A Appendix

Here, we report the proofs missing from the main text.

A.1 Details of Example 1

Consider the function $f(x) = \frac{1}{2}x^2$. The gradient in t-th iteration is $\nabla f(x_t) = x_t$. Let the stochastic gradient be defined as $\boldsymbol{g}_t = \nabla f(x_t) + \xi_t$, where $P(\xi_t = \sigma_t) = \frac{7}{15}$, $P(\xi_t = -\frac{3}{2}\sigma_t) = \frac{1}{5}$ and $P(\xi_t = -\frac{1}{2}\sigma_t) = \frac{1}{3}$.

Let $A \triangleq \sum_{i=1}^{t-1} g_i^2 + \beta$. Then

$$\langle \mathbb{E}_t \eta_{t+1} \boldsymbol{g}_t, \nabla f(x_t) \rangle = \alpha \left[\frac{7}{15} \frac{(x_t + \sigma_t)x_t}{[A + (x_t + \sigma_t)^2]^{\frac{1}{2}+\epsilon}} + \frac{1}{5} \frac{(x_t - \frac{3}{2}\sigma_t)x_t}{[A + (x_t - \frac{3}{2}\sigma_t)^2]^{\frac{1}{2}+\epsilon}} + \frac{1}{3} \frac{(x_t - \frac{1}{2}\sigma_t)x_t}{[A + (x_t - \frac{1}{2}\sigma_t)^2]^{\frac{1}{2}+\epsilon}} \right].$$

This expression can be negative, for example, setting $x_t = 1$, $\sigma_t = 10$, $A = 10$, $\epsilon = 0$ or $\epsilon = 0.1$.

A.2 Proof of Lemma 2

Lemma 9. *Let $a_i \geq 0, \cdots, T$ and $f : [0, +\infty) \to [0, +\infty)$ nonincreasing function. Then*

$$\sum_{t=1}^{T} a_t f\left(a_0 + \sum_{i=1}^{t} a_i\right) \leq \int_{a_0}^{\sum_{t=0}^{T} a_t} f(x)dx.$$

Proof. Denote by $s_t = \sum_{i=0}^{t} a_i$.

$$a_i f(s_i) = \int_{s_{i-1}}^{s_i} f(s_i)dx \leq \int_{s_{i-1}}^{s_i} f(x)dx.$$

Summing over $i = 1, \cdots, T$, we have the stated bound. □

Proof of Lemma 2. The proof is immediate from Lemma 9. □

A.3 Proofs of Section 6.1

Proof of Lemma 4. From (4), for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have

$$f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} \rangle + \frac{M}{2}\|\boldsymbol{y}\|^2.$$

Take $\boldsymbol{y} = -\frac{1}{M}\nabla f(\boldsymbol{x})$, to have

$$f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + \left(\frac{1}{2M} - \frac{1}{M}\right)\|\nabla f(\boldsymbol{x})\|^2.$$

Hence,

$$\|\nabla f(\boldsymbol{x})\|^2 \leq 2M(f(\boldsymbol{x}) - f(\boldsymbol{x} + \boldsymbol{y})) \leq 2M(f(\boldsymbol{x}) - \min_{\boldsymbol{u}} f(\boldsymbol{u})).$$ □

Proof of Lemma 5. If $A \leq Bx$, then $x \leq C(2Bx)^{\frac{1}{2}+\epsilon}$, so $x \leq \left[C(2B)^{\frac{1}{2}+\epsilon}\right]^{\frac{1}{1/2-\epsilon}}$. And if $A > Bx$, then $x < C(2A)^{\frac{1}{2}+\epsilon}$. Taking the maximum of the two cases, we have the stated bound. □

Proof of Lemma 6. Assume that $Bx > A$. We have that

$$x^2 \leq (A + Bx)(C + D\ln(A + Bx)) < 2Bx(C + D\ln(2Bx)) < 2Bx(C + 2D\sqrt{2Bx}),$$

that is

$$x < 2BC + 4BD\sqrt{2Bx}.$$

We can solve this inequality, to obtain

$$x < 32B^3D^2 + 2BC + 8B^2D\sqrt{C}.$$

On the other hand, if $Bx \le A$, we have $x \le \frac{A}{B}$. Taking the sum of these two case, we have the stated bound. □

Proof of Lemma 7. Let $f(x) = (x+y)^p - x^p - y^p$. We can see that $f'(x) = p(x+y)^{p-1} - px^{p-1} \le 0$ when $x, y \ge 0$. So $f(x) \le f(0) = 0$. The inequality holds. □

Lemma 10. *If $x > 0$, $\alpha > 0$, then $\ln(x) \le \alpha(x^{\frac{1}{\alpha}} - 1)$.*

Proof of Lemma 10. Let $f(x) = \ln(x) - \alpha x^{\frac{1}{\alpha}} + \alpha$. $f'(x) = \frac{1}{x} - x^{\frac{1}{\alpha}-1}$ is positive when $0 < x < 1$, $f'(1) = 0$ and $f'(x) < 0$ when $x > 1$. So $f(x) \le f(1) = 0$. The inequality holds. □

Proof of Lemma 8. Using the assumption on the noise, we have

$$\begin{aligned}
\exp\left(\frac{\mathbb{E}\left[\max_{1\le i\le T}\|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i,\xi_i)\|^2\right]}{\sigma^2}\right) &\le \mathbb{E}\left[\exp\left(\frac{\max_{1\le i\le T}\|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i,\xi_i)\|^2}{\sigma^2}\right)\right] \\
= \mathbb{E}\left[\max_{1\le i\le T}\exp\left(\frac{\|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i,\xi_i)\|^2}{\sigma^2}\right)\right] &\le \sum_{i=1}^T \mathbb{E}\left[\exp\left(\frac{\|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i,\xi_i)\|^2}{\sigma^2}\right)\right] \\
= \sum_{i=1}^T \mathbb{E}\left[\mathbb{E}_i\left[\exp\left(\frac{\|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i,\xi_i)\|^2}{\sigma^2}\right)\right]\right] &\le Te,
\end{aligned}$$

that implies

$$\mathbb{E}\left[\max_{1\le i\le T}\|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i,\xi_i)\|^2\right] \le \sigma^2(1+\ln T). \tag{12}$$

Hence, when $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \eta_t^2\|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2\right] &= \mathbb{E}\left[\sum_{t=1}^T \eta_{t+1}^2\|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2 + \sum_{t=1}^T \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2(\eta_t^2 - \eta_{t+1}^2)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \eta_{t+1}^2\|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2 + \sum_{t=1}^T \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2(\eta_t + \eta_{t+1})(\eta_t - \eta_{t+1})\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T \eta_{t+1}^2\|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2 + \sum_{t=1}^T 2\eta_t\|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2(\eta_t - \eta_{t+1})\right] \\
&\le \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 2\eta_1\mathbb{E}\left[\max_{1\le t\le T}\eta_t\|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2\right] \\
&\le \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 4\eta_1\mathbb{E}\left[\max_{1\le t\le T}\eta_t\left(\|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t) - \nabla f(\boldsymbol{x}_t)\|^2 + \|\nabla f(\boldsymbol{x}_t)\|^2\right)\right] \\
&\le \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 4\eta_1^2(1+\ln T)\sigma^2 + 4\eta_1\mathbb{E}\left[\sum_{t=1}^T \eta_t\|\nabla f(\boldsymbol{x}_t)\|^2\right] \\
&= \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \frac{4\alpha^2}{\beta^{1+2\epsilon}}(1+\ln T)\sigma^2 + \frac{4\alpha}{\beta^{\frac{1}{2}+\epsilon}}\mathbb{E}\left[\sum_{t=1}^T \eta_t\|\nabla f(\boldsymbol{x}_t)\|^2\right],
\end{aligned}$$

where in second inequality we used Lemma 2 and in fourth one we used (12). Note that the analysis after the second inequality also holds when $\epsilon = 0$.

And when $\epsilon = 0$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \eta_{t+1}^2 \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2\right] = \mathbb{E}\left[\sum_{t=1}^T \frac{\alpha^2 \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2}{(\beta + \sum_{i=1}^t \|\boldsymbol{g}(\boldsymbol{x}_i,\xi_t)\|^2)}\right] \\
&\leq 2\alpha^2 \mathbb{E}\left[\ln\left(\sqrt{\beta + \sum_{t=1}^T \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2}\right)\right] \\
&\leq 2\alpha^2 \mathbb{E}\left[\ln\left(\sqrt{\beta + 2\sum_{t=1}^T \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t) - \nabla f(\boldsymbol{x}_t)\|^2} + \sqrt{2\sum_{t=1}^T \|\nabla f(\boldsymbol{x}_t)\|^2}\right)\right] \\
&\leq 2\alpha^2 \ln\left(\sqrt{\beta + 2T\sigma^2} + \sqrt{2}\mathbb{E}\left[\sqrt{\sum_{t=1}^T \|\nabla f(\boldsymbol{x}_t)\|^2}\right]\right)
\end{aligned}$$

where in first inequality we used Lemma 10 and in the third one we used Jensen's inequality. Putting things together, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \eta_t^2 \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2\right] = \mathbb{E}\left[\sum_{t=1}^T \eta_{t+1}^2 \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2 + \sum_{t=1}^T \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2 (\eta_t^2 - \eta_{t+1}^2)\right] \\
&\leq 2\alpha^2 \ln\left(\sqrt{\beta + 2T\sigma^2} + \sqrt{2}\mathbb{E}\left[\sqrt{\sum_{t=1}^T \|\nabla f(\boldsymbol{x}_t)\|^2}\right]\right) + \frac{4\alpha^2}{\beta}(1+\ln T)\sigma^2 + \frac{4\alpha}{\beta^{\frac{1}{2}}}\mathbb{E}\left[\sum_{t=1}^T \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2\right]
\end{aligned}$$

□

A.4 Proofs of Section 5

Proof of Lemma 3. From (4), we have

$$\begin{aligned}
f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle + \frac{M}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 \\
&= f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t(\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t,\xi_t))\rangle - \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle + \frac{M}{2}\|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2.
\end{aligned}$$

Taking the conditional expectation with respect to $\xi_1, \cdots, \xi_{t-1}$, we have that

$$E_t[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t(\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t,\xi_t))\rangle] = \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t) - \boldsymbol{\eta}_t \mathbb{E}_t[\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)]\rangle = 0.$$

Hence, from the law of total expectation, we have

$$\mathbb{E}\left[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle\right] \leq \mathbb{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) + \frac{M}{2}\|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2\right].$$

Summing over $t = 1$ to T and lower bounding $f(\boldsymbol{x}_{T+1})$ with $f^\star$, we have the stated bound. □

Proof of Lemma 1. Since the series $\sum_{t=1}^\infty a_t$ diverges, given that $\sum_{t=1}^\infty a_t b_t$ converges, we necessarily have $\liminf_{t\to\infty} b_t = 0$. So there exists a subsequence $\{b_{i(t)}\}$ of $\{b_t\}$ such that $\lim_{t\to\infty} b_{i(t)} = 0$.

Let us proceed by contradiction and assume that there exists some $\alpha > 0$ and some other subsequence $\{b_{m(t)}\}$ of $\{b_t\}$ such that $b_{m(t)} \geq \alpha$ for all t. In this case, we can construct a third subsequence $\{b_{j(t)}\}$ of $\{b_t\}$ where the subindices $j(t)$ are chosen in the following way:

$$j(0) = \min\{l \geq 0 : b_l \geq \alpha\}$$

and, given $j(2t)$,

$$j(2t+1) = \min\{l \geq j(2t) : b_l \leq \frac{1}{2}\alpha\}, \tag{13}$$

$$j(2t+2) = \min\{l \geq j(2t+1) : b_l \leq \frac{1}{2}\alpha\}. \tag{14}$$

Note that the existence of $\{b_{i(t)}\}$ and $\{b_{m(t)}\}$ guarantees that $j(t)$ is well defined. Also by (13) and (14)

$$b_l \leq \frac{\alpha}{2} \text{for } j(2t) \leq l \leq j(2t+1) - 1.$$

Then, denoting $\phi_t = \sum_{l=2t}^{j(2t+1)-1} a_l$, we have

$$\infty > \sum_{t=1}^{\infty} a_t b_t \geq \sum_{t=1}^{\infty} \sum_{l=2t}^{j(2t+1)-1} a_l b_l \leq \frac{\alpha}{2} \sum_{t=1}^{\infty} \phi_t.$$

Therefore, we have $\lim_{t\to\infty} \phi_t = 0$.

On the other hand, by (13) and (14), we have $b_{j(2t)} \geq \alpha$, $b_{j(2t+1)} \leq \frac{1}{\alpha}$, so that

$$\frac{\alpha}{2} \leq b_{j(2t)} - b_{j(2t+1)} = \sum_{l=j(2t)}^{j(2t+1)-1} (b_l - b_{l+1}) \leq \sum_{l=j(2t)}^{j(2t+1)-1} K a_l = K\phi_t.$$

So $\phi_t \geq \frac{\alpha}{2K}$, which is in contradiction with $\lim_{t\to\infty} \phi_t = 0$. Therefore, b_t goes to zero.

□

Proof of Theorem 2. We proceed similarly to the proof of Theorem 1, to get

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle\right] \leq f(\boldsymbol{x}_1) - f(\boldsymbol{x}^\star) + \frac{M}{2}\mathbb{E}\left[\sum_{t=1}^{\infty} \|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|_2^2\right].$$

Observe that

$$\sum_{t=1}^{\infty} \|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 = \sum_{t=1}^{\infty}\sum_{i=1}^{d} \eta_{t,i}^2 \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)_i^2 = \sum_{i=1}^{d}\sum_{t=1}^{\infty} \eta_{t,i}^2 \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)_i^2 < \infty,$$

where the last inequality comes from the same reasoning in (5). Hence, we have

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle\right] < \infty.$$

Hence, with probability 1, we have

$$\sum_{t=1}^{\infty} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle = \sum_{t=1}^{\infty}\sum_{j=1}^{d} \eta_{t,j} \nabla f(\boldsymbol{x}_t)_j^2 = \sum_{j=1}^{d}\sum_{t=1}^{\infty} \eta_{t,j} \nabla f(\boldsymbol{x}_t)_j^2 < \infty.$$

and, for any $j = 1, \cdots, d$,

$$\sum_{t=1}^{\infty} \eta_{t,j} (\nabla f(\boldsymbol{x}_t))_j^2 < \infty.$$

Now, observe that the Lipschitzness of f and the bounded support of the noise on the gradients gives

$$\sum_{t=1}^{\infty} \eta_{t,j} = \sum_{t=1}^{\infty} \frac{\alpha}{(\beta + \sum_{i=1}^{t-1} (g(\boldsymbol{x}_i, \xi_i)_j)^2)^{1/2+\epsilon}} \geq \sum_{t=1}^{\infty} \frac{\alpha}{(\beta + 2(t-1)(L^2 + S^2))^{1/2+\epsilon}} = \infty.$$

Using the fact the f is L-Lipschitz and M-smooth, we also have

$$\begin{aligned}
&\left|((\nabla f(\boldsymbol{x}_{t+1}))_j)^2 - ((\nabla f(\boldsymbol{x}_t))_j)^2\right| = ((\nabla f(\boldsymbol{x}_{t+1}))_j + (\nabla f(\boldsymbol{x}_t))_j) \cdot |(\nabla f(\boldsymbol{x}_{t+1}))_j - (\nabla f(\boldsymbol{x}_t))_j| \\
&\quad \leq 2LM\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = 2LM\|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\| \leq 2LM(L+S)\eta_t.
\end{aligned}$$

Hence, we case use Lemma 1 to obtain

$$\lim_{t\to\infty} ((\nabla f(\boldsymbol{x}_t))_j)^2 = 0.$$

For the second statement, observe that, with probability 1,

$$\sum_{t=1}^{\infty} ((\nabla f(\boldsymbol{x}_t))_j)^2 t^{1/2-\epsilon} \frac{\alpha}{t(2L^2 + 2S^2 + \beta)^{1/2+\epsilon}} \leq \sum_{t=1}^{\infty} \eta_{t,j} (\nabla f(\boldsymbol{x}_t))_j)^2 < \infty.$$

Hence, noting that $\sum_{t=1}^{\infty} \frac{1}{t} = \infty$, we have that $\liminf_{t\to\infty} ((\nabla f(\boldsymbol{x}_t))_j)^2 t^{1/2-\epsilon} = 0$. □