OPTIMIZATION OVER SPARSE SUPPORT-PRESERVING SETS VIA TWO-STEP PROJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In sparse optimization, enforcing hard constraints using the ℓ_0 pseudo-norm offers advantages like controlled sparsity compared to convex relaxations. However, many real-world applications (e.g., portfolio optimization) demand not only sparsity constraints but also some extra constraint (such as limit of budget). While prior algorithms have been developed to address this complex scenario with mixed combinatorial and convex constraints, they typically require the closed form projection onto the mixed constraints which might not exist, and/or only provide local guarantees of convergence which is different from the global guarantees commonly sought in sparse optimization. To fill this gap, in this paper, we study the problem of sparse optimization with extra *support-preserving* constraints commonly encountered in the literature. We present a new variant of iterative hard-thresholding algorithm equipped with a two-step consecutive projection operator customized for these mixed constraints, serving as a simple alternative to the Euclidean projection onto the mixed constraint. By introducing a novel trade-off between sparsity relaxation and sub-optimality, we provide global guarantees in objective value for the output of our algorithm, in the deterministic, stochastic, and zeroth-order settings, under the conventional restricted strong-convexity/smoothness assumptions. As a fundamental contribution in proof techniques, we develop a novel extension of the classic three-point lemma to the considered two-step non-convex projection operator, which allows us to analyze the convergence in objective value in an elegant way that has not been possible with existing techniques. Finally, we illustrate the applicability of our method on several sparse learning tasks.

1 INTRODUCTION

In sparse optimization, directly enforcing sparsity with the ℓ_0 pseudo-norm has several advantages over its convex relaxation counterpart. In compressive sensing for instance (Foucart & Rauhut, 2013), one may seek to recover an unknown vector, which sparsity level is known to be at most k. Similarly, in portfolio optimization, due to transaction costs, one may seek to ensure hard constraints on the maximum number of assets invested in (Brodie et al., 2009; DeMiguel et al., 2009). However, in several use cases, one may also seek to enforce additional constraints, such as, for instance, a budget constraint in the case of portfolio optimization, which can be enforced through an extra ℓ_1 constraint, as in Takeda et al. (2013). As another example, in sparse non-negative matrix factorization, when estimating the hidden components, one seeks to enforce at the same time a norm constraint and a sparsity constraint Hoyer (2002). The problem of ℓ_0 empirical risk minimization (ERM) with additional constraints can be formulated as follows, where R is an empirical risk function, $\Gamma \subseteq \mathbb{R}^d$ denotes a convex constraint set, and $\|\cdot\|_0$ denotes the ℓ_0 pseudo-norm (number of non-zero components of a vector):

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} R(\boldsymbol{w}), \quad \text{s.t. } \|\boldsymbol{w}\|_0 \leq k \text{ and } \boldsymbol{w} \in \Gamma. \tag{1}$$

In the literature, several algorithms have been developed to address such a problem with mixed constraints, but they typically require the existence of a closed form for the projection onto the mixed constraint, and/or their convergence guarantees are only local, which makes it difficult to estimate the sub-optimality of the output of the algorithm. More precisely, on one hand, some works provide convergence analyses for variants of a (non-convex) projected gradient descent, explicitly for mixed sparse constraints (Metel, 2023; Pan et al., 2017; Lu, 2015; Beck & Hallak, 2016), or for general proximal terms (which encompasses our mixed constraints) (Frankel et al. (2014); Xu et al. (2019b); Attouch et al. (2013); De Marchi & Themelis (2022); Yang & Yu (2020); Gu et al. (2018); Yang & Li (2023); Bolte et al. (2014); Boţ et al. (2016); Xu et al. (2019a); Li & Lin (2015)), but such analyses are only local. On the other hand, several existing works on Iterative Hard Thresholding (IHT) provide global guarantees on sub-optimality gap (Jain et al., 2014; Nguyen et al., 2017; Li et al., 2016; Shen & Li, 2017; de Vazelhes et al., 2022), but they do not apply to the mixed constraint case we consider. In between the two approaches, one can also find Barber & Ha (2018) andLiu & Foygel Barber (2020) which give global guarantees for general non-convex constraints or projection operators, but which do not provide explicit convergence guarantees for the particular mixed constraint setting that we consider: their rates depend on some constants (the relative concavity or the local concavity constant) for which, up to our knowledge, an explicit form is still unknown for the mixed constraints we consider. We present a more detailed review of related works in Appendix B, and an overview of them in Table 1. To fill this gap, we focus on solving problem 1 in the case where Γ belongs to a general family of *support-preserving* sets, which encompasses many usual sets encountered in the literature. As will be described in more detail in Section 2, such sets are convex sets for which the projection of a k-sparse vector onto them gets its support preserved, such as for instance ℓ_p norm balls (for $p \geq 1$), or a broader family of *sign–free* convex sets described for instance in Lu (2015); Beck & Hallak (2016).

Adapted to the properties of such constraints, we propose a new variant of IHT, with a two-step projection operator, which, as a first step, identifies the set S of coordinates of the top k components of a given vector and sets the other components to 0 (hard-thresholding), and as a second step projects the resulting vector onto Γ. This two-step projection can offer a simpler alternative to Euclidean projection onto the mixed constraint in the cases where there is a closed form for the latter projection, and handle the cases where there is not. We then provide global sub-optimality guarantees without system error for the objective value, for such an algorithm as well as its stochastic and zeroth-order variants, under the restricted strong-convexity (RSC) and restricted smoothness (RSS) assumptions, in Theorems 1, 2, and 3. Key to our analysis is a novel extension of the three-point lemma to such non-convex setting with mixed constraints, which also allows, as a byproduct, to simplify existing proofs of convergence in objective value for IHT and its variants. In the zeroth-order case, such technique also allows to obtain, up to our knowledge, the first convergence in risk result without system error for a zeroth-order hard-thresholding algorithm. Additionally, our results highlight a compromise between sparsity and sub-optimality gap specific to the additional constraints setting: through a free parameter ρ, one can obtain smaller upper bounds in terms of risk but at the cost of relaxing further the sparsity level of the iterates, or, alternatively, enforce sparser iterates but at the cost of a larger upper bound on the risk.

Finally, we illustrate the applicability of our method on several sparse learning tasks, namely index tracking for portfolio selection, multiclass logistic regression, and adversarial attacks.

Contributions: We summarize the main contributions of our paper as follows:

1. We present a variant of IHT to solve hard sparsity problems with additional support-preserving constraints, using a novel two-step projection operator.

2. We describe a novel extension of the three-point lemma to such constraint which allows to simplify existing proofs for IHT and to provide global convergence guarantees in objective value without system error for the algorithm above, in the RSC/RSS setting, highlighting a novel trade-off between sparsity of iterates and sub-optimality gap in such mixed constraints setting.

3. We extend the above algorithm to the stochastic and zeroth-order optimization settings, obtaining similar global convergence guarantees in objective value (without system error) for such mixed constraints setting. In the zeroth-order case, this also provides, up to our knowledge, the first convergence result in objective value without system error for a zeroth-order hard-thresholding algorithm (with or without extra constraints).

Table 1: Comparison of results for Iterative Hard Thresholding with/without additional constraints. [1] $\mathcal{S}$: symmetric convex sets being sign-free or non-negative (Lu, 2015), $\mathcal{A}$: sets verifying Assumption 3. [2] If a paper reports both $\|\boldsymbol{w} - \bar{\boldsymbol{w}}\|$ and $R(\boldsymbol{w}) - R(\bar{\boldsymbol{w}})$, we report only the latter. $\hat{T}$: time index of the $\boldsymbol{w}$ returned by the method (e.g. $\hat{T} = \arg\min_{t \in [T]} R(\boldsymbol{w}_t)$). $\bar{\boldsymbol{w}}$: $\bar{k}$-sparse vector in Γ. Δ: System error (term which depends on the gradient at optimality (e.g. $\mathbb{E}_i \|\nabla R_i(\bar{\boldsymbol{w}})\|$, (see corresponding references))). [4]: $\kappa_s = \frac{L_s}{\nu_s}$ and $\kappa_{s'} = \frac{L_{s'}}{\nu_s}$ (cf. corresponding refs. for defs. of s and s'). [3] SM: Lipschitz-smooth, D: Deterministic. S: Stochastic, Z: Zeroth-Order, L: Lipschitz continuous.

Reference	Γ[1]	Convergence[2]	$\boldsymbol{k}$	Setting[3]
Jain et al. (2014)	$\mathbb{R}^d$	$R(\boldsymbol{w}_{\hat{T}}) \leq R(\bar{\boldsymbol{w}}) + \varepsilon$	$\Omega(\kappa_s^2 \bar{k})$	D, RSS, RSC
Nguyen et al. (2017)	$\mathbb{R}^d$	$\mathbb{E}\|\boldsymbol{w}_{\hat{T}} - \bar{\boldsymbol{w}}\| \leq \varepsilon + \mathcal{O}(\Delta)$	$\Omega(\kappa_s^2 \bar{k})$	S, RSS, RSC
Li et al. (2016)	$\mathbb{R}^d$	$\mathbb{E}R(\boldsymbol{w}_{\hat{T}}) \leq R(\bar{\boldsymbol{w}}) + \varepsilon + \mathcal{O}(\Delta)$	$\Omega(\kappa_s^2 \bar{k})$	S, RSS, RSC
Zhou et al. (2018)	$\mathbb{R}^d$	$\mathbb{E}R(\boldsymbol{w}_{\hat{T}}) \leq R(\bar{\boldsymbol{w}}) + \varepsilon$	$\Omega(\kappa_s^2 \bar{k})$	S, RSS, RSC
de Vazelhes et al. (2022)	$\mathbb{R}^d$	$\mathbb{E}\|\boldsymbol{w}_{\hat{T}} - \bar{\boldsymbol{w}}\| \leq \varepsilon + \mathcal{O}(\Delta) + \mathcal{O}(\mu)$	$\Omega(\kappa_{s'}^4 \bar{k})$	S, Z, RSS', RSC
Lu (2015), Beck & Hallak (2016)	$\Gamma \in \mathcal{S}$	local convergence	-	D, SM
Metel (2023)	ℓ_∞ ball around 0	local convergence	-	S, Z, L
IHT-TSP (Thm. 1)	$\Gamma \in \mathcal{A} \supset \mathcal{S}$	$R(\boldsymbol{w}_{\hat{T}}) \leq (1+2\rho) R(\bar{\boldsymbol{w}}) + \varepsilon$	$\Omega\left(\frac{\kappa_s^2 \bar{k}}{\rho^2}\right)$	D, RSS, RSC
HSG-HT-TSP (Thm. 2)	$\Gamma \in \mathcal{A} \supset \mathcal{S}$	$\mathbb{E}R(\boldsymbol{w}_{\hat{T}}) \leq (1+2\rho) R(\bar{\boldsymbol{w}}) + \varepsilon$	$\Omega\left(\frac{\kappa_s^2 \bar{k}}{\rho^2}\right)$	S, RSS, RSC
HZO-HT-TSP (Thm. 3)	$\Gamma \in \mathcal{A} \supset \mathcal{S}$	$\mathbb{E}R(\boldsymbol{w}_{\hat{T}}) \leq (1+2\rho) R(\bar{\boldsymbol{w}}) + \varepsilon + \mathcal{O}(\mu)$	$\Omega\left(\frac{\kappa_{s'}^2 \bar{k}}{\rho^2}\right)$	Z, RSS', RSC
HZO-HT (Thm. 6 in App. E.3.2)	$\mathbb{R}^d$	$\mathbb{E}[R(\boldsymbol{w}_{\hat{T}}) - R(\bar{\boldsymbol{w}})] \leq \varepsilon + \mathcal{O}(\mu)$	$\Omega(\kappa_{s'}^2 \bar{k})$	Z, RSS', RSC

2 Preliminaries

Throughout this paper, we adopt the following notations. For any $\boldsymbol{w} \in \mathbb{R}^d$, $\Pi_\Gamma(\boldsymbol{w})$ denotes a Euclidean projection of $\boldsymbol{w}$ onto Γ, that is $\Pi_\Gamma(\boldsymbol{w}) \in \arg\min_{\boldsymbol{z} \in \Gamma} \|\boldsymbol{w} - \boldsymbol{z}\|_2$, and w_i denotes the i-th component of $\boldsymbol{w}$. $\mathcal{B}_0(k)$ denotes the ℓ_0 pseudo-ball of radius k, i.e. $\mathcal{B}_0(k) = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_0 \leq k\}$, with $\|\cdot\|_0$ the ℓ_0 pseudo-norm (i.e. the number of nonzero components of a vector). $\mathcal{H}_k$ denotes the Euclidean projection onto $\mathcal{B}_0(k)$, also known as the hard-thresholding operator (which keeps the k largest (in magnitude) components of a vector, and sets the others to 0 (if there are ties, we can break them e.g. lexicographically)). $\|\cdot\|_p$ denotes the ℓ_p norm for $p \in [1, +\infty)$, and $\|\cdot\|$ the ℓ_2 norm (unless otherwise specified). $[n]$ denotes the set $\{1, ..., n\}$ for $n \in \mathbb{N}^*$. For any $S \subseteq [d]$, $|S|$ denotes its number of elements. For any $\boldsymbol{w} \in \mathbb{R}^d$, $\text{supp}(\boldsymbol{w})$ denotes its support, i.e. the set of coordinates of its non-zero components. We will also need the following assumptions on R.

Assumption 1 ((ν_s, s)-RSC, Jain et al. (2014); Negahban et al. (2009); Loh & Wainwright (2013); Yuan et al. (2017); Li et al. (2016); Shen & Li (2017); Nguyen et al. (2017))**.** *R is ν_s restricted strongly convex with sparsity parameter s, i.e. it is differentiable, and there exists a generic constant ν_s such that for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d$ with $\|\boldsymbol{x} - \boldsymbol{y}\|_0 \leq s$:*

$$R(\boldsymbol{y}) \geq R(\boldsymbol{x}) + \langle \nabla R(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\nu_s}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

Assumption 2 ((L_s, s)-RSS, Jain et al. (2014); Li et al. (2016); Yuan et al. (2017))**.** *R is L_s restricted smooth with sparsity level s, i.e. it is differentiable, and there exists a generic constant L_s such that for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d$ with $\|\boldsymbol{x} - \boldsymbol{y}\|_0 \leq s$:*

$$R(\boldsymbol{y}) \leq R(\boldsymbol{x}) + \langle \nabla R(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L_s}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

We then define the notion of support-preserving set that we will use throughout the paper. It essentially requires that projecting any k-sparse vector $\boldsymbol{w}$ onto Γ preserves its support. That is, the convex constraint Γ should be compatible to the sparsity level constraint $\|\boldsymbol{w}\|_0 \leq k$.

Assumption 3 (k-support-preserving set)**.** *$\Gamma \subseteq \mathbb{R}^d$ is k-support-preserving , i.e. it is convex and for any $\boldsymbol{w} \in \mathbb{R}^d$ such that $\|\boldsymbol{w}\|_0 \leq k$, supp$(\Pi_\Gamma(\boldsymbol{w})) \subseteq$ supp$(\boldsymbol{w})$.*

Remark 1. *Below we present some examples of usual sets that also verify Assumption 3 (see Appendix C for a proof of such statements):*

- *Elementwise decomposable constraints, such as box constraints of the form $\{\boldsymbol{w} \in \mathbb{R}^d : \forall i \in [d], l_i \leq w_i \leq u_i\}$.*
- *Group-wise separable constraints where the constraint on each group is k-support-preserving (such as our constraints in Section 5 for the index tracking problem).*
- *Sign-free convex sets (Lu, 2015; Beck & Hallak, 2016) (def. in App. C), e.g. ℓ_q norm-balls.*

3 Deterministic Case

3.1 Algorithm

Two-step projection In all the algorithms of this paper, we will make use of a *two-step projection* operator (TSP), which is different in general from the usual Euclidean projection (EP), in order to obtain, from an arbitrary vector $\boldsymbol{w} \in \mathbb{R}^d$, a vector in $\boldsymbol{w} \in \mathcal{B}_0(k) \cap \Gamma$. We consider such a TSP instead of EP since it enables the derivation a variant of three-point lemma (Lemma 1) which can handle our specific non-convex mixed constraints, and is key to obtaining the convergence analyses we present in Sections 3 and 4. In addition, the TSP can be more intuitive and efficient to implement than EP (see App. F.2 for more discussions about TSP vs EP). The TSP procedure, which we denote by $\bar{\Pi}^k_\Gamma$, is as follows: we first project $\boldsymbol{w}$ onto $\mathcal{B}_0(k)$ through the hard-thresholding operator $\mathcal{H}_k$, to obtain a k-sparse vector $\boldsymbol{v}_k = \mathcal{H}_k(\boldsymbol{w})$. Then, we project $\boldsymbol{v}_k$ onto Γ , to obtain a final vector $\boldsymbol{w}_S = \Pi_\Gamma(\boldsymbol{v}_k)$, where $S = \text{supp}(\boldsymbol{v}_k)$. Note that consequently, the obtained $\boldsymbol{w}_S$ is not necessarily the EP of $\boldsymbol{w}$ onto $\mathcal{B}_0(k) \cap \Gamma$, that is, we do not necessarily have $\boldsymbol{w}_S = \Pi_{\mathcal{B}_0(k)\cap\Gamma}(\boldsymbol{w})$. However, when Assumption 3 is verified, we have $\boldsymbol{w}_S \in \mathcal{B}_0(k) \cap \Gamma$ (since, because of Assumption 3, supp$(\boldsymbol{w}_S) \subseteq$ supp$(\boldsymbol{v}_k)$ and hence $\|\boldsymbol{w}_S\|_0 \leq \|\boldsymbol{v}_k\|_0 \leq k$), therefore each iteration remains feasible in the constraint. We illustrate such a two-step projection on Figure 1.

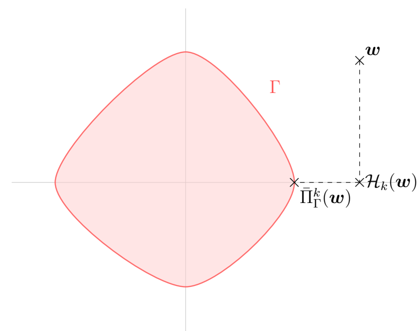


Figure 1: Support-preserving set and two-step projection ($d = 2, k = 1$).

We now present our full algorithm in the case where R is a deterministic function without further knowledge of its structure. It is similar to the usual (non-convex) projected gradient descent algorithm, that is, a gradient update step followed by a projection step, except that instead of projecting onto $\Gamma \cap \mathcal{B}_0(k)$ using the Euclidean projection, we obtain a vector $\boldsymbol{w}_k \in \Gamma \cap \mathcal{B}_0(k)$ through the two-step projection method described above. We describe the algorithm in Algorithm 1 below.

Algorithm 1: Deterministic IHT with extra constraints (IHT-TSP)

Input: $\boldsymbol{w}_0$: initial value, η: learning rate, T: number of iterations
for $t = 1$ *to* T **do**
  $\boldsymbol{w}_t \leftarrow \bar{\Pi}^k_\Gamma(\boldsymbol{w}_{t-1} - \eta\nabla R(\boldsymbol{w}_{t-1}))$;
end
Output: $\boldsymbol{w}_T$

Remark 2. *In the case where Γ is a symmetric sign-free convex set (we refer to Lu (2015) for the definition of such sets, which include for instance any ℓ_p norm constraint set for $p \in [1, +\infty)$), then the two-step projection is the closed form of an Euclidean projection onto the mixed constraint $\Gamma \cap \mathcal{B}_0(k)$ (see Theorem 2.1 from Lu (2015)). Therefore, in such cases, Algorithm 1 is identical to a vanilla (non-convex) projected gradient descent algorithm (for which up to now there was still no global convergence guarantees in such a mixed constraints setting in the literature).*

3.2 CONVERGENCE ANALYSIS

Before proceeding with the convergence analysis, we first present below a variant of the usual three-point lemma, which plays a key role in the proof. The three-point lemma is usually used in the proofs for projected gradient descent in the convex setting. However, due to the non-convexity of the ℓ_0 pseudo-ball, such proofs cannot apply, and to provide convergence in risk, some complex work-arounds are often taken via careful consideration of the sizes of the support of the iterates, such as in the proofs of Jain et al. (2014) or Zhou et al. (2018). However, such a modified three-point lemma below allows to obtain simpler proofs in such non-convex setting, remaining very close to usual convex optimization proofs, while also being able to take into account the additional constraint, which is important in our problem setting. More specifically, the common three-point lemma for a projection onto a convex set $\mathcal{E}$ relates the distance between a point $\boldsymbol{w} \in \mathbb{R}^d$, its projection $\Pi_{\mathcal{E}}(\boldsymbol{w})$, and any vector $\bar{\boldsymbol{w}}$ from the set $\mathcal{E}$, through the relation $\|\boldsymbol{w}-\bar{\boldsymbol{w}}\|^2 \geq \|\Pi_{\mathcal{E}}(\boldsymbol{w})-\boldsymbol{w}\|^2+\|\Pi_{\mathcal{E}}(\boldsymbol{w})-\bar{\boldsymbol{w}}\|^2$. However, in our case, our lemma relates together the four points involved in the two step projection ($\boldsymbol{w} \in \mathbb{R}^d$, $\mathcal{H}_k(\boldsymbol{w})$, $\bar{\Pi}^k_{\Gamma}(\boldsymbol{w})$, and $\bar{\boldsymbol{w}} \in \Gamma \cap \mathcal{B}_0(k)$), and additionally, it contains a constant β which takes into account the sparsity level k enforced in the algorithm and the sparsity $\bar{k}$ ($\leq k$) of a reference point $\bar{\boldsymbol{w}}$ (see e.g. Liu & Foygel Barber (2020) for a discussion regarding k and $\bar{k}$).

Lemma 1 (Constrained ℓ_0-Three-Point, proof in App. D.1)**.** *Suppose that Assumption 3 holds. Consider $\boldsymbol{w}, \bar{\boldsymbol{w}} \in \mathbb{R}^p$ with $\|\bar{\boldsymbol{w}}\|_0 \leq \bar{k}$ and $\bar{\boldsymbol{w}} \in \Gamma$. Then the following holds for any $k > \bar{k}$, with $\beta := \frac{\bar{k}}{k}$:*

$$\|\bar{\Pi}^k_{\Gamma}(\boldsymbol{w}) - \boldsymbol{w}\|^2 \leq \|\boldsymbol{w} - \bar{\boldsymbol{w}}\|^2 - \|\bar{\Pi}^k_{\Gamma}(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2 + \sqrt{\beta}\|\mathcal{H}_k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2.$$

In the case where $\Gamma = \mathbb{R}^d$, we have $\bar{\Pi}^k_{\Gamma}(\boldsymbol{w}) = \mathcal{H}_k(\boldsymbol{w})$, and we can observe that if $k \gg \bar{k}$, β tends to 0, and therefore we approach the usual three-point lemma from convex optimization. This is coherent with the literature on IHT, in which relaxing the sparsity degree (i.e. considering some $k \gg \bar{k}$) is known to make the problem easier to solve (we refer the reader to references in Appendix B.2 for more details). Equipped with such lemma, we can now present the convergence analysis of Algorithm 1 below, using the assumptions from Section 2, and we will describe how the results give rise to a trade-off between the sparsity of the iterates and the tightness of the sub-optimality bound, specific to our mixed constraints setting.

Theorem 1 (Proof in App. D.2)**.** *Suppose that Assumption 1, 2, and 3 hold, and that R is non-negative (without loss of generality). Let $s = 2k$, $\eta = \frac{1}{L_s}$, and $\bar{\boldsymbol{w}}$ be an arbitrary $\bar{k}$-sparse vector. Let $\rho \in (0, \frac{1}{2}]$ be an arbitrary scalar. Suppose that $k \geq \frac{4(1-\rho)^2 L_s^2}{\rho^2 \nu_s^2}\bar{k}$. Then for any $\varepsilon > 0$, for $T \geq \left\lceil \frac{L_s}{\nu_s} \log\left(\frac{(L_s-\nu_s)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2}{2\varepsilon(1-\rho)}\right)\right\rceil + 1 = \mathcal{O}(\kappa_s \log(\frac{1}{\varepsilon}))$, the iterates of IHT-TSP satisfy*

$$\min_{t\in[T]} R(\boldsymbol{w}_t) \leq (1+2\rho)R(\bar{\boldsymbol{w}}) + \varepsilon.$$

Further, if $\bar{\boldsymbol{w}}$ is a global minimizer of R over $\mathcal{B}_0(k) := \{\boldsymbol{w} : \|\boldsymbol{w}\|_0 \leq k\}$, then, with $\rho = 0.5$ in the expressions of k and T above: $\min_{t\in[T]} R(\boldsymbol{w}_t) \leq R(\bar{\boldsymbol{w}}) + \varepsilon$.

Proof Sketch. Our proof starts by deriving a novel convergence proof for IHT in the case where $\Gamma = \mathbb{R}^d$ (Theorem 4 in Appendix), greatly simplifying the one from Jain et al. (2014) (Proof of Thm. 1 in App. B.1), and much closer to usual constrained convex optimization proofs. Using the L_s-RSS of R and some algebraic manipulations, and denoting $\boldsymbol{g}_t = \nabla R(\boldsymbol{w}_t)$ and $\boldsymbol{v}_t := \mathcal{H}_k(\boldsymbol{w}_{t-1} - \frac{1}{L_s}\boldsymbol{g}_{t-1})$ ($= \boldsymbol{w}_t$ when $\Gamma = \mathbb{R}^d$), we have:

$$\begin{aligned}
&R(\boldsymbol{v}_t) \leq R(\boldsymbol{w}_{t-1}) + \frac{L_s}{2}\|\boldsymbol{v}_t - \boldsymbol{w}_{t-1} + \frac{1}{L_s}\boldsymbol{g}_{t-1}\|^2 - \frac{1}{2L_s}\|\nabla R(\boldsymbol{w}_{t-1})\|^2 \\
&\overset{(a)}{\leq} R(\boldsymbol{w}_{t-1}) + \frac{L_s}{2}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1} + \frac{1}{L_s}\boldsymbol{g}_{t-1}\|^2 - \frac{L_s}{2}(1-\sqrt{\beta})\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 - \frac{1}{2L_s}\|\nabla R(\boldsymbol{w}_{t-1})\|^2 \\
&\overset{(b)}{\leq} R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{L_s}{2}(1-\sqrt{\beta})\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2, \qquad (2)
\end{aligned}$$

where in (a) we used our new ℓ_0-three-point lemma (Lemma 3 in App. D.1.1), and in (b) we used the RSC of R with some rearrangements. At that stage, the proof for Theorem 4 can be concluded with telescopic sum arguments. To obtain the proof for general Γ (i.e. Theorem 1), we reiterate the above process but instead of Lemma 3 we use our more general Lemma 1, adapted to general Γ and to our two-step projection technique, to obtain:

$$R(\boldsymbol{w}_t) \leq R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{L_s}{2}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{L_s}{2}\sqrt{\beta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2. \tag{3}$$

Finally, taking a convex combination of equations 2 ($\times\rho$) and 3 ($\times(1-\rho)$) for $\rho \in (0, 0.5]$, using the bound $\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 \leq \|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2$ (non-expansiveness of convex projection onto Γ), and carefully tuning k depending on ρ (resulting in our final trade-off between sparsity and optimality), we can fall back to a telescopic sum and conclude the proof. $\square$

Remark 3. *Theorem 1 therefore provides a global convergence guarantee in objective value. However, contrary to usual guarantees for IHT algorithms under RSS/RSC conditions (which are bounds of the form $R(\boldsymbol{w}_t) \leq R(\bar{\boldsymbol{w}}) + \varepsilon$ for some t) , our bound is of the form $R(\boldsymbol{w}_t) \leq (1+2\rho)R(\bar{\boldsymbol{w}}) + \varepsilon$. There is a trade-off about the choice of $\rho \in (0, 0.5]$. On one hand, $\rho \to 0$ is preferred in view of the RHS of above bound. On the other hand, the sparsity-level relaxation condition $k \geq \frac{4(1-\rho)^2 L_s^2}{\rho^2 \nu_s^2}\bar{k}$ prefers $\rho \to 0.5$. We illustrate such a trade-off on some synthetic experiments in Section F.5.*

4 Extensions: Stochastic and Zeroth-Order Cases

In this section, we provide extensions of Algorithm 1 to the stochastic and zeroth-order sparse optimization problems, and provide the corresponding convergence guarantees in objective value without system error.

4.1 Stochastic Optimization

In this section, we consider the previous risk minimization problem, in a finite-sum setting, i.e. where $R(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n R_i(\boldsymbol{w})$, similarly to Zhou et al. (2018); Nguyen et al. (2017): in such case, stochastic algorithms allow to deal more easily with large-scale datasets where estimating the full $\nabla R(\boldsymbol{w})$ is expensive.

4.1.1 Algorithm

We describe the stochastic variant of our previous Algorithm 1 in Algorithm 2 below, which is an extension of the algorithm from Zhou et al. (2018), to the considered mixed constraints problem setting, using our two-step projection. More precisely, we approximate the gradient of R by a mini-batch stochastic gradient with a batch-size increasing exponentially along training, and following the gradient step, we apply our two-step projection operator.

Algorithm 2: Hybrid Stochastic IHT with Extra Constraints (HSG-HT-TSP)

Input: $\boldsymbol{w}_0$: initial point, η: learning rate, T: number of iterations, $\{s_t\}$: mini-batch sizes.
for $t = 1$ *to* T **do**
  Uniformly sample s_t indices $\mathcal{S}_t$ from $[n]$ without replacement ;
  Compute the approximate gradient $\boldsymbol{g}_{t-1} = \frac{1}{s_{t-1}}\sum_{i_t \in \mathcal{S}_t} \nabla R_{i_t}(\boldsymbol{w}_{t-1})$
  $\boldsymbol{w}_t = \bar{\Pi}_\Gamma^k(\boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_{t-1})$;
end
Output: $\hat{\boldsymbol{w}}_T = \arg\min_{\boldsymbol{w} \in \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T\}} R(\boldsymbol{w})$.

4.1.2 Convergence Analysis

Before proceeding with the convergence analysis, we make an additional assumption on the population variance of the stochastic gradients, similar to the one in Mishchenko et al. (2020).

Assumption 4 (Bounded stochastic gradient variance)**.** *For any $\boldsymbol{w}$, the population variance of the gradient estimator is bounded by B:*

$$\frac{1}{n}\sum_{i=1}^n \|\nabla R_i(\boldsymbol{w}) - \nabla R(\boldsymbol{w})\|^2 \leq B.$$

We now present our convergence analysis below:

Theorem 2 (Proof in App. E.1)**.** *Suppose that Assumptions 1 2, 3 and 4 hold, and that R is non-negative (without loss of generality). Let $s = 2k$. Let $\bar{\boldsymbol{w}}$ be an arbitrary $\bar{k}$-sparse vector. Let C be an arbitrary positive constant. Assume that we run HSG-HT-TSP (Algorithm 2) for T timesteps, with $\eta = \frac{1}{L_s+C}$, and denote $\alpha := \frac{C}{L_s} + 1$ and $\kappa_s := \frac{L_s}{\nu_s}$. Suppose that $k \geq 4\alpha^2 \frac{1}{\rho^2}\kappa_s^2 \bar{k}$ for some $\rho \in (0,1)$. Finally, assume that we take the following batch-size: $s_t := \left\lceil \frac{\tau}{\omega^t} \right\rceil$ with $\omega := 1 - \frac{1}{4\alpha\frac{1}{\rho}\kappa_s}$ and $\tau := \frac{\eta B}{C}$. Then, we have the following convergence rate:*

$$\mathbb{E} \min_{t\in[T]} R(\boldsymbol{w}_t) - (1+2\rho)R(\bar{\boldsymbol{w}}) \leq 2\frac{\alpha^2}{\rho(1-\rho)} L_s \kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right).$$

Further, if $\bar{\boldsymbol{w}}$ is a global minimizer of R over $\mathcal{B}_0(k) := \{\boldsymbol{w} : \|\boldsymbol{w}\|_0 \leq k\}$, then, with $\rho = 0.5$:

$$\mathbb{E} \min_{t\in[T]} R(\boldsymbol{w}_t) - R(\bar{\boldsymbol{w}}) \leq 8\alpha^2 L_s \kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right).$$

Corollary 1 (Proof in App. E.2.)**.** *Therefore, the number of calls to a gradient ∇R_i (#IFO), and the number of hard thresholding operations (#HT) such that the left-hand sides in Theorem 2 above are smaller than some $\varepsilon > 0$, are respectively: #HT $= \mathcal{O}(\kappa_s \log(\frac{1}{\varepsilon}))$ and #IFO $= \mathcal{O}\left(\frac{\kappa_s}{\nu_s \varepsilon}\right)$.*

4.2 Zeroth-Order Optimization

We now consider the zeroth-order (ZO) case (Nesterov & Spokoiny, 2017), in which one does not have access to the gradient $\nabla R(\boldsymbol{w})$, but only to function values $R(\boldsymbol{w})$, which arises for instance when the dataset is private as in distributed learning (Gratton et al., 2021; Zhang et al., 2021) or the model is private as in black-box adversarial attacks Liu et al. (2018), or when computing $\nabla R(\boldsymbol{w})$ is too expensive such as in certain graphical modeling tasks Wainwright et al. (2008). The idea is then to approximate $\nabla R(\boldsymbol{w})$ using finite differences. We refer the reader to Berahas et al. (2021) and Liu et al. (2020) for an overview of ZO methods.

4.2.1 Algorithm

In this section, we describe the ZO version of our algorithm. At its core, it uses the ZO estimator from de Vazelhes et al. (2022). We present the full algorithm in Algorithm 3, where $\mathcal{D}_{s_2}$ is a uniform probability distribution on the following set $\mathcal{B}$, which is the set of unit spheres supported on supports of size $s_2 \leq d$: $\mathcal{B} = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_0 \leq s_2, \|\boldsymbol{w}\|_2 \leq 1\}$. We can sample from this set by first sampling a random support of size s_2, and then sampling from the unit sphere on that support. Note that if we choose $s_2 := d$, this estimator simply becomes the vanilla ZO estimator with unit-sphere smoothing (Liu et al., 2020). Choosing $s_2 < d$ allows to avoid the full-smoothness assumption and can reduce memory consumption by allowing to sample random vectors of size s_2 instead of d. We refer to de Vazelhes et al. (2022) for more details on such a ZO estimator. The difference with de Vazelhes et al. (2022) (in addition to the mixed constraint setting and the use of the TSP) is that in our case we sample an exponentially increasing number of random directions, which allows us to obtain convergence in risk without system error (except the system error due to the smoothing μ).

Algorithm 3: Hybrid ZO IHT with Extra Constraints (HZO-HT-TSP)

Input: $\boldsymbol{w}_0$: initial point, η: learning rate, T: number of iterations, s_2: size of the random supports, $\{q_t\}$: number of random directions.
for $t = 1$ *to* T **do**
  Uniformly sample q_{t-1} i.i.d. random directions $\{\boldsymbol{u}_i\}_{i=1}^{q_{t-1}} \sim \mathcal{D}_{s_2}$;
  Compute the approximate gradient $\boldsymbol{g}_t = \frac{1}{q_{t-1}} \sum_{i=1}^{q_{t-1}} \frac{d}{\mu} \left(R(\boldsymbol{w}_{t-1} + \mu \boldsymbol{u}_i) - R(\boldsymbol{w}_{t-1})\right) \boldsymbol{u}_i$
  $\boldsymbol{w}_t = \bar{\Pi}^k_\Gamma(\boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_{t-1})$;
end
Output: $\hat{\boldsymbol{w}}_T = \arg\min_{\boldsymbol{w} \in \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T\}} R(\boldsymbol{w})$.

4.2.2 Convergence Analysis

Assumption 5 ((L_s, s)-RSS', Shen & Li (2017); Nguyen et al. (2017))**.** *R is L_s strongly restricted smooth with sparsity level s, i.e. it is differentiable, and there exist a generic constant L_s such that for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d$ with $\|\boldsymbol{x} - \boldsymbol{y}\|_0 \leq s$:*

$$\|\nabla R(\boldsymbol{x}) - \nabla R(\boldsymbol{y})\| \leq L_s \|\boldsymbol{x} - \boldsymbol{y}\|.$$

Note that if a function R is (L_s, s)-RSS', then it is (L_s, s)-RSS.

Such assumption is often simply called restricted smoothness, but we name it strong restricted smoothess to avoid any confusion with Assumption 2. Assumption 5 is slightly more restrictive than Assumption 2, but it is necessary when working with ZO gradient estimators (see more details in de Vazelhes et al. (2022)). We now present our main convergence theorem for the ZO setting.

Theorem 3 (Proof in App. E.3)**.** *Suppose that Assumptions 1, 3, and 5 hold, and that R is non-negative (without loss of generality). Let $s = 3k$, and let $\bar{\boldsymbol{w}}$ be an arbitrary $\bar{k}$-sparse vector. Let $s_2 \in \{1, ..., d\}$. Assume that R is $(L_{s'}, s')$-RSS' with $s' = \max(s_2, s)$, and ν_s-RSC. Denote $\kappa_s := \frac{L_{s'}}{\nu_s}$. Let C be an arbitrary positive constant, and denote $\varepsilon_F := \frac{2d}{(s_2+2)}\left(\frac{(s-1)(s_2-1)}{d-1} + 3\right)$, $\varepsilon_{abs} := 2dL_{s'}^2 s s_2 \left(\frac{(s-1)(s_2-1)}{d-1} + 1\right)$, and $\varepsilon_\mu := L_{s'}^2 sd$. Assume that we run HZO-HT-TSP (Algorithm 3) for T timesteps, with $\eta = \frac{1}{L_{s'}+C} = \frac{1}{\alpha L_{s'}}$, with $\alpha := \frac{C}{L_{s'}} + 1$. Suppose that $k \geq 16\frac{\alpha^2}{\rho^2}\kappa_s^2 \bar{k}$ for some $\rho \in (0,1)$. Finally, assume that we take q_t random directions at each iteration, with $q_t := \left\lceil \frac{\tau}{\omega^t} \right\rceil$ with $\omega := 1 - \frac{1}{8\frac{1}{\rho}\alpha\kappa_s}$ and $\tau := 16\kappa_s \frac{\varepsilon_F}{(\alpha-1)}$. Then, we have the following convergence rate:*

$$\mathbb{E} \min_{t\in[T]} R(\boldsymbol{w}_t) - (1+2\rho)R(\bar{\boldsymbol{w}}) \leq 4\frac{\alpha^2}{\rho(1-\rho)} L_{s'}\kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{1}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{\kappa_s L_{s'}} \right) + Z\mu^2,$$

with $Z = \frac{1}{1-\rho}\left(\varepsilon_\mu \left(\frac{2}{\nu_s} + \frac{1}{C}\right) + \frac{\varepsilon_{abs}}{C}\right)$. Further, if $\bar{\boldsymbol{w}}$ is a global minimizer of R over $\mathcal{B}_0(k) := \{\boldsymbol{w} : \|\boldsymbol{w}\|_0 \leq k\}$, then, with $\rho = 0.5$:

$$\mathbb{E} \min_{t\in[T]} R(\boldsymbol{w}_t) - R(\bar{\boldsymbol{w}}) \leq 16\alpha^2 L_{s'}\kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{1}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{\kappa_s L_{s'}} \right) + Z\mu^2.$$

Corollary 2 (Proof in App. E.4.)**.** *Additionally, the number of calls to the function R (#IZO), and the number of hard thresholding operations (#HT) such that the left-hand sides in Theorem 3 above are smaller than $\varepsilon + Z\mu^2$, for some $\varepsilon > 0$ are respectively: #HT $= \mathcal{O}(\kappa_s \log(\frac{1}{\varepsilon}))$ and #IZO $= \mathcal{O}\left(\varepsilon_F \frac{\kappa_s^3 L_s}{\varepsilon}\right)$. Note that if $s_2 = d$, we have $\varepsilon_F = \mathcal{O}(s) = \mathcal{O}(k)$, and therefore we obtain a query complexity that is dimension independent.*

Remark 4. *If $\Gamma = \mathbb{R}^d$, we name the corresponding algorithm HZO-HT, and we provide the convergence rate of HZO-HT in Theorem 6 in Appendix E.3.2, also recalled in Table 1. Such a result is novel, and can be seen as an independent contribution illustrating the power of proof techniques based on our three-point lemma. Up to our knowledge, it is the first global convergence guarantee without system error for a zeroth-order hard-thresholding algorithm (see Table 1), and as such, is a significant improvement over the result from de Vazelhes et al. (2022) .*

5 Experiments

Before describing our experiments, we provide a short discussion about the settings and algorithms that we will illustrate. For constraints Γ for which the Euclidean projection onto $\mathcal{B}_0(k) \cap \Gamma$ has a closed form equal to the TSP, our algorithm is identical to a vanilla non-convex projected gradient descent baseline (see Remark 2). In such case, our contribution in this paper is on the theoretical side, by providing some global guarantees on the optimization, instead of the local guarantees from existing work (cf. Table 1). Additionally, there are case in which there exists a closed form for projection onto $\Gamma \cap \mathcal{B}_0(k)$, different from the TSP (e.g. when $\Gamma = \mathbb{R}_+^d$, cf. Lu (2015)). Although our framework allows us to get approximate global convergence results when using the TSP, still, at the iteration level, a gradient step followed by Euclidean projection (not TSP) is optimal, since it minimizes a constrained quadratic upper bound on R. Therefore, we may not expect much improvement of the TSP over the Euclidean projection in such case, except on the computational side . For these reasons, we illustrate cases where, up to our knowledge, there is no known closed form for projection onto $\Gamma \cap \mathcal{B}_0(k)$, which we believe are the most interesting from the empirical perspective (since no algorithm was about to deal with such cases before). We present below an experiment on a real life index tracking use-case, and provide some extra experimental results in Appendix F, for the settings of multi-class logistic regression as well as adversarial attacks.

Setting: Index tracking. We consider the following index tracking problem, originally presented in Takeda et al. (2013), and used as well in Lu (2015); Beck & Hallak (2016). It is also similar to the portfolio optimization problem presented in Kyrillidis et al. (2013). We seek to reproduce the performance of an index fund (such as S&P500), by investing only in a few key k assets, in order to limit transaction costs. The general problem can be formulated as a linear regression problem:

$$\min_{\boldsymbol{w} \in \mathcal{B}_0(k) \cap \Gamma} \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{y}\|^2 \tag{4}$$

where $\boldsymbol{w}$ represents the amount invested in each asset. For each $i \in [n]$ denoting a timestep , the i-th row of $\boldsymbol{A}$ denotes the returns of the d stocks at timestep i, and y_i the return of the index fund. In our scenario, we seek to limit to a value $D > 0$ the amount of transactions in each of c activity sector (group) of the portfolio (e.g. Industrials, Healthcare, etc.), denoted as G_i for $i \in [c]$. We ensure such constraint through an ℓ_1 norm constraint on each group: $\Gamma = \{\boldsymbol{w} \in \mathbb{R}^d : \forall i \in [c], \|\boldsymbol{w}_{G_i}\|_1 \leq D\}$, where $\boldsymbol{w}_{G_i}$ is the restriction of $\boldsymbol{w}$ to group G_i (i.e. for $j \in [d]$, $\boldsymbol{w}_{G_i j} = \boldsymbol{w}_j$ if $j \in G_i$ and 0 otherwise). In our case, $\boldsymbol{y}$ denotes the daily returns of the S&P500 index from January 1, 2021, to December 31, 2022, and $\boldsymbol{A}$ the returns of the corresponding $d = 497$ assets (over $c = 11$ sectors) of the index during such period. We choose $k = 15$ and $D = 50$. We also apply our algorithms to additional financial indices (CSI300 and HSI) in Appendix F.1.

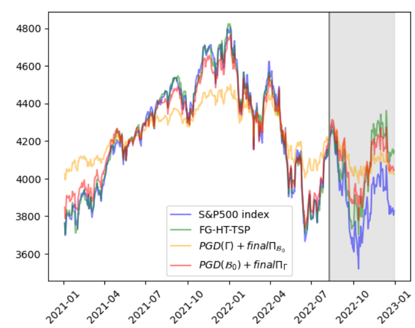


Figure 2: Index Tracking with Sector constraints

Results. Up to our knowledge, there are no closed form for the Euclidean projection onto $\mathcal{B}_0(k) \cap \Gamma$, but the two-step projection can easily be done by projecting onto the ℓ_1 ball for each sector independently. We compare our algorithm (FG-HT-TSP) to two naive baselines: (a) the first one. called "PGD(Γ) + final$\Pi_{\mathcal{B}_0}$", consists in only ensuring the constraints in Γ, followed at the end of training by a simple hard-thresholding step to keep the k largest components of $\boldsymbol{w}$ in absolute value, and (b) the second one, called "PGD($\mathcal{B}_0$) + finalΠ_Γ", consists in running vanilla IHT, followed at the end of training by a simple projection onto Γ to keep $\boldsymbol{w}$ in $\Gamma \cap \mathcal{B}_0$. We plot in Figure 2 the value of the returns for (i) the tracked index, (ii) our index (output of FG-HT-TSP), and (iii) our two baselines (a) and (b). We learn the weights of the portfolio on 80% of the considered period, and evaluate the out of sample (test set) performance on the remaining 20% (shaded area in the figure). As we can observe, the true index is successfully tracked by our method (FG-HT-TSP) (better than the two baselines as can be observed in particular on the train-set: the green curve is the one which is the closest to the blue one), and our algorithm solution spans 9 sectors, therefore it is well diversified, which illustrates the applicability of our method in practice.

6 CONCLUSION

In this paper, we provided global convergence guarantees for variants of Iterative Hard Thresholding which can handle extra convex constraints which are support-preserving, via a two-step projection algorithm. We provided our analysis in the deterministic, stochastic, and zeroth-order settings. To that end, we used a variant of the three-point lemma, adapted to such mixed constraints, which allowed to simplified existing proofs for vanilla constraints (and to provide a new kind of result in the ZO setting), as well as obtaining new proofs in such combined constraints setting. We illustrated the applicability of our algorithm on several sparse learning tasks. Finally, it would also be interesting to extend this work to a broader family of sparsity structures and constraints, for instance to matrices or graphs. We leave this for future work.

REFERENCES

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137:91–129, 2013.

Kyriakos Axiotis and Maxim Sviridenko. Iterative hard thresholding with adaptive regularization: Sparser solutions without sacrificing runtime. In *International Conference on Machine Learning*, pp. 1175–1197, 2022.

Rina Foygel Barber and Wooseok Ha. Gradient descent with non-convex constraints: local concavity determines convergence. *Information and Inference: A Journal of the IMA*, 7:755–806, 2018.

Amir Beck and Nadav Hallak. On the minimization over sparse symmetric sets: Projections, optimality conditions, and algorithms. *Mathematics of Operations Research*, 41:196–223, 2016.

Albert S Berahas, Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, pp. 1–54, 2021.

Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44:197–200, 1992.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146:459–494, 2014.

Radu Ioan Boţ, Ernö Robert Csetnek, and Szilárd Csaba László. An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, 4:3–25, 2016.

Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106:12267–12272, 2009.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

Saeed Damadi and Jinglai Shen. Gradient properties of hard thresholding operator. *arXiv preprint arXiv:2209.08247*, 2022.

Alberto De Marchi and Andreas Themelis. An interior proximal gradient method for nonconvex optimization. *arXiv preprint arXiv:2208.00799*, 2022.

William de Vazelhes, Hualin Zhang, Huimin Wu, Xiaotong Yuan, and Bin Gu. Zeroth-order hard-thresholding: Gradient error vs. expansivity. *Advances in Neural Information Processing Systems*, 35:22589–22601, 2022.

Victor DeMiguel, Lorenzo Garlappi, Francisco J Nogales, and Raman Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55:798–812, 2009.

Simon Foucart and Holger Rauhut. An invitation to compressive sensing. In *A Mathematical Introduction to Compressive Sensing*, pp. 1–39. 2013.

Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165:874–900, sep 2014.

Cristiano Gratton, Naveen KD Venkategowda, Reza Arablouei, and Stefan Werner. Privacy-preserved distributed learning with zeroth-order optimization. *IEEE Transactions on Information Forensics and Security*, 17:265–279, 2021.

Bin Gu, De Wang, Zhouyuan Huo, and Heng Huang. Inexact proximal gradient methods for non-convex and non-smooth optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Patrik O Hoyer. Non-negative sparse coding. In *Proceedings of the 12th IEEE workshop on neural networks for signal processing*, pp. 557–565. IEEE, 2002.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in Neural Information Processing Systems*, 27, 2014.

Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Anastasios Kyrillidis, Stephen Becker, Volkan Cevher, and Christoph Koch. Sparse projections onto the simplex. In *International Conference on Machine Learning*, pp. 235–243, 2013.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. *Advances in Neural Information Processing Systems*, 28, 2015.

Xingguo Li, Raman Arora, Han Liu, Jarvis Haupt, and Tuo Zhao. Nonconvex sparse learning via stochastic optimization with progressive variance reduction. *arXiv preprint arXiv:1605.02711*, 2016.

Haoyang Liu and Rina Foygel Barber. Between hard and soft thresholding: optimal iterative thresholding algorithms. *Information and Inference: A Journal of the IMA*, 9:899–933, 2020.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *arXiv preprint arXiv:1805.10367*, 2018.

Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37:43–54, 2020.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Advances in Neural Information Processing Systems*, 26, 2013.

Zhaosong Lu. Optimization over sparse symmetric sets via a nonmonotone projected gradient method. *arXiv preprint arXiv:1509.08581*, 2015.

Michael R Metel. Sparse training with lipschitz continuous loss functions and a weighted group l0-norm constraint. *Journal of Machine Learning Research*, 24:1–44, 2023.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagréou, Tom Dupré La Tour, Ghislain Durif, Cassio F Dantas, et al. Benchopt: Reproducible, efficient and collaborative optimization benchmarks. *Advances in Neural Information Processing Systems*, 35, 2022.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Advances in Neural Information Processing Systems*, 22, 2009.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, pp. 527–566, 2017.

Nam Nguyen, Deanna Needell, and Tina Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63:6869–6895, 2017.

Lili Pan, Shenglong Zhou, Naihua Xiu, and Hou-Duo Qi. A convergent iterative hard thresholding for nonnegative sparsity optimization. *Pacific Journal of Optimization*, 13:325–353, 2017.

Jie Shen and Ping Li. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18:7650–7691, 2017.

Akiko Takeda, Mahesan Niranjan, Jun-ya Gotoh, and Yoshinobu Kawahara. Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios. *Computational Management Science*, 10:21–49, 2013.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

Yi Xu, Rong Jin, and Tianbao Yang. Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems. *Advances in Neural Information Processing Systems*, 32, 2019a.

Yi Xu, Qi Qi, Qihang Lin, Rong Jin, and Tianbao Yang. Stochastic optimization for dc functions and non-smooth non-convex regularizers with non-asymptotic convergence. In *International Conference on Machine Learning*, pp. 6942–6951, 2019b.

Yingzhen Yang and Ping Li. Projective proximal gradient descent for a class of nonconvex nonsmooth optimization problems: Fast convergence without kurdyka-lojasiewicz (kl) property. *arXiv preprint arXiv:2304.10499*, 2023.

Yingzhen Yang and Jiahui Yu. Fast proximal gradient descent for a class of non-convex and non-smooth sparse learning problems. In *Uncertainty in Artificial Intelligence*, pp. 1253–1262, 2020.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18:6027–6069, 2017.

Qingsong Zhang, Bin Gu, Zhiyuan Dang, Cheng Deng, and Heng Huang. Desirable companion for vertical federated learning: New zeroth-order gradient based algorithm. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2598–2607, 2021.

Pan Zhou, Xiaotong Yuan, and Jiashi Feng. Efficient stochastic gradient hard thresholding. *Advances in Neural Information Processing Systems*, 31, 2018.

Appendix
Optimization over Sparse Support-Preserving Sets via Two-Step Projection

Contents

A NOTATIONS

Below we aggregate the various notations used throughout the paper, for ease of reference.

- $\Pi_\Gamma(\boldsymbol{w})$: Euclidean projection of $\boldsymbol{w}$ onto a set Γ, i.e. $\Pi_\Gamma(\boldsymbol{w}) \in \arg\min_{\boldsymbol{z}\in\Gamma} \|\boldsymbol{w} - \boldsymbol{z}\|_2$.
- w_i: i-th component of $\boldsymbol{w}$.
- $\mathcal{B}_0(k)$: ℓ_0 pseudo-ball of radius k, i.e. $\mathcal{B}_0(k) = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_0 \leq k\}$.
- $\mathcal{H}_k$: Euclidean projection onto $\mathcal{B}_0(k)$, also known as the hard-thresholding operator (which keeps the k largest (in magnitude) components of a vector, and sets the others to 0 (if there are ties, we can break them e.g. lexicographically)).
- $\bar{\Pi}_\Gamma^k$: Two-step projection of sparsity k onto the set Γ, i.e. $\bar{\Pi}_\Gamma^k(\cdot) = \Pi_\Gamma(\mathcal{H}_k(\cdot))$.
- $\|\cdot\|_p$: ℓ_p norm for $p \in [1, +\infty)$.
- $\|\cdot\|$: ℓ_2 norm.
- $[n]$: set $\{1, ..., n\}$ for $n \in \mathbb{N}^*$.
- $|S|$: number of elements of a set $S \subseteq [d]$.
- supp$(\boldsymbol{w})$: support of a vector $\boldsymbol{w} \in \mathbb{R}^d$, i.e. the set of coordinates of its non-zero components.
- TSP: two-step projection
- EP: Euclidean projection

B RELATED WORKS

Below we present a more detailed review of the related works.

B.1 LOCAL GUARANTEES FOR COMBINED CONSTRAINTS

Among the works considering optimization over the intersection of the ℓ_0 pseudo-ball of radius k and a set Γ, Metel (2023) analyze the convergence of a first-order and zeroth-order stochastic algorithm with a weighted ℓ_0 group norm constraint (which generalizes the ℓ_0 norm), combined with an ℓ_∞ ball constraint. Pan et al. (2017) provide a deterministic algorithm which can tackle extra positivity constraints. Lu (2015) and Beck & Hallak (2016) analyze the convergence of variants of hard-thresholding in the deterministic case, with extra constraints that are symmetric and sign-free or positive. Other line of works such as Frankel et al. (2014); Xu et al. (2019b); Attouch et al. (2013); De Marchi & Themelis (2022); Yang & Yu (2020); Gu et al. (2018); Yang & Li (2023); Bolte et al. (2014); Boţ et al. (2016); Xu et al. (2019a); Li & Lin (2015) have a general approach, and analyze the convergence of general proximal algorithms, for composite problems of the form $\min_{\boldsymbol{w}} R(\boldsymbol{w}) + h(\boldsymbol{w})$ where h is a more general non-convex regularizer which can include the ℓ_0 constraint combined with an additional constraint, as long as the closed form for the projection onto the mixed constraint is known (or an approximation of it in the case of Gu et al. (2018)). However, all of these works only provide guarantees of convergence towards a critical point, or at best, a local optimum. We provide an overview of those works in Table 1. More details about algorithms with local convergence specialized to ℓ_0 optimization can also be found in Table 1 from Damadi & Shen (2022).

B.2 GLOBAL GUARANTEES FOR IHT AND RSC FUNCTIONS

On the other hand, in the case of restricted strongly convex (RSC) and restricted smooth (RSS) functions, existing approximate global guarantees for the IHT algorithm do not apply to problems with such combined constraints. Indeed, several works have considered global convergence guarantees for IHT in various settings: the full gradient (deterministic) setting (IHT (Jain et al., 2014)), the stochastic setting (Nguyen et al., 2017; Li et al., 2016; Shen & Li, 2017), and the zeroth-order setting (de Vazelhes et al., 2022). However, they do not address the case where the extra constraint Γ is added to the original sparsity constraint. The works of Barber & Ha (2018); Liu & Foygel Barber (2020) tackle respectively general non-convex thresholding operators, and general non-convex constraints, in the full gradient (deterministic) setting but however they do not provide explicit convergence rates for the particular type of sets that we consider in this paper: their rates depend on some constants (the relative concavity or the local concavity constant) for which, up to our knowledge, an explicit form is still unknown for the sets we consider.

C Proof of Remark 1

Before proceeding with the proof of Remark 1, we recall the definition of sign-free convex sets from Lu (2015) and Beck & Hallak (2016) below. Essentially, sign-free convex sets are convex sets that are closed by swapping the sign of any coordinate.

Definition 1 (Lu (2015),Beck & Hallak (2016))**.** *A convex set Γ is sign-free if for all $\boldsymbol{y} \in \{-1, 1\}^d$ and for all $\boldsymbol{x} \in \Gamma$, $\boldsymbol{x} \odot \boldsymbol{y} \in \Gamma$, where $\odot$ denotes the element-wise vector multiplication (Hadamard product for vectors).*

We now proceed with the proof of Remark 1.

Proof. It is easy to show that any elementwise decomposable constraint such as box constraint is support-preserving (as projection can be done component-wise, independently). Similarly, for group-wise separable constraints where the constraint on each group is k-support-preserving (such as the constraint for the index tracking problem in our Section 5), for a k-sparse vector $\boldsymbol{x} \in \mathbb{R}^d$, one can project each group of coordinates independently, and each of such projection will have its support preserved (since each such group of coordinates also contains less than k non-zero elements, i.e. they are k-sparse). Therefore, we analyze in more detail the case of sign-free convex sets. Let Γ be a sign-free convex set, and let $\boldsymbol{x} \in \mathbb{R}^d$ be a k-sparse vector. Define $\boldsymbol{z} = \Pi_\Gamma(\boldsymbol{x})$ and assume that $\text{supp}(\boldsymbol{z}) \not\subseteq \text{supp}(\boldsymbol{x})$. This implies that there exist some non-empty set of coordinates $S \subseteq [d]$, such that for all $i \in S$: $z_i \neq 0$ and $x_i = 0$. Define $\boldsymbol{z}'$ such that $z'_k = \begin{cases} -z_k \text{ if } k \in S \\ z_k \text{ otherwise} \end{cases}$. Since Γ is sign-free, $\boldsymbol{z}' \in \Gamma$. Now, define $\boldsymbol{z}''$ such that $z''_k = \begin{cases} 0 \text{ if } k \in S \\ z_k \text{ if otherwise} \end{cases}$. Since Γ is convex and since $\boldsymbol{z}'' = \frac{1}{2}\boldsymbol{z}' + \frac{1}{2}\boldsymbol{z}$, we have $\boldsymbol{z}'' \in \Gamma$. Now, we have:

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{z}''\|_2^2 &= \sum_{k=1}^{d}(x_k - z''_k)^2 = \sum_{k \in [d]\setminus S}(x_k - z_k)^2 \\ &< \sum_{k \in [d]\setminus S}(x_k - z_k)^2 + \sum_{k \in S}(x_k - z_k)^2 = \sum_{k=1}^{d}(x_k - z_k)^2 = \|\boldsymbol{x} - \boldsymbol{z}\|_2^2\end{aligned}$$

Therefore, we encounter a contradiction since we have defined $\boldsymbol{z} = \Pi_\Gamma(\boldsymbol{x})$, and therefore, our assumption $\text{supp}(\boldsymbol{z}) \not\subseteq \text{supp}(\boldsymbol{x})$ is wrong, which means that $\text{supp}(\boldsymbol{z}) \subseteq \text{supp}(\boldsymbol{x})$. □

D Proofs of Section 3 (Deterministic optimization)

D.1 Proof of Lemma 1

Before providing the proof of Lemma 1, we first recall below some useful definitions and lemmas from the literature. We then proceed with the proof of Lemma 1 in Section D.1.2.

D.1.1 Useful Lemmas

In this section, as mentioned above, we first recall some useful definitions and lemmas from the literature.

Definition 2 (Relative concavity Liu & Foygel Barber (2020))**.** *The relative concavity coefficient $\gamma_{k,\beta}$ of a k-sparse projection operator $\mathcal{H}_k$, of relative sparsity $\beta := \frac{\bar{k}}{k}$ with $\bar{k} \leq k$ is defined as:*

$$\gamma_{k,\beta}\left(\mathcal{H}_k\right) = \sup\left\{\frac{\langle \boldsymbol{y}-\mathcal{H}_k(\boldsymbol{z}), \boldsymbol{z}-\mathcal{H}_k(\boldsymbol{z})\rangle}{\|\boldsymbol{y}-\mathcal{H}_k(\boldsymbol{z})\|_2^2}\ \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d, \|\boldsymbol{y}\|_0 \leq \beta k, \boldsymbol{y} \neq \mathcal{H}_k(\boldsymbol{z})\right\}.$$

Lemma 2 (Lemma 4.1 Liu & Foygel Barber (2020))**.** *When $\mathcal{H}_k$ is the hard-thresholding operator at sparsity level k, we have:*

$$\gamma_{k,\beta}\left(\mathcal{H}_k\right) = \frac{\sqrt{\beta}}{2} = \frac{1}{2}\sqrt{\frac{\bar{k}}{k}}.$$

Proof. Proof in Liu & Foygel Barber (2020). □

This allows us to derive the following 3 points lemma for hard-thresholding, without additional constraints first:

Lemma 3 (ℓ_0 three-point lemma)**.** *Consider $\boldsymbol{w}, \bar{\boldsymbol{w}} \in \mathbb{R}^p$ with $\|\bar{\boldsymbol{w}}\|_0 \leq \bar{k}$. For any $\bar{k} \leq k$ it holds that:*

$$\|\mathcal{H}_k(\boldsymbol{w}) - \boldsymbol{w}\|^2 \leq \|\boldsymbol{w} - \bar{\boldsymbol{w}}\|^2 - \left(1 - \sqrt{\beta}\right)\|\mathcal{H}_k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2.$$

Proof. We have:

$$\begin{aligned}
\|\boldsymbol{w} - \bar{\boldsymbol{w}}\|^2 &= \|\boldsymbol{w} - \mathcal{H}_k(\boldsymbol{w})\|^2 + \|\mathcal{H}_k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2 + 2\langle \boldsymbol{w} - \mathcal{H}_k(\boldsymbol{w}), \mathcal{H}_k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\rangle \\
&\overset{(a)}{\geq} \|\boldsymbol{w} - \mathcal{H}_k(\boldsymbol{w})\|^2 + \|\mathcal{H}_k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2 - 2\gamma_{k,\rho}\|\mathcal{H}_k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2 \\
&= \|\boldsymbol{w} - \mathcal{H}_k(\boldsymbol{w})\|^2 + (1 - 2\gamma_{k,\rho})\|\mathcal{H}_k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2 \\
&\overset{(b)}{=} \|\boldsymbol{w} - \mathcal{H}_k(\boldsymbol{w})\|^2 + \left(1 - \sqrt{\frac{\bar{k}}{k}}\right)\|\mathcal{H}_k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2,
\end{aligned}$$

where (a) follows from Definition 2 and (b) follows from Lemma 2. Therefore, rearranging, we obtain:

$$\|\mathcal{H}_k(\boldsymbol{w}) - \boldsymbol{w}\|^2 \leq \|\boldsymbol{w} - \bar{\boldsymbol{w}}\|^2 - \left(1 - \sqrt{\frac{\bar{k}}{k}}\right)\|\mathcal{H}_k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2.$$

The proof is completed. □

D.1.2 Proof of Lemma 1

Using the above lemmas, we can now proceed to the proof of Lemma 1.

Proof. Let us abbreviate $\boldsymbol{v}_k := \mathcal{H}_k(\boldsymbol{w})$. It can be verified that

$$\begin{aligned}
\|\bar{\Pi}_{\Gamma}^k(\boldsymbol{w}) - \boldsymbol{w}\|^2 &= \left\|\bar{\Pi}_{\Gamma}^k(\boldsymbol{w}) - \boldsymbol{v}_k + \boldsymbol{v}_k - \boldsymbol{w}\right\|^2 \\
&\overset{(a)}{=} \left\|\bar{\Pi}_{\Gamma}^k(\boldsymbol{w}) - \boldsymbol{v}_k\right\|^2 + \|\boldsymbol{v}_k - \boldsymbol{w}\|^2 \\
&\overset{(b)}{\leq} \|\boldsymbol{v}_k - \bar{\boldsymbol{w}}\|^2 - \|\bar{\Pi}_{\Gamma}^k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2 + \|\boldsymbol{w} - \bar{\boldsymbol{w}}\|^2 - \left(1 - \sqrt{\beta}\right)\|\boldsymbol{v}_k - \bar{\boldsymbol{w}}\|^2 \\
&= \|\boldsymbol{w} - \bar{\boldsymbol{w}}\|^2 - \|\bar{\Pi}_{\Gamma}^k(\boldsymbol{w}) - \bar{\boldsymbol{w}}\|^2 + \sqrt{\beta}\|\boldsymbol{v}_k - \bar{\boldsymbol{w}}\|^2,
\end{aligned}$$

where (a) is due to Assumption 3 and the definition of the two-step projection, which imply that $\bar{\Pi}_{\Gamma}^k(\boldsymbol{w}) - \boldsymbol{v}_k$ and $\boldsymbol{v}_k - \boldsymbol{w}$ have disjoint supporting sets, and (b) uses the three-point-lemma for projection onto a convex set Γ , as well as Lemma 3. The proof is completed. □

D.2 Proof of Theorem 1

Before proceeding with the proof of Theorem 1, we first present a result and proof for the convergence of Algorithm 1, without the additional constraint, which is needed for the proof of Theorem 1, but also, as a byproduct, illustrates how the three-points lemma simplifies previous proofs of Iterative Hard-Thresholding. Then, the full proof of Theorem 1 will be given in Section D.2.2.

D.2.1 CASE WITHOUT ADDITIONAL CONSTRAINT

In this section, we present a result and proof for the convergence of Algorithm 1, without the additional constraint, which as mentioned above, is needed for the proof of Theorem 1, but also, as a byproduct, illustrates how the three-points lemma simplifies previous proofs of Iterative Hard-Thresholding.

Theorem 4. *Assume that $\Gamma = \mathbb{R}^d$. Suppose that Assumption 1 and Assumption 2 holds. Let $s = 2k$. Let $\eta = \frac{1}{L_s}$. Let $\bar{\boldsymbol{w}}$ be an arbitrary $\bar{k}$-sparse vector. Suppose that $k \geq \frac{4L_s^2}{\nu_s^2}\bar{k}$. Then for any $\varepsilon > 0$, the iterate of IHT satisfies $R(\boldsymbol{w}_t) \leq R(\bar{\boldsymbol{w}}) + \varepsilon$ if*

$$t \geq \left\lceil \frac{2L_s}{\nu_s} \log\left(\frac{(L_s - \nu_s)\|\boldsymbol{w}_0 - \bar{\boldsymbol{w}}\|^2}{2\varepsilon}\right)\right\rceil + 1.$$

Proof. The L_s- restricted smoothness of R implies that

$$\begin{aligned}
&R(\boldsymbol{w}_t)\\
\leq& R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\\
=& R(\boldsymbol{w}_{t-1}) + \frac{L_s}{2}\left\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1} + \frac{1}{L_s}\nabla R(\boldsymbol{w}_{t-1})\right\|^2 - \frac{1}{2L_s}\|\nabla R(\boldsymbol{w}_{t-1})\|^2\\
\overset{(a)}{\leq}& R(\boldsymbol{w}_{t-1}) + \frac{L_s}{2}\left\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1} + \frac{1}{L_s}\nabla R(\boldsymbol{w}_{t-1})\right\|^2 - \frac{L_s}{2}(1 - \sqrt{\beta})\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\\
&- \frac{1}{2L_s}\|\nabla R(\boldsymbol{w}_{t-1})\|^2\\
=& R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{L_s}{2}(1 - \sqrt{\beta})\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\\
\overset{(b)}{\leq}& R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{L_s}{2}(1 - \sqrt{\beta})\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\\
\leq& R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{2L_s - \nu_s}{4}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2,
\end{aligned} \tag{5}$$

where (a) uses Lemma 3, (b) is due to the ν_s-restricted strong-convexity of R, while the last step is implied by the condition on the sparsity level k from the theorem ($k \geq \frac{4L_s^2}{\nu_s^2}\bar{k}$), and the definition of β ($\beta = \sqrt{\frac{\bar{k}}{k}}$).

The update rule composed of the gradient step and the projection from Algorithm 1 can be rewritten into the following (given that the learning rate is $\eta = \frac{1}{L_s}$, and by definition of a projection):

$$\begin{aligned}
\boldsymbol{w}_t &= \arg\min_{\boldsymbol{w} \text{ s.t.}\|\boldsymbol{w}\|_0 \leq k} \left\|\boldsymbol{w} - \left(\boldsymbol{w}_{t-1} - \frac{1}{L_s}\nabla R(\boldsymbol{w}_{t-1})\right)\right\|^2\\
&= \arg\min_{\boldsymbol{w} \text{ s.t.}\|\boldsymbol{w}\|_0 \leq k} \frac{2}{L_s}\langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w} - \boldsymbol{w}_{t-1}\rangle + \|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|^2 + \frac{1}{L_s^2}\|\nabla R(\boldsymbol{w}_{t-1})\|^2\\
&= \arg\min_{\boldsymbol{w} \text{ s.t.}\|\boldsymbol{w}\|_0 \leq k} R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w} - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|^2.
\end{aligned}$$

Therefore, by definition of an $\arg\min$, we have:

$$\begin{aligned}
&R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\\
&\leq R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_{t-1} - \boldsymbol{w}_{t-1}\|^2\\
&= R(\boldsymbol{w}_{t-1}).
\end{aligned} \tag{6}$$

And from the L_s smoothness of R, we also have:

$$R(\boldsymbol{w}_t) \leq R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2. \tag{7}$$

Therefore, combining equations 6 and 7, we obtain:

$$R(\boldsymbol{w}_t) \leq R(\boldsymbol{w}_{t-1}).$$

That is, the sequence $\{R(\boldsymbol{w}_t)\}_{t\geq 0}$ of risk is non-increasing.

Let us now consider

$$T := \left\lceil \frac{2L_s}{\nu_s} \log\left(\frac{(L_s-\nu_s)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2}{2\varepsilon}\right)\right\rceil.$$

We claim that $R(\boldsymbol{w}_t) \leq R(\bar{\boldsymbol{w}}) + \varepsilon$ for $t \geq T+1$. To show this, suppose that $\exists t \in [T]$ such that $R(\boldsymbol{w}_t) \leq R(\bar{\boldsymbol{w}}) + \varepsilon$. Then the claim is naturally true by monotonicity. Otherwise assume that $R(\boldsymbol{w}_t) > R(\bar{\boldsymbol{w}}) + \varepsilon$ for all $t \in [T]$. Then in view of the inequality equation 5 we know that

$$\begin{aligned}
\|\boldsymbol{w}_T-\bar{\boldsymbol{w}}\|^2 &\leq \frac{2L_s-2\nu_s}{2L_s-\nu_s}\|\boldsymbol{w}_{T-1}-\bar{\boldsymbol{w}}\|^2\\
&\leq \left(1-\frac{\nu_s}{2L_s}\right)\|\boldsymbol{w}_{T-1}-\bar{\boldsymbol{w}}\|^2\\
&\leq \left(1-\frac{\nu_s}{2L_s}\right)^T\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2\\
&= \exp\left(T\log\left(1-\frac{\nu_s}{2L_s}\right)\right)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2\\
&\leq \exp\left(\frac{2L_s}{\nu_s}\log\left(\frac{(L_s-\nu_s)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2}{2\varepsilon}+1\right)\log\left(1-\frac{\nu_s}{2L_s}\right)\right)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2\\
&= \left(1-\frac{\nu_s}{2L_s}\right)\exp\left(\frac{2L_s}{\nu_s}\log\left(\frac{(L_s-\nu_s)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2}{2\varepsilon}\right)\log\left(1-\frac{\nu_s}{2L_s}\right)\right)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2\\
&\overset{(a)}{\leq} \left(1-\frac{\nu_s}{2L_s}\right)\exp\left(\frac{2L_s}{\nu_s}\log\left(\frac{2\varepsilon}{(L_s-\nu_s)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2}\right)\frac{\nu_s}{2L_s}\right)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2\\
&= \left(1-\frac{\nu_s}{2L_s}\right)\frac{2\varepsilon}{L_s-\nu_s} \overset{(b)}{\leq} \frac{2\varepsilon}{L_s-\nu_s},
\end{aligned}$$

where (a) follows from the fact that for all x in $(-\infty, 1)$: $\log(1-x) \leq -x$, and (b) uses the fact that $\left(1-\frac{\nu_s}{2L_s}\right) \leq 1$.

Then according to equation 5 we must have

$$R(\boldsymbol{w}_{T+1}) \leq R(\bar{\boldsymbol{w}}) + \frac{L_s-\nu_s}{2}\|\boldsymbol{w}_T-\bar{\boldsymbol{w}}\|^2 \leq R(\bar{\boldsymbol{w}}) + \varepsilon,$$

which implies the desired claim. The proof is completed. □

Remark 5. *Theorem 4 recovers the result of Jain et al. (2014, Theorem 1). Our proof is shorter yet more intuitive than in that paper.*

D.2.2 Proof of Theorem 1

Using the above results, we can now proceed to the full proof of convergence of Theorem 1 below.

Proof. Denote $\boldsymbol{v}_t = \mathcal{H}_k(\boldsymbol{w}_{t-1} - \frac{1}{L_s}\nabla R(\boldsymbol{w}_{t-1}))$ for any $t \in \mathbb{N}$. Similar to the arguments for equation 5, based on the L_s-restricted smoothness of R we can show that:

$$\begin{aligned}
&R(\boldsymbol{w}_t)\\
\leq &R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_t-\boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|^2\\
= &R(\boldsymbol{w}_{t-1}) + \frac{L_s}{2}\left\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}+\frac{1}{L_s}\nabla R(\boldsymbol{w}_{t-1})\right\|^2 - \frac{1}{2L_s}\|\nabla R(\boldsymbol{w}_{t-1})\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} R(\boldsymbol{w}_{t-1}) + \frac{L_s}{2}\left\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1} + \frac{1}{L_s}\nabla R(\boldsymbol{w}_{t-1})\right\|^2 - \frac{L_s}{2}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 \\
&\quad + \frac{L_s}{2}\sqrt{\beta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 - \frac{1}{2L_s}\|\nabla R(\boldsymbol{w}_{t-1})\|^2 \\
&= R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{L_s}{2}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 \\
&\quad + \frac{L_s}{2}\sqrt{\beta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 \\
&\overset{(b)}{\leq} R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{L_s}{2}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{L_s}{2}\sqrt{\beta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 \\
&\leq R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{L_s}{2}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\rho\nu_s}{4(1-\rho)}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2, \qquad (8)
\end{aligned}
$$

where (a) uses Lemma 3, (b) is due to the ν_s-restricted strong-convexity of R, and the last step is due to the condition on sparsity level k from the theorem ($k \geq \frac{4L_s^2(1-\rho)^2}{\nu_s^2\rho^2}\bar{k}$), and the definition of $\beta = \sqrt{\frac{\bar{k}}{k}}$.

In view of equation 5, which is valid under the given conditions, we know that

$$
R(\boldsymbol{v}_t) \leq R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{2L_s - \nu_s}{4}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2. \qquad (9)
$$

After proper scaling and summing both sides of equation 8 and equation 9 yields that

$$
\begin{aligned}
&(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t) \\
\leq& R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{(1-\rho)L_s}{2}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 - \frac{\rho(L_s - \nu_s)}{2}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 \\
=& R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 - \frac{L_s - \rho\nu_s}{2}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2, \qquad (10)
\end{aligned}
$$

where in the second inequality we have used $\bar{\boldsymbol{w}} \in \Gamma$ and the non-expansiveness of projection over convex sets.

Let us now consider

$$
T := \left\lceil \frac{2L_s}{\nu_s}\log\left(\frac{(L_s - \nu_s)\|\boldsymbol{w}_0 - \bar{\boldsymbol{w}}\|^2}{2\varepsilon}\right)\right\rceil. \qquad (11)
$$

We claim that:

$$
\min_{t \in [T+1]} \{(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t)\} \leq R(\bar{\boldsymbol{w}}) + \varepsilon. \qquad (12)
$$

To show this, suppose that $\exists t \in [T]$ such that $(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t) \leq R(\bar{\boldsymbol{w}}) + \varepsilon$. Then the claim is naturally true. Otherwise assume that $(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t) > R(\bar{\boldsymbol{w}}) + \varepsilon$ for all $t \in [T]$. Then in view of the inequality equation 10 we know that

$$
\begin{aligned}
\|\boldsymbol{w}_T - \bar{\boldsymbol{w}}\|^2 \leq& \frac{L_s - \nu_s}{L_s - \rho\nu_s}\|\boldsymbol{w}_{T-1} - \bar{\boldsymbol{w}}\|^2 \leq \left(1 - \frac{(1-\rho)\nu_s}{L_s}\right)\|\boldsymbol{w}_{T-1} - \bar{\boldsymbol{w}}\|^2 \\
\leq& \left(1 - \frac{(1-\rho)\nu_s}{L_s}\right)^T\|\boldsymbol{w}_0 - \bar{\boldsymbol{w}}\|^2 \leq \frac{2\varepsilon}{L_s - \nu_s}.
\end{aligned}
$$

Then according to equation 10 we must have

$$
(1-\rho)R(\boldsymbol{w}_{T+1}) + \rho R(\boldsymbol{v}_{T+1}) \leq R(\bar{\boldsymbol{w}}) + \frac{L_s - \nu_s}{2}\|\boldsymbol{w}_T - \bar{\boldsymbol{w}}\|^2 \leq R(\bar{\boldsymbol{w}}) + \varepsilon, \qquad (13)
$$

which proves the claim from equation 12. Now, recall that we have assumed in the Assumptions of Theorem 1, without loss of generality, that R is non-negative (if not, we can redefine R by adding a constant, without modifying the gradient of R, keeping the algorithm untouched), which implies that $R(\boldsymbol{v}_t) \geq 0$. Plugging this in equation 12, for $T \geq \left\lceil \frac{2L_s}{\nu_s} \log\left(\frac{(L_s-\nu_s)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2}{2\varepsilon'(1-\rho)}\right)\right\rceil + 1$ implies that:

$$\min_{t\in[T]} R(\boldsymbol{w}_t) \leq \frac{1}{1-\rho} R(\bar{\boldsymbol{w}}) + \frac{\varepsilon}{1-\rho} \leq (1+2\rho)R(\bar{\boldsymbol{w}}) + \frac{\varepsilon}{1-\rho}. \tag{14}$$

Plugging the change of variable $\varepsilon' = \frac{\varepsilon}{1-\rho}$ into equation 14 above, and in 11, we obtain that when $T \geq \left\lceil \frac{2L_s}{\nu_s} \log\left(\frac{(L_s-\nu_s)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2}{2\varepsilon'(1-\rho)}\right)\right\rceil + 1$:

$$\min_{t\in[T]} R(\boldsymbol{w}_t) \leq (1+2\rho)R(\bar{\boldsymbol{w}}) + \varepsilon'.$$

Further, consider an ideal case where $\bar{\boldsymbol{w}}$ is a global minimizer of R over $\mathcal{B}_0(k) := \{\boldsymbol{w} : \|\boldsymbol{w}\|_0 \leq k\}$. Then $R(\boldsymbol{v}_t) \geq R(\bar{\boldsymbol{w}})$ is always true for all $t \geq 1$. It follows that the bound in equation 12 yields, for $T \geq \left\lceil \frac{2L_s}{\nu_s} \log\left(\frac{(L_s-\nu_s)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2}{2\varepsilon}\right)\right\rceil + 1$:

$$\min_{t\in[T]} \{(1-\rho)R(\boldsymbol{w}_t) + \rho R(\bar{\boldsymbol{w}})\} \leq \min_{t\in[T]} \{(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t)\} \leq R(\bar{\boldsymbol{w}}) + \varepsilon,$$

which implies: $\min_{t\in[T]} R(\boldsymbol{w}_t) \leq R(\bar{\boldsymbol{w}}) + \frac{\varepsilon}{1-\rho}$. In this case, we can simply set $\rho = 0.5$, and define $\varepsilon' = \frac{\varepsilon}{1-\rho} = 2\varepsilon$ similarly as above. This implies the desired claims. The proof is completed.

□

E PROOFS OF SECTION 4 (STOCHASTIC AND ZEROTH-ORDER OPTIMIZATION)

E.1 PROOF OF THEOREM 2

For the proof of Theorem 2, we use a similar technique as in Theorem 1 to deal with the extra constraint, starting from the case $\Gamma = \mathbb{R}^d$ (see Theorem 5, Appendix E.1.2). Based on our ℓ_0 three-point lemma (Lemma 3), such proof of Theorem 5 is much simpler than the corresponding proof of Zhou et al. (2018) (Proof of Theorem 2, Appendix B.3). Also, compared to the deterministic setting, here, we need to carefully incorporate the exponentially decreasing error of the gradient estimator into a propertly weighted telescopic sum containing terms in $\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2$.

Below we provide several intermediary results needed for the proof of Theorem 2. Then, the proof of Theorem 2 will be provided in Section E.1.3.

E.1.1 USEFUL LEMMA

Before starting the proof, we present the following lemma from Mishchenko et al. (2020), which relates the batch-size s_t and the error of the gradient estimator:

Lemma 4 (Mishchenko et al. (2020), Lemma 1)**.** *Let $\boldsymbol{w}_t \in \mathbb{R}^d$. Assume that $\boldsymbol{g}_t$ is the sampled gradient in Algorithm 2 and that the population variance of $R_i(\boldsymbol{w}_t)$ is bounded by B as in Assumption 4. Then the gradient estimate $\boldsymbol{g}_t$ is an unbiased estimate of $\nabla R(\boldsymbol{w}_t)$, and its variance is as follows:*

$$\mathbb{E}\|\boldsymbol{g}_t - \nabla R(\boldsymbol{w}_t)\|^2 \leq \frac{n-s_t}{n-1}\frac{1}{s_t}B, \tag{15}$$

Note that the original Lemma from Mishchenko et al. (2020) is written as an equality, in terms of the exact population variance of a random variable, denoted σ^2, but we rewrite it as an inequality here for simplicity, in order to have a general bound that applies at each iteration.

Proof. Proof in Mishchenko et al. (2020). □

E.1.2 CASE WITHOUT ADDITIONAL CONSTRAINT

Below we now first present some results (and their proofs) for the convergence of Algorithm 2 without the additional constraint, which is needed for the proof of Theorem 2, and also, as a byproduct, illustrates how the three-point lemma simplifies such proof.

Theorem 5. *Assume that $\Gamma = \mathbb{R}^d$. Suppose that Assumption 1, Assumption 2 and Assumption 4 hold. Let $s = 2k$. Let $\bar{\boldsymbol{w}}$ be an arbitrary $\bar{k}$-sparse vector. Let C be an arbitrary positive constant. Assume that we run HSG-HT-TSP (Algorithm 2) for T timesteps, with $\eta = \frac{1}{L_s+C}$, and denote $\alpha := \frac{C}{L_s} + 1$ and $\kappa_s := \frac{L_s}{\nu_s}$. Suppose that $k \geq 4\alpha^2\kappa_s^2\bar{k}$. Finally, assume that we take the following batch-size:*

$s_t := \left\lceil \frac{\tau}{\omega^t} \right\rceil$ with $\omega := 1 - \frac{1}{4\alpha\kappa_s}$ and $\tau := \frac{\eta B}{C}$.

Then, we have the following convergence rate:

$$\mathbb{E}R(\hat{\boldsymbol{w}}_T) - R(\bar{\boldsymbol{w}}) \leq 2\alpha^2 L_s \kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right). \tag{16}$$

Proof. The L_s-smoothness of R implies that

$$\begin{aligned}
&R(\boldsymbol{w}_t)\\
\leq& R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\\
=& R(\boldsymbol{w}_{t-1}) + \langle \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\boldsymbol{w}_t - (\boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_{t-1})\|^2 - \eta^2\|\boldsymbol{g}_{t-1}\|^2 - \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\right] + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\\
&+ \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\boldsymbol{w}_t - (\boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_{t-1})\|^2 - \frac{\eta}{2}\|\boldsymbol{g}_{t-1}\|^2 + \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\\
&+ \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
\overset{(a)}{\leq}& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\bar{\boldsymbol{w}} - (\boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_{t-1})\|^2 - (1-\sqrt{\beta})\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\right] - \frac{\eta}{2}\|\boldsymbol{g}_{t-1}\|^2\\
&+ \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 + \eta^2\|\boldsymbol{g}_{t-1}\|^2 - 2\langle \eta \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle\right] - \frac{1}{2\eta}(1-\sqrt{\beta})\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\\
&- \frac{\eta}{2}\|\boldsymbol{g}_{t-1}\|^2 + \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - 2\langle \eta \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle\right] - \frac{1}{2\eta}(1-\sqrt{\beta})\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\\
&+ \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
\overset{(b)}{=}& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle - \frac{1}{2\eta}(1-\sqrt{\beta})\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\\
&+ \left[\frac{L_s - \frac{1}{\eta} + C}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \frac{1}{2C}\|\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}\|^2,
\end{aligned}$$

where (a) follows from Lemma 3 and (b) follows from the inequality $\langle a, b\rangle \leq \frac{C}{2}a^2 + \frac{1}{2C}b^2$, for any $(a, b) \in (\mathbb{R}^d)^2$ with $C > 0$ an arbitrary strictly positive constant.

Let us now assume that $\eta = \frac{1}{L_s+C}$: therefore the term $\left[\frac{L_s - \frac{1}{\eta} + C}{2}\right] \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2$ above is 0. We now take the conditional expectation (conditioned on w_{t-1}, which is the random variable which realizations are $\boldsymbol{w}_{t-1}$), on both sides, and from Lemma 4 we obtain the inequality below (we slightly abuse notations and denote $\mathbb{E}[\cdot|\mathrm{w}_{t-1} = \boldsymbol{w}_{t-1}]$ by $\mathbb{E}[\cdot|\boldsymbol{w}_{t-1}]$):

$$\begin{aligned}
\mathbb{E}[R(\boldsymbol{w}_t)|\boldsymbol{w}_{t-1}] \leq & R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
& - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right] + \frac{B(n-s_{t-1})}{2Cs_{t-1}(n-1)} \\
\overset{(a)}{\leq} & R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 + \left[R(\bar{\boldsymbol{w}}) - R(\boldsymbol{w}_{t-1}) - \frac{\nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2\right] \\
& - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right] + \frac{B}{2Cs_{t-1}} \\
= & R(\bar{\boldsymbol{w}}) + \left[\frac{\frac{1}{\eta} - \nu_s}{2}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right] \\
& + \frac{B}{2Cs_{t-1}},
\end{aligned}$$

where (a) follows from the RSC condition, and the fact that $s_{t-1} \in \mathbb{N}^*$.

We recall that $\eta = \frac{1}{L_s+C}$. Let us define $\alpha := \frac{C}{L_s} + 1$. Then $C = (\alpha - 1)L_s$, and $\eta = \frac{1}{\alpha L_s}$. Also recall that $\kappa_s = \frac{L_s}{\nu_s}$.

We can simplify the inequality above into:

$$\begin{aligned}
\mathbb{E}[R(\boldsymbol{w}_t)|\boldsymbol{w}_{t-1}] - R(\bar{\boldsymbol{w}}) \leq & \frac{1}{2\eta}\left[\left(1 - \frac{1}{\alpha\kappa_s}\right)\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]\right. \\
& \left. + \frac{\eta B}{Cs_{t-1}}\right].
\end{aligned}$$

We now take the expectation over w_{t-1} of the above inequality (i.e. we take $\mathbb{E}_{\mathrm{w}_{t-1}}[\cdot]$): using the law of total expectation ($\mathbb{E}[\cdot] = \mathbb{E}_{\mathrm{w}_{t-1}}[\mathbb{E}[\cdot|\boldsymbol{w}_{t-1}]]$) we obtain:

$$\mathbb{E}R(\boldsymbol{w}_t) - R(\bar{\boldsymbol{w}}) \leq \frac{1}{2\eta}\left[\left(1 - \frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\eta B}{Cs_{t-1}}\right] \tag{17}$$

Similarly as in Liu & Foygel Barber (2020), we now take a weighted sum over $t = 1, ..., T$, to obtain:

$$\begin{aligned}
& \sum_{t=1}^{T} 2\eta \left(\frac{1 - \frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \mathbb{E}[R(\boldsymbol{w}_t) - R(\bar{\boldsymbol{w}})] \\
\leq & \sum_{t=1}^{T} \left(\frac{1 - \frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \left[\left(1 - \frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\eta B}{Cs_{t-1}}\right] \\
= & \sum_{t=1}^{T} \left(\frac{1 - \frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \left[\left(1 - \frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\right] \\
& + \sum_{t=1}^{T} \left(\frac{1 - \frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \frac{\eta B}{Cs_{t-1}} \\
= & (1-\sqrt{\beta})\sum_{t=1}^{T} \left[\left(\frac{1 - \frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t+1} \mathbb{E}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \left(\frac{1 - \frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\frac{\eta B}{Cs_{t-1}}\\
&\overset{(a)}{=}(1-\sqrt{\beta})\left[\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2-\mathbb{E}\|\boldsymbol{w}_T-\bar{\boldsymbol{w}}\|^2\right]+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\frac{\eta B}{Cs_{t-1}}\\
&\leq(1-\sqrt{\beta})\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\frac{\eta B}{Cs_{t-1}}\\
&\leq\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\frac{\eta B}{Cs_{t-1}},
\end{aligned}
\tag{18}
$$

where (a) follows from simplifying the telescopic sum.

We now choose k and s_t as follows: we choose $k \geq 4\alpha^2\kappa_s^2\bar{k}$, which implies that:

$$
\begin{aligned}
&\sqrt{\beta}\leq\frac{1}{2\alpha\kappa_s}\\
&\implies \sqrt{\beta}\leq\frac{1}{2\alpha\kappa_s-1}\\
&\implies 1-\sqrt{\beta}\geq 1-\frac{1}{2\alpha\kappa_s-1}=\frac{2\alpha\kappa_s-2}{2\alpha\kappa_s-1}=\frac{1-\frac{1}{\alpha\kappa_s}}{1-\frac{1}{2\alpha\kappa_s}}\\
&\implies \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)\leq 1-\frac{1}{2\alpha\kappa_s}.
\end{aligned}
\tag{19}
$$

And we choose $s_t := \left\lceil\frac{\tau}{\omega^t}\right\rceil$ with $\omega := 1-\frac{1}{4\alpha\kappa_s}$ and $\tau := \frac{\eta B}{C}$.

Let us call $\nu := 1-\frac{1}{2\alpha\kappa_s}$. Note that we have:

$$
\nu\leq\omega. \tag{20}
$$

And that we have the inequality below:

$$
\frac{\nu}{\omega}=\frac{1-\frac{1}{2\alpha\kappa_s}}{1-\frac{1}{4\alpha\kappa_s}}=\frac{4\alpha\kappa_s-2}{4\alpha\kappa_s-1}=1-\frac{1}{4\alpha\kappa_s-1}\leq 1-\frac{1}{4\alpha\kappa_s}=\omega. \tag{21}
$$

This allows us to simplify equation 18 into:

$$
\begin{aligned}
\mathbb{E}\sum_{t=1}^{T}2\eta\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}[R(\boldsymbol{w}_t)-R(\bar{\boldsymbol{w}})] &\leq \nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\nu^{T-t}\omega^{t-1}\\
&=\nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{\omega^T}{\omega}\sum_{t=1}^{T}\left(\frac{\nu}{\omega}\right)^{T-t}\\
&=\nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{\omega^T}{\omega}\frac{1-\left(\frac{\nu}{\omega}\right)^T}{1-\left(\frac{\nu}{\omega}\right)}\\
&\leq\nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{\omega^T}{\omega}\frac{1}{1-\left(\frac{\nu}{\omega}\right)}\\
&\overset{(a)}{\leq}\nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{\omega^T}{\omega}\frac{1}{1-\omega}\\
&\overset{(b)}{\leq}\nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\omega^T\frac{1}{1-\omega}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \omega^T \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\omega^T \frac{1}{1-\omega} \\
&\overset{(d)}{\leq} \frac{\omega^T}{1-\omega} \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\omega^T \frac{1}{1-\omega} \\
&= \frac{\omega^T}{1-\omega} \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right) \\
&= 4\alpha\kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right),
\end{aligned}
$$

where in the left hand side we have used the linearity of expectation, and where (a) uses equation 21, (b) uses the fact that $\frac{1}{\omega} = \frac{1}{1-\frac{1}{4\alpha\kappa_s}} \leq \frac{1}{1-\frac{1}{4}} = \frac{4}{3}$ (since $\kappa_s \geq 1$ and $\alpha \geq 1$ (indeed, from the theorem's assumption $\alpha = \frac{C}{L_s} + 1$ with $C > 0$)), (c) uses equation 20, and (d) uses the fact that $\omega < 1$ so $1 < \frac{1}{1-\omega}$.

Let us now normalize the above inequality:

$$
\mathbb{E} \frac{\sum_{t=1}^T 2\eta \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}} \right)^{T-t} R(\boldsymbol{w}_t)}{\sum_{t=1}^T 2\eta \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}} \right)^{T-t}} \leq R(\bar{\boldsymbol{w}}) + \frac{4\alpha\kappa_s\omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\sum_{t=1}^T 2\eta \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}} \right)^{T-t}}.
$$

The left hand side above is a weighted sum, which is an upper bound on the smallest term of the sum.

Regarding the right hand side, we can simplify it using the fact that $0 < \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}} \right)$, and therefore:

$$
\sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}} \right)^{T-t} \geq 1.
$$

Therefore, we obtain:

$$
\mathbb{E} \min_{t \in \{1,..,T\}} R(\boldsymbol{w}_t) - R(\bar{\boldsymbol{w}}) \leq \frac{4\alpha\kappa_s\omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{2\eta} = 2\alpha^2 L_s \kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)
$$

Which can be simplified into the expression below, using the definition of $\hat{\boldsymbol{w}}_T$:

$$
\mathbb{E}R(\hat{\boldsymbol{w}}_T) - R(\bar{\boldsymbol{w}}) \leq 2\alpha^2 L_s \kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right).
$$

The proof is completed. □

Corollary 3. *Under the assumptions of Theorem 5, let ε be a small enough positive number $\varepsilon > 0$. To achieve an error $\mathbb{E}R(\hat{\boldsymbol{w}}_T) - R(\bar{\boldsymbol{w}}) \leq \varepsilon$ using Algorithm 2 the number of calls to a gradient ∇R_i (#IFO), and the number of hard thresholding operations (#HT) are respectively:*

$$
\textit{\#HT} = \mathcal{O}(\kappa_s \log(\frac{1}{\varepsilon})), \quad \textit{\#IFO} = \mathcal{O}\left(\frac{\kappa_s}{\nu_s \varepsilon} \right).
$$

Proof. Let $\varepsilon \in \mathbb{R}_+^*$. Let us find T to ensure that $\mathbb{E}R(\hat{\boldsymbol{w}}_T) - R(\bar{\boldsymbol{w}}) \leq \varepsilon$. This will be enforced if:

$$
\begin{aligned}
&2\alpha^2 L_s \kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right) \leq \varepsilon \\
&\iff T \log(\omega) \leq \log \left(\frac{\varepsilon}{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)} \right)
\end{aligned}
$$

$$\iff T \geq \frac{1}{\log(\frac{1}{\omega})} \log \left(\frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \right).$$

Therefore, let us take:

$$T := \left\lceil \frac{1}{\log(\frac{1}{\omega})} \log \left(\frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \right) \right\rceil. \tag{22}$$

We can now derive the #IFO and #HT. First, we have one hard-thresholding operation at each iteration, therefore #HT= T. Using the fact that $\frac{1}{\log(\frac{1}{\omega})} = \frac{1}{-\log(\omega)} = \frac{1}{-\log(1-\frac{1}{4\alpha\kappa_s})} \leq \frac{1}{\frac{1}{4\alpha\kappa_s}} = 4\alpha\kappa_s$ (since by property of the logarithm, for all $x \in (-\infty, -1) : \log(1-x) \leq -x$), we obtain that #HT $= \mathcal{O}(\kappa_s \log\left(\frac{1}{\varepsilon}\right))$.

We now turn to computing the #IFO. At each iteration t we have s_t gradient evaluations, therefore:

$$\begin{aligned}
\#\text{IFO} &= \sum_{t=0}^{T-1} s_t \\
&\leq \sum_{t=0}^{T-1} \left(\frac{\tau}{\omega^t} + 1 \right) \\
&= T + \tau \frac{\left(\frac{1}{\omega}\right)^T - 1}{\frac{1}{\omega} - 1} \\
&\leq T + \frac{\tau}{\frac{1}{\omega} - 1} \left(\frac{1}{\omega} \right)^T \\
&= T + \frac{\tau}{\frac{1}{\omega} - 1} \exp\left(T \log\left(\frac{1}{\omega} \right) \right) \\
&\overset{(a)}{\leq} 1 + \frac{1}{\log(\frac{1}{\omega})} \log \left(\frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \right) \\
&\quad + \frac{\tau}{\frac{1}{\omega} - 1} \exp\left(\log\left(\frac{1}{\omega} \right) \left[\frac{1}{\log(\frac{1}{\omega})} \log \left(\frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \right) + 1 \right] \right) \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log \left(\frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \right) + \frac{\frac{\tau}{\omega}}{\frac{1}{\omega} - 1} \frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log \left(\frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \right) + \frac{\tau}{1 - \omega} \frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log \left(\frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \right) + \tau \frac{8\alpha^3 L_s \kappa_s^2 \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \\
&\overset{(b)}{=} 1 + \frac{1}{\log(\frac{1}{\omega})} \log \left(\frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \right) \\
&\quad + \frac{B}{\alpha L_s} \frac{1}{L_s(\alpha - 1)} \frac{8\alpha^3 L_s}{\varepsilon} \frac{L_s}{\nu_s} \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right) \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log \left(\frac{2\alpha^2 L_s \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{\varepsilon} \right) + \frac{8B\alpha^2 \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \right)}{(\alpha - 1)\nu_s} \frac{1}{\varepsilon},
\end{aligned}$$

where (a) follows from equation 22, and for (b) we recall that $\tau = \frac{\eta B}{C}$, $\eta = \frac{1}{\alpha L_s}$ and $C = L_s(\alpha - 1)$.

Therefore, overall, the IFO complexity is in $\mathcal{O}(\frac{\kappa_s}{\nu_s \varepsilon})$.

□

E.1.3 Proof of Theorem 2

We now proceed with the full proof of Theorem 2.

Proof. Similary as in the proof of Theorem 5 in Section E.1.2, let us take: $\eta := \frac{1}{L_s+C}$, and $\alpha := \frac{C}{L_s}+1$. Then $C = (\alpha-1)L_s$, and $\eta = \frac{1}{\alpha L_s}$. Recall that $\kappa_s := \frac{L_s}{\nu_s}$. Denote $\boldsymbol{v}_t = \mathcal{H}_k(\boldsymbol{w}_{t-1} - \eta\nabla R(\boldsymbol{w}_{t-1}))$ for any $t \in \mathbb{N}$.

Similarly as in Section E.1.2, the L_s-smoothness of R implies that

$$
\begin{aligned}
&R(\boldsymbol{w}_t)\\
\leq& R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\\
=& R(\boldsymbol{w}_{t-1}) + \langle \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\boldsymbol{w}_t - (\boldsymbol{w}_{t-1} - \eta\boldsymbol{g}_{t-1})\|^2 - \eta^2\|\boldsymbol{g}_{t-1}\|^2 - \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\right] + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\\
&+ \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\boldsymbol{w}_t - (\boldsymbol{w}_{t-1} - \eta\boldsymbol{g}_{t-1})\|^2 - \frac{\eta}{2}\|\boldsymbol{g}_{t-1}\|^2 + \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\\
&+ \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
\overset{(a)}{\leq}& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\bar{\boldsymbol{w}} - (\boldsymbol{w}_{t-1} - \eta\boldsymbol{g}_{t-1})\|^2 - \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \sqrt{\beta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2\right] - \frac{\eta}{2}\|\boldsymbol{g}_{t-1}\|^2\\
&+ \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 + \eta^2\|\boldsymbol{g}_{t-1}\|^2 - 2\langle \eta\boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle\right] - \frac{1}{2\eta}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\\
&+ \frac{\sqrt{\beta}}{2\eta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 - \frac{\eta}{2}\|\boldsymbol{g}_{t-1}\|^2 + \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - 2\langle \eta\boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle\right] - \frac{1}{2\eta}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\sqrt{\beta}}{2\eta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 \qquad (23)\\
&+ \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle\\
\overset{(b)}{=}& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle - \frac{1}{2\eta}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\sqrt{\beta}}{2\eta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2\\
&+ \left[\frac{L_s - \frac{1}{\eta} + C}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \frac{1}{2C}\|\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}\|^2, \qquad (24)
\end{aligned}
$$

where (a) follows from Lemma 1 and (b) follows from the inequality $\langle a, b\rangle \leq \frac{C}{2}a^2 + \frac{1}{2C}b^2$, for any $(a,b) \in (\mathbb{R}^d)^2$ with $C > 0$ an arbitrary strictly positive constant. Let us now take $\eta := \frac{1}{L_s+C}$: therefore the term $\left[\frac{L_s - \frac{1}{\eta} + C}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2$ above is 0. We now take the conditional expectation (conditioned on w_{t-1}, which is the random variable which realizations are $\boldsymbol{w}_{t-1}$), on both sides, and from Lemma 4 we obtain the inequality below (we slightly abuse notations and denote $\mathbb{E}[\cdot|\mathrm{w}_{t-1} = \boldsymbol{w}_{t-1}]$ by $\mathbb{E}[\cdot|\boldsymbol{w}_{t-1}]$):

$$
\mathbb{E}[R(\boldsymbol{w}_t)|\boldsymbol{w}_{t-1}] \leq R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle
$$

$$
\begin{aligned}
&-\frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]+\frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]+\frac{B(n-s_{t-1})}{2Cs_{t-1}(n-1)}\\
\overset{(a)}{\leq}&R(\boldsymbol{w}_{t-1})+\frac{1}{2\eta}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2+\left[R(\bar{\boldsymbol{w}})-R(\boldsymbol{w}_{t-1})-\frac{\nu_s}{2}\|\boldsymbol{w}_{t-1}-\bar{\boldsymbol{w}}\|^2\right]\\
&-\frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]+\frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]+\frac{B}{2Cs_{t-1}}\\
=&R(\bar{\boldsymbol{w}})+\left[\frac{\frac{1}{\eta}-\nu_s}{2}\right]\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-\frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]+\frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]\\
&+\frac{B}{2Cs_{t-1}}, \qquad (25)
\end{aligned}
$$

where (a) follows from the RSC condition, and the fact that $s_{t-1}\in\mathbb{N}^*$.

Now recall that we have taken $\eta=\frac{1}{L_s+C}$, and let us define $\alpha:=\frac{C}{L_s}+1$. Then $C=(\alpha-1)L_s$, and $\eta=\frac{1}{\alpha L_s}$. Also recall that $\kappa_s=\frac{L_s}{\nu_s}$.

We can simplify the inequality above into:

$$
\begin{aligned}
&\mathbb{E}[R(\boldsymbol{w}_t)|\boldsymbol{w}_{t-1}]-R(\bar{\boldsymbol{w}})\\
\leq&\frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha\kappa_s}\right)\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-\mathbb{E}\left[\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]+\sqrt{\beta}\mathbb{E}\left[\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]+\frac{\eta B}{Cs_{t-1}}\right].
\end{aligned}
$$

We now take the expectation over w_{t-1} of the above inequality (i.e. we take $\mathbb{E}_{\mathrm{w}_{t-1}}[\cdot]$): using the law of total expectation ($\mathbb{E}[\cdot]=\mathbb{E}_{\mathrm{w}_{t-1}}[\mathbb{E}[\cdot|\boldsymbol{w}_{t-1}]]$) we obtain:

$$
\mathbb{E}R(\boldsymbol{w}_t)-R(\bar{\boldsymbol{w}})\leq\frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2+\sqrt{\beta}\mathbb{E}\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2+\frac{\eta B}{Cs_{t-1}}\right].
$$

Additionally, in view of equation 17 applied at $\boldsymbol{v}_t$ instead of $\boldsymbol{w}_t$, (since $\boldsymbol{v}_t$ here corresponds to the $\boldsymbol{w}_t$ from Section E.1.2, i.e. $\boldsymbol{v}_t$ is the hard-thresholding of an iterate after a gradient step), we know that:

$$
\mathbb{E}R(\boldsymbol{v}_t)-R(\bar{\boldsymbol{w}})\leq\frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-(1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2+\frac{\eta B}{Cs_{t-1}}\right].
$$

We now take a convex combination similarly as in the case without additional constraint (section D.2), for some $\rho\in(0,1)$.

$$
\begin{aligned}
&\mathbb{E}(1-\rho)R(\boldsymbol{w}_t)+\rho R(\boldsymbol{v}_t)\\
\leq&R(\bar{\boldsymbol{w}})+\frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-(1-\rho)\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right.\\
&\qquad\left.+\left((1-\rho)\sqrt{\beta}-(1-\sqrt{\beta})\rho\right)\mathbb{E}\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2+\frac{\eta B}{Cs_{t-1}}\right]\\
=&R(\bar{\boldsymbol{w}})+\frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-(1-\rho)\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right.\\
&\qquad\left.-\left(\rho-\sqrt{\beta}\right)\mathbb{E}\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2+\frac{\eta B}{Cs_{t-1}}\right]\\
\overset{(b)}{\leq}&R(\bar{\boldsymbol{w}})+\frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-(1-\rho)\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \left(\rho - \sqrt{\beta}\right) \mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\eta B}{C s_{t-1}}\Bigg] \\
&= R(\bar{\boldsymbol{w}}) + \frac{1}{2\eta}\left[\left(1 - \frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\eta B}{C s_{t-1}}\right],
\end{aligned}
$$

where in (b), we have assumed that $\sqrt{\beta} \leq \rho$ (later we will verify that our choice of k ensures such a condition), and have used the fact that projection onto a convex set is non-expansive (which implies that $\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 \geq \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2$). Similarly as in E.1.2, we now take a weighted sum over $t = 1, ..., T$, to obtain:

$$
\begin{aligned}
&\sum_{t=1}^{T} 2\eta \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \mathbb{E}[(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t) - R(\bar{\boldsymbol{w}})] \\
&\leq \sum_{t=1}^{T} \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \left[\left(1 - \frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\eta B}{C s_{t-1}}\right] \\
&= \sum_{t=1}^{T} \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \left[\left(1 - \frac{1}{\alpha\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\right] \\
&\quad + \sum_{t=1}^{T} \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \frac{\eta B}{C s_{t-1}} \\
&= (1-\sqrt{\beta}) \sum_{t=1}^{T} \left[\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t+1} \mathbb{E}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\right] \\
&\quad + \sum_{t=1}^{T} \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \frac{\eta B}{C s_{t-1}} \\
&\overset{(a)}{=} (1-\sqrt{\beta}) \left[\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T} \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 - \mathbb{E}\|\boldsymbol{w}_T - \bar{\boldsymbol{w}}\|^2\right] + \sum_{t=1}^{T} \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \frac{\eta B}{C s_{t-1}} \\
&\leq (1-\sqrt{\beta}) \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T} \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \sum_{t=1}^{T} \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \frac{\eta B}{C s_{t-1}} \\
&\leq \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T} \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \sum_{t=1}^{T} \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \frac{\eta B}{C s_{t-1}},
\end{aligned} \tag{26}
$$

where (a) follows from simplifying the telescopic sum.

We now choose k and s_t as follows: we choose $k \geq 4\frac{1}{\rho^2}\alpha^2\kappa_s^2\bar{k}$, which implies that:

$\rho \geq \sqrt{\beta}$ (thereby verifying the assumption made earlier), and that:

$$
\begin{aligned}
&\sqrt{\beta} \leq \frac{1}{2\alpha\frac{1}{\rho}\kappa_s} \\
&\implies \sqrt{\beta} \leq \frac{1}{2\alpha\frac{1}{\rho}\kappa_s - 1} \\
&\implies 1 - \sqrt{\beta} \geq 1 - \frac{1}{2\alpha\frac{1}{\rho}\kappa_s - 1} = \frac{2\alpha\frac{1}{\rho}\kappa_s - 2}{2\alpha\frac{1}{\rho}\kappa_s - 1} = \frac{1 - \frac{1}{\alpha\frac{1}{\rho}\kappa_s}}{1 - \frac{1}{2\alpha\frac{1}{\rho}\kappa_s}} \overset{(a)}{\geq} \frac{1 - \frac{1}{\alpha\kappa_s}}{1 - \frac{1}{2\alpha\frac{1}{\rho}\kappa_s}} \\
&\implies \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right) \leq 1 - \frac{1}{2\alpha\frac{1}{\rho}\kappa_s},
\end{aligned} \tag{27}
$$

where (a) follows from the fact that $\rho \leq 1$.

And we now choose $s_t := \lceil \frac{\tau}{\omega^t} \rceil$, with $\omega := 1 - \frac{1}{4\alpha \frac{1}{\rho}\kappa_s}$ and $\tau := \frac{\eta B}{C}$.

Let us call $\nu := 1 - \frac{1}{2\alpha \frac{1}{\rho}\kappa_s}$. Note that we have:

$$\nu \leq \omega. \tag{28}$$

And that we have the inequality below:

$$\frac{\nu}{\omega} = \frac{1 - \frac{1}{2\alpha \frac{1}{\rho}\kappa_s}}{1 - \frac{1}{4\alpha \frac{1}{\rho}\kappa_s}} = \frac{4\alpha\frac{1}{\rho}\kappa_s - 2}{4\alpha\frac{1}{\rho}\kappa_s - 1} = 1 - \frac{1}{4\alpha\frac{1}{\rho}\kappa_s - 1} \leq 1 - \frac{1}{4\alpha\frac{1}{\rho}\kappa_s} = \omega. \tag{29}$$

This allows us to simplify equation 26 into:

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T} 2\eta \left(\frac{1 - \frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} [(1-\rho)R(w_t) + \rho R(v_t) - R(\bar{w})] \\
&\leq \nu^T \|\bar{w} - w_0\|^2 + \sum_{t=1}^{T} \nu^{T-t}\omega^{t-1} \\
&= \nu^T \|\bar{w} - w_0\|^2 + \frac{\omega^T}{\omega}\sum_{t=1}^{T}\left(\frac{\nu}{\omega}\right)^{T-t} \\
&= \nu^T \|\bar{w} - w_0\|^2 + \frac{\omega^T}{\omega}\frac{1 - \left(\frac{\nu}{\omega}\right)^T}{1 - \left(\frac{\nu}{\omega}\right)} \\
&\leq \nu^T \|\bar{w} - w_0\|^2 + \frac{\omega^T}{\omega}\frac{1}{1 - \left(\frac{\nu}{\omega}\right)} \\
&\overset{(a)}{\leq} \nu^T \|\bar{w} - w_0\|^2 + \frac{\omega^T}{\omega}\frac{1}{1-\omega} \\
&\overset{(b)}{\leq} \nu^T \|\bar{w} - w_0\|^2 + \frac{4}{3}\omega^T\frac{1}{1-\omega} \\
&\overset{(c)}{\leq} \omega^T \|\bar{w} - w_0\|^2 + \frac{4}{3}\omega^T\frac{1}{1-\omega} \\
&\overset{(d)}{\leq} \frac{\omega^T}{1-\omega} \|\bar{w} - w_0\|^2 + \frac{4}{3}\omega^T\frac{1}{1-\omega} \\
&= \frac{\omega^T}{1-\omega}\left(\|\bar{w} - w_0\|^2 + \frac{4}{3}\right) \\
&= 4\alpha\frac{1}{\rho}\kappa_s\omega^T\left(\|\bar{w} - w_0\|^2 + \frac{4}{3}\right),
\end{aligned}$$

where in the left hand side we have used the linearity of expectation, and where (a) uses equation 29, (b) uses the fact that $\frac{1}{\omega} = \frac{1}{1 - \frac{1}{4\alpha\frac{1}{\rho}\kappa_s}} \leq \frac{1}{1-\frac{1}{4}} = \frac{4}{3}$ (since $\kappa_s \geq 1$ and $\alpha \geq 1$ (indeed, from the theorem's assumption $\alpha = \frac{C}{L_s} + 1$ with $C > 0$), so consequently $\alpha\frac{1}{\rho} \geq 1$), (c) uses equation 28, and (d) uses the fact that $\omega < 1$ so $1 < \frac{1}{1-\omega}$.

Let us now normalize the above inequality:

$$\mathbb{E}\frac{\sum_{t=1}^{T} 2\eta\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t)}{\sum_{t=1}^{T} 2\eta\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}} \leq R(\bar{\boldsymbol{w}}) + \frac{4\alpha\frac{1}{\rho}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\sum_{t=1}^{T} 2\eta\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}}.$$

The left hand side above is a weighted sum, which is an upper bound on the smallest term of the sum.

Regarding the right hand side, we can simplify it using the fact that $0 < \left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)$, and therefore:

$$\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \geq 1.$$

Therefore, we obtain:

$$\begin{aligned}\mathbb{E}\min_{t\in\{1,..,T\}}(1-\rho)R(\boldsymbol{w}_t)+\rho R(\boldsymbol{v}_t)-R(\bar{\boldsymbol{w}}) &\leq \frac{4\alpha\frac{1}{\rho}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\right)}{2\eta}\\ &= 2\alpha^2\frac{1}{\rho}L_s\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\right). \end{aligned} \tag{30}$$

We denote by ε_T the right-hand side above:

$$\varepsilon_T = 2\alpha^2\frac{1}{\rho}L_s\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\right).$$

We now proceed similarly as in the proof of Theorem 1 above. Recall that we have assumed in the Assumptions of Theorem 2, without loss of generality, that R is non-negative, which implies that $R(\boldsymbol{v}_t) \geq 0$. Plugging this in equation 30 implies that:

$$\mathbb{E}\min_{t\in[T]} R(\boldsymbol{w}_t) \leq \frac{1}{1-\rho}R(\bar{\boldsymbol{w}}) + \frac{\varepsilon_T}{1-\rho} \leq (1+2\rho)R(\bar{\boldsymbol{w}}) + \frac{\varepsilon_T}{1-\rho}. \tag{31}$$

Plugging the change of variable $\varepsilon'_T = \frac{\varepsilon_T}{1-\rho}$ into equation 31 above, we obtain that:

$$\mathbb{E}\min_{t\in[T]} R(\boldsymbol{w}_t) \leq (1+2\rho)R(\bar{\boldsymbol{w}}) + \varepsilon'_T.$$

Further, consider an ideal case where $\bar{\boldsymbol{w}}$ is a global minimizer of R over $\mathcal{B}_0(k) := \{\boldsymbol{w} : \|\boldsymbol{w}\|_0 \leq k\}$. Then $R(\boldsymbol{v}_t) \geq R(\bar{\boldsymbol{w}})$ is always true for all $t \geq 1$. It follows that the bound in equation 31 yields:

$$\mathbb{E}\min_{t\in[T]}\{(1-\rho)R(\boldsymbol{w}_t)+\rho R(\bar{\boldsymbol{w}})\} \leq \mathbb{E}\min_{t\in[T]}\{(1-\rho)R(\boldsymbol{w}_t)+\rho R(\boldsymbol{v}_t)\} \leq R(\bar{\boldsymbol{w}})+\varepsilon_T,$$

which implies: $\mathbb{E}\min_{t\in[T]} R(\boldsymbol{w}_t) \leq R(\bar{\boldsymbol{w}}) + \frac{\varepsilon_T}{1-\rho}$. In this case, we can simply set $\rho = 0.5$, and define $\varepsilon'_T = \frac{\varepsilon_T}{1-\rho} = 2\varepsilon_T$ similarly as above.. The proof is completed. □

E.2 Proof of Corollary 1

Proof. We proceed similarly as in the proof of Corollary 3 in Section E.1.2:

Let $\varepsilon \in \mathbb{R}^*_+$. Let us find T to ensure that $\mathbb{E}\min_{t\in\{1,..,T\}}(1-\rho)R(\boldsymbol{w}_t)+\rho R(\boldsymbol{v}_t) - R(\bar{\boldsymbol{w}}) \leq \varepsilon$ This will be enforced if:

$$\begin{aligned}&2\alpha^2\frac{1}{\rho}L_s\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\right) \leq \varepsilon\\ &\iff T\log(\omega) \leq \log\left(\frac{\varepsilon}{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\right)}\right)\\ &\iff T \geq \frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\right)}{\varepsilon}\right).\end{aligned}$$

Therefore, let us take:

$$T := \left\lceil \frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\right)}{\varepsilon}\right)\right\rceil. \tag{32}$$

We can now derive the #IFO and #HT. First, we have one hard-thresholding operation at each iteration, therefore #HT$= T$. Using the fact that $\frac{1}{\log(\frac{1}{\omega})} = \frac{1}{-\log(\omega)} = \frac{1}{-\log(1-\frac{1}{4\alpha\frac{1}{\rho}\kappa_s})} \leq \frac{1}{\frac{1}{4\alpha\frac{1}{\rho}\kappa_s}} = 4\alpha\frac{1}{\rho}\kappa_s$ (since by property of the logarithm, for all $x \in (-\infty, -1) : \log(1-x) \leq -x$), we obtain that #HT $= \mathcal{O}(\kappa_s \log\left(\frac{1}{\varepsilon}\right))$.

We now turn to computing the #IFO. At each iteration t we have s_t gradient evaluations, therefore:

$$
\begin{aligned}
\#\text{IFO} &= \sum_{t=0}^{T-1} s_t \\
&\leq \sum_{t=0}^{T-1} \left(\frac{\tau}{\omega^t} + 1\right) \\
&= T + \tau \frac{\left(\frac{1}{\omega}\right)^T - 1}{\frac{1}{\omega} - 1} \\
&\leq T + \frac{\tau}{\frac{1}{\omega} - 1}\left(\frac{1}{\omega}\right)^T \\
&= T + \frac{\tau}{\frac{1}{\omega} - 1} \exp\left(T \log\left(\frac{1}{\omega}\right)\right) \\
&\overset{(a)}{\leq} 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon}\right) \\
&\quad + \frac{\tau}{\frac{1}{\omega} - 1} \exp\left(\log\left(\frac{1}{\omega}\right)\left[\frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon}\right) + 1\right]\right) \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon}\right) + \frac{\frac{\tau}{\omega}}{\frac{1}{\omega} - 1} \frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon} \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon}\right) + \frac{\tau}{1-\omega} \frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon} \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon}\right) + \tau \frac{8\alpha^3\frac{1}{\rho^2}L_s\kappa_s^2\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon} \\
&\overset{(b)}{=} 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon}\right) \\
&\quad + \frac{B}{\alpha L_s}\frac{1}{L_s(\alpha-1)}\frac{8\alpha^3\frac{1}{\rho^2}L_s}{\varepsilon}\frac{L_s}{\nu_s}\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right) \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{2\alpha^2\frac{1}{\rho}L_s\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{\varepsilon}\right) + \frac{8B\alpha^2\frac{1}{\rho^2}\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right)}{(\alpha-1)\nu_s}\frac{1}{\varepsilon},
\end{aligned}
$$

where (a) follows from equation 32, and for (b) we recall that $\tau = \frac{\eta B}{C}, \eta = \frac{1}{\alpha L_s}$ and $C = L_s(\alpha - 1)$.

Therefore, overall, the IFO complexity is in $\mathcal{O}(\frac{\kappa_s}{\nu_s \varepsilon})$.

$\square$

E.3 Proof of Theorem 3

Our proof is similar to the one for Theorem 2, though we needed to refine some results from de Vazelhes et al. (2022) to properly express the variance of the ZO gradient estimator and incorporate it into the telescopic sum. Before proving the main Theorem 3, below we provide several intermediary results needed for the proof of Theorem 3. Then, the proof of Theorem 2 will be provided in Section E.3.3.

E.3.1 USEFUL LEMMAS

We first recall the following results from de Vazelhes et al. (2022):

Proposition 1 (Proposition 1 (i) de Vazelhes et al. (2022))**.** *Let us consider any support $F \subseteq [d]$ of size s ($|F| = s$). For the Z0 gradient estimator $\boldsymbol{g}_t$ in Algorithm 3 at $\boldsymbol{w}_t$, with q_t random directions, and random supports of size s_2, and assuming that R is (L_{s_2}, s_2)-RSS' , we have, with $[\boldsymbol{u}]_F$ denoting the hard thresholding of a vector $\boldsymbol{u}$ on F (that is, we set all coordinates not in F to 0):*

$$\|[\mathbb{E}\boldsymbol{g}_t]_F - [\nabla R(\boldsymbol{w}_t)]_F\|^2 \leq \varepsilon_\mu \mu^2 \tag{33}$$

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

Proof. Proof in de Vazelhes et al. (2022). □

Lemma 5 (Lemma C.2 de Vazelhes et al. (2022))**.** *For any (L_{s_2}, s_2)-RSS' function R, using the gradient estimator $\boldsymbol{g}_t$ defined in Algorithm 3 with $q_t = 1$, we have, for any support $F \subseteq [d]$, with $|F| = s$, and $F^c := [d] \setminus F$:*

$$\mathbb{E}\|[\boldsymbol{g}_t]_F\|^2 = \varepsilon_F \left\|[\nabla R(\boldsymbol{w}_t)]_F\right\|^2 + \varepsilon_{F^c} \left\|[\nabla R(\boldsymbol{w}_t)]_{F^c}\right\|^2 + \varepsilon_{abs}\mu^2 \tag{34}$$

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) $\varepsilon_{abs} := 2dL_s^2 ss_2\left(\frac{(s-1)(s_2-1)}{d-1} + 1\right)$.

Proof. Proof in de Vazelhes et al. (2022). □

We now use the above lemma to bound the variance of the zeroth-order gradient estimator $\boldsymbol{g}_t$.

Lemma 6. *The gradient estimator $\boldsymbol{g}_t$ defined in Algorithm 3 verifies the following properties for any $q_t \in \mathbb{N}^*$:*

$$\mathbb{E}\|[\boldsymbol{g}_t]_F - \mathbb{E}[\boldsymbol{g}_t]_F\|^2 \leq \frac{\varepsilon_F}{q_t}\left\|\nabla R(\boldsymbol{w})\right\|^2 + \frac{\varepsilon_{abs}}{q_t}\mu^2 \tag{35}$$

with ε_F and ε_{abs} defined above in Lemma 5

Proof. If $q_t = 1$, we have:

$$\begin{aligned}
\mathbb{E}\|[\boldsymbol{g}_t]_F - \mathbb{E}[\boldsymbol{g}_t]_F\|^2 &\overset{(a)}{=} \mathbb{E}\|[\boldsymbol{g}_t]_F\|^2 - \|[\mathbb{E}\boldsymbol{g}]_F\|^2 \\
&\leq \mathbb{E}\|[\boldsymbol{g}_t]_F\|^2 \\
&\overset{(34)}{\leq} \varepsilon_F\|[\nabla R(\boldsymbol{w})]_F\|^2 + \varepsilon_{F^c}\|[\nabla R(\boldsymbol{w})]_{F^c}\|^2 + \varepsilon_{\text{abs}}\mu^2 \\
&\overset{(b)}{\leq} \varepsilon_F\|\nabla R(\boldsymbol{w})\|^2 + \varepsilon_{abs}\mu^2,
\end{aligned}$$

where (a) follows from the bias-variance formula $\mathbb{E}\|X - E[X]\|_2^2 = \mathbb{E}\|X\|_2^2 - \|\mathbb{E}X\|_2^2$ for a multi-dimensional random variable X, and (b) follows from the fact that

$$\varepsilon_F = \frac{2d}{s_2+2}\left(\frac{s(s_2-1)}{d-1} + 3 - \frac{s_2-1}{d}\right) > \frac{2d}{s_2+2}\left(\frac{s(s_2-1)}{d-1}\right) = \varepsilon_{F^c}$$

(since $s_2 \leq d$), and since $\|[\nabla R(\boldsymbol{w})]_F\|^2 + \|[\nabla R(\boldsymbol{w})]_{F^c}\|^2 = \|\nabla R(\boldsymbol{w})\|^2$ (by definition of the Euclidean norm).

Now, if $q_t \geq 1$, we know that the variance of an average of q_t i.i.d. realizations of a random variable of total variance σ^2 is $\frac{\sigma^2}{q_t}$ (and its expected value remains the same by linearity of expectation):

indeed, for any random multidimensional random variable X, for which we consider the q i.i.d. random variables X_i of same distribution, we have:

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{q_t}\sum_{i=1}^{q_t} X_i - \mathbb{E}\left[\frac{1}{q_t}\sum_{i=1}^{q_t} X_i\right]\right\|_2^2 &= \mathbb{E}\left\|\frac{1}{q_t}\sum_{i=1}^{q_t}(X_i - \mathbb{E}X_i)\right\|_2^2 \\
&= \frac{1}{q_t^2}\left(\sum_{i=1}^{q_t}(X_i - \mathbb{E}X_i)\right)^\top\left(\sum_{i=1}^{q_t}(X_i - \mathbb{E}X_i)\right) \\
&\overset{(a)}{=} \frac{1}{q_t^2}\sum_{i=1}^{q_t}\|X_i - \mathbb{E}X_i\|_2^2 \\
&= \frac{1}{q_t^2}\sum_{i=1}^{q_t}\|X - \mathbb{E}X\|_2^2 \\
&= \frac{1}{q_t^2} q_t \|X - \mathbb{E}X\|_2^2 \\
&= \frac{1}{q_t}\|X - \mathbb{E}X\|_2^2,
\end{aligned}$$

where (a) follows from the fact that X_i are i.i.d hence for $i \neq j$: $\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i - \mathbb{E}X_i)^\top(X_j - \mathbb{E}X_j) = 0$. Applying this to the random variable which realizations are $[\boldsymbol{g}_t]_F$, this concludes the proof. □

E.3.2 Case without additional constraint

Below we now first present some results (and their proofs) for the convergence of Algorithm 3 without the additional constraint, which is needed for the proof of Theorem 3, and also, as a byproduct, provides, up to our knowledge, the first convergence guarantee in objective value without system error for a zeroth-order hard-thresholding algorithm.

Theorem 6. *Assume that $\Gamma = \mathbb{R}^d$. Let $\bar{\boldsymbol{w}}$ be an arbitrary $\bar{k}$-sparse vector. Let $s = 3k$, and $s_2 \in \{1, ..., d\}$. Assume that R is $(L_{s'}, s')$-RSS' with $s' = \max(s_2, s)$, and ν_s-restricted strongly convex. Denote $\kappa_s := \frac{L_{s'}}{\nu_s}$. Let C be an arbitrary positive constant, and denote $\varepsilon_F := \frac{2d}{(s_2+2)}\left(\frac{(s-1)(s_2-1)}{d-1} + 3\right)$, $\varepsilon_{abs} := 2dL_{s'}^2 s s_2\left(\frac{(s-1)(s_2-1)}{d-1} + 1\right)$, and $\varepsilon_\mu := L_{s'}^2 sd$. Assume that we run HZO-HT-TSP (Algorithm 3) for T timesteps, with $\eta = \frac{1}{L_{s'}+C} = \frac{1}{\alpha L_{s'}}$, with $\alpha := \frac{C}{L_{s'}} + 1$. Suppose that $k \geq 16\alpha^2\kappa_s^2\bar{k}$. Finally, assume that we take the following number q_t of random directions at each iteration:*

$q_t := \left\lceil \frac{\tau}{\omega^t} \right\rceil$ with $\omega := 1 - \frac{1}{8\alpha\kappa_s}$ and $\tau := 16\kappa_s\frac{\varepsilon_F}{(\alpha-1)}$. Then, we have the following convergence rate:

$$\mathbb{E}R(\hat{\boldsymbol{w}}_T) - R(\bar{\boldsymbol{w}}) \leq 4\alpha^2 L_{s'}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{1}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{\kappa_s L_{s'}}\right) + Z\mu^2 \tag{36}$$

with $Z = \varepsilon_\mu\left(\frac{2}{\nu_s} + \frac{1}{C}\right) + \frac{\varepsilon_{abs}}{C}$

Proof. Let us denote for simplicity: $C_1 := \frac{\varepsilon_F}{q_t}$, $C_2 := \frac{\varepsilon_{abs}}{q_t}$, and $C_3 := \varepsilon_\mu\mu^2$. Moreover, let us denote $F := \mathrm{supp}(\boldsymbol{w}_t) \cup \mathrm{supp}(\boldsymbol{w}_{t-1}) \cup \mathrm{supp}(\bar{\boldsymbol{w}})$, where supp denotes the support of a vector, i.e. the set of coordinates of its non-zero components. Note that therefore we have $|F| \leq 2k + \bar{k} \leq 3k$. In addition $[\boldsymbol{u}]_F$ denotes the thresholding of $\boldsymbol{u}$ to the support F, that is, the vector $\boldsymbol{u}$ with its components that are not in F set to 0.

The fact that R is $(L_{s'}, s')$-RSS', therefore also $(L_{s'}, s)$-RSS', implies from the remark in 5 that it is also $(L_{s'}, s)$-RSS, therefore:

$R(\boldsymbol{w}_t)$

$$
\begin{aligned}
&\leq R(\boldsymbol{w}_{t-1}) + \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_t - \boldsymbol{w}_{t-1} \rangle + \frac{L_{s'}}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 \\
&= R(\boldsymbol{w}_{t-1}) + \langle \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1} \rangle + \frac{L_{s'}}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1} \rangle \\
&= R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta} \left[\|\boldsymbol{w}_t - (\boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_{t-1})\|^2 - \eta^2 \|\boldsymbol{g}_{t-1}\|^2 - \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 \right] + \frac{L_{s'}}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 \\
&\quad + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1} \rangle \\
&= R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta} \|\boldsymbol{w}_t - (\boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_{t-1})\|^2 - \frac{\eta}{2} \|\boldsymbol{g}_{t-1}\|^2 + \left[\frac{L_{s'} - \frac{1}{\eta}}{2} \right] \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 \\
&\quad + \langle [\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F, \boldsymbol{w}_t - \boldsymbol{w}_{t-1} \rangle \\
&\overset{(a)}{\leq} R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta} \left[\|\bar{\boldsymbol{w}} - (\boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_{t-1})\|^2 - (1 - \sqrt{\beta}) \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 \right] - \frac{\eta}{2} \|\boldsymbol{g}_{t-1}\|^2 \\
&\quad + \left[\frac{L_{s'} - \frac{1}{\eta}}{2} \right] \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle [\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F, \boldsymbol{w}_t - \boldsymbol{w}_{t-1} \rangle \\
&= R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta} \left[\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 + \eta^2 \|\boldsymbol{g}_{t-1}\|^2 - 2\langle \eta \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}} \rangle \right] - \frac{1}{2\eta} (1 - \sqrt{\beta}) \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 \\
&\quad - \frac{\eta}{2} \|\boldsymbol{g}_{t-1}\|^2 + \left[\frac{L_{s'} - \frac{1}{\eta}}{2} \right] \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle [\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F, \boldsymbol{w}_t - \boldsymbol{w}_{t-1} \rangle \\
&= R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta} \left[\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - 2\langle \eta \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}} \rangle \right] - \frac{1}{2\eta} (1 - \sqrt{\beta}) \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 \\
&\quad + \left[\frac{L_{s'} - \frac{1}{\eta}}{2} \right] \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle [\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F, \boldsymbol{w}_t - \boldsymbol{w}_{t-1} \rangle \\
&\overset{(b)}{=} R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta} \|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}} \rangle - \frac{1}{2\eta} (1 - \sqrt{\beta}) \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 \\
&\quad + \left[\frac{L_{s'} - \frac{1}{\eta} + C}{2} \right] \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \frac{1}{2C} \|[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F\|^2 \\
&= R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta} \|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}} \rangle + \langle [\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}} \rangle \\
&\quad - \frac{1}{2\eta} (1 - \sqrt{\beta}) \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \left[\frac{L_{s'} - \frac{1}{\eta} + C}{2} \right] \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \frac{1}{2C} \|[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F\|^2,
\end{aligned}
$$

where (a) follows from Lemma 3 and (b) follows from the inequality $\langle a, b \rangle \leq \frac{C}{2} a^2 + \frac{1}{2C} b^2$, for any $(a, b) \in (\mathbb{R}^d)^2$ with $C > 0$ an arbitrary strictly positive constant.

Let us now choose $\eta := \frac{1}{L_{s'} + C}$: therefore the term $\left[\frac{L_{s'} - \frac{1}{\eta} + C}{2} \right] \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2$ above is 0. We now take the conditional expectation (conditioned on w_{t-1}, which is the random variable which realizations are $\boldsymbol{w}_{t-1}$), on both sides, and from Lemma 4 we obtain the inequality below (we slightly abuse notations and denote $\mathbb{E}[\cdot | \mathrm{w}_{t-1} = \boldsymbol{w}_{t-1}]$ by $\mathbb{E}[\cdot | \boldsymbol{w}_{t-1}]$):

$$
\begin{aligned}
&\mathbb{E}[R(\boldsymbol{w}_t) | \boldsymbol{w}_{t-1}] \\
&\leq R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta} \|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}} \rangle \\
&\quad - \frac{1}{2\eta} (1 - \sqrt{\beta}) \mathbb{E} \left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1} \right] + \langle \left[\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E} \left[\boldsymbol{g}_{t-1} | \boldsymbol{w}_{t-1} \right] \right]_F, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}} \rangle \\
&\quad + \mathbb{E} \left[\frac{1}{2C} \|[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F\|^2 | \boldsymbol{w}_{t-1} \right]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(a)}{\leq} R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&\quad - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}\|[\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2 + \frac{1}{2G}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 \\
&\quad + \frac{1}{2C}\mathbb{E}\left[\|\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}\|^2 | \boldsymbol{w}_{t-1}\right] \\
&= R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&\quad - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}\|[\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2 \\
&\quad + \frac{1}{2C}\mathbb{E}\left[\|[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F\|^2 | \boldsymbol{w}_{t-1}\right] \\
&\overset{(b)}{\leq} R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&\quad - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}\|[\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2 \\
&\quad + \frac{1}{2C}\left(2\|[\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2 + 2\|[\boldsymbol{g}_{t-1} - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2\right) \\
&\overset{(33)+(35)}{\leq} R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&\quad - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 \\
&\quad + \frac{1}{2C}\left(2C_3 + 2C_1\|\nabla R(\boldsymbol{w}_{t-1})\|^2 + 2C_2\mu^2\right) \\
&\overset{(c)}{\leq} R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&\quad - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 \\
&\quad + \frac{1}{2C}\left(2C_1\left(2\|\nabla R(\boldsymbol{w}_{t-1}) - \nabla R(\bar{\boldsymbol{w}})\|^2 + 2\|\nabla R(\bar{\boldsymbol{w}})\|^2\right) + 2C_2\mu^2 + 2C_3\right) \\
&\overset{(d)}{\leq} R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&\quad - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 \\
&\quad + \frac{1}{2C}\left(2C_1\left(2L_{s'}^2\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 + 2\|\nabla R(\bar{\boldsymbol{w}})\|^2\right) + 2C_2\mu^2 + 2C_3\right) \\
&= R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G} + \frac{2C_1L_{s'}^2}{C}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&\quad - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2 + C_2\mu^2 + C_3\right) \\
&\overset{(e)}{\leq} R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G} + \frac{2C_1L_{s'}^2}{C}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 + \left[R(\bar{\boldsymbol{w}}) - R(\boldsymbol{w}_{t-1}) - \frac{\nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2\right] \\
&\quad - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2 + C_2\mu^2 + C_3\right) \\
&= R(\bar{\boldsymbol{w}}) + \left[\frac{\frac{1}{\eta} - \nu_s}{2} + \frac{1}{2G} + \frac{2C_1L_{s'}^2}{C}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 \\
&\quad - \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2 + C_2\mu^2 + C_3\right)
\end{aligned}
$$

$$\overset{(f)}{\leq} R(\bar{\boldsymbol{w}}) + \left[\frac{\frac{1}{\eta} - \nu_s}{2} + \frac{1}{2G} + \frac{2\varepsilon_F L_{s'}^2}{\tau C}\right] \|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2$$
$$- \frac{1}{2\eta}(1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2 + C_2\mu^2 + C_3\right), \tag{37}$$

where (a) follows from the inequality $\langle a, b\rangle \leq \frac{G}{2}a^2 + \frac{1}{2G}b^2$, for any $(a, b) \in (\mathbb{R}^d)^2$ with $G > 0$ an arbitrary strictly positive constant, (b) and (c) follow from the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any $(a, b) \in (\mathbb{R}^d)^2$, (d) follows from the fact that R is $(L_{s'}, s')$-RSS' (Assumption 5 with sparsity level s'), therefore it is also $(L_{s'}, s_2)$-RSS', (e) follows from the RSC condition, and for (f), we recall that $C_1 = \frac{\varepsilon_F}{q_t}$, and we define $q_t = \left\lceil \frac{\tau}{\omega^t} \right\rceil$, for some $\omega > 1$ and $\tau > 0$ that will be chosen later in the proof.

Recall that we have chosen $\eta = \frac{1}{L_{s'}+C}$. Let us define $\alpha := \frac{C}{L_{s'}} + 1$. Then $C = (\alpha - 1)L_{s'}$, and $\eta = \frac{1}{\alpha L_{s'}}$. Also recall that $\kappa_s = \frac{L_{s'}}{\nu_s}$.

We will now choose the constant G and C, in order to simplify the inequality above, such that it matches as much as possible the structure of the previous proofs:

We will seek to rewrite:
$\left[\frac{\frac{1}{\eta}-\nu_s}{2} + \frac{1}{2G} + \frac{2\frac{\varepsilon_F}{\tau}L_{s'}^2}{C}\right] \left(= \frac{1}{2\eta}\left[1 + \frac{1}{G\alpha L_{s'}} + \frac{4L_{s'}^2\frac{\varepsilon_F}{\tau}}{(\alpha-1)\alpha L_{s'}^2} - \frac{1}{\alpha\kappa_s}\right]\right)$, into :

$\frac{1}{2\eta}\left[1 - \frac{1}{\alpha'\kappa_s}\right]$ for some $\alpha' > 0$ (we will seek $\alpha' \propto \alpha$, with a dimensionless proportionality constant for simplicity).

Therefore, let us choose $G := \frac{4}{\nu_s}$, which implies:

$$\frac{1}{G\alpha L_{s'}} = \frac{1}{4\alpha\kappa_s}. \tag{38}$$

And let us choose $\tau := \frac{16\kappa_s\varepsilon_F}{(\alpha-1)}$, which implies:

$$\frac{4L_{s'}^2\frac{\varepsilon_F}{\tau}}{(\alpha-1)\alpha L_{s'}^2} = \frac{1}{4\alpha\kappa_s}. \tag{39}$$

Therefore, using equations 38 and 39, we obtain:

$$\begin{aligned}\left[\frac{\frac{1}{\eta}-\nu_s}{2} + \frac{1}{2G} + \frac{2\frac{\varepsilon_F}{\tau}L_{s'}^2}{C}\right] &= \frac{1}{2\eta}\left[1 + \frac{1}{G\alpha L_{s'}} + \frac{4L_{s'}^2\frac{\varepsilon_F}{\tau}}{(\alpha-1)\alpha L_{s'}^2} - \frac{1}{\alpha\kappa_s}\right] \\ &= \frac{1}{2\eta}\left[1 + \frac{1}{4\alpha\kappa_s} + \frac{1}{4\alpha\kappa_s} - \frac{1}{\alpha\kappa_s}\right] \\ &= \frac{1}{2\eta}\left[1 - \frac{1}{2\alpha\kappa_s}\right] = \frac{1}{2\eta}\left[1 - \frac{1}{\alpha'\kappa_s}\right],\end{aligned}$$

where for simplicity we have denoted $\alpha' = 2\alpha$.

We can therefore simplify (37) into:

$$\begin{aligned}\mathbb{E}[R(\boldsymbol{w}_t)|\boldsymbol{w}_{t-1}] - R(\bar{\boldsymbol{w}}) \leq &\frac{1}{2\eta}\left[\left(1 - \frac{1}{\alpha'\kappa_s}\right)\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]\right. \\ &\left. + 2\eta\left(\frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2 + C_2\mu^2 + C_3\right)\right)\right].\end{aligned}$$

We now take the expectation over w_{t-1} of the above inequality (i.e. we take $\mathbb{E}_{\mathrm{w}_{t-1}}[\cdot]$): using the law of total expectation ($\mathbb{E}[\cdot] = \mathbb{E}_{\mathrm{w}_{t-1}}[\mathbb{E}[\cdot|\boldsymbol{w}_{t-1}]]$) we obtain:

$$\mathbb{E}R(\boldsymbol{w}_t) - R(\bar{\boldsymbol{w}}) \leq \frac{1}{2\eta}\left[\left(1 - \frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2\right. \tag{40}$$

$$+2\eta\left(\frac{G}{2}C_3+\frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2+C_2\mu^2+C_3\right)\right)\Bigg] \tag{41}$$

Let us call $A := 2\eta\left(\frac{G}{2}C_3+\frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2+C_2\mu^2+C_3\right)\right)$ for simplicity. Similarly as in Liu & Foygel Barber (2020), we now take a weighted sum over $t = 1, ..., T$, to obtain:

$$\begin{aligned}
&\sum_{t=1}^{T}2\eta\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\mathbb{E}[R(\boldsymbol{w}_t)-R(\bar{\boldsymbol{w}})] \\
\le&\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-(1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2+A\right] \\
=&\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-(1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right] \\
&+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}A \\
=&(1-\sqrt{\beta})\sum_{t=1}^{T}\left[\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t+1}\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right] \\
&+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}A \\
\overset{(a)}{=}&(1-\sqrt{\beta})\left[\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2-\mathbb{E}\|\boldsymbol{w}_T-\bar{\boldsymbol{w}}\|^2\right]+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}A \\
\le&(1-\sqrt{\beta})\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}A \\
\le&\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}A \\
=&\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}2\eta\left(\frac{G}{2}C_3+\frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2+C_2\mu^2+C_3\right)\right) \\
=&\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}2\eta\left(\frac{G}{2}C_3+\frac{1}{C}\left(2\frac{\varepsilon_F}{q_t}\|\nabla R(\bar{\boldsymbol{w}})\|^2+\frac{\varepsilon_{abs}\mu^2}{q_t}+C_3\right)\right) \\
=&\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\frac{2\eta}{q_t}\left(\frac{2\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2+\varepsilon_{abs}\mu^2}{C}\right) \\
&+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}2\eta C_3\left(\frac{G}{2}+\frac{1}{C}\right) \\
=&\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\frac{2\eta}{q_t}\left(\frac{2\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}\right) \\
&+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}2\eta\mu^2\left(\varepsilon_\mu\left(\frac{G}{2}+\frac{1}{C}\right)+\frac{\varepsilon_{abs}}{Cq_t}\right) \\
\le&\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\frac{2\eta}{q_t}\left(\frac{2\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}\right)
\end{aligned}$$

$$+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta\mu^2\left(\varepsilon_\mu\left(\frac{G}{2}+\frac{1}{C}\right)+\frac{\varepsilon_{abs}}{C}\right), \tag{42}$$

where (a) follows from simplifying the telescopic sum. Let us denote for simplicity $\zeta := \frac{2\eta(2\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2)}{C} = \frac{4\eta\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}$ and $Z := \varepsilon_\mu\left(\frac{G}{2}+\frac{1}{C}\right)+\frac{\varepsilon_{abs}}{C}$.

We now choose k and q_t as follows: we choose $k \geq 4\alpha'^2\kappa_s^2\bar{k}$, which implies that:

$$\begin{aligned}
&\sqrt{\beta} \leq \frac{1}{2\alpha'\kappa_s} \\
&\implies \sqrt{\beta} \leq \frac{1}{2\alpha'\kappa_s - 1} \\
&\implies 1-\sqrt{\beta} \geq 1-\frac{1}{2\alpha'\kappa_s-1} = \frac{2\alpha'\kappa_s-2}{2\alpha'\kappa_s-1} = \frac{1-\frac{1}{\alpha'\kappa_s}}{1-\frac{1}{2\alpha'\kappa_s}} \\
&\implies \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right) \leq 1-\frac{1}{2\alpha'\kappa_s}.
\end{aligned} \tag{43}$$

We recall that we previously defined $q_t = \left\lceil \frac{\tau}{\omega^t}\right\rceil$, with $\tau := \frac{16\kappa_s\varepsilon_F}{(\alpha-1)}$. We now set the value of ω, to $\omega := 1-\frac{1}{4\alpha'\kappa_s}$.

Let us call $\nu := 1-\frac{1}{2\alpha'\kappa_s}$. Note that we have:

$$\nu \leq \omega. \tag{44}$$

And that we have the inequality below:

$$\frac{\nu}{\omega} = \frac{1-\frac{1}{2\alpha'\kappa_s}}{1-\frac{1}{4\alpha'\kappa_s}} = \frac{4\alpha'\kappa_s-2}{4\alpha'\kappa_s-1} = 1-\frac{1}{4\alpha'\kappa_s-1} \leq 1-\frac{1}{4\alpha'\kappa_s} = \omega. \tag{45}$$

This allows us to simplify equation 42 into:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} 2\eta\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}[R(\boldsymbol{w}_t)-R(\bar{\boldsymbol{w}})]\right] \\
&\leq \nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\sum_{t=1}^{T}\nu^{T-t}\omega^{t-1} + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&= \nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\frac{\omega^T}{\omega}\sum_{t=1}^{T}\left(\frac{\nu}{\omega}\right)^{T-t} + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&= \nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\frac{\omega^T}{\omega}\frac{1-\left(\frac{\nu}{\omega}\right)^T}{1-\left(\frac{\nu}{\omega}\right)} + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&\leq \nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\frac{\omega^T}{\omega}\frac{1}{1-\left(\frac{\nu}{\omega}\right)} + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&\overset{(a)}{\leq} \nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\frac{\omega^T}{\omega}\frac{1}{1-\omega} + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&\overset{(b)}{\leq} \nu^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\frac{4}{3}\omega^T\frac{1}{1-\omega} + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&\overset{(c)}{\leq} \omega^T\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\frac{4}{3}\omega^T\frac{1}{1-\omega} + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2
\end{aligned}$$

$$\begin{aligned}
&\overset{(d)}{\leq} \frac{\omega^T}{1-\omega}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\frac{4}{3}\omega^T\frac{1}{1-\omega} + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&= \frac{\omega^T}{1-\omega}\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\frac{4}{3}\right) + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&= 4\alpha'\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{\zeta}{\tau}\frac{4}{3}\right) + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2,
\end{aligned}$$

where in the left hand side we have used the linearity of expectation, and where (a) uses equation 45, (b) uses the fact that $\frac{1}{\omega} = \frac{1}{1-\frac{1}{4\alpha'\kappa_s}} \leq \frac{1}{1-\frac{1}{4}} = \frac{4}{3}$ (since $\kappa_s \geq 1$ and $\alpha' \geq 1$ (indeed, we have $\alpha' = 2\alpha = 2(\frac{C}{L_{s'}} + 1)$ with $C > 0$)), (c) uses equation 44, and (d) uses the fact that $\omega < 1$ so $1 < \frac{1}{1-\omega}$.

Let us now normalize the above inequality:

$$\mathbb{E}\frac{\sum_{t=1}^{T} 2\eta\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} R(\boldsymbol{w}_t)}{\sum_{t=1}^{T} 2\eta\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}} \leq R(\bar{\boldsymbol{w}}) + \frac{4\alpha'\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{4}{3}\frac{\zeta}{\tau}\right)}{\sum_{t=1}^{T} 2\eta\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}} + Z\mu^2.$$

The left hand side above is a weighted sum, which is an upper bound on the smallest term of the sum.

Regarding the right hand side, we can simplify it using the fact that $0 < \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)$, and therefore:

$$\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \geq 1.$$

Therefore, we obtain:

$$\begin{aligned}
\mathbb{E}\min_{t\in\{1,..,T\}} R(\boldsymbol{w}_t) - R(\bar{\boldsymbol{w}}) &\leq \frac{4\alpha'\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{4}{3}\frac{\zeta}{\tau}\right)}{2\eta} + Z\mu^2 \\
&= 4\alpha^2 L_{s'}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{4}{3}\frac{\zeta}{\tau}\right) + Z\mu^2.
\end{aligned}$$

Which can be simplified into the expression below, using the definition of $\hat{\boldsymbol{w}}_T$:

$$\mathbb{E}R(\hat{\boldsymbol{w}}_T) - R(\bar{\boldsymbol{w}}) \leq 4\alpha^2 L_{s'}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \frac{4}{3}\frac{\zeta}{\tau}\right) + Z\mu^2. \tag{46}$$

To simplify the above result, we recall the assumptions made earlier on: we have chosen $\tau = \frac{16\kappa_s\varepsilon_F}{(\alpha-1)}$, and $G = \frac{4}{\nu_s}$.

Therefore, to sum up, we have:

$$Z = \varepsilon_\mu\left(\frac{G}{2} + \frac{1}{C}\right) + \frac{\varepsilon_{abs}}{C} = \varepsilon_\mu\left(\frac{2}{\nu_s} + \frac{1}{C}\right) + \frac{\varepsilon_{abs}}{C}.$$

$$\omega = 1 - \frac{1}{4\alpha'\kappa_s} = 1 - \frac{1}{8\alpha\kappa_s}$$

$$\zeta = \frac{4\eta\varepsilon_F \|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}$$

The last inequality implies: $\frac{\zeta}{\tau} = \frac{\frac{4\eta\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}}{16\kappa_s L_{s'}\frac{\varepsilon_F}{C}} = \frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}$. □

Corollary 4. *Additionally, the number of calls to the function R (#IZO), and the number of hard thresholding operations (#HT) such that the upper bound in Theorem 2 above is smaller than $\varepsilon + Z\mu$, with $\varepsilon > 0$ are respectively: #HT $= \mathcal{O}(\kappa_s \log(\frac{1}{\varepsilon}))$ and #IZO $= \mathcal{O}\left(\frac{\varepsilon_F \kappa_s^3 L_s}{\varepsilon}\right)$. Note that if $s_2 = d$, we have $\varepsilon_F = \mathcal{O}(s) = \mathcal{O}(k)$, and therefore we obtain a query complexity that is dimension independent.*

Proof. Let $\varepsilon \in \mathbb{R}_+^*$. Let us find T to ensure that $\mathbb{E}R(\hat{\boldsymbol{w}}_T) - R(\bar{\boldsymbol{w}}) \leq \varepsilon + Z\mu^2$ This will be enforced if:

$$\begin{aligned}
&4\alpha^2 L_{s'}\kappa_s \omega^T \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right) \leq \varepsilon \\
&\iff T\log(\omega) \leq \log\left(\frac{\varepsilon}{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}\right) \\
&\iff T \geq \frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right).
\end{aligned}$$

Therefore, let us take:

$$T := \left\lceil \frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right)\right\rceil. \tag{47}$$

We can now derive the #IZO and #HT. First, we have one hard-thresholding operation at each iteration, therefore #HT$= T$. Using the fact that $\frac{1}{\log(\frac{1}{\omega})} = \frac{1}{-\log(\omega)} = \frac{1}{-\log(1-\frac{1}{8\alpha\kappa_s})} \leq \frac{1}{\frac{1}{8\alpha\kappa_s}} = 8\alpha\kappa_s$ (since by property of the logarithm, for all $x \in (-\infty, -1) : \log(1-x) \leq -x$), and the fact that $\alpha = \frac{C}{L_{s'}}$ is independent of κ_s, we obtain that #HT $= \mathcal{O}(\kappa_s \log\left(\frac{1}{\varepsilon}\right))$.

We now turn to computing the #IZO. At each iteration t we have q_t function evaluations, therefore:

$$\begin{aligned}
\#\text{IFO} &= \sum_{t=0}^{T-1} q_t \\
&\leq \sum_{t=0}^{T-1}\left(\frac{\tau}{\omega^t} + 1\right) \\
&= T + \tau\frac{\left(\frac{1}{\omega}\right)^T - 1}{\frac{1}{\omega} - 1} \\
&\leq T + \frac{\tau}{\frac{1}{\omega} - 1}\left(\frac{1}{\omega}\right)^T \\
&= T + \frac{\tau}{\frac{1}{\omega} - 1}\exp\left(T\log\left(\frac{1}{\omega}\right)\right) \\
&\overset{(a)}{\leq} 1 + \frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\tau}{\frac{1}{\omega}-1} \exp\left(\log\left(\frac{1}{\omega}\right)\left[\frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right)+1\right]\right) \\
&= 1+\frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right)+\frac{\frac{\tau}{\omega}}{\frac{1}{\omega}-1}\frac{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon} \\
&= 1+\frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right)+\frac{\tau}{1-\omega}\frac{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon} \\
&= 1+\frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{4\alpha^2 L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right)+\tau\frac{32\alpha^3 L_{s'}\kappa_s^2\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon},
\end{aligned}
$$

where (a) follows from equation 47.

And we recall that $\tau := \frac{16\kappa_s\varepsilon_F}{(\alpha-1)}$, which implies that:

$$
\tau\frac{32\alpha^3 L_{s'}\kappa_s^2\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{2\gamma\kappa_s L_{s'}}\right)}{\varepsilon} = \mathcal{O}\left(\frac{\varepsilon_F}{\varepsilon}\left(\kappa_s^3 L_{s'}+\frac{\kappa_s}{\nu_s}\right)\right).
$$

Therefore, overall, the # IZO complexity is in $\mathcal{O}\left(\frac{\varepsilon_F}{\varepsilon}\kappa_s^3 L_{s'}\right)$.

□

E.3.3 Proof of Theorem 3

Using the results above, we can now proceed to the proof of Theorem 3.

Proof. Let us denote for simplicity: $C_1 := \frac{\varepsilon_F}{q_t}$, $C_2 := \frac{\varepsilon_{abs}}{q_t}$, and $C_3 := \varepsilon_\mu \mu^2$. Moreover, let us denote $F := \text{supp}(\boldsymbol{w}_t) \cup \text{supp}(\boldsymbol{w}_{t-1}) \cup \text{supp}(\bar{\boldsymbol{w}})$, where supp denotes the support of a vector, i.e. the set of coordinates of its non-zero components. Note that therefore we have $|F| \leq 2k + \bar{k} \leq 3k$. In addition $[\boldsymbol{u}]_F$ denotes the thresholding of $\boldsymbol{u}$ to the support F, that is, the vector $\boldsymbol{u}$ with its components that are not in F set to 0. Since R is $L_{s'}$-RSS', with $s' = \max(s_2, s)$, R is also s-RSS' and s_2-RSS', with Lipschitz constant $L_{s'}$.

Denote $\boldsymbol{v}_t = \mathcal{H}_k(\boldsymbol{w}_{t-1} - \eta\nabla R(\boldsymbol{w}_{t-1}))$ for any $t \in \mathbb{N}$. The fact that R is $(L_{s'}, s')$-RSS', therefore also $(L_{s'}, s)$-RSS', implies from the remark in Assumption 5 that it is also $(L_{s'}, s)$-RSS, therefore:

$$
\begin{aligned}
&R(\boldsymbol{w}_t) \\
\leq& R(\boldsymbol{w}_{t-1}) + \langle\nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 \\
=& R(\boldsymbol{w}_{t-1}) + \langle\boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle \\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\boldsymbol{w}_t - (\boldsymbol{w}_{t-1} - \eta\boldsymbol{g}_{t-1})\|^2 - \eta^2\|\boldsymbol{g}_{t-1}\|^2 - \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2\right] + \frac{L_s}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 \\
&+ \langle\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle \\
=& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\boldsymbol{w}_t - (\boldsymbol{w}_{t-1} - \eta\boldsymbol{g}_{t-1})\|^2 - \frac{\eta}{2}\|\boldsymbol{g}_{t-1}\|^2 + \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 \\
&+ \langle[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle \\
\overset{(a)}{\leq}& R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\bar{\boldsymbol{w}} - (\boldsymbol{w}_{t-1} - \eta\boldsymbol{g}_{t-1})\|^2 - \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \sqrt{\beta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2\right] - \frac{\eta}{2}\|\boldsymbol{g}_{t-1}\|^2 \\
&+ \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle
\end{aligned}
$$

$$
\begin{aligned}
=&R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 + \eta^2\|\boldsymbol{g}_{t-1}\|^2 - 2\langle \eta \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle\right] - \frac{1}{2\eta}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\sqrt{\beta}}{2\eta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 \\
&- \frac{\eta}{2}\|\boldsymbol{g}_{t-1}\|^2 + \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle [\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle \\
=&R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\left[\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - 2\langle \eta \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle\right] - \frac{1}{2\eta}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\sqrt{\beta}}{2\eta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 \\
&+ \left[\frac{L_s - \frac{1}{\eta}}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \langle [\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle \\
\overset{(b)}{=}&R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle - \frac{1}{2\eta}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\sqrt{\beta}}{2\eta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 \\
&+ \left[\frac{L_s - \frac{1}{\eta} + C}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \frac{1}{2C}\|[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F\|^2 \\
=&R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle + \langle \nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&- -\frac{1}{2\eta}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 + \frac{\sqrt{\beta}}{2\eta}\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 + \left[\frac{L_{s'} - \frac{1}{\eta} + C}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2 + \frac{1}{2C}\|[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F\|^2,
\end{aligned}
$$

where (a) follows from Lemma 3 and (b) follows from the inequality $\langle a, b\rangle \leq \frac{C}{2}a^2 + \frac{1}{2C}b^2$, for any $(a, b) \in (\mathbb{R}^d)^2$ with $C > 0$ an arbitrary strictly positive constant.

Let us now assume that $\eta := \frac{1}{L_{s'}+C}$: therefore the term $\left[\frac{L_{s'} - \frac{1}{\eta} + C}{2}\right]\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2$ above is 0. We now take the conditional expectation (conditioned on w_{t-1}, which is the random variable which realizations are $\boldsymbol{w}_{t-1}$), on both sides, and from Lemma 4 we obtain the inequality below (we slightly abuse notations and denote $\mathbb{E}[\cdot|\mathrm{w}_{t-1} = \boldsymbol{w}_{t-1}]$ by $\mathbb{E}[\cdot|\boldsymbol{w}_{t-1}]$):

$$
\begin{aligned}
&\mathbb{E}[R(\boldsymbol{w}_t)|\boldsymbol{w}_{t-1}] \\
\leq&R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&- \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right] + \langle [\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E}\left[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}\right]]_F, \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&+ \mathbb{E}\left[\frac{1}{2C}\|[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F\|^2|\boldsymbol{w}_{t-1}\right] \\
\overset{(a)}{\leq}&R(\boldsymbol{w}_{t-1}) + \frac{1}{2\eta}\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&- \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right] + \\
&\frac{G}{2}\|[\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2 + \frac{1}{2G}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 + \frac{1}{2C}\mathbb{E}\left[\|\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}\|^2|\boldsymbol{w}_{t-1}\right] \\
=&R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
&- \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right] + \frac{G}{2}\|[\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2 \\
&+ \frac{1}{2C}\mathbb{E}\left[\|[\nabla R(\boldsymbol{w}_{t-1}) - \boldsymbol{g}_{t-1}]_F\|^2|\boldsymbol{w}_{t-1}\right] \\
\overset{(b)}{\leq}&R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle
\end{aligned}
$$

$$
\begin{aligned}
& - \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}\|[\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2 \\
& + \frac{1}{2C}\left(2\|[\nabla R(\boldsymbol{w}_{t-1}) - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2 + 2\|[\boldsymbol{g}_{t-1} - \mathbb{E}[\boldsymbol{g}_{t-1}|\boldsymbol{w}_{t-1}]]_F\|^2\right) \\
\overset{(33)+(35)}{\leq} & R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
& - \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 \\
& + \frac{1}{2C}\left(2C_3 + 2C_1\|\nabla R(\boldsymbol{w}_{t-1})\|^2 + 2C_2\mu^2\right) \\
\overset{(c)}{\leq} & R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
& - \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 \\
& + \frac{1}{2C}\left(2C_1\left(2\|\nabla R(\boldsymbol{w}_{t-1}) - \nabla R(\bar{\boldsymbol{w}})\|^2 + 2\|\nabla R(\bar{\boldsymbol{w}})\|^2\right) + 2C_2\mu^2 + 2C_3\right) \\
\overset{(d)}{\leq} & R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
& - \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 \\
& + \frac{1}{2C}\left(2C_1\left(2L_{s'}^2\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2 + 2\|\nabla R(\bar{\boldsymbol{w}})\|^2\right) + 2C_2\mu^2 + 2C_3\right) \\
= & R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G} + \frac{2C_1L_{s'}^2}{C}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 - \langle \nabla R(\boldsymbol{w}_{t-1}), \boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\rangle \\
& - \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2 + C_2\mu^2 + C_3\right) \\
\overset{(e)}{\leq} & R(\boldsymbol{w}_{t-1}) + \left[\frac{1}{2\eta} + \frac{1}{2G} + \frac{2C_1L_{s'}^2}{C}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 + \left[R(\bar{\boldsymbol{w}}) - R(\boldsymbol{w}_{t-1}) - \frac{\nu_s}{2}\|\boldsymbol{w}_{t-1} - \bar{\boldsymbol{w}}\|^2\right] \\
& - \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2 + C_2\mu^2 + C_3\right) \\
= & R(\bar{\boldsymbol{w}}) + \left[\frac{\frac{1}{\eta} - \nu_s}{2} + \frac{1}{2G} + \frac{2C_1L_{s'}^2}{C}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 \\
& - \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2 + C_2\mu^2 + C_3\right) \\
\overset{(f)}{\leq} & R(\bar{\boldsymbol{w}}) + \left[\frac{\frac{1}{\eta} - \nu_s}{2} + \frac{1}{2G} + \frac{2\varepsilon_F L_{s'}^2}{\tau C}\right]\|\bar{\boldsymbol{w}} - \boldsymbol{w}_{t-1}\|^2 \\
& - \frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t - \bar{\boldsymbol{w}}\|^2 | \boldsymbol{w}_{t-1}\right] + \frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2 + C_2\mu^2 + C_3\right)
\end{aligned} \tag{48}
$$

Where (a) follows from the inequality $\langle a, b\rangle \leq \frac{G}{2}a^2 + \frac{1}{2G}b^2$, for any $(a, b) \in (\mathbb{R}^d)^2$ with $G > 0$ an arbitrary strictly positive constant, (b) and (c) follow from the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any $(a, b) \in (\mathbb{R}^d)^2$, (d) follows from the fact that R is $(L_{s'}, s')$-RSS' (Assumption 5 with sparsity level s'), therefore it is also $(L_{s'}, s)$-RSS', (e) follows from the RSC condition, and for (f), we recall that $C_1 = \frac{\varepsilon_F}{q_t}$, and we define $q_t = \left\lceil \frac{\tau}{\omega^t} \right\rceil$, for some $\omega > 1$ and $\tau > 0$ that will be chosen later in the proof.

Recall that we have chosen $\eta := \frac{1}{L_{s'}+C}$. Let us define $\alpha := \frac{C}{L_{s'}} + 1$. Then $C = (\alpha - 1)L_{s'}$, and $\eta = \frac{1}{\alpha L_{s'}}$. Also recall that $\kappa_s = \frac{L_{s'}}{\nu_s}$.

We will now choose the constant G and C, in order to simplify the inequality above, such that it matches as much as possible the structure of the previous proofs:

We will seek to rewrite:

$\left[\frac{\frac{1}{\eta}-\nu_s}{2}+\frac{1}{2G}+\frac{2\frac{\varepsilon_F}{\tau}L_{s'}^2}{C}\right]\left(=\frac{1}{2\eta}\left[1+\frac{1}{G\alpha L_{s'}}+\frac{4L_{s'}^2\frac{\varepsilon_F}{\tau}}{(\alpha-1)\alpha L_{s'}^2}-\frac{1}{\alpha\kappa_s}\right]\right)$, into :

$\frac{1}{2\eta}\left[1-\frac{1}{\alpha'\kappa_s}\right]$ for some $\alpha'>0$ (we will seek $\alpha'\propto\alpha$, with a dimensionless proportionality constant for simplicity).

Therefore, let us choose $G:=\frac{4}{\nu_s}$, which implies:

$$\frac{1}{G\alpha L_{s'}}=\frac{1}{4\alpha\kappa_s}. \tag{49}$$

And let us choose $\tau:=\frac{16\kappa_s\varepsilon_F}{(\alpha-1)}$, which implies:

$$\frac{4L_{s'}^2\frac{\varepsilon_F}{\tau}}{(\alpha-1)\alpha L_{s'}^2}=\frac{1}{4\alpha\kappa_s}. \tag{50}$$

Therefore, using equations 49 and 50, we obtain:

$$\begin{aligned}\left[\frac{\frac{1}{\eta}-\nu_s}{2}+\frac{1}{2G}+\frac{2\frac{\varepsilon_F}{\tau}L_{s'}^2}{C}\right]&=\frac{1}{2\eta}\left[1+\frac{1}{G\alpha L_{s'}}+\frac{4L_{s'}^2\frac{\varepsilon_F}{\tau}}{(\alpha-1)\alpha L_{s'}^2}-\frac{1}{\alpha\kappa_s}\right]\\&=\frac{1}{2\eta}\left[1+\frac{1}{4\alpha\kappa_s}+\frac{1}{4\alpha\kappa_s}-\frac{1}{\alpha\kappa_s}\right]\\&=\frac{1}{2\eta}\left[1-\frac{1}{2\alpha\kappa_s}\right]=\frac{1}{2\eta}\left[1-\frac{1}{\alpha'\kappa_s}\right],\end{aligned}$$

where for simplicity we denote $\alpha'=2\alpha$.

We can therefore simplify (48) into:

$$\begin{aligned}\mathbb{E}[R(\boldsymbol{w}_t)|\boldsymbol{w}_{t-1}]-R(\bar{\boldsymbol{w}})\leq&\frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-\frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]\right.\\&+\frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2|\boldsymbol{w}_{t-1}\right]\\&\left.+2\eta\left(\frac{G}{2}C_3+\frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2+C_2\mu^2+C_3\right)\right)\right].\end{aligned}$$

We now take the expectation over w_{t-1} of the above inequality (i.e. we take $\mathbb{E}_{\mathrm{w}_{t-1}}[\cdot]$): using the law of total expectation ($\mathbb{E}[\cdot]=\mathbb{E}_{\mathrm{w}_{t-1}}[\mathbb{E}[\cdot|\boldsymbol{w}_{t-1}]]$) we obtain:

$$\begin{aligned}\mathbb{E}R(\boldsymbol{w}_t)-R(\bar{\boldsymbol{w}})\leq&\frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-\frac{1}{2\eta}\mathbb{E}\left[\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right]\right. &(51)\\&+\frac{\sqrt{\beta}}{2\eta}\mathbb{E}\left[\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2\right] &(52)\\&\left.+2\eta\left(\frac{G}{2}C_3+\frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2+C_2\mu^2+C_3\right)\right)\right]. &(53)\end{aligned}$$

Let us call $A:=2\eta\left(\frac{G}{2}C_3+\frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2+C_2\mu^2+C_3\right)\right)$ for simplicity.

This gives:

$$\mathbb{E}R(\boldsymbol{w}_t)-R(\bar{\boldsymbol{w}})\leq\frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2-\frac{1}{2\eta}\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2+\frac{\sqrt{\beta}}{2\eta}\mathbb{E}\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2+A\right]. \tag{54}$$

Additionally, in view of equation 40 applied at $\boldsymbol{v}_t$ instead of $\boldsymbol{w}_t$, (since $\boldsymbol{v}_t$ here corresponds to the $\boldsymbol{w}_t$ from Section E.1.2, i.e. $\boldsymbol{v}_t$ is the hard-thresholding of an iterate after a gradient step), we know that:

$$\mathbb{E}R(\boldsymbol{v}_t) - R(\bar{\boldsymbol{w}}) \leq \frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2 + A\right].$$

We now take a convex combination similarly as in the case without additional constraint (section D.2), for some $\rho \in (0,1)$.

$$\begin{aligned}
&\mathbb{E}(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t)\\
\leq &R(\bar{\boldsymbol{w}}) + \frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2 - (1-\rho)\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right.\\
&\left. + \left((1-\rho)\sqrt{\beta} - (1-\sqrt{\beta})\rho\right)\mathbb{E}\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2 + A\right]\\
= &R(\bar{\boldsymbol{w}}) + \frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2 - (1-\rho)\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right.\\
&\left. - \left(\rho - \sqrt{\beta}\right)\mathbb{E}\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2 + A\right]\\
\overset{(b)}{\leq} &R(\bar{\boldsymbol{w}}) + \frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2 - (1-\rho)\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right.\\
&\left. - \left(\rho - \sqrt{\beta}\right)\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2 + A\right]\\
= &R(\bar{\boldsymbol{w}}) + \frac{1}{2\eta}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2 + A\right].
\end{aligned}$$

where in (b), we have assumed that $\sqrt{\beta} \leq \rho$ (later we will verify that our choice of k ensures such a condition), and have used the fact that projection onto a convex set is non-expansive (which implies that $\|\boldsymbol{v}_t-\bar{\boldsymbol{w}}\|^2 \geq \|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2$).

Similarly as in Liu & Foygel Barber (2020), we now take a weighted sum over $t = 1, ..., T$, to obtain:

$$\begin{aligned}
&\sum_{t=1}^{T} 2\eta\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\mathbb{E}[(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t) - R(\bar{\boldsymbol{w}})]\\
\leq &\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2 + A\right]\\
= &\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\left[\left(1-\frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2 - (1-\sqrt{\beta})\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right]\\
&+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}A\\
= &(1-\sqrt{\beta})\sum_{t=1}^{T}\left[\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t+1}\mathbb{E}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_{t-1}\|^2 - \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}\mathbb{E}\|\boldsymbol{w}_t-\bar{\boldsymbol{w}}\|^2\right]\\
&+\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}A\\
\overset{(a)}{=} &(1-\sqrt{\beta})\left[\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T}\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 - \mathbb{E}\|\boldsymbol{w}_T-\bar{\boldsymbol{w}}\|^2\right] + \sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}A
\end{aligned}$$

$$\begin{aligned}
&\leq (1-\sqrt{\beta})\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^T \|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} A \\
&\leq \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^T \|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} A \\
&= \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^T \|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta\left(\frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla R(\bar{\boldsymbol{w}})\|^2\right.\right. \\
&\qquad \left.\left. + C_2\mu^2 + C_3\right)\right) \\
&= \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^T \|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta\left(\frac{G}{2}C_3 + \frac{1}{C}\left(2\frac{\varepsilon_F}{q_t}\|\nabla R(\bar{\boldsymbol{w}})\|^2\right.\right. \\
&\qquad \left.\left. + \frac{\varepsilon_{abs}\mu^2}{q_t} + C_3\right)\right) \\
&= \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^T \|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \frac{2\eta}{q_t}\left(\frac{2\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2 + \varepsilon_{abs}\mu^2}{C}\right) \\
&\quad + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta C_3\left(\frac{G}{2}+\frac{1}{C}\right) \\
&= \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^T \|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \frac{2\eta}{q_t}\left(\frac{2\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}\right) \\
&\quad + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta\mu^2\left(\varepsilon_\mu\left(\frac{G}{2}+\frac{1}{C}\right) + \frac{\varepsilon_{abs}}{Cq_t}\right) \\
&\leq \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^T \|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2 + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \frac{2\eta}{q_t}\left(\frac{2\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}\right) \\
&\quad + \sum_{t=1}^T \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta\mu^2\left(\varepsilon_\mu\left(\frac{G}{2}+\frac{1}{C}\right) + \frac{\varepsilon_{abs}}{C}\right),
\end{aligned} \tag{55}$$

where (a) follows from simplifying the telescopic sum. Let us denote for simplicity $\zeta := \frac{2\eta(2\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2)}{C} = \frac{4\eta\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}$ and $Z := \varepsilon_\mu\left(\frac{G}{2}+\frac{1}{C}\right) + \frac{\varepsilon_{abs}}{C}$.

We now choose k and s_t as follows: we choose $k \geq 4\frac{\alpha'^2}{\rho}\kappa_s^2\bar{k}$, which implies that:

$$\begin{aligned}
&\sqrt{\beta} \leq \frac{1}{2\frac{\alpha'}{\rho}\kappa_s} \\
&\implies \sqrt{\beta} \leq \frac{1}{2\frac{\alpha'}{\rho}\kappa_s - 1} \\
&\implies 1-\sqrt{\beta} \geq 1 - \frac{1}{2\frac{\alpha'}{\rho}\kappa_s - 1} = \frac{2\frac{\alpha'}{\rho}\kappa_s - 2}{2\frac{\alpha'}{\rho}\kappa_s - 1} = \frac{1-\frac{1}{\frac{\alpha'}{\rho}\kappa_s}}{1-\frac{1}{2\frac{\alpha'}{\rho}\kappa_s}} \\
&\implies \left(\frac{1-\frac{1}{\frac{\alpha'}{\rho}\kappa_s}}{1-\sqrt{\beta}}\right) \leq 1 - \frac{1}{2\frac{\alpha'}{\rho}\kappa_s}.
\end{aligned} \tag{56}$$

We recall that we previously defined $q_t = \left\lceil \frac{\tau}{\omega^t} \right\rceil$, with $\tau = 16\kappa_s\frac{\varepsilon_F}{(\alpha-1)}$. We now set the value of ω, to $\omega := 1 - \frac{1}{\frac{\alpha'}{\rho}\kappa_s}$.

Let us call $\nu := 1 - \frac{1}{2\frac{\alpha'}{\rho}\kappa_s}$. Note that we have:

$$\nu \leq \omega. \tag{57}$$

And that we have the inequality below:

$$\frac{\nu}{\omega} = \frac{1 - \frac{1}{2\frac{\alpha'}{\rho}\kappa_s}}{1 - \frac{1}{4\frac{\alpha'}{\rho}\kappa_s}} = \frac{4\frac{\alpha'}{\rho}\kappa_s - 2}{4\frac{\alpha'}{\rho}\kappa_s - 1} = 1 - \frac{1}{4\frac{\alpha'}{\rho}\kappa_s - 1} \leq 1 - \frac{1}{4\frac{\alpha'}{\rho}\kappa_s} = \omega. \tag{58}$$

This allows us to simplify equation 55 into:

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T} 2\eta \left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} [(1-\rho)R(\boldsymbol{w}_t) + \rho R(\boldsymbol{v}_t) - R(\bar{\boldsymbol{w}})] \\
&\leq \nu^T \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \sum_{t=1}^{T} \nu^{T-t}\omega^{t-1} + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&= \nu^T \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{\omega^T}{\omega}\sum_{t=1}^{T}\left(\frac{\nu}{\omega}\right)^{T-t} + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&= \nu^T \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{\omega^T}{\omega}\frac{1-\left(\frac{\nu}{\omega}\right)^T}{1-\left(\frac{\nu}{\omega}\right)} + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&\leq \nu^T \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{\omega^T}{\omega}\frac{1}{1-\left(\frac{\nu}{\omega}\right)} + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&\overset{(a)}{\leq} \nu^T \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{\omega^T}{\omega}\frac{1}{1-\omega} + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&\overset{(b)}{\leq} \nu^T \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\omega^T\frac{1}{1-\omega} + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&\overset{(c)}{\leq} \omega^T \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\omega^T\frac{1}{1-\omega} + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&\overset{(d)}{\leq} \frac{\omega^T}{1-\omega} \|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\omega^T\frac{1}{1-\omega} + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&= \frac{\omega^T}{1-\omega}\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right) + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2 \\
&= 4\frac{\alpha'}{\rho}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3}\right) + \frac{\zeta}{\tau}\sum_{t=1}^{T}\left(\frac{1 - \frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} 2\eta Z\mu^2,
\end{aligned}$$

where in the left hand side we have used the linearity of expectation, and where (a) uses equation 58, (b) uses the fact that $\frac{1}{\omega} = \frac{1}{1-\frac{1}{4\frac{\alpha'}{\rho}\kappa_s}} \leq \frac{1}{1-\frac{1}{4}} = \frac{4}{3}$ (since $\kappa_s \geq 1$ and $\alpha' \geq 1$ (indeed, we have $\alpha' = 2\alpha = 2(\frac{C}{L_{s'}} + 1)$ with $C > 0$), so consequently $\frac{\alpha'}{\rho} \geq 1$), (c) uses equation 57, and (d) uses the fact that $\omega < 1$ so $1 < \frac{1}{1-\omega}$.

Let us now normalize the above inequality:

$$\mathbb{E}\frac{\sum_{t=1}^{T} 2\eta \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}[(1-\rho)R(\boldsymbol{w}_t)+\rho R(\boldsymbol{v}_t)]}{\sum_{t=1}^{T} 2\eta \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}} \leq R(\bar{\boldsymbol{w}})+\frac{4\frac{\alpha'}{\rho}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\zeta}{\tau}\right)}{\sum_{t=1}^{T} 2\eta \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t}}+Z\mu^2.$$

The left hand side above is a weighted sum, which is an upper bound on the smallest term of the sum.

Regarding the right hand side, we can simplify it using the fact that $0 < \left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)$, and therefore:

$$\sum_{t=1}^{T}\left(\frac{1-\frac{1}{\alpha'\kappa_s}}{1-\sqrt{\beta}}\right)^{T-t} \geq 1.$$

Therefore, we obtain:

$$\begin{aligned}\mathbb{E}\min_{t\in\{1,..,T\}}[(1-\rho)R(\boldsymbol{w}_t)+\rho R(\boldsymbol{v}_t)-R(\bar{\boldsymbol{w}})] &\leq \frac{4\frac{\alpha'}{\rho}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\zeta}{\tau}\right)}{2\eta}+Z\mu^2\\ &= 4\frac{\alpha^2}{\rho}L_{s'}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\zeta}{\tau}\right)+Z\mu^2,\end{aligned}$$

which can be simplified into the expression below, using the definition of $\hat{\boldsymbol{w}}_T$:

$$\mathbb{E}[\min_{t\in[T]}(1-\rho)R(\boldsymbol{w}_t)+\rho R(\boldsymbol{v}_t)-R(\bar{\boldsymbol{w}})] \leq 4\frac{\alpha^2}{\rho}L_{s'}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\zeta}{\tau}\right)+Z\mu^2. \quad (59)$$

To simplify the above result, we recall the assumptions made earlier on: we have chosen
$\tau=\frac{16\kappa_s\varepsilon_F}{(\alpha-1)}$, and $G=\frac{4}{\nu_s}$.

Therefore, to sum up, we have:

$$Z=\varepsilon_\mu\left(\frac{G}{2}+\frac{1}{C}\right)+\frac{\varepsilon_{abs}}{C}=\varepsilon_\mu\left(\frac{2}{\nu_s}+\frac{1}{C}\right)+\frac{\varepsilon_{abs}}{C}$$
$$\omega=1-\frac{1}{4\frac{\alpha'}{\rho}\kappa_s}=1-\frac{1}{8\frac{\alpha}{\rho}\kappa_s}$$

$$\zeta=\frac{4\eta\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}$$

The last inequality implies: $\frac{\zeta}{\tau}=\frac{\frac{4\eta\varepsilon_F\|\nabla R(\bar{\boldsymbol{w}})\|^2}{C}}{16\kappa_s L_{s'}\frac{\varepsilon_F}{C}}=\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}$.

Let us denote by ε_T the right-hand side term from equation 59:

$$\varepsilon_T=4\frac{\alpha^2}{\rho}L_{s'}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)+Z\mu^2.$$

We now proceed similarly as in the proof of Theorem 2 above. Recall that we have assumed in the Assumptions of Theorem 3, without loss of generality, that R is non-negative, which implies that $R(\boldsymbol{v}_t)\geq 0$. Plugging this in equation 59 implies that:

$$\mathbb{E}\min_{t\in[T]}R(\boldsymbol{w}_t)\leq\frac{1}{1-\rho}R(\bar{\boldsymbol{w}})+\frac{\varepsilon_T}{1-\rho}+\frac{Z}{(1-\rho)}\mu^2\leq(1+2\rho)R(\bar{\boldsymbol{w}})+\frac{\varepsilon_T}{1-\rho}+\frac{Z}{1-\rho}\mu^2. \quad (60)$$

Plugging the change of variable $\varepsilon'_T = \frac{\varepsilon_T}{1-\rho}$ into equation 60 above, and redefining Z into $Z := \frac{1}{1-\rho}\left(\varepsilon_\mu\left(\frac{2}{\nu_s}+\frac{1}{C}\right)+\frac{\varepsilon_{abs}}{C}\right)$, we obtain that:

$$\mathbb{E}\min_{t\in[T]} R\left(\boldsymbol{w}_t\right) \leq (1+2\rho)R(\bar{\boldsymbol{w}}) + \varepsilon'_T + Z\mu^2.$$

Further, consider an ideal case where $\bar{\boldsymbol{w}}$ is a global minimizer of R over $\mathcal{B}_0(k) := \{\boldsymbol{w} : \|\boldsymbol{w}\|_0 \leq k\}$. Then $R\left(\boldsymbol{v}_t\right) \geq R(\bar{\boldsymbol{w}})$ is always true for all $t \geq 1$. It follows that the bound in equation 59 yields:

$$\mathbb{E}\min_{t\in[T]}\left\{(1-\rho)R\left(\boldsymbol{w}_t\right)+\rho R(\bar{\boldsymbol{w}})\right\} \leq \mathbb{E}\min_{t\in[T]}\left\{(1-\rho)R\left(\boldsymbol{w}_t\right)+\rho R\left(\boldsymbol{v}_t\right)\right\} \leq R(\bar{\boldsymbol{w}}) + \varepsilon_T,$$

which implies: $\mathbb{E}\min_{t\in[T]} R\left(\boldsymbol{w}_t\right) \leq R(\bar{\boldsymbol{w}}) + \frac{\varepsilon_T}{1-\rho}$. In this case, we can simply set $\rho = 0.5$, and define $\varepsilon'_T = \frac{\varepsilon_T}{1-\rho} = 2\varepsilon_T$ similarly as above. The proof is completed.

□

E.4 Proof of Corollary 2

Proof. Let $\varepsilon \in \mathbb{R}_+^*$. Let us find T to ensure that $\mathbb{E}\min_{t\in\{1,..,T\}}(1-\rho)R(\boldsymbol{w}_t)+\rho R(\boldsymbol{v}_t) - R(\bar{\boldsymbol{w}}) \leq \varepsilon + Z\mu^2$

This will be enforced if:

$$\begin{aligned}
&4\alpha^2\frac{1}{\rho}L_{s'}\kappa_s\omega^T\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)\leq\varepsilon\\
&\iff T\log(\omega)\leq\log\left(\frac{\varepsilon}{4\alpha^2\frac{1}{\rho}L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}\right)\\
&\iff T\geq\frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{4\alpha^2\frac{1}{\rho}L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right).
\end{aligned}$$

Therefore, let us take:

$$T := \left\lceil\frac{1}{\log(\frac{1}{\omega})}\log\left(\frac{4\alpha^2\frac{1}{\rho}L_{s'}\kappa_s\left(\|\bar{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{4}{3}\frac{\eta\|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right)\right\rceil. \tag{61}$$

We can now derive the #IZO and #HT. First, we have one hard-thresholding operation at each iteration, therefore #HT$=T$. Using the fact that $\frac{1}{\log(\frac{1}{\omega})} = \frac{1}{-\log(\omega)} = \frac{1}{-\log(1-\frac{1}{8\alpha\frac{1}{\rho}\kappa_s})} \leq \frac{1}{\frac{1}{8\alpha\frac{1}{\rho}\kappa_s}} = 8\alpha\frac{1}{\rho}\kappa_s$ (since by property of the logarithm, for all $x\in(-\infty,-1): \log(1-x)\leq -x$), and the fact that $\alpha = \frac{C}{L_{s'}}$ is independent of κ_s, we obtain that #HT $= \mathcal{O}(\kappa_s\log\left(\frac{1}{\varepsilon}\right))$.

We now turn to computing the #IZO. At each iteration t we have q_t function evaluations, therefore:

$$\begin{aligned}
\#\text{IZO} &= \sum_{t=0}^{T-1} q_t\\
&\leq \sum_{t=0}^{T-1}\left(\frac{\tau}{\omega^t}+1\right)\\
&= T + \tau\frac{\left(\frac{1}{\omega}\right)^T-1}{\frac{1}{\omega}-1}\\
&\leq T + \frac{\tau}{\frac{1}{\omega}-1}\left(\frac{1}{\omega}\right)^T
\end{aligned}$$

$$
\begin{aligned}
&= T + \frac{\tau}{\frac{1}{\omega} - 1} \exp\left(T \log\left(\frac{1}{\omega}\right)\right) \\
&\overset{(a)}{\leq} 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{4\alpha^2 \frac{1}{\rho} L_{s'} \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \frac{\eta \|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right) \\
&\quad + \frac{\tau}{\frac{1}{\omega} - 1} \exp\left(\log\left(\frac{1}{\omega}\right) \left[\frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{4\alpha^2 \frac{1}{\rho} L_{s'} \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \frac{\eta \|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right)\right.\right. \\
&\quad \left.\left. +1\right]\right) \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{4\alpha^2 \frac{1}{\rho} L_{s'} \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \frac{\eta \|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right) \\
&\quad + \frac{\frac{\tau}{\omega}}{\frac{1}{\omega} - 1} \frac{4\alpha^2 \frac{1}{\rho} L_{s'} \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \frac{\eta \|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon} \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{4\alpha^2 \frac{1}{\rho} L_{s'} \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \frac{\eta \|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right) \\
&\quad + \frac{\tau}{1 - \omega} \frac{4\alpha^2 \frac{1}{\rho} L_{s'} \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \frac{\eta \|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon} \\
&= 1 + \frac{1}{\log(\frac{1}{\omega})} \log\left(\frac{4\alpha^2 \frac{1}{\rho} L_{s'} \kappa_s \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \frac{\eta \|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon}\right) \\
&\quad + \tau \frac{32\alpha^3 \frac{1}{\rho^2} L_{s'} \kappa_s^2 \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \frac{\eta \|\nabla R(\bar{\boldsymbol{w}})\|^2}{4\kappa_s L_{s'}}\right)}{\varepsilon},
\end{aligned}
$$

where (a) follows from equation 61.

And we recall that $\tau = 16\kappa_s \frac{\varepsilon_F}{(\alpha - 1)}$, which implies that:

$$
\tau \frac{32\alpha^3 \frac{1}{\rho^2} L_{s'} \kappa_s^2 \left(\|\bar{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 + \frac{4}{3} \frac{\eta \|\nabla R(\bar{\boldsymbol{w}})\|^2}{2\gamma\kappa_s L_{s'}}\right)}{\varepsilon} = \mathcal{O}\left(\frac{\varepsilon_F}{\varepsilon}\left(\kappa_s^3 L_{s'} + \frac{\kappa_s}{\nu_s}\right)\right).
$$

Therefore, overall, the IZO (query complexity) is in $\mathcal{O}\left(\frac{\varepsilon_F}{\varepsilon} \kappa_s^3 L_{s'}\right)$. The proof is completed. $\square$

F Additional Experiments

F.1 Additional results and details for the index tracking problem

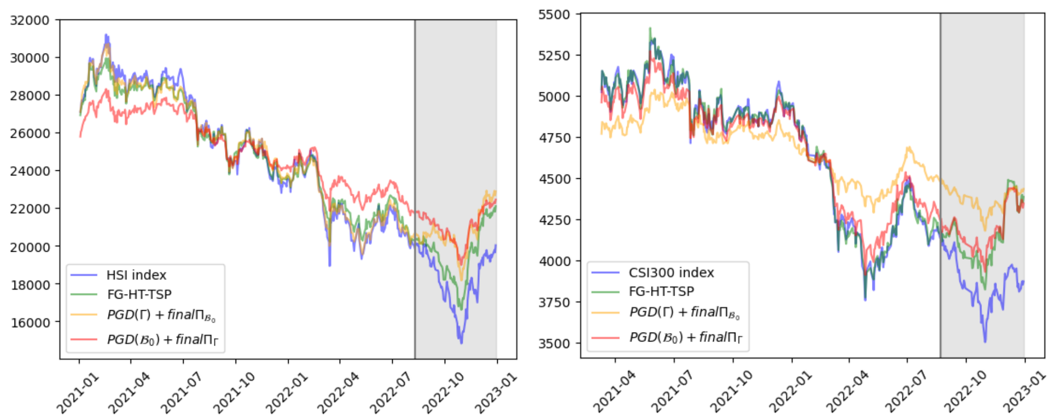
In section 5, we presented the performance of an index tracking strategy based on FG-HT-TSP, for the S&P500 index. In this Appendix, we also present the performance of the index tracking strategy on two additional indices: the CSI300 index in Figure 3a, and the HSI index in Figure 3b, over the same time period for HSI, and for CSI300 we start the period in March 2021 due to missing values. We keep the constraint $k = 15$ for both indices, and enforce a constraint on sector transactions of $D = 100$ for CSI300 and $D = 1000$ for HSI. We provide in Table 2 below the respective dimensions of the train-sets used for the experiments (which constitutes, as we recall, 80% of the total dataset).

After running FG-HT-TSP, the obtained weight vector for the CSI300 index spans 7 sectors out of the 10 total sectors of the index, and the one for the HSI index spans 3 of the 4 sectors of the index, which validates the diversifying effect of the enforced constraint. Additionally, we can observe that the index strategy based on FG-HT-TSP can more successfully track the true index than the two baselines.

INDEX	n	d
S&P500	402	497
CSI300	353	291
HSI	394	72

Table 2: Number of samples (n) and dimension (d) of the training sets for the index tracking experiment

The data for those three indices is scrapped from the web using the `beautifulsoup`[1] library to gather information about the index, and the `yfinance`[2] library to scrap the returns of such stocks during the considered time period.



(a) HSI (b) CSI300

Figure 3: Index tracking with sector constraints on HSI and CSI300

On the verification of Assumptions 1 to 3: Note that such index tracking experiments verify Assumptions 1, 2 and 3:

- **Assumption 1** is verified since the cost function is quadratic, with a design matrix of size $n > d$ (except in the case of S&P500). As can be expected with such matrices in general, the Hessian $\boldsymbol{H} = 2\boldsymbol{A}^\top \boldsymbol{A}$ is positive-definite (we have indeed verified in our code that it is). Therefore the RSC constant is bounded below by $\lambda_{\min}$ where $\lambda_{\min}$ is the smallest eigenvalue of $2\boldsymbol{A}^\top\boldsymbol{A}$. Note that for S&P500, strong convexity is not verified since $d > n$: however, since we take $k = 15$, with high probability (i.e. unless we can find $s = 2k = 30$ columns of $\boldsymbol{A}$ that are exactly linearly dependent), RSC should be verified.

- **Assumption 2 and Assumption 5** are both verified since the cost function is quadratic, therefore the (strong) RSS constant is bounded above by $2\|\boldsymbol{A}\|_s^2$, where $\|\cdot\|_s$ denotes the spectral norm.

- **Assumption 3** is verified since projection onto Γ can be done group-wise, and for each group the projection is onto an ℓ_1 ball, which is a convex symmetric set (which is support-preserving from Remark 1), therefore, overall, Γ is support-preserving).

[1] `https://pypi.org/project/beautifulsoup4/`
[2] `https://github.com/ranaroussi/yfinance`

F.2 Comparison between two-step projection and Euclidean Projection

F.2.1 Recall on the differences between two-step projection and Euclidean projection

In this section, we recall the differences between the two-step projection and the Euclidean projection onto the mixed constraints $\Gamma \cap \mathcal{B}_0(k)$. As described in the paper, one can encounter several possible cases:

- **Case (i):** the two-step projection (TSP) and the Euclidean projection onto $\Gamma \cap \mathcal{B}_0(k)$ are identical (see e.g. Remark 2): in that case, the contribution of our paper are on the theoretical side: Theorems 1 2, 3 give global convergence guarantee which therefore in this case apply to the usual (non-convex) projected gradient descent algorithm with Euclidean projection.
- **Case (ii):** the TSP and the Euclidean projection onto the mixed constraints are different: this case can be declined into several sub-cases as described below:
  - Case (a): the Euclidean projection onto the mixed constraint $\Gamma \cap \mathcal{B}_0(k)$ is unknown (such as for the constraints Γ used in the experiments from Section 5): in that case, the TSP can allow to fill such gap, since the TSP only requires the knowledge of the projection onto Γ, which is often known and easy to do.
  - Case (b): the Euclidean projection onto the mixed constraint $\Gamma \cap \mathcal{B}_0(k)$ is known, but computationally expensive: in that case, the TSP can provide a simpler and faster alternative to the Euclidean projection, while still enjoying some convergence guarantees as shown in this paper.
  - Case (c): the Euclidean projection onto the mixed constraint $\Gamma \cap \mathcal{B}_0(k)$ is known and is efficient enough (e.g. when Γ belongs to the set of positive symmetric sets such as in Lu (2015)). In such cases, it is unclear whether the TSP can improve upon Euclidean projection since, at the iteration level, using the Euclidean projection is optimal (indeed, a (Euclidean) projected gradient descent step minimizes a quadratic upper bound on the objective value under constraints (derived from the smoothness of R)), and the TSP is therefore suboptimal in that sense (at the iteration level). *This is the case that we will analyze in this section*, in order to evaluate in practice the extend of such differences between TSP and Euclidean projection in such case.

F.2.2 Setting

As mentioned above, we analyze in more details the case (ii,c) above. We consider a simple synthetic linear regression setting with a correlated design matrix, i.e. where the design matrix $\boldsymbol{X}$ is formed by n i.i.d. samples from d (we take $d = 1000$, and $n = 5000$) correlated Gaussian random variables $\{X_1, .., X_d\}$ of zero mean and unit variance, such that:

$$\forall i \in \{1, \dots, d\} : \mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = 1;$$

$$\forall (i, j) \in \{1, \dots, d\}^2, i \neq j : \mathbb{E}[X_i X_j] = \rho^{|i-j|}.$$

More precisely, we generate each feature X_i in an auto-regressive manner, from previous features, using a correlation $\rho \in [0, 1)$, in the following way: we have $X_1 \sim \mathcal{N}(0, 1)$ and $\sigma^2 = 1 - \rho^2$, and for all $j \in \{2, ..., d\}$: $X_{j+1} = \rho X_j + \epsilon_j$ where $\epsilon_j = \sigma \Delta$, with $\Delta \sim \mathcal{N}(0, 1)$. Additionally, the data is generated from a vector $\boldsymbol{w}^*$ of k^*-sparse support sampled uniformly at random, with $k^* = 20$, and with each non-zero entry sampled from a normal distribution, and $\boldsymbol{y}$ is obtained with a noise vector $\boldsymbol{\epsilon}$ created from i.i.d. samples from a normal distribution, rescaled to enforce a given signal to noise ratio (SNR), as follows:

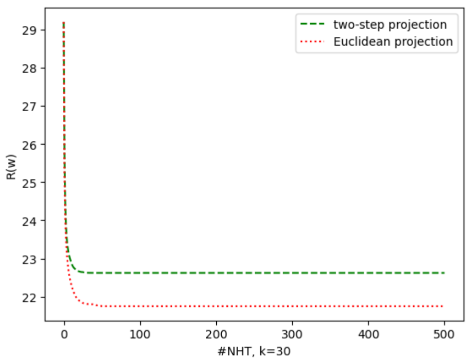
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{\epsilon}$$

with the signal to noise ratio defined as $\text{snr} = \frac{\|\boldsymbol{X}\boldsymbol{w}^*\|}{\|\boldsymbol{\epsilon}\|}$ (we choose $\text{snr} = 3$). We generate this dataset using the `make_correlated_data` function from the `benchopt` package Moreau et al. (2022). The problem that we solve is:

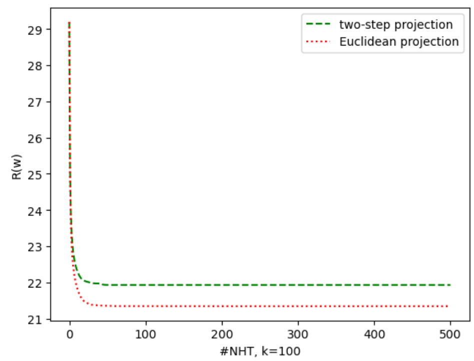
$$\min_{\boldsymbol{w} \in \Gamma \cap \mathbb{R}^d} \frac{1}{n} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

In such case, the Euclidean projection of $\boldsymbol{w} \in \mathbb{R}^d$ onto $\Gamma \cap \mathcal{B}_0(k)$ is given in Lu (2015), Beck & Hallak (2016), and consists in simply sorting the entries in $\boldsymbol{w}$, $(w_1, ..., w_d)$ (not in absolute value), keeping the k largest ones (and setting the others to 0) to obtain $\boldsymbol{w}'$ and then replacing each coordinate w'_i by $\max(0, w'_i)$. The two-step projection (TSP) in such case is simply hard-thresholding of $\boldsymbol{w}$ to obtain a vector $\boldsymbol{w}'$ followed by replacing each coordinate w'_i by $\max(0, w'_i)$).

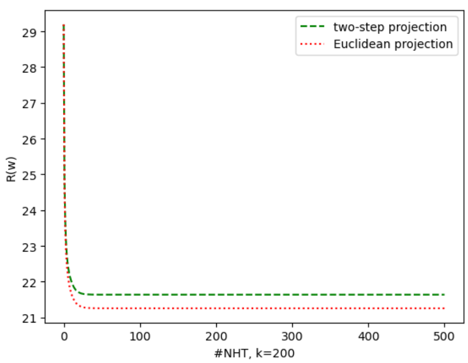
We plot the optimization curves for several values of k ($k \in \{30, 100, 200, 500, 800, 1000\}$ in Figure 4). In all curves, the learning rate is set to $1/L$ where L is the smoothness constant, equal to $\frac{2}{n}\|\boldsymbol{X}\|_s^2$ where $\|\boldsymbol{X}\|_s$ is the spectral norm of $\boldsymbol{X}$.



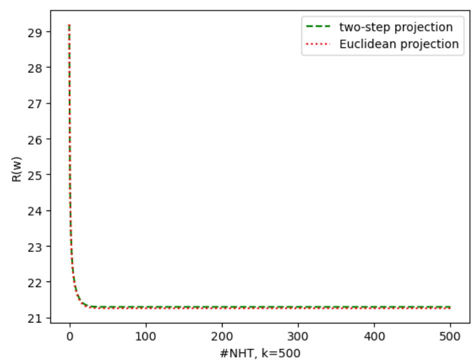
(a) $k = 30$



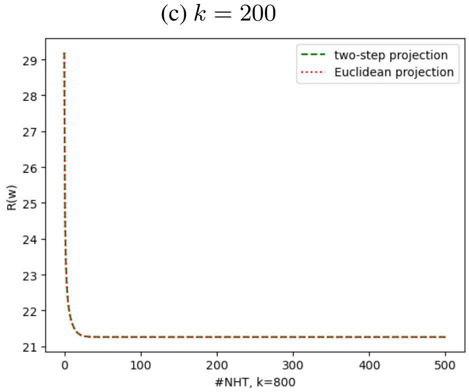
(b) $k = 100$



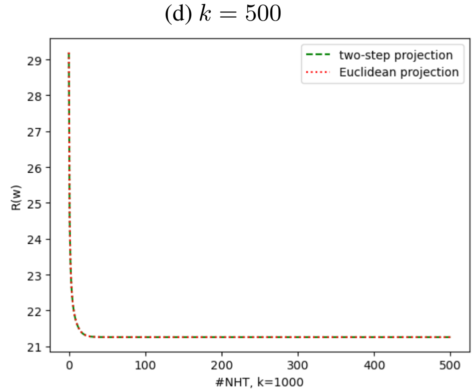
(c) $k = 200$



(d) $k = 500$



(e) $k = 800$



(f) $k = 1000$

Figure 4: Comparison of TSP vs. Euclidean projection for several k

F.2.3 Discussion

As we can observe in Figure 4, the Euclidean projection onto $\Gamma \cap \mathcal{B}_0(k)$ performs better in terms of objective value than the TSP in some cases. However, the gap between the two methods closes as the enforced sparsity of the iterates k increases. We interpret it in the following way. First, (non-convex) projected gradient descent (i.e. using Euclidean projection) is guaranteed to converge to a (non-convex constraints version of a) stationary point of the objective function (see e.g. Theorem 1 from Xu et al. (2019a)), whereas our method does not possess such guarantee (indeed, our guarantees are of the global kind: we give upper bounds on the objective value for the output of the algorithm), and therefore, the TSP may in some cases not converge to a stationary point, which may explain why Euclidean projection sometimes performs better than TSP. However, for larger k, in both cases the projections operators (TSP or Euclidean projection) become closer to a simple projection onto Γ (i.e. without sparsity constraints), which explains why as k grows, the gap between the two methods reduces. Finally, the improved performance of the TSP when k is larger is consistent with our Theorem 1, since for larger k, the upper bound on R from Theorem 1 can be made smaller, since considering larger k implies that ρ can be taken smaller as per Remark 3, reducing our upper bound on the objective value.

In conclusion, these results show that in case (ii) from Section F.2.1 above, the TSP introduced in this paper can be the most useful if the Euclidean projection onto $\Gamma \cap \mathcal{B}_0(k)$ is unknown, or too expensive computationally. Additionally, the gap between the two methods reduces if the enforced sparsity k of the iterates is large enough, or if the constraint forces iterates to stay close to 0.

F.3 Multiclass Logistic Regression

We consider the multiclass logistic regression problem with class group-wise ℓ_2 norm constraint as follows. We have $R_i(\boldsymbol{w}) = \sum_{j=1}^{c} \left[\frac{\lambda}{c} \|\boldsymbol{w}_j\|_2^2 - \mathbf{1}\{y_i = j\} \log \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{w}_j)}{\sum_{l=1}^{c} \exp(\boldsymbol{x}_i^\top \boldsymbol{w}_l)}\right]$, where $\boldsymbol{y}_i$ is the target output of $\boldsymbol{x}_i$, c is the number of classes, and $\boldsymbol{w}_j$ is the weight vector specific to class j. In addition to the sparsity constraint $\mathcal{B}_0(k)$, we enforce the following additional constraint $\Gamma = \{\boldsymbol{w} \in \mathbb{R}^d : \forall j \in [c] : \|\boldsymbol{w}_j\|_2 \leq D\}$, for some constant $D \in \mathbb{R}_+$, where $d = p \times c$, with p the number of features of the samples $\boldsymbol{x}_i$. More precisely, in such multiclass logistic regression, we seek to ensure an extra regularization not only on the whole global weight vector $\boldsymbol{w}$ (with the used squared ℓ_2 penalty), but also on each weight vector related to each class (through Γ), in order to prevent a potential class-wise overfitting.

Up to our knowledge, there is no known closed form for the Euclidean projection onto such $\Gamma \cap \mathcal{B}_0(k)$. However, the two-step projection (TSP) can be done easily: once the first projection is done (projection onto $\mathcal{B}_0(k)$, i.e. hard-thresholding) and the sparse support S is identified as per Section 3.1, the projection onto Γ restricted to S can be easily done since Γ is class-wise decomposable, and therefore it suffices to project, for each $j \in [c]$, each $\boldsymbol{w}_j$ onto the ℓ_2 ball of radius D.

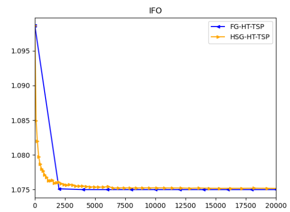
We have the smoothness constant L as below (see Böhning (1992) for a derivation):

$$L = \sigma_{\max}\left(\frac{1}{2n}\left(\boldsymbol{I}_{c\times c} - \frac{1}{c}\mathbf{1}_c\mathbf{1}_c^\top\right) \otimes \boldsymbol{X}^\top\boldsymbol{X} + 2\lambda \boldsymbol{I}_{d\times d}\right) \tag{62}$$

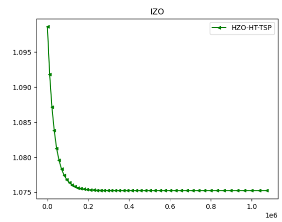
Where $\otimes$ denotes the Kronecker product, $\sigma_{\max}$ the largest singular value of a matrix, $\boldsymbol{I}_{m\times m}$ the identity matrix of size $m \times m$ for some m, and $\mathbf{1}_c$ the vector $[1, 1, .., 1]^\top \in \mathbb{R}^c$.

We consider the `dna` dataset from the LibSVM dataset repository (Chang & Lin, 2011), and we choose $D = 0.5$, $\lambda = 10$. For the stochastic case we take $B = 1e^5$, and for the stochastic and ZO case we take $\alpha = 2$. Note that in the stochastic case, if the growing batch-size required by Theorem 2 becomes larger than n, we keep it fixed to n (i.e. in such case we take the whole dataset at each step). In the zeroth-order case, we take $\mu = 1e - 6$. We set set all other hyperparameters as per Theorems 1, 2 and 3. In Figures 5, 6, 7 and 8, we plot the number of calls to a gradient ∇R_i (IFO: iterative first order oracle), and number of hard-thresholding operations (NHT), for various values of k and D (for the zeroth-order case, we plot the IZO (number of calls to the function R) instead of the IFO). We can observe that HSG-HT-TSP allows a smaller IFO than FG-HT-TSP in early iterations, since it does not need to compute a full gradient at each iteration.

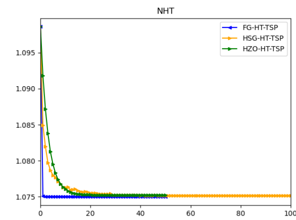
In addition, to illustrate the theoretical improvement of our results on zeroth-order, even in the case where there is no additional constraint, we compare in Figures 9, 10 and 11 our algorithm HZO-HT with ZOHT (de Vazelhes et al., 2022), choosing for both algorithm an initial number of random direction as prescribed by our Theorem 3, and choosing, for the learning rate, in our case the one prescribed by Theorem 3, and for ZOHT, the one prescribed by Theorem 1 from de Vazelhes et al. (2022) (and in both cases we fix $s = 3k$ as per Theorem 3): we can see that, in addition to being able to obtain a convergence in risk without system error, contrary to ZOHT (cf. Table 1), our Theorem 3 also prescribes a better (larger) learning rate (i.e. less conservative), leading to faster convergence.



(a) #IFO

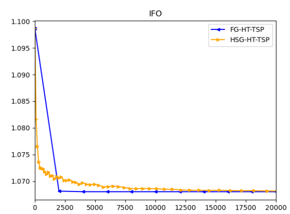


(b) #IZO

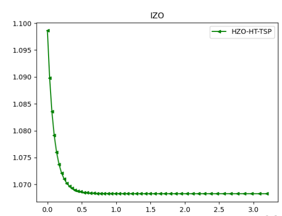


(c) #NHT

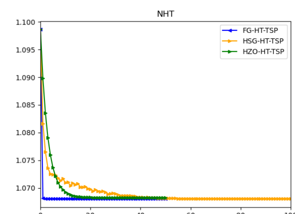
Figure 5: Multiclass Logistic Regression with TSP, $k = 50$, $D = 0.5$



(a) #IFO

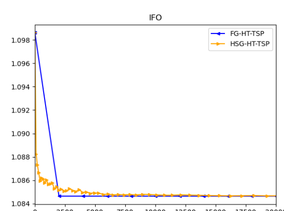


(b) #IZO

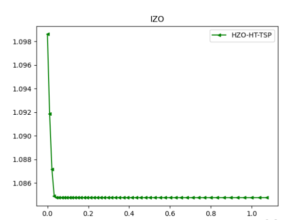


(c) #NHT

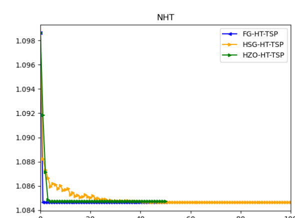
Figure 6: Multiclass Logistic Regression with TSP, $k = 150$, $D = 0.5$



(a) #IFO

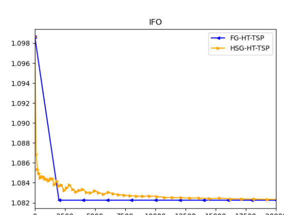


(b) #IZO

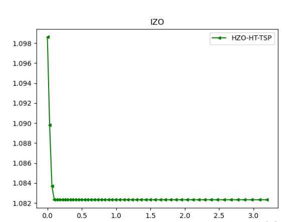


(c) #NHT

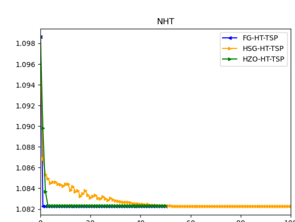
Figure 7: Multiclass Logistic Regression with TSP, $k = 50$, $D = 0.01$



(a) #IFO

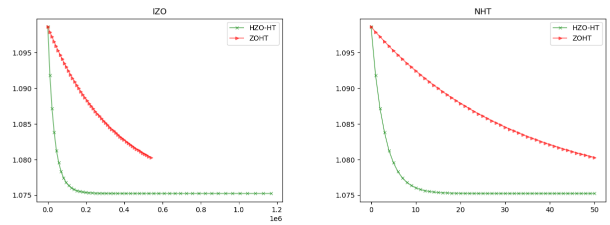


(b) #IZO



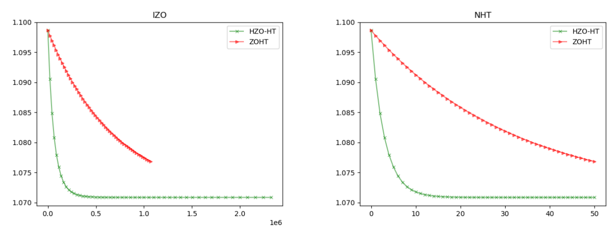
(c) #NHT

Figure 8: Multiclass Logistic Regression with TSP, $k = 150$, $D = 0.01$



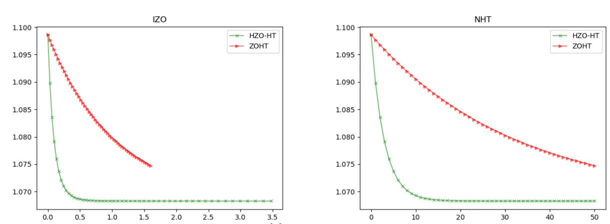
(a) #IZO (b) #NHT

Figure 9: Multiclass Logistic Regression: HZO-HT vs. ZOHT, $k = 50$



(a) #IZO (b) #NHT

Figure 10: Multiclass Logistic Regression: HZO-HT vs. ZOHT, $k = 100$



(a) #IZO (b) #NHT

Figure 11: Multiclass Logistic Regression: HZO-HT vs. ZOHT, $k = 150$

On the verification of Assumptions 1 to 3: Note that such logistic regression experiments verify Assumptions 1, 2, 5 and 3:

- **Assumption 1** is verified thanks to the added squared ℓ_2 regularization, which makes the problem strongly convex and hence also restricted strongly convex.
- **Assumption 2 and Assumption 5** are both verified since the problem is smooth with a constant L as described above in equation 62, and therefore such constant is also a valid (strong) restricted-smoothness constant.
- **Assumption 3** is verified since, since, similarly as in the index tracking experiments from Section F.1, projection onto Γ can be done group-wise, and for each group the projection is onto an ℓ_1 ball, which is a convex sign-free set (which is support-preserving from Remark 1), therefore, overall, Γ is support-preserving.

F.4 Adversarial Attacks

We consider the problem of adversarial attacks, where we seek to optimize a perturbation $\boldsymbol{\delta}$ applied to an image, such that a (previously trained) classifier (e.g. deep convolutional neural network) predicts the wrong class for the perturbed image. In our case we seek to enforce sparse constraints (i.e. where the number of pixels modified must be at most k). In addition, we seek to enforce an additional group constraint over a grid, similar to the constraints in the previous experiments (Section 5): $\Gamma = \{\boldsymbol{\delta} \in \mathbb{R}^d : \forall i \in r, \|\boldsymbol{\delta}_{G_i}\|_2 \leq D\}$, where r denotes the number of regions (16 in our case, see Figure 12), and where each group G_i corresponds to a region from a grid, shown in Figure 12 below. We consider the `CIFAR10` dataset (Krizhevsky et al., 2009), with $k = 50$, and $D = \frac{1}{4} = \frac{1}{\sqrt{16}}$. The motivation for using such constraint is that methods which enforce a global ℓ_2 constraint may still leave the freedom to the attack to be focused in a small region of the image, in which case it might be more detectable. Enforcing a maximum ℓ_2 norm over each region ensures that there are no region with an ℓ_2 norm being too large. We compare such method to a simple algorithm (baseline) which ensures a simple, global, ℓ_2 constraint of radius $D = 1$. Finally, we use the package `Torchattacks` by Kim (2020) to conduct those adversarial attacks.

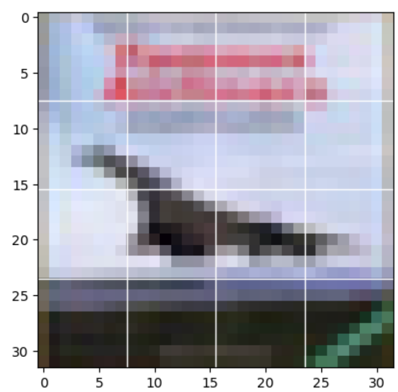
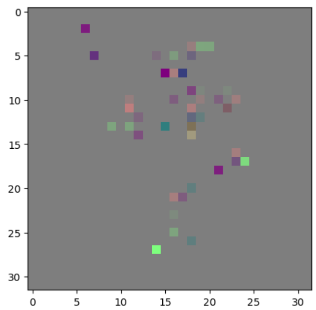


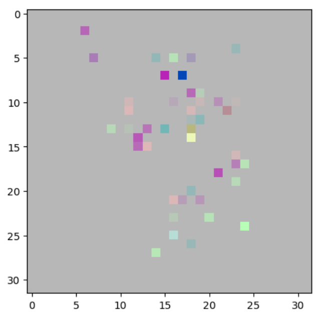
Figure 12: Original image, and the corresponding grid.

White-Box Adversarial Attacks In white-box adversarial attacks, one has access to the gradient of the objective function (cross-entropy of the prediction by the neural network, which we seek to maximize in order to attack the image). Therefore, we use FG-HT-TSP for the optimization in this case. We illustrate the attack and the baseline in Figure 16: we plot the learned attack $\boldsymbol{\delta}$ in both cases (Figure 13, the image after attack (Figure 14), and the intensity map, that is, the ℓ_2 norm of each RGB pixel seen as an $\mathbb{R}^3$ vector (Figure 15)). Note that in the displayed image, the attack was successful (for the two methods), misclassifying the 'airplane' into a 'ship'. As we can observe, enforcing an ℓ_2 norm constraint on each region allows to have potentially more imperceptible attacks than few-pixels adversarial attacks which only constrain the global ℓ_2 norm: the maximum ℓ_2 norm of the regions is 0.25 for our method, but 0.46 for the baseline.

Black-Box Adversarial Attacks In many practical adversarial attacks settings however, we do not have access to the gradient, and can rely only on function evaluations, also called Black-Box Adversarial Attacks. Therefore, in such a case, we use our algorithm HZO-HT-TSP to optimize the cost function. We take a smoothing radius $\mu = 0.01$, and consider the negative log-likelihood loss instead of cross-entropy loss, which is more stable for zeroth-order optimization, as it leads to less numerical imprecision. We illustrate the attack and the baseline in Figure 20: we plot the learned attack $\boldsymbol{\delta}$ in both cases (Figure 17, the image after attack (Figure 18), and the intensity map, that is, the ℓ_2 norm of each RGB pixel seen as an $\mathbb{R}^3$ vector (Figure 19)). Note that in the displayed image, the attack was successful (for the two methods), misclassifying the 'airplane' into a 'ship'. As we can observe, enforcing an ℓ_2 norm constraint on each region allows to have potentially more imperceptible attacks than few-pixels adversarial attacks which only constrain the global ℓ_2 norm: : the maximum ℓ_2 norm of the regions is 0.25 for our method, but 0.41 for the baseline.

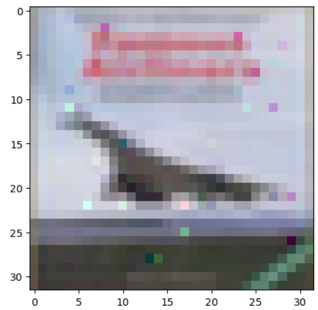


(a) Extra constraint: Γ

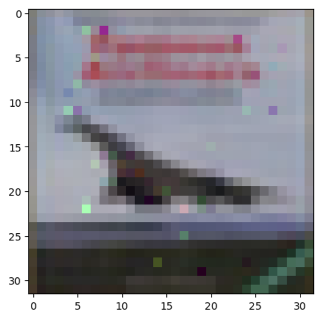


(b) Extra constraint: ℓ_2 ball

Figure 13: Attack $\boldsymbol{\delta}$

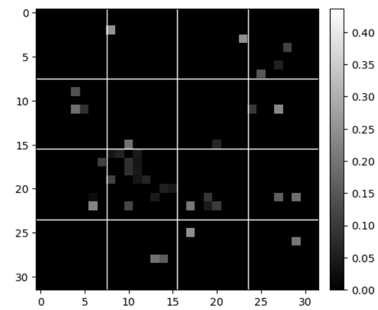


(a) Extra constraint: Γ

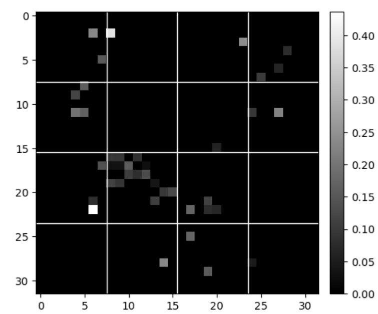


(b) Extra constraint: ℓ_2 ball

Figure 14: Image after attack



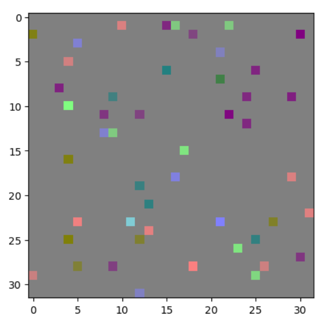
(a) Extra constraint: Γ



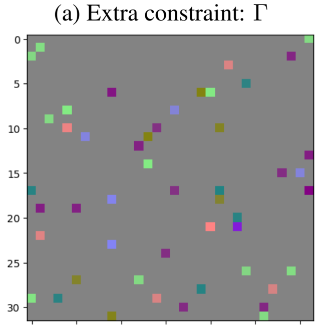
(b) Extra constraint: ℓ_2 ball

Figure 15: Intensity map

Figure 16: White-Box Adversarial Attacks

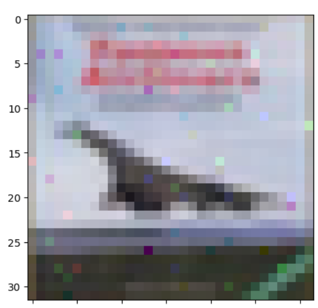


(a) Extra constraint: Γ

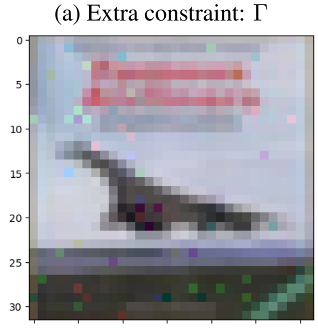


(b) Extra constraint: ℓ_2 ball

Figure 17: Attack $\boldsymbol{\delta}$

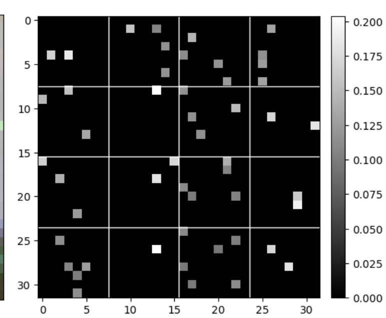


(a) Extra constraint: Γ

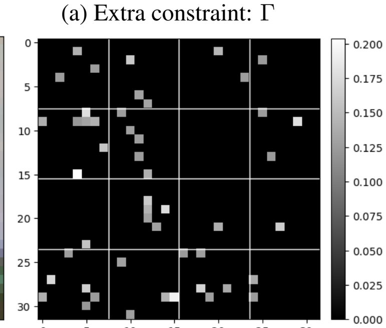


(b) Extra constraint: ℓ_2 ball

Figure 18: Image after attack



(a) Extra constraint: Γ



(b) Extra constraint: ℓ_2 ball

Figure 19: Intensity map

Figure 20: Black-Box Adversarial Attacks

F.5 SYNTHETIC EXPERIMENTS

In the section below, we provide a synthetic experiment to illustrate our Theorem 1, i.e. the trade-off between sparsity and optimality that is introduced by the extra constraint Γ, and that is measured by $\rho \in (0, 0.5]$. We consider the synthetic linear regression example from Axiotis & Sviridenko (2022) (Section E), with the risk below:

$$R(\boldsymbol{w}) := \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right\|_2^2,$$

and where $\boldsymbol{X}$ is diagonal with:

$$\boldsymbol{X}_{ii} = \begin{cases} 1 & \text{if } i \in I_1 \\ \sqrt{\kappa} & \text{if } i \in I_2 \\ 1 & \text{if } i \in I_3\,, \end{cases}$$

where $I_1 = [s]$, $I_2 = [s+1, s(\kappa+1)]$, $I_3 = [s(\kappa+1)+1, s(\kappa^2+\kappa+1)]$ for some $s \geq 1$ and $\kappa \geq 1$ (we choose $s = 50$ and $\kappa = 2$, which results in having $d = 350$) and $\boldsymbol{y}$ is defined as

$$y_i = \begin{cases} \kappa\sqrt{1-4\delta} & \text{if } i \in I_1 \\ \sqrt{\kappa}\sqrt{1-2\delta} & \text{if } i \in I_2 \\ 1 & \text{if } i \in I_3 \end{cases}$$

for some small $\delta > 0$ used for tie-breaking (we set it to $1e-4$). We chose such an example as it is used by Axiotis & Sviridenko (2022) to prove a lower bound on the fundamental trade-off between sparsity and optimality proper to IHT: they use it to show that the relaxation of the sparsity k, of the order $k = \Omega(\kappa^2\bar{k})$ (see also Table 1) is in fact unavoidable for IHT-type algorithms.

Case without extra constraints First, we illustrate our Theorem 4 which considers vanilla IHT, without extra constraints. In Figure 21, on the one hand, we plot in blue, for every $k \in [d]$, the value of $R(\hat{\boldsymbol{w}}_k)$ where $\hat{\boldsymbol{w}}_k$ is the result of running vanilla IHT with sparsity k up to convergence. Then, on the other hand, we go through every value of $\bar{k} \in [d]$, and for each of them, we plot a point $(K(\bar{k}), R(\bar{\boldsymbol{w}}_{\bar{k}}))$, where $K(\bar{k})$ denotes the value of k required in our Theorem 4, i.e.: $K(\bar{k}) := 4\kappa^2\bar{k}$, and $\bar{\boldsymbol{w}}_{\bar{k}} := \min_{\boldsymbol{w}\in\mathbb{R}^d : \|\boldsymbol{w}\|_0 \leq \bar{k}} R(\boldsymbol{w})$. Therefore, each of such point $R(\bar{\boldsymbol{w}}_{\bar{k}})$ constitutes an upper bound on the value of $R(\hat{\boldsymbol{w}}_{K(\bar{k})})$, as we can indeed observe on Figure 21.

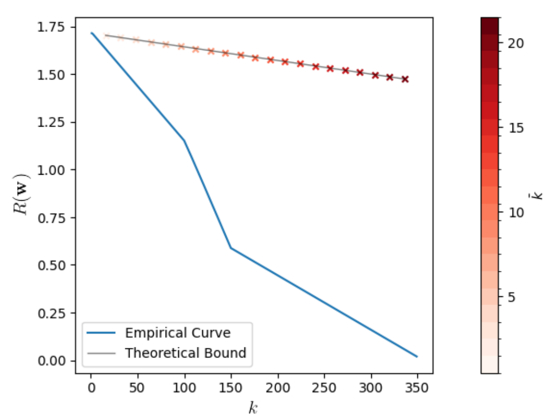
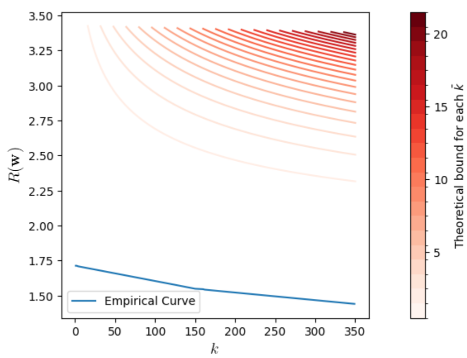


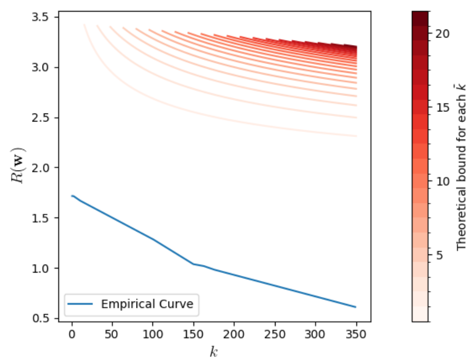
Figure 21: Illustration of Theorem 4 (i.e. $\Gamma = \mathbb{R}^d$).

Case with extra constraints We now illustrate the influence of the extra constraint Γ on the problem. We consider for Γ an ℓ_∞ norm constraint of radius $\lambda > 0$, that is: $\Gamma = \{\boldsymbol{w} \in \mathbb{R}^d : \forall i \in [d], |w_i| \leq \lambda\}$. In this new setting, we also go through every value of $\bar{k} \in [d]$, but this time, each of those values actually defines a curve parameterized by ρ, according to our Theorem 1: for each $\bar{k}$ we plot the parametric curve $(K(\bar{k}, \rho), (1 + 2\rho)R(\bar{\boldsymbol{w}}_{\bar{k}}))$, where, similarly as above, $K(\bar{k}, \rho)$ denotes the required value of k according to Theorem 1 (i.e., $K(\bar{k}, \rho) = \frac{4(1-\rho)^2\bar{k}\kappa^2}{\rho^2}$), and

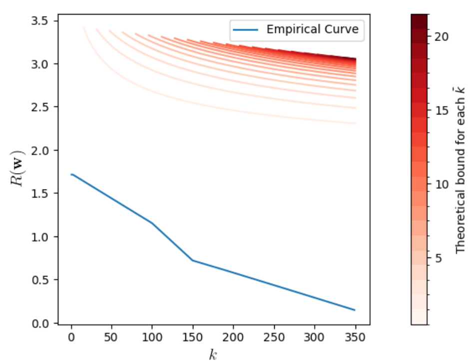
$\bar{\boldsymbol{w}}_{\bar{k}} := \min_{\boldsymbol{w}\in\mathbb{R}^d:\|\boldsymbol{w}\|_0\leq\bar{k}} R(\boldsymbol{w})$, and where ρ ranges in $(0, 0.5]$. We present the results for several values of λ in Figure 22 below. Note *that a priori*, the curves are allowed to cross, i.e. for a given k on the x-axis, one could have a point from a curve of small $\bar{k}$ (i.e. lighter shade of red) which could potentially also belong to a curve of larger $\bar{k}$ (let us denote it $\bar{k}'$) (darker shade of red), which would necessarily have a larger ρ (let us denote it ρ'), but for which the overall $(1+2\rho')R(\bar{\boldsymbol{w}}_{\bar{k}'})$ could be equal to $(1+2\rho)R(\bar{\boldsymbol{w}}_{\bar{k}})$ (since the problem will be less constrained with $\bar{k}'$ than with $\bar{k}$). However, interestingly, this is not the case here due to the simplicity of the structure of the example. We can also observe that similarly as in the case where $\Gamma = \mathbb{R}^d$, the bound is a bit tighter in the small k regime (i.e. when $k \in [50, 100]$).



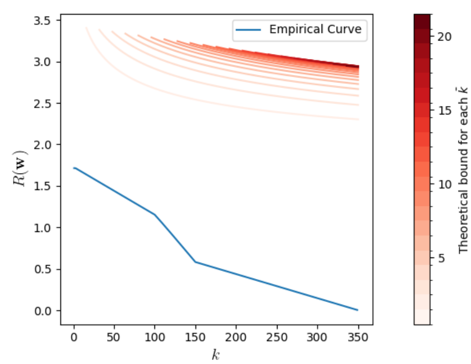
(a) $\lambda = 0.1$



(b) $\lambda = 0.5$



(c) $\lambda = 1$



(d) $\lambda = 2$

Figure 22: Illustration of Theorem 4 (with Γ an ℓ_∞ ball of radius λ).