A Technical details of the convergence analysis

We collect below some basic tools and definitions from convex analysis.

Definition A.1 (Bregman divergence)**.** Let $h : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ be differentiable strictly convex function. The *Bregman divergence* generated by h is

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y),\, x - y \rangle, \qquad x, y \in \mathcal{X}. \tag{A.1}$$

– **Fenchel conjugate:**

$$f^*(y) = \sup_{x \in \mathcal{X}} \langle x, y \rangle - f(x) \tag{A.2}$$

– **Prox operator:**

$$\operatorname{prox}_f(x) = \operatorname*{argmin}_{y \in \mathcal{X}} f(y) + \frac{1}{2}\|x - y\|_2^2, \qquad \forall\ x \in \mathcal{X} \tag{A.3}$$

– **Moreau decomposition:**

$$x = \operatorname{prox}_f(x) + \operatorname{prox}_{f^*}(x), \qquad \forall\ x \in \mathcal{X} \tag{A.4}$$

– **Fenchel-Young inequality:**

$$\langle x,\, y \rangle \le f(x) + f^*(y) \tag{A.5}$$

– **Projection lemma:**

$$\langle y - \Pi_{\mathcal{X}}(y),\, x - \Pi_{\mathcal{X}}(y) \rangle \le 0, \qquad \forall\ x \in \mathcal{X}. \tag{A.6}$$

– **Descent lemma:**

$$f(y) \le f(x) + \langle \nabla f(x),\, y - x \rangle + \tfrac{L}{2}\|y - x\|^2. \tag{A.7}$$

– **Four-point identity:** Bregman divergences satisfy the following *four point identity*:

$$\langle \nabla h(a) - \nabla h(b),\, c - d \rangle = D_h(d, a) - D_h(d, b) - D_h(c, a) + D_h(c, b). \tag{A.8}$$

A special case of (A.8) is the "three-point" identity

$$\langle \nabla h(a) - \nabla h(b),\, b - c \rangle = D_h(c, a) - D_h(c, b) - D_h(b, a). \tag{A.9}$$

A.1 Bounding the change $f(x_{t+1}) - f(x^*)$

We start the analysis by bounding the gap $f(x_{t+1}) - f(x^*)$. The lemma below is just a combination of several results of [1]. We present the details below in one place for easy reference. The impact of our delay sensitive step sizes shows up in subsequent lemmas, where we bound the individual terms that arise from Lemma A.2.

Lemma A.2. *At any time-point t, let the gradient error due to delays be*

$$e_t := \nabla f(x_t) - g(t - \tau_t). \tag{A.10}$$

Then, we have the following (deterministic) bound:

$$\begin{aligned}
&f(x_{t+1}) - f(x^*) \\
&= \frac{1}{2\alpha(t, \tau_t)}\left[\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2\right] + \langle e_t,\, x_{t+1} - x^* \rangle + \tfrac{L - 1/\alpha(t,\tau_t)}{2}\|x_t - x_{t+1}\|^2, \\
&\le \frac{1}{2\alpha(t, \tau_t)}\left[\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2\right] + \langle \nabla f(x_t) - \nabla f(x(t - \tau_t)),\, x_{t+1} - x^* \rangle \\
&\quad + \langle \nabla f(x(t - \tau_t)) - g(t - \tau_t),\, x_t - x^* \rangle + \tfrac{1}{2\eta(t,\tau_t)}\|\nabla f(x(t - \tau_t)) - g(t - \tau_t)\|^2.
\end{aligned} \tag{A.11}$$

Proof. Using convexity of f we have

$$f(x_t) - f(x^*) \leq \langle \nabla f(x_t),\, x_{t+1} - x^* \rangle + \langle \nabla f(x_t),\, x_t - x_{t+1} \rangle. \tag{A.12}$$

Now apply Lipschitz continuity of ∇f to the second term to obtain

$$\begin{aligned} & f(x_t) - f(x^*) \leq \langle \nabla f(x_t),\, x_{t+1} - x^* \rangle + f(x_t) - f(x_{t+1}) + \tfrac{L}{2}\|x_t - x_{t+1}\|^2, \\ \implies\ & f(x_{t+1}) - f(x^*) \leq \langle \nabla f(x_t),\, x_{t+1} - x^* \rangle + \tfrac{L}{2}\|x_t - x_{t+1}\|^2. \end{aligned} \tag{A.13}$$

Using the definition (A.10) of the gradient error e_t, we can rewrite (A.13) as

$$f(x_{t+1}) - f(x^*) \leq \underbrace{\langle g(t-\tau_t),\, x_{t+1} - x^* \rangle}_{T1} + \underbrace{\langle e_t,\, x_{t+1} - x^* \rangle}_{T2} + \tfrac{L}{2}\|x_t - x_{t+1}\|^2.$$

To complete the proof, we bound the terms $T1$ and $T2$ separately below.

Bounding T1: Since x_{t+1} is a minimizer in (2.1), from the projection inequality (A.6) we have

$$\langle x_t - \alpha(t,\tau_t) g(t-\tau_t) - x_{t+1},\, x - x_{t+1} \rangle \leq 0, \quad \forall x \in \mathcal{X}.$$

Choose $x = x^*$; then rewrite the above inequality and identity (A.9) with $h(x) = \frac{1}{2}\|x\|^2$ to get

$$\begin{aligned} \alpha(t,\tau_t)\langle g(t-\tau_t),\, x_{t+1} - x^* \rangle &\leq \langle x_t - x_{t+1},\, x_{t+1} - x^* \rangle \\ &= \tfrac{1}{2}\|x^* - x_t\|^2 - \tfrac{1}{2}\|x^* - x_{t+1}\|^2 - \tfrac{1}{2}\|x_{t+1} - x_t\|^2; \end{aligned}$$

Plugging in this bound for $T1$ and collecting the $\|x_{t+1} - x_t\|^2$ terms we obtain

$$\begin{aligned} & f(x_{t+1}) - f(x^*) \\ & \leq \tfrac{1}{2\alpha(t,\tau_t)}\left[\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2 - \|x_{t+1} - x_t\|^2\right] + \langle e_t,\, x_{t+1} - x^* \rangle + \tfrac{L}{2}\|x_t - x_{t+1}\|^2 \\ & = \tfrac{1}{2\alpha(t,\tau_t)}\left[\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2\right] + \langle e_t,\, x_{t+1} - x^* \rangle + \tfrac{L - 1/\alpha(t,\tau_t)}{2}\|x_t - x_{t+1}\|^2. \end{aligned} \tag{A.14}$$

Bounding T2: Adding and subtracting $\nabla f(x(t-\tau_t))$ we obtain

$$\begin{aligned} & \langle e_t,\, x_{t+1} - x^* \rangle = \langle \nabla f(x_t) - g(t-\tau_t),\, x_{t+1} - x^* \rangle \\ & = \langle \nabla f(x_t) - \nabla f(x(t-\tau_t)),\, x_{t+1} - x^* \rangle + \langle \nabla f(x(t-\tau_t)) - g(t-\tau_t),\, x_{t+1} - x^* \rangle \\ & = \langle \nabla f(x_t) - \nabla f(x(t-\tau_t)),\, x_{t+1} - x^* \rangle + \langle \nabla f(x(t-\tau_t)) - g(t-\tau_t),\, x_t - x^* \rangle \\ & \quad + \langle \nabla f(x(t-\tau_t)) - g(t-\tau_t),\, x_{t+1} - x_t \rangle \\ & \leq \langle \nabla f(x_t) - \nabla f(x(t-\tau_t)),\, x_{t+1} - x^* \rangle + \langle \nabla f(x(t-\tau_t)) - g(t-\tau_t),\, x_t - x^* \rangle \\ & \quad + \tfrac{1}{2\eta(t,\tau_t)}\|\nabla f(x(t-\tau_t)) - g(t-\tau_t)\|^2 + \tfrac{\eta(t,\tau_t)}{2}\|x_{t+1} - x_t\|^2, \end{aligned}$$

where the last inequality is an application of (A.5). Adding this inequality to (A.14) and using $1/\alpha(t,\tau_t) = L + \eta(t,\tau_t)$, we obtain (A.11). □

The next step is to take expectations over (A.11) and then further bound the resulting terms separately. Note that $\nabla f(x(t-\tau_t)) - g(t-\tau_t)$ is independent of x_t given $g(1), \ldots, g(t-\tau_t-1)$ (since x_t is a function of gradients up to time $t-\tau_t-1$). Thus, the third term in (A.11) has zero expectation. It remains to consider expectations over the following three quantities:

$$\Delta(t) := \frac{1}{2\alpha(t,\tau_t)}\left[\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2\right]; \tag{A.15}$$

$$\Gamma(t) := \langle \nabla f(x_t) - \nabla f(x(t-\tau_t)),\, x_{t+1} - x^* \rangle; \tag{A.16}$$

$$\Sigma(t) := \tfrac{1}{2\eta(t,\tau_t)}\|\nabla f(x(t-\tau_t)) - g(t-\tau_t)\|^2. \tag{A.17}$$

Lemma A.3 bounds (A.15) under Assumption 2.5(A), while Lemma A.4 provides a bound under the Assumption 2.5(B). Similarly, Lemmas A.5 and A.6 bound (A.16), while Lemmas A.7 bounds (A.17). Combining these bounds we obtain the theorem.

A.2 Bounding Δ, Γ, and Σ

Lemma A.3. *Let $\Delta(t)$ be given by (A.15), and let Assumption 2.5 (A) hold. Then,*

$$\sum_{t=1}^{T}\mathbb{E}[\Delta(t)] = \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\Big[\frac{1}{\alpha(t,\tau_t)}\left(\|x^*-x_t\|^2-\|x^*-x_{t+1}\|^2\right)\Big] \le \tfrac{1}{2}(L+c)R^2+\sqrt{2}cR^2\bar{\tau}\sqrt{T}.$$

Proof. Unlike the delay independent step sizes treated in [1], bounding $\Delta(t)$ requires some more work because $\alpha(t,\tau_t)$ depends on τ_t, which in turn breaks the monotonically decreasing nature of $\alpha(t,\tau_t)$ (we wish to avoid using a fixed worst case bound on the steps, to gain more precise insight into the impacts of being sensitive to delays), necessitating a more intricate analysis.

Let $r_t = \|x_t - x^*\|^2$. Observe that although $r_t \perp\!\!\!\perp \tau_t$, it is *not* independent of $\tau(t-1)$. Thus, with

$$z_t = \frac{1}{\alpha(t,\tau_t)} - \frac{1}{\alpha(t-1,\tau_{t-1})} = c(\sqrt{t+\tau_t}-\sqrt{t-1+\tau_{t-1}}),$$

we have

$$\sum_{t=1}^{T}\mathbb{E}[\Delta(t)] = \frac{1}{2}\mathbb{E}\Big[\frac{r_1}{\alpha(1,\tau(1))}+\sum_{t=2}^{T}z_t r_t\Big] \le \frac{1}{2}(L+c)R^2+\frac{1}{2}\mathbb{E}\Big[\sum_{t=2}^{T}z_t r_t\Big]. \tag{A.18}$$

Since $\alpha(t,\tau_t)$ is *not* monotonically decreasing with t, while upper-bounding $\mathbb{E}[\Delta(t)]$ we cannot simply discard the final term in (A.18).

When $\tau(t-1)\sim U(\{0,2\bar{\tau}\})$, r_t uniformly takes on at most $2\bar{\tau}+1$ values

$$r_{t,s} := \|x_{t,s}-x^*\|^2, \qquad s\in[2\bar{\tau}],$$

where $x_{t,s} = \Pi_{\mathcal{X}}[x_{t-1}-\alpha(t-1,\tau(t-1)=s)g(t-1,\tau(t-1))]$. Given a delay $\tau(t-1)=s$, r_t is just $r_{t,s}$. Using $z_t = \alpha(t)^{-1}-\alpha(t-1)^{-1} = c\sqrt{t+\tau_t}-c\sqrt{t-1+\tau_{t-1}}$, we have

$$z_{t,s} = c\left(\sqrt{t+\tau_t}-\sqrt{t-1+s}\right), \qquad s\in[2\bar{\tau}].$$

Using nested expectations $\mathbb{E}[z_t r_t] = \mathbb{E}_{\tau_t}[\mathbb{E}[z_t r_t|\tau_t]]$ we then see that

$$\begin{aligned}\mathbb{E}[z_t r_t] &= \frac{1}{2\bar{\tau}+1}\sum_{l=0}^{2\bar{\tau}}\left(\sum_{s=0}^{2\bar{\tau}}(2\bar{\tau}+1)^{-1}r_{t,s}c\left(\sqrt{t+l}-\sqrt{t-1+s}\right)\right)\\ &\le \frac{1}{2\bar{\tau}+1}\sum_{l=0}^{2\bar{\tau}}\left(\sum_{s=0}^{l-1}(2\bar{\tau}+1)^{-1}r_{t,s}c\left(\sqrt{t+l}-\sqrt{t-1+s}\right)\right),\end{aligned}$$

where we dropped the terms with $s\ge l$ as they are non-positive.

Consider now the inner summation above. We have

$$\begin{aligned}&\frac{c}{2\bar{\tau}+1}\sum_{s=0}^{l-1}r_{t,s}\left(\sqrt{t+l}-\sqrt{t-1+s}\right)\\ &\le \frac{cR^2}{2\bar{\tau}+1}\sum_{s=0}^{l-1}\left(\sqrt{t+l}-\sqrt{t-1+s}\right)\\ &= \frac{cR^2}{2\bar{\tau}+1}\sum_{s=0}^{l-1}\frac{l-s+1}{\sqrt{t+l}+\sqrt{t-1+s}}\\ &\le \frac{cR^2}{2\bar{\tau}+1}\frac{1}{\sqrt{2t-1}}\sum_{s=0}^{l-1}(l-s+1)\\ &= \frac{cR^2}{2\bar{\tau}+1}\frac{1}{\sqrt{2t-1}}\frac{3l+l^2}{2}.\end{aligned}$$

Thus, we now consider

$$\begin{aligned}\mathbb{E}[z_t r_t] &\leq \frac{1}{2\bar{\tau}+1}\sum_{l=0}^{2\bar{\tau}} \frac{cR^2}{2\bar{\tau}+1}\frac{1}{\sqrt{2t-1}}\frac{3l+l^2}{2} \\ &= \frac{cR^2}{(2\bar{\tau}+1)^2\sqrt{2t-1}}(2\bar{\tau}+1)(4\bar{\tau}+2.5)\bar{\tau} \\ &< \frac{2cR^2\bar{\tau}}{\sqrt{2t-1}}.\end{aligned}$$

Summing over $t=2$ to T, we finally obtain the upper bound

$$\sum_{t=2}^{T}\mathbb{E}[z_t r_t] \leq cR^2\bar{\tau}\sum_{t=2}^{T}\frac{1}{\sqrt{2t-1}} \leq 2cR^2\bar{\tau}\sqrt{2T}. \qquad \square$$

Lemma A.4. *Let Assumption (2.5) (B) hold. Then*

$$\sum_{t=1}^{T}\mathbb{E}[\Delta(t)] \leq \frac{1}{2}R^2(L+c) + \frac{1}{2}cR^2\sum_{t=2}^{T}\frac{\bar{\tau}_t+1}{\sqrt{2t-1}}.$$

Proof. Proceeding as for Lemma A.3, according to (A.18), the task reduces to bounding $\mathbb{E}[z_t r_t]$. Consider thus,

$$\mathbb{E}[z_t r_t] \leq \mathbb{E}[z_t^+ r_t] \leq R^2\mathbb{E}[z_t^+],$$

where we use z_t^+ to denote $\max(z_t, 0)$. Let us now control the last expectation. Let $P_t(l) = \mathbb{P}(\tau(t)=l)$, then

$$\begin{aligned}\mathbb{E}[z_t^+] &= \sum_{\tau_t,\tau_{t-1}} P(\tau_t,\tau_{t-1})\max(0,z_t) \\ &= c\sum_{l=0}^{t-1}\sum_{s=0}^{t-2} P_t(l)P_{t-1}(s)[\sqrt{t+l}-\sqrt{t-1+s}]^+ \\ &= c\sum_{l=0}^{t-1}\sum_{s=0}^{l} P_t(l)P_{t-1}(s)\frac{l+1-s}{\sqrt{t+l}+\sqrt{t-1+s}} \\ &\leq c\sum_{l=0}^{t-1}\sum_{s=0}^{l} P_t(l)P_{t-1}(s)\frac{l+1}{\sqrt{2t+l-1}} \\ &\leq c\sum_{l=0}^{t-1} P_t(l)\frac{l+1}{\sqrt{2t+l-1}} \\ &\leq c\sum_{l=0}^{t-1} P_t(l)\frac{l+1}{\sqrt{2t-1}} = c\frac{\bar{\tau}_t+1}{\sqrt{2t-1}}.\end{aligned}$$

So

$$\sum_{t=2}^{T} R^2\mathbb{E}[z_t^+] \leq cR^2\sum_{t=2}^{T}\frac{\bar{\tau}_t+1}{\sqrt{2t-1}}. \qquad \square$$

Lemma A.5.

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[\Gamma(t)] &= \sum_{t=1}^{T}\mathbb{E}\left[\langle \nabla f(x_t) - \nabla f(x(t-\tau_t)),\, x_{t+1}-x^*\rangle\right] \\ &\leq \bar{\tau}GR + \frac{LC_1}{2} + \frac{LC_2}{2}\log T\end{aligned}$$

where

$$C_1 = \frac{G^2\bar{\tau}(\bar{\tau}+1)(2\bar{\tau}+1)^2}{3(L^2+c^2)} \quad \textit{and} \quad C_2 = \frac{G^2(4\bar{\tau}+3)(\bar{\tau}+1)}{3c^2}$$

Proof. This proof is an adaptation of Lemma 4 and Corollary 1 of Agarwal and Duchi [1]. First, we exploit convexity of f to help analyze the gradient differences using the four-point identity (A.8):

$$\begin{aligned}&\langle \nabla f(x_t) - \nabla f(x(t-\tau_t)),\, x_{t+1} - x^*\rangle \\ &= D_f(x^*, x_t) - D_f(x^*, x(t-\tau_t)) - D_f(x_{t+1}, x_t) + D_f(x_{t+1}, x(t-\tau_t)).\end{aligned} \tag{A.19}$$

Since ∇f is L-Lipschitz, we further have

$$f(x_{t+1}) \le f(x(t-\tau_t)) + \langle \nabla f(x(t-\tau_t)),\, x_{t+1} - x(t-\tau_t)\rangle + \tfrac{L}{2}\|x(t-\tau_t) - x_{t+1}\|^2.$$

By definition of a Bregman divergence, we also have

$$D_f(x_{t+1}, x(t-\tau_t)) = f(x_{t+1}) - f(x(t-\tau_t)) - \langle \nabla f(x(t-\tau_t)),\, x_{t+1} - x(t-\tau_t)\rangle,$$

which, upon using using A.7, immediately yields the bound

$$D_f(x_{t+1}, x(t-\tau_t)) \le \tfrac{L}{2}\|x(t-\tau_t) - x_{t+1}\|^2.$$

Dropping the negative term $D_f(x_{t+1}, x_t)$ from (A.19) and summing over t, we then obtain

$$\begin{aligned}&\sum_{t=1}^{T} \langle \nabla f(x_t) - \nabla f(x(t-\tau_t)),\, x_{t+1} - x^*\rangle \\ &\le \sum_{t=1}^{T} \left[D_f(x^*, x_t) - D_f(x^*, x(t-\tau_t))\right] + \frac{L}{2}\sum_{t=1}^{T} \|x_{t+1} - x(t-\tau_t)\|^2.\end{aligned}$$

Notice that the first sum partially telescopes, leaving only the terms not received by the server within the first T iterations. Thus, we obtain the bound

$$\sum_{t:t+\tau_t>T} D_f(x^*, x_t) + \frac{L}{2}\sum_{t=1}^{T} \|x_{t+1} - x(t-\tau_t)\|^2. \tag{A.20}$$

We bound both each of the terms in (A.20) in turn below.

To bound the contribution of the first term in expectation, compute the expected cardinality

$$\mathbb{E}[|\{t : t + \tau_t > T\}|] = \sum_{t=1}^{T} \Pr(\tau_t > T - t), \tag{A.21}$$

Assuming delays uniform on $\{0, 2\bar{\tau}\}$ bounding this cardinality is easy, since

$$\Pr(\tau_t > T - t) = \begin{cases} 0 & T - t > 2\bar{\tau}, \\ \frac{2\bar{\tau} - T + t}{2\bar{\tau}+1} & \text{otherwise.} \end{cases}$$

Assuming that $2\bar{\tau} + 1 < T$, (A.21) becomes (unsurprisingly)

$$\sum_{s=1}^{2\bar{\tau}} \frac{2\bar{\tau} - s}{2\bar{\tau}+1} = \frac{(4\bar{\tau} - 2\bar{\tau})(2\bar{\tau}+1)}{2(2\bar{\tau}+1)} = \bar{\tau}.$$

From definition of a Bregman divergence we immediately see that

$$0 \le D_f(x^*, x_t) \le -\langle \nabla f(x_t),\, x^* - x_t\rangle \le \|\nabla f(x_t)\|\|x^* - x_t\| \le GR.$$

Thus, the contribution of the first term in (A.20) is bounded in expectation by by $\bar{\tau}GR$.

To bound the contribution of the second term, use convexity of $\|\cdot\|^2$ to obtain

$$\begin{aligned}
&\|x_{t+1} - x(t-\tau_t)\| \\
=&\|x_{t+1} - x_t + x_t - x(t-1) + \cdots + x(t-\tau_t+1) - x(t-\tau_t)\|^2 \\
\leq&(\tau_t+1)^2 \sum_{s=0}^{\tau_t} \tfrac{1}{\tau_t+1} \|x_{t+1-s} - x_{t-s}\|^2 \\
=&(\tau_t+1) \sum_{s=0}^{\tau_t} \|\Pi_{\mathcal{X}}\big(x(t-s) - \alpha(t-s,\tau_{t-s}) g(t-s,\tau_{t-s})\big) - \Pi_{\mathcal{X}}(x(t-s))\|^2 \\
\leq&(\tau_t+1) G^2 \sum_{s=0}^{\tau_t} \alpha(t-s,\tau_{t-s})^2.
\end{aligned}$$

Conditioned on the delay τ_t we have

$$\mathbb{E}[\|x_{t+1} - x(t-\tau_t)\|^2 | \tau_t] \leq (\tau_t+1) G^2 \textstyle\sum_{s=0}^{\tau_t} \mathbb{E}[\alpha(t-s,\tau_{t-s})^2].$$

Under the uniform or scaled assumptions on delays, we obtain similar bounds on the above quantity. Consider now the expectation

$$\begin{aligned}
\mathbb{E}[\alpha(t-s,\tau(t-s))^2] = \mathbb{E}[\frac{1}{L^2 + c^2((t-s)+\tau(t-s)) + 2Lc\sqrt{t-s+\tau(t-s)}}] \leq \frac{1}{L^2 + c^2(t-s)} \\
\implies \text{ if } \tau_t = l, \ \sum_{s=0}^{\tau_t} \mathbb{E}[\alpha(t-s,\tau_{t-s})^2] \leq \sum_{s=0}^{l} \frac{1}{L^2 + c^2(t-l)} = \frac{l+1}{L^2 + c^2(t-l)}.
\end{aligned}$$

Thus, for $t > 2\bar{\tau}$, we have the following bound

$$\begin{aligned}
\mathbb{E}[\|x_{t+1} - x(t-\tau_t)\|^2] &\leq G^2 \sum_{l=0}^{2\bar{\tau}} \frac{1}{2\bar{\tau}+1} \frac{(l+1)^2}{L^2 + c^2(t-l)} \\
&\leq \frac{G^2}{(2\bar{\tau}+1)(L^2 + c^2(t-2\bar{\tau}))} \sum_{l=0}^{2\bar{\tau}} (l+1)^2 \\
&= \frac{G^2(4\bar{\tau}+3)(\bar{\tau}+1)}{3(L^2 + c^2(t-2\bar{\tau}))}.
\end{aligned}$$

and for $t \leq 2\bar{\tau}$, we have

$$\begin{aligned}
\mathbb{E}[\|x_{t+1} - x(t-\tau_t)\|^2] &\leq G^2 \sum_{l=0}^{t-1} P_t(l) \frac{(l+1)^2}{L^2 + c^2(t-l)} \\
&\leq G^2 \sum_{l=0}^{t-1} \frac{(l+1)^2}{L^2 + c^2} \\
&= \frac{G^2 t(t+1)(2t+1)}{6(L^2 + c^2)}.
\end{aligned}$$

Now adding up over $t = 1$ to T, we have

$$\sum_{t=1}^{T} \mathbb{E}[\|x_{t+1} - x(t-\tau_t)\|^2] \leq C_1 + C_2 \log T$$ □

Lemma A.6. *Assuming scaled delays, we have the bound*

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\Gamma(t)] &= \sum_{t=1}^{T} \mathbb{E}\left[\langle \nabla f(x_t) - \nabla f(x(t-\tau_t)),\, x_{t+1} - x^* \rangle\right] \\
&\leq GR\left(1 + \sum_{t=1}^{T-1} \frac{B_t^2}{(T-t)^2}\right) + LG^2 \sum_{t=1}^{T} \frac{B_t^2 + 1 + \bar{\tau}_t}{L^2 + c^2(1-\theta_t)t}.
\end{aligned}$$

Proof. We build on Corollary 1 of [1], and proceed as in Lemma A.5 to bound the terms in (A.20) separately. For the first term, we bound the expected cardinality using Chebyshev's inequality and Assumption 2.5 (B):

$$\mathbb{E}[|\{t : t + \tau_t > T\}|] = \sum_{t=1}^{T} \Pr(\tau_t > T - t) \leq 1 + \sum_{t=1}^{T-1} \frac{\mathbb{E}[\tau_t^2]}{(T-t)^2} = 1 + \sum_{t=1}^{T-1} \frac{B_t^2}{(T-t)^2}$$

To bound the second term, we again follow Lemma A.5 to obtain

$$\mathbb{E}[\|x_{t+1} - x(t - \tau_t)\|^2 | \tau_t] \leq (\tau_t + 1) G^2 \sum\nolimits_{s=0}^{\tau_t} \mathbb{E}[\alpha(t - s, \tau_{t-s})^2].$$

$$\begin{aligned}\mathbb{E}[\alpha(t-s, \tau(t-s))^2] &= \mathbb{E}[\frac{1}{L^2 + c^2((t-s) + \tau(t-s)) + 2Lc\sqrt{t - s + \tau(t-s)}}] \\ &\leq \frac{1}{L^2 + c^2(t-s)},\end{aligned}$$

which yields the bound (since $0 \leq s \leq \tau_t$)

$$\mathbb{E}[\|x_{t+1} - x(t - \tau_t)\|^2 | \tau_t] \leq \frac{G^2(\tau_t + 1)^2}{L^2 + c^2(t - \tau_t)}$$

Now adding up over $t = 1$ to T consider

$$G^2 \sum\nolimits_{t=1}^{T} \frac{(\tau_t + 1)^2}{L^2 + c^2(t - \tau_t)},$$

so that taking expectation (over τ_t) we then obtain

$$\sum_{t=1}^{T} \mathbb{E}[\|x_{t+1} - x(t - \tau_t)\|^2] \leq G^2 \sum_{t=1}^{T} \mathbb{E}\left[\frac{(\tau_t + 1)^2}{L^2 + c^2(t - \tau_t)}\right].$$

Using our assumption that $\tau_t < \theta_t t$ for $\theta_t \in (0, 1)$, we have in particular that

$$\begin{aligned}&G^2 \sum_{t=1}^{T} \mathbb{E}\left[\frac{(\tau_t + 1)^2}{L^2 + c^2(t - \tau_t)}\right] \\ \leq &G^2 \sum_{t=1}^{T} \frac{1}{L^2 + c^2(1 - \theta_t)t} \mathbb{E}[(\tau_t + 1)^2] \\ \leq &G^2 \sum_{t=1}^{T} \frac{B_t^2 + 1 + \bar{\tau}_t}{L^2 + c^2(1 - \theta_t)t}\end{aligned}$$

□

Lemma A.7. *Let the step-offsets be $\eta(t, \tau_t) = c\sqrt{t + \tau_t}$. For any delay distribution we have*

$$\sum_{t=1}^{T} \mathbb{E}[\Sigma(t)] \leq \frac{\sigma^2}{c}\sqrt{T}.$$

Proof. From Assumption 2.2 on the variance of stochastic gradients, it follows that

$$\mathbb{E}[\Sigma(t)] = \mathbb{E}\left[\tfrac{1}{2\eta(t,\tau_t)}\|\nabla f(x(t - \tau_t)) - g(t - \tau_t)\|^2\right] \leq \frac{\sigma^2}{2}\mathbb{E}\left[\eta(t, \tau_t)^{-1}\right].$$

Plugging in $\eta(t, \tau_t) = c\sqrt{t + \tau_t}$, clearly the bound

$$\frac{1}{c}\mathbb{E}[(t + \tau_t)^{-1/2}] = \frac{1}{c}\sum_{s=0}^{t-1} P(s)\frac{1}{\sqrt{t+s}} \leq \frac{1}{c\sqrt{t}}, \tag{A.22}$$

holds for any delay distribution. Summing up over t, we then obtain

$$\sum_{t=1}^{T} \mathbb{E}[\Sigma(t)] \leq \frac{\sigma^2}{2c}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq \frac{\sigma^2}{c}\sqrt{T}.$$

□

B More general step-sizes

If we use the offsets $\eta_t = c(t+\tau_t)^\beta$, where $0 < \beta < 1$, we obtain slightly more general step sizes that fit within our framework. The *only* benefit of considering stepsizes other than $\beta = 1/2$ is because they allow us to tradeoff the contributions of the various terms in the bounds, and for a larger value of β for instance, we will obtain smaller step sizes, which can be beneficial in high noise regimes, at least in the initial iterations. The theoretical sweet-spot (in terms of dependence on T), is, however $\beta = 1/2$, the choice analyzed above. We summarize below the impact of these steps sizes for non-uniform scaled delays; the uniform case is even simpler. For simplicity, we do not bound the terms as tightly as for the special case $\beta = 1/2$.

Lemma B.1. *Assume that τ_t satisfies Assumption 2.5 (B) and $\eta_t = c(t+\tau_t)^\beta$ and $0 < \beta < 1$. Then,*

$$\mathbb{E}[z_t^+] \leq \frac{cR^2\beta(\bar{\tau}_t+1)}{(t-1)^{1-\beta}} \tag{B.1}$$

$$\mathbb{E}[\|x_{t+1} - x(t-\tau_t)\|^2] \leq \frac{G^2(\tau_t+1)^2}{L^2 + c^2(t-\tau_t)^{2\beta}} \tag{B.2}$$

$$\mathbb{E}[\eta(t,\tau_t)^{-1}] \leq \frac{1}{ct^\beta}. \tag{B.3}$$

Proof. Proceeding as in Lemma A.4 we bound

$$\begin{aligned}
\mathbb{E}[z_t^+] &= c\sum_{l=0}^{t-1}\sum_{s=0}^{l} P_t(l)P_{t-1}(s)\left((t+l)^\beta - (t-1+s)^\beta\right) \\
&\leq c\sum_{l=0}^{t-1}\sum_{s=0}^{l} P_t(l)P_{t-1}(s)\beta\frac{l+1-s}{(t-1+s)^{1-\beta}} \\
&\leq c\beta\sum_{l=0}^{t-1}\sum_{s=0}^{l} P_t(l)P_{t-1}(s)\frac{l+1}{(t-1)^{1-\beta}} \\
&\leq c\beta\sum_{l=0}^{t-1} P_t(l)\frac{l+1}{(t-1)^{1-\beta}} = \frac{c\beta(\bar{\tau}_t+1)}{(t-1)^{1-\beta}}.
\end{aligned}$$

where the first inequality follows from concavity if t^β, the second one since $\frac{l+1-s}{(t-1+s)^{1-\beta}}$ is decreasing in s, while the third is clear as P_{t-1} is a probability.

Next, we bound (B.2). Proceeding as in Lemma A.6, we obtain the bounds

$$\begin{aligned}
&\mathbb{E}[\alpha(t-s,\tau_{t-s})^2] \leq \frac{1}{L^2 + c^2(t-s)^{2\beta}} \\
\implies &\mathbb{E}[\|x_{t+1} - x(t-\tau_t)\|^2|\tau_t] \leq \frac{G^2(\tau_t+1)^2}{L^2 + c^2(t-\tau_t)^{2\beta}}
\end{aligned}$$

Finally, the bound on (B.3) is trivial; since $\eta_t^{-1} = c^{-1}(t+\tau_t)^{-\beta}$, we have

$$\frac{1}{c}\mathbb{E}[(t+\tau_t)^{-\beta}] = \frac{1}{c}\sum_{s=0}^{t-1} P_t(s)\frac{1}{(t+s)^\beta} \leq \frac{1}{ct^\beta}. \qquad \square$$

Using these key bounds, we can defined full versions of Lemmas A.4, A.6, and A.7, where we finally we will need a bound of the form

$$\sum_{t=1}^{T}\frac{1}{t^\beta} \leq 1 + \int_0^T t^{-\beta}dt = 1 + \frac{\left(T^{1-\beta}-1\right)}{1-\beta} \leq \frac{1}{1-\beta}T^{1-\beta}.$$