Appendix

6 Proofs for Section 2

Lemma 1 (Duality Between Smoothness and Convexity for Convex Functions). *Let $\mathcal{K}$ be a convex set and $f : \mathcal{K} \to \mathbb{R}$ be a convex function. Suppose f is 1-strongly convex at x_0. Then f^*, the Legendre transform of f, is 1-strongly smooth at $y_0 = \nabla f(x_0)$.*

Proof. Notice first that for any pair of convex functions $f, g : \mathcal{K} \to \mathbb{R}$, the fact that $f(x_0) \geq g(x_0)$ for some $x_0 \in \mathcal{K}$ implies that $f^*(y_0) \leq g^*(y_0)$ for $y_0 = \nabla f(x_0)$.

Now, f being 1-strongly convex at x_0 means that $f(x) \geq h(x) = f(x_0) + g_0^\top(x - x_0) + \frac{\sigma}{2}\|x - x_0\|_2^2$. Thus, it suffices to show that $h^*(y) = f^*(y_0) + x_0^\top(y - y_0) + \frac{1}{2}\|y - y_0\|_2^2$, since $x_0 = \nabla(h^*)(y_0)$.

To see this, we can compute that

$$\begin{aligned}
h^*(y) &= \max_x y^\top x - h(x) \\
&= y^\top(y - y_0 + x_0) - h(x) \\
&\quad (\text{max attained at } y_0 + (x - x_0) = \nabla h(x) = y) \\
&= y^\top(y - y_0 + x_0) \\
&\quad - \left[f(x_0) + y_0^\top(x - x_0) + \frac{1}{2}\|x - x_0\|_2^2\right] \\
&= \frac{1}{2}\|y - y_0\|_2^2 + y^\top x_0 - f(x_0) \\
&= -f(x_0) + x_0^\top y_0 + x_0^\top(y - y_0) + \frac{1}{2}\|y - y_0\|_2^2 \\
&= f^*(y_0) + x_0^\top(y - y_0) + \frac{1}{2}\|y - y_0\|_2^2
\end{aligned}$$

□

Theorem 2 (AO-FTRL-Gen). *Let $\{r_t\}$ be a sequence of non-negative functions, and let $\tilde{g}_t$ be the learner's estimate of g_t given the history of functions $f_1, \dots, f_{t-1}$ and points $x_1, \dots, x_{t-1}$. Assume further that the function $h_{0:t} \colon x \mapsto g_{1:t}^\top x + \tilde{g}_{t+1}^\top x + r_{0:t}(x)$ is 1-strongly convex with respect to some norm $\|\cdot\|_{(t)}$ (i.e. $r_{0:t}$ is 1-strongly convex with respect to $\|\cdot\|_{(t)}$). Then, the following regret bound holds for AO-FTRL (Algorithm 1):*

$$\sum_{t=1}^T f_t(x_t) - f_t(x) \leq r_{0:T-1}(x) + \sum_{t=1}^T \|g_t - \tilde{g}_t\|_{(t-1),*}^2$$

Proof. Recall that $x_{t+1} = \operatorname{argmin}_x x^\top(g_{1:t} + \tilde{g}_{t+1}) + r_{0:t}(x)$, and let $y_t = \operatorname{argmin}_x x^\top g_{1:t} + r_{0:t-1}(x)$. Then by convexity,

$$\begin{aligned}
\sum_{t=1}^T f_t(x_t) - f_t(x) &\leq \sum_{t=1}^T g_t^\top(x_t - x) \\
&= \sum_{t=1}^T (g_t - \tilde{g}_t)^\top(x_t - y_t) \\
&\qquad + \tilde{g}_t^\top(x_t - y_t) + g_t^\top(y_t - x)
\end{aligned}$$

Now, we first show via induction that $\forall x \in \mathcal{K}$, the following holds:

$$\sum_{t=1}^T \tilde{g}_t^\top(x_t - y_t) + g_t^\top y_t \leq \sum_{t=1}^T g_t^\top x + r_{0:T-1}(x).$$

For $T = 1$, the fact that $r_t \geq 0$, $\tilde{g}_1 = 0$, and the definition of y_t imply the result.

Now suppose the result is true for time T. Then

$$\begin{aligned}
&\sum_{t=1}^{T+1} \tilde{g}_t^\top(x_t - y_t) + g_t^\top y_t \\
&= \left[\sum_{t=1}^T \tilde{g}_t^\top(x_t - y_t) + g_t^\top y_t\right] \\
&\qquad + \tilde{g}_{T+1}^\top(x_{T+1} - y_{T+1}) + g_{T+1}^\top y_{T+1} \\
&\leq \left[\sum_{t=1}^T g_t^\top x_{T+1} + r_{0:T-1}(x_{T+1})\right] \\
&\qquad + \tilde{g}_{T+1}^\top(x_{T+1} - y_{T+1}) + g_{T+1}^\top y_{T+1} \\
&\quad (\text{by the induction hypothesis for } x = x_{T+1}) \\
&\leq \left[(g_{1:T} + \tilde{g}_{T+1})^\top x_{T+1} + r_{0:T}(x_{T+1})\right] \\
&\qquad + \tilde{g}_{T+1}^\top(-y_{T+1}) + g_{T+1}^\top y_{T+1} \\
&\quad (\text{since } r_t \geq 0, \forall t) \\
&\leq \left[(g_{1:T} + \tilde{g}_{T+1})^\top y_{T+1} + r_{0:T}(y_{T+1})\right] \\
&\qquad + \tilde{g}_{T+1}^\top(-y_{T+1}) + g_{T+1}^\top y_{T+1} \\
&\quad (\text{by definition of } x_{T+1}) \\
&\leq g_{1:T+1}^\top y + r_{0:T}(y), \text{ for any } y. \\
&\quad (\text{by definition of } y_{T+1})
\end{aligned}$$

Thus, we have that $\sum_{t=1}^T f_t(x_t) - f_t(x) \leq r_{0:T-1}(x) + \sum_{t=1}^T (g_t - \tilde{g}_t)^\top(x_t - y_t)$ and it suffices to bound $\sum_{t=1}^T (g_t - \tilde{g}_t)^\top(x_t - y_t)$. By duality again, one can immediately get $(g_t - \tilde{g}_t)^\top(x_t - y_t) \leq \|g_t - \tilde{g}_t\|_{(t-1),*}\|x_t - y_t\|_{(t-1)}$. To bound $\|x_t - y_t\|_{(t)}$ in terms of the gradient, recall first that

$$\begin{aligned}
x_t &= \operatorname*{argmin}_x h_{0:t-1}(x) \\
y_t &= \operatorname*{argmin}_x h_{0:t-1}(x) + (g_t - \hat{g}_t)^\top x.
\end{aligned}$$

The fact that $r_{0:t-1}(x)$ is 1-strongly convex with respect to the norm $\|\cdot\|_{(t-1)}$ implies that $h_{0:t-1}$ is as well. In particular, it is strongly convex at the points x_t and y_t. But, this then implies that the conjugate function is smooth at $\nabla(h_{0:t-1})(x_t)$ and $\nabla(h_{0:t-1})(y_t)$, so that

$$\begin{aligned}&\|\nabla(h^*_{0:t-1})(-(g_t-\tilde{g}_t))\\ &\quad -\nabla(h^*_{0:t-1})(0)\|_{(t)} \leq \|g_t-\tilde{g}_t\|_{(t-1),*}\end{aligned}$$

Since $\nabla(h^*_{0:t-1})(-(g_t-\tilde{g}_t)) = y_t$ and $\nabla(h^*_{0:t-1})(0) = x_t$, we have that $\|x_t-y_t\|_{(t-1)} \leq \|g_t-\tilde{g}_t\|_{(t-1),*}$.

□

Theorem 3 (CAO-FTRL-Prox)**.** *Let $\{r_t\}$ be a sequence of proximal non-negative functions, such that $\text{argmin}_{x\in\mathcal{K}} r_t(x) = x_t$, and let $\tilde{g}_t$ be the learner's estimate of g_t given the history of functions $f_1,\ldots,f_{t-1}$ and points $x_1,\ldots,x_{t-1}$. Let $\{\psi_t\}_{t=1}^{\infty}$ be a sequence of non-negative convex functions, such that $\psi_1(x_1)=0$. Assume further that the function $h_{0:t}: x\mapsto g_{1:t}^\top x + \tilde{g}_{t+1}^\top x + r_{0:t}(x) + \psi_{1:t+1}(x)$ is 1-strongly convex with respect to some norm $\|\cdot\|_{(t)}$. Then the following regret bounds hold for CAO-FTRL (Algorithm 2):*

$$\begin{aligned}&\sum_{t=1}^T f_t(x_t)-f_t(x)\\ &\quad\leq \psi_{1:T-1}(x)+r_{0:T-1}(x)+\sum_{t=1}^T\|g_t-\tilde{g}_t\|^2_{(t-1),*}\\ &\sum_{t=1}^T [f_t(x_t)+\psi_t(x_t)]-[f_t(x)+\psi_t(x)]\\ &\quad\leq r_{0:T}(x)+\sum_{t=1}^T\|g_t-\tilde{g}_t\|^2_{(t),*}\,.\end{aligned}$$

Proof. For the first regret bound, define the auxiliary regularization functions $\tilde{r}_t(x) = r_t(x)+\psi_t(x)$, and apply Theorem 2 to get

$$\begin{aligned}&\sum_{t=1}^T f_t(x_t)-f_t(x)\\ &\quad\leq \tilde{r}_{0:T-1}(x)+\sum_{t=1}^T\|g_t-\tilde{g}_t\|^2_{(t-1),*}\\ &\quad= \psi_{1:T-1}(x)+r_{0:T-1}(x)+\sum_{t=1}^T\|g_t-\tilde{g}_t\|^2_{(t-1),*}\end{aligned}$$

Notice that while r_t is proximal, $\tilde{r}_t$, in general, is not, and so we must apply the theorem with general regularizers instead of the one with proximal regularizers.

For the second regret bound, we can follow the prescription of Theorem 1 while keeping track of the additional composite terms:

Recall that $x_{t+1} = \text{argmin}_x x^\top(g_{1:t}+\tilde{g}_{t+1}) + r_{0:t+1}(x)+\psi_{1:t+1}(x)$, and let $y_t = \text{argmin}_x x^\top g_{1:t} + r_{0:t}(x)+\psi_{1:t}(x)$.

We can compute that:

$$\begin{aligned}&\sum_{t=1}^T f_t(x_t)+\alpha_t\psi(x_t)-[f_t(x)+\psi_t(x)]\\ &\quad\leq \sum_{t=1}^T g_t^\top(x_t-x)+\psi_t(x_t)-\psi_t(x)\\ &\quad= \sum_{t=1}^T (g_t-\tilde{g}_t)^\top(x_t-y_t)\\ &\qquad+\tilde{g}_t^\top(x_t-y_t)+g_t^\top(y_t-x)+\psi_t(x_t)-\psi_t(x)\end{aligned}$$

Similar to before, we show via induction that $\forall x\in\mathcal{K}$, $\sum_{t=1}^T \tilde{g}_t^\top(x_t-y_t)+g_t^\top y_t+\psi_t(x_t) \leq r_{0:T}(x)+\sum_{t=1}^T g_t^\top x+\psi_t(x)$.

For $T=1$, the fact that $r_t\geq 0$, $\hat{g}_1=0$, $\psi_1(x_1)=0$, and the definition of y_t imply the result.

Now suppose the result is true for time T. Then

$$\sum_{t=1}^{T+1} \tilde{g}_t^\top (x_t - y_t) + g_t^\top y_t + \psi_t(x_t)$$
$$= \left[\sum_{t=1}^{T} \tilde{g}_t^\top (x_t - y_t) + g_t^\top y_t + \psi_t(x_t)\right] + \tilde{g}_{T+1}^\top (x_{T+1} - y_{T+1}) + g_{T+1}^\top y_{T+1} + \psi_{T+1}(x_{T+1})$$
$$\leq \left[\sum_{t=1}^{T} g_t^\top x_{T+1} + r_{0:T}(x_{T+1}) + \psi_t(x_{T+1})\right] + \tilde{g}_{T+1}^\top (x_{T+1} - y_{T+1}) + g_{T+1}^\top y_{T+1} + \psi_{T+1}(x_{T+1})$$
(by the induction hypothesis for $x = x_{T+1}$)
$$\leq (g_{1:T} + \tilde{g}_{T+1})^\top x_{T+1} + r_{0:T+1}(x_{T+1}) + \psi_t(x_{T+1}) + \tilde{g}_{T+1}^\top (-y_{T+1}) + g_{T+1}^\top y_{T+1} + \psi_{T+1}(x_{T+1})$$
(since $r_t \geq 0$, $\forall t$)
$$\leq (g_{1:T} + \tilde{g}_{T+1})^\top y_{T+1} + r_{0:T+1}(y_{T+1}) + \psi_t(y_{T+1}) + \tilde{g}_{T+1}^\top (-y_{T+1}) + g_{T+1}^\top y_{T+1} + \psi_{T+1}(y_{T+1})$$
(by definition of x_{T+1})
$$\leq g_{1:T+1}^\top y + r_{0:T+1}(y) + \psi_{1:T+1}(y), \text{ for any } y$$
(by definition of y_{T+1})

Thus, we have that

$$\sum_{t=1}^{T} f_t(x_t) + \psi_t(x_t) - [f_t(x) + \psi_t(x)] \leq r_{0:T}(x) + \sum_{t=1}^{T} (g_t - \tilde{g}_t)^\top (x_t - y_t),$$

and we can bound the sum in the same way as before, since the strong convexity properties of $h_{0:t}$ are retained due to the convexity of ψ_t.

□

Theorem 6 (CAO-FTRL-Gen). *Let $\{r_t\}$ be a sequence of non-negative functions, and let $\tilde{g}_t$ be the learner's estimate of g_t given the history of functions $f_1, \ldots, f_{t-1}$ and points $x_1, \ldots, x_{t-1}$. Let $\{\psi_t\}_{t=1}^{\infty}$ be a sequence of non-negative convex functions such that $\psi_1(x_1) = 0$. Assume further that the function $h_{0:t} : x \mapsto g_{1:t}^\top x + \tilde{g}_{t+1}^\top x + r_{0:t}(x) + \psi_{1:t+1}(x)$ is 1-strongly convex with respect to some norm $\|\cdot\|_{(t)}$. Then, the following regret bound holds for CAO-FTRL (Algorithm 2):*

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x) \leq \psi_{1:T-1}(x) + r_{0:T-1}(x) + \sum_{t=1}^{T} \|g_t - \tilde{g}_t\|_{(t-1),*}^2$$
$$\sum_{t=1}^{T} f_t(x_t) + \psi_t(x_t) - [f_t(x) + \psi_t(x)] \leq r_{0:T-1}(x) + \sum_{t=1}^{T} \|g_t - \tilde{g}_t\|_{(t),*}^2 .$$

Proof. For the first regret bound, define the auxiliary regularization functions $\tilde{r}_t(x) = r_t(x) + \alpha_t \psi(x)$, and apply Theorem 2 to get

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x) \leq \tilde{r}_{0:T-1}(x) + \sum_{t=1}^{T} \|g_t - \hat{g}_t\|_{(t),*}^2 = \psi_{1:T-1}(x) + r_{0:T-1}(x) + \sum_{t=1}^{T} \|g_t - \hat{g}_t\|_{(t-1),*}^2$$

For the second bound, we can proceed as in the original proof, but now keep track of the additional composite terms.

Recall that $x_{t+1} = \operatorname{argmin}_x x^\top (g_{1:t} + \tilde{g}_{t+1}) + r_{0:t}(x) + \psi_{1:t+1}(x)$, and let $y_t = \operatorname{argmin}_x x^\top g_{1:t} + r_{0:t-1}(x) + \psi_{1:t}(x)$. Then

$$\sum_{t=1}^{T} f_t(x_t) + \psi_t(x_t) - f_t(x) - \psi_t(x) \leq \sum_{t=1}^{T} g_t^\top (x_t - x) + \psi_t(x_t) - \psi_t(x)$$
$$= \sum_{t=1}^{T} (g_t - \tilde{g}_t)^\top (x_t - y_t) + \tilde{g}_t^\top (x_t - y_t) + g_t^\top (y_t - x) + \psi_t(x_t) - \psi_t(x)$$

Now, we show via induction that $\forall x \in \mathcal{K}$, $\sum_{t=1}^{T} \tilde{g}_t^\top (x_t - y_t) + g_t^\top y_t + \alpha_t \psi(x_t) \leq \sum_{t=1}^{T} g_t^\top x + \psi_t(x) + r_{0:T-1}(x)$.

For $T = 1$, the fact that $r_t \geq 0$, $\hat{g}_1 = 0$, $\psi_1(x_1) = 0$, and the definition of y_t imply the result.

Now suppose the result is true for time T. Then

$$\begin{aligned}
&\sum_{t=1}^{T+1} \tilde{g}_t^\top (x_t - y_t) + g_t^\top y_t + \psi_t(x_t) \\
&= \left[\sum_{t=1}^{T} \tilde{g}_t^\top (x_t - y_t) + g_t^\top y_t + \psi_t(x_t)\right] \\
&\qquad + \tilde{g}_{T+1}^\top (x_{T+1} - y_{T+1}) + g_{T+1}^\top y_{T+1} \\
&\qquad + \psi_{T+1}(x_{T+1}) \\
&\le \left[\sum_{t=1}^{T} g_t^\top x_{T+1} + r_{0:T-1}(x_{T+1}) + \psi_t(x_{T+1})\right] \\
&\qquad + \tilde{g}_{T+1}^\top (x_{T+1} - y_{T+1}) + g_{T+1}^\top y_{T+1} \\
&\qquad + \psi_{T+1}(x_{T+1}) \\
&\qquad \text{(by the induction hypothesis for } x = x_{T+1}) \\
&\le \left[(g_{1:T} + \tilde{g}_{T+1})^\top x_{T+1} + r_{0:T}(x_{T+1}) + \psi_t(x_{T+1})\right] \\
&\qquad + \tilde{g}_{T+1}^\top (-y_{T+1}) + g_{T+1}^\top y_{T+1} \\
&\qquad + \psi_{T+1}(x_{T+1}) \\
&\qquad \text{(since } r_t \ge 0, \forall t) \\
&\le g_{1:T+1}^\top y_{T+1} + \tilde{g}_{T+1}^\top y_{T+1} + r_{0:T}(y_{T+1}) \\
&\qquad + \psi_{1:T+1}(y_{T+1}) \\
&\qquad + \tilde{g}_{T+1}^\top (-y_{T+1}) + g_{T+1}^\top y_{T+1} \\
&\qquad \text{(by definition of } x_{T+1}) \\
&\le g_{1:T+1}^\top y + r_{0:T}(y) + \psi_{1:T+1}(y), \text{ for any } y \\
&\quad \text{(by definition of } y_{T+1})
\end{aligned}$$

Thus, we have that $\sum_{t=1}^{T} f_t(x_t) + \psi_t(x_t) - f_t(x) - \psi_t(x) \le r_{0:T-1}(x) + \sum_{t=1}^{T} (g_t - \tilde{g}_t)^\top (x_t - y_t)$ and the remainder follows as in the non-composite setting since the strong convexity properties are retained.

□

7 Proofs for Section 2.2.1

The following lemma is central to the derivation of regret bounds for many algorithms employing adaptive regularization. Its proof, via induction, can be found in Auer et al (2002).

Lemma 2. *Let $\{a_j\}_{j=1}^\infty$ be a sequence of non-negative numbers. Then $\sum_{j=1}^{t} \frac{a_j}{\sum_{k=1}^{j} a_k} \le 2\sqrt{\sum_{j=1}^{t} a_j}$.*

Corollary 2 (AO-GD). *Let $\mathcal{K} \subset \times_{i=1}^{n} [-R_i, R_i]$ be an n-dimensional rectangle, and denote $\Delta_{s,i} = \sqrt{\sum_{a=1}^{s} (g_{a,i} - \tilde{g}_{a,i})^2}$. Set*

$$r_{0:t} = \sum_{i=1}^{n} \sum_{s=1}^{t} \frac{\Delta_{s,i} - \Delta_{s-1,i}}{2R_i} (x_i - x_{s,i})^2.$$

Then, if we use the martingale-type gradient prediction $\tilde{g}_{t+1} = g_t$, the following regret bound holds:

$$\mathit{Reg}_T(x) \le 4 \sum_{i=1}^{n} R_i \sqrt{\sum_{t=1}^{T} (g_{t,i} - g_{t-1,i})^2}.$$

Moreover, this regret bound is nearly a posteriori optimal over a family of quadratic regularizers :

$$\begin{aligned}
&\max_i R_i \sum_{i=1}^{n} \sqrt{\sum_{t=1}^{T} (g_{t,i} - g_{t-1,i})^2} \\
&= \max_i R_i \sqrt{n \inf_{s \succcurlyeq 0, \langle s, 1 \rangle \le n} \sum_{t=1}^{T} \|g_t - g_{t-1}\|^2_{diag(s)^{-1}}}
\end{aligned}$$

Proof. $r_{0:t}$ is 1-strongly convex with respect to the norm:

$$\|x\|_{(t)}^2 = \sum_{i=1}^{n} \frac{\sqrt{\sum_{a=1}^{t} (g_{a,i} - \tilde{g}_{a,i})^2}}{R_i} x_i^2,$$

which has corresponding dual norm:

$$\|x\|_{(t),*}^2 = \sum_{i=1}^{n} \frac{R_i}{\sqrt{\sum_{a=1}^{t} (g_{a,i} - \tilde{g}_{a,i})^2}} x_i^2.$$

By the choice of this regularization, the prediction $\tilde{g}_t = g_{t-1}$, and Theorem 3, the following holds:

$$\begin{aligned}
&\mathrm{Reg}_T(\mathcal{A}, x) \\
&\le \sum_{i=1}^{n} \sum_{s=1}^{T} \\
&\qquad \frac{\sqrt{\sum_{a=1}^{s} (g_{a,i} - \tilde{g}_{a,i})^2} - \sqrt{\sum_{a=1}^{s-1} (g_{a,i} - \tilde{g}_{a,i})^2}}{2R_i} \\
&\qquad (x_i - x_{s,i})^2 \\
&\quad + \sum_{t=1}^{T} \|g_t - g_{t-1}\|_{(t),*}^2 \\
&= \sum_{i=1}^{n} 2R_i \sqrt{\sum_{t=1}^{T} (g_{t,i} - g_{t-1,i})^2} \\
&\quad + \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{R_i (g_{t,i} - g_{t-1,i})^2}{\sqrt{\sum_{a=1}^{t} (g_{a,i} - g_{a-1,i})^2}} \\
&\le \sum_{i=1}^{n} 2R_i \sqrt{\sum_{t=1}^{T} (g_{t,i} - g_{t-1,i})^2} \\
&\quad + \sum_{i=1}^{n} 2R_i \sqrt{\sum_{t=1}^{T} (g_{t,i} - g_{t-1,i})^2} \\
&\quad \text{by Lemma 2}
\end{aligned}$$

The last statement follows from the fact that

$$\inf_{s \succcurlyeq 0, \langle s, 1, \rangle \leq n} \sum_{t=1}^{T} \sum_{i=1}^{n} \frac{g_{t,i}^2}{s_i} = \frac{1}{n} \left(\sum_{i=1}^{n} \|g_{1:T}, i\|_2 \right)^2,$$

since the infimum on the left hand side is attained when $s_i \propto \|g_{1:T,i}\|_2$. □

8 Proofs for Section 3

Theorem 4 (CAOS-FTRL-Prox). *Let $\{r_t\}$ be a sequence of proximal non-negative functions, such that $\operatorname{argmin}_{x \in \mathcal{K}} r_t(x) = x_t$, and let $\tilde{g}_t$ be the learner's estimate of $\hat{g}_t$ given the history of noisy gradients $\hat{g}_1, \ldots, \hat{g}_{t-1}$ and points $x_1, \ldots, x_{t-1}$. Let $\{\psi_t\}_{t=1}^{\infty}$ be a sequence of non-negative convex functions, such that $\psi_1(x_1) = 0$. Assume further that the function*

$$h_{0:t}(x) = \hat{g}_{1:t}^\top x + \tilde{g}_{t+1}^\top x + r_{0:t}(x) + \psi_{1:t+1}(x)$$

is 1-strongly convex with respect to some norm $\|\cdot\|_{(t)}$. Then, the update $x_{t+1} = \operatorname{argmin}_x h_{0:t}(x)$ of Algorithm 3 yields the following regret bounds:

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) - f_t(x)\right] \leq \mathbb{E}\left[\psi_{1:T-1}(x) + r_{0:T-1}(x) + \sum_{t=1}^{T} \|\hat{g}_t - \tilde{g}_t\|_{(t-1),*}^2\right]$$

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) + \psi_t(x_t) - f_t(x) - \alpha_t \psi_t(x)\right] \leq \mathbb{E}\left[r_{0:T}(x) + \sum_{t=1}^{T} \|\hat{g}_t - \tilde{g}_t\|_{(t),*}^2\right].$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) - f_t(x)\right] &\leq \sum_{t=1}^{T} \mathbb{E}\left[g_t^\top (x_t - x)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}[\hat{g}_t | \hat{g}_1, \ldots, \hat{g}_{t-1}, x_1, \ldots, x_t]^\top (x_t - x)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}[\hat{g}_t^\top (x_t - x) | \hat{g}_1, \ldots, \hat{g}_{t-1}, x_1, \ldots, x_t]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\hat{g}_t^\top (x_t - x)\right]
\end{aligned}$$

This implies that upon taking an expectation, we can freely upper bound the difference $f_t(x_t) - f_t(x)$ by the noisy linearized estimate $\hat{g}_t^\top (x_t - x)$. After that, we can apply Algorithm 2 on the gradient estimates to get the bounds:

$$\mathbb{E}\left[\sum_{t=1}^{T} \hat{g}_t^\top (x_t - x)\right] \leq \mathbb{E}\left[\psi_{1:T-1}(x) + r_{0:T-1}(x) + \sum_{t=1}^{T} \|\hat{g}_t - \tilde{g}_t\|_{(t-1),*}^2\right]$$

$$\mathbb{E}\left[\sum_{t=1}^{T} \hat{g}_t^\top (x_t - x) + \psi_t(x_t) - \psi_t(x)\right] \leq \mathbb{E}\left[r_{0:T}(x) + \sum_{t=1}^{T} \|\hat{g}_t - \tilde{g}_t\|_{(t),*}^2\right]$$

□

Theorem 7 (CAOS-FTRL-Gen). *Let $\{r_t\}$ be a sequence of non-negative functions, and let $\tilde{g}_t$ be the learner's estimate of $\hat{g}_t$ given the history of noisy gradients $\hat{g}_1, \ldots, \hat{g}_{t-1}$ and points $x_1, \ldots, x_{t-1}$. Let $\{\psi_t\}_{t=1}^{\infty}$ be a sequence of non-negative convex functions, such that $\psi_1(x_1) = 0$. Assume furthermore that the function $h_{0:t}(x) = \hat{g}_{1:t}^\top x + \tilde{g}_{t+1}^\top x + r_{0:t}(x) + \psi_{1:t+1}(x)$ is 1-strongly convex with respect to some norm $\|\cdot\|_{(t)}$. Then, the update $x_{t+1} = \operatorname{argmin}_x h_{0:t}(x)$ of Algorithm 3 yields the regret bounds:*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) - f_t(x)\right] \leq \mathbb{E}\left[\psi_{1:T-1}(x) + r_{0:T-1}(x) + \sum_{t=1}^{T} \|\hat{g}_t - \tilde{g}_t\|_{(t-1),*}^2\right]$$

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) + \psi_t(x_t) - f_t(x) - \psi_t(x)\right] \leq \mathbb{E}\left[r_{0:T-1}(x) + \sum_{t=1}^{T} \|\hat{g}_t - \tilde{g}_t\|_{(t-1),*}^2\right]$$

Proof. The argument is the same as for Theorem 4, except that we now apply the bound of Theorem 6 at the end. □

9 Proofs for Section 3.2.1

Theorem 5 (CAO-RCD). *Assume $\mathcal{K} \subset \times_{i=1}^{n} [-R_i, R_i]$. Let i_t be a random variable sampled according to the distribution p_t, and let*

$$\hat{g}_t = \frac{(g_t^\top e_{i_t}) e_{i_t}}{p_{t,i_t}}, \quad \hat{\tilde{g}}_t = \frac{(\tilde{g}_t^\top e_{i_t}) e_{i_t}}{p_{t,i_t}},$$

be the estimated gradient and estimated gradient prediction. Denote $\Delta_{s,i} = \sqrt{\sum_{a=1}^{s}(\hat{g}_{a,i} - \hat{\tilde{g}}_{a,i})^2}$, *and let*

$$r_{0:t} = \sum_{i=1}^{n}\sum_{s=1}^{t}\frac{\Delta_{s,i} - \Delta_{s-1,i}}{2R_i}(x_i - x_{s,i})^2$$

be the adaptive regularization. Then the regret of the resulting algorithm is bounded by:

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) + \alpha_t\psi(x_t) - f_t(x) - \alpha_t\psi(x)\right]$$
$$\leq 4\sum_{i=1}^{n} R_i\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\frac{(g_{t,i} - \tilde{g}_{t,i})^2}{p_{t,i}}\right]}.$$

Proof. We can first compute that

$$\mathbb{E}\left[\hat{g}_t\right] = \mathbb{E}\left[\frac{(g_t^\top e_{i_t})e_{i_t}}{p_{t,i_t}}\right] = \sum_{i=1}^{n}\frac{(g_t^\top e_i)e_i}{p_{t,i}}p_{t,i} = g_t$$

and similarly for the gradient prediction $\tilde{g}_t$.

Now, as in Corollary 2, the choice of regularization ensures us a regret bound of the form:

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) + \alpha_t\psi(x_t) - f_t(x) - \alpha_t\psi(x)\right]$$
$$\leq 4\sum_{i=1}^{n} R_i\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}(\hat{g}_{t,i} - \tilde{g}_{t,i})^2}\right]$$

Moreover, we can compute that:

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}(\hat{g}_{t,i} - \tilde{g}_{t,i})^2}\right] \leq \sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{i_t}[(\hat{g}_{t,i} - \tilde{g}_{t,i})^2]\right]}$$
$$= \sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\frac{(g_{t,i} - \tilde{g}_{t,i})^2}{p_{t,i}}\right]}$$

□

10 Further Discussion for Section 3.2.2

We present here Algorithm 5, a mini-batch version of Algorithm 4, with an accompanying guarantee.

Corollary 6. *Assume* $\mathcal{K} \subset \times_{i=1}^{n}[-R_i, R_i]$. *Let* $\cup_{j=1}^{l}\{\Pi_j\} = \{1, \ldots, n\}$ *be a partition of the functions* f_i, *and let* $e_{\Pi_j} = \sum_{i\in\Pi_j} e_i$. *Denote* $\Delta_{s,i} = \sqrt{\sum_{a=1}^{s}(\hat{g}_{a,i} - \tilde{g}_{a,i})^2}$, *and let* $r_{0:t} = \sum_{i=1}^{n}\sum_{s=1}^{t}\frac{\Delta_{s,i} - \Delta_{s-1,i}}{2R_i}(x_i - x_{s,i})^2$ *be the adaptive regularization.*

Then the regret of Algorithm 5 is bounded by:

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) + \alpha\psi(x_t) - f_t(x) - \alpha\psi(x)\right]$$
$$\leq \sum_{i=1}^{n} 4R_i\sqrt{\sum_{s=1}^{k}\sum_{t=(s-1)(T/k)+1}^{(s-1)(T/k)+T/k}\sum_{a=1}^{l}\frac{\left|\sum_{j\in\Pi_j} g_{t,i}^j - \bar{g}_{s,i}^j\right|^2}{p_{t,a}}}.$$

Moreover, if $\|\nabla f_j\|_\infty \leq L_j \ \forall j$, *then setting* $p_{t,j} = \frac{L_i}{\sum_{j=1}^{m} L_j}$ *yields a worst-case bound of:* $8\sum_{i=1}^{n} R_i\sqrt{T\left(\sum_{j=1}^{m} L_j\right)^2}$.

Algorithm 5 CAOS-Reg-ERM-Epoch-Mini-Batch

Input: scaling constant $\alpha > 0$, composite term ψ, $r_0 = 0$, partitions $\cup_{j=1}^{l}\{\Pi_j\} = \{1, \ldots, m\}$.
Initialize: initial point $x_1 \in \mathcal{K}$, distribution p_1 over $\{1, \ldots, l\}$.
Sample j_1 according to p_1, and set $t = 1$.
for $s = 1, \ldots, k$: **do**
  Compute $\bar{g}_s^j = \nabla f_j(x_1)\ \forall j \in \{1, \ldots, m\}$.
  for $a = 1, \ldots, T/k$: **do**
    If $T \mod k = 0$, compute $g^j = \nabla f_j(x_t)\ \forall j$.
    Set $\hat{g}_t = \frac{\sum_{j\in\Pi_{j_t}} g_t^j}{p_{t,j_t}}$, and construct $r_t \geq 0$.
    Sample $j_{t+1} \sim p_{t+1}$.
    Set $\tilde{g}_{t+1} = \frac{\sum_{j\in\Pi_{j_t}} \bar{g}_s^j}{p_{t,j_t}}$.
    Update $x_{t+1} = \operatorname{argmin}_{x\in\mathcal{K}} \hat{g}_{1:t}^\top x + \tilde{g}_{t+1}^\top x + r_{0:t}(x) + (t+1)\alpha\psi(x)$ and $t = t + 1$.
  end for
end for

A similar approach to Regularized ERM was developed independently by (Zhao and Zhang, 2014). However, the one here improves upon that algorithm through the incorporation of adaptive regularization, optimistic gradient predictions, and the fact that we do not assume higher regularity conditions such as strong convexity for our loss functions.