A Auxiliary Results

In this appendix, we state and collect lemmas and propositions required to prove the main results.

Notation. In the following sections, we denote with $\mathcal{F}_k$ the filtration $\sigma(G_1, \cdots, G_{k-1})$. Moreover, to simplify the notation, we define g_k as the gradient surrogate in eq.(2) at time-step k i.e. $g_k := g_{(G_k,h_k)}(x_k)$ and $g(\cdot) := g_{(G,h)}(\cdot)$ for an arbitrary $G \in O(d)$ and $h > 0$. We denote the normalized Haar measure [37] by μ. We define the unit ball $\mathbb{B}^d$ and the unit sphere $\mathbb{S}^{d-1}$ as follow

$$\mathbb{B}^d := \{v \in \mathbb{R}^d \mid \|v\| \leq 1\} \qquad \text{and} \qquad \mathbb{S}^{d-1} := \{v \in \mathbb{R}^d \mid \|v\| = 1\}.$$

We denote by σ and σ_N the spherical measure and the normalized spherical measure on $\mathbb{S}^{d-1}$, respectively. Moreover, we denote with $I_{d,\ell} \in \mathbb{R}^{d\times \ell}$ the (truncated) identity matrix.

Lemma 2. *Let $\beta(\mathbb{S}^{d-1})$ be the surface area of $\mathbb{S}^{d-1}$ and let $I \in \mathbb{R}^{d\times d}$ be the identity matrix. Then,*

$$\int_{\mathbb{S}^{d-1}} vv^\intercal \, d\sigma(v) = \frac{\beta(\mathbb{S}^{d-1})}{d} I.$$

Proof. This result is proved in [21, Lemma 7.3, point (b)]. ☐

Lemma 3. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a L-Lipschitz function . If u is uniformly distributed on $\mathbb{S}^{d-1}$, then*

$$(\mathbb{E}[\phi(u) - \mathbb{E}[\phi(u)]])^2 \leq c\frac{L^2}{d},$$

for some numerical constant $c > 0$.

Proof. The proof follows the same line as [46, Lemma 9]. ☐

A.1 Smoothing Lemma & Properties

In this appendix, we provide the proof of the Smoothing Lemma (i.e. Lemma 1).

Proof of Smoothing Lemma. By eq. (2),

$$\mathbb{E}_G[g_{(G,h)}(x)] = \frac{d}{\ell}\sum_{i=1}^{\ell} \int_{O(d)} \frac{f(x + hGe_i) - f(x - hGe_i)}{2h} Ge_i \, d\mu(G).$$

By [37, Theorem 3.7],

$$\mathbb{E}_G[g_{(G,h)}(x)] = \frac{d}{2\ell h}\sum_{i=1}^{\ell} \int_{\mathbb{S}^{d-1}} (f(x + hv^{(i)}) - f(x - hv^{(i)}))v^{(i)} \, d\sigma_N(v^{(i)}).$$

Since $v^{(i)}$ is uniformly distributed on the sphere, which is symmetric with respect to the origin, we have

$$\mathbb{E}_G[g_{(G,h)}(x)] = \frac{d}{\ell h}\sum_{i=1}^{\ell} \int_{\mathbb{S}^{d-1}} f(x + hv^{(i)})v^{(i)} \, d\sigma_N(v^{(i)}).$$

As a consequence of Stokes' Theorem (details in [18, Lemma 1] and [1, Theorem A8.8]), we get

$$\mathbb{E}[g_{(G,h)}(x)] = \frac{1}{\ell}\sum_{i=1}^{\ell} \nabla f_h(x) \qquad \text{with} \qquad f_h(x) := \frac{1}{\text{vol}(\mathbb{B}^d)} \int_{\mathbb{B}^d} f(x + hu) \, du.$$

Rearranging terms, we get the claim. ☐

Proposition 1 (Smoothing properties). *Let f_h be the smooth approximation of f defined in eq.* (4). *Then the following hold:*
If f is convex then f_h is convex and, for every $x \in \mathbb{R}^d$,

$$f(x) \leq f_h(x).$$

If f is L_0-Lipschitz continuous - i.e. $\forall x, y \in \mathbb{R}^d$, $|f(x) - f(y)| \leq L_0\|x - y\|$, then f_h is L_0-Lipschitz continuous, differentiable and for every $x, y \in \mathbb{R}^d$

$$\|\nabla f_h(x) - \nabla f_h(y)\| \leq \frac{L_0\sqrt{d}}{h}\|x - y\| \quad \textit{and} \quad f_h(x) \leq f(x) + L_0 h.$$

If f is L_1-smooth - i.e. f is differentiable and $\forall x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \leq L_1\|x - y\|$ then f_h is L_1-smooth and for every $x \in \mathbb{R}^d$,

$$\|\nabla f_h(x) - \nabla f(x)\| \leq \frac{hdL_1}{2} \quad \textit{and} \quad f_h(x) \leq f(x) + \frac{L_1}{2}h^2.$$

Proof. These are standard results proposed and proved in different works - see for example [16, Lemma 8],[21, Proposition 7.5],[34, Proposition 2.2],[49]. □

Lemma 4 (Approximation Error)**.** *Let $g(\cdot)$ be the surrogate defined in eq.* (2) *for arbitrary $h > 0$ and $G \in O(d)$. Then the following hold:*

(i) If f is L_0-Lipschitz (see Assumption 1), then, for every $x \in \mathbb{R}^d$,

$$\mathbb{E}_G[\|g(x)\|^2] \leq 2c\frac{dL_0^2}{\ell},$$

where c is a numerical constant.

(ii) If f is L_1-smooth (see Assumption 3), then, for every $x \in \mathbb{R}^d$,

$$E_G[\|g(x)\|^2] \leq \frac{2d}{\ell}\|\nabla f(x)\|^2 + \frac{L_1^2 d^2}{2\ell}h^2.$$

Proof. Note that, since directions are orthogonal, we have

$$\mathbb{E}_G[\|g(x)\|^2] = \frac{d^2}{4\ell^2h^2}\sum_{i=1}^{\ell}\mathbb{E}_G[(f(x + hGe_i) - f(x - hGe_i))^2\|Ge_i\|^2].$$

By [37, Theorem 3.7],

$$\mathbb{E}_G[\|g(x)\|^2] = \frac{d^2}{4\ell^2h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v_i}[(f(x + hv^{(i)}) - f(x - hv^{(i)}))^2\|v^{(i)}\|^2], \tag{5}$$

where each $v^{(i)}$ is uniformly distributed on $\mathbb{S}^{d-1}$.
(i): Set $\gamma = \mathbb{E}_{v^{(i)}}[f(x + hv^{(i)})]$ for every i (this expectation does not depend on i). Then

$$\begin{aligned}
\mathbb{E}_G[\|g(x)\|^2] &= \frac{d^2}{4\ell^2h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v^{(i)}}[(f(x + hv^{(i)}) - f(x - hv^{(i)}) + \gamma - \gamma)^2\|v^{(i)}\|^2] \\
&= \frac{d^2}{4\ell^2h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v^{(i)}}[((f(x + hv^{(i)}) - \gamma) - (f(x - hv^{(i)}) - \gamma))^2\|v^{(i)}\|^2] \\
&\leq \frac{d^2}{2\ell^2h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v^{(i)}}[((f(x + hv^{(i)}) - \gamma)^2 + (f(x - hv^{(i)}) - \gamma)^2)\|v^{(i)}\|^2] \\
&= \frac{d^2}{2\ell^2h^2}\sum_{i=1}^{\ell}\Big[\mathbb{E}_{v^{(i)}}[(f(x + hv^{(i)}) - \gamma)^2\|v^{(i)}\|^2] \\
&\quad + \mathbb{E}_{v^{(i)}}[(f(x - hv^{(i)}) - \gamma)^2\|v^{(i)}\|^2]\Big].
\end{aligned}$$

Since $v^{(i)}$ is uniformly distributed on $\mathbb{S}^{d-1}$, it satisfies $\|v^{(i)}\|^2 = 1$ and by symmetry we have

$$\mathbb{E}_G[\|g(x)\|^2] \leq \frac{d^2}{\ell^2h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v^{(i)}}[(f(x + hv^{(i)}) - \gamma)^2].$$

The definition of γ yields

$$\begin{aligned}\mathbb{E}_G[\|g(x)\|^2] &\leq \frac{d^2}{\ell^2 h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v^{(i)}}[((f(x+hv^{(i)})-\gamma)^2]\\ &= \frac{d^2}{\ell^2 h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v^{(i)}}[(f(x+hv^{(i)})-\mathbb{E}_{v^{(i)}}[f(x+hv^{(i)})])^2].\end{aligned}$$

The claim follows by Lemma 3 and the fact that $f(x+hv^{(i)})$ is hL_0-Lipschitz continuous w.r.t to $v^{(i)}$.
(ii): Equation (5) yields

$$\begin{aligned}\mathbb{E}_G[\|g(x)\|^2] &= \frac{d^2}{4\ell^2 h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v^{(i)}}[(f(x+hv^{(i)})-f(x-hv^{(i)})-f(x)+f(x))^2\|v^{(i)}\|^2]\\ &\leq \frac{d^2}{2\ell^2 h^2}\sum_{i=1}^{\ell}\Big[\mathbb{E}_{v^{(i)}}[(f(x+hv^{(i)})-f(x))^2\|v^{(i)}\|^2]\\ &+\mathbb{E}_{v^{(i)}}[(f(x-hv^{(i)})-f(x))^2\|v^{(i)}\|^2]\Big]\\ &= \frac{d^2}{\ell^2 h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v^{(i)}}[(f(x+hv^{(i)})-f(x))^2],\end{aligned}$$

where the last equation follows by symmetry. Adding and subtracting $\left\langle \nabla f(x), hv^{(i)}\right\rangle$ we derive

$$\begin{aligned}\mathbb{E}_G[\|g(x)\|^2] &\leq \frac{d^2}{\ell^2 h^2}\sum_{i=1}^{\ell}\mathbb{E}_{v^{(i)}}\left[\left(f(x+hv^{(i)})-f(x)-\left\langle \nabla f(x), hv^{(i)}\right\rangle+\left\langle \nabla f(x), hv^{(i)}\right\rangle\right)^2\right]\\ &\leq \frac{2d^2}{\ell^2 h^2}\sum_{i=1}^{\ell}\left(\mathbb{E}_{v^{(i)}}\left[\left(f(x+hv^{(i)})-f(x)-\left\langle \nabla f(x), hv^{(i)}\right\rangle\right)^2\right]\right.\\ &\left.+\mathbb{E}_{v^{(i)}}\left[\left(\left\langle \nabla f(x), hv^{(i)}\right\rangle\right)^2\right]\right).\end{aligned}$$

Denote by $\beta(\mathbb{S}^{d-1})$ the surface area of $\mathbb{S}^{d-1}$. The Descent Lemma [41] implies

$$\begin{aligned}\mathbb{E}_G[\|g(x)\|^2] &\leq \frac{2d^2}{\ell^2 h^2}\sum_{i=1}^{\ell}\left[\left(\frac{L_1^2}{4}h^4\right)+\mathbb{E}\left[\left(\left\langle \nabla f(x), hv^{(i)}\right\rangle\right)^2\right]\right]\\ &= \frac{L_1^2 d^2}{2\ell}h^2+\frac{2d^2}{\ell^2 h^2}\sum_{i=1}^{\ell}\mathbb{E}\left[\left(\left\langle \nabla f(x), hv^{(i)}\right\rangle\right)^2\right]\\ &= \frac{L_1^2 d^2}{2\ell}h^2+\frac{2d^2}{\ell^2\beta(\mathbb{S}^{d-1})}\sum_{i=1}^{\ell}\int_{\mathbb{S}^{d-1}}\nabla f(x)^{\intercal}v^{(i)}v^{(i)\intercal}\nabla f(x)\,d\sigma(v).\end{aligned}$$

By Lemma 2, we get the claim. Indeed,

$$\begin{aligned}\mathbb{E}_G[\|g(x)\|^2] &\leq \frac{L_1^2 d^2}{2\ell}h^2+\frac{2d^2}{\ell^2\beta(\mathbb{S}^{d-1})}\sum_{i=1}^{\ell}\left(\frac{\beta(\mathbb{S}^{d-1})}{d}\|\nabla f(x)\|^2\right)\\ &= \frac{2d}{\ell}\|\nabla f(x)\|^2+\frac{L_1^2 d^2}{2\ell}h^2.\end{aligned}$$

□

A.2 Auxiliary results and proofs for the nonsmooth setting, convex, and nonconvex.

In this subsection, for every k, we will denote by $\mathcal{F}_k$ the σ-algebra $\sigma(G_0,\ldots,G_{k-1})$.

Lemma 5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a lower semi-continuous function and denote with $S = \arg\min f$ and $f^* = \min f$. Then,*

$$\begin{cases} \text{(A)} \quad \forall x^* \in S, \, \exists \lim_k \|x_k - x^*\| \\ \text{(B)} \quad \liminf_k f(x_k) = f^* \end{cases} \implies \exists x_\infty \in S \quad \textit{s.t.} \quad x_k \to x_\infty.$$

Proof. Since (B) holds, we have that exists $(x_{k_j})_{j\in\mathbb{N}}$ subsequence of $(x_k)_{k\in\mathbb{N}}$ such that $f(x_{k_j}) \to f^*$. Since $S \neq \emptyset$ and (A) we have that

$$\exists x^* \in S \quad \text{and} \quad \exists \lim_k \|x_k - x^*\|.$$

Thus, the sequence $(x_k)_{k\in\mathbb{N}}$ is bounded and, therefore, also $(x_{k_j})_{j\in\mathbb{N}}$ is bounded. Taking a convergent subsequence $(x_{k_{j_n}})_{n\in\mathbb{N}}$ of $(x_{k_j})_{j\in\mathbb{N}}$, we have that exists x_∞ s.t.

$$x_{k_{j_n}} \to x_\infty.$$

Since f is assumed to be lower semi-continuous, we have that

$$f(x_\infty) \leq \liminf_n f(x_{k_{j_n}}) = f^* = \lim_j f(x_{k_j}).$$

Thus, we have that $x_\infty \in S$ which implies, by (A), that

$$\exists \lim_k \|x_k - x_\infty\| \quad \text{and} \quad \lim_n \|x_{k_{j_n}} - x_\infty\| = 0.$$

Hence, since $x_{k_{j_n}}$ is a subsequence of x_k,

$$\lim_k \|x_k - x_\infty\| = 0,$$

and, therefore, $x_k \to x_\infty \in S$. ☐

Lemma 6 (Convergence: convex non-smooth)**.** *Assume that f is convex and L_0 Lipschitz continuous. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1 and let $x^* \in \arg\min f$. Then, for every $k \in \mathbb{N}$, the following inequality holds:*

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k] - \|x_k - x^*\|^2 + 2\alpha_k(f(x_k) - f(x^*)) \leq 2c\frac{L_0^2 d}{\ell}\alpha_k^2 + 2L_0\alpha_k h_k,$$

where c is some non-negative constant independent from the dimension. Moreover, if the stepsizes satisfy Assumption 2, we have

$$\lim_{k\to+\infty} f(x_k) = f(x^*) \quad \textit{a.s,}$$

and there exists a random variable $\hat{x}$ taking values in in $\arg\min f$ such that $x_k \to \hat{x}$ a.s.

Proof. Let $k \in \mathbb{N}$. By Algorithm 1,

$$\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 = \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle. \tag{6}$$

Since f_{h_k} is convex by Proposition 1 and $\mathbb{E}[g_k | \mathcal{F}_k] = \nabla f_{h_k}(x_k)$ (see Lemma 1), we have

$$-\langle \nabla f_{h_k}(x_k), x_k - x^* \rangle \leq f_{h_k}(x^*) - f_{h_k}(x_k).$$

Thus, taking the conditional expectation with respect to $\mathcal{F}_k$, by Lemma 4, we get,

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k] - \|x_k - x^*\|^2 \leq \underbrace{2c\frac{L_0^2 d}{\ell}\alpha_k^2}_{=:C_k} - 2\alpha_k(f_{h_k}(x_k) - f_{h_k}(x^*)).$$

Then, by Proposition 1,

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k] - \|x_k - x^*\|^2 \leq C_k - 2\alpha_k(f(x_k) - f(x^*)) + 2L_0\alpha_k h_k.$$

Next suppose that Assumption 2 holds. Rearranging the terms,

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k] - \|x_k - x^*\|^2 + 2\alpha_k(f(x_k) - f(x^*)) \leq C_k + 2L_0\alpha_k h_k,$$

with $C_k \in \ell^1$ and $\alpha_k h_k \in \ell^1$. Therefore, Robbins-Siegmund Theorem [43] implies that $(\|x_k - x^*\|)_{k\in\mathbb{N}}$ is a.s. convergent, $\alpha_k(f(x_k) - f(x^*)) \in \ell^1$ a.s. and thus, since $\alpha_k \notin \ell^1$,

$$\liminf_{k\to\infty} f(x_k) = f(x^*) \quad \text{a.s.} \tag{7}$$

We derive from [32, Lemma 9.9] and Lemma 5 that there exists a random variable $\hat{x}$ taking values in $\arg\min f$ such that $x_k \to \hat{x}$ a.s. Finally, continuity of f yields that $\lim_k f(x_k) = f(x_*)$ a.s. ☐

In the next Lemma, to derive bounds on function values, we study the sequence $(f_{h_k}(x_{k+1}) - f_{h_k}(x_k))_{k\in\mathbb{N}}$. It is the difference between the smoothed function at iteration k evaluated at x_k and at x_{k+1}. It corresponds to the function value decrease between the iterations $k+1$ and k if h_k is constant.

Lemma 7 (Function Value decrease: nonconvex non-smooth setting)**.** *Under Assumption 1, let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Then,*

$$\mathbb{E}[f_{h_k}(x_{k+1})|\mathcal{F}_k] - f_{h_k}(x_k) \leq -\alpha_k \|\nabla f_{h_k}(x_k)\|^2 + c\frac{L_0^3 d\sqrt{d}}{\ell}\frac{\alpha_k^2}{h_k},$$

where c is a numerical constant.

Proof. By Lemma 1, we have that f_{h_k} is $L_0\sqrt{d}/h_k$-smooth. Thus, by the Descent Lemma [41],

$$f_{h_k}(x_{k+1}) - f_{h_k}(x_k) \leq -\alpha_k \langle \nabla f_{h_k}(x_k), g_k\rangle + \frac{L_0\sqrt{d}}{2h_k}\alpha_k^2\|g_k\|^2.$$

Taking the conditional expectation with respect to $\mathcal{F}_k$,

$$\mathbb{E}[f_{h_k}(x_{k+1})|\mathcal{F}_k] - f_{h_k}(x_k) \leq -\alpha_k \|\nabla f_{h_k}(x_k)\|^2 + \frac{L_0\sqrt{d}}{2h_k}\alpha_k^2\, \mathbb{E}[\|g_k\|^2|\mathcal{F}_k]. \tag{8}$$

The claim follows from Lemma 4. ☐

A.3 Auxiliary results for smooth setting.

Lemma 8 (Function value decrease: convex smooth setting)**.** *Under Assumption 3 , let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Then the following holds:*

$$\mathbb{E}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) \leq -\alpha_k\Big(\frac{1}{2} - \frac{L_1 d}{\ell}\alpha_k\Big)\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2 \alpha_k h_k^2}{8} + \frac{L_1^3 d^2}{4\ell}\alpha_k^2 h_k^2.$$

Proof. By the Descent Lemma [41] and Algorithm 1,

$$f(x_{k+1}) - f(x_k) \leq -\alpha_k \langle \nabla f(x_k), g_k\rangle + \frac{L_1}{2}\alpha_k^2\|g_k\|^2.$$

Taking the conditional expectation and by Lemma 4,

$$\mathbb{E}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) \leq -\alpha_k \langle \nabla f(x_k), \nabla f_{h_k}(x_k)\rangle + \frac{L_1}{2}\alpha_k^2\Big[\frac{2d}{\ell}\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}h_k^2\Big].$$

Adding and subtracting $\nabla f(x_k)$,

$$\begin{aligned}\mathbb{E}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) \leq & -\alpha_k \langle \nabla f(x_k), \nabla f_{h_k}(x_k) - \nabla f(x_k)\rangle - \alpha_k\|\nabla f(x_k)\|^2 \\ & + \frac{L_1}{2}\alpha_k^2\Big[\frac{2d}{\ell}\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}h_k^2\Big].\end{aligned}$$

By Cauchy-Schwarz inequality and Proposition 1,

$$\begin{aligned}\mathbb{E}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) \leq & \,\alpha_k\Big(\frac{L_1 d}{2}h_k\Big)\|\nabla f(x_k)\| - \alpha_k\|\nabla f(x_k)\|^2 \\ & + \frac{L_1}{2}\alpha_k^2\Big[\frac{2d}{\ell}\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}h_k^2\Big].\end{aligned}$$

By Young's inequality,

$$\begin{aligned}\mathbb{E}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) \leq & \,\frac{L_1^2 d^2 \alpha_k h_k^2}{8} + \frac{\alpha_k}{2}\|\nabla f(x_k)\|^2 - \alpha_k\|\nabla f(x_k)\|^2 \\ & + \frac{L_1}{2}\alpha_k^2\Big[\frac{2d}{\ell}\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}h_k^2\Big]. \\ = & -\alpha_k\Big(\frac{1}{2} - \frac{L_1 d}{\ell}\alpha_k\Big)\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2 \alpha_k h_k^2}{8} + \frac{L_1^3 d^2}{4\ell}\alpha_k^2 h_k^2.\end{aligned}$$

This concludes the proof. ☐

Lemma 9 (Convergence in smooth setting)**.** *Let* $(x_k)_{k\in\mathbb{N}}$ *be the sequence generated by Algorithm 1 and let* $x^* \in \underset{x\in\mathbb{R}^d}{\arg\min} f(x)$*. Then, under Assumption 3, the following inequality holds*

$$\begin{aligned}\mathbb{E}[\|x_{k+1}-x^*\|^2|\mathcal{F}_k] - \|x_k - x^*\|^2 \leq \frac{2d}{\ell}\alpha_k^2\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2 \\ + L_1 d\alpha_k h_k\|x_k - x^*\| - 2\alpha_k\left\langle \nabla f(x_k), x_k - x^*\right\rangle .\end{aligned}$$

Moreover, if f *is convex, Assumption 4 holds and* $\alpha_k \leq \bar{\alpha} < \ell/(2dL_1)$*. Then*

- $(\alpha_k\|\nabla f(x_k)\|^2)_{k\in\mathbb{N}} \in \ell^1$ *a.s.*
- $(\|x_k - x^*\|)_{k\in\mathbb{N}}$ *is a.s. convergent.*
- $(\alpha_k(f(x_k) - f(x^*)))_{k\in\mathbb{N}} \in \ell^1$ *a.s.*
- *there exists a random variable* $\hat{x}$ *taking values in* $\arg\min f$ *such that* $x_k \to \hat{x}$ *a.s. and* $\lim_{k\to\infty} f(x_k) = \min f$*.*

Proof. We have

$$\|x_{k+1}-x^*\|^2 - \|x_k - x^*\|^2 = \alpha_k^2\|g_k\|^2 - 2\alpha_k\left\langle g_k, x_k - x^*\right\rangle .$$

Taking the conditional expectation,

$$\mathbb{E}[\|x_{k+1}-x^*\|^2|\mathcal{F}_k] - \|x_k - x^*\|^2 = \alpha_k^2\,\mathbb{E}[\|g_k\|^2|\mathcal{F}_k] - 2\alpha_k\left\langle \nabla f_{h_k}(x_k), x_k - x^*\right\rangle .$$

For every k, set $u_k = \|x_k - x^*\|$. By Lemma 4,

$$\mathbb{E}[u_{k+1}^2|\mathcal{F}_k] - u_k^2 \leq \frac{2d}{\ell}\alpha_k^2\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2 - 2\alpha_k\left\langle \nabla f_{h_k}(x_k), x_k - x^*\right\rangle .$$

Note that

$$-2\alpha_k\left\langle \nabla f_{h_k}(x_k), x_k - x^*\right\rangle = 2\alpha_k\left\langle \nabla f_{h_k}(x_k) - \nabla f(x_k), x^* - x_k\right\rangle - 2\alpha_k\left\langle \nabla f(x_k), x_k - x^*\right\rangle .$$

Thus,

$$\begin{aligned}\mathbb{E}[u_{k+1}^2|\mathcal{F}_k] - u_k^2 \leq \frac{2d}{\ell}\alpha_k^2\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2 \\ + 2\alpha_k\left\langle \nabla f_{h_k}(x_k) - \nabla f(x_k), x^* - x_k\right\rangle - 2\alpha_k\left\langle \nabla f(x_k), x_k - x^*\right\rangle .\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}[u_{k+1}^2|\mathcal{F}_k] - u_k^2 \leq \frac{2d}{\ell}\alpha_k^2\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2 \\ + 2\alpha_k\|\nabla f_{h_k}(x_k) - \nabla f(x_k)\|u_k - 2\alpha_k\left\langle \nabla f(x_k), x_k - x^*\right\rangle .\end{aligned}$$

The first claim follows from Proposition 1. By Proposition 1 and Young's inequality with parameter $\tau_k = \alpha_k h_k$, we get

$$\begin{aligned}\mathbb{E}[u_{k+1}^2|\mathcal{F}_k] - u_k^2 &\leq \frac{2d}{\ell}\alpha_k^2\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2 \\ &\quad + \frac{L_1 d}{2\tau_k}\alpha_k^2 h_k^2 + \frac{L_1 d\tau_k}{2}u_k^2 - 2\alpha_k\left\langle \nabla f(x_k), x_k - x^*\right\rangle . \\ &= \frac{2d}{\ell}\alpha_k^2\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2 \\ &\quad + \frac{L_1 d}{2}\alpha_k h_k + \frac{L_1 d}{2}\alpha_k h_k u_k^2 \\ &\quad - 2\alpha_k\left\langle \nabla f(x_k), x_k - x^*\right\rangle .\end{aligned} \tag{9}$$

Since f is convex, by Baillon-Haddad Theorem [3], we derive that

$$\begin{aligned}\mathbb{E}[u_{k+1}^2|\mathcal{F}_k] - u_k^2 \leq -2\Big(\frac{1}{L_1} - \frac{d}{\ell}\alpha_k\Big)\alpha_k\|\nabla f(x_k)\|^2 + \frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2 \\ + \frac{L_1 d}{2}\alpha_k h_k + \frac{L_1 d}{2}\alpha_k h_k u_k^2.\end{aligned}$$

By Assumption 4,

$$\mathbb{E}[u_{k+1}^2|\mathcal{F}_k] - u_k^2 \leq -2\underbrace{\left(\frac{1}{L_1} - \frac{d}{\ell}\bar{\alpha}\right)}_{=:\Delta}\alpha_k\|\nabla f(x_k)\|^2 + \underbrace{\frac{L_1 d}{2}\alpha_k h_k}_{=:\rho_k} u_k^2$$
$$+ \underbrace{\frac{L_1 d}{2}\alpha_k h_k + \frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2}_{=:C_k}.$$

Note that $\Delta > 0$. Thus, rearranging the terms

$$\mathbb{E}[u_{k+1}^2|\mathcal{F}_k] - (1+\rho_k)u_k^2 + 2\Delta\alpha_k\|\nabla f(x_k)\|^2 \leq C_k.$$

Since $\rho_k, C_k \in \ell^1$ by Assumption 4, Robbins-Siegmund Theorem [43] ensures that $(u_k^2)_{k\in\mathbb{N}}$ is convergent and $(\alpha_k\|\nabla f(x_k)\|^2)_{k\in\mathbb{N}} \in \ell^1$ a.s. Since f is convex, it follows from (9) that

$$\mathbb{E}[u_{k+1}^2|\mathcal{F}_k] - (1+\rho_k)u_k^2 \leq \frac{2d}{\ell}\alpha_k^2\|\nabla f(x_k)\|^2 - 2\alpha_k(f(x_k) - f(x^*)) + C_k.$$

Robbins-Siegmund Theorem [43] implies that $(\alpha_k(f(x_k) - f(x^*)))_{k\in\mathbb{N}} \in \ell^1$ a.s. Assumption 4 implies that $\alpha_k \notin \ell^1$ therefore

$$\liminf_k f(x_k) - f(x^*) = 0 \text{ a.s.} \tag{10}$$

By Lemma 8 and Assumption 4, we have that the sequence $\mathbb{E}[f(x_{k+1}) - f(x^*)|\mathcal{F}_k] - (f(x_k) - f(x^*))$ is upper-bounded by a sequence in ℓ^1. Thus, by Robbins-Siegmund Theorem [43], $\lim_k(f(x_k) - f(x^*))$ exists a.s. Then, it follows from (10) that

$$\lim_{k\to\infty} f(x_k) = f(x^*) \quad a.s.$$

Moreover, as we saw before, $(\|x_k - x^*\|)_{k\in\mathbb{N}}$ is convergent a.s. for every $x^* \in \arg\min f$. Then, by Opial's Lemma [40], there exists a random variable $\hat{x}$ taking values in $\arg\min f$ such that $x_k \to \hat{x}$ a.s. □

Lemma 10 (Gradient bound: convex smooth setting)**.** *Suppose that Assumptions 3 and 4 hold, and assume f to be convex. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Then, for every $k \in \mathbb{N}$ and every $x^* \in \arg\min f$,*

$$\sum_{i=0}^{k} \alpha_i\, \mathbb{E}[\|\nabla f(x_i)\|^2] \leq \frac{1}{2\Delta}\Big(S_k + \sum_{i=0}^{k}\rho_i\sqrt{\mathbb{E}[\|x_i - x^*\|^2]}\Big),$$

and

$$\sqrt{\mathbb{E}[\|x_k - x^*\|^2]} \leq \sqrt{S_{k-1}} + \sum_{i=0}^{k}\rho_i,$$

where

$$\Delta := \left(\frac{1}{L_1} - \frac{d}{\ell}\bar{\alpha}\right), \quad S_k := \|x_0 - x^*\| + \sum_{i=0}^{k} C_i$$
$$C_k := \frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2 \quad \text{and} \quad \rho_k := L_1 d\alpha_k h_k.$$

Proof. By Lemma 9 we derive

$$\mathbb{E}[\|x_{k+1} - x^*\|^2|\mathcal{F}_k] - \|x_k - x^*\|^2 \leq \frac{2d}{\ell}\alpha_k^2\|\nabla f(x_k)\|^2 + C_k$$
$$+ \rho_k\|x_k - x^*\| - 2\alpha_k\langle\nabla f(x_k), x_k - x^*\rangle.$$

By Baillon-Haddad Theorem and Assumption 4,

$$\mathbb{E}[\|x_{k+1} - x^*\|^2|\mathcal{F}_k] - \|x_k - x^*\|^2 \leq -2\Delta\alpha_k\|\nabla f(x_k)\|^2 + C_k + \rho_k\|x_k - x^*\|.$$

Let $u_k := \sqrt{\mathbb{E}[\|x_k - x^*\|^2]}$. Taking the full expectation, by Jensen inequality we have

$$u_{k+1}^2 - u_k^2 \leq -2\Delta\alpha_k\,\mathbb{E}[\|\nabla f(x_k)\|^2] + \rho_k u_k + C_k.$$

Summing the previous inequality from $i = 0, \cdots, k$, we get

$$u_{k+1}^2 + 2\Delta\sum_{i=0}^{k}\alpha_i\,\mathbb{E}[\|\nabla f(x_i)\|^2] \leq \underbrace{u_0^2 + \sum_{i=0}^{k} C_i}_{=:S_k} + \sum_{i=0}^{k}\rho_i u_i. \tag{11}$$

Since u_k is non-negative, the first claim of the lemma follows. Since $\Delta > 0$, $\rho_k \geq 0$, S_k is non decreasing, and $S_k \geq u_0^2$ in (11), then

$$u_{k+1}^2 \leq S_k + \sum_{i=0}^{k}\rho_i u_i.$$

Thus, the (discrete) Bihari's Lemma [32, Lemma 9.8] yields

$$u_{k+1} \leq \frac{1}{2}\sum_{i=0}^{k}\rho_i + \Big[S_k + \Big(\frac{1}{2}\sum_{i=0}^{k}\rho_i\Big)^2\Big]^{1/2} \leq \sqrt{S_k} + \sum_{i=0}^{k}\rho_i,$$

concluding the proof. □

B Proofs of Main Results

B.1 Proof of Theorem 1

By Lemma 6,

$$\mathbb{E}[\|x_{k+1} - x^*\|^2|\mathcal{F}_k] - \|x_k - x^*\|^2 + 2\alpha_k(f(x_k) - f(x^*)) \leq 2c\frac{L_0^2 d}{\ell}\alpha_k^2 + 2L_0\alpha_k h_k.$$

Rearranging the terms, taking the full expectation, and summing the first k iterations

$$\sum_{i=0}^{k}\alpha_i\,\mathbb{E}[(f(x_i) - f(x^*))] \leq \frac{\|x_0 - x^*\|^2}{2} + c\frac{dL_0^2}{\ell}\sum_{i=0}^{k}\alpha_i^2 + L_0\sum_{i=0}^{k}\alpha_i h_i.$$

Let $\bar{x}_k := \sum\limits_{i=0}^{k}\alpha_i x_i/(\sum\limits_{i=0}^{k}\alpha_i)$. Dividing by $\sum\limits_{i=0}^{k}\alpha_i$ and observing that by convexity we have

$$\mathbb{E}[f(\bar{x}_k) - \min f] \leq \frac{\sum\limits_{i=0}^{k}\alpha_i\,\mathbb{E}[(f(x_i) - f(x^*))]}{\sum\limits_{i=0}^{k}\alpha_i},$$

we get the first claim. Under Assumption 2, the second claim holds by Lemma 6.

B.2 Proof of Corollary 1

By Theorem 1,

$$\mathbb{E}[f(\bar{x}_k) - \min f] \leq \frac{1}{\sum\limits_{i=0}^{k}\alpha_i}\left(\frac{\|x_0 - x^*\|^2}{2} + c\frac{dL_0^2}{\ell}\sum_{i=0}^{k}\alpha_i^2 + L_0\sum_{i=0}^{k}\alpha_i h_i\right).$$

Replacing α_k and h_k with the sequences in the statement,

$$\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leq \frac{C_1}{\alpha k^{1-\theta}} + \frac{C_2}{k^{\rho}}h + \frac{d}{\ell}\frac{C_3}{k^{\theta}}\alpha,$$

with

$$C_1 := \frac{(1-\theta)\|x_0 - x^*\|^2}{2}, \qquad C_2 := \frac{L_0(1-\theta)}{(1-\theta-\rho)} \quad \text{and} \quad C_3 := \frac{cL_0^2(1-\theta)}{(1-2\theta)}.$$

The second point of the corollary can be proved replacing $\alpha_k = \alpha$ and $h_k = h$. Now, to prove the third point, fix $\varepsilon \in (0,1)$. Since we want $\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leq \varepsilon$, we impose

$$\frac{\|x_0 - x^*\|^2}{2\alpha k} + \frac{cdL_0^2}{\ell}\alpha + L_0 h \leq \varepsilon.$$

Choosing $h_k = h \leq \frac{\varepsilon}{2L_0}$, to get the previous inequality it is sufficient to impose

$$\frac{\|x_0 - x^*\|^2}{2\alpha k} + \frac{cdL_0^2}{\ell}\alpha \leq \frac{\varepsilon}{2}.$$

We fix a priori a number of iterations K and we minimize the left handside with respect to α, obtaining

$$\alpha = \sqrt{\frac{\ell}{d}}\frac{\|x_0 - x^*\|}{\sqrt{2cK}L_0}.$$

Thus, for $h_k = h \leq \frac{\varepsilon}{2L_0}$, α as above and

$$K \geq \frac{8\|x_0 - x^*\|^2 L_0^2 cd}{\ell \varepsilon^2},$$

we have $\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leq \varepsilon$. Note that, since the computation of the surrogate requires 2ℓ function evaluations, to ensure an error of ε we need to perform a number of function evaluations of the order

$$\mathcal{O}(d\varepsilon^{-2}).$$

This concludes the proof.

B.3 Proof of Theorem 2

By Lemma 7,

$$\mathbb{E}[f_h(x_{k+1})|\mathcal{F}_k] - f_h(x_k) \leq -\alpha_k\|\nabla f_h(x_k)\|^2 + c\frac{L_0^3 d\sqrt{d}}{\ell}\frac{\alpha_k^2}{h}.$$

Taking the full expectation and rearranging the terms,

$$\alpha_k\,\mathbb{E}[\|\nabla f_h(x_k)\|^2] \leq \mathbb{E}[f_h(x_k) - f_h(x_{k+1})] + c\frac{L_0^3 d\sqrt{d}}{\ell}\frac{\alpha_k^2}{h}.$$

Next sum from $i = 0$ to $i = k$. By definition of f_h, we have $f_h(x) \geq \min f$ for every $x \in \mathbb{R}^d$, thus,

$$\sum_{i=0}^{k}\alpha_i\,\mathbb{E}[\|\nabla f_h(x_i)\|^2] \leq \mathbb{E}[f_h(x_0) - \min f] + c\frac{L_0^3 d\sqrt{d}}{\ell}\sum_{i=0}^{k}\frac{\alpha_i^2}{h}. \tag{12}$$

The claim follows.

B.4 Proof of Corollaries 2 and 3

By Theorem 2,

$$\eta_k^{(h)} \leq \Big((f_h(x_0) - f(x^*)) + c\frac{L_0^3 d\sqrt{d}}{\ell}\sum_{i=0}^{k}\frac{\alpha_i^2}{h}\Big)/\Big(\sum_{i=0}^{k}\alpha_i\Big).$$

Due to the choice of $\alpha_k = \alpha(k+1)^{-\theta}$ with $\theta \in (1/2, 1)$ and $\alpha > 0$, we get

$$\eta_k^{(h)} \leq \frac{C_1}{\alpha(k+1)^{1-\theta}} + \frac{C_2 d\sqrt{d}\alpha}{\ell h}\frac{1}{(k+1)^{\theta}},$$

where

$$C_1 := \|x_0 - x^*\|^2(1-\theta) \quad \text{and} \quad C_2 := \frac{cL_0^3(1-\theta)}{(1-2\theta)}.$$

If we choose $\alpha_k = \alpha$, we derive

$$\eta_k^{(h)} \leq \frac{f_h(x_0) - \min f}{\alpha k} + \frac{cL_0^3 d\sqrt{d}\alpha}{\ell h}. \tag{13}$$

If we fix a priori a number of iteration K and we minimize the right handside with respect to α, we get

$$\hat{\alpha} = \sqrt{\frac{(f_h(x_0) - f(x^*))\ell h}{KcL_0^3 d\sqrt{d}}}.$$

Let $\varepsilon \in (0,1)$. Choosing $\alpha = \hat{\alpha}$, we get $\eta_K^{(h)} \leq \varepsilon$ for

$$K \geq 4\frac{(f_h(x_0) - f(x^*))cL_0^3 d\sqrt{d}}{\ell h}\varepsilon^{-2}. \tag{14}$$

This concludes the proof of Corollary 2. To prove Corollary 3, we fix a maximum number of iterations $K \in \mathbb{N}$ and consider the random variable I of the statement. Let $\partial_h f$ be the h-Goldstein subdifferential defined in Definition 1. It follows from [34, Theorem 3.1] that $\nabla f_h(x_I) \in \partial_h f(x_I)$ almost surely, therefore

$$\mathbb{E}_I \min[\|\eta\|^2 \,:\, \eta \in \partial_h f(x_I)] \leq \mathbb{E}_I\, \mathbb{E}[\|\nabla f_h(x_I)\|^2].$$

In addition, Theorem 2 yields

$$\begin{aligned}
\mathbb{E}_I\, \mathbb{E}_G[\|\nabla f_h(x_I)\|^2] &= \Big(\sum_{j=0}^{K-1} \alpha_j\, \mathbb{E}_G[\|\nabla f_h(x_j)\|^2]\Big) / \sum_{j=0}^{K-1} \alpha_j \\
&\leq \mathbb{E}[f_h(x_0) - \min f] + c\frac{L_0^3 d\sqrt{d}}{\ell}\sum_{i=0}^{k} \frac{\alpha_i^2}{h}.
\end{aligned}$$

Thus,

$$\mathbb{E}_I[\|\eta\|^2 \,:\, \eta \in \partial_h f(x_I)] \leq \mathbb{E}_I\, \mathbb{E}[\|\nabla f_h(x_I)\|^2] = \eta_k^{(h)}.$$

Hence, for $\alpha = \bar{\alpha}$ and K chosen s.t. inequality (14) holds, we have

$$\mathbb{E}_I[\|\eta\|^2 \,:\, \eta \in \partial_h f(x_I)] \leq \varepsilon.$$

This concludes the proof.

B.5 Proof of Theorem 3

By Lemma 9,

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k] - \|x_k - x^*\|^2 \leq{}& \frac{2d}{\ell}\alpha_k^2 \|\nabla f(x_k)\|^2 + 2\alpha_k \langle \nabla f(x_k), x^* - x_k\rangle \\
&+ \underbrace{L_1 d\alpha_k h_k}_{=:\rho_k} \|x^* - x_k\| + \underbrace{\frac{L_1^2 d^2}{2\ell}\alpha_k^2 h_k^2}_{=:C_k}.
\end{aligned}$$

By convexity,

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k] - \|x_k - x^*\|^2 \leq{}& \frac{2d}{\ell}\alpha_k^2 \|\nabla f(x_k)\|^2 - 2\alpha_k(f(x_k) - f(x^*)) \\
&+ \rho_k \|x^* - x_k\| + C_k.
\end{aligned}$$

Rearranging the terms and taking the full expectation,

$$\begin{aligned}
2\,\mathbb{E}[\alpha_k(f(x_k) - f(x^*))] \leq{}& \mathbb{E}[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2] + \frac{2d}{\ell}\alpha_k^2\, \mathbb{E}[\|\nabla f(x_k)\|^2] \\
&+ \rho_k\, \mathbb{E}[\|x^* - x_k\|] + C_k.
\end{aligned}$$

Since $\mathbb{E}[\|x^* - x_k\|] = \mathbb{E}[\sqrt{\|x^* - x_k\|^2}]$, Jensen's inequality implies that

$$2\,\mathbb{E}[\alpha_k(f(x_k) - f(x^*))] \leq \mathbb{E}[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2] + \frac{2d}{\ell}\alpha_k^2\,\mathbb{E}[\|\nabla f(x_k)\|^2] + \rho_k\sqrt{\mathbb{E}[\|x^* - x_k\|^2]} + C_k.$$

Denoting with $u_k = \mathbb{E}[\|x_k - x^*\|^2]$ and taking the sum from $i = 0$ to $i = k$,

$$\begin{aligned} 2\sum_{i=0}^{k}\alpha_i\,\mathbb{E}[f(x_i) - f(x^*)] &\leq \underbrace{u_0^2 + \sum_{i=0}^{k}C_i}_{=:S_k} + \frac{2d}{\ell}\sum_{i=0}^{k}\alpha_i^2\,\mathbb{E}[\|\nabla f(x_i)\|^2] + \sum_{i=0}^{k}\rho_i u_i \\ &\leq S_k + \frac{2d}{\ell}\bar{\alpha}\sum_{i=0}^{k}\alpha_i\,\mathbb{E}[\|\nabla f(x_i)\|^2] + \sum_{i=0}^{k}\rho_i u_i, \end{aligned}$$

where the last inequality holds by Assumption 4. Let $\Delta := (1/L_1 - (d/\ell)\bar{\alpha})$. By Lemma 10, we have

$$\begin{aligned} \sum_{i=0}^{k}\alpha_i\,\mathbb{E}[f(x_i) - f(x^*)] &\leq \frac{1}{2}\Big(S_k + \frac{d\bar{\alpha}}{\Delta\ell}\Big[S_k + \sum_{i=0}^{k}\rho_i u_i\Big] + \sum_{i=0}^{k}\rho_i u_i\Big) \\ &= \frac{\ell\Delta + d\bar{\alpha}}{2\ell\Delta}\Big(S_k + \sum_{i=0}^{k}\rho_i u_i\Big) \\ &\leq \frac{\ell\Delta + d\bar{\alpha}}{2\ell\Delta}\Big(S_k + \sum_{i=0}^{k}\rho_i(\sqrt{S_i} + \sum_{j=0}^{i}\rho_j)\Big). \end{aligned}$$

Let $\bar{x}_k := \sum_{i=0}^{k}\alpha_i x_i / (\sum_{i=0}^{k}\alpha_i)$. Dividing both sides by $\sum_{i=0}^{k}\alpha_i$, convexity yields

$$\mathbb{E}[f(\bar{x}_k) - \min f] \leq \frac{\sum_{i=0}^{k}\alpha_i\,\mathbb{E}[(f(x_i) - f(x^*))]}{\sum_{i=0}^{k}\alpha_i}.$$

B.6 Proof of Corollary 4

In this proof, we use the same notation as the one in the proof of Theorem 3. By the choices of the parameters, we have

$$\sum_{i=0}^{k}\rho_i \leq C_1 d\alpha h \quad \text{with} \quad C_1 := \frac{L_1\theta}{\theta - 1},$$

$$S_k \leq \|x_0 - x^*\|^2 + C_2\frac{d^2}{\ell}\alpha^2 h^2 \quad \text{with} \quad C_2 := \frac{L_1^2\theta}{2\theta - 1}.$$

Thus, using these inequalities in Theorem 3, we get

$$D_k \leq \frac{\ell\Delta + d\bar{\alpha}}{2\ell\Delta}\Big(\|x_0 - x^*\|^2 + C_2\frac{d^2\alpha^2h^2}{\ell} + \sqrt{\|x_0 - x^*\|^2}C_3 d\alpha h + C_4\frac{d\alpha h}{\sqrt{\ell}} + C_5 d^2\alpha^2h^2\Big),$$

with

$$C_3 := \frac{L_1^2\theta}{\theta - 1}, \quad C_4 := \frac{L_1^2}{\sqrt{2}}, \quad C_5 := \frac{L_1^2\theta}{(\theta - 1)^2}.$$

Dividing by $\sum_{i=0}^{k}\alpha_i$, we get

$$\mathbb{E}[f(\bar{x}_k) - \min f] \leq \frac{C}{\alpha k}.$$

Note that by Assumption 4, $\alpha < \ell/(dL_1)$, thus $1/\alpha > (dL_1)/\ell$. The algorithm performs 2ℓ function evaluations at each iteration. Thus, to guarantee $\mathbb{E}[f(\bar{x}_k) - \min f] \leq \varepsilon$ for $\varepsilon \in (0,1)$, the algorithm has to perform a number of function evaluations in the order of

$$\mathcal{O}(d\varepsilon^{-1}).$$

Assuming, instead, $\alpha_k \leq \bar{\alpha} < \ell/(2dL_1)$, by Lemma 9 we get the last claim; i.e, there exists a random variable $\hat{x}$ taking values in $\arg\min f$ s.t. $x_k \to \hat{x}$ a.s.

B.7 Proof of Theorem 4

Set $C_1 = (dL_1)/2$. It follows from Lemma 8 that

$$\mathbb{E}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) \leq -\Big(\frac{1}{2} - \frac{L_1 d}{\ell}\bar{\alpha}\Big)\alpha_k\|\nabla f(x_k)\|^2 + \frac{C_1^2\alpha_k h_k^2}{2} + \frac{L_1^3 d^2}{4\ell}\alpha_k^2 h_k^2.$$

Taking the full expectation and rearranging the terms, and recalling the definition of Δ,

$$\Delta\alpha_k\,\mathbb{E}[\|\nabla f(x_k)\|^2] \leq \mathbb{E}[f(x_k) - f(x_{k+1})] + \frac{C_1^2\alpha_k h_k^2}{2} + \frac{L_1^3 d^2}{4\ell}\alpha_k^2 h_k^2.$$

Summing for $i = 0, \cdots, k$ and observing that $\min f \leq f(x)$ for every x,

$$\Delta\sum_{i=0}^{k}\alpha_i\,\mathbb{E}[\|\nabla f(x_i)\|^2] \leq f(x_0) - \min f + \sum_{i=0}^{k}\frac{C_1^2\alpha_i h_i^2}{2} + \frac{L_1^3 d^2}{4\ell}\sum_{i=0}^{k}\alpha_i^2 h_i^2.$$

Divinding by $\Delta\sum\limits_{i=0}^{k}\alpha_i$ we get the claim.

B.8 Proof of Corollary 5

(i): From the choice of α_k and h_k, we have

$$\sum_{i=0}^{k}\alpha_i h_i^2 \leq \frac{2\theta\alpha h^2}{2\theta - 1} \qquad \sum_{i=0}^{k}\alpha_i^2 h_i^2 \leq \frac{2\theta\alpha^2 h^2}{2\theta - 1}.$$

It follows from Theorem 4 that

$$\eta_k \leq \frac{1}{\Delta\alpha k}\left(f(x_0) - \min f + C_1 d^2\alpha h^2 + \frac{C_2\alpha^2 h^2 d^2}{\ell}\right),$$

with $C_1 = \frac{L_1^2\theta}{4(2\theta-1)}$ and $C_2 = \frac{L_1^3\theta}{2(2\theta-1)}$.

(ii): It follows directly from Theorem 4 taking into account that

$$\sum_{i=0}^{k}\alpha_i h_i^2 = k\alpha h^2, \qquad \sum_{i=0}^{k}\alpha_i^2 h_i^2 = k\alpha^2 h^2,$$

and setting $C_1 = L_1^2/8$ and $C_2 = L_1^3/4$.

C Experimental Details

In this appendix, we report details on the experiments performed. We implemented every script in Python3 (version 3.9.11) and used numpy (version 1.22.2) [27] and matplotlib (version 3.5.1) [29] libraries.

Machine used to perform the experiments. In the following table, we describe the features of the machine used to perform the experiments in Section 4.

Table 1: Machine used to perform the experiments

Feature	
OS	Debian GNU/Linux 11
CPU(s)	4 x Intel(R) Core(TM) i7-1165G7 11th Gen @ 2.80GHz
CPU Core(s)	4
RAM	8 GB

Target Functions. We considered two synthetic target functions: a convex smooth function f_1 and a convex non-smooth function f_2 defined as follows

$$\text{(Convex Smooth)} \quad f_1(x) := \frac{1}{2}\|Ax\|^2 \quad \text{with} \quad A \in \mathbb{R}^{d \times d}$$

$$\text{(Convex Non-smooth)} \quad f_2(x) := \|x - \bar{v}\|_1$$

where A is a random Gaussian matrix (i.e. $A_{i,j} \sim \mathcal{N}(0,1)$) and $\bar{v} := [0, 1, \cdots, d-1]^\intercal$.

Choice of the number of directions. We report here the details of the first experiment of Section 4. For these experiments, we consider $d = 50$ and we use, for the smooth convex case, the following parameters

$$\alpha_k = 0.99\frac{\ell}{dL_1} \qquad \text{and} \qquad h_k = \frac{10^{-5}}{k+1}.$$

The constant L_1 is computed as the maximum eigenvalue of the matrix $A^\intercal A$. Note that this parameter choice satisfies Assumption 4. For the non-smooth target, we used

$$\alpha_k = \sqrt{\frac{\ell}{d}}k^{-1/2-10^{-5}} \qquad \text{and} \qquad h_k = \frac{10^{-7}}{k+1}.$$

Note that this parameter configuration satisfies Assumption 2. The maximum number of function evaluations considered is 4000. The direction matrices G_k are generated with the QR method - see Appendix D.

Comparison with Finite-difference methods. In Section 4, we compare finite-difference method with different choice of directions. In order to make a fair comparison we consider only central finite-differences. However, note that Algorithm 1 can be modified (in practice) considering computationally cheaper gradient estimators - see Remark 1. For these experiments, we consider $d = 10$ and $\ell = d$ for methods with multiple directions. The maximum number of function evaluations is 1000 for both smooth and non-smooth targets and the direction matrices G_k for Algorithm 1 are generated with the QR method - see Appendix D. To solve the smooth problem we consider the following parameter choice for every method

$$\alpha_k = c\frac{\ell}{dL_1} \qquad \text{and} \qquad h_k = \frac{10^{-7}}{d^2(k+1)},$$

where L_1 is computed taking the maximum eigenvalue of $A^\intercal A$. For Algorithm 1 and finite-difference with single and multiple spherical directions $c = 0.99$ while it is equal to $c = 0.11$ for finite-difference with single and multiple Gaussian directions. We made this choice since for finite-difference methods with Gaussian directions we observed divergence for larger choices of c - see Figure 3.

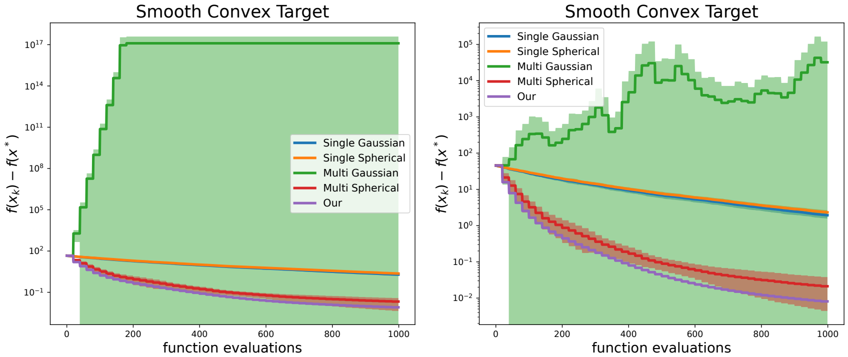

Figure 3: From left to right, comparison of finite-difference methods for smooth convex target with $c = 0.99$ and $c = 0.2$ for methods with Gaussian directions.

For the non-smooth convex target, we considered the following parameter choice

$$\alpha_k = c\frac{\ell}{d}k^{-1/2-10^{-5}} \qquad \text{and} \qquad h_k = \frac{1}{d^2(k+1)}.$$

For every method, we selected $c = 0.65$ except for the method with multiple Gaussian directions in which we selected $c = 0.08$ since it provided better performances - see Figure 4.

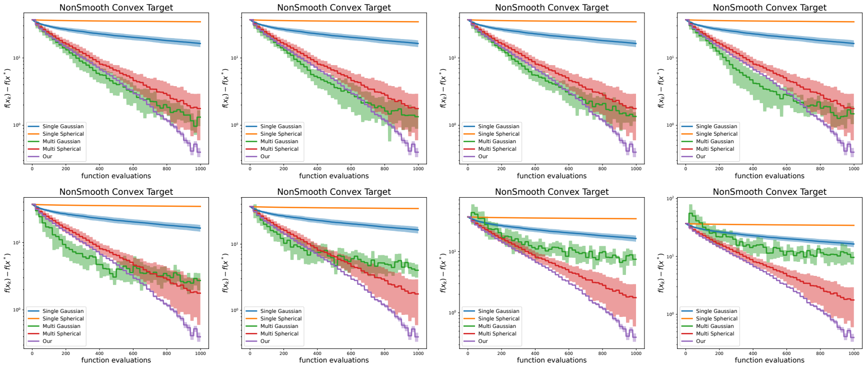

Figure 4: From left to right and up to down, comparison of finite difference method with different directions and different values of c for multiple Gaussian directions. The values of c considered are the following $[0.085, 0.089, 0.09, 0.1, 0.2, 0.3, 0.5, 0.65]$

D Techniques to Generate Orthogonal Direction Matrices

In the literature, different algorithms were proposed to generate orthogonal matrices - see for instance [23, 38, 11, 28, 7, 2, 4, 44, 8] and references therein. Such methods can be used to generate the direction matrices G_k required for the iteration proposed in Algorithm (1). In this appendix, we briefly discuss three of them.

QR factorization. As observed in [32, 42], a way to generate orthogonal consists in generating a random Gaussian matrix $A \in \mathbb{R}^{d \times d}$ with $A_{i,j} \sim \mathcal{N}(0, 1)$ and perform the QR factorization i.e. $A = QR$. Then, the direction matrix is the truncation of the Q matrix i.e. $QI_{d,\ell}$.

Householder Reflection. To obtain a direction matrix, we can use a Householder reflector. This can be done by sampling a vector v from the unit sphere $\mathbb{S}^{d-1}$. The direction matrix G is defined as a Householder reflector, given by

$$G := I - 2vv^{\intercal},$$

with $I \in \mathbb{R}^{d \times d}$ identity matrix. To obtain the desired matrix, we compute the product of G with $I_{d,\ell}$, i.e., we take the first ℓ columns. The (truncated) identity matrix can be generated and stored offline (note that since it is very sparse, it can be stored using a sparse format (e.g. the COO format proposed in scikit-learn library[9]). In this way, we can save resources in high-dimensional settings. In order to quantify the time-cost of this procedure, we compared the time of generating this kind of matrix with random matrices with different dimensions. For this experiment, we consider the $\ell = d$ case i.e. the most expensive. Matrices are computed in CPU and the details of the machine used are described in Appendix C. We report the mean and standard deviation of the time using 500 repetitions. In Figure 5, we compare the time-cost of generating orthogonal matrices with this procedure against generating random matrices while in Table 2 we report the mean and standard deviation of the results.

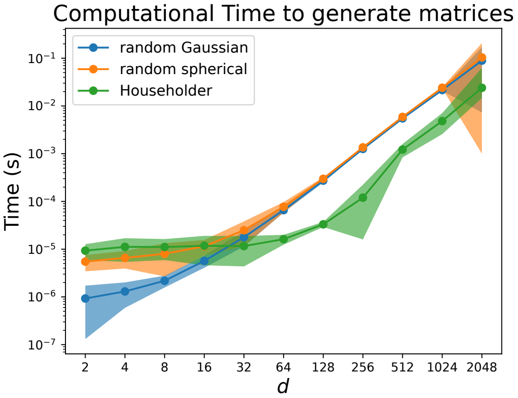

Figure 5: Time comparison in CPU of different methods to generate direction matrices.

In Figure 5, we can observe that using this strategy we can limit the cost of generating random orthogonal matrices. In particular, for dimensions larger than 32, our method is faster than random gaussian and spherical directions.

Table 2: Comparison of the time-cost (seconds) of generating random and orthogonal matrices with different dimensions

d	Random Gaussian	Random Spherical	Householder
2	$9.27 \times 10^{-7} \pm 7.96 \times 10^{-7}$	$5.49 \times 10^{-6} \pm 2.05 \times 10^{-6}$	$9.32 \times 10^{-6} \pm 3.34 \times 10^{-6}$
4	$1.30 \times 10^{-6} \pm 7.21 \times 10^{-7}$	$6.56 \times 10^{-6} \pm 2.63 \times 10^{-6}$	$1.12 \times 10^{-5} \pm 5.79 \times 10^{-6}$
8	$2.18 \times 10^{-6} \pm 6.06 \times 10^{-7}$	$8.01 \times 10^{-6} \pm 5.32 \times 10^{-6}$	$1.11 \times 10^{-5} \pm 5.20 \times 10^{-6}$
8	$2.18 \times 10^{-6} \pm 6.06 \times 10^{-7}$	$8.01 \times 10^{-6} \pm 5.32 \times 10^{-6}$	$1.11 \times 10^{-5} \pm 5.20 \times 10^{-6}$
16	$5.69 \times 10^{-6} \pm 1.61 \times 10^{-6}$	$1.15 \times 10^{-5} \pm 4.10 \times 10^{-6}$	$1.18 \times 10^{-5} \pm 7.20 \times 10^{-6}$
32	$1.78 \times 10^{-5} \pm 6.42 \times 10^{-6}$	$2.49 \times 10^{-5} \pm 1.33 \times 10^{-5}$	$1.16 \times 10^{-5} \pm 7.25 \times 10^{-6}$
64	$6.58 \times 10^{-5} \pm 7.03 \times 10^{-6}$	$7.74 \times 10^{-5} \pm 1.95 \times 10^{-5}$	$1.62 \times 10^{-5} \pm 3.79 \times 10^{-6}$
128	$2.73 \times 10^{-4} \pm 2.37 \times 10^{-5}$	$2.98 \times 10^{-4} \pm 2.45 \times 10^{-5}$	$3.32 \times 10^{-5} \pm 4.02 \times 10^{-6}$
256	$1.26 \times 10^{-3} \pm 2.79 \times 10^{-5}$	$1.36 \times 10^{-3} \pm 2.90 \times 10^{-5}$	$1.20 \times 10^{-4} \pm 1.04 \times 10^{-4}$
512	$5.50 \times 10^{-3} \pm 1.63 \times 10^{-4}$	$5.91 \times 10^{-3} \pm 1.22 \times 10^{-4}$	$1.22 \times 10^{-3} \pm 3.82 \times 10^{-4}$
1024	$2.16 \times 10^{-2} \pm 6.92 \times 10^{-4}$	$2.41 \times 10^{-2} \pm 7.35 \times 10^{-4}$	$4.83 \times 10^{-3} \pm 2.26 \times 10^{-3}$
2048	$8.92 \times 10^{-2} \pm 8.19 \times 10^{-2}$	$1.04 \times 10^{-1} \pm 1.03 \times 10^{-1}$	$2.40 \times 10^{-2} \pm 3.87 \times 10^{-2}$

Moreover, if more computational resources are available, we can build $m > 1$ Householder reflectors $G_1, \cdots, G_m$ using m random vectors $v_1, \cdots, v_m$ sampled i.i.d from $\mathbb{S}^{d-1}$ and define the direction matrix as

$$G_1 G_2 \cdots G_m I_{d,\ell}.$$

It is important to note that when $m = d$, this procedure is equivalent to using the QR factorization.

Haar Butterfly matrices. We can build orthogonal matrices using Butterfly matrices [48]. Let $G^{(0)} := [1]$, we can build an orthogonal matrix of dimension $d = 2^n$ with the following recursion

$$G^{(n)} = \begin{bmatrix} \cos(\theta_n) G^{(n-1)} & \sin(\theta_n) G^{(n-1)} \\ -\sin(\theta_n) G^{(n-1)} & \cos(\theta_n) G^{(n-1)} \end{bmatrix}$$

where θ_n is sampled uniformly in $[0, 2\pi]$. Then we compute $G I_{d,\ell}$ (we take the first ℓ columns). The construction of Haar butterfly matrices is faster than previous methods because it only requires simple operations. However, this procedure allows to build only matrices with $d = 2^n$ for $n \geq 0$. In literature, different methods were proposed to cope with this limitation e.g. [23].

E Limitations

In this appendix, we discuss the main practical limitations of Algorithm 1. Like all finite-difference methods with multiple directions, O-ZD requires multiple function evaluations to execute a single step. In many practical applications, function evaluations can be time-consuming, leading to the use of a small number of directions ℓ. This may result in poor performance as observed in numerical experiments. As for the subgradient method, in O-ZD the step size significantly affects performance, and tuning it can be challenging. To address this limitation, an adaptive stepsize selection method could be proposed. Furthermore, decreasing the sequence h_k too quickly can lead to numerical instability, as noted in [42].

F Other Experiments

We performed other experiments in minimizing convex functions. We considered the targets defined in Table 3 and, for each experiment, we reported the mean and standard deviation using 20 repetitions.

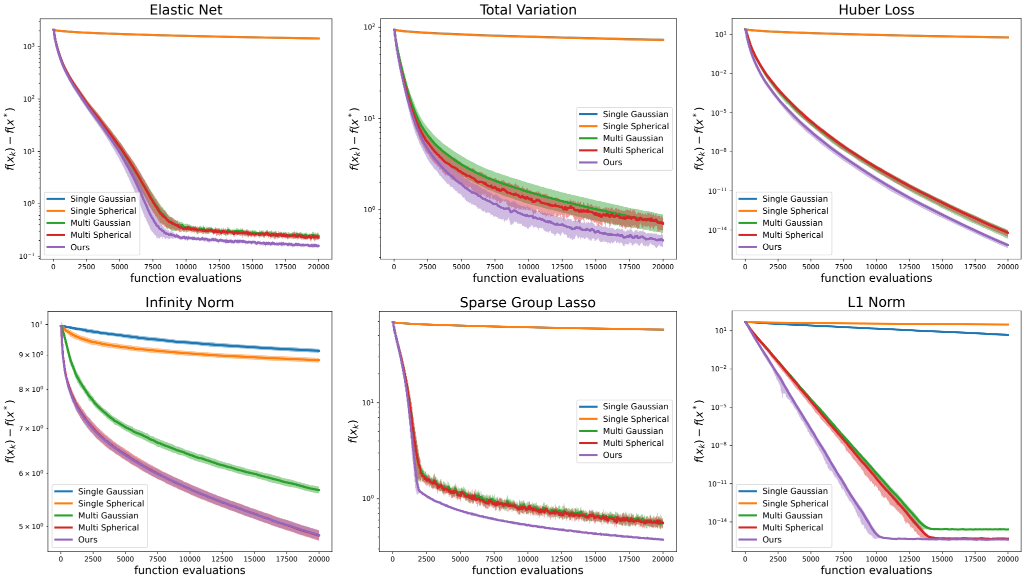

Figure 6: Function values per function evaluation in optimizing functions with different algorithms.

In Figure 6, we can observe that structured finite-difference performs better than unstructured methods.

Table 3: Functions used and relative dimension and number of directions considered.

Name	Definition		d	ℓ
Sparse Group Lasso	$f(x) := \sum_{i=1}^{p} \lVert x^{(\beta_i)} \rVert$		50	25
Huber Loss	$f(x) := \begin{cases} 0.5\lVert x\rVert_2^2 & \lVert x\rVert_2 \leq \delta \\ \delta\lVert x\rVert_2 - 0.5\delta^2 & \text{otherwise} \end{cases}$	for $\delta > 0$	50	25
Elastic Net	$f(x) := \alpha\lVert x\rVert_1 + 0.5\beta\lVert x\rVert_2^2$		50	25
L1	$f(x) := \lVert x\rVert_1$		50	25
Infinity Norm	$f(x) := \lVert x\rVert_\infty$		50	20
Total Variation	$f(x) := \lVert x\rVert_{\text{TV}}$		50	25

In Table 3, we define the function used for the experiments. In particular:

- Sparse Group Lasso: p is set to 3 and given an $x \in \mathbb{R}^d$, $x^{(\beta_i)}$ is a vector obtained by taking 3 entries of x.
- Huber Loss: δ is set to 0.5.
- Elastic Net: α, β are set to 0.5.