Supplementary material

A Notations and Definitions

Throughout this appendix, we will use the following notations:

- we denote the vectors in bold letters.
- $\nabla f(\boldsymbol{x})$ denotes the gradient of f at $\boldsymbol{x}$.
- $[d]$ denotes the set of all integers between 1 and d: $\{1, .., d\}$.
- $\boldsymbol{u}_i$ denotes the i-th coordinate of vector $\boldsymbol{u}$, and $\nabla_i f(\boldsymbol{x})$ the i-th coordinate of $\nabla f(\boldsymbol{x})$.
- $\|\cdot\|_0$ denotes the ℓ_0 norm (which is not a proper norm).
- $\|\cdot\|$ denotes the ℓ_2 norm.
- $\|\cdot\|_\infty$ denotes the maximum absolute component of a vector.
- $\boldsymbol{x} \sim \mathcal{P}$ denotes that the random variable $\boldsymbol{X}$ (denoted as $\boldsymbol{x}$), of realization $\boldsymbol{x}$, follows a probability distribution $\mathcal{P}$ (we abuse notation by denoting similarly a random variable and its realization).
- $\boldsymbol{x}_1, .., \boldsymbol{x}_n \overset{i.i.d}{\sim} \mathcal{P}$ denotes that we draw n i.i.d. samples of a random variable $\boldsymbol{x}$, each from the distribution $\mathcal{P}$.
- $P(\boldsymbol{x})$ denotes the value of the probability of $\boldsymbol{x}$ according to its probability distribution.
- $\mathbb{E}_{\boldsymbol{x} \sim \mathcal{P}}$ (or simply $\mathbb{E}_{\boldsymbol{x}}$ if there is no possible confusion) to denote the expectation of $\boldsymbol{x}$ which follows the distribution $\mathcal{P}$.
- We denote by $\text{supp}(\boldsymbol{x})$ the support of a vector $\boldsymbol{x}$, that is the set of its non-zero coordinates.
- $|F|$ the cardinality (number of elements) of a set F.
- All the sets we consider are subsets of $[d]$. So for a given set F, F^c denotes the complement of F in $[d]$
- $\mathcal{S}^d(R)$ (or $\mathcal{S}^d(R)$ for simplicity if $R = 1$) denotes the d-sphere of radius R, that is $\mathcal{S}^d(R) = \{\boldsymbol{u} \in \mathbb{R}^d / \|\boldsymbol{u}\| = R\}$.
- $\mathcal{U}(\mathcal{S}^d)$ the uniform distribution on that unit sphere.
- $\beta(d)$ is the surface area of the unit d-sphere defined above.
- $\mathcal{S}^d_S$ denotes a set that we call the restricted d-sphere on S, described as: $\{\boldsymbol{u}_S / \boldsymbol{u} \in \{\boldsymbol{v} \in \mathbb{R}^d / \|\boldsymbol{v}_S\| = 1\}\}$, that is the set of unit vectors supported by S.
- $\mathcal{U}(\mathcal{S}^d_S)$ denotes the uniform distribution on that restricted sphere above.
- We denote by $\boldsymbol{u}_F$ (resp. $\nabla_F f(\boldsymbol{x})$) the hard-thresholding of $\boldsymbol{u}$ (resp. $\nabla f(\boldsymbol{x})$) over the support F, that is, a vector which keeps $\boldsymbol{u}$ (resp. $\nabla f(\boldsymbol{x})$) untouched for the set of coordinates in F, but sets all other coordinates to 0.
- $\binom{[d]}{s}$ denotes the set of all subsets of $[d]$ that contain s elements: $\binom{[d]}{s} = \{S : |S| = s, S \subseteq [d]\}$.
- $\mathcal{U}(\binom{[d]}{s})$ denotes the uniform distribution on the set above.
- $\boldsymbol{I}$ denotes the identity matrix $\boldsymbol{I}_{d \times d}$.
- $\boldsymbol{I}_S$ denotes the identity matrix with 1 on the diagonal only at indices belonging to the support S: $\boldsymbol{I}_{i,i} = 1$ if $i \in S$, and 0 elsewhere.
- $S \ni e$ denotes that set S contains the element e.
- $(\boldsymbol{u}_i)_{i=1}^n$ denotes the n-uple of elements $\boldsymbol{u}_1, .., \boldsymbol{u}_n$.
- Γ denotes the Gamma function [1].
- $\int_A f(\boldsymbol{u}) d\boldsymbol{u}$ denotes the integral of f over the set A.
- $\log$ denotes the natural logarithm (in base e).

B Auxilliary Lemmas

Lemma B.1 ([38] (10))**.** *Let $\boldsymbol{p} \in \mathbb{N}^d$, and denote $p := \sum_{i=1}^d \boldsymbol{p}_i$, we have:*

$$\int_{\mathcal{S}^d} \prod_{i=1}^d \boldsymbol{u}_i^{\boldsymbol{p}_i} d\boldsymbol{u} = 2\frac{\prod_{i=1}^n \Gamma(\boldsymbol{p}_i + 1/2)}{\Gamma(p + d/2)}$$

Proof. The proof is given in [38]. ☐

Lemma B.2. *Let F be a subset of $[d]$, of size s, with $(s, d) \in \mathbb{N}_*^2$. We have the following:*

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\|\boldsymbol{u}_F\| \leq \sqrt{\frac{s}{d}} \tag{5}$$

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\|\boldsymbol{u}_F\|^2 = \frac{s}{d} \tag{6}$$

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\|\boldsymbol{u}_F\|^4 = \frac{(s+2)s}{(d+2)d} \tag{7}$$

Proof. We start by proving (6). Decomposing the norm onto every component, we get:

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\|\boldsymbol{u}_F\|^2 = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\sum_{i\in F}\boldsymbol{u}_i^2 = \sum_{i\in F}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\boldsymbol{u}_i^2 \tag{8}$$

By symmetry, each $\boldsymbol{u}_i$ has the same marginal probability distribution, so:

$$\forall i \in [d] : \qquad \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\boldsymbol{u}_i^2 = \frac{1}{d}\sum_{i=1}^d \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\boldsymbol{u}_i^2 \tag{9}$$

We also know, from the definition of the ℓ_2 norm, and the fact that $\boldsymbol{u}$ is a unit vector, that:

$$\sum_{i=1}^d \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\boldsymbol{u}_i^2 = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\sum_{i=1}^d \boldsymbol{u}_i^2 = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\|\boldsymbol{u}\|^2 = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}1 = 1 \tag{10}$$

Therefore, combining (9) and (10):

$$\forall i \in [d] : \qquad \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\boldsymbol{u}_i^2 = \frac{1}{d}$$

Plugging this into (8), we get (6):

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\|\boldsymbol{u}_F\|^2 = \frac{s}{d}$$

Using Jensen's inequality, (5) follows from (6). Let us now prove (7). By definition of the expectation for a uniform distribution on the unit sphere:

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\|\boldsymbol{u}_F\|^4 = \frac{1}{\beta(d)}\int_{\mathcal{S}^d}\|\boldsymbol{u}_F\|^4 d\boldsymbol{u}$$

We further develop the integral as follows:

$$\begin{aligned}\int_{\mathcal{S}^d}\|\boldsymbol{u}_F\|^4 d\boldsymbol{u} &= \int_{\mathcal{S}^d}(\|\boldsymbol{u}_F\|^2)^2 d\boldsymbol{u} = \int_{\mathcal{S}^d}(\sum_{i\in F}\boldsymbol{u}_i^4 + \sum_{(i,j)\in F, j\neq i}\boldsymbol{u}_i^2\boldsymbol{u}_j^2)d\boldsymbol{u} \\ &= s\int_{\mathcal{S}^d}\boldsymbol{u}_1^4 d\boldsymbol{u} + 2\binom{s}{2}\int_{\mathcal{S}^d}\boldsymbol{u}_1^2\boldsymbol{u}_2^2 d\boldsymbol{u} \quad \text{(by symmetry)}\end{aligned}$$

Using Lemma B.1 in the expression above, with $\boldsymbol{p}^{(a)} := (2, 0, ..., 0)$, and $\boldsymbol{p}^{(b)} := (1, 1, 0, ..., 0)$, we obtain:

$$\begin{aligned}\int_{\mathcal{S}^d}\|\boldsymbol{u}_F\|^4 d\boldsymbol{u} &= s\frac{\prod_{i=1}^d \Gamma(\boldsymbol{p}_k^{(a)} + \frac{1}{2})}{\Gamma(2 + d/2)} + 2\frac{s(s-1)}{2}2\frac{\prod_{i=1}^d \Gamma(\boldsymbol{p}_k^{(b)} + 1/2)}{\Gamma(2 + d/2)} \\ &\overset{(a)}{=} \frac{6s\sqrt{\pi}^d}{(d+2)d\Gamma(d/2)} + \frac{2s(s-1)\sqrt{\pi}^d}{(d+2)d\Gamma(d/2)} = \frac{2(s+2)s\sqrt{\pi}^d}{(d+2)d\Gamma(d/2)}\end{aligned}$$

Where in (a) we used the fact that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$. So:

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\|\boldsymbol{u}_F\|^4 = \frac{1}{\beta(d)}\int_{S^d}\|\boldsymbol{u}_F\|^4 d\boldsymbol{u} \overset{(b)}{=} \frac{s+2}{d+2}\frac{s}{d}$$

Where (b) comes from the closed form for the area of a d unit sphere: $\beta(d) = \frac{2\sqrt{\pi}^d}{\Gamma(\frac{d}{2})}$ ☐

Lemma B.3 ([12], Lemma 7.3.b)**.**

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\boldsymbol{u}\boldsymbol{u}^T = \frac{1}{d}\boldsymbol{I}$$

Proof. The proof is given in [12]. ☐

Lemma B.4 ([36], Theorem 1; [45], Lemma 17)**.** *Let $\boldsymbol{b} \in \mathbb{R}^d$ be an arbitrary d-dimensional vector and $\boldsymbol{a} \in \mathbb{R}^d$ be any k-sparse vector. Denote $\bar{k} = \|\boldsymbol{a}\|_0 \leq k$, and $\boldsymbol{b}_k$ the vector $\boldsymbol{b}$ with all the $d-k$ smallest components set to 0 (that is, $\boldsymbol{b}_k$ is the best k-sparse approximation of $\boldsymbol{b}$). Then, we have the following bound:*

$$\|\boldsymbol{b}_k - \boldsymbol{a}\|^2 \leq \delta\|\boldsymbol{b} - \boldsymbol{a}\|^2, \quad \delta = 1 + \frac{\beta + \sqrt{(4+\beta)\beta}}{2}, \quad \beta = \frac{\min\{\bar{k}, d-k\}}{k - \bar{k} + \min\{\bar{k}, d-k\}}$$

Proof. The proof is given in [36]. ☐

Corollary B.1. *With the notations and variables above in Lemma B.4, we also have the following, simpler bound, from [45]:*

$$\|\boldsymbol{b}_k - \boldsymbol{a}\| \leq \gamma\|\boldsymbol{b} - \boldsymbol{a}\|$$

with

$$\gamma = \sqrt{1 + \left(\bar{k}/k + \sqrt{\left(4 + \bar{k}/k\right)\bar{k}/k}\right)/2}$$

Proof. There are two possibilities for β in Lemma B.4: either $\beta = \frac{\bar{k}}{k}$ (if $d - k > \bar{k}$) or $\beta = \frac{d-k}{d-\bar{k}}$ (if $d - k \leq \bar{k}$). In the latter case:

$$d-k \leq \bar{k} \implies d-\bar{k} \leq k \implies \frac{k-\bar{k}}{d-\bar{k}} \geq \frac{k-\bar{k}}{k} \implies 1 - \frac{k-\bar{k}}{d-\bar{k}} \leq 1 - \frac{k-\bar{k}}{k} \implies \frac{d-k}{d-\bar{k}} \leq \frac{\bar{k}}{k}$$

Therefore, in both cases, $\beta \leq \frac{\bar{k}}{k}$, which, plugging into Lemma B.4, gives Corollary B.1. ☐

C Proof of Proposition 1

With an abuse of notation, let us denote by f any function $f_{\boldsymbol{\xi}}$ for some given value of the noise $\boldsymbol{\xi}$. First, we derive in section C.1 the error of the gradient estimate if we sample only one direction ($q = 1$). Then, in section C.2, we show how sampling q directions reduces the error of the gradient estimator, producing the results of Proposition 1.

C.1 One direction estimator

Throughout all this section, we assume that $q = 1$ for the gradient estimator $\hat{\nabla} f(x)$ defined in (2).

C.1.1 Expected deviation from the mean

Lemma C.1. *For any (L_{s_2}, s_2)-RSS function f, using the gradient estimator $\hat{\nabla} f(x)$ defined in* (2) *with $q = 1$, we have, for any support $F \in [d]$, with $|F| = s$:*

$$\left\|\mathbb{E}\left[\hat{\nabla}_F f(\boldsymbol{x})\right] - \nabla_F f(\boldsymbol{x})\right\|^2 \leq \varepsilon_\mu \mu^2$$

with $\varepsilon_\mu = L_{s_2}^2 sd$

Proof. From the definition of the gradient estimator in (2):

$$\|\mathbb{E}[\hat{\nabla}_F f(\boldsymbol{x})] - \nabla_F f(\boldsymbol{x})\| = \left\| \mathbb{E} d \frac{f(\boldsymbol{x} + \mu \boldsymbol{u}) - f(\boldsymbol{x})}{\mu} \boldsymbol{u}_F - \nabla_F f(\boldsymbol{x}) \right\|$$

Now, (L_{s_2}, s_2)-RSS implies continuous differentiability over an s_2-sparse direction (since (L_{s_2}, s_2)-RSS actually equals Lipschitz continuity of the gradient over any s_2-sparse set, which implies continuity of the gradient over those sets). Therefore, from the mean value theorem, , we have, for some $c \in [0, \mu]$: $\frac{f(\boldsymbol{x}+\mu\boldsymbol{u}) - f(\boldsymbol{x})}{\mu} = \langle \nabla f(\boldsymbol{x} + c\boldsymbol{u}), \boldsymbol{u} \rangle$. We now use the following result:

$$\mathbb{E}\boldsymbol{u}\boldsymbol{u}^T = \mathbb{E}_{S \sim \binom{[d]}{s_2}} \mathbb{E}_{\boldsymbol{u} \sim \mathcal{U}(\mathcal{S}^d_S)} \boldsymbol{u}\boldsymbol{u}^T \overset{(a)}{=} \mathbb{E}_{S \sim \binom{[d]}{s_2}} \frac{1}{s_2} \boldsymbol{I}_S = \frac{1}{s_2} \mathbb{E}_{S \sim \binom{[d]}{s_2}} \boldsymbol{I}_S \overset{(b)}{=} \frac{1}{s_2} \frac{s_2}{d} \boldsymbol{I} = \frac{1}{d} \boldsymbol{I}$$

Where for (a) comes from applying Lemma B.3 to the unit sub-sphere on the support S, and (b) follows by observing that each diagonal element of index i actually follows a Bernoulli distribution of parameter $\frac{s_2}{d}$, since there are $\binom{d-1}{s_2-1}$ arrangements of the support which contain i, over $\binom{d}{s_2}$ total arrangements, which gives a probability $p = \frac{\binom{d-1}{s_2-1}}{\binom{d}{s_2}} = \frac{(d-1)! s_2! (d-s_2)!}{(s_2-1)!(d-1-(s_2-1))! d!} = \frac{s_2}{d}$ to get the value 1 at i.

This allows to factor the true gradient into the scalar product:

$$\begin{aligned} \|\mathbb{E}[\hat{\nabla}_F f(\boldsymbol{x})] - \nabla_F f(\boldsymbol{x})\| &= d \|\mathbb{E} \langle \nabla f(\boldsymbol{x} + c\boldsymbol{u}) - \nabla f(\boldsymbol{x}), \boldsymbol{u} \rangle \boldsymbol{u}_F\| \\ &\leq d \mathbb{E} \|\boldsymbol{u}_F \boldsymbol{u}^T [\nabla f(\boldsymbol{x} + c\boldsymbol{u}) - \nabla f(\boldsymbol{x})]\| \end{aligned}$$

where the last inequality follows from the property $\mathbb{E}\|\boldsymbol{X} - \mathbb{E}\boldsymbol{X}\|^2 = \mathbb{E}\|\boldsymbol{X}\|^2 - \|\mathbb{E}\boldsymbol{X}\|^2$, which implies $\|\mathbb{E}\boldsymbol{X}\| = \sqrt{\mathbb{E}\|\boldsymbol{X}\|^2 - \mathbb{E}\|(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})\|^2} \leq \mathbb{E}\|\boldsymbol{X}\|$, for any multidimensional random variable $\boldsymbol{X}$. Using the Cauchy-Schwarz inequality, we obtain:

$$\|\mathbb{E}[\hat{\nabla}_F f(\boldsymbol{x})] - \nabla_F f(\boldsymbol{x})\| \leq \mathbb{E}_{S \sim \binom{[d]}{s_2}} \mathbb{E}_{\boldsymbol{u} \sim \mathcal{U}(\mathcal{S}^d_S)} \|\boldsymbol{u}_F\| \|\boldsymbol{u}\| \|\nabla_S f(\boldsymbol{x} + c\boldsymbol{u}) - \nabla_S f(\boldsymbol{x})\|$$

Since $f \in (L_{s_2}, s_2)$-RSS and $\|\boldsymbol{u}_s\|_0 \leq s_2$, we have: $\|\nabla_S f(\boldsymbol{x} + c\boldsymbol{u}) - \nabla_S f(\boldsymbol{x})\| \leq L_{s_2} \|c\boldsymbol{u}\|$. We also have $c \in [0, \mu]$, which implies $\|c\boldsymbol{u}\| \leq \mu \|\boldsymbol{u}\|$. Therefore:

$$\begin{aligned} \|\mathbb{E}[\hat{\nabla}_F f(\boldsymbol{x})] - \nabla_F f(\boldsymbol{x})\| &\leq \mathbb{E}_S \mathbb{E}_{\boldsymbol{u}} d L_{s_2} \mu \|\boldsymbol{u}_F\| \|\boldsymbol{u}\| \|\boldsymbol{u}\| = \mathbb{E}_S \mathbb{E}_{\boldsymbol{u}} d L_{s_2} \mu \|\boldsymbol{u}_F\| \|\boldsymbol{u}\|^2 = \mathbb{E}_S \mathbb{E}_{\boldsymbol{u}} d L_{s_2} \mu \|\boldsymbol{u}_F\| \\ &\overset{(a)}{\leq} d L_{s_2} \mu \mathbb{E}_S \mathbb{E}_{\boldsymbol{u}} \sqrt{\frac{|S \cap F|}{s_2}} \\ &\overset{(b)}{\leq} d L_{s_2} \mu \sqrt{\mathbb{E}_S \frac{|S \cap F|}{s_2}} = d L_{s_2} \mu \sqrt{\mathbb{E}_k \mathbb{E}_{S \mid |S \cap F| = k} \frac{k}{s_2}} \\ &= d L_{s_2} \mu \sqrt{\frac{s s_2}{d s_2}} = L_{s_2} \mu \sqrt{sd} \end{aligned}$$

Where (a) follows from Lemma B.2, restricted to the support S, and (b) follows from Jensen's inequality. □

C.1.2 Expected norm

Lemma C.2. *For any (L_{s_2}, s_2)-RSS function f, using the gradient estimator $\hat{\nabla} f(x)$ defined in (2) with $q = 1$, we have, for any support $F \in [d]$, with $|F| = s$:*

$$\mathbb{E}\|\hat{\nabla}_F f(\boldsymbol{x})\|^2 = \varepsilon_F \|\nabla_F f(\boldsymbol{x})\|^2 + \varepsilon_{F^c} \|\nabla_{F^c} f(\boldsymbol{x})\|^2 + \varepsilon_{abs} \mu^2$$

with:
(i) $\varepsilon_F = \frac{2d}{(s_2+2)} \left(\frac{(s-1)(s_2-1)}{d-1} + 3 \right)$
(ii) $\varepsilon_{F^c} = \frac{2d}{(s_2+2)} \left(\frac{s(s_2-1)}{d-1} \right)$
(iii) $\varepsilon_{abs} = 2 d L_s^2 s s_2 \left(\frac{(s-1)(s_2-1)}{d-1} + 1 \right)$

Proof.

$$\begin{aligned}\mathbb{E}\|\hat{\nabla}_F f(\boldsymbol{x})\|^2 &= \mathbb{E}\left\| d\frac{f(\boldsymbol{x}+\mu\boldsymbol{u})-f(\boldsymbol{x})}{\mu}\boldsymbol{u}_F\right\|^2\\ &= \mathbb{E}\frac{d^2}{\mu^2}|f(\boldsymbol{x}+\mu\boldsymbol{u})-f(\boldsymbol{x})|^2\|\boldsymbol{u}_F\|^2\\ &= \frac{d^2}{\mu^2}\mathbb{E}[f(\boldsymbol{x}+\mu\boldsymbol{u})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}),\mu\boldsymbol{u}\rangle+\langle\nabla f(\boldsymbol{x}),\mu\boldsymbol{u}\rangle]^2\|\boldsymbol{u}_F\|^2\end{aligned}$$

Using the mean value theorem, we obtain that for a certain $c \in (0, \mu)$, we have:

$$f(\boldsymbol{x}+\mu\boldsymbol{u})-f(\boldsymbol{x}) = \langle\nabla f(\boldsymbol{x}+c), \mu\boldsymbol{u}\rangle$$

Therefore, plugging this in the above:

$$\begin{aligned}\mathbb{E}\|\hat{\nabla}_F f(\boldsymbol{x})\|^2 &\leq d^2\mathbb{E}[\langle\nabla f(\boldsymbol{x}+c\boldsymbol{u})-\nabla f(\boldsymbol{x}),\boldsymbol{u}\rangle+\langle\nabla f(\boldsymbol{x}),\boldsymbol{u}\rangle]^2\|\boldsymbol{u}_F\|^2\\ &\overset{(a)}{\leq} d^2\mathbb{E}\left[2\langle\nabla f(\boldsymbol{x}+c\boldsymbol{u})-\nabla f(\boldsymbol{x}),\boldsymbol{u}\rangle^2\|\boldsymbol{u}_F\|^2+\langle\nabla f(\boldsymbol{x}),\boldsymbol{u}\rangle^2\|\boldsymbol{u}_F\|^2\right]\\ &\leq 2d^2\mathbb{E}[\|\nabla f(\boldsymbol{x}+c\boldsymbol{u})-\nabla f(\boldsymbol{x})\|^2\|\boldsymbol{u}\|^2\|\boldsymbol{u}_F\|^2+\langle\nabla f(\boldsymbol{x}),\boldsymbol{u}\rangle^2\|\boldsymbol{u}_F\|^2]\\ &\overset{\leq}{(b)}2d^2\mathbb{E}[L_s^2\mu^2\|\boldsymbol{u}\|^2\|\boldsymbol{u}\|^2\|\boldsymbol{u}_F\|^2+\langle\nabla f(\boldsymbol{x}),\boldsymbol{u}\rangle^2\|\boldsymbol{u}_F\|^2]\\ &\overset{(c)}{=} 2d^2\mathbb{E}[L_s^2\mu^2\|\boldsymbol{u}_F\|^2+\langle\nabla f(\boldsymbol{x}),\boldsymbol{u}\rangle^2\|\boldsymbol{u}_F\|^2]\\ &= 2d^2[L_s^2\mu^2\mathbb{E}\|\boldsymbol{u}_F\|^2+\nabla f(\boldsymbol{x})^T\left(\mathbb{E}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_F\|^2\right)\nabla f(\boldsymbol{x})]\\ &= 2d^2[L_{s_2}^2\mu^2\mathbb{E}\|\boldsymbol{u}_F\|^2+\nabla f(\boldsymbol{x})^T(\mathbb{E}_{S\sim\binom{[d]}{s_2}}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_F\|^2)\nabla f(\boldsymbol{x})]\\ &\overset{(d)}{=} 2d^2[L_{s_2}^2\mu^2\mathbb{E}\|\boldsymbol{u}_F\|^2+\mathbb{E}_{S\sim\binom{[d]}{s_2}}[\nabla f(\boldsymbol{x})^T(\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_F\|^2)\nabla f(\boldsymbol{x})]] \quad (11)\end{aligned}$$

Where (a) follows from the fact that for any $(a, b) \in \mathbb{R}^2 : (a+b)^2 \leq 2a^2 + 2b^2$, (b) follows from the Cauchy-Schwarz inequality, (c) follows from the fact that $\|\boldsymbol{u}\| = 1$ since $\boldsymbol{u} \in \mathcal{S}_S^d$, and (d) follows by linearity of expectation. Let us turn to computing the following expression above: $\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_F\|^2$. We start by distinguishing the indices that belong to F and those that do not. By symmetry, denoting $i_1, ..., i_s$ the elements of F:

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_{i_1}^2\|\boldsymbol{u}_F\|^2 = ... = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_{i_s}^2\|\boldsymbol{u}_F\|^2$$

Therefore, for all $i \in F$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_i^2\|\boldsymbol{u}_F\|^2 &= \frac{1}{|S\cap F|}\sum_{j=1}^{s}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_{i_j}^2\|\boldsymbol{u}_F\|^2\\ &= \frac{1}{|S\cap F|}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\sum_{j=1}^{s}\boldsymbol{u}_{i_j}^2\|\boldsymbol{u}_F\|^2 = \frac{1}{|S\cap F|}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\|\boldsymbol{u}_F\|^4 \quad (12)\end{aligned}$$

By definition of the restricted d-sphere on F (see section A), for all $\boldsymbol{u} \in \mathcal{S}_S^d$, if $i \notin S$: $\boldsymbol{u}_i = 0$. Therefore, since the exact indices of the elements of F do not matter in the expected value (12), but only their cardinality, (12) can be rewritten using a simpler expectation over a unit $|S|$-sphere as follows :

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\|\boldsymbol{u}_F\|^4 = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^{|S|})}\|\boldsymbol{u}_{[|S\cap F|]}\|^4$$

Using Lemma B.2 to get a closed form expression of the expected value above, we further obtain:

$$\forall i \in F : \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\boldsymbol{u}_i^2\|\boldsymbol{u}_F\|^2 = \frac{1}{|S\cap F|}\frac{|S\cap F|(|S\cap F|+2)}{d(d+2)} = \frac{|S\cap F|+2}{d(d+2)} \quad (13)$$

Similarly, by symmetry, denoting $i_1, ..., i_{d-s}$ the elements of F^c:

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_{i_j}^2\|\boldsymbol{u}_F\|^2 = ... = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_{i_j}^2\|\boldsymbol{u}_F\|^2$$

Therefore, for all $i \notin F$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_i^2\|\boldsymbol{u}_F\|^2 &= \frac{1}{d-s}\sum_{j=1}^{d-s}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_{i_j}^2\|\boldsymbol{u}_F\|^2 = \frac{1}{d-s}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\sum_{j=1}^{d-s}\boldsymbol{u}_{i_j}^2\|\boldsymbol{u}_F\|^2 \\ &\overset{(a)}{=} \frac{1}{d-s}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}(\|\boldsymbol{u}\|^2 - \|\boldsymbol{u}_F\|^2)\|\boldsymbol{u}_F\|^2 \\ &\overset{(b)}{=} \frac{1}{d-s}(\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\|\boldsymbol{u}_F\|^2 - \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\|\boldsymbol{u}\|^4)\end{aligned}$$

Where (a) follows from the Pythagorean theorem and (b) follows from $\|\boldsymbol{u}\| = 1$. Similarly as before, rewriting those expected values and using Lemma B.2, we obtain:

$$\forall i \notin F : \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\boldsymbol{u}_i^2\|\boldsymbol{u}_F\|^2 = \frac{1}{d-|S\cap F|}\frac{|S\cap F|(d+2-(|S\cap F|+2))}{d(d+2)} = \frac{|S\cap F|}{d(d+2)} \tag{14}$$

Finally, by symmetry of the distribution $\mathcal{U}(\mathcal{S}_S^d)$, we have, for all $(i,j)\in[d]^2$ with $i\neq j$:

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_i\boldsymbol{u}_j\|\boldsymbol{u}_F\|^2 = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}(-\boldsymbol{u}_i)\boldsymbol{u}_j\|\boldsymbol{u}_F\|^2 = -\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_i\boldsymbol{u}_j\|\boldsymbol{u}_F\|^2$$

Therefore, for all $(i,j)\in[d]^2, i\neq j$:

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}_i\boldsymbol{u}_j\|\boldsymbol{u}_F\|^2 = 0 \tag{15}$$

Therefore, combining (13), (14) and (15), we obtain:

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_F\|^2 = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_d \end{bmatrix}$$

With, for all $i\in[d]: a_i = \begin{cases}\frac{|S\cap F|+2}{d(d+2)} \text{ if } i\in F \\ \frac{|S\cap F|}{d(d+2)} \text{ if } i\notin F\end{cases}$. Plugging this back into (11), we obtain:

$$\begin{aligned}A &:= \mathbb{E}_{S\sim\binom{[d]}{s_2}}[\nabla f(\boldsymbol{x})^T\left(\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_F\|^2\right)\nabla f(\boldsymbol{x})] \\ &= \mathbb{E}_{S\sim\binom{[d]}{s_2}}\left[\frac{|S\cap F|+2}{s_2(s_2+2)}\|\nabla_{S\cap F}f(\boldsymbol{x})\|^2 + \frac{|S\cap F|}{s_2(s_2+2)}\|\nabla_{S\setminus(S\cap F)}f(\boldsymbol{x})\|^2\right] \\ &= \frac{1}{s_2(s_2+2)}\Bigg[\mathbb{E}_{S\sim\binom{[d]}{s_2}}\left[|S\cap F|\ \|\nabla_{F\cap S}f(\boldsymbol{x})\|^2\right] \\ &\quad +2\mathbb{E}_{S\sim\binom{[d]}{s_2}}\left[\|\nabla_{F\cap S}f(\boldsymbol{x})\|^2 + |S\cap F|\ \|\nabla_{S\setminus(S\cap F)}f(\boldsymbol{x})\|^2\right]\Bigg]\end{aligned} \tag{16}$$

We will now develop the expected values above using the law of total expectation, to exhibit the role of the random variable k denoting the size of $S\cap F$. Given that we sample s_2 indices from $[d]$ without replacement, k follows a hypergeometric distribution with, as parameters, population size d, number of success states s and number of draws s_2, which we denote $\mathcal{H}(d,s,s_2)$. For simplicity, we will use the following notations for the expected values: $\mathbb{E}_k[\cdot] := \mathbb{E}_{k\sim\mathcal{H}(d,s,s_2)}[\cdot]$, and $\mathbb{E}_{S||S\cap F|=k}[\cdot] = \mathbb{E}_{S\sim\binom{[d]}{s_2}||S\cap F|=k}[\cdot]$. Therefore, rewriting (16) using the law of total expectation, we obtain:

$$\begin{aligned}A &= \frac{1}{s_2(s_2+2)}\big[\mathbb{E}_k\mathbb{E}_{S||S\cap F|=k}k\|\nabla_{S\cap F}f(\boldsymbol{x})\|^2 + 2\mathbb{E}_k\mathbb{E}_{S||S\cap F|=k}\|\nabla_{S\cap F}f(\boldsymbol{x})\|^2 \\ &\quad +\mathbb{E}_k\mathbb{E}_{S||S\cap F|=k}k\|\nabla_{S\setminus(S\cap F)}f(\boldsymbol{x})\|^2\big] \\ &= \frac{1}{s_2(s_2+2)}\big[\mathbb{E}_k k\mathbb{E}_{S||S\cap F|=k}\|\nabla_{S\cap F}f(\boldsymbol{x})\|^2 + 2\mathbb{E}_S\mathbb{E}_{S||S\cap F|=k}\|\nabla_{S\cap F}f(\boldsymbol{x})\|^2 \\ &\quad +\mathbb{E}_k k\mathbb{E}_{S||S\cap F|=k}\|\nabla_{S\setminus(S\cap F)}f(\boldsymbol{x})\|^2\big]\end{aligned} \tag{17}$$

To compute the conditional expectations above, let us consider the first of them (the other ones will follow similarly) : $\mathbb{E}_{S||S\cap F|=k}\|\nabla_{S\cap F} f(\boldsymbol{x})\|^2$. Given some k, from the multiplication principle in combinatorics, we can have $\binom{d}{k}\binom{d-s}{s_2-k}$ arrangements of supports such that k elements of that support are in F (because it means there are k elements in F and $s_2 - k$ elements outside of F). So the conditional probability of each of those supports S, assuming they indeed have at least one element in common with F, is $\left(\binom{d}{k}\binom{d-s}{s_2-k}\right)^{-1}$. Otherwise it is 0. To rewrite it:

$$P(S||S\cap F| = k) = \begin{cases} \left(\binom{d}{k}\binom{d-s}{s_2-k}\right)^{-1} \text{ if } S\cap F \neq \varnothing \\ 0 \text{ if } S\cap F \neq \varnothing \end{cases}$$

So, developing $\mathbb{E}_{S||S\cap F|=k}\|\nabla_{S\cap F} f(\boldsymbol{x})\|^2$ using the definition of conditional probability, we have:

$$\begin{aligned}
\mathbb{E}_{S||S\cap F|=k}\|\nabla_{S\cap F} f(\boldsymbol{x})\|^2 &= \sum_{S} P(S|\ |S\cap F| = k) \sum_{i\in S\cap F} \nabla_i f(\boldsymbol{x})^2 \\
&= \sum_{S/|S\cap F|=k} \left(\binom{d}{k}\binom{d-s}{s_2-k}\right)^{-1} \sum_{i\in S\cap F} \nabla_i f(\boldsymbol{x})^2 \\
&= \left(\binom{d}{k}\binom{d-s}{s_2-k}\right)^{-1} \sum_{S/|S\cap F|=k}\ \sum_{i\in S\cap F} \nabla_i f(\boldsymbol{x})^2 \\
&\overset{(a)}{=} \left(\binom{d}{k}\binom{d-s}{s_2-k}\right)^{-1} \sum_{i\in F}\ \sum_{S/((|S\cap F|=k),(S\ni i))} \nabla_i f(\boldsymbol{x})^2 \\
&\overset{(b)}{=} \left(\binom{d}{k}\binom{d-s}{s_2-k}\right)^{-1} \sum_{i\in F} \binom{s-1}{k-1}\binom{d-s}{s_2-k} \nabla_i f(\boldsymbol{x})^2 \\
&= \frac{s}{k}\sum_{i\in F} \nabla_i f(\boldsymbol{x})^2 \\
&= \frac{s}{k}\|\nabla_F f(\boldsymbol{x})\|^2 \qquad (18)
\end{aligned}$$

Where (a) follows by re-arranging the sum, and (b) follows by observing that by the multiplication principle, there are $\binom{s-1}{k-1}\binom{d-s}{s_2-k}$ possible arrangements of support such that: $(|S\cap F| = k), (S \ni i)$, since one element of S is already fixed to be i, so there remains $k-1$ indices to arrange over $s-1$ possibilities, and still $s_2 - k$ indices to arrange over $d - s$ possibilities. Similarly, to (18) we have, for the second expectation:

$$\mathbb{E}_{S||S\cap F|=k}\|\nabla_{S\setminus(S\cap F)} f(\boldsymbol{x})\|^2 = \frac{s_2-k}{d-s}\|\nabla_{F^c} f(\boldsymbol{x})\|^2 \qquad (19)$$

Therefore, plugging (18) and (19) into (17)

$$\begin{aligned}
A &= \frac{1}{s_2(s_2+2)}\left[\mathbb{E}_k k\frac{k}{s}\|\nabla_F f(\boldsymbol{x})\|^2 + 2\mathbb{E}_k\frac{k}{s}\|\nabla_F f(\boldsymbol{x})\|^2 + \mathbb{E}_k k\frac{s_2-k}{d-s}\|\nabla_{F^c} f(\boldsymbol{x})\|^2\right] \\
&= \frac{1}{s_2(s_2+2)}\left[\frac{1}{s}\|\nabla_F f(\boldsymbol{x})\|^2\left[\mathbb{E}_k k^2 + 2\mathbb{E}_k k\right] + \|\nabla_{F^c} f(\boldsymbol{x})\|^2\left[\frac{s_2}{d-s}\left(\mathbb{E}_k k\right) - \frac{1}{d-s}\mathbb{E}_k k^2\right]\right] \qquad (20)
\end{aligned}$$

Since k follows a hypergeometric distribution $\mathcal{H}(d, s, s_2)$, its expected value is given in closed form by: $\mathbb{E}_k k = \frac{ss_2}{d}$ (see [43], section 2.1.3). We can also express the non-centered moment of order 2, using the formula for Var$(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, which holds for a random variable X, where $Var(X)$ denotes the variance of X:

$$\begin{aligned}
\mathbb{E}_k k^2 &= Var(k) + (\mathbb{E}_k[k])^2 \overset{(a)}{=} \frac{ss_2}{d}\frac{d-s}{d}\frac{d-s_2}{d-1} + \left(\frac{ss_2}{d}\right)^2 = \frac{ss_2}{d}\left(\frac{d-s}{d}\frac{d-s_2}{d-1} + \frac{ss_2}{d}\right) \\
&= \frac{ss_2}{d}\left(\frac{d^2 - sd - s_2 d + ss_2 + ss_2 d - ss_2}{d(d-1)}\right) = \frac{ss_2}{d}\left(\frac{d - s - s_2 + ss_2}{d-1}\right) \\
&= \frac{ss_2}{d}\left(\frac{(s-1)(s_2-1)}{d-1} + 1\right)
\end{aligned}$$

Where (a) follows by the closed form for the variance of a hypergeometric variable given in [43]. Therefore, plugging in into (20):

$$
\begin{aligned}
&\mathbb{E}_S \nabla f(\boldsymbol{x})^T \left(\mathbb{E}_{\mathcal{U}_S|S} \boldsymbol{u}\boldsymbol{u}^T \|\boldsymbol{u}_F\|^2\right) \nabla f(\boldsymbol{x}) \\
&= \frac{1}{s_2(s_2+2)} \left[\frac{1}{s}\|\nabla_F f(\boldsymbol{x})\|^2 \left[\frac{ss_2}{d}\left(\frac{(s-1)(s_2-1)}{d-1}+1\right) + 2\frac{ss_2}{d}\right]\right] \\
&\quad + \frac{1}{s_2(s_2+2)} \|\nabla_{F^c} f(\boldsymbol{x})\|^2 \left[\frac{s_2}{d-s}\frac{ss_2}{d} - \frac{1}{d-s}\frac{ss_2}{d}\left(\frac{(s-1)(s_2-1)}{d-1}+1\right)\right] \\
&= \frac{1}{s_2+2}\left[\|\nabla_F f(\boldsymbol{x})\|^2 \left[\frac{1}{d}\left(\frac{(s-1)(s_2-1)}{d-1}+3\right)\right]\right. \\
&\quad \left. + \|\nabla_{F^c} f(\boldsymbol{x})\|^2 \left[\frac{s}{(d-s)d}\left(s_2 - \left(\frac{(s-1)(s_2-1)}{d-1}+1\right)\right)\right]\right] \\
&= \frac{1}{d(s_2+2)}\left[\|\nabla_F f(\boldsymbol{x})\|^2 \left[\left(\frac{(s-1)(s_2-1)}{d-1}+3\right)\right]\right. \\
&\quad \left. + \|\nabla_{F^c} f(\boldsymbol{x})\|^2 \left[\frac{s}{(d-s)}\left(s_2 - \left(\frac{(s-1)(s_2-1)}{d-1}+1\right)\right)\right]\right] \qquad (21)
\end{aligned}
$$

Let us simplify the rightmost term:

$$
\begin{aligned}
\frac{s}{(d-s)}\left(s_2 - \left(\frac{(s-1)(s_2-1)}{d-1}+1\right)\right) &= \frac{s(s_2-1)}{d-s}\left[1 - \frac{s-1}{d-1}\right] \\
&= \frac{s(s_2-1)}{(d-s)}\left[\frac{d-s}{d-1}\right] = \frac{s(s_2-1)}{d-1}
\end{aligned}
$$

Plugging it back into (21):

$$
\begin{aligned}
&\mathbb{E}_S \nabla f(\boldsymbol{x})^T \left(\mathbb{E}_{\mathcal{U}_S|S} \boldsymbol{u}\boldsymbol{u}^T \|\boldsymbol{u}_F\|^2\right) \nabla f(\boldsymbol{x}) \\
&= \frac{1}{d(s_2+2)}\left[\|\nabla_F f(\boldsymbol{x})\|^2 \left(\frac{(s-1)(s_2-1)}{d-1}+3\right) + \|\nabla_{F^c} f(\boldsymbol{x})\|^2 \left(\frac{s(s_2-1)}{d-1}\right)\right]
\end{aligned}
$$

Finally, plugging this back into (11):

$$
\begin{aligned}
\mathbb{E}\|\hat{\nabla}_F f(\boldsymbol{x})\|^2 &= 2d^2\left[L_{s_2}^2\mu^2\mathbb{E}\|\boldsymbol{u}_F\|^2 + \nabla f(\boldsymbol{x})^T\left(\mathbb{E}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_F\|^2\right)\nabla f(\boldsymbol{x})\right] \\
&= 2d^2\left[L_{s_2}^2\mu^2\mathbb{E}_k\mathbb{E}_{\boldsymbol{u}||S\cap F|=k}\|\boldsymbol{u}_F\|^2 + \nabla f(\boldsymbol{x})^T\left(\mathbb{E}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_F\|^2\right)\nabla f(\boldsymbol{x})\right] \\
&= 2d^2\left[L_{s_2}^2\mu^2\mathbb{E}_k k^2 + \nabla f(\boldsymbol{x})^T\left(\mathbb{E}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_F\|^2\right)\nabla f(\boldsymbol{x})\right] \\
&= d2L_{s_2}^2\mu^2 ss_2\left(\frac{(s-1)(s_2-1)}{d-1}+1\right) \\
&\quad + \frac{2d}{(s_2+2)}\left[\|\nabla_F f(\boldsymbol{x})\|^2\left(\frac{(s-1)(s_2-1)}{d-1}+3\right) + \|\nabla_{F^c} f(\boldsymbol{x})\|^2\left(\frac{s(s_2-1)}{d-1}\right)\right]
\end{aligned}
$$

□

C.2 Batched-version of the one-direction estimator

We now describe how sampling $q \geq 1$ random directions improves the gradient estimate. Our proof is similar to the proof of Lemma 2 in [24], however we make sure that it works for our random support gradient estimator, and with our new expression in C.2, which depends on the two terms $\|\nabla_F f(\boldsymbol{x})\|^2$ and $\|\nabla_{F^c} f(\boldsymbol{x})\|^2$. We express our results here in the form of a general lemma, depending only on the general bounding factors ε_F, ε_{F^c}, ε_{abs} and ε_μ defined below, in such a way that the proof of Proposition 1 follows immediately from plugging the results of Lemma C.1 and C.2 into Lemma C.3 below.

Lemma C.3. *For any (L_{s_2}, s_2)-RSS function f, we use the gradient estimator $\hat{\nabla} f(x)$ defined in* (2) *with $q \geq 1$. Let us suppose that the estimator $\hat{\nabla} f(x)$ is such that for $q = 1$, it verifies the following bounds for some ε_F, ε_{F^c}, ε_{abs} and ε_μ in $\mathbb{R}_+^*$, for any support $F \in [d]$, with $|F| = s$:*

(i) $\|\mathbb{E}\hat{\nabla}_F f(\boldsymbol{x}) - \nabla_F f(\boldsymbol{x})\|^2 \leq \varepsilon_\mu \mu^2$, *and*

(ii) $\|\mathbb{E}\hat{\nabla}_F f(\boldsymbol{x})\|^2 \leq \varepsilon_F \|\nabla_F f(\boldsymbol{x})\|^2 + \varepsilon_{F^c} \|\nabla_{F^c} f(x)\|^2 + \varepsilon_{abs} \mu^2$

Then, the estimator $\hat{\nabla} f(x)$ also verifies, for arbitrary $q \geq 1$:

(a) $\|\mathbb{E}\hat{\nabla}_F f(\boldsymbol{x}) - \nabla_F f(\boldsymbol{x})\|^2 \leq \varepsilon_\mu \mu^2$

(b) $\mathbb{E}\left\|\hat{\nabla}_F f(\boldsymbol{x})\right\|^2 \leq \left(\frac{\varepsilon_F}{q} + 2\right) \|\nabla_F f(\boldsymbol{x})\|^2 + \frac{\varepsilon_{F^c}}{q} \|\nabla_{F^c} f(\boldsymbol{x})\|^2 + \left(\frac{\varepsilon_{abs}}{q} + 2\varepsilon_\mu\right) \mu^2$

Proof. Let us denote by $\hat{\nabla} f(\boldsymbol{x}; (\boldsymbol{u}_i)_{i=1}^q)$ the gradient estimate from (2) along the i.i.d. sampled directions $(\boldsymbol{u}_i)_{i=1}^q$ (we simplify it into $\hat{\nabla} f(\boldsymbol{x}; \boldsymbol{u})$ if there is only one direction $\boldsymbol{u}$). We can first see that, since the random directions $\boldsymbol{u}_i$ are independent identically distributed (i.i.d.) we have:

$$\mathbb{E}\hat{\nabla} f(\boldsymbol{x}; (\boldsymbol{u}_i)_{i=1}^q) = \mathbb{E}\frac{1}{q}\sum_{i=1}^q \hat{\nabla} f(\boldsymbol{x}; \boldsymbol{u}_i) = \frac{1}{q}\sum_{i=1}^q \mathbb{E}\hat{\nabla} f(\boldsymbol{x}; \boldsymbol{u}_1) = \mathbb{E}\hat{\nabla} f(\boldsymbol{x}; \boldsymbol{u}_1)$$

This proves C.3 (a). Let us now turn to C.3 (b). We have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\nabla}_F f(\boldsymbol{x}; (\boldsymbol{u}_i)_{i=1}^q)\right\|^2\right] &= \mathbb{E}\left\|\frac{1}{q}\sum_{i=1}^q \hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_i)\right\|^2 \\
&= \frac{1}{q^2}\mathbb{E}\left(\sum_{i=1}^q \hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_i)\right)^\top \left(\sum_{i=1}^q \hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_i)\right) \\
&= \frac{1}{q^2}\sum_{i=1}^q \sum_{j=1}^q \mathbb{E}\left[\hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_i)^\top \hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_j)\right] \\
&\overset{(a)}{=} \frac{1}{q^2}\left[q\mathbb{E}\|\hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_1)\|^2 + \sum_{i=1}^q \sum_{j=1 (j\neq i)}^q (\mathbb{E}\hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_i))^\top (\mathbb{E}\hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_j))\right] \\
&= \frac{1}{q^2}\left[q\mathbb{E}||\hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_1)\|^2 + q(q-1)\|\mathbb{E}\hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_1)||^2\right] \\
&\overset{(b)}{\leq} \frac{1}{q^2}\left[q\left[\varepsilon_F \|\nabla_F f(\boldsymbol{x})\|^2 + \varepsilon_{F^c}\|\nabla_{F^c} f(\boldsymbol{x})\|^2 + \varepsilon_{abs}\mu^2\right] + q\,(q-1)\left\|\mathbb{E}\hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_1)\right\|^2\right] \qquad (22)
\end{aligned}$$

Where (a) comes from the fact that the random directions are i.i.d. and (b) comes from assumptions (i) and (ii) of the current Lemma (Lemma C.3). Assumption (ii) also allows to bound the last term above in the following way:

$$\begin{aligned}
\|\mathbb{E}\hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_1)\|^2 &\leq 2\|\nabla_F f(\boldsymbol{x}; \boldsymbol{u}_1) - \mathbb{E}\hat{\nabla}_F f(\boldsymbol{x}; \boldsymbol{u}_1)\|^2 + 2\|\nabla_F f(\boldsymbol{x}; \boldsymbol{u}_1)\|^2 \\
&\leq 2\varepsilon_\mu \mu^2 + 2\|\nabla_F f(\boldsymbol{x}; \boldsymbol{u}_1)\|^2 \qquad (23)
\end{aligned}$$

Plugging (23) into (22), we obtain:

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\nabla}_F f(\boldsymbol{x})\right\|^2\right] &\leq \frac{1}{q}\left[\varepsilon_F + 2(q-1)\right] \|\nabla_F f(\boldsymbol{x})\|^2 + \frac{\varepsilon_{F^c}}{q}\|\nabla_{F^c} f(\boldsymbol{x})\|^2 \\
&\quad + \frac{1}{q}\left[\varepsilon_{abs}\mu^2 + 2\,(q-1)\,\varepsilon_\mu \mu^2\right] \\
&\leq \left(\frac{\varepsilon_F}{q} + 2\right) \|\nabla_F f(\boldsymbol{x})\|^2 + \frac{\varepsilon_{F^c}}{q}\|\nabla_{F^c} f(\boldsymbol{x})\|^2 + \left(\frac{\varepsilon_{abs}}{q} + 2\varepsilon_\mu\right)\mu^2
\end{aligned}$$

□

C.3 Proof of Proposition 1

Proof. Proposition 1 (a) and (b) follow by plugging the values of ε_F, ε_{F^c}, ε_{abs} and ε_μ from Lemma C.1 and Lemma C.2 into Lemma C.3. Proposition (c) follows from the inequality $\|\boldsymbol{a} + \boldsymbol{b}\|^2 \leq 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$, for $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R}^p$ with $p \in \mathbb{N}^*$. □

D Proofs of section 4

D.1 Proof of Theorem 1

Proof. We will combine the proof from [45] and [30], using ideas of the proof of Theorem 8 from Nesterov to deal with zeroth order gradient approximations, and ideas from the proof of [45] (Theorem 2 and 5, Lemma 19), to deal with the hard thresholding operation in the convergence rate. Let us call η an arbitrary learning rate, that will be fixed later in the proof. Let us call F the following support $F = F^{(t-1)} \cup F^{(t)} \cup \text{supp}(\boldsymbol{x}^*)$, with $F^{(t)} = \text{supp}(\boldsymbol{x}^t)$. We have, for a given random direction $\boldsymbol{u}$ and function noise $\boldsymbol{\xi}$, at a given timestep t of SZOHT:

$$\begin{aligned}\|\boldsymbol{x}^t - \boldsymbol{x}^* - \eta\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) + \eta\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 = \|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 &- 2\eta\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\rangle \\ &+ \eta^2\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2\end{aligned}$$

Taking the expectation with respect to $\boldsymbol{\xi}$ and to the possible random directions $\boldsymbol{u}_1, ..., \boldsymbol{u}_q$ (that we denote with a simple $\boldsymbol{u}$, abusing notations) at step t, we get:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\boldsymbol{x}^t - \boldsymbol{x}^* - \eta\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) + \eta\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&= \|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 - 2\eta\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}[\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)]\rangle + \eta^2\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&= \|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 - 2\eta\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}[\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)]\rangle \\
&\quad - 2\eta\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \mathbb{E}_{\boldsymbol{\xi}}[\mathbb{E}_{\boldsymbol{u}}\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t)]\rangle + \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\eta^2\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&= \|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 - 2\eta\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \nabla_F f(\boldsymbol{x}^t) - \nabla_F f(\boldsymbol{x}^*)\rangle \\
&\quad - 2\eta\langle\sqrt{\eta}L_{s'}\left(\boldsymbol{x}^t - \boldsymbol{x}^*\right), \frac{1}{\sqrt{\eta}L_{s'}}(\mathbb{E}_{\boldsymbol{\xi}}\mathbb{E}_{\boldsymbol{u}}[\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t))]\rangle \\
&\quad + \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\eta^2\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&\overset{(a)}{\leq} \|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 - 2\eta\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \nabla_F f(\boldsymbol{x}^t) - \nabla_F f(\boldsymbol{x}^*)\rangle + \eta^2 L_{s'}^2\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 \\
&\quad + \frac{1}{L_{s'}^2}\mathbb{E}_{\boldsymbol{\xi}}\|\mathbb{E}_{\boldsymbol{u}}\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t))\|^2 + \eta^2\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \qquad (24)\end{aligned}$$

Where (a) follows from the inequality $2\langle\boldsymbol{u}, \boldsymbol{v}\rangle \leq \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2$ for any $(\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2$. From Proposition 1 (b), since almost each $f_{\boldsymbol{\xi}}$ is $(L_{s'}, s')$-RSS (hence also $(L_{s'}, s_2)$-RSS), we know that for the ε_F, ε_{F^c} and ε_{abs} defined in Proposition 1 (b), we have for almost all $\boldsymbol{\xi}$: $\mathbb{E}_{\boldsymbol{u}}\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t)\|^2 \leq \varepsilon_F\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t)\|^2 + \varepsilon_{F^c}\|\nabla_{F^c} f_{\boldsymbol{\xi}}(\boldsymbol{x}^t)\|^2 + \varepsilon_{\text{abs}}\mu^2$. This allows to develop the last term of (24) into the following:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}.\boldsymbol{u}}\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 &\leq 2\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t)\|^2 + 2\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&\leq 2\varepsilon_F\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t)\|^2 + 2\varepsilon_{F^c}\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_{F^c} f_{\boldsymbol{\xi}}(\boldsymbol{x}^t)\|^2 \\
&\quad + 2\varepsilon_{\text{abs}}\mu^2 + 2\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&\leq 2\varepsilon_F\left[2\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 + 2\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2\right] \\
&\quad + 2\varepsilon_{F^c}\left[2\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_{F^c} f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_{F^c} f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 + 2\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_{F^c} f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2\right] \\
&\quad + 2\varepsilon_{\text{abs}}\mu^2 + 2\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2\end{aligned}$$

Just like the proof in [45], we will express our result in terms of the infinity norm of $\nabla f(\boldsymbol{x}^*)$. For that, we will plug above the two following inequalites: Same as their proof of Lemma 19, we have $\|\nabla_F f(\boldsymbol{x}^*)\| \leq \|\nabla_s f(\boldsymbol{x}^*)\|$ (that is because we will have equality if the sets in the definition of F, namely $F^{(t-1)}$, $F^{(t)}$ and supp$(\boldsymbol{x}^*)$, are disjoints (because their cardinality is respectively k, k and k^*), but they may intersect). And we also have $\|\nabla_s f(\boldsymbol{x}^*)\|_2^2 \leq s\|\nabla f(\boldsymbol{x}^*)\|_\infty^2$ (by definition of the ℓ_2 norm and of the ℓ_∞ norm). Similarly, we also have: $\|\nabla_{F^c} f(\boldsymbol{x}^*)\|_2^2 \leq (d-k)\|\nabla f(\boldsymbol{x}^*)\|_\infty^2$, since $|F^c| \leq d - k$.

Therefore, we obtain:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&\leq 4\varepsilon_F \mathbb{E}_{\boldsymbol{\xi}}\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 + 4\varepsilon_{F^c}\mathbb{E}_{\boldsymbol{\xi}}\|\nabla_{F^c} f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&\quad + ((4\varepsilon_F s + 2) + \varepsilon_{F^c}(d-k))\mathbb{E}_{\boldsymbol{\xi}}\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|_\infty^2 + 2\varepsilon_{\text{abs}}\mu^2 \\
&\overset{(a)}{\leq} 4\varepsilon_F \mathbb{E}_{\boldsymbol{\xi}}\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 + ((4\varepsilon_F s + 2) + \varepsilon_{F^c}(d-k))\mathbb{E}_{\boldsymbol{\xi}}\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|_\infty^2 + 2\varepsilon_{\text{abs}}\mu^2
\end{aligned}$$

Where (a) follows by observing in Proposition 1 (b) that $\varepsilon_{F^c} \leq \varepsilon_F$, and using the definition of the Euclidean norm. Let us plug the above into (24), and use the fact that, from Proposition 1 (a), since each $f_{\boldsymbol{\xi}}$ is $(L_{s'}, s' := \max(s_2, s))$-RSS, it is also $(L_{s'}, s_2)$-RSS, so for the ε_μ from Proposition 1 (a), we have, for almost any given $\boldsymbol{\xi}$: $\|\mathbb{E}_{\boldsymbol{u}}\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t))\|^2 \leq \varepsilon_\mu \mu^2$, and let us also use the fact that since each $f_{\boldsymbol{\xi}}$ is $(L_{s'}, \max(s_2, s))$-RSS , it is also $(L_{s'}, |F|)$-RSS (since $|F| \leq s$) which gives that for almost any $\boldsymbol{\xi}$: $f_{\boldsymbol{\xi}}$: $\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \leq L_{s'}^2\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2$, to finally obtain:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\boldsymbol{x}^t - \boldsymbol{x}^* - \eta\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) + \eta\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&\leq (1 + \eta^2 L_{s'}^2 + 4\varepsilon_F \eta^2 L_{s'}^2)\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 - 2\eta\langle \boldsymbol{x}^t - \boldsymbol{x}^*, \mathbb{E}_{\boldsymbol{\xi}}[\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) - \nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)]\rangle \\
&\quad + \frac{\varepsilon_\mu}{L_{s'}^2}\mu + 2\eta^2\varepsilon_{\text{abs}}\mu^2 + \eta^2((4\varepsilon_F s + 2) + \varepsilon_{F^c}(d-k))\mathbb{E}_{\boldsymbol{\xi}}\|\nabla f(\boldsymbol{x}^*)\|_\infty^2 \\
&= (1 + \eta^2 L_{s'}^2 + 4\varepsilon_F \eta^2 L_{s'}^2)\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 - 2\eta\langle \boldsymbol{x}^t - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^*)\rangle \\
&\quad + \frac{\varepsilon_\mu}{L_{s'}^2}\mu + 2\eta^2\varepsilon_{\text{abs}}\mu^2 + \eta^2((4\varepsilon_F s + 2) + \varepsilon_{F^c}(d-k))\mathbb{E}_{\boldsymbol{\xi}}\|\nabla f(\boldsymbol{x}^*)\|_\infty^2
\end{aligned}$$

Since f is (ν_s, s)-RSC, it is also $(\nu_s, |F|)$-RSC, since $|F| \leq 2k + k^* \leq s$, therefore, we have: $\langle \boldsymbol{x}^t - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^*)\rangle \geq \nu_s\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2$ (this can be proven by adding together the definition of (ν_s, s)-RSC written respectively at $\boldsymbol{x} = \boldsymbol{x}^t$,$\boldsymbol{y} = \boldsymbol{x}^*$, and at $\boldsymbol{x} = \boldsymbol{x}^*$,$\boldsymbol{y} = \boldsymbol{x}^t$). Plugging this into the above:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\boldsymbol{x}^t - \boldsymbol{x}^* - \eta\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) + \eta\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&\leq \left(1 - 2\eta\nu_s + \left(4\varepsilon_F + 1\right)L_{s'}^2\eta^2\right)\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 \\
&\quad + \frac{\varepsilon_\mu}{L_{s'}^2}\mu^2 + 2\eta^2\varepsilon_{\text{abs}}\mu^2 + \eta^2((4\varepsilon_F s + 2) + \varepsilon_{F^c}(d-k))\mathbb{E}_{\boldsymbol{\xi}}\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|_\infty^2
\end{aligned}$$

The value of η that minimizes the left term in η is equal to $\frac{\nu_s}{(4\varepsilon_F + 1)L_{s'}^2}$ (because the optimum of the quadratic function $ax^2 + bx + c$ is attained in $-\frac{b}{2a}$ and its value is $-\frac{b^2}{4a} + c$). Let us choose it, that is, we fix $\eta = \frac{\nu_s}{(4\varepsilon_F + 1)L_{s'}^2}$. Let us now define the following ρ:

$$\rho^2 = 1 - \frac{4\nu_s^2}{4(4\varepsilon_F + 1)L_{s'}^2} = 1 - \frac{\nu_s^2}{(4\varepsilon_F + 1)L_{s'}^2}$$

We therefore have:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\boldsymbol{x}^t - \boldsymbol{x}^* - \eta\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) + \eta\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|^2 \\
&\leq \rho^2\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + \frac{\varepsilon_\mu}{L_{s'}^2}\mu^2 + 2\eta^2\varepsilon_{\text{abs}}\mu^2 + \eta^2((4\varepsilon_F s + 2) + \varepsilon_{F^c}(d-k))\mathbb{E}_{\boldsymbol{\xi}}\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|_\infty^2
\end{aligned}$$

We can now use the fact that for all $(a, b) \in (\mathbb{R}_+)^2 : \sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$, as well as Jensen's inequality, to obtain:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\boldsymbol{x}^t - \boldsymbol{x}^* - \eta\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^t) + \eta\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\| \\
&\leq \rho\|\boldsymbol{x}^t - \boldsymbol{x}^*\| + \frac{\sqrt{\varepsilon_\mu}}{L_{s'}}\mu^2 + \eta\sqrt{2\varepsilon_{\text{abs}}\mu^2} + \eta\sqrt{((4\varepsilon_F s + 2) + \varepsilon_{F^c}(d-k)))\mathbb{E}_{\boldsymbol{\xi}}\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|_\infty^2}
\end{aligned}$$

We can now formulate a first decrease-rate type of result, before the hard thresholding operation, as follows, using for η the value previously defined, and with:

$$\boldsymbol{y}^t := \boldsymbol{x}^t - \eta\hat{\nabla}_F f_{\boldsymbol{\xi}}\left(\boldsymbol{x}^t\right) \tag{25}$$

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\boldsymbol{y}^t-\boldsymbol{x}^*\| &= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\left\|\boldsymbol{x}^t-\eta\hat{\nabla}_F f_{\boldsymbol{\xi}}\left(\boldsymbol{x}^t\right)-\boldsymbol{x}^*\right\| \\
&\leq \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\left\|\boldsymbol{x}^t-\boldsymbol{x}^*-\eta\hat{\nabla}_F f_{\boldsymbol{\xi}}\left(\boldsymbol{x}^t\right)+\eta\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\right\| + \eta\mathbb{E}_{\boldsymbol{\xi}}\left\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\right\| \\
&= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\left\|\boldsymbol{x}^t-\boldsymbol{x}^*-\eta\hat{\nabla}_F f_{\boldsymbol{\xi}}\left(\boldsymbol{x}^t\right)+\eta\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\right\| + \eta\mathbb{E}_{\boldsymbol{\xi}}\sqrt{\left\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\right\|^2} \\
&\leq \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\left\|\boldsymbol{x}^t-\boldsymbol{x}^*-\eta\hat{\nabla}_F f_{\boldsymbol{\xi}}\left(\boldsymbol{x}^t\right)+\eta\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\right\| + \eta\sqrt{\mathbb{E}_{\boldsymbol{\xi}}\left\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\right\|^2} \\
&\leq \rho\left\|\boldsymbol{x}^t-\boldsymbol{x}^*\right\| + \eta(\sqrt{((4\varepsilon_F s+2)+\varepsilon_{F^c}(d-k))\mathbb{E}_{\boldsymbol{\xi}}\left\|\nabla f(\boldsymbol{x}^*)\right\|_\infty^2} \\
&\quad + \sqrt{s}\sqrt{\mathbb{E}_{\boldsymbol{\xi}}\left\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\right\|_\infty^2}) + \frac{\sqrt{\varepsilon_\mu}}{L_{s'}}\mu^2 + \eta\sqrt{2\varepsilon_{\text{abs}}\mu^2} \\
&= \rho\left\|\boldsymbol{x}^t-\boldsymbol{x}^*\right\| + \eta(\sqrt{(4\varepsilon_F s+2)+\varepsilon_{F^c}(d-k)}+\sqrt{s})\sqrt{\mathbb{E}_{\boldsymbol{\xi}}\left\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\right\|_\infty^2} \\
&\quad + \frac{\sqrt{\varepsilon_\mu}}{L_{s'}}\mu + \eta\sqrt{2\varepsilon_{\text{abs}}\mu^2} \\
&\overset{(a)}{\leq} \rho\left\|\boldsymbol{x}^t-\boldsymbol{x}^*\right\| + \eta(\sqrt{(4\varepsilon_F s+2)+\varepsilon_{F^c}(d-k)}+\sqrt{s})\sigma \\
&\quad + \frac{\sqrt{\varepsilon_\mu}}{L_{s'}}\mu + \eta\sqrt{2\varepsilon_{\text{abs}}\mu^2} \\
&\leq \rho\left\|\boldsymbol{x}^t-\boldsymbol{x}^*\right\| + \eta(\sqrt{(4\varepsilon_F s+2)+\varepsilon_{F^c}(d-k)}+\sqrt{s})\sigma \\
&\quad + \frac{\sqrt{\varepsilon_\mu}}{L_{s'}}\mu + \eta\sqrt{2\varepsilon_{\text{abs}}\mu^2} \qquad (26)
\end{aligned}
$$

Where (a) follows from the σ-FGN assumption. We now consider $\boldsymbol{x}^{t+1}$, that is, the best-k-sparse approximation of $\boldsymbol{z}^t := \boldsymbol{x}^t - \eta\hat{\nabla} f_{\boldsymbol{\xi}}\left(\boldsymbol{x}^t\right)$ from the hard thresholding operation in SZOHT. We can notice that $\boldsymbol{x}_F^t = \boldsymbol{x}^t$ (because supp($\boldsymbol{x}^t$) $= F^{(t)} \subset F$), which gives $\boldsymbol{y}^t = \boldsymbol{z}_F^t$. Since $F^{(t+1)} \subset F$, the coordinates of the top k magnitude components of $\boldsymbol{z}^t$ are in F, so they are also those of the top k magnitude components of $\boldsymbol{z}_F^t = \boldsymbol{y}^t$. Therefore, $\boldsymbol{x}^{t+1}$ is also the best k-sparse approximation of $\boldsymbol{y}^t$. Therefore, using Corollary B.1, we obtain:

$$
\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^*\| \leq \gamma\|\boldsymbol{y}^t-\boldsymbol{x}^*\|
$$

with:

$$
\gamma := \sqrt{1+\left(k^*/k+\sqrt{\left(4+k^*/k\right)k^*/k}\right)/2} \qquad (27)
$$

Where $k^* = \|\boldsymbol{x}^*\|_0$. Plugging this into (26) gives:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{u}}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^*\| \leq \gamma\rho\left\|\boldsymbol{x}^t-\boldsymbol{x}^*\right\| &+ \gamma\eta(\sqrt{(4\varepsilon_F s+2)+\varepsilon_{F^c}(d-k)})+\sqrt{s})\sigma \\
&+ \gamma\frac{\sqrt{\varepsilon_\mu}}{L_{s'}}\mu + \eta\sqrt{2\varepsilon_{\text{abs}}}\mu
\end{aligned}
$$

This will allow us to obtain the following final result:

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^*\| \leq \gamma\rho\left\|\boldsymbol{x}^t-\boldsymbol{x}^*\right\| &+ \gamma\underbrace{\eta\left(\sqrt{(4\varepsilon_F s+2)+\varepsilon_{F^c}(d-k)}+\sqrt{s}\right)\sigma}_{:=a} \\
&+ \gamma\underbrace{\left(\frac{\sqrt{\varepsilon_\mu}}{L_{s'}}+\eta\sqrt{2\varepsilon_{\text{abs}}}\right)}_{:=b}\mu \qquad (28)
\end{aligned}
$$

with $\eta = \frac{\nu_s}{(4\varepsilon_F+1)L_{s'}^2}$ and $\rho^2 = 1-\frac{2\nu_s^2}{(4\varepsilon_F+1)L_{s'}^2}$. We need to have $\rho\gamma < 1$ in order to have a contraction at each step. Let us suppose that $k \geq \rho^2 k^*/(1-\rho^2)^2$: we will show that this value for k allows to verify that condition on $\rho\gamma$. That implies $\frac{k^*}{k} \leq \frac{(1-\rho^2)^2}{\rho^2}$. We then have, from the definition of γ in

(27):

$$
\begin{aligned}
\gamma^2 &\leq 1 + \left(\frac{(1-\rho^2)^2}{\rho^2} + \sqrt{\left(4 + \frac{(1-\rho^2)^2}{\rho^2}\right) \frac{(1-\rho^2)^2}{\rho^2}} \right) \frac{1}{2} \\
&= 1 + \left(\frac{(1-\rho^2)^2}{\rho^2} + \sqrt{\left(\frac{4\rho^2 + 1 + \rho^4 - 2\rho^2}{\rho^2}\right) \frac{(1-\rho^2)^2}{\rho^2}} \right) \frac{1}{2} \\
&= 1 + \left(\frac{(1-\rho^2)^2}{\rho^2} + \sqrt{\frac{(1+\rho^2)^2(1-\rho^2)^2}{\rho^4}} \right) \frac{1}{2} \\
&= 1 + \left(\frac{(1-\rho^2)^2}{\rho^2} + \frac{(1+\rho^2)(1-\rho^2)}{\rho^2} \right) \frac{1}{2} = 1 + \left(\frac{(1-\rho^2)(1-\rho^2+1+\rho^2)}{\rho^2} \right) \frac{1}{2} \\
&= 1 + \frac{(1-\rho^2)}{\rho^2} = \frac{1}{\rho^2} \qquad (29)
\end{aligned}
$$

Therefore, we indeed have $\rho\gamma \leq 1$ when choosing $k \geq \rho^2 k^*/(1-\rho^2)^2$.

Unrolling inequality (28) through time, we then have, at iteration $t+1$, and denoting by $\boldsymbol{\xi}^{t+1}$ the noise drawn at time step $t+1$ and $\boldsymbol{u}^{t+1}$ the random directions $\boldsymbol{u}_1, ..., \boldsymbol{u}_q$ chosen at time step $t+1$, from the law of total expectations:

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\| &= \mathbb{E}_{\boldsymbol{\xi}^t, \boldsymbol{u}^t, .., \boldsymbol{\xi}^1, \boldsymbol{u}^1} \mathbb{E}_{\boldsymbol{\xi}^{t+1}, \boldsymbol{u}^{t+1} | \boldsymbol{\xi}^t, \boldsymbol{u}^t, .., \boldsymbol{\xi}^1, \boldsymbol{u}^1} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\| \\
&\leq \mathbb{E}_{\boldsymbol{\xi}^t, \boldsymbol{u}^t, .., \boldsymbol{\xi}^1, \boldsymbol{u}^1} [\gamma\rho \|\boldsymbol{x}^t - \boldsymbol{x}^*\| + \gamma a \sigma + \gamma b \mu] \\
&= \gamma\rho \mathbb{E}_{\boldsymbol{\xi}^t, \boldsymbol{u}^t, .., \boldsymbol{\xi}^1, \boldsymbol{u}^1} [\|\boldsymbol{x}^t - \boldsymbol{x}^*\|] + \gamma a \sigma + \gamma b \mu \\
&\leq (\gamma\rho)^2 \mathbb{E}_{\boldsymbol{\xi}^{t-1}, \boldsymbol{u}^{t-1}, .., \boldsymbol{\xi}^1, \boldsymbol{u}^1} [\|\boldsymbol{x}^{t-1} - \boldsymbol{x}^*\|] + (\gamma\rho)^2 a \sigma \\
&\quad + \gamma a \sigma + (\gamma\rho)^2 b \mu + \gamma b \mu \\
&\leq (\gamma\rho)^{t+1} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\| + \left(\sum_{i=0}^{t} (\gamma\rho)^i \right) \gamma a \sigma + \left(\sum_{i=0}^{t} (\gamma\rho)^i \right) \gamma b \mu \\
&= (\gamma\rho)^{t+1} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\| + \frac{1 - (\gamma\rho)^t}{1 - \gamma\rho} \gamma a \sigma + \frac{1 - (\gamma\rho)^t}{1 - \gamma\rho} \gamma b \mu \\
&\leq (\gamma\rho)^{t+1} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\| + \frac{1}{1 - \gamma\rho} \gamma a \sigma + \frac{1}{1 - \gamma\rho} \gamma b \mu
\end{aligned}
$$

Where the last inequality follows from the fact that $\rho\gamma < 1$. ☐

D.2 Proof of Remark 4

Proof. We show below that, due to the complex impact of q and k on the convergence analysis in our ZO + HT (hard-thresholding) setting (compared to ZO only), q cannot be taken as small as we want here (in particular we can never take $q = 1$, which is different from classical ZO algorithms such as [23, Corollary 3]), if we want Theorem 1 to apply with $\rho\gamma < 1$. In other words, there is a necessary (but not sufficient) minimal (i.e. > 1) value for q.

A necessary condition for Theorem 1 to describe convergence of SZOHT is that $\rho\gamma < 1$. From the expressions of ρ and γ We have $\rho = \rho(q, k)$, and $\gamma = \gamma(k)$. We recall those expressions below:

$\gamma = \sqrt{1 + \left(k^*/k + \sqrt{(4 + k^*/k)\, k^*/k}\right)/2}$

$\rho^2 = 1 - \frac{\nu_s^2}{(4\varepsilon_F + 1) L_{s'}^2} = 1 - \frac{1}{(4\varepsilon_F + 1)\kappa^2}$ with $\kappa = \frac{L_{s'}}{\nu_s}$.

with: $\varepsilon_F = \frac{2d}{q(s_2+2)} \left(\frac{(s-1)(s_2-1)}{d-1} + 3 \right) + 2$, with $s = 2k + k^*$ (we consider the smallest s possible from Theorem 1)

So therefore:

$$\begin{aligned}\rho^2 &= 1 - \frac{1}{\left[\frac{8d}{q(s_2+2)}(\frac{(s-1)(s_2-1)}{d-1} + 3) + 9\right]\kappa^2}\\ &= 1 - \frac{1}{\left[\frac{8d}{q(s_2+2)}(\frac{(2k+k^*-1)(s_2-1)}{d-1} + 3) + 9\right]\kappa^2}\end{aligned}$$

Let us define $a := \frac{16d\kappa^2(s_2-1)}{q(s_2+2)(d-1)}$ and $b := \kappa^2\left[\frac{8d}{q(s_2+2)}[\frac{(s_2-1)(k^*-1)}{d-1} + 3] + 9\right]$

We then have:

$$\rho^2 = 1 - \frac{1}{ak+b}$$

To ensure convergence, we need to have $\rho\gamma < 1$, therefore (following the same derivation as in (29)) a necessary condition that we need to verify is $k \geq \rho^2 k^*/(1-\rho^2)^2$.

Which means we need:

$$k \geq \frac{\left(1 - \frac{1}{ak+b}\right)k^*}{\left(\frac{1}{ak+b}\right)^2} \tag{30}$$

$$k \geq \left[(ak+b)^2 - (ak+b)\right]k^* \tag{31}$$

$$k \geq k^*\left[a^2k^2 + 2abk + b^2 - ak - b\right] \tag{32}$$

$$0 \geq k^*a^2k^2 + \left(2ab - \frac{1}{k^*} - a\right)k^*k + \left(b^2 - b\right)k^* \tag{33}$$

If we want that there exist a k such that this is true, we need (since $k^* \geq 0$):

$$\Delta \geq 0$$

with:

$$\begin{aligned}\Delta &:= k^{*2}(2ab - \frac{1}{k^*} - a)^2 - 4k^{*^2}a^2\left(b^2 - b\right)\\ &= k^{*^2}\left(4a^2b^2 + \left(\frac{1}{k^*} + a\right)^2 - 4ab\left(\frac{1}{k^*} + a\right)\right) - 4k^{*2}a^2\left(b^2 - b\right)\\ &= k^{*2}\left[4a^2b^2 + \frac{1}{k^{*^2}} + a^2 + \frac{2a}{k^*} - \frac{4ab}{k^*} - 4a^2b - 4a^2b^2 + 4a^2b\right]\\ &= 1 + a^2k^{*2} + 2ak^* - 4abk^*\end{aligned}$$

$$\Delta \geq 0 \Rightarrow 1 + a^2k^{*2} + 2ak^* \geq 4abk^* \tag{34}$$

Let us express a and b in terms of q, as:

$$a = \frac{A}{q} \quad \text{with} \quad A = \frac{16d\kappa^2\left(s_2 - 1\right)}{\left(s_2 + 2\right)\left(d - 1\right)} \tag{35}$$

$$b = \frac{B}{q} + C \quad \text{with} \quad B = \kappa^2\left[\frac{8d}{\left(s_2 + 2\right)}\left(\frac{\left(s_2 - 1\right)\left(k^* - 1\right)}{d - 1} + 3\right)\right] \tag{36}$$

$$\text{and with } C = 9\kappa^2 \tag{37}$$

So plugging in (34), what we need is:

$$\begin{aligned}1 + \frac{A^2}{q^2}k^{*^2} + 2\frac{A}{q}k^* &\geq 4\frac{A}{q}\left(\frac{B}{q} + C\right)k^*\\ q^2 + A^2k^{*2} + 2Ak^*q &\geq 4ABk^* + 4CAqk^*\\ q^2 + q\left(2Ak^* - 4CAk^*\right) + A^2k^{*2} - 4ABk^* &\geq 0\end{aligned}$$

To ensure that, we need to compute Δ', defined as:

$$\begin{aligned}
\Delta' &:= \left(2Ak^* - 4CAk^*\right)^2 - 4\left(A^2k^{*2} - 4ABk^*\right) \\
&= 4A^2k^{*2} + 16C^2A^2k^{*2} - 16CA^2k^{*2} - 4A^2k^{*2} + 16ABk^* \\
&= 16CA^2k^{*2}(C-1) + 16ABk^* = 16Ak^*\left[k^*C(C-1)A + B\right]
\end{aligned}$$

We now have:

$$C = 9\kappa^2 \Rightarrow C \geq 1 \Rightarrow \Delta' \geq 0$$

Therefore, there is a minimal value for q, and it is:

$$q \geq q_{\text{min}}$$

With:

$$\begin{aligned}
q_{\text{min}} &= \frac{-\left(2Ak^* - 4CAk^*\right) + \sqrt{16CA^2k^{*2}(C-1) + 16ABk^*}}{2} && (38) \\
&= \frac{2Ak^*\left(2C-1\right) + \sqrt{16A^2k^{*2}\left[C(C-1) + \frac{B}{Ak^*}\right]}}{2} && (39)
\end{aligned}$$

Case $s_2 > 1$: Assuming $s_2 > 1$ gives $A > 0$, and since $A = \frac{16d\kappa^2(s_2-1)}{(s_2+2)(d-1)}$ and $B = \frac{8\kappa^2 d}{s_2+2}(\frac{(s_2-1)(k^*-1)}{d-1} + 3)$
This gives: $\frac{B}{Ak^*} = \frac{1}{2} - \frac{1}{2k^*} + \frac{3}{2}\frac{d-1}{k^*(s_2-1)}$
Therefore: $q_{\text{min}} = Ak^*\left[2C - 1 + 2\sqrt{C(C-1) + \frac{1}{2} - \frac{1}{2k^*} + \frac{3}{2}\frac{d-1}{k^*(s_2-1)}}\right]$
with $C = 9\kappa^2$, which reads:

$$q_{\text{min}} = \frac{16d(s_2-1)k^*\kappa^2}{\left(s_2+2\right)\left(d-1\right)}\left[18\kappa^2 - 1 + 2\sqrt{9\kappa^2(9\kappa^2-1) + \frac{1}{2} - \frac{1}{2k^*} + \frac{3}{2}\frac{d-1}{k^*(s_2-1)}}\right]$$

Case $s_2 = 1$: In the case $s_2 = 1$, we have $A = 0$, so therefore, from (38), $q_{\text{min}} = 0$, so the necessary condition on q as above so that there exist k such that: $k \geq \rho^2k^*/(1-\rho^2)^2$ does not apply here. We may therefore think that it may be possible to take $q = 1$ in that case. However, there is another condition on k that should also be enforced, which is that $k \leq d$ (since we cannot keep more components than d). And in that $s_2 = 1$ case, we have $a = 0$, and $b = \kappa^2[8\frac{d}{q} + 9]$ (from (35) and (36)). Now, enforcing the condition $k \geq k^*[(ak+b)^2 - (ak+b)] = k^*b(b-1)$ leads to the following chain of implications (i.e. each downstream assertion is a necessary condition for the upstream assertion):

$$\begin{aligned}
\frac{k}{k^*} \geq b(b-1) \quad \text{and} \quad k \leq d \implies \frac{d}{k^*} \geq (b-1)^2 \implies \sqrt{\frac{d}{k^*}} + 1 \geq b \implies \sqrt{\frac{d}{k^*}} + 1 \geq \frac{B}{q} + C& \\
\implies \sqrt{\frac{d}{k^*}} + 1 - C \geq \frac{B}{q}& \\
\implies q \geq \frac{B}{\sqrt{\frac{d}{k^*}} + 1 - C} \quad \text{and} \quad C - \sqrt{\frac{d}{k^*}} + 1 > 0& \\
\implies q \geq \frac{B}{\sqrt{\frac{d}{k^*}} + 1} \implies q \geq \frac{8\kappa^2 d}{\sqrt{\frac{d}{k^*}} + 1}& \qquad (40)
\end{aligned}$$

Where the last inequality follows from the expression of B in (36) when $s_2 = 1$.

So the right hand side in (40) is also a minimal necessary value for q in this case, though for a different reason than in the case $s_2 > 1$.

□

D.3 Proof of Corollary 1

Proof. We first restrict the result of Theorem 1 to a particular q. By inspection of Proposition 1 (b), we choose q such that the part of ε_F that depends on q becomes 1: we believe this will allow to better understand the dependence between variables in our convergence rate result, although other choices of q are possible. Therefore, we choose:

$$q' := \frac{2d}{s_2+2}\left(\frac{(s-1)(s_2-1)}{d-1}+3\right) \tag{41}$$

so that we obtain: $\varepsilon'_F := 1 + 2 = 3$ (from Proposition 1 (b)), which also implies :

$$\eta' := \frac{\nu_s}{(4\varepsilon'_F+1)L^2_{s'}} = \frac{\nu_s}{13L^2_{s'}}$$

and:

$$\rho'^2 := 1 - \frac{2\nu_s^2}{(4\varepsilon'_F+1)L^2_{s'}} = 1 - \frac{2\nu_s^2}{13L^2_{s'}} \tag{42}$$

Now, regarding the value of q, we also note that any value of random directions $q'' \geq q'$ can be taken too, since the bound in Proposition 1 (b) would then still be verified for ε'_F (that is, we would still have $\mathbb{E}\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x})\|^2 \leq \varepsilon'_F\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x})\|^2 + \varepsilon'_{F^c}\|\nabla_{F^c} f_{\boldsymbol{\xi}}(\boldsymbol{x})\|^2 + \varepsilon_{abs}\mu^2$) (with ε'_{F^c} the value of ε_{F^c} for $q = q'$).
Therefore, we will choose a value q'' so that our result is simpler. First, notice that $s \leq d \implies 1 - \frac{1}{s} \leq 1 - \frac{1}{d} \implies \frac{s-1}{s} \leq \frac{d-1}{d} \implies \frac{s-1}{d-1} \leq \frac{s}{d}$. Therefore, if we take $q \geq 2s + 6\frac{d}{s_2}$, we will also have $q \geq \frac{2d}{s_2+2}\left(\frac{(s-1)(s_2-1)}{d-1}+3\right) = q'$.

Let us now impose a lower bound on k that is slightly (twice) bigger than the lower bound from Theorem 1. As will become clear below, this allows us to have a $\rho\gamma$ enough bounded away from 1, which guarantees a reasonable constant in the $\mathcal{O}$ notation for the query complexity (see the end of the proof). Let us therefore take:

$$k \geq 2k^* \frac{\rho^2}{(1-\rho^2)^2} \tag{43}$$

and plug the value of ρ above into the expression:

$$k \geq 2k^* \frac{\rho'^2}{(1-\rho'^2)^2} \iff k \geq 2k^* \frac{1 - \frac{2\nu_s^2}{13L^2_{s'}}}{(\frac{2\nu_s^2}{13L^2_{s'}})^2} \iff k \geq 2k^* \left(\left(\frac{13L^2_{s'}}{2\nu_s^2}\right)^2 - \frac{13L^2_{s'}}{2\nu_s^2}\right)$$
$$\iff k \geq 2k^*(\frac{13}{2}\kappa^2)(\frac{13}{2}\kappa^2 - 1)$$

With κ denoting $\frac{L_{s'}}{\nu_s}$. Therefore, if we take:

$$k \geq (86\kappa^4 - 12\kappa^2)k^*$$

we will indeed verify the formula above $k \geq 2k^*(\frac{13}{2}\kappa^2)(\frac{13}{2}\kappa^2 - 1)$.
We now turn to describing the query complexity of the algorithm: To ensure that $(\gamma\rho)^t\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\| \leq \varepsilon$, we need:

$$t \geq \frac{1}{\log\frac{1}{\gamma\rho}} \log(\frac{1}{\varepsilon}) \log(\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|) \tag{44}$$

with $\gamma\rho$ belonging to the interval $(0, 1)$. Let us compute more precisely an upper bound to $\rho\gamma$ in this case, to show that it is reasonably enough bounded away from 1: Taking k as described in (43), and plugging that value into the expression of γ from Theorem 1, we obtain:

$$\begin{aligned}
\gamma^2 &= 1 + \left(\frac{(1-\rho^2)^2}{2\rho^2} + \sqrt{\left(4 + \frac{(1-\rho^2)^2}{2\rho^2}\right)\frac{(1-\rho^2)^2}{2\rho^2}}\right)/2 \\
&\leq 1 + \frac{1}{\sqrt{2}}\left(\frac{(1-\rho^2)^2}{\rho^2} + \sqrt{\left(4 + \frac{(1-\rho^2)^2}{\rho^2}\right)\frac{(1-\rho^2)^2}{\rho^2}}\right)/2 \\
&\overset{(a)}{=} 1 + \frac{1}{\sqrt{2}}\frac{1-\rho^2}{\rho^2}
\end{aligned}$$

Where the simplification in (a) above follows similarly to (29). Therefore, in that case, we have:

$$\begin{aligned}
\rho^2\gamma^2 &\leq \rho^2 + \frac{1}{\sqrt{2}}(1-\rho^2) = \frac{1}{\sqrt{2}} + \rho^2(1 - \frac{1}{\sqrt{2}}) \\
&= \frac{1}{\sqrt{2}} + (1 - \frac{2}{13\kappa^2})(1 - \frac{1}{\sqrt{2}}) = 1 - \frac{2(1-\frac{1}{\sqrt{2}})}{13\kappa^2} \overset{(a)}{\leq} 1 - \frac{1}{26\kappa^2}
\end{aligned}$$

Where (a) follows because $(1 - \frac{1}{\sqrt{2}}) \approx 0.29 \geq 1/4$ Therefore:

$$\frac{1}{(\rho\gamma)^2} \geq \frac{1}{1 - \frac{1}{26\kappa^2}} \tag{45}$$

Given that $\log(\frac{1}{1-x}) \geq x$ for all $x \in [0, 1)$, we have:

$$\log\left(\frac{1}{(\rho\gamma)^2}\right) \geq \frac{1}{26\kappa^2}$$

Therefore:

$$\frac{1}{\log(\frac{1}{\rho\gamma})} = \frac{2}{\log(\frac{1}{(\rho\gamma)^2})} \leq 52\kappa^2$$

Therefore, plugging this into (44), we obtain that with $t \geq 52\kappa^2\log(\frac{1}{\varepsilon})\log(\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|) = \mathcal{O}(\kappa^2\log(\frac{1}{\varepsilon}))$ iterations, we can get $(\gamma\rho)^t\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \varepsilon$.

To obtain the query complexity (QC), we therefore just need to multiply the number of iterations by the number of queries per iteration $q = 2s + 6\frac{d}{s_2}$: to ensure $(\gamma\rho)^t\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \varepsilon$, we need to query the zeroth-order oracle at least the following number of times: $(2s + 6\frac{d}{s_2})52\kappa^2\log(\frac{1}{\varepsilon})\log(\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|) = \mathcal{O}((k + \frac{d}{s_2})\kappa\log(\frac{1}{\varepsilon}))$, since $s = 2k + k^*$.

D.4 Proof of Corollary 2

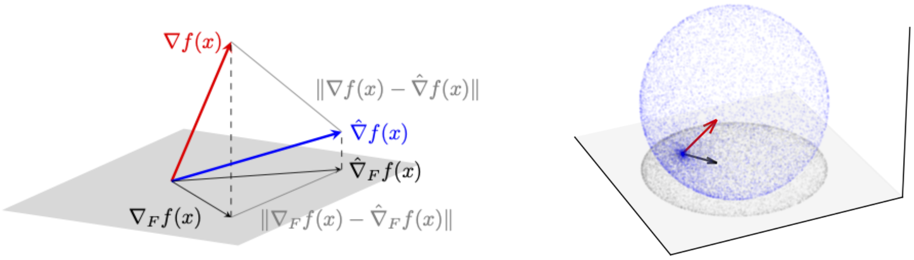
Almost all $f_{\boldsymbol{\xi}}$ are L-smooth, which is equivalent to saying that they are (L, d)-RSS. So we can directly plug $s_2 = d$ in equation (41), which gives a necessary value for q of:

$$q = \frac{2d}{d+2}(s+2) \tag{46}$$

Since any value of q larger than the one in (46) is valid, we choose $q \geq 2(s+2)(\geq \frac{2d}{d+2}(s+2))$ for simplicity. The query complexity is obtained similarly as in the proof of Corollary 1 above, with that new value for q (the number of iterations needed is unchanged from the proof of Corollary 1), only the query complexity q per iteration changes), which means we need to query the zeroth-order oracle the following number of times: $2(s+2)52\kappa^2\log(\frac{1}{\varepsilon})\log(\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|) = \mathcal{O}(k\kappa\log(\frac{1}{\varepsilon}))$ $\square$

E Projection of the gradient estimator onto a sparse support

Below we plot the true gradient $\nabla f(x)$ and its estimator $\hat{\nabla} f(x)$ (for $q = 1$), as well as their respective projections $\nabla_F f(x)$ and $\hat{\nabla}_F f(x)$, with $F = \{0, 1\}$ (i.e. F is the hyperplane $z = 0$), for n_{dir} random directions. In Figure 5(b), due to the large number of random directions, we plot them as points not vectors. For simplicity, the figure is plotted for $\mu \to 0$, and $s_2 = d$. We can see that even though gradient estimates $\hat{\nabla} f(x)$ are poor estimates of $\nabla f(x)$, $\hat{\nabla}_F f(x)$ is a better estimate of $\nabla_F f(x)$.


$\nabla f(x)$
$\|\nabla f(x) - \hat{\nabla} f(x)\|$
$\hat{\nabla} f(x)$
$\hat{\nabla}_F f(x)$
$\nabla_F f(x)$
$\|\nabla_F f(x) - \hat{\nabla}_F f(x)\|$

(a) $n_{\text{dir}} = 1$ (b) $n_{\text{dir}} = 10^6$

Figure 5: $\nabla f(x)$ and $\hat{\nabla} f(x)$ and their projections $\nabla_F f(x)$ and $\hat{\nabla}_F f(x)$ onto F

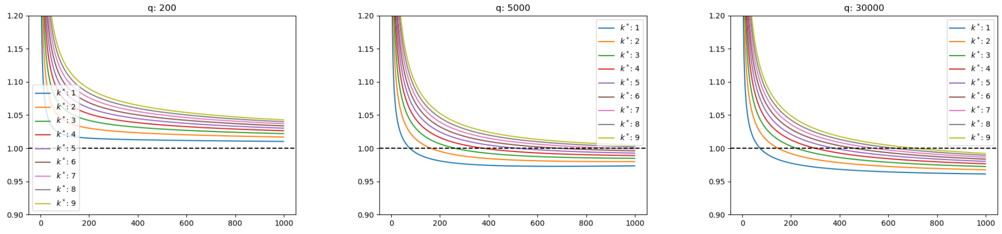
Remark 5. *An interesting fact that can be observed in Figure 5(b) above is that when $\mu \to 0$ and $s_2 = d$, the ZO gradient estimates belong to a sphere. This comes from the fact that, in that case, the ZO estimate using the random direction $\boldsymbol{u}$ is actually a directional derivative (scaled by d): $\hat{\nabla} f(\boldsymbol{x}) = d\langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle \boldsymbol{u}$, for which we have :*

$$\begin{aligned}
\|\hat{\nabla} f(\boldsymbol{x}) - \frac{d}{2}\nabla f(\boldsymbol{x})\|^2 &= d^2(\langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle)^2 \langle \boldsymbol{u}, \boldsymbol{u}\rangle + \frac{d^2}{4}\|\nabla f(\boldsymbol{x})\|^2 \\
&\quad - d^2 \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle \langle \boldsymbol{u}, \nabla f(\boldsymbol{x})\rangle \\
&= \frac{d^2}{4}\|\nabla f(\boldsymbol{x})\|^2
\end{aligned}$$

(since $\|\boldsymbol{u}\| = 1$). That is, gradient estimates belong to a sphere of center $\frac{d}{2}\nabla f(\boldsymbol{x})$ and radius $\frac{d}{2}\|\nabla f(\boldsymbol{x})\|$. However, the distribution of $\hat{\nabla} f(\boldsymbol{x})$ is not uniform on that sphere: it is more concentrated around $\mathbf{0}$ as we can observe in Figure 5(b).

F Value of $\rho\gamma$ depending on q and k^*

In this section, we further illustrate the importance on the value of q as discussed in Remark 4, by showing in Figure 6 that if q is too small, then there does not exist any k that verifies the condition $k \geq \frac{k^* \rho^2}{(1-\rho^2)^2}$, no matter how small is k^* (i.e., even if $k^* = 1$). However, if q is large enough, then there exist some k^* such that this condition is true. To generate the curves below, we simply use the formulas for $\gamma = \gamma(k, k^*)$ and $\rho = \rho(s, q)$ with $s = 2k + k^*$ from Theorem 1, and with $d = 30000$ and $s_2 = d$.



(a) $q = 200$ (b) $q = 5000$ (c) $q = 30000$

Figure 6: $\rho\gamma$ (y axis) as a function of k (x axis) for several values of q and k^*

G Dimension independence/weak-dependence

In this section, we show the dependence of SZOHT on the dimension. To that end, we consider minimizing the following synthetic problem:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{s.t.} \quad \|\boldsymbol{x}\|_0 \leq k$$

with $k = 500$, and f chosen as: $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$, with $\boldsymbol{y}_i = 0$ if $i < d - k^*$ and $\boldsymbol{y}_i = \frac{1}{(k^* - (d-i))}$ if $i > d - k^*$ with $k^* = 5$. In other words, the k^* last components of $\boldsymbol{y}$ are regularly spaced from $1/k^*$ to 1: in a way, this simulates the recovery of a k^*-sparse vector $\boldsymbol{y}$ by observing only the squared deviation of some queries $\boldsymbol{x}$. In that case, we can easily check that f verifies the following properties:

- f is L-smooth with $L = 1$, as well as $(L_{s'}, s')$-RSS for any s' such that $1 \leq s' \leq d$, with $L_{s'} = 1$, and (ν_s, s)-RSC with $s = 2k + k^*$ and $\nu_s = 1$ (so $\kappa = \frac{L}{\nu_s} = \frac{L_{s'}}{\nu_s} = 1$)
- $\boldsymbol{y} = \boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad s.t. \quad \|\boldsymbol{x}\|_0 \leq k^*$
- $f(\boldsymbol{y}) = f(\boldsymbol{x}^*) = 0$
- $\nabla f(\boldsymbol{y}) = \mathbf{0}$ so f is σ-FGN with $\sigma = 0$

We also note that the above setting of k and k^* verifies $k \geq (86\kappa^4 - 12\kappa^2)k^*$ (since $\kappa = 1$). Finally, we initialize $\boldsymbol{x}^0$ such that ${\boldsymbol{x}^0}_i = 1/d$ if $d - k^* \geq i$ and 0 otherwise. We choose this initialization and not $\boldsymbol{x}^0 = \mathbf{0}$, just to ensure that $\nabla f(\boldsymbol{x}^0)_i \neq 0$ for any i: this way the optimization is really done over all d variables, not just the k^* last ones. In addition, this initialization ensures that $\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|$ is constant no matter the d, which makes the convergence curves comparable. We consider several settings of s_2 to showcase the dependence on the dimension below.

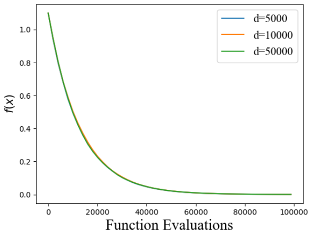
Dimension Independence

- $s_2 = d$: As from Corollary 2, we take $q = 2(s + 2)$ with $s = 2k + k^*$ (i.e. $q = 2014$). We choose $\mu = 1e - 8$, to have the smallest possible system error due to zeroth-order approximations. As we can see in Figure 7, all curves are superimposed, which shows that the query complexity is indeed dimension independent, as described by Corollary 2
- $s_2 = \mathcal{O}(\frac{d}{k})$ (We choose $s_2 = \lfloor \frac{d}{k} \rfloor$): As from Corollary 1, we take $q = 2s + 6\frac{d}{s_2}$ with $s = 2k + k^*$. In that case, from Corollary 1, the query complexity will still be $\mathcal{O}(k)$ (i.e. dimension independent), as a sum of two $\mathcal{O}(k)$ terms, although larger than in the case $s_2 = d$ above (since the constant from the $\mathcal{O}$ notation in Corollary 1 will be larger here). We can observe that this is indeed the case in Figure 8.

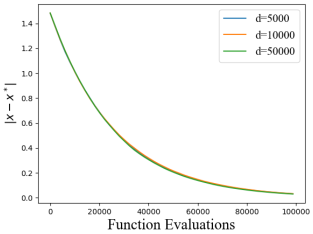
Dimension weak-dependence We now turn to the case where s_2 is fixed. We choose q as in Corollary 1 ($q = 2s + 6\frac{d}{s_2}$ with $s = 2k + k^*$): the query now depends on d in that case, as predicted by Corollary 1, which can indeed be observed in Figure 9.

H Additional results on adversarial attacks

In this section, we provide additional results for the adversarial attacks problem in 5.3, in Figure 11. The parameters we used for SZOHT to generate that table are the same as in 5.3, except for MNIST, for which we choose $k = 20$, $q = 10$, and $s_2 = 10$, and for ImageNet, for which we choose $k = 100000, s_2 = 20000$ and $q = 100$. As we can see, SZOHT allows to obtain sparse attacks, contrary to the other algorithms, and with a smaller ℓ_2 distance and a larger success rate, using less iterations: this shows that SZOHT allows to enforce sparsity, and efficiently exploits that sparsity in order to have a lower query complexity than vanilla sparsity constrained ZO algorithms.

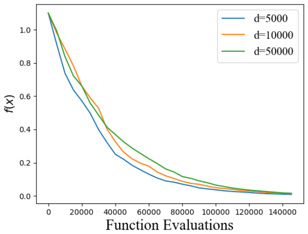


(a) $f(\boldsymbol{x})$

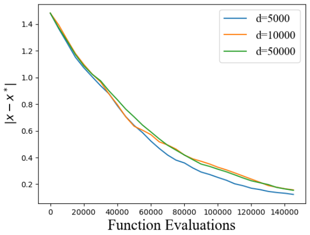


(b) $\|\boldsymbol{x} - \boldsymbol{x}^*\|$

Figure 7: $s_2 = d$

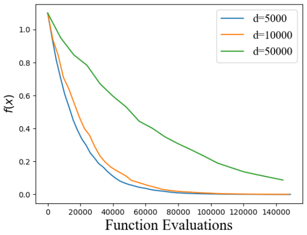


(a) $f(\boldsymbol{x})$

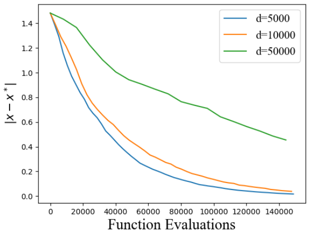


(b) $\|\boldsymbol{x} - \boldsymbol{x}^*\|$

Figure 8: $s_2 = \lfloor \frac{d}{k} \rfloor$



(a) $f(\boldsymbol{x})$



(b) $\|\boldsymbol{x} - \boldsymbol{x}^*\|$

Figure 9: $s_2 = 50$

Figure 10: Dependence on the dimensionality of the query complexity

Method	ASR	ℓ_0 dist.	ℓ_2 dist.	Iter
RSPGF	78%	100%	10.9	67
ZORO	75%	100%	15.1	550
ZSCG	79%	100%	10.3	252
SZOHT	79%	2.5%	8.5	36

(a) MNIST

Method	ASR	ℓ_0 dist.	ℓ_2 dist.	Iter
RSPGF	83%	100%	4.1	326
ZORO	86%	100%	62.9	592
ZSCG	86%	100%	8.4	126
SZOHT	91%	1.9%	2.6	26

(b) CIFAR

Method	ASR	ℓ_0 dist.	ℓ_2 dist.	Iter.
RSPGF	91%	100%	19.9	137
ZORO	90%	100%	111.9	674
ZSCG	76%	100%	111.3	277
SZOHT	95%	37.3%	10.5	61

(c) ImageNet

Figure 11: Summary of results on adversarial attacks