathcal{D}}\,\ell$ *being* α*-QG w.r.t.* $\|\cdot\|$ *and let* $\theta^* \in \arg\min_{\theta\in\Theta} \sum_{i=1}^m \ell_i(\theta)$ *be any of the optimal actions in hindsight. Then w.p.* $1-\delta$ *the closest minimum* $\theta' \in \Theta$ *of* $\mathbb{E}_{\ell\sim\mathcal{D}}\,\ell$ *to* θ^* *satisfies*

$$\frac{1}{2}\|\theta^* - \theta'\|^2 \le \sqrt{\frac{8}{\alpha^2 m}\log\frac{2}{\delta}}$$

Proof. By definition of θ^* and θ' we have w.p. $1-\delta$ that

$$\begin{aligned}
\frac{\alpha}{2}\|\theta^* - \theta'\|^2 &\le \frac{1}{m}\underset{\{\ell_i\}\sim\mathcal{D}^m}{\mathbb{E}}\sum_{i=1}^m \ell_i(\theta^*) - \frac{1}{m}\underset{\{\ell_i\}\sim\mathcal{D}^m}{\mathbb{E}}\sum_{i=1}^m \ell_i(\theta') && \text{(apply } \alpha\text{-QG)}\\
&\le \frac{1}{m}\sum_{i=1}^m \ell_i(\theta^*) - \frac{1}{m}\sum_{i=1}^m \ell_i(\theta') + \sqrt{\frac{8}{m}\log\frac{2}{\delta}} && \text{(apply Prp. A.2 twice)}\\
&\le \sqrt{\frac{8}{m}\log\frac{2}{\delta}} && \text{(definition of } \theta^*\text{)}
\end{aligned}$$

□

Lemma B.5. *Suppose* $\forall\, t \in [T]$ *the r.v.* Q_t *satisfies* $0 \le Q_t \le B$ *a.s. and* $Q_t \le \sqrt{\frac{8}{m_t}\log\frac{2}{\delta}}$ *w.p.* $1-\delta$ *for any* $\delta \in (0,1)$*. Then for nonnegative* $\alpha_1, \dots, \alpha_T$ *we have w.p.* $1-\gamma$ *for any* $\gamma \in (0,1)$ *that*

$$\sum_{t=1}^T \alpha_t Q_t \le \frac{2B\alpha_{\max}}{3}\log\frac{1}{\gamma} + 2\sum_{t=1}^T \alpha_t\sqrt{\frac{1+4\log(Bm_t)}{m_t}\left(1+\log\frac{1}{\gamma}\right)}$$

Proof. Define convenience coefficients $\beta_t = \frac{\alpha_t}{\alpha_{1:T}}$, the auxiliary sequence $Z_t = \beta_t Q_t \;\forall\, t \in [T]$, the martingale sequence $Y_0 = 0, Y_t = Z_{1:t} - \mathbb{E}\, Z_{1:t} \;\forall\, t \in [T]$ and the associated martingale difference sequence $X_t = Y_t - Y_{t-1} \;\forall\, t \in [T]$. By substituting $\delta = \frac{2}{Bm_t}$ we then have

$$\mathbb{E}_{t-1} X_t^2 = \mathbb{E}_{t-1}(Y_t - Y_{t-1})^2 = \beta_t^2\,\mathbb{E}(Q_t - \mathbb{E}\,Q_t)^2 \le \beta_t^2\,\mathbb{E}\,Q_t^2 \le \beta_t^2\left(\frac{8}{m_t}\log\frac{2}{\delta} + \delta B\right) \le \frac{2+8\log(Bm_t)}{m_t}\beta_t^2$$

Note further that using $\delta = \frac{2}{\sqrt{Bm_t}}$ and Jensen's inequality we have

$$\mathbb{E}\,Q_t \le \sqrt{\frac{8}{m_t}\log\frac{2}{\delta}} + \delta B \le \sqrt{\frac{4+8\log(Bm_t)}{m_t}}$$

Noting that $Q_t \le B$ a.s. $\implies X_t \le B$ a.s., we have by Freedman's inequality (Freedman, 1975, Theorem 1.6) that

$$\mathbb{P}\left(\sum_{t=1}^T \beta_t Q_t \ge \tau + 2\sum_{t=1}^T \beta_t\sqrt{\frac{1+2\log(Bm_t)}{m_t}}\right) \le \mathbb{P}\left(\sum_{t=1}^T \beta_t Q_t \ge \tau + \sum_{t=1}^T \beta_t\,\mathbb{E}\,Q_t\right) \le \exp\left(-\frac{\tau^2}{2\sigma^2 + \frac{2B\beta_{\max}}{3}\tau}\right)$$

for $\tau \ge 0, \sigma^2 = \sum_{t=1}^T \frac{2+8\log(Bm_t)}{m_t}\beta_t^2$. Substituting $\tau = \frac{2\beta_{\max}}{3}\log\frac{1}{\gamma} + \sqrt{2\sigma^2\log\frac{1}{\gamma}}$ yields

$$\mathbb{P}\left(\sum_{t=1}^T \beta_t Q_t \ge \frac{2B\beta_{\max}}{3}\log\frac{1}{\gamma} + 2\sum_{t=1}^T \beta_t\sqrt{\frac{1+2\log(Bm_t)}{m_t}} + \sqrt{2\log\frac{1}{\gamma}\sum_{t=1}^T \frac{2+8\log(Bm_t)}{m_t}\beta_t^2}\right) \le \gamma$$

□

Proposition B.2. *In Setting B.1, if for each task $t \in [T]$ the losses $\ell_{t,1}, \dots, \ell_{t,m_t}$ and reference parameter θ'_t satisfy the α-QG condition as in Lemma B.4, then for $\hat{\theta}_t = \theta^*_t$ set according to the FAL algorithm we have w.p. $1-\delta$ that $\kappa = 1, \nu = 3\beta$,*

$$\Delta^*_t = 0\ \forall\, t \in [T], \quad \Delta' = \frac{4\beta}{\alpha}\left(\sigma_{\max}\log\frac{2}{\delta} + 3\sum_{t=1}^T \sigma_t \sqrt{\frac{1+4\log m_t}{m_t}\left(1+\log\frac{2}{\delta}\right)}\right), \quad \Delta_{\max} = \frac{4}{\alpha}\sqrt{\frac{1}{m_{\min}}\log\frac{2T}{\delta}}$$

Proof. $\kappa = 1$ and $\Delta^*_t = 0\ \forall\, t \in [T]$ because $\hat{\theta}_t = \theta^*_t\ \forall\, t \in [T]$. Applying Titu's lemma as in the proof of Proposition B.1 yields the values of ν and Δ' w.p. $1-2\delta$:

$$\begin{aligned}\sum_{t=1}^T \sigma_t \mathcal{B}_R(\hat{\theta}_t||\hat{\phi}) &\le \frac{3\beta}{2}\sum_{t=1}^T \sigma_t(\|\theta^*_t - \theta'_t\|^2 + \|\theta'_t - \phi'\|^2 + \|\phi' - \hat{\phi}\|^2)\\ &\le 3\beta \sum_{t=1}^T \sigma_t\left(\|\theta^*_t - \theta'_t\|^2 + \mathcal{B}_R(\theta'_t||\phi')\right)\\ &\le \frac{4\beta\sigma_{\max}}{\alpha}\log\frac{1}{\delta} + \frac{12\beta}{\alpha}\sum_{t=1}^T \sigma_t\sqrt{\frac{1+4\log m_t}{m_t}\left(1+\log\frac{2}{\delta}\right)} + 3\beta\sum_{t=1}^T \sigma_t \mathcal{B}_R(\theta'_t||\phi')\end{aligned}$$

Here in the last step we applied Lemma B.5 on $Q_t = \frac{\alpha}{2}\|\theta^*_t - \theta'_t\|^2$, which is 1-bounded by Lemma B.4. The value of $\Delta_{\max}$ follows directly by Lemma B.4 w.p. $1-2\delta$. □

The following upper bound yields the FAL result in Theorem 3.1:

Corollary B.3. *In the Approx. case of Setting B.1, if $G_t = G, m_t = m\ \forall\, t \in [T], \gamma = \frac{1+\log T}{\log T}$, and $\varepsilon = D\frac{1+\log T}{T}$ then the FAL variant of Algorithm 2 has TAR*

$$\bar{\mathbf{R}} = \mathcal{O}\left(\frac{D}{D^*}\left(\frac{\log T}{T} + \sqrt{\frac{1}{\sqrt[3]{m}}\log\frac{Tm}{\delta}}\right) + D^*\right) G\sqrt{m}$$

Proof. Substitute Proposition B.2 into Theorem B.1 and simplify. □

Lemma B.6. *Let $\ell_1, \dots, \ell_m : \Theta \mapsto [0,1]$ be a sequence of G_i-Lipschitz convex losses on Θ drawn i.i.d. from some distribution $\mathcal{D}$ with risk $\mathbb{E}_{\ell\sim\mathcal{D}}\,\ell$ being α-QG w.r.t. $\|\cdot\|$ and define $\hat{\theta} = \frac{1}{m}\theta_{1:m}$ to be the the average iterate of running $\mathrm{FTRL}^{(R)}_{\eta,\phi}$ or $\mathrm{OMD}^{(R)}_{\eta,\phi}$ on $\ell_1, \dots, \ell_m$ for $\eta > 0, \phi \in \Theta$, and $R : \Theta \mapsto \mathbb{R}$ 1-strongly convex w.r.t. $\|\cdot\|$. Then w.p. $1-\delta$ the closest minimum $\theta' \in \Theta$ of $\mathbb{E}_{\ell\sim\mathcal{D}}\,\ell$ to $\hat{\theta}$ satisfies*

$$\frac{1}{2}\|\theta' - \hat{\theta}\|^2 \le \frac{\mathcal{B}_R(\theta'||\phi) + \eta^2 G^2 m + \eta\sqrt{8m\log\frac{2}{\delta}}}{\alpha\eta m}$$

where $G^2 = \frac{1}{m}\sum_{i=1}^m G_i^2$.

Proof. By definition of $\hat{\theta}$ and θ' we have w.p. $1-\delta$ that

$$\begin{aligned}\frac{\alpha}{2}\|\hat{\theta} - \theta'\|^2 &\le \frac{1}{m}\mathop{\mathbb{E}}_{\{\ell_i\}\sim\mathcal{D}^m}\sum_{i=1}^m \ell_i(\theta_i) - \frac{1}{m}\mathop{\mathbb{E}}_{\{\ell_i\}\sim\mathcal{D}^m}\sum_{i=1}^m \ell_i(\theta') && \text{(apply } \alpha\text{-QG and Jensen's inequality)}\\ &\le \frac{1}{m}\sum_{i=1}^m \ell_i(\theta_i) - \frac{1}{m}\sum_{i=1}^m \ell_i(\theta') + \sqrt{\frac{8}{m}\log\frac{2}{\delta}} && \text{(apply Prp. A.2 twice)}\\ &\le \frac{\frac{1}{\eta}\mathcal{B}_R(\theta'||\phi) + \eta G^2 m}{m} + \sqrt{\frac{8}{m}\log\frac{2}{\delta}} && \text{(substitute the regret of FTRL/OMD)}\end{aligned}$$

□

Proposition B.3. *In Setting B.1, if for each task $t \in [T]$ the losses $\ell_{t,1}, \dots, \ell_{t,m_t}$ and reference parameter θ'_t satisfy the α-QG condition as in Lemma B.6 and $\varepsilon \geq \max_t \frac{24\beta G_t}{\alpha\sqrt{m_t}}$, then for $\hat{\theta}_t$ set according to the FLI-Batch algorithm we have w.p. $1-\delta$ that $\kappa = 12\beta, \nu = 3\beta$,*

$$\Delta_t^* = \frac{3\beta}{\alpha}\left(1 + \frac{4\beta G_t}{\alpha\varepsilon}\right)\left(\frac{2\alpha_{\max}}{3\alpha_t T}\log\frac{3}{\delta} + 2\sqrt{\frac{1+4\log m_t}{m_t}\left(1+\log\frac{3}{\delta}\right)}\right) + \frac{12\beta G_t^2(D+\varepsilon)}{\alpha m_t}\ \forall\, t \in [T]$$

$$\Delta' = \frac{4\beta\sigma_{\max}}{\alpha}\log\frac{3}{\delta} + \frac{12\beta}{\alpha}\sum_{t=1}^T\left(\sqrt{\frac{1+4\log m_t}{m_t}\left(1+\log\frac{3}{\delta}\right)} + \frac{(D^2+\varepsilon)G_t}{2\varepsilon\sqrt{m_t}}\right)\sigma_t, \quad \Delta_{\max} = \frac{1}{\alpha}\sqrt{\frac{8}{m_{\min}}\log\frac{6T}{\delta}}$$

Proof. Applying the triangle inequality, Jensen's inequality, Lemma B.4, and Lemma B.6 yields w.p. $1-\delta$

$$\begin{aligned}
\|\theta_t^* - \phi_t\|^2 &\leq 3\|\theta_t^* - \theta'_t\|^2 + 3\|\theta'_t - \hat{\theta}_t\|^2 + 3\|\hat{\theta}_t - \phi_t\|^2 \\
&\leq \frac{3}{\alpha}\sqrt{\frac{8}{m_t}\log\frac{2}{\delta}} + \frac{6}{\alpha\eta_t m_t}\left(\mathcal{B}_R(\theta'_t||\phi_t) + \eta_t^2 G_t^2 m_t + \eta_t\sqrt{8m_t\log\frac{2}{\delta}}\right) + 6\mathcal{B}_R(\hat{\theta}_t||\phi_t) \\
&\leq \frac{9}{\alpha}\sqrt{\frac{8}{m_t}\log\frac{2}{\delta}} + \frac{12\beta}{\alpha\eta_t m_t}(\|\theta'_t - \theta_t^*\|^2 + \|\theta_t^* - \phi_t\|^2) + \frac{6G_t^2(D+\varepsilon)}{\alpha m_t} + 6\mathcal{B}_R(\hat{\theta}_t||\phi_t) \\
&\leq \frac{3}{\alpha}\left(1 + \frac{4\beta G_t}{\alpha\varepsilon}\right)\sqrt{\frac{8}{m_t}\log\frac{2}{\delta}} + \frac{12\beta}{\alpha\eta_t m_t}\|\theta_t^* - \phi_t\|^2 + \frac{6G_t^2(D+\varepsilon)}{\alpha m_t} + 6\mathcal{B}_R(\hat{\theta}_t||\phi_t)
\end{aligned}$$

where we have used the uniqueness of the reference parameter θ'_t. The above implies

$$\mathcal{B}_R(\theta_t^*||\phi_t) \leq \frac{\beta}{2}\|\theta_t^* - \phi_t\|^2 \leq \frac{3\beta}{\alpha}\left(1 + \frac{4\beta G_t}{\alpha\varepsilon}\right)\sqrt{\frac{8}{m_t}\log\frac{2}{\delta}} + \frac{12\beta G^2(D+\varepsilon)}{\alpha m_t} + 12\beta\mathcal{B}_R(\hat{\theta}_t||\phi_t)$$

Here in the last step we used the fact that $\varepsilon \geq \frac{24\beta G_t}{\alpha\sqrt{m_{\min}}} \implies \eta_t \geq \frac{24\beta}{\alpha m_t}\ \forall\, t \in [T]$. Thus by Lemma B.5 w.p. $1-3\delta$

$$\begin{aligned}
\sum_{t=1}^T \alpha_t\mathcal{B}_R(\theta_t^*||\phi_t) \leq \frac{3\beta}{\alpha}\left(1 + \frac{4\beta G_t}{\alpha\varepsilon}\right)&\left(\frac{2\alpha_{\max}}{3}\log\frac{3}{\delta} + 2\sum_{t=1}^T\alpha_t\sqrt{\frac{1+4\log m_t}{m_t}\left(1+\log\frac{3}{\delta}\right)}\right) \\
&+ \frac{12\beta G_t^2(D+\varepsilon)}{\alpha}\sum_{t=1}^T\frac{\alpha_t}{m_t} + 12\beta\sum_{t=1}^T\alpha_t\mathcal{B}_R(\hat{\theta}_t||\phi_t)
\end{aligned}$$

This yields the values of κ and $\Delta_t^*\ \forall\, t \in [T]$. We next have by applying Titu's lemma as in the proof of Proposition B.1

$$\begin{aligned}
\sum_{t=1}^T\sigma_t\mathcal{B}_R(\hat{\theta}_t||\hat{\phi}) &\leq 3\beta\sum_{t=1}^T\sigma_t(\|\hat{\theta}_t - \theta'_t\|^2 + \mathcal{B}_R(\theta'_t||\phi')) \\
&\leq \frac{6\beta}{\alpha}\sum_{t=1}^T\frac{\sigma_t\mathcal{B}_R(\theta'_t||\phi_t)}{\eta_t m_t} + \sigma_t\eta_t G^2 + \sigma_t\sqrt{\frac{8}{m_t}\log\frac{2}{\delta}} + 3\beta\sum_{t=1}^T\sigma_t\mathcal{B}_R(\theta'_t||\phi')) \\
&\leq \frac{4\beta\sigma_{\max}}{\alpha}\log\frac{3}{\delta} + \frac{12\beta}{\alpha}\sum_{t=1}^T\left(\sqrt{\frac{1+4\log m_t}{m_t}\left(1+\log\frac{3}{\delta}\right)} + \frac{(D^2+\varepsilon)G_t}{2\varepsilon\sqrt{m_t}}\right)\sigma_t + 3\beta\sum_{t=1}^T\sigma_t\mathcal{B}_R(\theta'_t||\phi')
\end{aligned}$$

This yields the values of ν and Δ'. The value of $\Delta_{\max}$ follows directly by Lemma B.4 w.p. $1-3\delta$. □

The following final upper bound yields the FLI-Batch result in Theorem 3.1:

Corollary B.4. *In the Approx. case of Setting B.1, if $G_t = G, m_t = m\ \forall\, t \in [T], \gamma = \frac{1+\log T}{\log T}$, and $\varepsilon = \frac{24\beta G}{\alpha\sqrt{m}} + D\frac{1+\log T}{T}$ then the FLI-Batch variant of Algorithm 2 has TAR*

$$\bar{\mathbf{R}} = \mathcal{O}\left(\frac{D}{D^*}\left(\frac{\log T}{T} + \sqrt{\frac{1}{\sqrt[3]{m}}\log\frac{Tm}{\delta}}\right) + D^*\right)G\sqrt{m}$$

Proof. Substitute Proposition B.3 into Theorem B.1 and simplify. □

B.4. Online-to-Batch Conversion for Task-Averaged Regret

The following yields a bound on the expected transfer risk when randomizing over the output of any TAR-minimizing algorithm when in the setting of statistical LTL.

Theorem B.3. *Let $\mathcal{Q}$ be a distribution over distributions $\mathcal{P}$ over convex loss functions $\ell : \Theta \mapsto [0,1]$. A sequence of sequences of loss functions $\{\ell_{t,i}\}_{t\in[T],i\in[m]}$ is generated by drawing m loss functions i.i.d. from each in a sequence of distributions $\{\mathcal{P}_t\}_{t\in[T]}$ themselves drawn i.i.d. from $\mathcal{Q}$. If such a sequence is given to an meta-learning algorithm with task-averaged regret bound $\bar{\mathbf{R}}$ that has states $\{s_t\}_{t\in[T]}$ at the beginning of each task t then we have w.p. $1-\delta$ for any $\theta^*\in\Theta$ that*

$$\mathop{\mathbb{E}}_{t\sim\mathcal{U}[T]}\mathop{\mathbb{E}}_{\mathcal{P}\sim\mathcal{Q}}\mathop{\mathbb{E}}_{\mathcal{P}^m}\mathop{\mathbb{E}}_{\ell\sim\mathcal{P}}\ell(\bar{\theta}) \le \mathop{\mathbb{E}}_{\mathcal{P}\sim\mathcal{Q}}\mathop{\mathbb{E}}_{\ell\sim\mathcal{P}}\ell(\theta^*) + \frac{\bar{\mathbf{R}}}{m} + \sqrt{\frac{8}{T}\log\frac{1}{\delta}}$$

where $\bar{\theta} = \frac{1}{m}\theta_{1:m}$ is generated by randomly sampling $t\in\mathcal{U}[T]$, running the online algorithm with state s_t, and averaging the actions $\{\theta_i\}_{i\in[m]}$.

Proof. Applying Proposition A.1, linearity of expectations, the fact that the regret over 1-bounded loss functions is m-bounded, and Proposition A.2 yields

$$\begin{aligned}
\mathop{\mathbb{E}}_{t\sim\mathcal{U}[T]}\mathop{\mathbb{E}}_{\mathcal{P}\sim\mathcal{Q}}\mathop{\mathbb{E}}_{\mathcal{P}^m}\mathop{\mathbb{E}}_{\ell\sim\mathcal{P}}\ell(\bar{\theta}) \le \mathop{\mathbb{E}}_{\mathcal{P}\sim\mathcal{Q}}\left(\mathop{\mathbb{E}}_{\ell\sim\mathcal{P}}\ell(\theta^*) + \frac{\mathbf{R}_m(s_t)}{m}\right) &\le \mathop{\mathbb{E}}_{\mathcal{P}\sim\mathcal{Q}}\mathop{\mathbb{E}}_{\ell\sim\mathcal{P}}\ell(\theta^*) + \frac{1}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{\mathcal{P}\sim\mathcal{Q}}\left(\frac{\mathbf{R}_m(s_t)}{m}\right)\\
&= \mathop{\mathbb{E}}_{\mathcal{P}\sim\mathcal{Q}}\mathop{\mathbb{E}}_{\ell\sim\mathcal{P}}\ell(\theta^*) + \frac{2}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{\mathcal{P}\sim\mathcal{Q}}\left(\frac{\mathbf{R}_m(s_t)}{2m}+\frac{1}{2}\right) - 1\\
&\le \mathop{\mathbb{E}}_{\mathcal{P}\sim\mathcal{Q}}\mathop{\mathbb{E}}_{\ell\sim\mathcal{P}}\ell(\theta^*) + \frac{\bar{\mathbf{R}}}{m} + \sqrt{\frac{8}{T}\log\frac{1}{\delta}}
\end{aligned}$$

$\square$

C. Computing the Quadratic Growth Factor

For our analysis of the FLI variants of Algorithm 2 we consider a class of functions related to strongly convex functions that satisfy the quadratic growth (QG) condition:

$$\frac{\alpha}{2}\|\theta-\theta^*\|^2 \le f(\theta)-f(\theta^*) \tag{6}$$

By Theorem 2 of Karimi et al. (2016), in the convex case QG is equivalent, up to multiplicative constants, with the Polyak-Łojaciewicz (PL) inequality (Polyak, 1963). Using the latter condition, Karimi et al. (2016) further show that functions of form $f(A\theta)$ for f strongly-convex satisfy the PL inequality, and thus also QG, with constant $\alpha=\Omega(\sigma_{\min}(A))$. This provides data-dependent guarantees for a variety of practical problems, including least-squares and logistic regression. Garber (2019) shows a similar result for expectations of such functions with the QG constant depending now on $\lambda_{\min}(\mathbb{E}\,A^TA)$; in order to do so they assume the constraint set is a polytope, e.g. an ℓ_1 or ℓ_∞ ball.

For our results we require a stronger condition, namely that if L is a sum of m convex losses then L satisfies αm-QG. While this additive property holds directly if the losses are strongly-convex, in the general case it does not. Furthermore, the spectral lower bound on α studied by Karimi et al. (2016) and Garber (2019) is an underestimate; for example, in the strongly-convex case, where A^TA is the identity, the lower bound will be 1 even though their sum is m-QG.

Here we derive an alternative approach for verifying α-QG for a convex Lipschitz function f constrained to a ball of radius B. Note that since the functions are Lipschitz, we can focus on computing the minimal difference between $f(\theta)$ and $f(\theta^*)$ over all θ located some fixed distance δ away from any minimizer θ^* of f over the ball:

$$\begin{aligned}\varepsilon_\delta = \min \quad & f(\theta)-f(\theta^*)\\ \text{s.t.} \quad & \|\theta-\theta^*\|_2^2 \ge \delta^2\\ & \|\theta\|_2 \le B\end{aligned}$$

Then if f is α-QG, Equation 6 implies that $\alpha_\delta=\frac{2\varepsilon_\delta}{\delta^2}$ should be a constant, or equivalently that $\varepsilon_\delta=\Omega(\delta^2)$. While the above problem is non-convex due to the first constraint, note that

$$\delta^2 \le \|\theta-\theta^*\|_2^2 = \|\theta\|_2^2 - 2\langle\theta,\theta^*\rangle + \|\theta^*\|_2^2 \le B^2 - 2\langle\theta,\theta^*\rangle + \|\theta^*\|_2^2$$

which is a linear constraint since θ^* is constant. Therefore we have

$$\begin{aligned}\varepsilon_\delta \ge \min \quad & f(\theta)-f(\theta^*)\\ \text{s.t.} \quad & 2\langle\theta^*,\theta\rangle \le B^2-\delta^2+\|\theta^*\|_2^2\\ & \|\theta\|_2 \le B\end{aligned}$$

which is a convex program amenable to standard solvers; we employ the Frank-Wolfe method (Frank & Wolfe, 1956).

D. Experimental Details

D.1. Constructing Mini-Wikipedia

We briefly describe the construction of Mini-Wiki. Starting with the raw corpus of the Wiki3029 dataset of Arora et al. (2019), we select those Wikipedia pages whose titles correspond to lemmas in the WordNet corpus (Fellbaum, 1998). We then use the hypernymy structure in this corpus to separate the pages into four semantically meaningful meta-classes; this is necessary when using linear classification as the task similarity only depends on the classifier and not the representation. Finally, we take the longest sentences from each page to construct m-shot tasks of $4m$ samples each, for $m = 1, 2, 4, \dots, 32$. We have made MiniWiki available here: `https://github.com/mkhodak/FMRL/blob/master/data/miniwikipedia.tar.gz`.

D.2. Complete Deep Learning Results

Below are plots for all evaluations on Omniglot and Mini-ImageNet. As our algorithm generalizes the Reptile method of Nichol et al. (2018), we use code they make available at `https://github.com/openai/supervised-reptile` and vary the parameters `train-shots` and `inner-iters`.

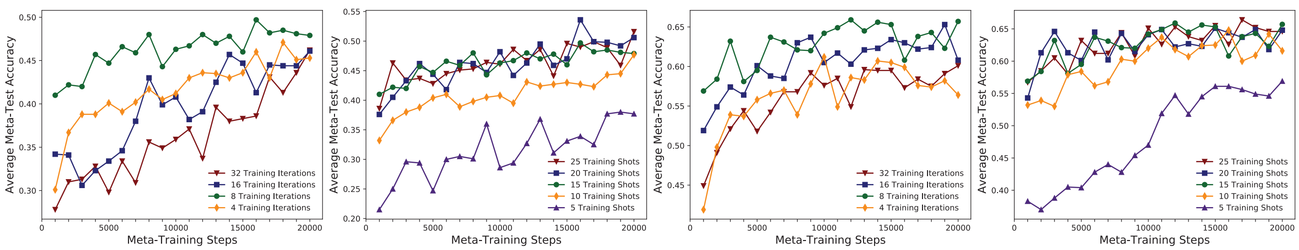

Figure 7. Performance of the FLI variant of Ephemeral with OGD within-task (Reptile) on 5-way Mini-ImageNet when varying the number of task samples and the number of iterations per training task. In the left-hand plots we use 1-shot at meta-test time; in the right-hand plots we use 5-shots. 50 iterations are used at meta-test time in both cases.

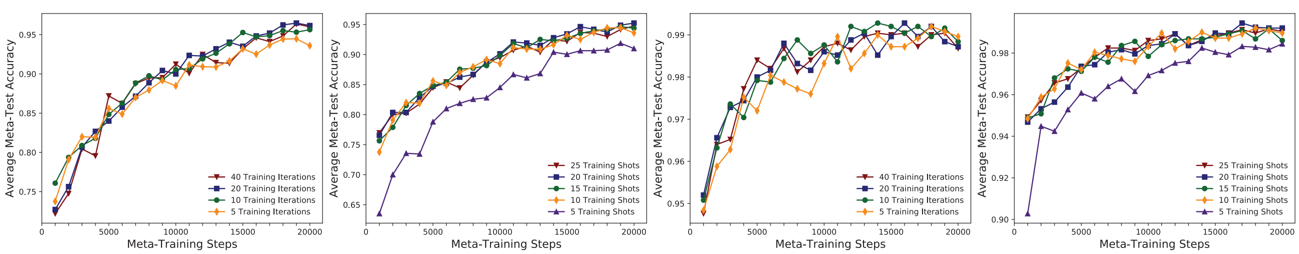

Figure 8. Performance of the FLI variant of Ephemeral with OGD within-task (Reptile) on 5-way Omniglot when varying the number of task samples and the number of iterations per training task. In the left-hand plots we use 1-shot at meta-test time; in the right-hand plots we use 5-shots. 50 iterations are used at meta-test time in both cases.

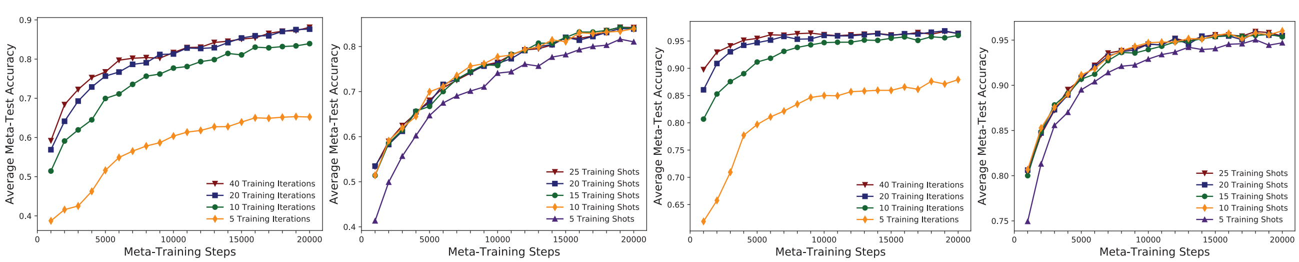

Figure 9. Performance of the FLI variant of Ephemeral with OGD within-task (Reptile) on 20-way Omniglot when varying the number of task samples and the number of iterations per training task. In the left-hand plots we use 1-shot at meta-test time; in the right-hand plots we use 5-shots. 50 iterations are used at meta-test time in both cases.