A. proofs

A.1. proof of Lemma 1

Proof. For any t and $i \leq t$,

$$\begin{aligned}
\lim_{\beta_1 \to 1} w_{i,t} &= \lim_{\beta_1 \to 1} \frac{(1-\beta_1)\beta_1^{t-i}}{1-\beta_1^t} \\
&= \lim_{\beta_1 \to 1} \frac{1-\beta_1}{1-\beta_1^t} \lim_{\beta_1 \to 1} \beta_1^{t-i} \\
&= \lim_{\beta_1 \to 1} \frac{1-\beta_1}{1-\beta_1^t} \\
&= \lim_{\beta_1 \to 1} \frac{1}{t\beta_1^{t-1}} \\
&= \frac{1}{t}.
\end{aligned}$$

here, the second equality holds by the limit properties. The last second equality holds by L'Hôpital's rule. □

A.2. proof of Proposition 1

Proof. from (11):

$$\begin{aligned}
\sum_{i=1}^{t} w_{i,t}Q_i - \sum_{i=1}^{t-1} w_{i,t-1}Q_i &= \sum_{i=1}^{t-1}(w_{i,t} - w_{i,t-1})Q_i + w_{t,t}Q_t \\
&= \sum_{i=1}^{t-1} \left(\frac{(1-\beta_1)\beta_1^{t-i}}{1-\beta_1^t} - \frac{(1-\beta_1)\beta_1^{t-1-i}}{1-\beta_1^{t-1}} \right) Q_i + w_{t,t}Q_t \\
&= \sum_{i=1}^{t-1} \left(\frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1^t} - 1 \right) w_{i,t-1}Q_i + w_{t,t}Q_t \\
&= \frac{\beta_1 - 1}{1-\beta_1^t} \sum_{i=1}^{t-1} w_{i,t-1}Q_i + w_{t,t}Q_t \\
&= -w_{t,t} \sum_{i=1}^{t-1} w_{i,t-1}Q_i + w_{t,t}Q_t.
\end{aligned}$$

Rearranging, we obtain

$$\begin{aligned}
w_{t,t}Q_t &= \sum_{i=1}^{t} w_{i,t}Q_i - (1-w_{t,t}) \sum_{i=1}^{t-1} w_{i,t-1}Q_i \\
&= \text{diag}\left(\frac{d_t}{1-\beta_1^t} \right) - (1-w_{t,t})\text{diag}\left(\frac{d_{t-1}}{1-\beta_1^{t-1}} \right) \\
&= \text{diag}\left(\frac{d_t - \beta_1 d_{t-1}}{1-\beta_1^t} \right).
\end{aligned}$$

Thus, $Q_t = \text{diag}\left(\frac{d_t - \beta_1 d_{t-1}}{1-\beta_1} \right)$. □

A.3. proof of Proposition 2

Proof. (13) can be rewritten as (apart from a constant)

$$\min_{\theta \in \Theta} \left\langle \sum_{i=1}^{t} w_{i,t} \left(g_i - \frac{\sigma_i}{1-\beta_1}\theta_{i-1} \right), \theta \right\rangle + \frac{1}{2}\|\theta\|^2_{\text{diag}\left(\frac{d_t}{1-\beta_1^t}\right)}. \tag{17}$$

Let $z_t = (1-\beta_1^t)\sum_{i=1}^t w_{i,t}\left(g_i - \frac{\sigma_i}{1-\beta_1}\theta_{i-1}\right)$. By (10), we have a simple recursive update rule:

$$\begin{aligned} z_t &= \beta_1 z_{t-1} + (1-\beta_1)\left(g_t - \frac{\sigma_t}{1-\beta_1}\theta_{t-1}\right) \\ &= \beta_1 z_{t-1} + (1-\beta_1)g_t - \sigma_t\theta_{t-1}. \end{aligned}$$

Substituting z_t into (17), we have

$$\min_{\theta\in\Theta}\left\langle \frac{z_t}{1-\beta_1^t}, \theta\right\rangle + \frac{1}{2}\|\theta\|^2_{\text{diag}\left(\frac{d_t}{1-\beta_1^t}\right)}.$$

Rearranging, we obtain

$$\min_{\theta\in\Theta}\frac{1}{2}\|\theta + z_t/d_t\|^2_{\text{diag}\left(\frac{d_t}{1-\beta_1^t}\right)},$$

with optimal solution $\Pi_\Theta^{\text{diag}\left(d_t/(1-\beta_1^t)\right)}(-z_t/d_t)$. □

A.4. proof of Proposition 3

Proof. When $\beta_1 = 0$, we have $w_{t,t} = 1$ and $w_{i,t} = 0$ for all $i < t$. Thus, $\sigma_t = d_t$, and (13) reduces to:

$$\min_{\theta\in\Theta}\langle g_t, \theta\rangle + \frac{1}{2}\|\theta - \theta_{t-1}\|^2_{\text{diag}\left(\frac{1}{\eta_t}\left(\sqrt{\frac{v_t}{1-\beta_2^t}}+\epsilon_t\mathbf{1}\right)\right)},$$

We can rewrite above as

$$\min_{\theta\in\Theta}\frac{1}{2}\left\|\theta - \left(\theta_{t-1} - \frac{\eta_t}{\sqrt{\frac{v_t}{1-\beta_2^t}}+\epsilon_t\mathbf{1}}g_t\right)\right\|^2_{\text{diag}\left(\frac{1}{\eta_t}\left(\sqrt{\frac{v_t}{1-\beta_2^t}}+\epsilon_t\mathbf{1}\right)\right)},$$

with optimal solution

$$\Pi_\Theta^{\text{diag}(d_t/(1-\beta_1^t))}\left(\theta_{t-1} - \text{diag}\left(\frac{\eta_t}{\sqrt{\frac{v_t}{1-\beta_2^t}}+\epsilon_t\mathbf{1}}\right)g_t\right). \tag{18}$$

analogous to (10) and Lemma 1,

$$\lim_{\beta_2\to 1}\frac{v_t}{1-\beta_2^t} = \lim_{\beta_2\to 1}\sum_{i=1}^t\frac{(1-\beta_2)\beta_2^{t-i}}{1-\beta_2^t}g_i^2 = \frac{1}{t}\sum_{i=1}^t g_i^2. \tag{19}$$

Combining with $\eta_t = \eta/\sqrt{t}$ and $\epsilon_t = \epsilon/\sqrt{t}$, we obtain

$$\lim_{\beta_2\to 1}\frac{\eta_t}{\sqrt{\frac{v_t}{1-\beta_2^t}}+\epsilon_t\mathbf{1}} = \frac{\eta}{\sqrt{g_{1:t}^2}+\epsilon\mathbf{1}},$$

and (18) reduces to below

$$\Pi_\Theta^{\text{diag}((\sqrt{g_{1:t}^2}+\epsilon\mathbf{1})/\eta)}\left(\theta_{t-1} - \text{diag}\left(\frac{\eta}{\sqrt{g_{1:t}^2}+\epsilon\mathbf{1}}\right)g_t\right),$$

□

A.5. proof of Proposition 4

Proof. When $\beta_1 \to 1$, we have

$$\begin{aligned}\lim_{\beta_1\to 1}\frac{\sigma_t}{1-\beta_1} &= \lim_{\beta_1\to 1}\left[\frac{d_t}{1-\beta_1}-\frac{\beta_1 d_{t-1}}{1-\beta_1}\right]\\ &= \lim_{\beta_1\to 1}\left[\frac{1-\beta_1^t}{1-\beta_1}\frac{\sqrt{\frac{v_t}{1-\beta_2^t}}+\epsilon_t\mathbf{1}}{\eta_t}-\frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1}\frac{\sqrt{\frac{v_{t-1}}{1-\beta_2^{t-1}}}+\epsilon_{t-1}\mathbf{1}}{\eta_{t-1}}\right]\\ &= t\frac{\sqrt{\frac{v_t}{1-\beta_2^t}}+\epsilon_t\mathbf{1}}{\eta_t}-(t-1)\frac{\sqrt{\frac{v_{t-1}}{1-\beta_2^{t-1}}}+\epsilon_{t-1}\mathbf{1}}{\eta_{t-1}}.\end{aligned}$$

Substituting this into (13), we obtain

$$\min_{\theta\in\Theta}\sum_{i=1}^{t}\left(\langle g_i,\theta\rangle+\frac{1}{2}\|\theta-\theta_{i-1}\|^2_{\text{diag}(m_i)}\right), \tag{20}$$

where $m_i = \frac{t}{\eta_t}\left(\sqrt{\frac{v_t}{1-\beta_2^t}}+\epsilon_t\mathbf{1}\right)-\frac{t-1}{\eta_{t-1}}\left(\sqrt{\frac{v_{t-1}}{1-\beta_2^{t-1}}}+\epsilon_{t-1}\mathbf{1}\right)$.

Combining with $\eta_t = \eta\sqrt{t}$, $\epsilon_t = \epsilon/\sqrt{t}$, and (19), we further obtain

$$\begin{aligned}\lim_{\beta_2\to 1} m_i &= \lim_{\beta_2\to 1} t\frac{\sqrt{\frac{v_t}{1-\beta_2^t}}+\epsilon_t\mathbf{1}}{\eta_t}-(t-1)\frac{\sqrt{\frac{v_{t-1}}{1-\beta_2^{t-1}}}+\epsilon_{t-1}\mathbf{1}}{\eta_{t-1}} = \frac{\sqrt{g_{1:t}^2}+\epsilon\mathbf{1}}{\eta}-\frac{\sqrt{g_{1:t-1}^2}+\epsilon\mathbf{1}}{\eta}\\ &= \frac{\sqrt{g_{1:t}^2}-\sqrt{g_{1:t-1}^2}}{\eta}.\end{aligned}$$

Substituting back into (20), we recover FTRL with adaptive learning rate. by using the equivalence theorem in (McMahan, 2011), we obtain $\theta_t \leftarrow \theta_{t-1}-\text{diag}\left(\frac{\eta}{\sqrt{g_{1:t}^2}+\epsilon\mathbf{1}}\right)g_t$. □

A.6. proof of Theorem 1

Proof. Note that $w_{i,t} = \frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1^t}w_{i,t-1}$. with $\Theta = \mathbb{R}^d$, consider the first term in the objective of (17): with z_t defined in proposition 2

$$\begin{aligned}&\left\langle\sum_{i=1}^{t}w_{i,t}\left(g_i-\frac{\sigma_i}{1-\beta_1}\theta_{i-1}\right),\theta\right\rangle\\ &= \frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1^t}\left\langle\sum_{i=1}^{t-1}w_{i,t-1}\left(g_i-\frac{\sigma_i}{1-\beta_1}\theta_{i-1}\right),\theta\right\rangle+\left\langle w_{t,t}\left(g_t-\frac{\sigma_t}{1-\beta_1}\theta_{t-1}\right),\theta\right\rangle\\ &= \frac{\beta_1}{1-\beta_1^t}\langle z_{t-1},\theta\rangle+\left\langle w_{t,t}\left(g_t-\frac{\sigma_t}{1-\beta_1}\theta_{t-1}\right),\theta\right\rangle\\ &= -\frac{\beta_1}{1-\beta_1^t}\langle d_{t-1}\theta_{t-1},\theta\rangle+\frac{1-\beta_1}{1-\beta_1^t}\left\langle g_t-\frac{\sigma_t}{1-\beta_1}\theta_{t-1},\theta\right\rangle\\ &= \frac{1}{1-\beta_1^t}\langle(1-\beta_1)g_t-\theta_{t-1}(\sigma_t+\beta_1 d_{t-1}),\theta\rangle\\ &= \frac{1}{1-\beta_1^t}\langle(1-\beta_1)g_t-d_t\theta_{t-1},\theta\rangle,\end{aligned}$$

where the second equality follows from the definition of z_t. The third equality holds since $\Theta = \mathbb{R}^d$ and therefore $\theta_t = -z_t/d_t$ by Proposition 2. Thus, combing this expression into (17), we obtain

$$\min_{\theta\in R^d}\frac{1}{1-\beta_1^t}\langle(1-\beta_1)g_t-d_t\theta_{t-1},\theta\rangle+\frac{1}{2}\|\theta\|^2_{\text{diag}(\frac{d_t}{1-\beta_1^t})},$$

With the definition of d_t, it can be seen that solving above problem (taking gradient w.r.t. θ and setting it to zero) leads to a gradient descent style update rule:

$$\theta_t \leftarrow \theta_{t-1} - \text{diag}\left(\frac{1-\beta_1}{1-\beta_1^t} \frac{\eta_t}{\left(\sqrt{\frac{v_t}{(1-\beta_2^t)}} + \epsilon_t \mathbf{1}\right)} \right) g_t.$$

which concludes the proof. □

A.7. proof of Proposition 5

Proof. Note that (15) can be rewritten as

$$\begin{aligned}
&\min_{\theta\in\Theta} \left(\left\langle \sum_{i=1}^{t} w_{i,t} g_i, \theta \right\rangle + \frac{1}{2}\|\theta - \theta_{t-1}\|^2_{\sum_{i=1}^{t} w_{i,t} \text{diag}\left(\frac{\sigma_i}{1-\beta_1}\right)} \right) \\
&= \min_{\theta\in\Theta} \left(\left\langle \sum_{i=1}^{t} w_{i,t} g_i, \theta \right\rangle + \frac{1}{2}\|\theta - \theta_{t-1}\|^2_{\text{diag}\left(\frac{d_t}{1-\beta_1^t}\right)} \right).
\end{aligned}$$

Thus, with the definition of d_t, solving above problem, we obtain (16). □