Zeroth-Order Stochastic Compositional Gradient Descent: Towards Black-Box Sparse AUC Maximization

Wenkang Wang
Jilin University

Dongxu Liu
Jilin University

Bin Gu✉
Jilin University

Abstract

The area under the ROC curve (AUC) is a key metric for classification tasks, valued for its robustness to class imbalance. Sparse models trained with ℓ_0 constraints further enhance interpretability and generalization. Building on prior work that reformulates nonlinear AUC maximization as a pointwise compositional optimization problem, we revisit this formulation as the basis for addressing the black-box setting, where only function evaluations are available. A central challenge arises from integrating zeroth-order gradient estimation with hard-thresholding operators in the compositional framework, which has remained unresolved. To overcome this difficulty, we propose the Zeroth-Order Stochastic Compositional Hard-Thresholding (ZO-SCHT) algorithm, which, to the best of our knowledge, is the first method for black-box sparse AUC maximization. We establish that ZO-SCHT achieves linear convergence up to a tolerance bound under a fixed step size. Extensive experiments on both black-box sparse AUC maximization and black-box adversarial attack tasks demonstrate the effectiveness and versatility of our approach.

1 INTRODUCTION

In classification tasks with class imbalance, accuracy-based metrics often disproportionately favor the majority class, leading to poor generalization for minority instances (Gultekin et al., 2020). In contrast, the area under the ROC curve (AUC) represents the probability that a randomly selected positive sample has a higher prediction score than a randomly selected negative sample, providing intrinsic robustness to imbalanced label distributions (Hanley and McNeil, 1982). This inherent advantage has made AUC maximization a preferred strategy in mission-critical domains such as medical image diagnosis (Yuan et al., 2021), molecular property prediction (Wu et al., 2018), and financial risk assessment (Zhou et al., 2009).

Model sparsity is a crucial objective in machine learning, especially in high-dimensional settings. An ℓ_0 constraint not only reduces memory and computational costs but also mitigates overfitting and ensures statistically stable estimation (Yuan and Li, 2021). A widely used method for ℓ_0-constrained problems is the hard-thresholding gradient method (Jain et al., 2014; Nguyen et al., 2017; Yuan et al., 2018), which alternates between gradient updates and hard-thresholding projections to enforce exact sparsity throughout the optimization process. However, in many practical scenarios, first-order gradients may be unavailable or computationally expensive. To address this challenge, de Vazelhes et al. (2022) introduced a combination of zeroth-order gradient estimation and hard-thresholding, resulting in the SZOHT algorithm. Subsequently, Yuan et al. (2024) enhanced this approach by incorporating the Stochastic Variance-Reduced Gradient (SVRG) technique into the zeroth-order hard-thresholding framework, yielding improved convergence guarantees and enhanced practical performance.

In the batch learning setting, AUC maximization is typically formulated as a convex empirical risk minimization problem using a surrogate loss. However, the inherent pairwise structure of the objective makes it challenging to scale to large datasets. To address this, Ying et al. (2016) reformulated AUC maximization as a stochastic min-max saddle point problem and developed a primal-dual style algorithm with a convergence rate of $\widetilde{O}(1/\sqrt{t})$. Despite its efficiency, such saddle point methods do not naturally accommodate hard-thresholding operations, limiting their applicability in ℓ_0-constrained sparse learning. To overcome this,

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. ✉ Correspondence to jsgubin@gmail.com

Yang et al. (2020) reformulated the linear AUC objective as a standard pointwise empirical risk minimization, yielding an optimization framework that readily integrates hard-thresholding projections. Wang et al. (2026) also extends this reformulation to the nonlinear case, where nonlinear AUC maximization can be equivalently cast as a pointwise compositional optimization problem. Building on this, our work presents the first approach to sparse AUC maximization in black-box settings.

To address zeroth-order stochastic compositional optimization, Liu et al. (2024) proposed gradient-free compositional optimization methods (GFCOM) for nonconvex nonsmooth problems via randomized smoothing, establishing non-asymptotic convergence rates. Chen et al. (2024) investigated the theoretical effects of black-box settings on SCGD and SCSC algorithms, providing stability-based generalization bounds and optimization guarantees for their zeroth-order variants. Despite these advances, how to effectively integrate zeroth-order gradient estimation with hard-thresholding operators within a compositional framework to simultaneously handle black-box and sparsity constraints remains an open challenge.

To address this challenge, we propose a novel Zeroth-Order Stochastic Compositional Hard-Thresholding (ZO-SCHT) algorithm. Specifically, we build upon the Stochastic Compositional Gradient Descent (SCGD) framework (Wang et al., 2017) and further incorporate the SZOHT algorithm (de Vazelhes et al., 2022), enabling the effective solution of black-box sparse optimization problems where explicit gradient information is unavailable.

The main contributions of this paper are summarized as follows:

- We propose a Zeroth-Order Stochastic Compositional Hard-Thresholding (ZO-SCHT) algorithm. To the best of our knowledge, this is the first approach specifically designed for black-box sparse AUC maximization.
- By combining compositional optimization with zeroth-order hard-thresholding, ZO-SCHT effectively enforces solution sparsity in black-box settings. Moreover, we rigorously establish that ZO-SCHT achieves linear convergence up to a tolerance bound under a fixed step size.
- Extensive experiments on black-box sparse AUC maximization and black-box adversarial attack tasks demonstrate that ZO-SCHT achieves superior performance compared to existing methods, validating its effectiveness and versatility in handling black-box sparse optimization problems.

2 RELATED WORK

AUC Maximization. In the linear setting, as previously discussed, Ying et al. (2016) proposed a primal–dual stochastic gradient method to address the scalability challenges of AUC maximization in large-scale applications, achieving a convergence rate of $\widetilde{O}(1/\sqrt{t})$. Building upon this, Natole et al. (2018) introduced an ℓ_1 constraint to promote sparsity and employed a stochastic proximal gradient method, which improved the convergence rate to $\widetilde{O}(1/t)$. Subsequently, Yang et al. (2020) reformulated the AUC objective and applied hard-thresholding gradient descent, thereby enabling effective AUC maximization under ℓ_0 constraints. In the nonlinear setting, Dang et al. (2020) adopted kernel approximation via Random Fourier Features and developed a triply stochastic gradient optimization scheme, attaining a convergence rate of $O(1/t)$. Furthermore, Liu et al. (2019) formulated the problem as a nonconvex–concave min–max optimization with deep neural networks and proposed two stochastic primal-dual algorithms—PPD-SG and PPD-AdaGrad—achieving an iteration complexity of $\widetilde{O}(1/\epsilon)$ under the Polyak-Łojasiewicz condition.

Compositional optimization. In the field of machine learning, for stochastic compositional optimization problems of the form $\min_{\boldsymbol{x}} \mathbb{E}_i[f_i(\mathbb{E}_j[g_j(\boldsymbol{x})])]$, Wang et al. (2017) pioneered the SCGD method and established convergence guarantees for convex, strongly convex, and nonconvex settings, along with preliminary insights into zeroth-order extensions. Building on this, recent work (Chen et al., 2024) developed a more effective stability analysis framework, deriving sharper generalization bounds for both SCGD and the stochastically corrected variant SCSC under convex and nonconvex assumptions, and further provided learning guarantees for three black-box variants (outer, inner, and full black-box). Moreover, Liu et al. (2024) proposed a gradient-free compositional optimization method (GFCOM) tailored for nonconvex and nonsmooth stochastic compositional problems, establishing non-asymptotic complexity bounds for finding (δ, ϵ)-Goldstein stationary points. Collectively, these advances significantly broaden the theoretical and algorithmic foundation of stochastic compositional optimization.

Hard Thresholding. The ℓ_0-sparse optimization problem plays a central role in sparse learning and compressed sensing. A seminal contribution in this domain is the Iterative Hard Thresholding (IHT) algorithm (Blumensath and Davies, 2009), which alternates between a gradient descent step and a hard-thresholding operation that preserves only the top-s entries in magnitude, thus enforcing exact s-sparsity.

Building upon this foundation, a stochastic variant known as Stochastic IHT (StoIHT) (Nguyen et al., 2017) was introduced to reduce the per-iteration computational cost by leveraging randomized gradient estimates. de Vazelhes et al. (2022) further extended this approach by proposing a zeroth-order stochastic hard-thresholding method for black-box ℓ_0-constrained problems, revealing an intrinsic trade-off between gradient approximation error and the expansive nature of the hard-thresholding operator. More recently, Yuan et al. (2024) developed a generalized variance-reduced zeroth-order hard-thresholding framework, which effectively alleviates this conflict and achieves faster convergence rates.

3 METHOD

3.1 Preliminaries and Notations

Let $n \geq 1$ be an integer, and denote the index set by $[n] = \{1, 2, \ldots, n\}$. For any vector $\boldsymbol{v} = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, the ℓ_0-norm $\|\boldsymbol{v}\|_0$ is defined as the number of nonzero entries in $\boldsymbol{v}$. Given $d \in \mathbb{N}$ and a subset $\mathcal{I} \subseteq [d]$, the orthogonal projection operator $\mathcal{P}_{\mathcal{I}}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is defined by:

$$\big(\mathcal{P}_{\mathcal{I}}(\boldsymbol{v})\big)_i = \begin{cases} v_i, & i \in \mathcal{I}, \\ 0, & i \notin \mathcal{I}. \end{cases}$$

In particular, if Γ denotes the index set of the k largest entries of $\boldsymbol{v}$ in magnitude, the hard-thresholding operator $\mathcal{H}_k(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is defined as:

$$\mathcal{H}_k(\boldsymbol{v}) = \mathcal{P}_{\Gamma}(\boldsymbol{v}).$$

Moreover, the projection of $\boldsymbol{v}$ onto the ℓ_2-ball of radius $\omega > 0$ is given by:

$$\Pi_\omega(\boldsymbol{v}) = \frac{\boldsymbol{v}}{\max\{1, \|\boldsymbol{v}\|_2/\omega\}}.$$

Consider an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and a label set $\mathcal{Y} = \{\pm 1\}$. The training dataset is given by $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [n]\}$, consisting of independent and identically distributed samples drawn from an unknown distribution over $\mathcal{X} \times \mathcal{Y}$. An instance is termed positive if $y_i = 1$ and negative otherwise. Let n_+ and n_- denote the numbers of positive and negative instances, respectively, and define the imbalance ratio as $r = n_+/n$.

For a scoring function $h : \mathcal{X} \to \mathbb{R}$, the population AUC is defined as:

$$\mathrm{AUC}(h) = \Pr(h(\boldsymbol{x}) \geq h(\boldsymbol{x}') \mid y = 1,\, y' = -1)\,,$$

where $(\boldsymbol{x}, y)$ and $(\boldsymbol{x}', y')$ are independent random pairs drawn from the distribution. Following common practice (Ying et al., 2016; Liu et al., 2019), we adopt the squared loss as a surrogate for the indicator function and learn a nonlinear scoring model $h(\boldsymbol{w}; \boldsymbol{x})$ parameterized by $\boldsymbol{w}$. The nonlinear sparse AUC maximization problem on the dataset $\mathcal{S}$ is formulated as:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & F(\boldsymbol{w}) = \frac{1}{n_+ n_-} \sum_{i=1}^{n} \sum_{j=1}^{n} \big(1 - h(\boldsymbol{w}; \boldsymbol{x}_i) \\ & \qquad + h(\boldsymbol{w}; \boldsymbol{x}_j)\big)^2 \mathbb{I}_{[y_i=1]} \mathbb{I}_{[y_j=-1]}. \\ \text{s.t.} \quad & \|\boldsymbol{w}\|_0 \leq k \end{aligned} \tag{1}$$

We introduce two commonly used assumptions in sparse learning theory, namely Restricted Strong Convexity (RSC) and Restricted Strong Smoothness (RSS) (Nguyen et al., 2017; Zhou et al., 2018; Shen and Li, 2018), which are defined as follows.

Definition 1. ***(Restricted Strong Convexity)*** *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy the property of restricted strong convexity with parameter $\alpha_s > 0$, if for all vectors $\boldsymbol{w}$, $\boldsymbol{w}' \in \mathbb{R}^d$ with $\|\boldsymbol{w} - \boldsymbol{w}'\|_0 \leq s$, it holds that*

$$f(\boldsymbol{w}') - f(\boldsymbol{w}) - \langle \nabla f(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w} \rangle \geq \frac{\alpha_s}{2} \|\boldsymbol{w}' - \boldsymbol{w}\|_2^2.$$

Definition 2. ***(Restricted Strong Smoothness)*** *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy the property of restricted strong smoothness with parameter $L_s > 0$, if for all vectors $\boldsymbol{w}$, $\boldsymbol{w}' \in \mathbb{R}^d$ with $\|\boldsymbol{w} - \boldsymbol{w}'\|_0 \leq s$, it holds that*

$$\|\nabla f(\boldsymbol{w}') - \nabla f(\boldsymbol{w})\|_2 \leq L_s \|\boldsymbol{w}' - \boldsymbol{w}\|_2.$$

3.2 Equivalent Reformulation

Direct optimization of $F(\boldsymbol{w})$ is challenging due to its pairwise nature, especially in large-scale settings (Yang et al., 2020). To overcome this, we reformulate the objective $F(\boldsymbol{w})$ as a compositional optimization problem. Define the mean scores for positive and negative instances as:

$$\begin{aligned} \mu_+(\boldsymbol{w}) &= \frac{1}{n_+} \sum_{i=1}^{n} h(\boldsymbol{w}; \boldsymbol{x}_i) \mathbb{I}_{[y_i=1]}, \\ \mu_-(\boldsymbol{w}) &= \frac{1}{n_-} \sum_{i=1}^{n} h(\boldsymbol{w}; \boldsymbol{x}_i) \mathbb{I}_{[y_i=-1]}, \end{aligned} \tag{2}$$

Next, define the vector-valued function $g_j(\boldsymbol{w})$:

$$g_j(\boldsymbol{w}) = \begin{pmatrix} \boldsymbol{w} \\ \frac{1}{r} h(\boldsymbol{w}; \boldsymbol{x}_j) \mathbb{I}_{[y_j=1]} \\ \frac{1}{1-r} h(\boldsymbol{w}; \boldsymbol{x}_j) \mathbb{I}_{[y_j=-1]} \end{pmatrix}, \quad j = 1, \ldots, n \tag{3}$$

and its mean function $g(\boldsymbol{w})$:

$$g(\boldsymbol{w}) = \frac{1}{n} \sum_{j=1}^{n} g_j(\boldsymbol{w}) = \begin{pmatrix} \boldsymbol{w} \\ \mu_+(\boldsymbol{w}) \\ \mu_-(\boldsymbol{w}) \end{pmatrix}. \tag{4}$$

Then, we have the following proposition.

Proposition 1. *(Proof in Appendix C) The AUC maximization objective $F(\boldsymbol{w})$ defined in Eq. (1) can be equivalently reformulated as:*

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & F(\boldsymbol{w}) = f(g(\boldsymbol{w})) = \frac{1}{n}\sum_{i=1}^{n} f_i\left(\frac{1}{n}\sum_{j=1}^{n} g_j(\boldsymbol{w})\right), \\ s.t. \quad & \|\boldsymbol{w}\|_0 \le k \end{aligned} \tag{5}$$

where the outer component function f_i is defined as:

$$\begin{aligned} f_i\left(g(\boldsymbol{w})\right) = & \left(1 + \mu_-(\boldsymbol{w}) - \mu_+(\boldsymbol{w})\right)^2 \\ & + \frac{1}{r}\left(h(\boldsymbol{w};\boldsymbol{x}_i) - \mu_+(\boldsymbol{w})\right)^2 \mathbb{I}_{y_i=1} \\ & + \frac{1}{1-r}\left(h(\boldsymbol{w};\boldsymbol{x}_i) - \mu_-(\boldsymbol{w})\right)^2 \mathbb{I}_{y_i=-1}. \end{aligned} \tag{6}$$

3.3 Zeroth-Order Gradient Estimator

In the black-box setting, where gradients are not directly accessible, we employ zeroth-order gradient estimation for the objective function $F(\boldsymbol{w}) = f(g(\boldsymbol{w}))$. Specifically, the Jacobian of the inner function $g_j : \mathbb{R}^d \to \mathbb{R}^{d+2}$ is estimated stochastically as follows:

$$\widehat{\nabla} g(\boldsymbol{w}) = \frac{d}{q\mu}\sum_{s=1}^{q} \boldsymbol{u}_s \left(g(\boldsymbol{w} + \mu \boldsymbol{u}_s) - g(\boldsymbol{w})\right)^\top, \tag{7}$$

where $\mu > 0$ is a smoothing parameter, and $\{\boldsymbol{u}_s\}_{s=1}^q$ are q random directions sampled from the sparse unit sphere $\mathcal{S}_{s_2}^d = \left\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_0 \le s_2, \|\boldsymbol{u}\| = 1\right\}$. To generate these vectors, we first randomly select a subset S of size s_2 from the index set $[d]$. Then, we sample a random vector $\boldsymbol{u}$, which is nonzero only at the coordinates in S, meaning $\boldsymbol{u}$ is uniformly selected from the set $\left\{\boldsymbol{u} \in \mathbb{R}^d : \boldsymbol{u}_{[d]-S} = \boldsymbol{0}, \|\boldsymbol{u}\| = 1\right\}$. In particular, when $s_2 = d$, $\widehat{\nabla} g(\boldsymbol{w})$ is the usual vanilla estimator with uniform smoothing on the sphere (Gao et al., 2018).

To estimate the gradient of the outer function $f_i : \mathbb{R}^{d+2} \to \mathbb{R}$, we apply a similar zeroth-order gradient estimation approach. Let $\{\boldsymbol{z}_s\}_{s=1}^q$ be q random directions uniformly sampled from the unit sphere $\mathcal{S}^{d+2} = \left\{\boldsymbol{z} \in \mathbb{R}^{d+2} : \|\boldsymbol{z}\| = 1\right\}$. The gradient estimator is then given by:

$$\widehat{\nabla} f(g) = \frac{d+2}{q\mu}\sum_{s=1}^{q} \left(f(g + \mu \boldsymbol{z}_s) - f(g)\right) \boldsymbol{z}_s, \tag{8}$$

where $\mu > 0$ is the smoothing parameter. This corresponds to the usual vanilla gradient estimator with uniform smoothing over the unit sphere.

We then quantify the estimation error of $\widehat{\nabla} f(g)$ and $\widehat{\nabla} g(\boldsymbol{w})$ in the following proposition.

Proposition 2. *(Proof in Appendix D) Let $\mathcal{I} \subset [d]$ be an arbitrary support set of size s (i.e., $|\mathcal{I}| = s$). For the zeroth-order gradient estimator defined in equations (7) and (8), which employs q random directions and random supports of size s_2, we assume that each outer function f_i is Lipschitz smooth with constant $L_f \ge 0$, and each scoring function $h_j(\boldsymbol{w})$ is (L_{s_2}, s_2)-RSS. Let $\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w})$ denote the hard-thresholding of the matrix $\hat{\nabla} g_j(\boldsymbol{w})$ on the support $\mathcal{I}$, meaning that all rows corresponding to indices not in $\mathcal{I}$ are set to zero. Under these assumptions, we have:*

(a) $\|\mathbb{E}\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w}) - \nabla_{\mathcal{I}} g_j(\boldsymbol{w})\|^2 \le \varepsilon_\mu \mu^2$

(b) $\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w})\right\|^2 \le \varepsilon_{\mathcal{I}} \|\nabla g_j(\boldsymbol{w})\|^2 + \varepsilon_g \mu^2$

(c) $\|\mathbb{E}\hat{\nabla} f_i(g) - \nabla f_i(g)\|^2 \le \varepsilon_{abs} \mu^2$

(d) $\mathbb{E}\|\hat{\nabla} f_i(g)\|^2 \le \varepsilon \|\nabla f_i(g)\|^2 + \varepsilon_f \mu^2$

with

$$a_j = \frac{\mathbb{I}_{[y_j=1]}}{r}, \quad b_j = \frac{\mathbb{I}_{[y_j=-1]}}{1-r}, \quad \varepsilon_\mu = ds(a_j^2 + b_j^2)L_{s_2}^2$$

$$\varepsilon_{abs} = (d+2)^2 L_f^2, \quad \varepsilon_g = 2\left(\frac{1}{q} + 1\right) ds(a_j^2 + b_j^2)L_{s_2}^2$$

$$\varepsilon_f = 2\left(\frac{1}{q} + 1\right)(d+2)^2 L_f^2, \quad \varepsilon = \frac{2(d+2)}{q} + 2$$

$$\varepsilon_{\mathcal{I}} = \frac{2d}{q(s_2+2)}\left(\frac{(s-1)(s_2-1)}{d-1} + 3\right) + 2.$$

3.4 Algorithm

We now present the complete Zeroth-Order Stochastic Compositional Hard-Thresholding (ZO-SCHT) algorithm. Building upon the SCGD framework (Wang et al., 2017), ZO-SCHT integrates zeroth-order gradient estimation with hard-thresholding to address the black-box sparse AUC maximization problem. The algorithm proceeds as follows:

At each iteration t, an index j_t is randomly selected, and the function value $g_{j_t}(\boldsymbol{w}_t)$ along with its zeroth-order Jacobian estimate $\widehat{\nabla} g_{j_t}(\boldsymbol{w}_t)$ are computed using Equations (3) and (7). An auxiliary variable $\boldsymbol{v}_{t+1}$ is updated via an exponential moving average with parameter $\beta \in (0, 1]$ to track $g(\boldsymbol{w}_t)$. Next, an index i_t is sampled, and the zeroth-order gradient estimate $\widehat{\nabla} f_{i_t}(\boldsymbol{v}_{t+1})$ is computed using Equation (8). The composite gradient is then approximated by the product $\widehat{\nabla} g_{j_t}(\boldsymbol{w}_t)\widehat{\nabla} f_{i_t}(\boldsymbol{v}_{t+1})$, which is used to update $\boldsymbol{w}_t$ with a step size η. Subsequently, the hard-thresholding operator $\mathcal{H}_k(\cdot)$ and the projection operator $\Pi_\omega(\cdot)$ are applied to enforce the ℓ_0 constraint and ensure boundedness. The complete ZO-SCHT algorithm is outlined in Algorithm 1.

Algorithm 1: Zeroth-Order Stochastic Compositional Hard-Thresholding

Input: Initial solution $\boldsymbol{w}_0$, learning rates η and β, sparsity level k, number of iterations T, number of random directions q, size of the random directions support s_2, projection radius ω.
Output: $\boldsymbol{w}_T$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Randomly sample $j_t \in [n]$, obtain $g_{j_t}(\boldsymbol{w}_t)$ and $\hat{\nabla} g_{j_t}(\boldsymbol{w}_t)$
  if $t \geq 1$ **then**
    $\boldsymbol{v}_{t+1} \leftarrow (1-\beta)\boldsymbol{v}_t + \beta g_{j_t}(\boldsymbol{w}_t)$
  else
    $\boldsymbol{v}_{t+1} \leftarrow g_{j_t}(\boldsymbol{w}_0)$
  end
  Randomly sample $i_t \in [n]$, obtain $\hat{\nabla} f_{i_t}(\boldsymbol{v}_{t+1})$
  $\boldsymbol{b}_{t+1} \leftarrow \boldsymbol{w}_t - \eta \hat{\nabla} g_{j_t}(\boldsymbol{w}_t) \hat{\nabla} f_{i_t}(\boldsymbol{v}_{t+1})$
  $\boldsymbol{r}_{t+1} \leftarrow \mathcal{H}_k(\boldsymbol{b}_{t+1})$
  $\boldsymbol{w}_{t+1} \leftarrow \Pi_\omega(\boldsymbol{r}_{t+1})$
end

4 CONVERGENCE ANALYSIS

This section presents the convergence results of Algorithm 1. Due to space limitations, the complete proofs are provided in the appendix. Before presenting the main results, we first introduce some general assumptions commonly adopted in sparse learning and stochastic compositional optimization.

4.1 Basic assumptions

Assumption 1: Let C_g, V_g, C_f, L_f be positive constants.

(i) The Jacobian of the inner function $g_j(\boldsymbol{w})$ has a bounded squared norm, and the sample estimates $g_j(\boldsymbol{w})$ exhibit bounded variance. Specifically, for all $\boldsymbol{w}$ and $j \in [n]$, the following inequalities hold:

$$\|\nabla g_j(\boldsymbol{w})\|^2 \leq C_g, \tag{9}$$

$$\|g_j(\boldsymbol{w}) - g(\boldsymbol{w})\|^2 \leq V_g. \tag{10}$$

(ii) The gradients of the outer functions f_i are bounded and Lipschitz continuous. Specifically, for all $\boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^{d+2}$ and $i \in [n]$, the following conditions hold:

$$\|\nabla f_i(\boldsymbol{v})\|^2 \leq C_f, \tag{11}$$

$$\|\nabla f_i(\boldsymbol{v}) - \nabla f_i(\boldsymbol{v}')\| \leq L_f \|\boldsymbol{v} - \boldsymbol{v}'\|. \tag{12}$$

Assumption 2: RSC and RSS conditions.

(i) The objective function $f(\boldsymbol{w})$ satisfies the RSC condition with parameter $\alpha_s > 0$ at sparsity level $s = 2k + k^*$, where k^* is the sparsity of the optimal solution $\boldsymbol{w}^*$, i.e., $\|\boldsymbol{w}^*\|_0 = k^*$.

(ii) Each scoring function $h_j(\boldsymbol{w})$ satisfies the (L_{s_2}, s_2)-RSS condition.

For convenience, we let $\alpha := \alpha_s$.

4.2 Main Theorem

Next, we establish the main convergence theorem for the ZO-SCHT algorithm. The following theorem and its corollary are derived based on the assumptions stated earlier.

Theorem 1. *(Proof in Appendix E.1) For Algorithm 1, assuming that Assumptions 1 and 2 hold, and that* $\frac{d-k^*}{2} \geq k > \rho k^*/(1-\rho)^2$, *where* ρ *is defined as follows, we obtain a geometric convergence rate for ZO-SCHT:*

$$\mathbb{E}\|\boldsymbol{w}_T - \boldsymbol{w}^*\|^2 \leq (\gamma\rho)^T \|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + \left(\frac{1}{1-\gamma\rho}\right)\xi_\eta + \left(\frac{1}{1-\gamma\rho}\right) L_\mu, \tag{13}$$

where

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

$$\xi_\eta = 4\eta\gamma\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f + \mathcal{O}\left(\frac{\eta\gamma\beta V_g C_g L_f^2}{\alpha}\right) + \mathcal{O}\left(\frac{\eta^3\gamma^2\varepsilon_{\mathcal{I}}\varepsilon C_g^3 C_f L_f^2}{\beta^2\alpha}\right)$$

and

$$L_\mu = \left[\eta^2\gamma\left(\varepsilon_{\mathcal{I}} C_g \varepsilon_f + \varepsilon C_f \varepsilon_g\right) + \frac{4\eta\gamma}{\alpha}\left(C_f\varepsilon_\mu + C_g\varepsilon_{abs}\right) + \mathcal{O}\left(\frac{\eta^3\gamma^2 C_g^2 L_f^2}{\beta^2\alpha}\right)\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right)\right]\mu^2 + \left[\eta^2\gamma\varepsilon_g\varepsilon_f + \frac{4\eta\gamma\varepsilon_\mu\varepsilon_{abs}}{\alpha} + \mathcal{O}\left(\frac{\eta^3\gamma^2 C_g^2 L_f^2}{\beta^2\alpha}\right)\varepsilon_g\varepsilon_f\right]\mu^4$$

Remark 1. *Theorem 1 demonstrates that, with a constant step size, Algorithm 1 achieves linear convergence up to a tolerance bound. This error is composed of two key components: The first is the term* ξ_η, *which is influenced by the step size and reflects a common phenomenon observed in stochastic gradient descent (SGD) frameworks when a constant step size is used (Schmidt, 2014; Gower et al., 2019). The second term is associated with the error introduced by zeroth-order gradient estimation, which depends on the smoothing radius* μ.

Remark 2 (Necessary condition on the η, proof in Appendix E.2)**.** *From the conditions in Theorem 1, we have $k > \rho k^*/(1-\rho)^2$ (which ensures that $\gamma\rho < 1$), and $k \leq (d-k^*)/2$ (which ensures that $s = 2k + k^* \leq d$). These conditions imply the following necessary (but not sufficient) condition on η:*

$$\frac{2}{\alpha\left(1+\sqrt{1+\frac{2(d-k^*)}{k^*}}\right)} \leq \eta \leq \frac{1}{\alpha}$$

Corollary 1 (Proof in Appendix E.3)**.** *Under the assumptions of Theorem 1, suppose the sparsity parameter satisfies $k > 2\rho k^*/(1-\rho)^2$, and the step size η is selected from the interval*

$$\frac{2}{\alpha\left(1+\sqrt{1+\frac{d-k^*}{k^*}}\right)} \leq \eta \leq \frac{1}{\alpha},$$

Then, the total function query complexity required to ensure that $\mathbb{E}\|\boldsymbol{w}_T - \boldsymbol{w}^*\|^2 \leq \varepsilon + \left(\frac{1}{1-\gamma\rho}\right)\xi_\eta + \left(\frac{1}{1-\gamma\rho}\right)L_\mu$ is $\mathcal{O}\left(\frac{q}{\alpha\eta}\log\left(\frac{1}{\varepsilon}\right)\right)$.

5 EXPERIMENTS

To assess the performance of the proposed algorithm, ZO-SCHT, we conducted comprehensive experiments in two representative black-box optimization scenarios: (a) sparse AUC maximization involving proprietary or non-differentiable feature transformations (e.g., third-party embedding APIs), where optimization is restricted to function-value queries only; and (b) black-box adversarial attacks under sparsity constraints, where the target classifier is a deployed proprietary system accessible solely via input-output queries, necessitating gradient-free optimization for security auditing. Both settings rigorously emulate real-world scenarios where gradient access is prohibited by system opacity or intellectual property restrictions.

5.1 Baselines

In both experimental scenarios, we compare ZO-SCHT against two distinct categories of baselines:

- Min-Max Baselines (Non-Compositional): ZO-Min-Max (Liu et al., 2020) and Acc-Semi-ZOMDA (abbreviated as ZOMDA) (Huang et al., 2022). These methods follow the standard saddle-point formulation for AUC (Ying et al., 2016).
- Compositional Baselines: GFCOM (Liu et al., 2024), Black-box SCGD (Chen et al., 2024), and Black-box SCSC (Chen et al., 2024). These methods represent generic gradient-free compositional optimization approaches.

Table 1: Statistics of Experimental Datasets

ID	Datasets	Instances	Features	Imbalance Rate
1	ionosphere	351	32	0.56
2	diabetes	768	8	0.536
3	german	1,000	24	0.302
4	usps	9,298	256	0.78
5	gassensor	13,910	128	0.132
6	a9a	32,561	123	0.240
7	covtype	58,102	54	0.486
8	acoustic	78,823	50	0.231

5.2 Black-Box Sparse AUC Maximization

In this experiment, we consider a black-box setting where the scoring function $h(\boldsymbol{w}; \boldsymbol{x}_i)$ is expressed as a linear combination of basis functions:

$$h(\boldsymbol{w}; \boldsymbol{x}_i) = \sum_{l \in A} \boldsymbol{w}(l)\, \phi_l(\boldsymbol{x}_i),$$

where $\phi_l(\boldsymbol{x}_i)$ represents a nonlinear feature mapping of the input. Crucially, under the black-box assumption, only the scalar output $h(\boldsymbol{w}; \boldsymbol{x}_i)$ is accessible, while the values of the internal feature mappings $\phi_l(\boldsymbol{x}_i)$ remain hidden. Consequently, first-order gradient information is unavailable, and optimization must rely exclusively on zeroth-order (gradient-free) estimation methods.

5.2.1 Experimental Setup

To evaluate the proposed algorithm, we selected eight benchmark datasets from LIBSVM (Chang and Lin, 2011) and the UCI Machine Learning Repository (Asuncion et al., 2007). Basic statistics for these datasets are summarized in Table 1. All features were scaled to the range $[-1, 1]$, and multiclass datasets were converted into imbalanced binary classification tasks. Each dataset was split into training (80%) and testing (20%) sets for model optimization and performance evaluation, respectively. To ensure reproducibility, data partitioning and parameter updates were controlled using fixed random seeds.

To identify the optimal hyperparameters for each algorithm–dataset pair, we employed 5-fold cross-validation. After selecting the best settings, each configuration was rerun three times with different random seeds to report the mean test AUC and its standard deviation, as summarized in Table 2. Due to space limitations, the detailed hyperparameter tuning settings for all algorithms are provided in Appendix F.

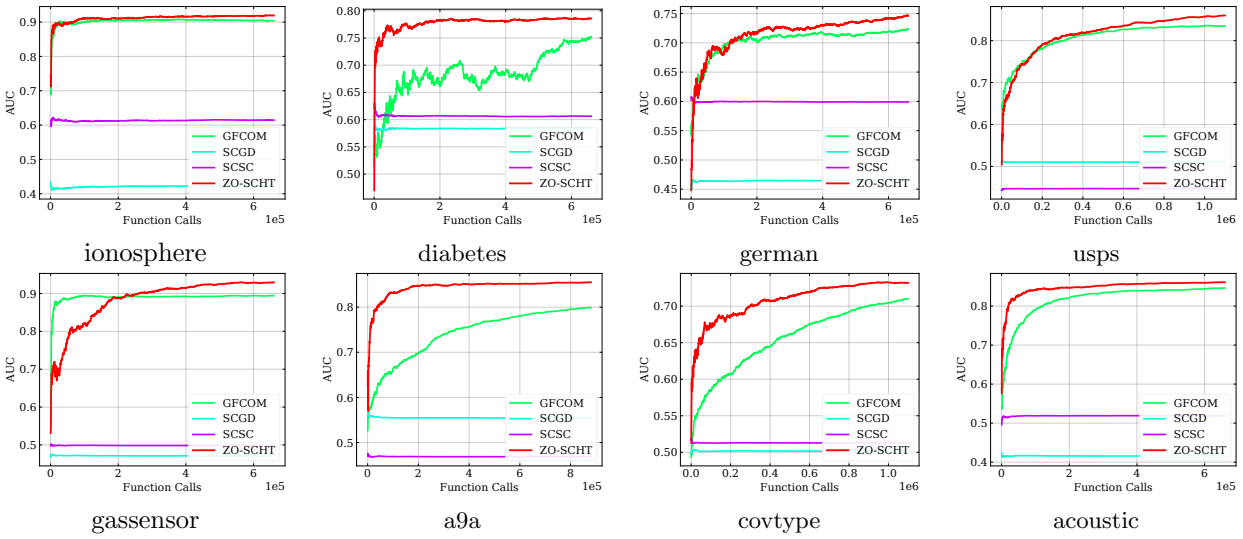

Figure 1: Test AUC vs. Function Calls.

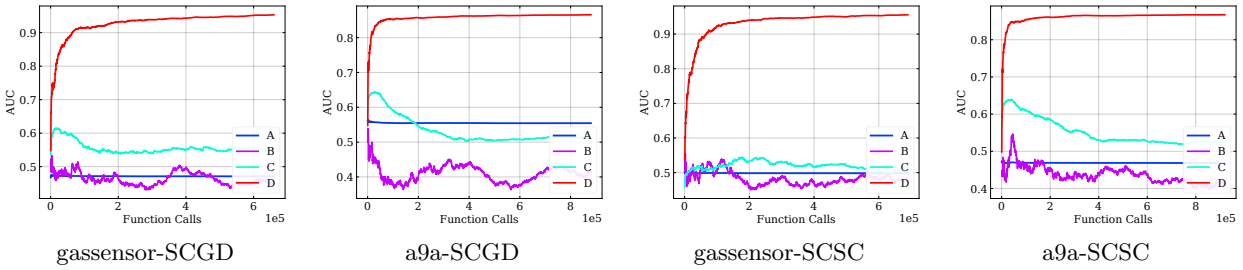

Figure 2: Ablation Experiment

Table 2: Comparison of the testing AUC values (mean ± std).

Dataset	ZO-Min-Max	ZOMDA	GFCOM	SCGD	SCSC	ZO-SCHT
ionosphere	0.905 ± 0.058	0.903 ± 0.023	0.905 ± 0.024	0.424 ± 0.141	0.615 ± 0.074	$\mathbf{0.920 \pm 0.046}$
diabetes	0.771 ± 0.021	0.776 ± 0.013	0.720 ± 0.042	0.584 ± 0.123	0.606 ± 0.056	$\mathbf{0.786 \pm 0.013}$
german	0.717 ± 0.020	0.720 ± 0.016	0.720 ± 0.053	0.465 ± 0.098	0.599 ± 0.063	$\mathbf{0.746 \pm 0.026}$
usps	0.839 ± 0.009	0.836 ± 0.008	0.836 ± 0.003	0.510 ± 0.099	0.447 ± 0.058	$\mathbf{0.860 \pm 0.003}$
gassensor	0.912 ± 0.007	0.906 ± 0.010	0.893 ± 0.005	0.471 ± 0.018	0.498 ± 0.091	$\mathbf{0.930 \pm 0.018}$
a9a	0.821 ± 0.008	0.846 ± 0.008	0.776 ± 0.009	0.554 ± 0.077	0.468 ± 0.143	$\mathbf{0.855 \pm 0.006}$
covtype	0.704 ± 0.004	0.722 ± 0.010	0.683 ± 0.021	0.502 ± 0.059	0.513 ± 0.080	$\mathbf{0.732 \pm 0.014}$
acoustic	0.829 ± 0.010	0.848 ± 0.006	0.809 ± 0.038	0.416 ± 0.172	0.519 ± 0.111	$\mathbf{0.861 \pm 0.003}$

5.2.2 Results and Discussion

From Table 2, we observe that ZO-SCHT achieves the best overall performance. Notably, the compositional baseline (GFCOM) underperforms the Min-Max methods, indicating that compositional structure alone does not ensure superior results. The motivation for adopting the compositional framework lies in its compatibility with sparsity constraints. Furthermore, the results of ZO-SCHT show that, under zeroth-order optimization, combining a compositional objective with hard-thresholding is essential for effective sparse AUC maximization. Moreover, we further assess algorithmic efficiency through the convergence curves of compositional optimization methods in Figure 1, which demonstrate that ZO-SCHT converges more efficiently and achieves the best overall performance.

Another notable observation from Table 2 is that Black-box SCGD and Black-box SCSC perform substantially worse. Although they share the SCGD framework with our approach, two key differences ex-

Table 3: Mean and Standard Deviation of Fooling Ratio (%) for Different Algorithms

	ZO-Min-Max	ZOMDA	GFCOM	SCGD	SCSC	ZO-SCHT
Fooling Ratio	28.56 ± 0.16	29.14 ± 0.25	26.75 ± 0.35	9.11 ± 3.52	10.34 ± 1.89	$\mathbf{30.20 \pm 0.38}$

plain this gap. Our zeroth-order gradient estimator incorporates dimension scaling of random directions, while Black-box SCGD and SCSC do not, and ZO-SCHT uses a constant step size, whereas Black-box SCGD and SCSC rely on decaying step sizes. To investigate which factor predominantly influences algorithmic performance, we conducted an ablation study under four configurations:

- A: no dimension scaling + decaying learning rate.
- B: no dimension scaling + constant learning rate.
- C: dimension scaling + decaying learning rate.
- D: dimension scaling + constant learning rate.

As illustrated in Figure 2, which reports results on the gassensor and a9a datasets, configurations A, B, and C fail to achieve effective optimization in Black-box SCGD and SCSC. In contrast, configuration D combines dimension-scaled gradient estimation with a constant learning rate and leads to substantial performance improvements. These results indicate that dimension scaling plays a crucial role in controlling the variance of zeroth-order gradient estimation, while a constant step size helps maintain sufficiently large update magnitudes throughout training. Only when both components are jointly incorporated can the algorithm effectively handle the challenging optimization landscape of black-box sparse AUC maximization. Additional sensitivity analyses of hyperparameters are provided in Appendix F.

5.3 Black-Box Adversarial Attacks

Universal Adversarial Perturbations (UAP) (Moosavi-Dezfooli et al., 2017) are input-agnostic: a single perturbation can cause a model to misclassify most natural images. Specifically, we aim to find a vector δ such that

$$F(\boldsymbol{x} + \delta) \neq F(\boldsymbol{x}) \quad \text{for most } \boldsymbol{x} \sim \mathcal{X},$$

where F is a classifier pretrained on the Cat&Dog dataset and is only accessible in a black-box manner, and $\boldsymbol{x}$ denotes an input image. The perturbation δ is treated as the parameter vector $\boldsymbol{w}$ optimized in (5). Since the goal is to induce misclassification (i.e., label flipping), the task is formulated as an AUC minimization problem, and the gradient descent step in the algorithm is replaced by gradient ascent.

5.3.1 Experimental Setup

Following the approach of (Liu et al., 2019), we adopt the post-Softmax probability of the positive class as the scoring function. We set the hyperparameters of ZO-SCHT as follows: $k = 1158, \eta = 5, \beta = 0.1, \mu = 0.01, q = 10, s_2 = d$(total number of pixels). Due to space limitations, the detailed hyperparameter tuning for other algorithms is provided in Appendix F. All methods are trained on 1,000 images and evaluated on a separate 1,000-image test set. Performance is measured by the fooling rate, i.e., the proportion of misclassified images post-perturbation. Each method is run three times with different seeds to report mean fooling ratio and standard deviation.

5.3.2 Results and Discussion

Table 3 reports the mean fooling ratios and standard deviations of different attack methods on the target black-box model. Consistent with Section 5.2.2, SCGD and SCSC perform poorly due to gradient estimation and decaying step sizes, while ZO-SCHT's sparse perturbation patterns achieve superior attack performance. Visualizations of the perturbations and attack outcomes are provided in Appendix G. All experiments are conducted on a server equipped with six NVIDIA 3090-24G GPUs.

6 CONCLUSION

In this paper, we propose a novel Zeroth-Order Stochastic Compositional Hard-Thresholding (ZO-SCHT) algorithm for black-box sparse AUC maximization. Specifically, by revisiting the compositional reformulation of nonlinear AUC objectives, we effectively integrate zeroth-order gradient estimation with hard-thresholding operators within the SCGD framework. Furthermore, we provide a rigorous convergence analysis of ZO-SCHT and show that it achieves linear convergence up to a tolerance bound under a fixed step size. Extensive experiments on black-box sparse AUC maximization and black-box adversarial attack tasks demonstrate the effectiveness and versatility of the proposed approach, highlighting the necessity of jointly incorporating compositional objectives, dimension-scaled zeroth-order gradient estimation, and hard-thresholding operators for efficient black-box sparse AUC optimization.

References

Asuncion, A., Newman, D., et al. (2007). Uci machine learning repository.

Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Chen, J., Chen, H., and Gu, B. (2024). How does black-box impact the learning guarantee of stochastic compositional optimization? *Advances in Neural Information Processing Systems*, 37:107745–107794.

Dang, Z., Li, X., Gu, B., Deng, C., and Huang, H. (2020). Large-scale nonlinear auc maximization via triply stochastic gradients. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1385–1398.

de Vazelhes, W., Zhang, H., Wu, H., Yuan, X., and Gu, B. (2022). Zeroth-order hard-thresholding: Gradient error vs. expansivity. *Advances in Neural Information Processing Systems*, 35:22589–22601.

Gao, X., Jiang, B., and Zhang, S. (2018). On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR.

Gultekin, S., Saha, A., Ratnaparkhi, A., and Paisley, J. (2020). Mba: mini-batch auc optimization. *IEEE transactions on neural networks and learning systems*, 31(12):5561–5574.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Huang, F., Gao, S., Pei, J., and Huang, H. (2022). Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70.

Jain, P., Tewari, A., and Kar, P. (2014). On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in neural information processing systems*, 27.

Liu, M., Yuan, Z., Ying, Y., and Yang, T. (2019). Stochastic auc maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*.

Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., and O'Reilly, U.-M. (2020). Min-max optimization without gradients: Convergence and applications to black-box evasion and poisoning attacks. In *International conference on machine learning*, pages 6282–6293. PMLR.

Liu, Z., Luo, L., and Low, B. K. H. (2024). Gradient-free methods for nonconvex nonsmooth stochastic compositional optimization. *Advances in Neural Information Processing Systems*, 37:45438–45461.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773.

Natole, M., Ying, Y., and Lyu, S. (2018). Stochastic proximal algorithms for auc maximization. In *International Conference on Machine Learning*, pages 3710–3719. PMLR.

Nguyen, N., Needell, D., and Woolf, T. (2017). Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63(11):6869–6895.

Schmidt, M. (2014). Convergence rate of stochastic gradient with constant step size.

Shen, J. and Li, P. (2018). A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18(208):1–42.

Wang, M., Fang, E. X., and Liu, H. (2017). Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1):419–449.

Wang, W., Liu, D., and Gu, B. (2026). Towards nonlinear sparse auc maximization via compositional stochastic hard thresholding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26489–26497.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Yang, Z., Zhou, B., Lei, Y., and Ying, Y. (2020). Stochastic hard thresholding algorithms for auc maximization. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 741–750. IEEE.

Ying, Y., Wen, L., and Lyu, S. (2016). Stochastic online auc maximization. *Advances in neural information processing systems*, 29.

Yuan, X., de Vazelhes, W., Gu, B., and Xiong, H. (2024). New insight of variance reduce in zero-order hard-thresholding: Mitigating gradient error and

expansivity contradictions. In *The Twelfth International Conference on Learning Representations*.

Yuan, X. and Li, P. (2021). Stability and risk bounds of iterative hard thresholding. In *International conference on artificial intelligence and statistics*, pages 1702–1710. PMLR.

Yuan, X.-T., Li, P., and Zhang, T. (2018). Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43.

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. (2021). Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049.

Zhou, L., Lai, K. K., and Yen, J. (2009). Credit scoring models with auc maximization based on weighted svm. *International journal of information technology & decision making*, 8(04):677–696.

Zhou, P., Yuan, X., and Feng, J. (2018). Efficient stochastic gradient hard thresholding. *Advances in Neural Information Processing Systems*, 31.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Notations and Definitions

Throughout this appendix, we will use the following notations:

- we denote the vectors in bold letters.
- $\nabla f(g)$ denotes the gradient of f at g, and $\nabla g(\boldsymbol{w})$ denotes the transpose of the Jacobian matrix of g at $\boldsymbol{w}$.
- $[d]$ denotes the set of all integers between 1 and d: $\{1, .., d\}$.
- u_i denotes the i-th coordinate of $\boldsymbol{u}$, and $\nabla g^i(\boldsymbol{w})$ represents the i-th column of $\nabla g(\boldsymbol{w})$.
- $\|\cdot\|_0$ denotes the ℓ_0 norm (which is not a proper norm).
- $\|a\|$ denotes the ℓ_2 norm when a is a vector, and the Frobenius norm when a is a matrix.
- $\text{supp}(\boldsymbol{w})$ denotes the support of a vector $\boldsymbol{w}$, which is the set of its non-zero coordinates.
- $|\mathcal{I}|$ denotes the cardinality (number of elements) of a set $\mathcal{I}$.
- All the sets we consider are subsets of $[d]$. So for a given set $\mathcal{I}$, $\mathcal{I}^c$ denotes the complement of $\mathcal{I}$ in $[d]$.
- $\mathcal{S}^d(R)$ denotes the d-sphere of radius R, that is $\mathcal{S}^d(R) = \{\boldsymbol{u} \in \mathbb{R}^d \mid \|\boldsymbol{u}\| = R\}$.
- $\mathcal{U}(\mathcal{S}^d)$ the uniform distribution on that unit sphere.
- $\mathcal{S}^d_S$ denotes a set called the restricted d-sphere on S, defined as: $\{\boldsymbol{u}_S \mid \boldsymbol{u} \in \{\boldsymbol{v} \in \mathbb{R}^d \mid \|\boldsymbol{v}_S\| = 1\}\}$, which is the set of unit vectors supported by S.
- $\mathcal{U}(\mathcal{S}^d_S)$ denotes the uniform distribution on the restricted sphere above.
- We denote by $\boldsymbol{u}_\mathcal{I}$ the hard-thresholding of the vector $\boldsymbol{u}$ over the support $\mathcal{I}$, that is, a vector that retains only the entries of $\boldsymbol{u}$ corresponding to the coordinates in $\mathcal{I}$ and sets all other coordinates to zero. Similarly, we denote by $\nabla_\mathcal{I} g(\boldsymbol{w})$ the hard-thresholding of the matrix $\nabla g(\boldsymbol{w})$ over the set $\mathcal{I}$, which retains only the rows whose indices belong to $\mathcal{I}$ and sets all other rows to zero.
- $\binom{[d]}{s}$ denotes the set of all subsets of $[d]$ that contain s elements: $\binom{[d]}{s} = \{S : |S| = s, S \subseteq [d]\}$.
- $\mathcal{U}(\binom{[d]}{s})$ denotes the uniform distribution on the set above.
- $\boldsymbol{I}$ denotes the identity matrix $\boldsymbol{I}_{d\times d}$.
- $\boldsymbol{I}_S$ denotes the identity matrix with 1 on the diagonal only at indices belonging to the support S: $\boldsymbol{I}_{i,i} = 1$ if $i \in S$, and 0 elsewhere.
- $(\boldsymbol{u}_i)_{i=1}^n$ denotes the n-uple of elements $\boldsymbol{u}_1, .., \boldsymbol{u}_n$.
- log denotes the natural logarithm (in base e).

B Auxilliary Lemmas

Lemma B.1 (Gao et al. (2018), Lemma 7.3.b)**.**

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}^d)}\boldsymbol{u}\boldsymbol{u}^T = \frac{1}{d}\boldsymbol{I}. \tag{14}$$

Lemma B.2 (de Vazelhes et al. (2022), Lemma C.1)**.** *Let $\mathcal{I}$ be a subset of $[d]$ with size s, and let $\boldsymbol{u} \sim \mathcal{U}(\mathcal{S}_S^d)$, where $|S| = s_2$. The following holds:*

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

Lemma B.3 (Shen and Li (2018), Theorem 1)**.** *Let $\boldsymbol{b} \in \mathbb{R}^d$ be an arbitrary vector and $\boldsymbol{x} \in \mathbb{R}^d$ be any K-sparse signal. For any $k \geq K$, we have the following bound:*

$$\begin{aligned} \|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{x}\|_2^2 &\leq \nu\|\boldsymbol{b} - \boldsymbol{x}\|_2^2, \\ \nu = 1 + \frac{\delta + \sqrt{(4+\delta)\delta}}{2}, \quad \delta &= \frac{\min\{K, d-k\}}{k - K + \min\{K, d-k\}}. \end{aligned} \tag{16}$$

Corollary B.1. *With the notations and variables above in Lemma B.3, we also have the following, simpler bound, from Shen and Li (2018):*

$$\|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{x}\|_2^2 \leq \gamma\|\boldsymbol{b} - \boldsymbol{x}\|_2^2$$

with

$$\gamma = 1 + \left(K/k + \sqrt{(4 + K/k)\,K/k}\right)/2 \tag{17}$$

Proof. There are two possibilities for δ in Lemma B.3: either $\delta = \frac{K}{k}$ (if $d - k > K$) or $\delta = \frac{d-k}{d-K}$ (if $d - k \leq K$). In the latter case:

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

Therefore, in both cases, we have $\delta \leq \frac{K}{k}$, which, when substituted into Lemma B.3, yields Corollary B.1. □

Lemma B.4 (de Vazelhes et al. (2022), Lemma C.2)**.** *Let $\mathcal{I}$ be a subset of $[d]$ with size s, and let $\boldsymbol{u} \sim \mathcal{U}(\mathcal{S}_S^d)$, where $|S| = s_2$. The following holds:*

$$\begin{aligned} &\mathbb{E}_{S\sim\binom{[d]}{s_2}}\left[\nabla f(\boldsymbol{x})^T\left(\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}\boldsymbol{u}^T\|\boldsymbol{u}_{\mathcal{I}}\|^2\right)\nabla f(\boldsymbol{x})\right] \\ =&\frac{1}{d(s_2+2)}\left[\|\nabla_{\mathcal{I}} f(\boldsymbol{x})\|^2\left(\frac{(s-1)(s_2-1)}{d-1} + 3\right) + \|\nabla_{\mathcal{I}^c} f(\boldsymbol{x})\|^2\left(\frac{s(s_2-1)}{d-1}\right)\right] \end{aligned} \tag{18}$$

Lemma B.5 (Wang et al. (2017), Lemma 2)**.** *Let $(\boldsymbol{w}_t, \boldsymbol{v}_t)$ be generated by Algorithm 1. We have:*

$$\begin{aligned} &\mathbb{E}\left\|\boldsymbol{v}_{t+1} - g(\boldsymbol{w}_t)\right\|^2 \\ \leq&(1-\beta)\left\|\boldsymbol{v}_t - g\left(\boldsymbol{w}_{t-1}\right)\right\|^2 + \beta^{-1}C_g\left\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\right\|^2 + 2V_g\beta^2 \end{aligned} \tag{19}$$

C Proof of Proposition 1

The objective function for AUC maximization in Equation (1) can be decomposed into three terms:

$$
\begin{aligned}
F(\boldsymbol{w}) =& \frac{1}{n_+}\frac{1}{n_-}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(1-h(\boldsymbol{w};\boldsymbol{x}_i)+h(\boldsymbol{w};\boldsymbol{x}_j)\right)^2\mathbb{I}_{[y_i=1]}\mathbb{I}_{[y_j=-1]} \\
=& \frac{1}{n_+}\frac{1}{n_-}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[(1+\mu_-(\boldsymbol{w})-\mu_+(\boldsymbol{w})-h(\boldsymbol{w};\boldsymbol{x}_i)+\mu_+(\boldsymbol{w}) \\
&\qquad +h(\boldsymbol{w};\boldsymbol{x}_j)-\mu_-(\boldsymbol{w}))^2\Big]\mathbb{I}_{[y_i=1]}\mathbb{I}_{[y_j=-1]} \\
=& \underbrace{\frac{1}{n_+}\frac{1}{n_-}\sum_{i=1}^{n}\sum_{j=1}^{n}(1+\mu_-(\boldsymbol{w})-\mu_+(\boldsymbol{w}))^2\mathbb{I}_{[y_i=1]}\mathbb{I}_{[y_j=-1]}}_{=:R_1} \\
&+\underbrace{\frac{1}{n_+}\frac{1}{n_-}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(h(\boldsymbol{w};\boldsymbol{x}_i)-\mu_+(\boldsymbol{w})-h(\boldsymbol{w};\boldsymbol{x}_j)+\mu_-(\boldsymbol{w})\right)^2\mathbb{I}_{[y_i=1]}\mathbb{I}_{[y_j=-1]}}_{=:R_2} \\
&-2\frac{1}{n_+}\frac{1}{n_-}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[\Big(1+\mu_-(\boldsymbol{w})-\mu_+(\boldsymbol{w})\Big)\cdot \\
&\qquad \Big(h(\boldsymbol{w};\boldsymbol{x}_i)-\mu_+(\boldsymbol{w})-h(\boldsymbol{w};\boldsymbol{x}_j)+\mu_-(\boldsymbol{w})\Big)\Big]\mathbb{I}_{[y_i=1]}\mathbb{I}_{[y_j=-1]}
\end{aligned}
$$

The last intersection term is denoted as R_3. It is sufficient to estimate each of the terms individually. To this end, the first term has n_+n_- same terms, so we have

$$R_1 = (1+\mu_-(\boldsymbol{w})-\mu_+(\boldsymbol{w}))^2.$$

For the second term, note that the cross term

$$
\begin{aligned}
R_4 &= \frac{1}{n_+}\frac{1}{n_-}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(h(\boldsymbol{w};\boldsymbol{x}_i)-\mu_+(\boldsymbol{w})\right)\left(h(\boldsymbol{w};\boldsymbol{x}_j)-\mu_-(\boldsymbol{w})\right)\mathbb{I}_{[y_i=1]}\mathbb{I}_{[y_j=-1]} \\
&= \left[\frac{1}{n_+}\sum_{i=1}^{n}(h(\boldsymbol{w};\boldsymbol{x}_i)-\mu_+(\boldsymbol{w}))\mathbb{I}_{[y_i=1]}\right]\left[\frac{1}{n_-}\sum_{j=1}^{n}(h(\boldsymbol{w};\boldsymbol{x}_j)-\mu_-(\boldsymbol{w}))\mathbb{I}_{[y_j=-1]}\right] \\
&= \left[\frac{1}{n_+}\sum_{i=1}^{n}h(\boldsymbol{w};\boldsymbol{x}_i)\mathbb{I}_{[y_i=1]}-\mu_+(\boldsymbol{w})\right]\left[\frac{1}{n_-}\sum_{j=1}^{n}h(\boldsymbol{w};\boldsymbol{x}_j)\mathbb{I}_{[y_j=-1]}-\mu_-(\boldsymbol{w})\right] \\
&\overset{(2)}{=} \left[\mu_+(\boldsymbol{w})-\mu_+(\boldsymbol{w})\right]\left[\mu_-(\boldsymbol{w})-\mu_-(\boldsymbol{w})\right] \\
&= 0,
\end{aligned}
$$

It follows that

$$
\begin{aligned}
R_2 =& \frac{1}{n_+}\frac{1}{n_-}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(h(\boldsymbol{w};\boldsymbol{x}_i)-\mu_+(\boldsymbol{w})\right)^2+\left(h(\boldsymbol{w};\boldsymbol{x}_j)-\mu_-(\boldsymbol{w})\right)^2\right]\mathbb{I}_{[y_i=1]}\mathbb{I}_{[y_j=-1]} \\
=& \frac{1}{n_+}\sum_{i=1}^{n}(h(\boldsymbol{w};\boldsymbol{x}_i)-\mu_+(\boldsymbol{w}))^2\mathbb{I}_{[y_i=1]}+\frac{1}{n_-}\sum_{j=1}^{n}(h(\boldsymbol{w};\boldsymbol{x}_j)-\mu_-(\boldsymbol{w}))^2\mathbb{I}_{[y_j=-1]}.
\end{aligned}
$$

For the third term, using a similar derivation to that of the cross term in the second term, we obtain $R_3 = 0$; thus,

$$\begin{aligned}
F(\boldsymbol{w}) =&(1+\mu_-(\boldsymbol{w})-\mu_+(\boldsymbol{w}))^2+\frac{1}{n_+}\sum_{i=1}^{n}(h(\boldsymbol{w};\boldsymbol{x}_i)-\mu_+(\boldsymbol{w}))^2\mathbb{I}_{[y_i=1]}\\
&+\frac{1}{n_-}\sum_{j=1}^{n}(h(\boldsymbol{w};\boldsymbol{x}_j)-\mu_-(\boldsymbol{w}))^2\mathbb{I}_{[y_j=-1]}\\
=&\frac{1}{n}\sum_{i=1}^{n}\left[(1+\mu_-(\boldsymbol{w})-\mu_+(\boldsymbol{w}))^2+\frac{n}{n_+}(h(\boldsymbol{w};\boldsymbol{x}_i)-\mu_+(\boldsymbol{w}))^2\mathbb{I}_{[y_i=1]}\right.\\
&\left.+\frac{n}{n_-}(h(\boldsymbol{w};\boldsymbol{x}_i)-\mu_-(\boldsymbol{w}))^2\mathbb{I}_{[y_i=-1]}\right]\\
\overset{(6)}{=}&\frac{1}{n}\sum_{i=1}^{n}f_i(g(\boldsymbol{w})).
\end{aligned}$$

□

D Proof of Proposition 2

First, we derive in section D.1 the error of the gradient estimate if we sample only one direction ($q = 1$). Then, in section D.2, we show how sampling q directions reduces the error of the gradient estimator, producing the results of Proposition 2.

D.1 One direction estimator

Throughout this section, we assume that $q = 1$ for the gradient estimator $\hat{\nabla} g(\boldsymbol{w})$ and $\hat{\nabla} f(g)$.

D.1.1 The internal function $g(\boldsymbol{w})$.

Lemma D.1. *For any (L_{s_2}, s_2)-RSS function $h_j(\boldsymbol{w})$, using the gradient estimator $\widehat{\nabla} g_j(\boldsymbol{w})$ defined in (7) with $q = 1$, we have, for any support $\mathcal{I} \in [d]$, with $|\mathcal{I}| = s$:*

(i) $\|\mathbb{E}\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w}) - \nabla_{\mathcal{I}} g_j(\boldsymbol{w})\|^2 \leq \varepsilon_\mu \mu^2$

(ii) $\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w})\right\|^2 \leq \varepsilon_{\mathcal{I}} \|\nabla g_j(\boldsymbol{w})\|^2 + 2\varepsilon_\mu \mu^2$

with $a_j = \frac{\mathbb{I}_{[y_j=1]}}{r}, b_j = \frac{\mathbb{I}_{[y_j=-1]}}{1-r}, \varepsilon_\mu = ds(a_j^2+b_j^2)L_{s_2}^2, \varepsilon_{\mathcal{I}} = \frac{2d}{(s_2+2)}\left(\frac{(s-1)(s_2-1)}{d-1}+3.\right)$

Proof. We first prove (i). From the definition of the gradient estimator, we have

$$\begin{aligned}
\mathbb{E}\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w}) =& \mathbb{E}d\frac{\boldsymbol{u}_{\mathcal{I}}\left(g_j(\boldsymbol{w}+\mu\boldsymbol{u})-g_j(\boldsymbol{w})\right)^\top}{\mu}\\
=&\mathbb{E}\frac{d}{\mu}\boldsymbol{u}_{\mathcal{I}}\left[\begin{pmatrix}\boldsymbol{w}+\mu\boldsymbol{u}\\ \frac{1}{r}h(\boldsymbol{w}+\mu\boldsymbol{u};\boldsymbol{x}_j)\mathbb{I}_{[y_j=1]}\\ \frac{1}{1-r}h(\boldsymbol{w}+\mu\boldsymbol{u};\boldsymbol{x}_j)\mathbb{I}_{[y_j=-1]}\end{pmatrix}-\begin{pmatrix}\boldsymbol{w}\\ \frac{1}{r}h(\boldsymbol{w};\boldsymbol{x}_j)\mathbb{I}_{[y_j=1]}\\ \frac{1}{1-r}h(\boldsymbol{w};\boldsymbol{x}_j)\mathbb{I}_{[y_j=-1]}\end{pmatrix}\right]^\top\\
=&\mathbb{E}\frac{d}{\mu}\boldsymbol{u}_{\mathcal{I}}\begin{pmatrix}\mu\boldsymbol{u}\\ \frac{\mathbb{I}_{[y_j=1]}}{r}\left(h(\boldsymbol{w}+\mu\boldsymbol{u};\boldsymbol{x}_j)-h(\boldsymbol{w};\boldsymbol{x}_j)\right)\\ \frac{\mathbb{I}_{[y_j=-1]}}{1-r}\left(h(\boldsymbol{w}+\mu\boldsymbol{u};\boldsymbol{x}_j)-h(\boldsymbol{w};\boldsymbol{x}_j)\right)\end{pmatrix}^\top\\
=&\mathbb{E}d\left(\boldsymbol{u}_{\mathcal{I}}\boldsymbol{u}^\top\quad\frac{\mathbb{I}_{[y_j=1]}}{r}\frac{h_j(\boldsymbol{w}+\mu\boldsymbol{u})-h_j(\boldsymbol{w})}{\mu}\boldsymbol{u}_{\mathcal{I}}\quad\frac{\mathbb{I}_{[y_j=-1]}}{1-r}\frac{h_j(\boldsymbol{w}+\mu\boldsymbol{u})-h_j(\boldsymbol{w})}{\mu}\boldsymbol{u}_{\mathcal{I}}\right)
\end{aligned}$$

$$\begin{aligned}
&\overset{(\xi_1)}{=} d\left(\mathbb{E}\boldsymbol{u}_{\mathcal{I}}\boldsymbol{u}^\top \quad \frac{\mathbb{I}_{[y_j=1]}}{r}\mathbb{E}\langle\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}} \quad \frac{\mathbb{I}_{[y_j=-1]}}{1-r}\mathbb{E}\langle\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}}\right)\\
&\overset{(\xi_2)}{=} d\left(\frac{1}{d}\boldsymbol{I}_{\mathcal{I}} \quad \frac{\mathbb{I}_{[y_j=1]}}{r}\mathbb{E}\langle\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}} \quad \frac{\mathbb{I}_{[y_j=-1]}}{1-r}\mathbb{E}\langle\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}}\right)\\
&= \left(\boldsymbol{I}_{\mathcal{I}} \quad \frac{\mathbb{I}_{[y_j=1]}}{r}d\mathbb{E}\langle\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}} \quad \frac{\mathbb{I}_{[y_j=-1]}}{1-r}d\mathbb{E}\langle\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}}\right)
\end{aligned}$$

ξ_1 comes from the (L_{s_2}, s_2)-RSS condition, which implies continuous differentiability in any s_2-sparse direction (since (L_{s_2}, s_2)-RSS is equivalent to the Lipschitz continuity of the gradient over any s_2-sparse set, ensuring that the gradient is continuous on these sets). Therefore, by the mean value theorem, for some $c \in [0, \mu]$, we have: $\frac{f(x+\mu u)-f(x)}{\mu} = \langle\nabla f(x+cu), u\rangle$. And ξ_2 comes from

$$\begin{aligned}
\mathbb{E}\boldsymbol{u}_{\mathcal{I}}\boldsymbol{u}^T = \mathbb{E}\mathcal{P}_{\mathcal{I}}\left(\boldsymbol{u}\boldsymbol{u}^T\right) &= \mathcal{P}_{\mathcal{I}}\left(\mathbb{E}\boldsymbol{u}\boldsymbol{u}^T\right)\\
&= \mathcal{P}_{\mathcal{I}}\left(\mathbb{E}_{S\sim\binom{[d]}{s_2}}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{U}(\mathcal{S}_S^d)}\boldsymbol{u}\boldsymbol{u}^T\right) \overset{(\xi_3)}{=} \mathcal{P}_{\mathcal{I}}\left(\mathbb{E}_{S\sim\binom{[d]}{s_2}}\frac{1}{s_2}\boldsymbol{I}_S\right)\\
&= \mathcal{P}_{\mathcal{I}}\left(\frac{1}{s_2}\mathbb{E}_{S\sim\binom{[d]}{s_2}}\boldsymbol{I}_S\right) \overset{(\xi_4)}{=} \mathcal{P}_{\mathcal{I}}\left(\frac{1}{s_2}\frac{s_2}{d}\boldsymbol{I}\right) = \frac{1}{d}\boldsymbol{I}_{\mathcal{I}}
\end{aligned}$$

Where ξ_3 comes from applying Lemma B.1 to the unit sub-sphere on the support S, and ξ_4 follows from noting that each diagonal element with index i follows a Bernoulli distribution with parameter $\frac{s_2}{d}$. Indeed, there are $\binom{d-1}{s_2-1}$ support configurations containing i, out of $\binom{d}{s_2}$ total configurations, giving a probability $p = \frac{\binom{d-1}{s_2-1}}{\binom{d}{s_2}} = \frac{(d-1)!s_2!(d-s_2)!}{(s_2-1)!(d-1-(s_2-1))!d!} = \frac{s_2}{d}$ of the value 1 occurring at position i. By the definition of the inner function of $g(\boldsymbol{w})$, we have:

$$\begin{aligned}
\nabla_{\mathcal{I}} g_j(\boldsymbol{w}) = P_{\mathcal{I}}\left(\left(\partial g_j(\boldsymbol{w})\right)^\top\right) &= \left(\boldsymbol{I} \quad \frac{\mathbb{I}_{[y_j=1]}}{r}\nabla h_j(\boldsymbol{w}) \quad \frac{\mathbb{I}_{[y_j=-1]}}{1-r}\nabla h_j(\boldsymbol{w})\right)_{\mathcal{I}}\\
&\overset{(14)}{=} \left(\boldsymbol{I} \quad \frac{\mathbb{I}_{[y_j=1]}}{r}d\mathbb{E}\boldsymbol{u}\boldsymbol{u}^T\nabla h_j(\boldsymbol{w}) \quad \frac{\mathbb{I}_{[y_j=-1]}}{1-r}d\mathbb{E}\boldsymbol{u}\boldsymbol{u}^T\nabla h_j(\boldsymbol{w})\right)_{\mathcal{I}}\\
&= \left(\boldsymbol{I}_{\mathcal{I}} \quad \frac{\mathbb{I}_{[y_j=1]}}{r}d\mathbb{E}\langle\nabla h_j(\boldsymbol{w}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}} \quad \frac{\mathbb{I}_{[y_j=-1]}}{1-r}d\mathbb{E}\langle\nabla h_j(\boldsymbol{w}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}}\right)
\end{aligned}$$

Let $a_j = \frac{\mathbb{I}_{[y_j=1]}}{r}, b_j = \frac{\mathbb{I}_{[y_j=-1]}}{1-r}$, we have

$$\begin{aligned}
&\left\|\mathbb{E}\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w}) - \nabla_{\mathcal{I}} g_j(\boldsymbol{w})\right\|^2\\
=&d^2\left\|a_j\mathbb{E}\langle\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}) - \nabla h_j(\boldsymbol{w}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}}\right\|^2 + d^2\left\|b_j\mathbb{E}\langle\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}) - \nabla h_j(\boldsymbol{w}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}}\right\|^2\\
=&d^2\left(a_j^2+b_j^2\right)\left\|\mathbb{E}\langle\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}) - \nabla h_j(\boldsymbol{w}),\boldsymbol{u}\rangle\boldsymbol{u}_{\mathcal{I}}\right\|^2\\
\leq&d^2\left(a_j^2+b_j^2\right)\mathbb{E}\|\nabla h_j(\boldsymbol{w}+c\boldsymbol{u}) - \nabla h_j(\boldsymbol{w})\|^2\|\boldsymbol{u}\|^2\|\boldsymbol{u}_{\mathcal{I}}\|^2\\
\leq&d^2\left(a_j^2+b_j^2\right)L_{s_2}^2\mathbb{E}\|c\boldsymbol{u}\|^2\|\boldsymbol{u}\|^2\|\boldsymbol{u}_{\mathcal{I}}\|^2\\
\leq&d^2\left(a_j^2+b_j^2\right)L_{s_2}^2\mu^2\mathbb{E}\|\boldsymbol{u}_{\mathcal{I}}\|^2\\
\overset{(15)}{=}&d^2\left(a_j^2+b_j^2\right)L_{s_2}^2\mu^2\frac{s}{d}\\
=&ds\left(a_j^2+b_j^2\right)L_{s_2}^2\mu^2
\end{aligned}$$

□

Now, we prove (ii).

$$\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w})\right\|^2 = \mathbb{E}\left\|d\boldsymbol{u}_{\mathcal{I}}\frac{\left(g_j(\boldsymbol{w}+\mu\boldsymbol{u}) - g_j(\boldsymbol{w})\right)^\top}{\mu}\right\|^2$$

Let $N = (d+2)$. Thus, we have

$$g_j(\boldsymbol{w}) = \begin{pmatrix} g_j^1(\boldsymbol{w}) \\ \cdots \\ g_j^N(\boldsymbol{w}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{w} \\ a_j h(\boldsymbol{w}; \boldsymbol{x}_j) \\ b_j h(\boldsymbol{w}; \boldsymbol{x}_j) \end{pmatrix}$$

For $1 \leq i \leq d$, $g_j^i(\boldsymbol{w}) = w_i$; for $i = d+1, g_j^i(\boldsymbol{w}) = a_j h(\boldsymbol{w}; \boldsymbol{x}_j)$; and for $i = d+2, g_j^i(\boldsymbol{w}) = b_j h(\boldsymbol{w}; \boldsymbol{x}_j)$. Therefore, we have:

$$\frac{g_j(\boldsymbol{w}+\mu\boldsymbol{u}) - g_j(\boldsymbol{w})}{\mu} = \begin{pmatrix} u_1 \\ \cdots \\ u_d \\ \frac{a_j(h_j(\boldsymbol{w}+\mu\boldsymbol{u}) - h_j(\boldsymbol{w}))}{\mu} \\ \frac{b_j(h_j(\boldsymbol{w}+\mu\boldsymbol{u}) - h_j(\boldsymbol{w}))}{\mu} \end{pmatrix} \overset{(\xi_5)}{=} \begin{pmatrix} \langle \nabla g_j^1(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \\ \cdots \\ \langle \nabla g_j^d(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \\ \langle \nabla g_j^{d+1}(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \\ \langle \nabla g_j^{d+2}(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \end{pmatrix}$$

Where ξ_5 originates from the fact that, for $1 \leq i \leq d$, $\nabla g_j^i(\boldsymbol{w}) = \boldsymbol{e}_i$ (where only the i-th element equals 1 and all others are 0); for $i = d+1, \nabla g_j^i(\boldsymbol{w}) = a_j \nabla h_j(\boldsymbol{w})$; and for $i = d+2, \nabla g_j^i(\boldsymbol{w}) = b_j \nabla h_j(\boldsymbol{w})$. It follows that

$$\frac{g_j(\boldsymbol{w}+\mu\boldsymbol{u}) - g_j(\boldsymbol{w})}{\mu} = \begin{pmatrix} \langle \nabla g_j^1(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \\ \cdots \\ \langle \nabla g_j^N(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \end{pmatrix}$$

Therefore, we have

$$\begin{aligned} \mathbb{E} \left\| \hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w}) \right\|^2 =& d^2 \mathbb{E} \left\| \langle \nabla g_j^1(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \boldsymbol{u}_{\mathcal{I}}, \cdots, \langle \nabla g_j^N(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \boldsymbol{u}_{\mathcal{I}} \right\|^2 \\ =& d^2 \mathbb{E} \sum_{i=1}^{N} \left\| \langle \nabla g_j^i(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \boldsymbol{u}_{\mathcal{I}} \right\|^2 \end{aligned}$$

For $\mathbb{E} \left\| \langle \nabla g_j^i(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \boldsymbol{u}_{\mathcal{I}} \right\|^2$, we have

$$\begin{aligned} & \mathbb{E} \left\| \langle \nabla g_j^i(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle \boldsymbol{u}_{\mathcal{I}} \right\|^2 \\ =& \mathbb{E} \left[\langle \nabla g_j^i(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle - \langle \nabla g_j^i(\boldsymbol{w}), \boldsymbol{u} \rangle + \langle \nabla g_j^i(\boldsymbol{w}), \boldsymbol{u} \rangle \right]^2 \|\boldsymbol{u}_{\mathcal{I}}\|^2 \\ \leq& 2\mathbb{E} \left[\nabla g_j^i(\boldsymbol{w}+c\boldsymbol{u}), \boldsymbol{u} \rangle - \langle \nabla g_j^i(\boldsymbol{w}), \boldsymbol{u} \rangle \right]^2 \|\boldsymbol{u}_{\mathcal{I}}\|^2 + 2\mathbb{E} \langle \nabla g_j^i(\boldsymbol{w}), \boldsymbol{u} \rangle^2 \|\boldsymbol{u}_{\mathcal{I}}\|^2 \\ \leq& 2\mathbb{E} \|\nabla g_j^i(\boldsymbol{w}+c\boldsymbol{u}) - \nabla g_j^i(\boldsymbol{w})\|^2 \|\boldsymbol{u}\|^2 \|\boldsymbol{u}_{\mathcal{I}}\|^2 + 2\mathbb{E} \left(\nabla g_j^i(\boldsymbol{w}) \right)^\top \boldsymbol{u}\boldsymbol{u}^\top \nabla g_j^i(\boldsymbol{w}) \|\boldsymbol{u}_{\mathcal{I}}\|^2 \end{aligned}$$

For $1 \leq i \leq d$, we have $\nabla g_j^i(\boldsymbol{w}) = \boldsymbol{e}_i$; therefore,

$$\mathbb{E} \left\| \nabla g_j^i(\boldsymbol{w}+c\boldsymbol{u}) - \nabla g_j^i(\boldsymbol{w}) \right\|^2 = 0,$$

and for $i = d+1$, we have

$$\begin{aligned} & 2\mathbb{E} \|\nabla g_j^i(\boldsymbol{w}+c\boldsymbol{u}) - \nabla g_j^i(\boldsymbol{w})\|^2 \|\boldsymbol{u}\|^2 \|\boldsymbol{u}_{\mathcal{I}}\|^2 \\ =& 2\mathbb{E} \|a_j \nabla h_j(\boldsymbol{w}+c\boldsymbol{u}) - a_j \nabla h_j(\boldsymbol{w})\|^2 \|\boldsymbol{u}\|^2 \|\boldsymbol{u}_{\mathcal{I}}\|^2 \\ \leq& 2\mathbb{E} a_j^2 L_{s_2}^2 \|c\boldsymbol{u}\|^2 \|\boldsymbol{u}\|^2 \|\boldsymbol{u}_{\mathcal{I}}\|^2 \\ \overset{(15)}{=}& \frac{2sa_j^2 L_{s_2}^2 \mu^2}{d} \end{aligned}$$

Similarly, for $i = d+2$, we have

$$2\mathbb{E}\left[\|\nabla g_j^i(\boldsymbol{w}+c\boldsymbol{u}) - \nabla g_j^i(\boldsymbol{w})\|^2 \|\boldsymbol{u}\|^2 \|\boldsymbol{u}_{\mathcal{I}}\|^2\right] \leq \frac{2sb_j^2 L_{s_2}^2 \mu^2}{d}.$$

For $2\mathbb{E}\left[\left(\nabla g_j^i(\boldsymbol{w})\right)^\top \boldsymbol{u}\,\boldsymbol{u}^\top \nabla g_j^i(\boldsymbol{w})\,\|\boldsymbol{u}_{\mathcal{I}}\|^2\right]$, by Lemma B.4, we have

$$\begin{aligned}
&2\mathbb{E}\left(\nabla g_j^i(\boldsymbol{w})\right)^\top \boldsymbol{u}\boldsymbol{u}^\top \nabla g_j^i(\boldsymbol{w}) \|\boldsymbol{u}_{\mathcal{I}}\|^2 \\
=&\frac{2}{d(s_2+2)}\left[\|\nabla_{\mathcal{I}} g_j^i(\boldsymbol{w})\|^2\left(\frac{(s-1)(s_2-1)}{d-1}+3\right) + \|\nabla_{\mathcal{I}^c} g_j^i(\boldsymbol{w})\|^2\left(\frac{s(s_2-1)}{d-1}\right)\right] \\
\overset{\xi_6}{\leq}&\frac{2}{d(s_2+2)}\left(\frac{(s-1)(s_2-1)}{d-1}+3\right)\|\nabla g_j^i(\boldsymbol{w})\|^2
\end{aligned}$$

Where ξ_6 follows from the inequality

$$\frac{s(s_2-1)}{d-1} \leq \frac{(s-1)(s_2-1)}{d-1} + 3.$$

It then follows that

$$\begin{aligned}
\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w})\right\|^2 \leq& 2ds(a_j^2+b_j^2)L_{s_2}^2\mu^2 + \frac{2d}{(s_2+2)}\left(\frac{(s-1)(s_2-1)}{d-1}+3\right)\sum_{i=1}^N \|\nabla g_j^i(\boldsymbol{w})\|^2 \\
=&2ds(a_j^2+b_j^2)L_{s_2}^2\mu^2 + \frac{2d}{(s_2+2)}\left(\frac{(s-1)(s_2-1)}{d-1}+3\right)\|\nabla g_j(\boldsymbol{w})\|^2
\end{aligned}$$

D.1.2 The outer function $f(g)$.

Lemma D.2. *We assume that each outer function f_i is Lipschitz smooth with constant $L_f \geq 0$. Using the gradient estimator $\widehat{\nabla} f_i(g)$ defined in (8) with $q = 1$, we have:*

(i) $\|\mathbb{E}\hat{\nabla} f_i(g) - \nabla f_i(g)\|^2 \leq \varepsilon_{abs}\mu^2$

(ii) $\mathbb{E}\|\hat{\nabla} f_i(g)\|^2 \leq \varepsilon\|\nabla f_i(g)\|^2 + 2\varepsilon_{abs}\mu^2$

with $\varepsilon_{abs} = (d+2)^2 L_f^2$, $\varepsilon = 2(d+2)$

Proof. We first prove (i). From the definition of the gradient estimator, we have

$$\begin{aligned}
\|\mathbb{E}\hat{\nabla} f_i(g) - \nabla f_i(g)\|^2 &= \left\|\mathbb{E}(d+2)\frac{f_i(g+\mu\boldsymbol{z})-f_i(g)}{\mu}\boldsymbol{z} - \nabla f_i(g)\right\|^2 \\
&\overset{(14)}{=} \left\|\mathbb{E}(d+2)\frac{f_i(g+\mu\boldsymbol{z})-f_i(g)}{\mu}\boldsymbol{z} - (d+2)\mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top\nabla f_i(g)\right\|^2 \\
&\overset{\xi_7}{=} (d+2)^2\left\|\mathbb{E}\langle\nabla f_i(g+c\boldsymbol{z}),\boldsymbol{z}\rangle\boldsymbol{z} - \langle\nabla f_i(g),\boldsymbol{z}\rangle\boldsymbol{z}\right\|^2 \\
&\leq (d+2)^2\mathbb{E}\|\nabla f_i(g+c\boldsymbol{z}) - \nabla f_i(g)\|^2\|\boldsymbol{z}\|^2\|\boldsymbol{z}\|^2 \\
&\leq (d+2)^2 L_f^2\mu^2
\end{aligned}$$

ξ_7 follows from the fact that each outer function $f_i(g)$ is Lipschitz smooth with constant $L_f \geq 0$, which implies that ∇f_i is Lipschitz continuous, and hence f_i is continuously differentiable. Therefore, by the mean value theorem, there exists some $c \in [0,\mu]$ such that $\frac{f_i(g+\mu\mathbf{z})-f_i(g)}{\mu} = \langle\nabla f_i(g+c\mathbf{z}),\mathbf{z}\rangle$.

We now prove (ii).

$$\begin{aligned}
\mathbb{E}\|\hat{\nabla} f_i(g)\|^2 =& \mathbb{E}\left\|(d+2)\frac{f_i(g+\mu\boldsymbol{z})-f_i(g)}{\mu}\boldsymbol{z}\right\|^2 \\
=&(d+2)^2\mathbb{E}\|\langle\nabla f_i(g+c\boldsymbol{z}),\boldsymbol{z}\rangle\boldsymbol{z}\|^2 \\
=&(d+2)^2\mathbb{E}\left[\langle\nabla f_i(g+c\boldsymbol{z}),\boldsymbol{z}\rangle-\langle\nabla f_i(g),\boldsymbol{z}\rangle+\langle\nabla f_i(g),\boldsymbol{z}\rangle\right]^2\|\boldsymbol{z}\|^2 \\
\leq&2(d+2)^2\mathbb{E}\left[\langle\nabla f_i(g+c\boldsymbol{z})-\nabla f_i(g),\boldsymbol{z}\rangle^2+\langle\nabla f_i(g),\boldsymbol{z}\rangle^2\right] \\
\leq&2(d+2)^2\mathbb{E}\left[\|\nabla f_i(g+c\boldsymbol{z})-\nabla f_i(g)\|^2\|\boldsymbol{z}\|^2+\nabla f_i(g)^\top\boldsymbol{z}\boldsymbol{z}^\top\nabla f_i(g)\right] \\
\leq&2(d+2)^2\left[L_f^2\mu^2+\frac{1}{d+2}\|\nabla f_i(g)\|^2\right] \\
\leq&2(d+2)^2L_f^2\mu^2+2(d+2)\|\nabla f_i(g)\|^2
\end{aligned}$$

□

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

We now describe how sampling $q\geq 1$ random directions can enhance the accuracy of the gradient estimation.

Lemma D.3. *Let $\mathcal{I}\subset[d]$ be an arbitrary support set of size s (i.e., $|\mathcal{I}|=s$). We assume that each outer function f_i is Lipschitz smooth with constant $L_f\geq 0$, and each scoring function $h_j(\boldsymbol{w})$ is (L_{s_2},s_2)-RSS. By Lemma D.1 and D.2, for the zeroth-order gradient estimator defined in equations (7) and (8) with $q=1$, we have*

(i) $\|\mathbb{E}\hat{\nabla}_{\mathcal{I}}g_j(\boldsymbol{w})-\nabla_{\mathcal{I}}g_j(\boldsymbol{w})\|^2\leq\varepsilon_\mu\mu^2$

(ii) $\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}}g_j(\boldsymbol{w})\right\|^2\leq\varepsilon_{\mathcal{I}}\|\nabla g_j(\boldsymbol{w})\|^2+2\varepsilon_\mu\mu^2$

(iii) $\|\mathbb{E}\hat{\nabla} f_i(g)-\nabla f_i(g)\|^2\leq\varepsilon_{abs}\mu^2$

(iv) $\mathbb{E}\|\hat{\nabla} f_i(g)\|^2\leq\varepsilon\|\nabla f_i(g)\|^2+2\varepsilon_{abs}\mu^2$

Then, the estimator $\hat{\nabla}_{\mathcal{I}}g_j(\boldsymbol{w})$ and $\hat{\nabla} f_i(g)$ also satisfy, for arbitrary $q\geq 1$:

(a) $\|\mathbb{E}\hat{\nabla}_{\mathcal{I}}g_j(\boldsymbol{w})-\nabla_{\mathcal{I}}g_j(\boldsymbol{w})\|^2\leq\varepsilon_\mu\mu^2$

(b) $\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}}g_j(\boldsymbol{w})\right\|^2\leq\left(\frac{\varepsilon_{\mathcal{I}}}{q}+2\right)\|\nabla g_j(\boldsymbol{w})\|^2+2\left(\frac{1}{q}+1\right)\varepsilon_\mu\mu^2$

(c) $\|\mathbb{E}\hat{\nabla} f_i(g)-\nabla f_i(g)\|^2\leq\varepsilon_{abs}\mu^2$

(d) $\mathbb{E}\|\hat{\nabla} f_i(g)\|^2\leq\left(\frac{\varepsilon}{q}+2\right)\|\nabla f_i(g)\|^2+2\left(\frac{1}{q}+1\right)\varepsilon_{abs}\mu^2$

with

$$a_j=\frac{\mathbb{I}_{[y_j=1]}}{r},b_j=\frac{\mathbb{I}_{[y_j=-1]}}{1-r},\varepsilon_\mu=ds(a_j^2+b_j^2)L_{s_2}^2,\varepsilon_{abs}=(d+2)^2L_f^2,$$
$$\varepsilon_{\mathcal{I}}=\frac{2d}{(s_2+2)}\left(\frac{(s-1)(s_2-1)}{d-1}+3\right),\varepsilon=2(d+2).$$

Proof. Let us denote by $\hat{\nabla}g_j\left(\boldsymbol{w};(\boldsymbol{u}_i)_{i=1}^q\right)$ the gradient estimate from (7) along the i.i.d. sampled directions $(\boldsymbol{u}_i)_{i=1}^q$ (we simplify it to $\hat{\nabla}g_j(\boldsymbol{w};\boldsymbol{u})$ if there is only one direction $\boldsymbol{u}$). We can first see that, since the random directions $\boldsymbol{u}_i$ are independent and identically distributed (i.i.d.), we have:

$$\mathbb{E}\hat{\nabla}g_j\left(\boldsymbol{w};(\boldsymbol{u}_i)_{i=1}^q\right)=\mathbb{E}\frac{1}{q}\sum_{i=1}^q\hat{\nabla}g_j(\boldsymbol{w};\boldsymbol{u}_i)=\frac{1}{q}\sum_{i=1}^q\mathbb{E}\hat{\nabla}g_j(\boldsymbol{w};\boldsymbol{u}_1)=\mathbb{E}\hat{\nabla}g_j(\boldsymbol{w};\boldsymbol{u}_1)$$

This proves Lemma D.3 (a). Similarly, we have $\mathbb{E}\hat{\nabla} f_i\left(g;(\boldsymbol{z}_i)_{i=1}^q\right)=\mathbb{E}\hat{\nabla} f_i(g;\boldsymbol{v}_1)$, which proves D.3 (c).

Let us now turn to D.3 (b). We have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\nabla}_{\mathcal{I}} g_j\left(\boldsymbol{w};(\boldsymbol{u}_i)_{i=1}^q\right)\right\|^2\right] &= \mathbb{E}\left\|\frac{1}{q}\sum_{i=1}^q \hat{\nabla}_{\mathcal{I}} g_j\left(\boldsymbol{w};\boldsymbol{u}_i\right)\right\|^2 \\
&= \frac{1}{q^2}\mathbb{E}\left(\sum_{i=1}^q \hat{\nabla}_{\mathcal{I}} g_j\left(\boldsymbol{w};\boldsymbol{u}_i\right)\right)^\top\left(\sum_{l=1}^q \hat{\nabla}_{\mathcal{I}} g_j\left(\boldsymbol{w};\boldsymbol{u}_l\right)\right) \\
&= \frac{1}{q^2}\sum_{i=1}^q\sum_{l=1}^q \mathbb{E}\left[\hat{\nabla}_{\mathcal{I}} g_j\left(\boldsymbol{w};\boldsymbol{u}_i\right)^\top \hat{\nabla}_{\mathcal{I}} g_j\left(\boldsymbol{w};\boldsymbol{u}_l\right)\right] \\
&\overset{\xi_8}{=} \frac{1}{q^2}\left[q\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w};\boldsymbol{u}_1)\right\|^2 + \sum_{i=1}^q\sum_{l=1(l\neq i)}^q\left(\mathbb{E}\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w};\boldsymbol{u}_i)\right)^\top\left(\mathbb{E}\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w};\boldsymbol{u}_l)\right)\right] \\
&= \frac{1}{q^2}\left[q\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w};\boldsymbol{u}_1)\right\|^2 + q(q-1)\left\|\mathbb{E}\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w};\boldsymbol{u}_1)\right\|^2\right] \\
&\leq \frac{1}{q}\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w};\boldsymbol{u}_1)\right\|^2 + \left\|\mathbb{E}\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w};\boldsymbol{u}_1) - \nabla_{\mathcal{I}} g_j(\boldsymbol{w}) + \nabla_{\mathcal{I}} g_j(\boldsymbol{w})\right\|^2 \\
&\leq \frac{1}{q}\left(\varepsilon_{\mathcal{I}}\|\nabla g_j(\boldsymbol{w})\|^2 + 2\varepsilon_\mu\mu^2\right) + 2\|\mathbb{E}\hat{\nabla}_{\mathcal{I}} g_j(\boldsymbol{w};\boldsymbol{u}_1) - \nabla_{\mathcal{I}} g_j(\boldsymbol{w})\|^2 + 2\|\nabla_{\mathcal{I}} g_j(\boldsymbol{w})\|^2 \\
&\leq \left(\frac{\varepsilon_{\mathcal{I}}}{q}+2\right)\|\nabla g_j(\boldsymbol{w})\|^2 + 2\left(\frac{1}{q}+1\right)\varepsilon_\mu\mu^2
\end{aligned}$$

Where ξ_8 comes from the fact that the random directions are i.i.d. Similarly, we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\nabla} f_i(g;(\boldsymbol{z}_l)_{l=1}^q)\right\|^2\right] &\leq \frac{1}{q}\mathbb{E}\left\|\hat{\nabla} f_i(g;\boldsymbol{z}_1)\right\|^2 + \left\|\mathbb{E}\hat{\nabla} f_i(g;\boldsymbol{z}_1) - \nabla f_i(g) + \nabla f_i(g)\right\|^2 \\
&\leq \frac{1}{q}\left(\varepsilon\|\nabla f_i(g)\|^2 + 2\varepsilon_{abs}\mu^2\right) + 2\varepsilon_{abs}\mu^2 + 2\|\nabla f_i(g)\|^2 \\
&\leq \left(\frac{\varepsilon}{q}+2\right)\|\nabla f_i(g)\|^2 + 2\left(\frac{1}{q}+1\right)\varepsilon_{abs}\mu^2
\end{aligned}$$

□

E Proofs of section 4

E.1 Proof of Theorem 1

Let $\mathcal{I}$ denote the support set defined as $\mathcal{I}=\mathcal{I}^{(t)}\cup\mathcal{I}^{(t+1)}\cup\operatorname{supp}(\boldsymbol{w}^*)$, where $\mathcal{I}^{(t)}=\operatorname{supp}(\boldsymbol{w}_t)$. For Algorithm 1, we have:

$$\begin{aligned}
\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2 \leq \|\boldsymbol{r}_{t+1}-\boldsymbol{w}^*\|^2 &= \|\mathcal{H}_k(\boldsymbol{b}_{t+1})-\boldsymbol{w}^*\|^2 \\
&= \|\mathcal{H}_k(\mathcal{P}_{\mathcal{I}}(\boldsymbol{b}_{t+1}))-\boldsymbol{w}^*\|^2 \overset{(17)}{\leqslant} \gamma\|\mathcal{P}_{\mathcal{I}}(\boldsymbol{b}_{t+1})-\boldsymbol{w}^*\|^2. \qquad (20)
\end{aligned}$$

Taking the expectation with respect to i_t, j_t, as well as the random directions $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_q$ and $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_q$ at step t, we obtain:

$$\begin{aligned}
&\mathbb{E}\|\mathcal{P}_{\mathcal{I}}(\boldsymbol{b}_{t+1})-\boldsymbol{w}^*\|^2 \\
=&\mathbb{E}\left\|\boldsymbol{w}_t-\boldsymbol{w}^*-\eta\hat{\nabla}_{\mathcal{I}} g_{j_t}(\boldsymbol{w}_t)\hat{\nabla} f_{i_t}(\boldsymbol{v}_{t+1})\right\|^2
\end{aligned}$$

$$=\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2+\eta^2\underbrace{\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})\right\|^2}_{=:T_1}\underbrace{-2\eta\mathbb{E}\left\langle\boldsymbol{w}_t-\boldsymbol{w}^*,\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})\right\rangle}_{=:T_2}. \tag{21}$$

We then bound T_1.

$$\begin{aligned}
T_1=&\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})\right\|^2\\
\leq&\mathbb{E}\left\|\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\right\|^2\mathbb{E}\left\|\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})\right\|^2\\
\leq&\left[\varepsilon_{\mathcal{I}}\|\nabla g_{j_t}(\boldsymbol{w})\|^2+\varepsilon_g\mu^2\right]\cdot\left[\varepsilon\|\nabla f_{i_t}(\boldsymbol{v}_{t+1})\|^2+\varepsilon_f\mu^2\right]\\
\leq&\varepsilon_{\mathcal{I}}\varepsilon C_gC_f+\left(\varepsilon_{\mathcal{I}}C_g\varepsilon_f+\varepsilon C_f\varepsilon_g\right)\mu^2+\varepsilon_g\varepsilon_f\mu^4
\end{aligned} \tag{22}$$

Next, we bound T_2.

$$\begin{aligned}
T_2=&2\eta\mathbb{E}\left\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})\right\rangle\\
=&2\eta\mathbb{E}\left\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\left[\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)-\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)+\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\right]\cdot\left[\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})-\nabla f_{i_t}(\boldsymbol{v}_{t+1})+\nabla f_{i_t}(\boldsymbol{v}_{t+1})\right]\right\rangle\\
=&2\eta\mathbb{E}\left\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\left[\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)-\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\right]\cdot\left[\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})-\nabla f_{i_t}(\boldsymbol{v}_{t+1})\right]\right\rangle\\
&+2\eta\mathbb{E}\left\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\left[\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)-\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\right]\cdot\nabla f_{i_t}(\boldsymbol{v}_{t+1})\right\rangle\\
&+2\eta\mathbb{E}\left\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\cdot\left[\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})-\nabla f_{i_t}(\boldsymbol{v}_{t+1})\right]\right\rangle\\
&+2\eta\mathbb{E}\left\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\cdot\nabla f_{i_t}(\boldsymbol{v}_{t+1})\right\rangle\\
\overset{\xi_9}{\leq}&\eta\left[\lambda_1\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2+\frac{1}{\lambda_1}\left\|\mathbb{E}\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)-\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\right\|^2\cdot\left\|\mathbb{E}\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})-\nabla f_{i_t}(\boldsymbol{v}_{t+1})\right\|^2\right]\\
&+\eta\left[\lambda_2\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2+\frac{1}{\lambda_2}\left\|\mathbb{E}\hat{\nabla}_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)-\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\right\|^2\cdot\|\nabla f_{i_t}(\boldsymbol{v}_{t+1})\|^2\right]\\
&+\eta\left[\lambda_3\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2+\frac{1}{\lambda_3}\|\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\|^2\left\|\mathbb{E}\hat{\nabla}f_{i_t}(\boldsymbol{v}_{t+1})-\nabla f_{i_t}(\boldsymbol{v}_{t+1})\right\|^2\right]\\
&+2\eta\mathbb{E}\left\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\cdot(\nabla f_{i_t}(\boldsymbol{v}_{t+1})-\nabla f_{i_t}(g(\boldsymbol{w}_t)))\right\rangle+2\eta\mathbb{E}\left\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\nabla_{\mathcal{I}}g_{j_t}(\boldsymbol{w}_t)\cdot\nabla f_{i_t}(g(\boldsymbol{w}_t))\right\rangle\\
\leq&(\lambda_1+\lambda_2+\lambda_3)\eta\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2+\frac{\eta}{\lambda_1}\varepsilon_\mu\varepsilon_{abs}\mu^4+\eta\left(\frac{1}{\lambda_2}C_f\varepsilon_\mu+\frac{1}{\lambda_3}C_g\varepsilon_{abs}\right)\mu^2\\
&+\lambda_4\eta\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2+\frac{\eta}{\lambda_4}C_gL_f^2\|\boldsymbol{v}_{t+1}-g(\boldsymbol{w}_t)\|^2+2\eta\mathbb{E}\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\nabla_{\mathcal{I}}f(\boldsymbol{w}_t)\rangle
\end{aligned} \tag{23}$$

ξ_9 follows from the fact that for any $\lambda>0$, we have

$$2|\langle x,y\rangle|\leq\lambda x^2+\frac{1}{\lambda}y^2$$

Since the function $f(\boldsymbol{w})$ satisfies the Restricted Strong Convexity (RSC) property, we have that

$$\begin{aligned}
&\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\nabla_{\mathcal{I}}f(\boldsymbol{w}_t)\rangle\\
=&\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\nabla_{\mathcal{I}}f(\boldsymbol{w}_t)-\nabla_{\mathcal{I}}f(\boldsymbol{w}^*)\rangle+\langle\boldsymbol{w}^*-\boldsymbol{w}_t,\nabla_{\mathcal{I}}f(\boldsymbol{w}^*)\rangle\\
\leq&-\alpha\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2+2\omega\|\nabla_{\mathcal{I}}f(\boldsymbol{w}^*)\|
\end{aligned} \tag{24}$$

The last inequality follows from the boundedness of the parameters. Substituting (24) into (23), we have

$$\begin{aligned}
T_2\leq&\eta(\lambda_1+\lambda_2+\lambda_3+\lambda_4-2\alpha)\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2+4\eta\omega\|\nabla_{\mathcal{I}}f(\boldsymbol{w}^*)\|+\frac{\eta}{\lambda_4}C_gL_f^2\|\boldsymbol{v}_{t+1}-g(\boldsymbol{w}_t)\|^2\\
&+\frac{\eta}{\lambda_1}\varepsilon_\mu\varepsilon_{abs}\mu^4+\eta\left(\frac{1}{\lambda_2}C_f\varepsilon_\mu+\frac{1}{\lambda_3}C_g\varepsilon_{abs}\right)\mu^2
\end{aligned}$$

Let $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \frac{\alpha}{4}$, which implies that:

$$\begin{aligned} T_2 \leq & -\alpha\eta\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 + 4\eta\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \frac{4\eta C_g L_f^2}{\alpha}\|\boldsymbol{v}_{t+1} - g(\boldsymbol{w}_t)\|^2 \\ & + \frac{4\eta\varepsilon_\mu\varepsilon_{abs}}{\alpha}\mu^4 + \frac{4\eta}{\alpha}\left(C_f\varepsilon_\mu + C_g\varepsilon_{abs}\right)\mu^2 \end{aligned} \tag{25}$$

Substituting (22) and (25) into (21), we have

$$\begin{aligned} \mathbb{E}\left\|\mathcal{P}_{\mathcal{I}}(\boldsymbol{b}_{t+1}) - \boldsymbol{w}^*\right\|^2 =& \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 + \eta^2 T_1 + T_2 \\ =& (1-\alpha\eta)\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 + 4\eta\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \frac{4\eta C_g L_f^2}{\alpha}\|\boldsymbol{v}_{t+1} - g(\boldsymbol{w}_t)\|^2 + \eta^2\varepsilon_{\mathcal{I}}\varepsilon C_g C_f \\ & + \underbrace{\left[\eta^2\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right) + \frac{4\eta}{\alpha}\left(C_f\varepsilon_\mu + C_g\varepsilon_{abs}\right)\right]}_{=:L_1}\mu^2 + \underbrace{\left[\eta^2\varepsilon_g\varepsilon_f + \frac{4\eta\varepsilon_\mu\varepsilon_{abs}}{\alpha}\right]}_{=:L_2}\mu^4 \end{aligned} \tag{26}$$

Putting this bound back into (20), we obtain

$$\begin{aligned} & \mathbb{E}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 \\ \leq & \gamma\mathbb{E}\|\mathcal{P}_{\mathcal{I}}(\boldsymbol{b}_{t+1}) - \boldsymbol{w}^*\|^2 \\ \leq & \gamma\underbrace{(1-\alpha\eta)}_{\rho}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 + 4\eta\gamma\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \frac{4\eta\gamma C_g L_f^2}{\alpha}\|\boldsymbol{v}_{t+1} - g(\boldsymbol{w}_t)\|^2 + \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f + \gamma L_1\mu^2 + \gamma L_2\mu^4 \end{aligned} \tag{27}$$

We multiply Eq. (19) by $\left(\Lambda + \frac{4\eta\gamma C_g L_f^2}{\alpha}\right)$ and sum it with Eq. (27), where $\Lambda = \max\left\{\frac{1}{\gamma\rho - 1 + \beta}\cdot\frac{4\eta\gamma C_g L_f^2}{\alpha}, 0\right\}$. We obtain

$$\begin{aligned} & \mathbb{E}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 + \Lambda\mathbb{E}\|\boldsymbol{v}_{t+1} - g(\boldsymbol{w}_t)\|^2 \\ \leq & \gamma\rho\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 + 4\eta\gamma\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f + \gamma L_1\mu^2 + \gamma L_2\mu^4 \\ & + (1-\beta)\left(\Lambda + \frac{4\eta\gamma C_g L_f^2}{\alpha}\right)\|\boldsymbol{v}_t - g(\boldsymbol{w}_{t-1})\|^2 + \left(\Lambda + \frac{4\eta\gamma C_g L_f^2}{\alpha}\right)\left(\beta^{-1} C_g\left\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\right\|^2 + 2V_g\beta^2\right) \end{aligned}$$

Let $U_{t+1} = \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 + \Lambda\|\boldsymbol{v}_{t+1} - g(\boldsymbol{w}_t)\|^2$, we have

$$\begin{aligned} \mathbb{E}U_{t+1} \leq & \gamma\rho U_t + 4\eta\gamma\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f + \gamma L_1\mu^2 + \gamma L_2\mu^4 \\ & + \mathcal{O}\left(\frac{\eta\gamma C_g L_f^2}{\beta\alpha}\right)\left(\beta^{-1} C_g\left\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\right\|^2 + 2V_g\beta^2\right) \end{aligned} \tag{28}$$

For $\left\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\right\|^2$, we have

$$\begin{aligned} \left\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\right\|^2 \leq & \|\boldsymbol{r}_t - \boldsymbol{w}_{t-1}\|^2 = \left\|\mathcal{H}_k\left(\mathcal{P}_{\mathcal{I}}(\boldsymbol{b}_t)\right) - \boldsymbol{w}_{t-1}\right\|^2 \\ \leq & \gamma\|\mathcal{P}_{\mathcal{I}}(\boldsymbol{b}_t) - \boldsymbol{w}_{t-1}\|^2 = \eta^2\gamma\|\hat{\nabla}_{\mathcal{I}} g_{j_t}(\boldsymbol{w}_{t-1})\hat{\nabla} f_{i_t}(\boldsymbol{v}_t)\|^2 \\ \leq & \eta^2\gamma\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} g_{j_t}(\boldsymbol{w}_{t-1})\|^2 \cdot \mathbb{E}\|\hat{\nabla} f_{i_t}(\boldsymbol{v}_t)\|^2 \\ \leq & \eta^2\gamma\left[\varepsilon_{\mathcal{I}}\|\nabla g_{j_t}(\boldsymbol{w}_{t-1})\|^2 + \varepsilon_g\mu^2\right]\cdot\left[\varepsilon\|\nabla f_{i_t}(\boldsymbol{v}_t)\|^2 + \varepsilon_f\mu^2\right] \\ \leq & \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f + \eta^2\gamma\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right)\mu^2 + \eta^2\gamma\varepsilon_g\varepsilon_f\mu^4 \end{aligned} \tag{29}$$

Substituting (29) into (28), we obtain

$$\begin{aligned}
\mathbb{E}U_{t+1} \leq & \gamma\rho U_t + 4\eta\gamma\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f + \mathcal{O}\left(\frac{\eta\gamma V_g \beta C_g L_f^2}{\alpha}\right) + \gamma L_1\mu^2 + \gamma L_2\mu^4 \\
& + \mathcal{O}\left(\frac{\eta\gamma C_g^2 L_f^2}{\beta^2\alpha}\right)\left(\eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f + \eta^2\gamma\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right)\mu^2 + \eta^2\gamma\varepsilon_g\varepsilon_f\mu^4\right) \\
\leq & \gamma\rho U_t + 4\eta\gamma\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f + \mathcal{O}\left(\frac{\eta\gamma V_g \beta C_g L_f^2}{\alpha}\right) + \mathcal{O}\left(\frac{\eta^3\gamma^2\varepsilon_{\mathcal{I}}\varepsilon C_g^3 C_f L_f^2}{\beta^2\alpha}\right) \\
& + \gamma L_1\mu^2 + \gamma L_2\mu^4 + \mathcal{O}\left(\frac{\eta\gamma C_g^2 L_f^2}{\beta^2\alpha}\right)\left(\eta^2\gamma\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right)\mu^2 + \eta^2\gamma\varepsilon_g\varepsilon_f\mu^4\right)
\end{aligned}$$

Let

$$\xi_\eta = 4\eta\gamma\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f + \mathcal{O}\left(\frac{\eta\gamma V_g \beta C_g L_f^2}{\alpha}\right) + \mathcal{O}\left(\frac{\eta^3\gamma^2\varepsilon_{\mathcal{I}}\varepsilon C_g^3 C_f L_f^2}{\beta^2\alpha}\right)$$

and

$$\begin{aligned}
L_\mu = & \gamma L_1\mu^2 + \gamma L_2\mu^4 + \mathcal{O}\left(\frac{\eta\gamma C_g^2 L_f^2}{\beta^2\alpha}\right)\left(\eta^2\gamma\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right)\mu^2 + \eta^2\gamma\varepsilon_g\varepsilon_f\mu^4\right) \\
= & \left[\eta^2\nu\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right) + \frac{4\eta\gamma}{\alpha}\left(C_f\varepsilon_\mu + C_g\varepsilon_{abs}\right)\right]\mu^2 + \mathcal{O}\left(\frac{\eta^3\gamma^2 C_g^2 L_f^2}{\beta^2\alpha}\right)\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right)\mu^2 \\
& + \left[\eta^2\nu\varepsilon_g\varepsilon_f + \frac{4\eta\gamma\varepsilon_\mu\varepsilon_{abs}}{\alpha}\right]\mu^4 + \mathcal{O}\left(\frac{\eta^3\gamma^2 C_g^2 L_f^2}{\beta^2\alpha}\right)\varepsilon_g\varepsilon_f\mu^4 \\
= & \left[\eta^2\gamma\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right) + \frac{4\eta\gamma}{\alpha}\left(C_f\varepsilon_\mu + C_g\varepsilon_{abs}\right) + \mathcal{O}\left(\frac{\eta^3\gamma^2 C_g^2 L_f^2}{\beta^2\alpha}\right)\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right)\right]\mu^2 \\
& + \left[\eta^2\gamma\varepsilon_g\varepsilon_f + \frac{4\eta\gamma\varepsilon_\mu\varepsilon_{abs}}{\alpha} + \mathcal{O}\left(\frac{\eta^3\gamma^2 C_g^2 L_f^2}{\beta^2\alpha}\right)\varepsilon_g\varepsilon_f\right]\mu^4
\end{aligned}$$

Therefore, we have

$$\mathbb{E}U_{t+1} \leq \gamma\rho U_t + \xi_\eta + L_\mu \tag{30}$$

We need to have $\gamma\rho < 1$ in order to ensure a contraction at each step. Suppose $k \geq \rho k^*/(1-\rho)^2$; we will show that this choice of k allows us to verify the condition $\gamma\rho < 1$. This implies $\frac{k^*}{k} \leq \frac{(1-\rho)^2}{\rho}$. We then have, from the definition of γ in Eq. (17):

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

$$
\begin{aligned}
&= 1 + \left(\frac{(1-\rho)(1-\rho+1+\rho)}{\rho} \right) \frac{1}{2} \\
&= 1 + \frac{(1-\rho)}{\rho} = \frac{1}{\rho}
\end{aligned} \tag{31}
$$

Therefore, we indeed have $\gamma\rho \leq 1$ when choosing $k \geq \rho k^*/(1-\rho)^2$. Unrolling inequality (30) through time, we then have

$$
\begin{aligned}
\mathbb{E}U_{t+1} \leq & \gamma\rho\mathbb{E}U_t + \xi_\eta + L_\mu \\
\leq & \gamma\rho[\gamma\rho\mathbb{E}U_{t-1} + \xi_\eta + L_\mu] + \xi_\eta + L_\mu \\
= & (\gamma\rho)^2\mathbb{E}U_{t-1} + \gamma\rho\xi_\eta + \xi_\eta + \gamma\rho L_\mu + L_\mu \\
& \cdots\cdots\cdots \\
\leq & (\gamma\rho)^t\mathbb{E}U_1 + \left[\sum_{i=0}^{t-1}(\gamma\rho)^i\right]\xi_\eta + \left[\sum_{i=0}^{t-1}(\gamma\rho)^i\right]L_\mu
\end{aligned} \tag{32}
$$

where

$$
\begin{aligned}
\mathbb{E}U_1 &= \mathbb{E}\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2 + \Lambda\mathbb{E}\|\boldsymbol{v}_1 - g(\boldsymbol{w}_0)\|^2 \\
&= \mathbb{E}\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2 + \Lambda\mathbb{E}\|g_{j_t}(\boldsymbol{w}_0) - g(\boldsymbol{w}_0)\|^2 \\
&\overset{(10)}{\leq} \mathbb{E}\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2 + \Lambda V_g
\end{aligned} \tag{33}
$$

For $\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2$, by a similar proof to that of Eq. (27), we have:

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2 \leq & \gamma\mathbb{E}\|\boldsymbol{w}_0 - \boldsymbol{w}^* - \eta\hat{\nabla}_{\mathcal{I}} g_{j_t}(\boldsymbol{w}_0)\hat{\nabla} f_{i_t}(\boldsymbol{v_1})\|^2 \\
\leq & \gamma\rho\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + 4\eta\gamma\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \frac{4\eta\gamma C_g L_f^2 V_g}{\alpha} + \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f \\
& + \left[\eta^2\gamma\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right) + \frac{4\eta\gamma}{\alpha}\left(C_f\varepsilon_\mu + C_g\varepsilon_{abs}\right)\right]\mu^2 + \left[\eta^2\gamma\varepsilon_g\varepsilon_f + \frac{4\eta\gamma\varepsilon_\mu\varepsilon_{abs}}{\alpha}\right]\mu^4
\end{aligned} \tag{34}
$$

Substituting (34) and (33) into (32), we obtain

$$
\begin{aligned}
\mathbb{E}U_{t+1} &\leq (\gamma\rho)^t\mathbb{E}U_1 + \left[\sum_{i=0}^{t-1}(\gamma\rho)^i\right]\xi_\eta + \left[\sum_{i=0}^{t-1}(\gamma\rho)^i\right]L_\mu \\
&\leq (\gamma\rho)^t\left[\gamma\rho\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + 4\eta\gamma\omega\|\nabla_{\mathcal{I}} f(\boldsymbol{w}^*)\| + \frac{4\eta\gamma C_g L_f^2 V_g}{\alpha} + \Lambda V_g + \eta^2\gamma\varepsilon_{\mathcal{I}}\varepsilon C_g C_f\right] + \left[\sum_{i=0}^{t-1}(\gamma\rho)^i\right]\xi_\eta \\
&\quad + (\gamma\rho)^t\left[\left(\eta^2\gamma\left(\varepsilon_{\mathcal{I}} C_g\varepsilon_f + \varepsilon C_f\varepsilon_g\right) + \frac{4\eta\gamma}{\alpha}\left(C_f\varepsilon_\mu + C_g\varepsilon_{abs}\right)\right)\mu^2 + \left(\eta^2\gamma\varepsilon_g\varepsilon_f + \frac{4\eta\gamma\varepsilon_\mu\varepsilon_{abs}}{\alpha}\right)\mu^4\right] \\
&\quad + \left[\sum_{i=0}^{t-1}(\gamma\rho)^i\right]L_\mu \\
&\leq (\gamma\rho)^{t+1}\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + + \left[\sum_{i=0}^{t}(\gamma\rho)^i\right]\xi_\eta + \left[\sum_{i=0}^{t}(\gamma\rho)^i\right]L_\mu \\
&\leq (\gamma\rho)^{t+1}\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + \frac{1}{1-\gamma\rho}\xi_\eta + \frac{1}{1-\gamma\rho}L_\mu
\end{aligned} \tag{35}
$$

Where the last inequality follows from the fact that $\gamma\rho < 1$. Therefore, we have

$$\mathbb{E}\|\boldsymbol{w}_T - \boldsymbol{w}^*\|^2 \leq \mathbb{E}U_T \leq (\gamma\rho)^T\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + \frac{1}{1-\gamma\rho}\xi_\eta + \frac{1}{1-\gamma\rho}L_\mu \tag{36}$$

E.2 Proof of Remark 2

To guarantee convergence, it is required that $\gamma\rho < 1$. Therefore, based on the derivation in equation (31), the condition that must be satisfied is:

$$\frac{k^*}{k} \leq \frac{(1-\rho)^2}{\rho} = \frac{(\alpha\eta)^2}{1-\alpha\eta},$$

which is equivalent to:

$$k \geq \frac{1}{\alpha\eta}\left(\frac{1}{\alpha\eta} - 1\right)k^*.$$

Here, $\eta < \frac{1}{\alpha}$. Additionally, k must satisfy the constraint $k \leq \frac{d-k^*}{2}$, i.e.,

$$\frac{1}{\alpha\eta}\left(\frac{1}{\alpha\eta} - 1\right)k^* \leq \frac{d-k^*}{2}. \tag{37}$$

Let $x = \frac{1}{\alpha\eta}$, then equation (37) can be written as:

$$x(x-1) \leq \frac{d-k^*}{2k^*} \Longleftrightarrow x^2 - x - \frac{d-k^*}{2k^*} \leq 0 \tag{38}$$

The discriminant Δ is given by:

$$\Delta = 1 - 4\left(-\frac{d-k^*}{2k^*}\right) = 1 + \frac{2(d-k^*)}{k^*} > 0$$

Thus, the maximum value of x satisfying Equation (38) is:

$$x_{\max} = \frac{1 + \sqrt{1 + \frac{2(d-k^*)}{k^*}}}{2},$$

which implies

$$\frac{1}{\alpha\eta} \leq \frac{1 + \sqrt{1 + \frac{2(d-k^*)}{k^*}}}{2}$$

Therefore, we have:

$$\frac{2}{\alpha\left(1 + \sqrt{1 + \frac{2(d-k^*)}{k^*}}\right)} \leq \eta \leq \frac{1}{\alpha} \tag{39}$$

E.3 Proof of Corollary 1

Let us now impose a lower bound on k that is slightly larger (specifically, twice as large) than the lower bound provided in Theorem 1. As will become clear below, this ensures that $\gamma\rho$ remains sufficiently bounded away from 1, which leads to a favorable constant factor in the big-$\mathcal{O}$ notation for the query complexity (see the end of the proof for details). Therefore, we take:

$$k \geq 2k^*\frac{\rho}{(1-\rho)^2} \tag{40}$$

Substituting the value of ρ into (40) yields:

$$k \geq 2k^*\frac{1-\alpha\eta}{(\alpha\eta)^2}$$

By a similar proof to that of Eq. (39), we have:

$$\frac{2}{\alpha\left(1+\sqrt{1+\frac{d-k^*}{k^*}}\right)} \leq \eta \leq \frac{1}{\alpha}$$

To ensure that $(\gamma\rho)^T\|\boldsymbol{w}_0-\boldsymbol{w}^*\|^2 \leq \varepsilon$, we need:

$$T \geq \frac{1}{\log\frac{1}{\gamma\rho}}\left[\log\left(\frac{1}{\varepsilon}\right)+\log\left(\|\boldsymbol{w}_0-\boldsymbol{w}^*\|^2\right)\right] \tag{41}$$

Where $\gamma\rho$ belongs to the interval $(0,1)$. Substituting (40) into (17), we obtain

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

Therefore, we have:

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

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

$$\frac{1}{\gamma\rho} \geq \frac{1}{1-\frac{\alpha\eta}{4}}$$

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

$$\log\left(\frac{1}{\gamma\rho}\right) \geq \log\left(\frac{1}{1-\frac{\alpha\eta}{4}}\right)\overset{(b)}{\geq}\frac{\alpha\eta}{4}$$

Where (b) follows because $\eta \leq 1/\alpha$. Therefore, we have:

$$\frac{1}{\log\left(\frac{1}{\gamma\rho}\right)} \leq \frac{4}{\alpha\eta} \tag{42}$$

Substituting (42) into (41), we obtain that after $T \geq \frac{4}{\alpha\eta}\left[\log\left(\frac{1}{\varepsilon}\right)+\log\left(\|\boldsymbol{w}_0-\boldsymbol{w}^*\|^2\right)\right] = \mathcal{O}\left(\frac{1}{\alpha\eta}\log\left(\frac{1}{\varepsilon}\right)\right)$ iterations, we can achieve $(\gamma\rho)^T\|\boldsymbol{w}_0-\boldsymbol{w}^*\|^2 \leq \varepsilon$.

By Algorithm 1, the function query complexity per iteration is $2q+2$. Therefore, to ensure $(\gamma\rho)^T\|\boldsymbol{w}_0-\boldsymbol{w}^*\|^2 \leq \varepsilon$, the total function query complexity is: $T(2q+2) = \mathcal{O}\left(\frac{q}{\alpha\eta}\log\left(\frac{1}{\varepsilon}\right)\right)$.

F Experimental Parameter Settings

F.1 Experimental parameter settings for Black-Box Sparse AUC Maximization.

To facilitate the reproducibility of our results, we provide the parameter-tuning details for each method below:

- ZO-SCHT involves seven parameters.The sparsity level k is set as $k = \lfloor d/s \rfloor$, where the scaling factor s is selected from the set $\{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 10, 20\}$, and d represents the dimension of $\boldsymbol{w}$. The learning rate η is selected from the set $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.7, 1, 2, 5\}$, and β is selected from $\{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.7, 0.8, 0.99\}$. The number of iterations T is chosen from $\{10000, 20000, 30000, 40000, 50000\}$, while the projection parameter ω is selected from $\{5, 10, 15, 20\}$. The number of random directions q is tuned from $\{10, 20, 30, 40, 50\}$, and the size of the random directions support is set to $s_2 = k$.
- GFCOM involves four parameters. The learning rate η is tuned in the same way as in ZO-SCHT. The number of iterations T is tuned from $\{1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000\}$, while the mini-batch sizes of the gradient estimators b_f and b_g are tuned from $\{10, 20, 30, 40, 50\}$.
- Black-box SCGD involves four parameters. The initial learning rate η is tuned from $\{0.01, 0.02, 0.05, 0.08, 0.1, 0.2, 0.5, 0.8, 1, 2, 5, 8, 10, 20, 50, 80, 100\}$, while the learning rate β is tuned in the same way as in ZO-SCHT. The number of iterations T is tuned from $\{10000, 20000, 30000, 40000, 50000\}$, and the number of random directions b is tuned from $\{10, 20, 30, 40, 50\}$.
- Black-box SCSC involves four parameters, which are tuned in the same way as in Black-box SCGD.
- ZO-Min-Max involves two learning rate parameters. The learning rate α controls the update of the min variable, while β governs the update of the max variable, and both parameters are selected from the set $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 0.01, 0.02, 0.05\}$.

F.2 Experimental parameter settings of Black-Box Adversarial Attacks.

Likewise, in order to facilitate the reproducibility of our results in the black-box adversarial attacks, we provide the parameter-tuning details for each method below:

- For ZO-SCHT, the learning rate η is selected from the set $\{0.01, 0.02, 0.05, 0.07, 0.1, 0.2, 0.5, 0.7, 1, 2, 5, 7, 10\}$, and β is selected from $\{0.01, 0.05, 0.1, 0.2, 0.5, 0.7, 0.99\}$. The number of iterations T is chosen from $\{20, 50, 100, 200, 300\}$, while the projection parameter ω is selected from $\{3000, 5000, 8000, 10000, 13000, 15000, 17000, 20000, 23000, 25000, 27000, 30000\}$. The number of random directions q is tuned from $\{10, 20, 30, 40, 50\}$, and the size of the random directions support is set to $s_2 = d(\text{total\_pixels})$. The sparsity level k is set as $k = \lfloor \text{total\_pixels}/s \rfloor$, where the scaling factor s is selected from the set $\{100, 130, 150, 170, 190, 210, 230, 250\}$, and total_pixels denotes the number of pixels per image.
- For GFCOM, the learning rate η is tuned in the same way as in ZO-SCHT. The number of iterations T is tuned from $\{5, 10, 15, 20, 25, 30\}$, and the mini-batch sizes of the gradient estimators b_f and b_g are tuned from $\{10, 20, 30, 40, 50\}$. The projection parameter R is tuned in the same way as ω in ZO-SCHT.
- For Black-box SCGD, the initial learning rate η is tuned from $\{1, 2, 5, 7, 10, 13, 15\}$, while the learning rate β is tuned in the same way as in ZO-SCHT. The number of iterations T is selected from $\{20, 50, 100, 200, 300\}$, and the number of random directions b is tuned from $\{10, 20, 30, 40, 50\}$. The projection parameter R is tuned in the same way as ω in ZO-SCHT.
- For Black-box SCSC, all of its parameters are tuned in the same way as in Black-box SCGD.
- For ZO-Min-Max, both learning rate parameters α and β are selected from the set $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.7, 1, 2, 5\}$.

F.3 Sensitivity analysis of hyperparameters.

To better understand how each hyperparameter influences the model's performance, we conducted a comprehensive sensitivity analysis, and the corresponding results are presented in Table 4. Note that s is the parameter such that k = int(d / s).

The experimental results demonstrate that ZO-SCHT maintains stable performance across reasonable ranges of all hyperparameters. As discussed in Appendix F.1, the sparsity parameter k can be set as a certain proportion of the dimensionality d; in practice, choosing k to be 25%–33% of d generally achieves good performance while reducing the number of model parameters. In most cases, the parameter μ can be selected from $\{1 \times 10^{-3}, 5 \times$

Table 4: Sensitivity analysis of hyperparameters.

μ	**gassensor**	**a9a**	q	**gassensor**	**a9a**	s_2	**gassensor**	**a9a**
0.001	0.932	0.855	10	0.910	0.852	30	0.926	0.853
0.005	0.930	0.857	15	0.925	0.854	50	0.935	0.856
0.01	0.927	0.852	20	0.930	0.855	k	0.932	0.855
0.05	0.927	0.846	30	0.933	0.858	d	0.933	0.859

Table 5: Sensitivity analysis of hyperparameters (continued).

s	**gassensor**	**a9a**	η	**gassensor**	**a9a**	β	**gassensor**	**a9a**
1	0.946	0.866	1×10^{-5}	0.937	0.850	1×10^{-5}	0.936	0.862
2	0.942	0.863	5×10^{-5}	0.937	0.861	5×10^{-5}	0.937	0.861
3	0.939	0.859	1×10^{-4}	0.930	0.855	1×10^{-4}	0.930	0.855
4	0.930	0.855	5×10^{-4}	0.935	0.853	5×10^{-4}	0.937	0.851
5	0.924	0.850	1×10^{-3}	0.923	0.846	1×10^{-3}	0.931	0.854
10	0.911	0.852	5×10^{-3}	0.807	0.823	5×10^{-3}	0.925	0.848

$10^{-3}\}$, and although a larger value of q yields more stable gradient estimates, a setting of 20 typically provides a good balance between stability and computational efficiency. To further minimize tuning efforts, s_2 can simply be set to k or d, which has proven effective across all experimental scenarios. For η and β, we recommend using smaller values for models with lower complexity and adopting larger values in more complex settings, such as black-box adversarial attacks, where they tend to yield better empirical performance.

G Black-Box Adversarial Perturbation Visualization.

Figure 3 illustrates the perturbation patterns generated by different algorithms on black-box models. Compared with the perturbations produced by other baseline methods, the perturbations generated by the ZO-SCHT algorithm exhibit significant sparsity. Figure 4 presents the adversarial examples generated by different algorithms on black-box models.

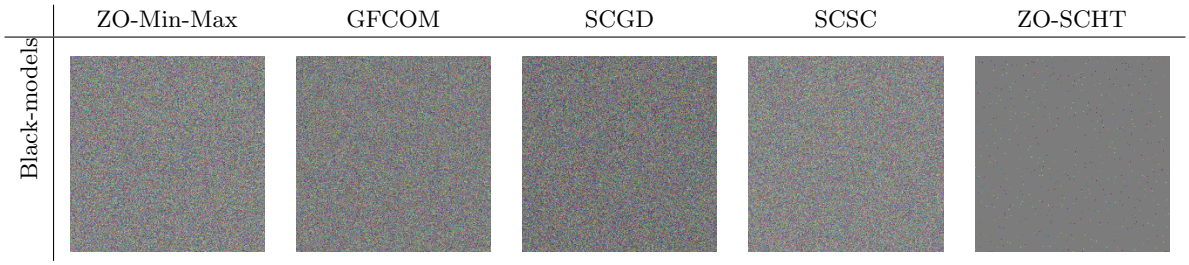

Figure 3: Universal Perturbations of Different Algorithms Across Black-box Network

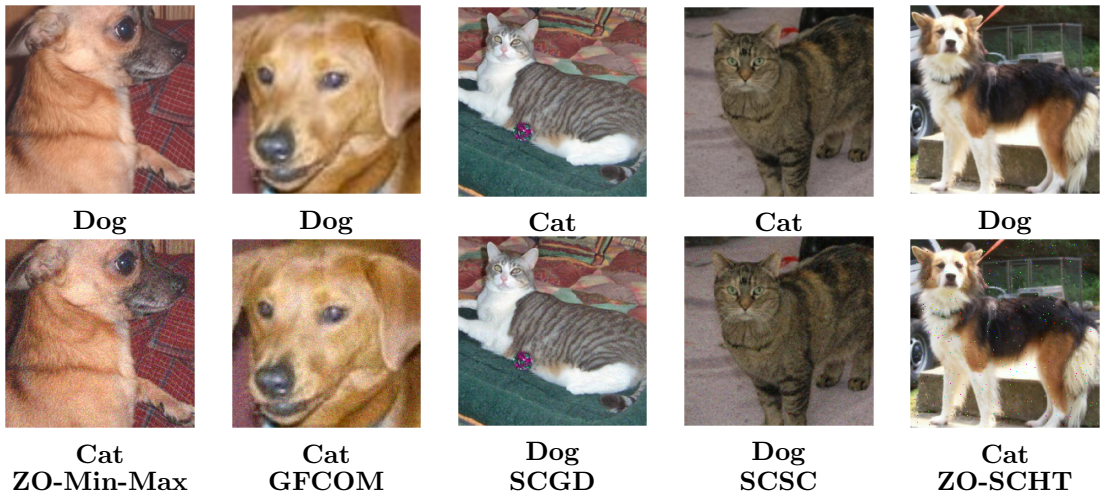

Figure 4: Attack results on Black-box Models (top: clean images; bottom: perturbed images).