MINI-EXTRAGRADIENT METHODS*

XIAOZHI LIU† AND YONG XIA†‡

Abstract. The Extragradient (EG) method stands as a cornerstone algorithm for solving monotone nonlinear equations but faces two important unresolved challenges: (i) how to select stepsizes without relying on the global Lipschitz constant or expensive line-search procedures, and (ii) how to reduce the two full evaluations of the mapping required per iteration to effectively one, without compromising convergence guarantees or computational efficiency. To address the first challenge, we propose the Greedy Mini-Extragradient (Mini-EG) method, which updates only the coordinate associated with the dominant component of the mapping at each extragradient step. This design capitalizes on componentwise Lipschitz constants that are far easier to estimate than the classical global Lipschitz constant. To further lower computational cost, we introduce a Random Mini-EG variant that replaces full mapping evaluations by sampling only a single coordinate per extragradient step. Although this resolves the second challenge from a theoretical standpoint, its practical efficiency remains limited. To bridge this gap, we develop the Watchdog-Max strategy, motivated by the slow decay of dominant component magnitudes. Instead of evaluating the full mapping, Watchdog-Max identifies and tracks only two coordinates at each extragradient step, dramatically reducing per-iteration cost while retaining strong practical performance. We establish convergence guarantees and rate analyses for all proposed methods. In particular, Greedy Mini-EG achieves enhanced convergence rates that surpass the classical guarantees of the vanilla EG method in several standard application settings. Numerical experiments on regularized decentralized logistic regression and compressed sensing show speedups exceeding 13× compared with the classical EG method on both synthetic and real datasets.

Key words. extragradient method, coordinate descent, monotone nonlinear equations, ergodic convergence

MSC codes. 90C25, 90C30, 90C47, 90C60, 65K05, 65K15

1. Introduction.

Many fundamental problems in optimization and scientific computing, including min-max problem and variational inequalities, can be cast as monotone nonlinear equations. These formulations underpin a broad range of applications, such as decentralized logistic regression [9], generative adversarial networks [8], image processing [6], and robust learning [18]. However, solving them efficiently remains a significant challenge, especially in regimes where the feature dimension substantially exceeds the number of samples, a scenario that has become increasingly prominent in modern few-sample scenarios [17].

One of the most widely studied methods for solving monotone nonlinear equations is the Extragradient (EG) method [11], which guarantees convergence under L-Lipschitz continuity. It has inspired a broad class of EG-type methods, each enhancing performance through different strategies. For example, EG+ [5] adopts a two-time-scale framework that allows more flexible stepsize selection. The Extra-Anchored-Gradient (EAG) method [21] achieves faster convergence by combining an initial and the most recent iterate at each step, achieving the optimal rate under monotonicity and Lipschitz continuity assumptions. Building on both EG+ and EAG, the Fast Extragradient (FEG) method [12] further integrates two-time-scale and anchoring techniques to further accelerate convergence. Another important variant is

* **Funding:** This work was funded by by National Key Research and Development Program of China under grant 2021YFA1003303 and National Natural Science Foundation of China under grant 12171021.

†School of Mathematical Sciences, Beihang University, Beijing 100191, People's Republic of China, {xzliu, yxia}@buaa.edu.cn.

‡Corresponding author.

the Past Extragradient (PEG) method [15], also known as the Optimistic Gradient method, which reduces computational cost by storing and reusing historical iterates.

Despite their broad applicability, EG-type methods suffer from two fundamental limitations: (i) Their stepsize rules depend on the global Lipschitz constant L, which is often unavailable or expensive to estimate in practice. As a result, practitioners typically resort to manual tuning or computationally intensive line-search procedures [20], undermining efficiency and practicality. (ii) Each iteration requires either two full evaluations of the mapping or storage of the entire previous iterate. Both options incur significant computational and memory costs, especially in large-scale settings.

These challenges naturally lead to two fundamental and unresolved questions:

1. Can we design a line-search-free EG variant that requires no knowledge of the global Lipschitz constant L, yet still enjoys rigorous convergence guarantees?
2. Can such a variant retain its convergence properties while reducing the per-iteration cost from two full evaluations of the mapping to effectively one, thereby improving overall computational efficiency?

Motivated by these questions, we develop three Mini-Extragradient (Mini-EG) methods: Greedy Mini-EG, Random Mini-EG, and Watchdog-Max. These methods eliminate the dependence on the global Lipschitz constant L for stepsize selection and updates only a single coordinate at each extragradient step, thereby avoiding full evaluations of the mapping.

Our main contributions are summarized as follows:

- These Mini-EG variants determine stepsizes based solely on componentwise Lipschitz constants, which are typically easier to compute and smaller than the global Lipschitz constant L.
- A key contribution of this work is Watchdog-Max, a novel Mini-EG method inspired by the slow decay of dominant component magnitudes. By incorporating randomized sampling, it requires only two coordinate evaluations per extragradient step, significantly reducing computational cost while maintaining convergence performance comparable to the greedy selection strategy.
- We provide rigorous convergence analysis and rate guarantees for all proposed methods. Notably, in several standard application settings, the Greedy Mini-EG method attains enhanced convergence rates that go beyond the classical guarantees established for the vanilla EG method.
- We evaluate the proposed methods on two representative tasks: regularized decentralized logistic regression and compressed sensing. On both synthetic and real datasets, the methods achieve speedups exceeding 13× relative to the standard EG algorithm.

The remainder of the paper is organized as follows. Section 2 introduces the problem formulation, basic assumptions, and a brief review of the EG method. In section 3, we present three novel Mini-EG methods and establish convergence analysis and rate guarantees. Extensive numerical experiments in section 4 demonstrate substantial acceleration of the proposed methods over the standard EG algorithm on both regularized decentralized logistic regression and compressed sensing tasks. Finally, section 5 concludes the paper and discusses several potential directions for future research.

Notation. Matrices are denoted by bold uppercase letters $\boldsymbol{A}$, vectors by bold lowercase letters $\boldsymbol{a}$, and scalars by lowercase letters a; $\boldsymbol{A}^T$ is the transpose of $\boldsymbol{A}$, and $\|\boldsymbol{a}\|$, $\|\boldsymbol{a}\|_1$ and $\|\boldsymbol{a}\|_\infty$ denote the ℓ_2-, ℓ_1-, and ℓ_∞-norms of $\boldsymbol{a}$, respectively.

2. Preliminaries. In this paper, we consider the problem of solving the following constrained nonlinear equations:

$$F(\boldsymbol{x}) = \mathbf{0}, \quad \boldsymbol{x} \in \Omega, \tag{2.1}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is a mapping of the form $F(\boldsymbol{x}) = [F_1(\boldsymbol{x}), \ldots, F_n(\boldsymbol{x})]^T$, and $\Omega \subseteq \mathbb{R}^n$ is a nonempty, closed, and convex set.

Most existing works assume that the mapping F is monotone and L-Lipschitz continuous, as defined below:

ASSUMPTION 2.1. *The mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is monotone, i.e.,*

$$\langle F(\boldsymbol{x}) - F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

ASSUMPTION 2.2. *The mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz continuous, i.e., there exists a constant L such that*

$$\|F(\boldsymbol{x}) - F(\boldsymbol{y})\| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Introduced by [11], the EG method is one of the most established approaches for solving monotone equations as in (2.1). Starting from an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, the method alternates between an extragradient step and an update step, as described below:

$$\begin{aligned} \boldsymbol{y}_k &= \boldsymbol{x}_k - \alpha F(\boldsymbol{x}_k), \\ \boldsymbol{x}_{k+1} &= P_\Omega \left(\boldsymbol{x}_k - \alpha F(\boldsymbol{y}_k) \right), \end{aligned}$$

where k is the iteration index, $\alpha > 0$ is the stepsize, and P_Ω denotes the projection operator onto the set Ω. In practice, the choice of α is critical, as it directly affects both convergence speed and numerical stability.

To leverage a more advanced and flexible benchmark, we adopt a two-time-scale EG variant in which the update step uses an adaptive stepsize β_k. This modification alleviates the rigid dependence on the global Lipschitz constant required by the classical EG method:

$$\boxed{\begin{aligned} \boldsymbol{y}_k &= \boldsymbol{x}_k - \alpha F(\boldsymbol{x}_k), \\ \boldsymbol{x}_{k+1} &= P_\Omega \left(\boldsymbol{x}_k - \beta_k F(\boldsymbol{y}_k) \right), \end{aligned}} \tag{2.2}$$

where $\alpha = \rho / L$ with a fixed parameter $\rho \in (0, 1)$. The stepsize β_k is computed as

$$\beta_k = \frac{F(\boldsymbol{y}_k)^T (\boldsymbol{x}_k - \boldsymbol{y}_k)}{\|F(\boldsymbol{y}_k)\|^2}. \tag{2.3}$$

Despite its popularity, the EG method has two notable limitations:

(i) The convergence of EG requires the stepsize α to satisfy $0 < \alpha < 1/L$, where L is the global Lipschitz constant of the mapping F [11]. In practice, L is often difficult to compute or estimate reliably. As a result, manual tuning or line-search procedures are typically employed to select a suitable stepsize, which introduces additional computational overhead and reduces practicality.

(ii) Each iteration of EG requires two full evaluations of the mapping F, which can be computationally intensive or even prohibitive in large-scale problems.

In this paper, we propose a strategy that eliminates the stepsize dependency on the global Lipschitz constant L and requires only a single full evaluation of the mapping F per iteration. This strategy is based on the assumption that F is componentwise Lipschitz continuous [13], which is a strictly weaker condition than Theorem 2.2. The formal definition is given below:

DEFINITION 2.3. *A mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is said to be componentwise Lipschitz continuous if there exist constants $l_1, \ldots, l_n > 0$ such that for all $i = 1, \ldots, n$,*

$$|F_i(\boldsymbol{x} + te_i) - F_i(\boldsymbol{x})| \leq l_i|t|, \quad \forall \boldsymbol{x} \in \mathbb{R}^n, t \in \mathbb{R}, \tag{2.4}$$

where e_i denotes the i-th standard basis vector in $\mathbb{R}^n$.

Remark 2.4. It can be readily verified that if F is L-Lipschitz continuous, then it is also componentwise Lipschitz continuous, with constants $l_i \leq L$ for all $i = 1, \ldots, n$. Moreover, the constants l_i are often easier to compute, especially in large-scale problems. We demonstrate this observation in the numerical experiments section.

Throughout the paper, we assume that both Assumptions 2.1 and 2.2 hold. In addition, we further assume that the mapping F satisfies the following condition:

ASSUMPTION 2.5. *The mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is componentwise Lipschitz continuous with constants $l_1, \ldots, l_n > 0$.*

As discussed in Theorem 2.4, Theorem 2.5 is implied by Theorem 2.2. Therefore, our analysis introduces no additional assumptions.

Remark 2.6. Theorem 2.2 is imposed solely for the purpose of convergence analysis. In practice, the proposed algorithms do not require access to the global Lipschitz constant L; instead, they rely only on the componentwise constants $l_1, \ldots, l_n$, which are substantially easier to compute and are generally much smaller than the global constant L.

3. Mini-Extragradient Methods and Convergence Analysis. Compared with the gradient descent method, the EG method offers the notable advantage of guaranteeing convergence under only Assumptions 2.1 and 2.2. In this work, we further demonstrate that convergence can still be ensured even when the extragradient step updates only a single coordinate at each iteration, offering a more efficient alternative for high-dimensional problems. We refer to these variants as Mini-Extragradient (Mini-EG) methods.

3.1. Greedy Mini-EG. The simplest and most intuitive coordinate selection approach is the greedy scheme, resulting in a variant we refer to as Greedy Mini-EG (G-Mini-EG):

$$\boxed{\begin{aligned} i_k &= \arg\max_{1 \leq i \leq n} |F_i(\boldsymbol{x}_k)|, \\ \boldsymbol{y}_k &= \boldsymbol{x}_k - \frac{\rho}{l_{i_k}} F_{i_k}(\boldsymbol{x}_k) e_{i_k}, \\ \boldsymbol{x}_{k+1} &= P_\Omega\left(\boldsymbol{x}_k - \beta_k F(\boldsymbol{y}_k)\right), \end{aligned}} \tag{3.1}$$

where $\rho \in (0, 1)$ is a fixed parameter, and the stepsize β_k is identical to the EG stepsize rule in (2.3).

Remark 3.1. (i) The proposed procedure relies only on the componentwise Lipschitz constants $l_1, \ldots, l_n$, which are typically much easier to compute and substantially smaller than the global constant L.

(ii) The computation of β_k can be simplified due to the componentwise update structure. Compared with the EG method in (2.3), it avoids full vector products, and the resulting expression is

$$\beta_k = \frac{\rho F_{i_k}(\boldsymbol{y}_k) F_{i_k}(\boldsymbol{x}_k)}{l_{i_k} \|F(\boldsymbol{y}_k)\|^2}. \tag{3.2}$$

We now proceed to establish the convergence of the G-Mini-EG algorithm.

It is well known that the projection operator onto a convex set is nonexpansive; see, for instance, [3]. For completeness, we restate this fundamental property below.

LEMMA 3.2. *let $\Omega \subseteq \mathbb{R}^n$ be a nonempty, closed, and convex set. Then the projection operator P_Ω is 1-Lipschitz continuous; that is,*

$$\|P_\Omega(\boldsymbol{x}) - P_\Omega(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

For the general mini-extragradient step

$$\boldsymbol{y} = \boldsymbol{x} - \frac{\rho}{l_i} F_i(\boldsymbol{x}) e_i, \quad \forall i = 1, \ldots, n, \tag{3.3}$$

we establish the following results:

LEMMA 3.3. *Suppose that Theorem 2.2 holds. For the update in (3.3), it holds that*

$$\|F(\boldsymbol{y})\| \leq \|F(\boldsymbol{x})\| + \frac{\rho L}{l_i} |F_i(\boldsymbol{x})|, \quad \forall i = 1, \ldots, n.$$

Proof. By Theorem 2.2 and the update in (3.3), for any $i = 1, \ldots, n$, we have

$$\|F(\boldsymbol{y}) - F(\boldsymbol{x})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\| = \frac{\rho L}{l_i} |F_i(\boldsymbol{x})|.$$

The result follows immediately by the triangle inequality. ∎

To facilitate the subsequent analysis, we introduce the following relaxed bound.

COROLLARY 3.4. *Under the same assumption as in Lemma 3.3, the update (3.3) satisfies*

$$\|F(\boldsymbol{y})\| \leq \left(1 + \frac{\rho L}{l_i}\right) \|F(\boldsymbol{x})\|, \quad \forall i = 1, \ldots, n.$$

LEMMA 3.5. *Under Theorem 2.5, consider the update in (3.3). Then, for any $i = 1, \ldots, n$, it holds that*

$$F(\boldsymbol{y})^T (\boldsymbol{x} - \boldsymbol{y}) \geq \frac{\rho(1-\rho)}{l_i} |F_i(\boldsymbol{x})|^2.$$

Proof. Given Theorem 2.5 and the form of the update in (3.3), it follows that for any $i = 1, \ldots, n$:

$$\begin{aligned}
(F(\boldsymbol{x}) - F(\boldsymbol{y}))^T (\boldsymbol{x} - \boldsymbol{y}) &= \frac{\rho}{l_i} (F_i(\boldsymbol{x}) - F_i(\boldsymbol{y})) F_i(\boldsymbol{x}) \\
&\leq \frac{\rho}{l_i} |F_i(\boldsymbol{x}) - F_i(\boldsymbol{y})| |F_i(\boldsymbol{x})| \\
&\leq \frac{\rho^2}{l_i} |F_i(\boldsymbol{x})|^2,
\end{aligned}$$

where the last inequality follows from the componentwise Lipschitz condition defined in (2.4).

Rearranging terms, we obtain

$$\begin{aligned}
F(\boldsymbol{y})^T (\boldsymbol{x} - \boldsymbol{y}) &\geq F(\boldsymbol{x})^T (\boldsymbol{x} - \boldsymbol{y}) - \frac{\rho^2}{l_i} |F_i(\boldsymbol{x})|^2 \\
&= \frac{\rho}{l_i} |F_i(\boldsymbol{x})|^2 - \frac{\rho^2}{l_i} |F_i(\boldsymbol{x})|^2 \\
&= \frac{\rho(1-\rho)}{l_i} |F_i(\boldsymbol{x})|^2,
\end{aligned}$$

which completes the proof. ∎

let $\boldsymbol{x}^* \in \Omega$ be a solution to problem (2.1), i.e., $F(\boldsymbol{x}^*) = \mathbf{0}$, and define $l_{\max} = \max_{1\le i\le n} l_i$. With the preparatory results established above, we are ready to present the global convergence analysis of the G-Mini-EG method.

LEMMA 3.6. *Suppose that Assumptions 2.1, 2.2, and 2.5 hold. Let $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$ be the sequence generated by the method* (3.1)*. Then, for all $k \ge 0$, it holds that*

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \ge \frac{\rho^2(1-\rho)^2}{(\sqrt{n}l_{\max} + \rho L)^2}\|F(\boldsymbol{x}_k)\|_\infty^2. \tag{3.4}$$

Proof. From the update rule in (3.1) and Lemma 3.2, we obtain

$$\begin{aligned} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 =& \|P_\Omega\left(\boldsymbol{x}_k - \beta_k F(\boldsymbol{y}_k)\right) - P_\Omega\left(\boldsymbol{x}^*\right)\|^2 \\ \le& \|\boldsymbol{x}_k - \boldsymbol{x}^* - \beta_k F(\boldsymbol{y}_k)\|^2 \\ =& \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \beta_k^2\|F(\boldsymbol{y}_k)\|^2 - 2\beta_k F(\boldsymbol{y}_k)^T(\boldsymbol{x}_k - \boldsymbol{x}^*). \end{aligned} \tag{3.5}$$

By the monotonicity of the mapping F (Theorem 2.1), we have

$$\begin{aligned} F(\boldsymbol{y}_k)^T(\boldsymbol{x}_k - \boldsymbol{x}^*) =& F(\boldsymbol{y}_k)^T(\boldsymbol{x}_k - \boldsymbol{y}_k) + \left(F(\boldsymbol{y}_k) - F(\boldsymbol{x}^*)\right)^T\left(\boldsymbol{y}_k - \boldsymbol{x}^*\right) \\ \ge& F(\boldsymbol{y}_k)^T(\boldsymbol{x}_k - \boldsymbol{y}_k). \end{aligned} \tag{3.6}$$

Substituting (3.6) into (3.5) yields

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \beta_k^2\|F(\boldsymbol{y}_k)\|^2 - 2\beta_k F(\boldsymbol{y}_k)^T(\boldsymbol{x}_k - \boldsymbol{y}_k).$$

Using the definition of β_k in (2.3), we further obtain

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \frac{\left(F(\boldsymbol{y}_k)^T(\boldsymbol{x}_k - \boldsymbol{y}_k)\right)^2}{\|F(\boldsymbol{y}_k)\|^2}.$$

Finally, invoking Lemmas 3.3 and 3.5 yields

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \ge \left(\frac{\rho(1-\rho)|F_{i_k}(\boldsymbol{x}_k)|^2}{l_{i_k}\|F(\boldsymbol{x}_k)\| + \rho L|F_{i_k}(\boldsymbol{x}_k)|}\right)^2.$$

Under the greedy selection rule in (3.1), which ensures $|F_{i_k}(\boldsymbol{x}_k)| = \|F(\boldsymbol{x}_k)\|_\infty$, and using the relation $\|F(\boldsymbol{x}_k)\| \le \sqrt{n}\,\|F(\boldsymbol{x}_k)\|_\infty$, we obtain

$$\begin{aligned} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &\ge \frac{\rho^2(1-\rho)^2}{(\sqrt{n}l_{i_k} + \rho L)^2}\|F(\boldsymbol{x}_k)\|_\infty^2 \\ &\ge \frac{\rho^2(1-\rho)^2}{(\sqrt{n}l_{\max} + \rho L)^2}\|F(\boldsymbol{x}_k)\|_\infty^2, \end{aligned}$$

where the last inequality follows from $l_{i_k} \le l_{\max}$. ∎

THEOREM 3.7. *Suppose Assumptions 2.1, 2.2, and 2.5 hold, and let the sequences $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$ be generated by the method* (3.1)*. Then, for any integer $K \ge 0$, we have*

$$\min_{0\le k\le K}\|F(\boldsymbol{x}_k)\|_\infty^2 \le \frac{(\sqrt{n}l_{\max} + \rho L)^2}{\rho^2(1-\rho)^2(K+1)}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2. \tag{3.7}$$

In addition, the residual sequence vanishes asymptotically:

$$\liminf_{k\to\infty}\|F(\boldsymbol{x}_k)\|_\infty = 0. \tag{3.8}$$

Proof. Summing inequality (3.4) from $k = 0$ to K, we obtain

$$\frac{\rho^2(1-\rho)^2}{(\sqrt{n}l_{\max} + \rho L)^2} \sum_{k=0}^{K} \|F(\boldsymbol{x}_k)\|_\infty^2 \leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{K+1} - \boldsymbol{x}^*\|^2 \leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2.$$

On the other hand, the minimum is bounded by the average:

$$\min_{0\leq k\leq K} \|F(\boldsymbol{x}_k)\|_\infty^2 \leq \frac{1}{K+1} \sum_{k=0}^{K} \|F(\boldsymbol{x}_k)\|_\infty^2.$$

Combining the two inequalities yields (3.7). Taking the limit as $K \to \infty$ gives (3.8), which completes the proof. ∎

Define the coupling factor $\kappa = L/l_{\max}$, which measures how much the global Lipschitz constant can exceed the directional derivative bounds along coordinate axes. We use this quantity to compare the convergence behavior of the G-Mini-EG method with that of the classical EG algorithm. It is well known that $\kappa \in [1, n]$ [19]. A key observation is that, unlike the convergence guarantees for classical coordinate-descent methods [13], the coefficient of L in Theorem 3.7 is independent of the problem dimension n. Moreover, in regimes where $\kappa > \sqrt{n}$, which frequently occur in least absolute shrinkage and selection operator (LASSO) problems [16] (see (4.6) and (4.7) in subsection 4.2), the worst-case complexity of the G-Mini-EG method coincides with that of the standard EG algorithm.[1]

3.2. Random Mini-EG. Although we have rigorously established the convergence of the G-Mini-EG, where only a single coordinate is updated at each extragradient step and the stepsize is determined solely by componentwise Lipschitz constants, the method still requires two full evaluations of the mapping F per iteration to identify the dominant component. This becomes computationally prohibitive in large-scale settings.

To address this limitation, we introduce a more efficient strategy inspired by random coordinate descent [13]. At each iteration k, the index i_k is sampled according to a probability distribution, leading to the following variant, which we refer to as the Random Mini-EG (R-Mini-EG):

$$\boxed{\begin{aligned} i_k &= \mathcal{A}_\gamma, \\ \boldsymbol{y}_k &= \boldsymbol{x}_k - \frac{\rho}{l_{i_k}} F_{i_k}(\boldsymbol{x}_k) e_{i_k}, \\ \boldsymbol{x}_{k+1} &= P_\Omega\left(\boldsymbol{x}_k - \beta_k F(\boldsymbol{y}_k)\right), \end{aligned}} \tag{3.9}$$

Here, the operation $i_k = \mathcal{A}_\gamma$ indicates that the index $i_k \in \{1, \ldots, n\}$ is drawn independently according to the probability distribution

$$p_\gamma(i) = \frac{l_i^\gamma}{\sum_{j=1}^n l_j^\gamma}, \quad \forall i = 1, \ldots, n,$$

[1] For a random matrix with independent, zero-mean entries of finite variance, random matrix theory provides an average-case estimate [2]. The spectral norm (global Lipschitz constant) scales as $\Theta(\sqrt{n})$, whereas the maximum diagonal element typically scales only as $\Theta(\sqrt{\log n})$. Consequently, the expected coupling factor grows as $\Theta(\sqrt{n/\log n})$.

with parameter $\gamma \geq 0$ is a tunable parameter. The scalar $\rho \in (0,1)$, and the stepsize β_k is computed as in (3.2).

To establish the convergence of the R-Mini-EG, we first introduce the following weighted norm [13]:

$$\|\boldsymbol{x}\|_{[\gamma]} = \sqrt{\sum_{i=1}^{n} l_i^{\gamma} x_i^2}. \tag{3.10}$$

LEMMA 3.8. *Let Assumptions 2.1, 2.2, and 2.5 hold. Consider the iterates $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$ generated by the method* (3.9)*. Then, for all $k \geq 0$, the following inequality holds:*

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - E_{i_k}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right] \geq \frac{\rho^2(1-\rho)^2\|F^2(\boldsymbol{x}_k)\|_{[\gamma]}^2}{\sum_{i=1}^{n} l_i^{\gamma}(l_{\max}+\rho L)^2\|F(\boldsymbol{x}_k)\|^2}. \tag{3.11}$$

where $F^2(\boldsymbol{x}) = [F_1^2(\boldsymbol{x}), \ldots, F_n^2(\boldsymbol{x})]^T$.

Proof. Proceeding analogously to the proof of Lemma 3.6 and using the looser bound in Corollary 3.4 instead of Lemma 3.3, we arrive at

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \geq \frac{\rho^2(1-\rho)^2|F_{i_k}(\boldsymbol{x}_k)|^4}{(l_{\max}+\rho L)^2\|F(\boldsymbol{x}_k)\|^2}. \tag{3.12}$$

Taking expectation with respect to the random variable i_k, we have

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - E_{i_k}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right] \geq \frac{\rho^2(1-\rho)^2}{(l_{\max}+\rho L)^2\|F(\boldsymbol{x}_k)\|^2}\sum_{i=1}^{n}\frac{l_i^{\gamma}}{\sum_{j=1}^{n} l_j^{\gamma}}|F_i(\boldsymbol{x}_k)|^4,$$

which proves inequality (3.11). ∎

THEOREM 3.9. *Assume that Assumptions 2.1, 2.2, and 2.5 hold. Let $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$ be the iterates generated by the method* (3.9)*. Then, for any integer $K \geq 0$, the following bound holds:*

$$\min_{0\leq k\leq K} E_{i_{k-1}}\left[\frac{\|F^2(\boldsymbol{x}_k)\|_{[\gamma]}^2}{\|F(\boldsymbol{x}_k)\|^2}\right] \leq \frac{\sum_{i=1}^{n} l_i^{\gamma}(l_{\max}+\rho L)^2}{\rho^2(1-\rho)^2(K+1)}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2. \tag{3.13}$$

Furthermore, the expected residual converges to zero:

$$\liminf_{k\to\infty} E_{i_{k-1}}\left[\|F(\boldsymbol{x}_k)\|\right] = 0. \tag{3.14}$$

Proof. Taking expectation with respect to i_{k-1} on both sides of inequality (3.11), we obtain

$$\begin{aligned} &E_{i_{k-1}}\left[\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2\right] - E_{i_k}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right] \\ &\geq \frac{\rho^2(1-\rho)^2}{\sum_{i=1}^{n} l_i^{\gamma}(l_{\max}+\rho L)^2} E_{i_{k-1}}\left[\frac{\|F^2(\boldsymbol{x}_k)\|_{[\gamma]}^2}{\|F(\boldsymbol{x}_k)\|^2}\right]. \end{aligned}$$

Summing over $k = 0$ to K yields the bound in (3.13), following a similar argument as in the proof of Theorem 3.7.

Taking the limit as $K \to \infty$, we obtain

$$\liminf_{k\to\infty} E_{i_{k-1}}\left[\frac{\|F^2(\boldsymbol{x}_k)\|_{[\gamma]}^2}{\|F(\boldsymbol{x}_k)\|^2}\right] = 0,$$

which directly implies the result in (3.14). ∎

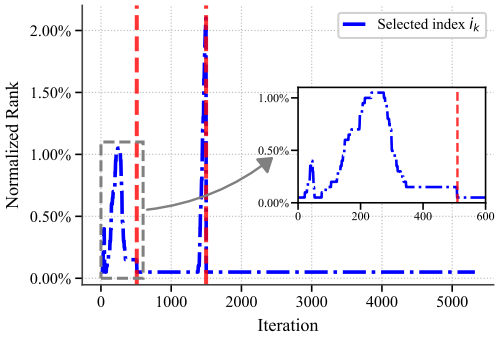


Fig. 1: Slow decay of dominant component magnitudes in Watchdog-Max on colon-cancer dataset. The y-axis shows the normalized rank, computed as the rank of the component magnitude at the selected index i_k divided by the feature dimension n. The red dashed line marks the detection of condition (3.15), triggering adaptive window reset.

3.3. Watchdog-Max. Although R-Mini-EG overcomes the need for two full evaluations of the mapping F per iteration in G-Mini-EG, empirical evidence shows that its purely random coordinate selection often requires substantially more iterations, resulting in higher overall computation time. This raises a natural question: how can we design a strategy that retains the computational efficiency of single-evaluation schemes while preserving the fast convergence behavior of greedy approaches?

An important observation is that the magnitude of the mapping component corresponding to the dominant coordinate tends to decay slowly. Specifically, if at k_0 the index $i_{k_0} = \arg\max_{1\le i\le n} |F_i(\boldsymbol{x}_{k_0})|$ is selected, then over a relatively long window of size T, the component $|F_{i_{k_0}}(\boldsymbol{x}_k)|$ consistently remains among the top-ranked entries for all $k_0 \le k \le k_0 + T$.

Remark 3.10. Although the function value corresponding to i_{k_0} may not remain the exact top-1 entry throughout the window $k_0 \le k \le k_0+T$, it typically stays among the top-ranked coordinates. As a result, the per-iteration efficiency is comparable to that of strictly applying the greedy scheme at every step. The key advantage of this approach is that it requires only a single full evaluation of $F(\boldsymbol{x}_{k_0})$ at the beginning of the window, while avoiding the need to compute $F(\boldsymbol{x}_k)$ at subsequent iteration within the window. This enables the selection of approximately greedy-optimal coordinates with significantly reduced computational cost.

The window size T can be adaptively adjusted by incorporating a random scheme. Specifically, during each window period $k_0 \le k \le k_0 + T$, we compute one additional coordinate value $F_{i_k^r}(\boldsymbol{x}_k)$ at each iteration, where $i_k^r = \mathcal{A}_\gamma$. If at any iteration k it is observed that

$$|F_{i_{k_0}}(\boldsymbol{x}_k)| < |F_{i_k^r}(\boldsymbol{x}_k)|, \tag{3.15}$$

then iteration k is designated as the new starting point of the next window, and the reference coordinate is updated to i_k^r.

We refer to this adaptive strategy as Watchdog-Max, and its iteration procedure is summarized as follows. Starting from an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, we first compute

the initial index:

$$i_0 = \arg\max_{1\le i\le n} |F_i(\boldsymbol{x}_0)|.$$

Then, for each iteration $k \ge 1$, we perform:

$$\boxed{\begin{aligned} i_k^r &= \mathcal{A}_\gamma, \\ i_k &= \arg\max_{i_{k-1}, i_k^r} \left\{|F_{i_{k-1}}(\boldsymbol{x}_k)|, |F_{i_k^r}(\boldsymbol{x}_k)|\right\}, \\ \boldsymbol{y}_k &= \boldsymbol{x}_k - \frac{\rho}{l_{i_k}} F_{i_k}(\boldsymbol{x}_k) e_{i_k}, \\ \boldsymbol{x}_{k+1} &= P_\Omega\left(\boldsymbol{x}_k - \beta_k F(\boldsymbol{y}_k)\right). \end{aligned}} \tag{3.16}$$

Here, $\rho \in (0,1)$ is a scalar parameter, and β_k is the stepsize defined in (3.2).

Remark 3.11. (i) The Watchdog-Max method performs a full evaluation of the mapping F only once at the initial point to determine the initial greedy-optimal index. In all subsequent iterations, it evaluates only two coordinate values to execute the mini-extragradient step. The inclusion of a randomly sampled coordinate i_k^r enables the algorithm to adaptively update the candidate for the greedy-optimal index. This design substantially reduces the computational cost per iteration and effectively balances the greedy and random selection strategies.

(ii) Empirical evidence indicates that computing the exact greedy-optimal index at every iteration is unnecessary. In fact, the index selection strategy of Watchdog-Max can, in some cases, lead to fewer iterations than G-Mini-EG (see the numerical results in section 4).

According to the index selection rule in (3.16), each iteration satisfies

$$|F_{i_k}(\boldsymbol{x}_k)| \ge |F_{i_k^r}(\boldsymbol{x}_k)|, \tag{3.17}$$

where $i_k^r = \mathcal{A}_\gamma$. Substituting (3.17) into (3.12) immediately yields the following result.

COROLLARY 3.12. *Assume that Assumptions 2.1, 2.2, and 2.5 hold. Then, the convergence guarantees established in Theorem* 3.9 *also apply to the Watchdog-Max method.*

With these results, we fully resolve the two challenges posed in the beginning. Although R-Mini-EG reduces the per-extragradient-step cost from a full mapping evaluation to essentially a single coordinate evaluation from a theoretical standpoint, the Watchdog-Max strategy further enhances practical efficiency by requiring only one additional coordinate evaluation per iteration.

4. Numerical Experiments. In this section, we evaluate the meta-level improvements achieved by the proposed Mini-EG variants over the vanilla EG method on two representative tasks: regularized decentralized logistic regression and compressed sensing. The comparison is carried out against the two-time-scale EG baseline in (2.2). Additional comparisons with recent baseline methods are provided in Table 2 of section A.

To ensure a fair comparison, the same stepsize strategy is applied to both the EG and Mini-EG variants. For all algorithms, we set $\rho = 0.999$ to allow a relatively large stepsize. The sensitivity of ρ for both EG and Mini-EG is further examined in section B. In addition, we set $\gamma = 0$ for the Watchdog-Max method. All algorithms terminate when $\|F(\boldsymbol{y}_k)\| \le 10^{-8}$ or when the iteration count exceeds 500,000.

Table 1: Average performance on three real-world datasets. Best results are shown in **bold**. The last column reports the *speedup* of Watchdog-Max over EG, calculated as the ratio of their Tcpu values.

Method	Tcpu	Itr	NF	$\|F^*\|$	*Speedup*
Colon-cancer [1]					
EG	2.54	15606.50	31213.00	9.87E-09	
G-Mini-EG	0.87	**5989.40**	11978.80	9.98E-09	*3.59*×
Watchdog-Max	**0.71**	6137.90	**6146.44**	9.99E-09	
Duke breast-cancer [18]					
EG	34.44	130408.60	260817.20	9.94E-09	
G-Mini-EG	4.49	**20225.00**	40450.00	9.91E-09	*9.25*×
Watchdog-Max	**3.72**	20584.10	**20593.17**	9.89E-09	
Leukemia [7]					
EG	23.61	91100.40	182200.80	9.83E-09	
G-Mini-EG	5.00	**24479.40**	48958.80	9.69E-09	*5.90*×
Watchdog-Max	**4.01**	25384.10	**25394.62**	9.71E-09	

Each experiment is repeated over 100 independent trials. We report the average number of iterations (`Itr`), function evaluations (`NF`)[2], CPU time in seconds (`Tcpu`), and the final residual $\|F(\boldsymbol{x})\|$ ($\|F^*\|$).

All experiments were performed in Python 3.9.12 on a Windows system, using a laptop equipped with an Intel i7-12700H processor (2.30 GHz) and 16 GB RAM.

4.1. Regularized Decentralized Logistic Regression. We first consider regularized decentralized logistic regression [9], which arises in regimes where the feature dimension n far exceeds the sample size N. The problem is formulated as:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1+\exp\left(-b_i \boldsymbol{a}_i^T \boldsymbol{x}\right)\right) + \frac{\tau}{2}\|\boldsymbol{x}\|^2, \tag{4.1}$$

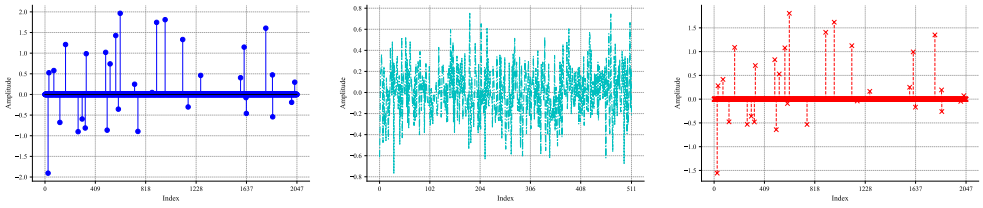
where $\tau > 0$ is a regularization parameter, and $(\boldsymbol{a}_i, b_i) \in \mathbb{R}^n \times \{-1, 1\}$ for $i = 1, \ldots, N$ are data samples drawn from a given dataset or distribution. In all experiments, we fix the regularization parameter as $\tau = 0.1$.

It is well known that the objective function f in (4.1) is smooth and strongly convex, owing to the presence of both the smooth logistic loss and the ℓ_2 regularization term. Therefore, solving (4.1) is equivalent to finding the root of the following system of nonlinear monotone equations:

$$F(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{-b_i \exp\left(-b_i \boldsymbol{a}_i^T \boldsymbol{x}\right)}{1+\exp\left(-b_i \boldsymbol{a}_i^T \boldsymbol{x}\right)} \boldsymbol{a}_i + \tau \boldsymbol{x} = \boldsymbol{0}.$$

Let $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]$ and define $\boldsymbol{H} = \boldsymbol{A}\boldsymbol{A}^T$. Through standard analysis, it can be verified that the mapping F satisfies both Assumptions 2.2 and 2.5. Specifically, the

[2]For Mini-EG methods, each component function evaluation counts as $1/n$ toward `NF`.



(a) (b) (c)

Fig. 2: Signal recovery visualization using Watchdog-Max with $n = 2048$, $N = 512$, $K = 32$, and SNR $= 20$ dB. (a) Original signal; (b) Observation; (c) Reconstruction.

global Lipschitz constant of F is given by

$$L = \frac{1}{4N}\lambda_1(\boldsymbol{H}) + \tau, \tag{4.2}$$

where $\lambda_1(\boldsymbol{H})$ denotes the largest eigenvalue of $\boldsymbol{H}$.

In contrast, the componentwise Lipschitz constants take the form

$$l_i = \frac{1}{4N}h_{ii} + \tau, \quad i = 1, \ldots, n, \tag{4.3}$$

where h_{ii} is the (i,i)-th diagonal entry of $\boldsymbol{H}$.

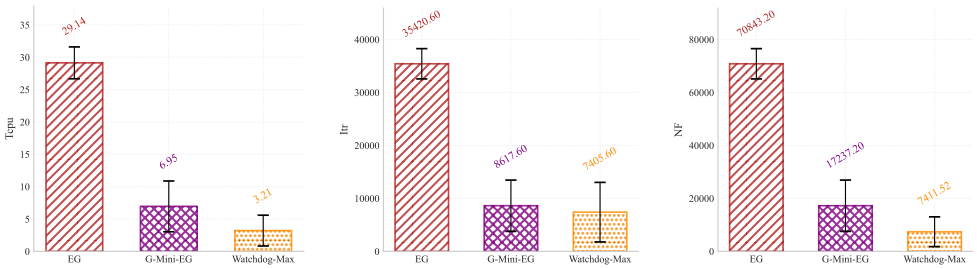
Remark 4.1. Comparing (4.2) and (4.3), computing the constant L requires estimating the dominant eigenvalue of $\boldsymbol{H}$, which is computationally expensive when n is large. In contrast, computing each l_i only needs a diagonal entry of $\boldsymbol{H}$, rendering its computation considerably more efficient.

Our experiments use three real-world datasets from the LIBSVM repository[3] [4]: colon-cancer ($N = 62$, $n = 2{,}000$), duke breast-cancer ($N = 44$, $n = 7{,}129$), and leukemia ($N = 38$, $n = 7{,}129$).

Figure 1 illustrates the slow decay of dominant component magnitudes in the Watchdog-Max method on the colon-cancer dataset. Throughout the iterations, the component selected at index i_k consistently ranks within the top 2%, with only two window resets triggered. In the final window, it persistently coincides with the top-1 component. This observation indicates that performing full evaluations at every extragradient step is unnecessary for identifying the dominant component.

Table 1 demonstrates that the Mini-EG variants outperform the standard EG method, with Watchdog-Max achieving a 9.25× speedup on duke breast-cancer data. Notably, Mini-EG methods reduce both the per-iteration cost and the total number of iterations compared with EG. Among them, G-Mini-EG attains the lowest iteration count, whereas Watchdog-Max requires slightly more iterations but updates only two components per extragradient step, which leads to substantially lower `NF` and `Tcpu`. These results indicate that full mapping evaluations at every extragradient step are not necessary, and that updating only a single coordinate per extragradient step already provides significant computational gains.

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/



(a) (b) (c)

Fig. 3: Performance comparison of different methods is shown, in terms of (a) `Tcpu`, (b) `Itr`, and (c) `NF`.

4.2. Compressed Sensing. We consider the LASSO problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \tau\|\boldsymbol{x}\|_1, \tag{4.4}$$

where $\boldsymbol{A}\in\mathbb{R}^{N\times n}$ $(N\le n)$ is the sensing matrix, $\boldsymbol{b}\in\mathbb{R}^N$ is the observation vector, and $\tau>0$ is a regularization parameter promoting sparsity. Following [10], we set $\tau = 0.1\|\boldsymbol{A}^T\boldsymbol{b}\|_\infty$.

By decomposing $\boldsymbol{x}$ into its positive and negative parts, i.e., $\boldsymbol{x}=\boldsymbol{u}-\boldsymbol{v}$, problem (4.4) can be reformulated as a system of nonlinear monotone equations [6]:

$$F(\boldsymbol{z}) = \min\{\boldsymbol{z}, \boldsymbol{H}\boldsymbol{z}+\boldsymbol{c}\} = \boldsymbol{0}, \tag{4.5}$$

where

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}, \quad \boldsymbol{H} = \begin{bmatrix} \boldsymbol{A}^T\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{A} \\ -\boldsymbol{A}^T\boldsymbol{A} & \boldsymbol{A}^T\boldsymbol{A} \end{bmatrix},$$

and

$$\boldsymbol{c} = \tau\boldsymbol{1}_{2n} + \begin{bmatrix} -\boldsymbol{A}^T\boldsymbol{b} \\ \boldsymbol{A}^T\boldsymbol{b} \end{bmatrix},$$

with $\boldsymbol{1}_{2n}$ denoting the all-one vector in $\mathbb{R}^{2n}$. The operator $\min\{\cdot\}$ is applied componentwise.

Following [14, Lemma 3], it can be shown that the mapping F in (4.5) has a global Lipschitz constant

$$L = \sqrt{n}\max\{\lambda_1(\boldsymbol{H}), 1\}, \tag{4.6}$$

and componentwise Lipschitz constants

$$l_i = \max\{h_{ii}, 1\}, \quad i = 1,\dots,2n. \tag{4.7}$$

Remark 4.2. From (4.6) and (4.7), we observe that l_i is easier to compute and typically smaller than L. Furthermore, leveraging the structure of $\boldsymbol{H}$, it suffices to compute only the first n diagonal entries h_{ii}, i.e., the diagonal elements of $\boldsymbol{A}^T\boldsymbol{A}$.

Table 2: Average performance comparison of Mini-EG and recent EG-type methods in compressed sensing under different configurations. The best results are highlighted in **bold**, and the *speedup* of Watchdog-Max over EG is also reported.

Method	$K = 8$	$K = 16$	$K = 32$
	TCPU/Itr/NF/$\|F^*\|$	**TCPU/Itr/NF/$\|F^*\|$**	**TCPU/Itr/NF/$\|F^*\|$**
	$N = 512$		
EG	6.24/24830.80/49661.60/1.00E-08	6.94/28310.80/56621.60/1.00E-08	8.32/35517.50/71035.00/1.00E-08
PEG	4.03/24851.10/24851.10/1.00E-08	4.23/28331.80/28331.80/1.00E-08	5.07/35538.60/35538.60/1.00E-08
EAG-C	11.44/50000.00/100000.00/4.96E-04	11.51/50000.00/100000.00/6.64E-04	11.41/50000.00/100000.00/9.27E-04
FEG	11.55/50000.00/100000.00/1.70E-01	11.75/50000.00/100000.00/2.29E-01	11.64/50000.00/100000.00/3.21E-01
G-Mini-EG	0.92/4707.50/9415.00/1.00E-08	1.55/**7007.80**/14015.60/1.00E-08	1.45/7508.20/15016.40/1.00E-08
Watchdog-Max	**0.46**/**4020.60**/**4024.36**/1.00E-08	**0.83**/7595.90/**7601.01**/1.00E-08	**0.73**/**6347.90**/**6352.60**/1.00E-08
Speedup	*13.51×*	*8.35×*	*11.36×*
	$N = 768$		
EG	5.09/18946.60/37893.20/1.00E-08	5.44/20763.70/41527.40/1.00E-08	5.85/23137.00/46274.00/1.00E-08
PEG	3.36/18966.90/18966.90/9.99E-09	3.54/20784.60/20784.60/1.00E-08	3.81/23158.10/23158.10/1.00E-08
EAG-C	11.32/50000.00/100000.00/4.29E-04	11.29/50000.00/100000.00/7.33E-04	11.20/50000.00/100000.00/9.20E-04
FEG	11.49/50000.00/100000.00/1.13E-01	11.55/50000.00/100000.00/1.94E-01	11.40/50000.00/100000.00/2.44E-01
G-Mini-EG	0.70/**3626.50**/7253.00/1.00E-08	0.74/3957.60/7915.20/1.00E-08	1.34/6860.20/13720.40/1.00E-08
Watchdog-Max	**0.62**/5134.80/**5138.71**/1.00E-08	**0.45**/**3486.90**/**3490.00**/1.00E-08	**0.70**/**5939.70**/**5943.80**/1.00E-08
Speedup	*8.24×*	*12.19×*	*8.31×*
	$N = 1024$		
EG	4.52/15660.20/31320.40/1.00E-08	4.68/16771.10/33542.20/1.00E-08	5.08/18511.30/37022.60/1.00E-08
PEG	3.12/15680.60/15680.60/9.99E-09	3.10/16791.60/16791.60/1.00E-08	3.36/18532.40/18532.40/9.99E-09
EAG-C	11.25/50000.00/100000.00/5.06E-04	11.47/50000.00/100000.00/6.72E-04	11.69/50000.00/100000.00/9.65E-04
FEG	11.46/50000.00/100000.00/1.13E-01	11.67/50000.00/100000.00/1.49E-01	11.77/50000.00/100000.00/2.14E-01
G-Mini-EG	1.03/**5481.90**/10963.80/1.00E-08	1.12/**5556.30**/11112.60/1.00E-08	1.31/7032.70/14065.40/1.00E-08
Watchdog-Max	**0.63**/5634.40/**5638.75**/1.00E-08	**0.73**/6082.20/**6086.77**/1.00E-08	**0.77**/**6428.20**/**6432.74**/1.00E-08
Speedup	*7.14×*	*6.40×*	*6.58×*

We generate synthetic datasets by randomly sampling sparse signals with sparsity level K. The sensing matrix $\boldsymbol{A}$ is drawn from a standard Gaussian distribution, and the observation signal-to-noise ratio (SNR) is set to 20 dB.

Under the setting $n = 2048$, $N = 512$, and $K = 32$ with a zero initial point, Figure 2 illustrates the signal recovery performance of Watchdog-Max. Figure 3 compares methods via bar plots (mean±std) for `Tcpu`, `Itr`, and `NF`. Watchdog-Max outperforms all baselines, achieving over 9× and 2× speedup in `Tcpu` versus EG and G-Mini-EG, respectively. Notably, it requires fewer iterations than G-Mini-EG, indicating that the greedy index-selection scheme in G-Mini-EG admits further improvement.

5. Conclusion and Future Work. This work introduces a family of Mini-EG methods for solving monotone nonlinear equations. By updating only the dominant coordinate and employing componentwise Lipschitz constants, the proposed G-Mini-EG method achieves more efficient iterations together with stepsizes that are substantially easier to compute. To further reduce computational overhead, the Watchdog-Max strategy incorporates randomized sampling so that only two coordinates are tracked at each extragradient step, while still maintaining both convergence guarantees and computational efficiency. Theoretical results are established for all proposed variants. Notably, in a range of commonly encountered problem settings, the Greedy Mini-EG method admits sharper convergence bounds than those available for the standard EG method. Empirical evaluations on decentralized logistic regression and compressed sensing further demonstrate runtime improvements of more than 13× over the classical EG algorithm.

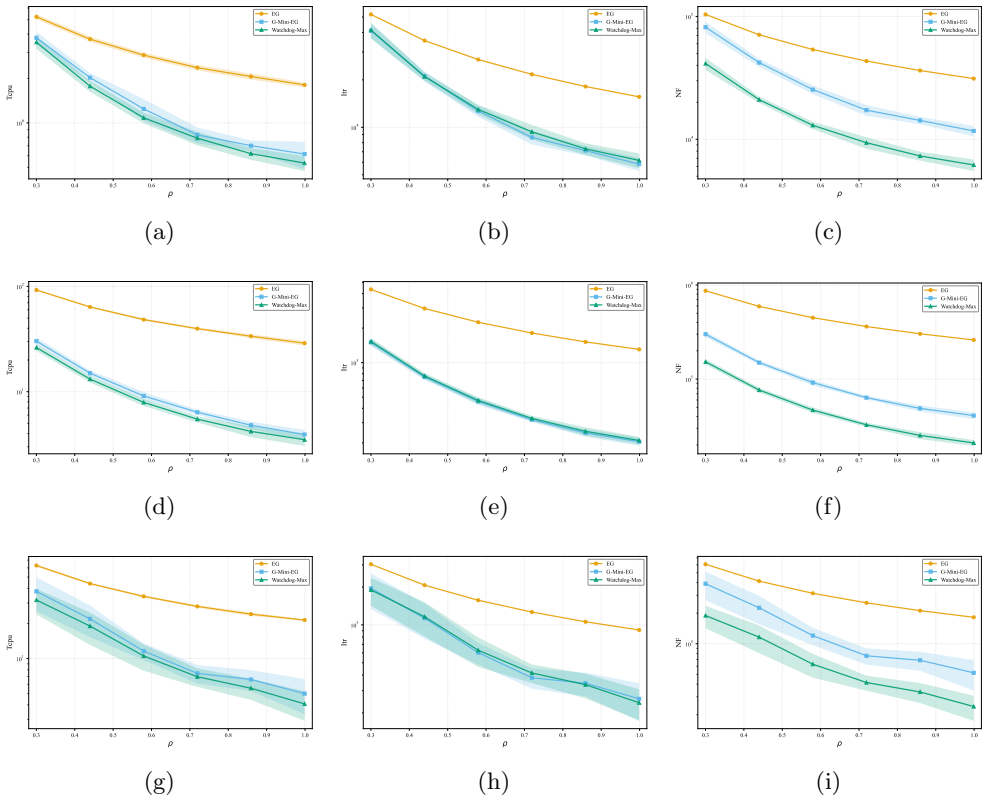


Fig. 4: Sensitivity of ρ for EG and Mini-EG on regularized decentralized logistic regression, measured by `Tcpu`, `Itr`, and `NF`. Curves show average results, and shaded areas indicate one standard deviation. Datasets: colon-cancer (a-c), duke breast-cancer (d-f), leukemia (g-i).

Future work. These findings significantly advance the applicability of EG-type methods to problems in optimization and scientific computing. Future work will investigate Mini-Batch extensions by selecting a subset of coordinates at each step, and explore integrating the mini-update strategy into both stages of the algorithm, paving the way for a potential Double-Mini-EG framework. Additionally, we plan to analyze the last-iterate convergence rate to address a gap in the convergence theory of Mini-EG methods.

Appendix A. Additional Comparisons with Recent Baselines.

In this section, we present additional comparisons between the proposed Mini-EG algorithms and several recent baseline methods, including PEG [15], EAG [21], and FEG [12], in the context of compressed sensing. We vary $N \in \{512, 768, 1024\}$ and $K \in \{8, 16, 32\}$ under $n = 2048$ and SNR $= 20$ dB. The maximum number of iterations is set to 50,000. The same stepsize is adopted for PEG, EG, and Mini-EG. For EAG with constant stepsize (EAG-C), the stepsize is tuned to the largest value within its convergence range. All parameters of FEG follow the default settings recommended in [12].

Table 2 reports the averaged results over all configurations. As observed, the Mini-EG variants consistently outperform existing EG-type methods across all performance metrics. In particular, Watchdog-Max achieves the best results in terms of `Tcpu` in every case, with a maximum speedup exceeding 13× relative to EG. It also demonstrates a clear improvement over G-Mini-EG, which is mainly due to the fact that each extragradient step in Watchdog-Max updates only two coordinates, thereby considerably reducing the `NF` while maintaining comparable iteration counts. Moreover, both EAG-C and FEG fail to reach the prescribed accuracy within the iteration limit, converging only to low-accuracy solutions.

Appendix B. Sensitivity Analysis of the Parameter ρ.

In this section, we analyze the sensitivity of the parameter ρ and its impact on the performance of the EG and Mini-EG methods. As shown in Figure 4, for different values of $\rho \in (0, 1)$, both G-Mini-EG and Watchdog-Max consistently outperform EG in terms of `Tcpu`, `Itr`, and `NF`. Moreover, as ρ increases, the acceleration achieved by Mini-EG relative to EG becomes more pronounced. These observations suggest that selecting ρ close to 1 can be beneficial for achieving faster convergence in certain scenarios.[4]

REFERENCES

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 6745–6750.

[2] A. S. Bandeira and R. van Handel, *Sharp nonasymptotic bounds on the norm of random matrices with independent entries*, Ann. Probab., 44 (2016), pp. 2479–2506.

[3] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011.

[4] C.-C. Chang and C.-J. Lin, *Libsvm: A library for support vector machines*, ACM Trans. Intell. Syst. Technol., 2 (2011), pp. 1–27.

[5] J. Diakonikolas, C. Daskalakis, and M. I. Jordan, *Efficient methods for structured nonconvex-nonconcave min-max optimization*, in Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 2746–2754.

[6] M. A. Figueiredo, R. D. Nowak, and S. J. Wright, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, IEEE J. Sel. Top. Signal Process., 1 (2008), pp. 586–597.

[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, Science, 286 (1999), pp. 531–537.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Proceedings of the 28th International Conference on Neural Information Processing Systems, vol. 27, Curran Associates, Inc., 2014.

[9] J. Jian, J. Yin, C. Tang, and D. Han, *A family of inertial derivative-free projection methods for constrained nonlinear pseudo-monotone equations with applications*, Comput. Appl. Math., 41 (2022), p. 309.

[10] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, *An interior-point method for large-scale ℓ_1-regularized least squares*, IEEE J. Sel. Top. Signal Process., 1 (2008), pp. 606–617.

[11] G. M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon, 12 (1976), pp. 747–756.

[4] Although, in theory, setting ρ near 1 does not necessarily yield the tightest convergence rate bound, in practice faster convergence may arise because l_{i_k} remains an overestimate.

[12] S. Lee and D. Kim, *Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems*, in Proceedings of the 35th International Conference on Neural Information Processing Systems, vol. 34, 2021, pp. 22588–22600.
[13] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM J. Optim., 22 (2012), pp. 341–362.
[14] J.-S. Pang, *Inexact newton methods for the nonlinear complementarity problem*, Math. Program., 36 (1986), pp. 54–71.
[15] L. D. Popov, *A modification of the arrow-hurwicz method for search of saddle points*, Math. Notes, 28 (1980), pp. 845–848.
[16] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B, 58 (1996), pp. 267–288.
[17] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, *Generalizing from a few examples: A survey on few-shot learning*, ACM Comput. Surv., 53 (2020), pp. 1–34.
[18] J. Wen, C.-N. Yu, and R. Greiner, *Robust learning under uncertain test distributions: Relating covariate shift to model misspecification*, in Proceedings of the 31st International Conference on Machine Learning, PMLR, 2014, pp. 631–639.
[19] S. J. Wright, *Coordinate descent algorithms*, Math. Program., 151 (2015), pp. 3–34.
[20] Y. Xiao, Q. Wang, and Q. Hu, *Non-smooth equations based method for ℓ_1-norm problems with applications to compressed sensing*, Nonlinear Anal. Theor., 74 (2011), pp. 3570–3577.
[21] T. Yoon and E. K. Ryu, *Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm*, in Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 12098–12109.