A. Additional Proofs

A.1. Bounded smooth interpolation

We start by recalling a recent and fundamental theorem which provide necessary and sufficient conditions under which a set $\{\mathbf{x}_t, \mathbf{g}_t, f_t\}_{i\in[T]}$ can be *interpolated* (or *extended* using the terminology from the classical text (Whitney, 1934)) by a convex function f with L-Lipschitz gradient such that $f(\mathbf{x}_t) = f_t$ and $\nabla f(\mathbf{x}_t) = \mathbf{g}_t$ for all $t \in [T]$. The theorem was established in (Taylor et al., 2017b) and also independently in (Azagra & Mudarra, 2017) for a more general setting in Hilbert spaces.

Theorem 5. *Let $L > 0$, $d \in \mathbb{N}$ and suppose $\{(\mathbf{x}_t, \mathbf{g}_t, f_t)\}_{t\in[T]}$ is some finite subset of $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$. Then there exists a convex function f with L-Lipschitz gradient that satisfies $f(\mathbf{x}_t) = f_t$ and $\nabla f(\mathbf{x}_t) = \mathbf{g}_t$ for all $t \in [T]$ if and only if*

$$\frac{1}{2L}\|\mathbf{g}_i - \mathbf{g}_j\|^2 \leq f_i - f_j - \langle \mathbf{g}_j, \mathbf{x}_i - \mathbf{x}_j\rangle, \quad \forall i, j \in [T]. \tag{9}$$

A similar result that provides necessary and sufficient conditions for non-convex interpolation is also known.

Theorem 6 ((Taylor et al., 2017a), Theorem 3.10)**.** *Let $L > 0$, $d \in \mathbb{N}$ and suppose $\{(\mathbf{x}_t, \mathbf{g}_t, f_t)\}_{t\in[T]}$ is some finite subset of $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$. Then there exists an function f with L-Lipschitz gradient that satisfies $f(\mathbf{x}_t) = f_t$ and $\nabla f(\mathbf{x}_t) = \mathbf{g}_t$ for all $t \in [T]$ if and only if*

$$\frac{1}{2L}\|\mathbf{g}_i - \mathbf{g}_j\|^2 - \frac{L}{4}\|\mathbf{x}_i - \mathbf{x}_j - \frac{1}{L}(\mathbf{g}_i - \mathbf{g}_j)\|^2 \leq f_i - f_j - \langle \mathbf{g}_j, \mathbf{x}_i - \mathbf{x}_j\rangle, \quad \forall i, j \in [T]. \tag{10}$$

Here we strengthen the results of Thm. 5 and Thm. 6, showing that interpolation can be performed in such a way that the resulting function is bounded from below and attains its minimum value. The proof is based on an explicit construction of a convex interpolating function developed in (Drori, 2017). This resolves an open question raised by (Fefferman et al., 2017) for the case where the interpolation set is finite.

Theorem 7. *Let $L > 0$, $d \in \mathbb{N}$ and suppose $\{(\mathbf{x}_t, \mathbf{g}_t, f_t)\}_{t\in[T]}$ is some finite subset of $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ that satisfies (9) (alternatively, (10)). Then there exists a convex (alternatively, nonconvex) function f with L-Lipschitz gradient that satisfies $f(\mathbf{x}_t) = f_t$, $\nabla f(\mathbf{x}_t) = \mathbf{g}_t$ and in addition, setting $j \in \arg\min_{t\in[T]} f_t - \frac{1}{2L}\|\mathbf{g}_t\|^2$, the function f also satisfies*

$$f^* := \min_{x\in\mathbb{R}^d} f(x) = f(\mathbf{x}_j - \frac{1}{L}\mathbf{g}_j) = f_j - \frac{1}{2L}\|\mathbf{g}_j\|^2.$$

Proof. The convex case follows directly from Theorem 1 in (Drori, 2017). Indeed, taking $C \leftarrow \{0\}$ and $\mathcal{T} \leftarrow \{(\mathbf{x}_t, \mathbf{g}_t, f_t)\}_{t\in[T]}$, the primal interpolation function $W_{\mathcal{T}}^C$ (see Definition 2.1 in (Drori, 2017)) can be written as

$$W(y) := \min_{\boldsymbol{\alpha}\in\Delta_T} \left[\frac{L}{2}\|y - \sum_{t\in T}\alpha_t(\mathbf{x}_t - \frac{1}{L}\mathbf{g}_t)\|^2 + \sum_{t\in T}\alpha_t(f_t - \frac{1}{2L}\|\mathbf{g}_t\|^2)\right], \tag{11}$$

where Δ_T is the T-dimensional unit simplex

$$\Delta_T := \{\boldsymbol{\alpha} \in \mathbb{R}^T : \sum_{t\in T}\alpha_t = 1,\ \alpha_t \geq 0,\ \forall t \in T\}.$$

By the assumption that $\mathcal{T}$ satisfies (9), Theorem 1 in (Drori, 2017) implies that W is convex, its gradient is L-Lipschitz and that $W(\mathbf{x}_t) = f_t$, $\nabla W(\mathbf{x}_t) = \mathbf{g}_t$. The lower bound on W then immediately follows from (11), as

$$W(y) \geq \min_{\boldsymbol{\alpha}\in\Delta_T}\left[\sum_{t\in T}\alpha_t(f_t - \frac{1}{2L}\|\mathbf{g}_t\|^2)\right] = \min_{t\in[T]}(f_t - \frac{1}{2L}\|\mathbf{g}_t\|^2) = f_j - \frac{1}{2L}\|\mathbf{g}_j\|^2,$$

and

$$W(\mathbf{x}_j - \frac{1}{L}\mathbf{g}_j) \leq f_j - \frac{1}{2L}\|\mathbf{g}_j\|^2,$$

which follows by taking $\boldsymbol{\alpha} = \mathbf{e}_j$ in (11). Finally, combining these two bounds completes the proof for the convex case:

$$f_j - \frac{1}{2L}\|\mathbf{g}_j\|^2 \leq \inf_y W(y) \leq W(\mathbf{x}_j - \frac{1}{L}\mathbf{g}_j) = f_j - \frac{1}{2L}\|\mathbf{g}_j\|^2.$$

For the non-convex case, consider the function

$$Z(y) := \min_{\boldsymbol{\alpha}\in\Delta_T} \left[L\|y - \sum_{t\in T}\alpha_t(\mathbf{x}_t - \frac{1}{2L}(\mathbf{g}_t + L\mathbf{x}_t))\|^2 + \sum_{t\in T}\alpha_t(f_t + \frac{L}{2}\|\mathbf{x}_t\|^2 - \frac{1}{4L}\|\mathbf{g}_t + L\mathbf{x}_t\|^2) \right].$$

It is straightforward to verify that this function is the primal interpolation function $W_{\mathcal{T}}^C$ taking $C \leftarrow \{0\}$, $L \leftarrow 2L$, and

$$\mathcal{T} \leftarrow \{(\mathbf{x}_t, \mathbf{g}_t + L\mathbf{x}_t, f_t + \frac{L}{2}\|\mathbf{x}_t\|^2)\}_{t\in[T]}.$$

As $\mathcal{T}$ satisfies (9) with a Lipschitz constant $2L$, by Theorem 1 in (Drori, 2017) it follows that Z is convex, has a $2L$-Lipschitz gradient and satisfies

$$\begin{aligned} Z(\mathbf{x}_t) &= f_t + \frac{L}{2}\|\mathbf{x}_t\|^2, \\ \nabla Z(\mathbf{x}_t) &= \mathbf{g}_t + L\mathbf{x}_t. \end{aligned}$$

Now let $\hat{W}$ be defined by $\hat{W}(y) := Z(y) - \frac{L}{2}\|y\|^2$. Clearly, $\hat{W}$ has L-Lipschitz gradient (see e.g., Lemma 3.9 in (Taylor et al., 2017a)), satisfies

$$\begin{aligned} \hat{W}(\mathbf{x}_t) &= f_t, \\ \nabla\hat{W}(\mathbf{x}_t) &= \mathbf{g}_t, \end{aligned}$$

and by basic algebra it is straightforward to show that

$$\begin{aligned} \hat{W}(y) = \min_{\boldsymbol{\alpha}\in\Delta_T} \Bigg[&\frac{L}{2}\|y - \sum_{t\in[T]}\alpha_t(\mathbf{x}_t - \frac{1}{L}\mathbf{g}_t)\|^2 - \frac{L}{4}\|\sum_{t\in[T]}\alpha_t(\mathbf{x}_t - \frac{1}{L}\mathbf{g}_t)\|^2 \\ &+ \sum_{t\in[T]}\alpha_t(f_t - \frac{1}{2L}\|\mathbf{g}_t\|^2 + \frac{L}{4}\|\mathbf{x}_t - \frac{1}{L}\mathbf{g}_t\|^2) \Bigg]. \end{aligned} \tag{12}$$

We have

$$\begin{aligned} \hat{W}(y) &\geq \min_{\boldsymbol{\alpha}\in\Delta_T} -\frac{L}{4}\sum_{t\in[T]}\|\alpha_t(\mathbf{x}_t - \frac{1}{L}\mathbf{g}_t)\|^2 + \sum_{t\in[T]}\alpha_t(f_t - \frac{1}{2L}\|\mathbf{g}_t\|^2 + \frac{L}{4}\|\mathbf{x}_t - \frac{1}{L}\mathbf{g}_t\|^2) \\ &\geq \min_{\boldsymbol{\alpha}\in\Delta_T} -\frac{L}{4}\sum_{t\in[T]}\alpha_t\|\mathbf{x}_t - \frac{1}{L}\mathbf{g}_t\|^2 + \sum_{t\in[T]}\alpha_t(f_t - \frac{1}{2L}\|\mathbf{g}_t\|^2 + \frac{L}{4}\|\mathbf{x}_t - \frac{1}{L}\mathbf{g}_t\|^2) \\ &= \min_{\boldsymbol{\alpha}\in\Delta_T}\sum_{t\in[T]}\alpha_t(f_t - \frac{1}{2L}\|\mathbf{g}_t\|^2) = \min_{t\in[T]}(f_t - \frac{1}{2L}\|\mathbf{g}_t\|^2) = f_j - \frac{1}{2L}\|\mathbf{g}_j\|^2, \end{aligned}$$

where the second inequality follows from the convexity of the squared norm. Finally we conclude the proof by establishing an upper bound that matches the lower bound on $\hat{W}^*$. Indeed,

$$\hat{W}(\mathbf{x}_j - \frac{1}{L}\mathbf{g}_j) \leq f_j - \frac{1}{2L}\|\mathbf{g}_j\|^2,$$

where the inequality follows, as in the convex case, by taking $\boldsymbol{\alpha} = \mathbf{e}_j$ in (12). □

A.2. Proof of Lemma 1

Theorem 7 will be the main tool used in the proof of Lemma 1 below.

Proof of Lemma 1. By Theorem 7, it is sufficient to show that there is a choice β for the value of $\hat{f}(\mathbf{y}_1),\ldots,\hat{f}(\mathbf{y}_m)$ with $\beta \geq \min_{i\in[n]} f(\mathbf{z}_i) - \frac{1}{L}\|\boldsymbol{\gamma}\|^2$ such that the set

$$\{(\mathbf{y}_j,\boldsymbol{\gamma},\beta)\}_{j\in[m]} \cup \{(\mathbf{z}_i,\boldsymbol{\gamma},f(\mathbf{z}_i)\}_{i\in[n]},$$

satisfies the interpolation conditions (10). The lower bound on $\hat{f}$ will then immediately follow since

$$\hat{f}(\mathbf{y}_j) - \frac{1}{2L}\|\boldsymbol{\gamma}\|^2 = \beta - \frac{1}{2L}\|\gamma\|^2 \geq \min_{k\in[n]} f(\mathbf{z}_k) - \frac{3}{2L}\|\boldsymbol{\gamma}\|^2, \quad \forall j\in[m],$$

and

$$\hat{f}(\mathbf{z}_i) - \frac{1}{2L}\|\boldsymbol{\gamma}\|^2 \geq \min_{k\in[n]} f(\mathbf{z}_k) - \frac{3}{2L}\|\boldsymbol{\gamma}\|^2, \quad \forall i\in[n].$$

In order to establish that the interpolation conditions hold, first note that all the interpolation conditions involving two points from $\{\mathbf{z}_i\}$ are naturally satisfied by the assumption that there exists some function with L-Lipschitz gradient (namely f) that interpolates $\{(\mathbf{z}_i,\nabla f(\mathbf{z}_i),f(\mathbf{z}_i)\} = \{(\mathbf{z}_i,\boldsymbol{\gamma},f(\mathbf{z}_i)\}$, and further note that by the assumptions, it follows that $\langle\boldsymbol{\gamma},\mathbf{y}_i-\mathbf{y}_j\rangle = 0$, hence the interpolation conditions involving both points in $\mathbf{y}_i$ are also trivially satisfied. We conclude that we only need to consider (10) for cases where one of the points is $\mathbf{z}_i$ and the other is $\mathbf{y}_j$, i.e., we are left the following set of inequalities:

$$-\frac{L}{4}\|\mathbf{z}_i-\mathbf{y}_j\|^2 \leq f(\mathbf{z}_i) - \beta - \langle\boldsymbol{\gamma},\mathbf{z}_i-\mathbf{y}_j\rangle, \quad i\in[n],\ j\in[m],$$
$$-\frac{L}{4}\|\mathbf{y}_j-\mathbf{z}_i\|^2 \leq \beta - f(\mathbf{z}_i) - \langle\boldsymbol{\gamma},\mathbf{y}_j-\mathbf{z}_i\rangle, \quad i\in[n],\ j\in[m].$$

Clearly, these inequalities hold if and only if

$$\beta \in [\max_{i,j}\left(f(\mathbf{z}_i) + \langle\boldsymbol{\gamma},\mathbf{y}_j-\mathbf{z}_i\rangle - \frac{L}{4}\|\mathbf{y}_j-\mathbf{z}_i\|^2\right), \min_{i,j}\left(f(\mathbf{z}_i) + \langle\boldsymbol{\gamma},\mathbf{y}_j-\mathbf{z}_i\rangle + \frac{L}{4}\|\mathbf{z}_i-\mathbf{y}_j\|^2\right)].$$

Now, this range is non-empty since it contains $f(\mathbf{y}_1)$ (recall that $f(\mathbf{y}_1)=\cdots=f(\mathbf{y}_m)$, $\nabla f(\mathbf{y}_1)=\cdots=\nabla f(\mathbf{y}_m)=\boldsymbol{\gamma}$, and that the interpolation conditions for the set $\{(\mathbf{y}_j,\nabla f(\mathbf{y}_j),f(\mathbf{y}_j)\}\cup\{(\mathbf{z}_i,\nabla f(\mathbf{z}_i),f(\mathbf{z}_i)\}$ naturally hold), hence there exits some i,j such that the choice

$$\hat{\beta} := f(\mathbf{z}_i) + \langle\boldsymbol{\gamma},\mathbf{y}_j-\mathbf{z}_i\rangle + \frac{L}{4}\|\mathbf{z}_i-\mathbf{y}_j\|^2$$

is a feasible choice for β. We get

$$\begin{aligned}\hat{\beta} &= f(\mathbf{z}_i) + \frac{L}{4}\|\mathbf{z}_i-\mathbf{y}_j-\frac{2}{L}\boldsymbol{\gamma}\|^2 - \frac{1}{L}\|\boldsymbol{\gamma}\|^2 \\ &\geq \min_k f(\mathbf{z}_k) - \frac{1}{L}\|\boldsymbol{\gamma}\|^2.\end{aligned}$$

which concludes the proof, as all interpolation conditions are satisfied, hence a function with the claimed properties exists. □

A.3. Proof of Thm. 2, adaptive step-size case

Consider the general, adaptive step-size case:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_{\mathbf{x}_1,\ldots,\mathbf{x}_t}\cdot(\nabla f(\mathbf{x}_t)+\boldsymbol{\xi}_t), \quad t\in[T-1],$$
$$\mathbf{x}_{\text{out}} = \sum_{t=1}^{T}\zeta^{(t)}_{\mathbf{x}_1,\ldots,\mathbf{x}_T}\mathbf{x}_i.$$

Here, our goal is to show that constants η_t and ζ_t from the proof of the fixed step-size case can be chosen in such a way that the method, when applied on $f_{\{\eta_t\},\{\zeta_t\}}$ constructed above, chooses step sizes and aggregation coefficients that are almost surely equal to the selected constants, i.e,

$$\eta_{\mathbf{x}_1,\ldots,\mathbf{x}_t} \overset{\text{a.s.}}{=} \eta_t, \quad \zeta^{(t)}_{\mathbf{x}_1,\ldots,\mathbf{x}_T} \overset{\text{a.s.}}{=} \zeta_t,$$

and thus the proof for the fixed-step case can proceed without change.

We use the following procedure to select η_t and ζ_t. We start by executing the first step of the algorithm on the initial point $\mathbf{x}_1 = 0$ and f, where the constants $\{\eta_t\}$ and $\{\zeta_t\}$ are set to arbitrarily values. Note that 1. the first-order information of f at $\mathbf{x}_1$ is independent of the choice for η_t, ζ_t, and 2. the norm of the noise vector $\boldsymbol{\xi}_1$ and its inner product with $\mathbf{x}_1$ and $\nabla f(\mathbf{x}_1)$ are independent of the specific direction chosen for the noise; therefore, by the assumption on the step size $\eta_{\mathbf{x}_1}$, it is independent of the specific value for $\boldsymbol{\xi}_1$, i.e., it is almost surely a constant. We denote this constant by η_1.

We continue by executing the second step of the algorithm on f, using the value of η_1 chosen above while keeping the constants $\eta_2, \ldots, \eta_{T-1}$, $\{\zeta_t\}_{t\in[T]}$ set to arbitrary values. As in the first iteration, 1. the first-order information of f at $\mathbf{x}_2$ is independent of the specific choice for $\eta_2, \ldots, \eta_{T-1}$, $\{\zeta_t\}_{t\in[T]}$ and 2. the norm of the noise vector $\boldsymbol{\xi}_2$ and its inner product with $\mathbf{x}_1, \mathbf{x}_2, \nabla f(\mathbf{x}_1), \nabla f(\mathbf{x}_2)$ and $\boldsymbol{\xi}_1$ are independent of the specific direction chosen for the noise; therefore by the assumption on $\eta_{\mathbf{x}_1,\mathbf{x}_2}$ it is almost surely a constant. As before, we denote the step size performed by the algorithm η_2.

Continuing in this fashion, we obtain a set of constants $\eta_1, \ldots, \eta_{T-1}$ with the property that when applying the method on $f = f_{\{\eta_t\},\{\zeta_t\}}$, then for any choice of aggregation coefficients $\{\zeta_t\}$ the step-sizes chosen by the method are almost surely equal to $\{\eta_t\}$. Finally, executing the aggregation step, by the assumption on the aggregation function, the coefficients are almost surely constants, which we denote by $\zeta_1, \ldots, \zeta_T$. To conclude, we have found a function $f = f_{\{\eta_t\},\{\zeta_t\}}$ such that the step sizes performed by the method on f are almost surely $\eta_1, \ldots, \eta_{T-1}$ and the aggregation coefficients chosen by the method are almost surely $\zeta_1, \ldots, \zeta_T$, hence the proof can continue as in the fixed-step case.

A.4. Proof of Proposition 1

We will utilize the following function:

$$f(\mathbf{x}) = \frac{1}{4\max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}} \cdot \langle \mathbf{x}, \mathbf{e}_1\rangle^2$$

and assume that the initialization $\mathbf{x}_1$ is

$$\mathbf{x}_1 := \left(\sqrt{\Delta \cdot \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}},\ 0, 0, \ldots, 0\right) .$$

It is easily verified that f has L-Lipschitz gradient, and that $f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x}) < \Delta$. Moreover,

$$\|\nabla f(\mathbf{x})\| = \frac{|\langle \mathbf{x}, \mathbf{e}_1\rangle|}{2\max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}. \tag{13}$$

Hereafter, for the sake of simplicity we drop the subscript indicating the coordinate number, and let x_t denote the first coordinate of iterate t, and ξ_t the first coordinate of the noise at iteration t.

We now turn to show that when d is large enough

$$\min_{t\in[T]} |x_t| \geq \frac{2}{5}\sqrt{\Delta \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}} \tag{14}$$

holds with arbitrarily high probability, which together with (13) implies the desired result.

The dynamics of SGD on the first coordinate is as follows: we initially have

$$x_1 = \sqrt{\Delta \cdot \max\{1/L, \sum_{t=1}^{T-1}\eta_t\}},$$

and

$$x_{t+1} = \left(1 - \frac{\eta_t}{2\max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}\right) x_t - \eta_t \xi_t.$$

Unrolling this recurrence, we have for any t

$$\begin{aligned} x_t = &\sqrt{\Delta \cdot \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}} \cdot \prod_{j=1}^{t-1}\left(1 - \frac{\eta_j}{2\max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}\right) \\ &- \sum_{j=1}^{t-1}\eta_j\xi_j \prod_{i=j+1}^{t-1}\left(1 - \frac{\eta_j}{2\max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}\right), \end{aligned} \tag{15}$$

where we use the convention that $\prod_{i=a}^{b} c_i$ is always 1 if $b < a$. Since each ξ_j is a zero-mean independent Gaussian, x_t is also Gaussian with

$$\begin{aligned} \mathbb{E}[x_t] &= \sqrt{\Delta \cdot \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}} \cdot \prod_{j=1}^{t-1}\left(1 - \frac{\eta_j}{2\max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}\right) \\ &\geq \sqrt{\Delta \cdot \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}} \cdot \exp\left(\ln\frac{1}{2} \cdot \sum_{j=1}^{t-1}\frac{\eta_j}{\max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}\right) \\ &\geq \frac{1}{2}\sqrt{\Delta \cdot \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}, \end{aligned}$$

here we used the assumption $\eta_t \geq 0$ and the fact that $1 - z/2 \geq \exp(\ln\frac{1}{2} \cdot z)$ for all $z \in [0, 1]$. In addition,

$$\begin{aligned} \mathbb{V}[x_t] &= \sum_{j=1}^{t-1}\eta_j^2\mathbb{V}[\xi_j]\prod_{i=j+1}^{t-1}\left(1 - \frac{\eta_j}{2\max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}\right)^2 \\ &\leq \sum_{j=1}^{t-1}\eta_j^2\mathbb{V}[\xi_j] \leq \frac{\sigma^2(T-1)}{L^2 d}, \end{aligned}$$

which follows since each ξ_j is independent and with variance at most σ^2/d, and $0 \leq \eta_j \leq 1/L$. Choosing

$$d \geq d_0 := \frac{\Phi^{-1}(1-\delta/T)^2\sigma^2(T-1)}{\left(\frac{1}{2} - \frac{2}{5}\right)^2 L^2\Delta \cdot \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}} = \mathcal{O}(\log(T/\delta)\sigma^2 T/(L^2\Delta)),$$

where Φ^{-1} is the inverse CDF of the normal distribution, we get that for all t with $\mathbb{V}[x_t] > 0$

$$\begin{aligned} &\Pr\left(x_t \geq \frac{2}{5}\sqrt{\Delta \cdot \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}\right) \\ &= \Pr\left(\frac{x_t - \mathbb{E}x_t}{\sqrt{\mathbb{V}[x_t]}} \geq -\frac{\left(\frac{1}{2} - \frac{2}{5}\right)\sqrt{\Delta \cdot \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}}{\sqrt{\sigma^2(T-1)/(L^2 d)}}\right) \\ &\geq \Pr\left(\frac{x_t - \mathbb{E}x_t}{\sqrt{\mathbb{V}[x_t]}} \geq -\frac{\left(\frac{1}{2} - \frac{2}{5}\right)\sqrt{\Delta \cdot \max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}}}{\sqrt{\sigma^2(T-1)/(L^2 d_0)}}\right) = 1 - \delta/T, \end{aligned}$$

and furthermore, the same bound holds almost surely for all t with $\mathbb{V}[x_t] = 0$. Finally, taking a union bound over t, we conclude that this lower bound holds for all x_t with probability $1-\delta$, which implies (14) as required.

A.5. Proof of Proposition 2

To prove the proposition, we will need the following Lemma, which formalizes the fact that the norm of high-dimensional Gaussian random variables tend to be concentrated around a fixed value:

Lemma 2. *Let $M, \gamma > 0$ be fixed. For any d, let $\mathbf{x}_d$ be a random variable normally distributed with $\mathbf{x}_d \sim \mathcal{N}(\mathbf{u}, \frac{\gamma}{d} I_d)$, where $\mathbf{u}$ is some vector in $\mathbb{R}^d$ with $\|\mathbf{u}\|^2 = M$. Then for any $\epsilon \in (0,1)$,*

$$\Pr\left(\left|\frac{\|\mathbf{x}_d\|^2}{M+\gamma} - 1\right| \leq \epsilon\right) \;\geq\; 1 - 4\exp\left(-\frac{d\epsilon^2}{24}\right).$$

Proof. Consider $\mathbf{x}_d$ for some fixed d. We can decompose it as $\mathbf{u} + \sqrt{\frac{\gamma}{d}}\mathbf{n}$, where $\mathbf{n}$ has a standard Gaussian distribution in $\mathbb{R}^d$ (zero mean and covariance matrix being the identity). Thus,

$$\begin{aligned}\frac{\|\mathbf{x}_d\|^2}{M+\gamma} - 1 \;&=\; \frac{\|\mathbf{u}\|^2 + 2\sqrt{\frac{\gamma}{d}}\mathbf{u}^\top\mathbf{n} + \frac{\gamma}{d}\|\mathbf{n}\|^2}{M+\gamma} - 1 \;=\; \frac{2\sqrt{\frac{\gamma}{d}}\mathbf{u}^\top\mathbf{n} + \gamma\left(\frac{1}{d}\|\mathbf{n}\|^2 - 1\right)}{M+\gamma} \\ &=\; \frac{2\sqrt{\gamma/d}}{M+\gamma}\mathbf{u}^\top\mathbf{n} + \frac{\gamma}{M+\gamma}\left(\frac{1}{d}\|\mathbf{n}\|^2 - 1\right). \end{aligned} \tag{16}$$

The first term in the sum above is distributed as a Gaussian in $\mathbb{R}$ with zero mean and variance $\frac{4\gamma}{d(M+\gamma)^2}\|\mathbf{u}\|^2 = \frac{4\gamma M}{d(M+\gamma)^2} \leq \frac{4\gamma M}{d\cdot 2\gamma M} = \frac{2}{d}$. By a standard Gaussian tail bound, it follows that the probability that it exceeds $\epsilon/2$ in absolute value is at most $2\exp(-d\epsilon^2/16)$. Similarly, for the second term, we have by a standard tail bound for Chi-squared random variables (see for example (Shalev-Shwartz & Ben-David, 2014), Lemma B.12) that

$$\begin{aligned}&\Pr\left(\frac{\gamma}{M+\gamma}\left|\frac{1}{d}\|\mathbf{n}\|^2 - 1\right| \geq \frac{\epsilon}{2}\right) \\ &\leq \Pr\left(\left|\frac{1}{d}\|\mathbf{n}\|^2 - 1\right| \geq \frac{\epsilon}{2}\right) \leq 2\exp(-d\epsilon^2/24).\end{aligned}$$

Combining the above with a union bound, it follows that (16) has absolute value more than ϵ with probability at most

$$2\exp(-d\epsilon^2/16) + 2\exp(-d\epsilon^2/24) \;\leq\; 4\exp(-d\epsilon^2/24)\,.$$

□

Proof of Proposition 2. We will utilize the function

$$f(\mathbf{x}) = \frac{L}{2}\|\mathbf{x}\|^2,$$

where $\mathbf{x}_1$ is some vector such that $\|\mathbf{x}_1\| = \sqrt{\Delta/L}$. Using a derivation similar to the one used in (15), we have

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \cdot (L\mathbf{x}_t + \boldsymbol{\xi}_t) = (1 - L\eta_t)\mathbf{x}_t - \eta_t\boldsymbol{\xi}_t,$$

hence

$$\mathbf{x}_t = \prod_{j=1}^{t-1}(1 - L\eta_j)\,\mathbf{x}_1 - \sum_{j=1}^{t-1}\eta_j \prod_{i=j+1}^{t-1}(1 - L\eta_i)\,\boldsymbol{\xi}_j. \tag{17}$$

Since each $\boldsymbol{\xi}_j$ is an independent zero-mean Gaussian with covariance matrix $\frac{\sigma^2}{d}I_d$, we get that $\mathbf{x}_t$ has a Gaussian distribution with mean $\prod_{j=1}^{t-1}(1 - L\eta_j)\,\mathbf{x}_1$ and covariance matrix $\frac{\gamma_t}{d}I_d$, where

$$\gamma_t = \sigma^2 \sum_{j=1}^{t-1}\eta_j^2 \prod_{i=j+1}^{t-1}(1 - L\eta_i)^2\,.$$

By Lemma 2, taking $\epsilon = 1/2$ and $d \geq d_0$ with

$$d_0 := 96 \log \frac{4T}{\delta} = \mathcal{O}(\log(T/\delta)),$$

it follows that $\|\nabla f(\mathbf{x}_t)\|^2 = \|L\mathbf{x}_t\|^2$ is at least

$$\frac{L}{2} \cdot \left(\Delta \prod_{j=1}^{t-1} (1 - L\eta_j)^2 + L\sigma^2 \sum_{j=1}^{t-1} \eta_j^2 \prod_{i=j+1}^{t-1} (1 - L\eta_i)^2 \right) \tag{18}$$

with probability at least $1 - \delta/T$. Our goal now will be to lower bound (18) under the conditions in the proposition. Plugging this lower bound and applying a union bound over all $t \in [T]$ will result in our proposition.

- If $\eta_t = \eta$ and $\eta \in [0, 1/L)$, we can lower bound (18) by

$$\begin{aligned}
&\frac{L}{2} \left(\Delta(1 - L\eta)^{2(t-1)} + L\sigma^2 \sum_{j=1}^{t-1} \eta^2 (1 - L\eta)^{2(t-1-j)} \right) \\
&= \frac{L}{2} \left(\Delta(1 - L\eta)^{2(t-1)} + L\sigma^2\eta^2 \cdot \frac{1 - (1 - L\eta)^{2(t-1)}}{1 - (1 - L\eta)^2} \right) \\
&= \frac{L}{2} \left(\Delta(1 - L\eta)^{2(t-1)} + \frac{\eta\sigma^2}{2 - L\eta} \left(1 - (1 - L\eta)^{2(t-1)} \right) \right).
\end{aligned}$$

  For any t, this is a convex combination of $\frac{L}{2}\Delta$ and $\frac{L}{2}\frac{\eta\sigma^2}{2-L\eta}$, hence is at least the minimum between them.

- If there exists some constant $c \geq 0$ such that $\eta_t \geq c/L$ for all t, we can lower bound (18) by $\frac{L^2}{2}\sigma^2\eta_{t-1}^2 \geq \frac{\sigma^2 c^2}{2}$ (i.e., accounting for the noise at the last iterate).

- If $\eta_t = \frac{a}{L(b+t^\theta)}$ (where $a > 0, b \geq 0, \theta \in (0, 1/2)$), then it is easily verified that for a certain constant $\tau_{a,b,\theta}$ depending only on a, b, θ,

$$1 \leq \frac{1}{L\eta_t} \leq \frac{t}{2} \text{ for all } t \geq \tau_{a,b,\theta}.$$

  In that case, we can lower bound (18) by

$$\begin{aligned}
&\frac{L^2\sigma^2}{2} \sum_{j=t-\lfloor 1/(L\eta_t) \rfloor}^{t-1} \eta_j^2 \prod_{i=j+1}^{t-1} (1 - L\eta_i)^2 \\
&\geq \frac{L^2\sigma^2\eta_t^2}{2} \cdot \left\lfloor \frac{1}{L\eta_t} \right\rfloor \left(1 - L\eta_{\lfloor t/2 \rfloor}\right)^{2\lfloor 1/(L\eta_t) \rfloor} \\
&\geq \frac{\sigma^2 L\eta_t}{4} \left(1 - \frac{a}{b + \lfloor t/2 \rfloor^\theta} \right)^{2\left\lfloor \frac{b+t^\theta}{a} \right\rfloor},
\end{aligned}$$

  which is at least $c_{a,b,\theta}\sigma^2 L\eta_t \geq c_{a,b,\theta}\sigma^2 L\eta_T$ if $t \geq \tau'_{a,b,\theta}$ (for some parameters $c_{a,b,\theta}, \tau'_{a,b,\theta}$ depending on a, b, θ). Moreover, if $t < \tau'_{a,b,\theta}$, then (18) is at least

$$\frac{L^2\sigma^2}{2}\eta_{t-1}^2 \geq \frac{L^2\sigma^2}{2}\eta_{\tau'_{a,b,\theta}}^2 = \frac{\sigma^2}{2} \cdot \left(\frac{a}{b + (\tau'_{a,b,\theta})^\theta} \right)^2.$$

  Combining both cases, we get that (18) is at least $c'_{a,b,\theta}\sigma^2 \cdot \min\{1, L\eta_T\}$, where $c'_{a,b,\theta}$ is again some constant dependent on a, b, θ, implying the stated result.

□

B. Upper Bounds for SGD

In order to place our lower bounds in perspective, we state and prove a rather standard $\mathcal{O}(\epsilon^{-4})$ complexity bound for SGD, which unlike the result discussed in the introduction, does not assume anything special about the Hessians or the noise, and is completely independent of the dimension.

We start the analysis with a technical lemma that we will use to derive bounds both in the stochastic and deterministic settings.

Lemma 3. *Consider the Stochastic Gradient Descent*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \left(\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t\right), \quad t \in [T-1],$$

where $0 < \eta_t < 1/L$, f is a non-convex function with L-Lipschitz gradient, and $\boldsymbol{\xi}_t$ is a random noise with $\mathbb{E}(\boldsymbol{\xi}_t) = 0$, $V(\boldsymbol{\xi}_t) = \sigma^2$. Then for any choice of κ_t, $t \in [T-1]$ such that $1 - L\eta_t \leq \kappa_t \leq (1 - L\eta_t)^{-1}$ we have

$$\min_{t\in[T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{4L(f(\mathbf{x}_1) - f(\mathbf{x}_*)) + \sum_{t=1}^{T-1} \frac{L^2\eta_t^2(1-L\eta_t+\kappa_t)}{1-L\eta_t}\sigma^2}{3(T-1) - \sum_{t=1}^{T-1}(1-L\eta_t)(1-L\eta_t+\kappa_t+\frac{1}{\kappa_t})},$$

where x_* is a stationary point with $f(\mathbf{x}_*) \leq f(\mathbf{x}_T)$.

Proof. By Thm. 6 we have

$$\begin{aligned}
&\frac{1}{2L}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1})\|^2 - \frac{L}{4}\|\mathbf{x}_t - \mathbf{x}_{t+1} - \frac{1}{L}(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1}))\|^2 \\
&\quad \overset{\text{a.s.}}{\leq} f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) - \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_t - \mathbf{x}_{t+1}\rangle, \quad t \in [T-1],
\end{aligned}$$

which by the definition of $\mathbf{x}_{t+1}$ becomes

$$\begin{aligned}
&\frac{1}{2L}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1})\|^2 - \frac{1}{4L}\|L\eta_t\boldsymbol{\xi}_t - (1-L\eta_t)\nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_{t+1})\|^2 \\
&\quad \overset{\text{a.s.}}{\leq} f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) - \langle \nabla f(\mathbf{x}_{t+1}), \eta_t\left(\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t\right)\rangle, \quad t \in [T-1],
\end{aligned}$$

Adding up the inequality above for all $t \in [T-1]$ brings us to

$$\begin{aligned}
&\frac{1}{2L}\sum_{t=1}^{T-1}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1})\|^2 - \frac{1}{4L}\sum_{t=1}^{T-1}\|L\eta_t\boldsymbol{\xi}_t - (1-L\eta_t)\nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_{t+1})\|^2 \\
&\quad + \sum_{t=1}^{T-1}\eta_t\langle \nabla f(\mathbf{x}_{t+1}), \nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t\rangle \overset{\text{a.s.}}{\leq} f(\mathbf{x}_1) - f(\mathbf{x}_T),
\end{aligned}$$

which, after adding

$$\frac{1}{4}\sum_{t=1}^{T-1}\eta_t(1-L\eta_t+\kappa_t)\left(\frac{L\eta_t}{1-L\eta_t}\|\boldsymbol{\xi}_t\|^2 + 2\langle\nabla f(\mathbf{x}_t), \boldsymbol{\xi}_t\rangle\right)$$

to both sides and rearranging the terms, brings us to

$$\begin{aligned}
&\frac{1}{4L}\sum_{t=1}^{T-1}\left(2 - (1-L\eta_t)(1-L\eta_t+\kappa_t)\right)\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{4L}\sum_{t=1}^{T-1}\left(1 - \frac{1-L\eta_t}{\kappa_t}\right)\|\nabla f(\mathbf{x}_{t+1})\|^2 \\
&\quad + \frac{1}{4L}\sum_{t=0}^{T-1}(1-L\eta_t)\kappa_t\left\|\nabla f(\mathbf{x}_t) - \frac{1}{\kappa_t}\nabla f(\mathbf{x}_{t+1}) - \frac{L\eta_t}{1-L\eta_t}\boldsymbol{\xi}_t\right\|^2 \\
&\quad \overset{\text{a.s.}}{\leq} f(\mathbf{x}_1) - f(\mathbf{x}_T) + \frac{1}{4}\sum_{t=0}^{T-1}\eta_t(1-L\eta_t+\kappa_t)\left(\frac{L\eta_t}{1-L\eta_t}\|\boldsymbol{\xi}_t\|^2 + 2\langle\nabla f(\mathbf{x}_t), \boldsymbol{\xi}_t\rangle\right)
\end{aligned}$$

Finally, taking the expected value of both side, and noting that $\mathbb{E}\|\boldsymbol{\xi}_t\| = \sigma$, $\mathbb{E}\langle \nabla f(\mathbf{x}_t), \boldsymbol{\xi}_t\rangle = 0$, and $\mathbb{E}f(\mathbf{x}_T) \geq f(\mathbf{x}_*)$, we reach

$$\frac{1}{4L}\left(3(T-1) - \sum_{t=1}^{T-1}(1-L\eta_t)(1-L\eta_t+\kappa_t) - \sum_{t=1}^{T-1}\frac{1-L\eta_t}{\kappa_t}\right)\min_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2$$
$$\leq f(\mathbf{x}_1) - f(\mathbf{x}_*) + \frac{1}{4}\sum_{t=1}^{T-1}\eta_t(1-L\eta_t+\kappa_t)\frac{L\eta_t}{1-L\eta_t}\sigma^2,$$

concluding the proof. □

An explicit optimal expression for κ_i appears to be complex in the general case, however, for two important cases a good approximation can be obtained. First, when σ is large, the term in the numerator dominates the expression, thus the optimal value for κ_t approaches $1 - L\eta_t$ as $\sigma \to \infty$, recovering the following result by Ghadimi and Lan:

Theorem 8 ((Ghadimi & Lan, 2013), Theorem 2.1)**.** *Consider the fixed-step Stochastic Gradient Descent*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t\left(\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t\right), \quad t \in [T-1],$$

where f is a nonconvex function with L-Lipschitz gradient, $\boldsymbol{\xi}_t$ is a random noise with $\mathbb{E}(\boldsymbol{\xi}_t) = 0$, $V(\boldsymbol{\xi}_t) = \sigma^2$ and $0 < L\eta_t < 1$. Then

$$\min_{t\in[T]}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2(f(\mathbf{x}_1) - f(\mathbf{x}_*)) + L\sum_{t=1}^{T-1}\eta_t^2\sigma^2}{\sum_{t=1}^{T-1}\eta_t(2-L\eta_t)}. \tag{19}$$

where x_* is a stationary point with $f(\mathbf{x}_*) \leq f(\mathbf{x}_T)$.

Proof. The result follows directly from Lemma 3, taking $\kappa_t = 1 - L\eta_t$. □

A second case where a simple expression for κ_t can be easily attained is when $\sigma = 0$, i.e., in the deterministic case. Here an optimal choice for κ is $\kappa_i = 1$, giving the following result which appears to be a new and slightly improved version of the classical result by Nesterov (Nesterov, 2004), eq. (1.2.15):

Corollary 1. *Consider the fixed-step Gradient Descent*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t\nabla f(\mathbf{x}_t), \quad t \in [T-1],$$

where f is a nonconvex function with L-Lipschitz gradient and $0 < \eta_t < 1/L$. Then

$$\min_{t\in[T]}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{4(f(\mathbf{x}_1) - f(\mathbf{x}_*))}{\sum_{t=1}^{T-1}\eta_t(4-L\eta_t)},$$

where x_* is a stationary point with $f(\mathbf{x}_*) \leq f(\mathbf{x}_T)$.

Remark 3. The discovery of the proof of Lemma 3 was guided by numerically solving an optimization problem called the Performance Estimation Problem, whose solution captures the worst-case performance of the SGD method. This technique was first introduced in (Drori & Teboulle, 2014) and was later shown in (Taylor et al., 2017b) to achieve tight bounds for a wide range of methods in the deterministic case. This, in conjunction with the nearly matching lower bound established in Thm. 2, motivates us to raise the conjecture that Lemma 3 gives a tight bound (including the constant) in the stochastic case.