A Single-Loop Accelerated Extra-Gradient Difference Algorithm with Improved Complexity Bounds for Constrained Minimax Optimization

Yuanyuan Liu[1], Fanhua Shang[2], Weixin An[1], Junhao Liu[1], Hongying Liu[2] and Zhouchen Lin[3]

[1]Xidian University
[2]Tianjin University
[3]Peking University

Optimization Problems

■ Constrained Minimax Optimization Problem

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y),$$

where $\mathcal{X} \subseteq \mathbb{R}^m$ and $\mathcal{Y} \subseteq \mathbb{R}^n$ are nonempty closed and convex feasible sets, and $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is a smooth function. In recent years, this problem has drawn considerable interest from machine learning and other engineering communities such as:

- Generative adversarial networks
- Adversarial machine learning
- Game theory
- Reinforcement learning
- Empirical risk minimization
- Robust optimization

Motivations

- For NC-C minimax problems, can we design a single-loop directly accelerated algorithm with the gradient complexity lower than the best-known result $\tilde{O}\ (\epsilon^{-2.5})$?

- Can we propose a single-loop directly accelerated algorithm with the complexity lower than the best-known result O(ϵ^{-4}) for constrained C-NC and NC-NC minimax problems?

Contributions

- We design a new single-loop accelerating algorithm for solving constrained NC-NC problems. In the proposed algorithm, we design a new extra-gradient difference scheme, and combine the gradient ascent and momentum acceleration steps for the dual variable update.

- We analyze the convergence properties of the proposed algorithm for constrained NC-NC problems. Theorem 1 shows that to find an ϵ-stationary point of f, the proposed algorithm can obtain the gradient complexity O(ϵ^{-2}), which is the first time to attains the complexity bound in constrained NC-NC setting, as shown in Table 1.

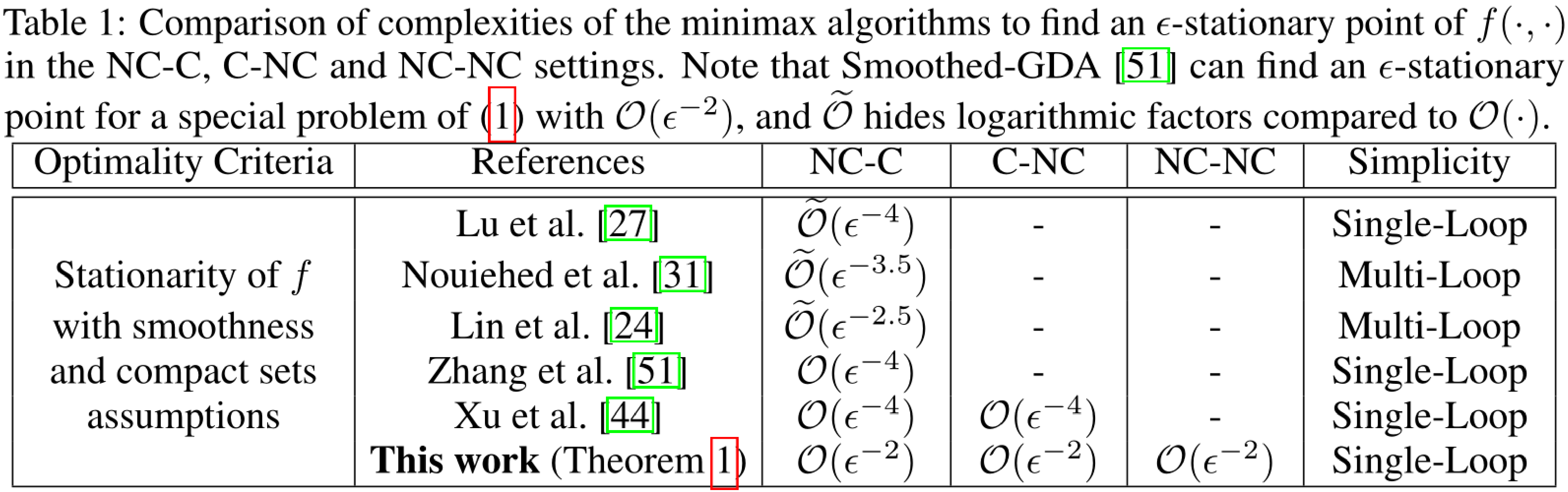


Table 1: Comparison of complexities of the minimax algorithms to find an ϵ-stationary point of $f(\cdot,\cdot)$ in the NC-C, C-NC and NC-NC settings. Note that Smoothed-GDA [51] can find an ϵ-stationary point for a special problem of (1) with $\mathcal{O}(\epsilon^{-2})$, and $\widetilde{\mathcal{O}}$ hides logarithmic factors compared to $\mathcal{O}(\cdot)$.

Optimality Criteria	References	NC-C	C-NC	NC-NC	Simplicity
Stationarity of f with smoothness and compact sets assumptions	Lu et al. [27]	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	-	-	Single-Loop
	Nouiehed et al. [31]	$\widetilde{\mathcal{O}}(\epsilon^{-3.5})$	-	-	Multi-Loop
	Lin et al. [24]	$\widetilde{\mathcal{O}}(\epsilon^{-2.5})$	-	-	Multi-Loop
	Zhang et al. [51]	$\mathcal{O}(\epsilon^{-4})$	-	-	Single-Loop
	Xu et al. [44]	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-4})$	-	Single-Loop
	This work (Theorem 1)	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	Single-Loop

Contributions

- We also provide the theoretical analysis of our algorithm in terms of another convergence criteria (i.e., anϵ-stationary point of φ) for constrained NC-C problems. The result shows that our algorithm improves the best-known result from $\tilde{O}\left(\epsilon^{-3}\right)$ to O(ϵ^{-2}), as shown in Table 2.

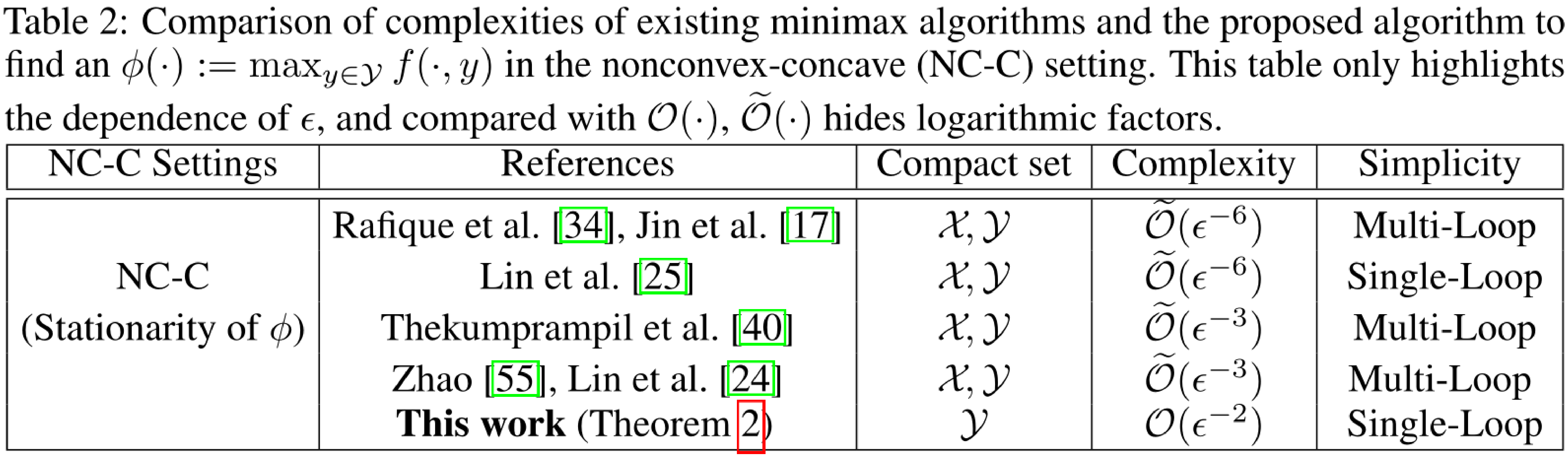


Table 2: Comparison of complexities of existing minimax algorithms and the proposed algorithm to find an $\phi(\cdot) := \max_{y \in \mathcal{Y}} f(\cdot, y)$ in the nonconvex-concave (NC-C) setting. This table only highlights the dependence of ϵ, and compared with $\mathcal{O}(\cdot)$, $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors.

NC-C Settings	References	Compact set	Complexity	Simplicity
NC-C (Stationarity of ϕ)	Rafique et al. [34], Jin et al. [17]	$\mathcal{X}, \mathcal{Y}$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	Multi-Loop
	Lin et al. [25]	$\mathcal{X}, \mathcal{Y}$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	Single-Loop
	Thekumprampil et al. [40]	$\mathcal{X}, \mathcal{Y}$	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	Multi-Loop
	Zhao [55], Lin et al. [24]	$\mathcal{X}, \mathcal{Y}$	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	Multi-Loop
	This work (Theorem 2)	$\mathcal{Y}$	$\mathcal{O}(\epsilon^{-2})$	Single-Loop

Main Ideas of Our Method

- Gradient descent:

$$x_{t+1} = \arg\min_{x\in\mathcal{X}} \{\langle \nabla_x f(x_t, y_t),\, x\rangle + \|x - x_t\|^2/\eta_x\}.$$

- Extra-gradient difference prediction:

$$\text{Prediction point} \rightarrow \boxed{u_{t+1/2} = y_t + \beta[\ \nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})]}.$$

- Gradient ascent correction:

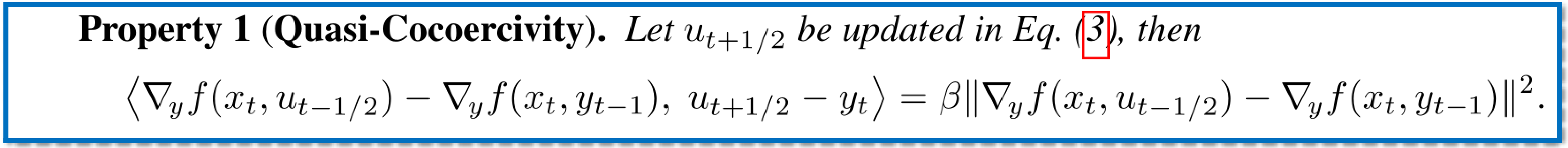
$$u_{t+1} = \arg\max_{u\in\mathcal{Y}} \{\langle \nabla_y f(x_t, u_{t-1/2}),\, u\rangle - \|u - y_t\|^2/\eta_y^t\}.$$

- Momentum acceleration:

$$y_{t+1} = \tau y_t + (1-\tau) u_{t+1}.$$

Advantages of Our Algorithm and Comparison to Related Work

- **Prediction Point:** The monotonicity and co-coercivity properties of gradient operators play a crucial role for conference analysis. However, the important properties do not hold for nonconvex problems. Some researchers developed this research in some special nonconvex settings, such as structured nonconvex and weakly convex, which require a weaker condition such as weakly monotone, pseudo-monotone, and MVI. However, such conditions seriously limit the application scope of Problem. To address this challenge, we design a new prediction point scheme, which can help us obtain a useful quasi-cocoercivity property. As a result, it does not require any monotone or structural assumption. Specifically, we find that we only require a weaker property in our theoretical analysis, that is, the co-coercivity is required at some special points $\{u_{t+1/2}, y_t\}$ ($\langle \nabla_y f(x_t, u_{t+1/2}) - \nabla_y f(x_t, y_t), u_{t+1/2} - y_t \rangle \geq \rho \|\nabla_y f(x_t, u_{t+1/2}) - \nabla_y f(x_t, y_t)\|^2$ with $\rho > 0$). Thus, we develop a decoupling idea to construct the prediction point $u_{t+1/2}$. That is, we use the gradients w.r.t. y at $u_{t-1/2}, y_{t-1}$ instead of those at the points $u_{t+1/2}, y_t$. We can obtain a property (called the quasi-co-coercivity), which plays a key role in our theoretical analysis.

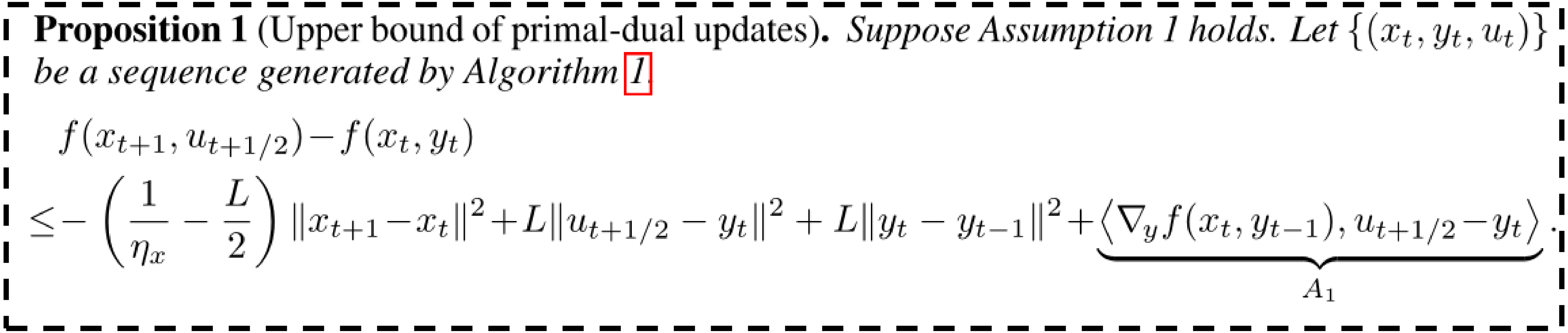


Property 1 (Quasi-Cocoercivity). *Let $u_{t+1/2}$ be updated in Eq. (3), then*

$$\left\langle \nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1}),\ u_{t+1/2} - y_t \right\rangle = \beta \|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2.$$

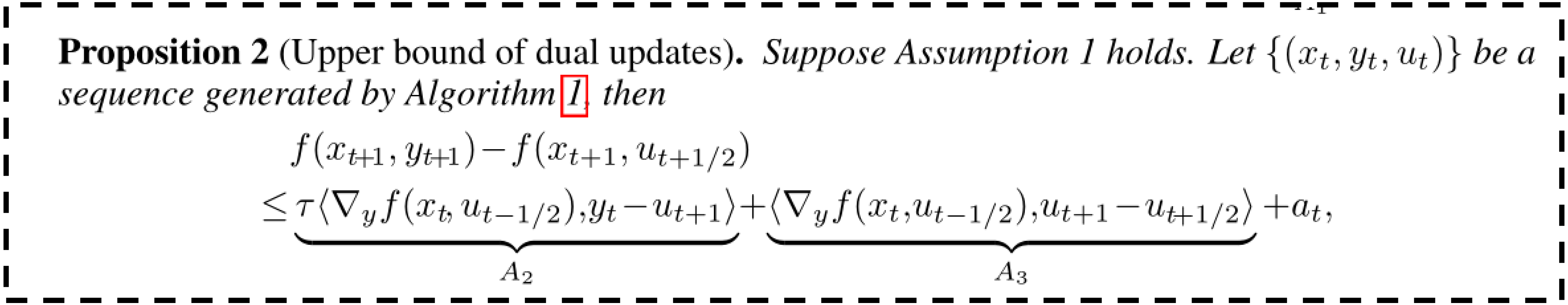
Convergence Guarantees

Core Propositions



Proposition 1 (Upper bound of primal-dual updates). *Suppose Assumption 1 holds. Let* $\{(x_t, y_t, u_t)\}$ *be a sequence generated by Algorithm 1.*

$$f(x_{t+1},u_{t+1/2})-f(x_t,y_t)$$
$$\le -\left(\frac{1}{\eta_x}-\frac{L}{2}\right)\|x_{t+1}-x_t\|^2+L\|u_{t+1/2}-y_t\|^2+L\|y_t-y_{t-1}\|^2+\underbrace{\langle\nabla_y f(x_t,y_{t-1}),u_{t+1/2}-y_t\rangle}_{A_1}.$$



Proposition 2 (Upper bound of dual updates). *Suppose Assumption 1 holds. Let* $\{(x_t, y_t, u_t)\}$ *be a sequence generated by Algorithm 1, then*

$$f(x_{t+1},y_{t+1})-f(x_{t+1},u_{t+1/2})$$
$$\le \underbrace{\tau\langle\nabla_y f(x_t,u_{t-1/2}),y_t-u_{t+1}\rangle}_{A_2}+\underbrace{\langle\nabla_y f(x_t,u_{t-1/2}),u_{t+1}-u_{t+1/2}\rangle}_{A_3}+a_t,$$

Convergence Guarantees

$$G_t := f(x_t, y_t) + 9L\|u_t - y_{t-1}\|^2 + 8L\beta^2\|\nabla_y f(x_{t-1}, u_{t-3/2}) - \nabla_y f(x_{t-1}, y_{t-2})\|^2.$$

Next, we need to prove that our defined potential function can make sufficient decrease at each iteration, i.e., $G_t - G_{t+1} > 0$ as in Lemma 1 below. To prove Lemma 1, we will provide and prove the following upper bounds.

$$G_{t+1} - G_t := \underbrace{f(x_{t+1}, y_{t+1}) - f(x_{t+1}, u_{t+1/2})}_{\text{Proposition 1}} + \underbrace{f(x_{t+1}, u_{t+1/2}) - f(x_t, y_t)}_{\text{Proposition 2}} + \text{other terms}$$

$$= \underbrace{A_1 + A_2 + A_3}_{\text{Quasi-Cocoercivity in Property 1}} + \text{other terms.}$$

Lemma 1 (Descent estimate of G). *Suppose Assumption 1 holds. Let $\{(x_t, y_t, u_t)\}$ be a sequence generated by Algorithm 1, then*

$$\begin{aligned} & G_0 - G_T \\ = & \sum_{t=0}^{T-1} (G_t - G_{t+1}) \\ \geq & \sum_{t=0}^{T-1} \left(\frac{1}{2\beta}\|u_{t+1} - y_t\|^2 + \frac{\beta - 30L\beta^2}{2}\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + \frac{1}{2\eta_x}\|x_{t+1} - x_t\|^2 \right). \end{aligned}$$

Convergence Guarantees

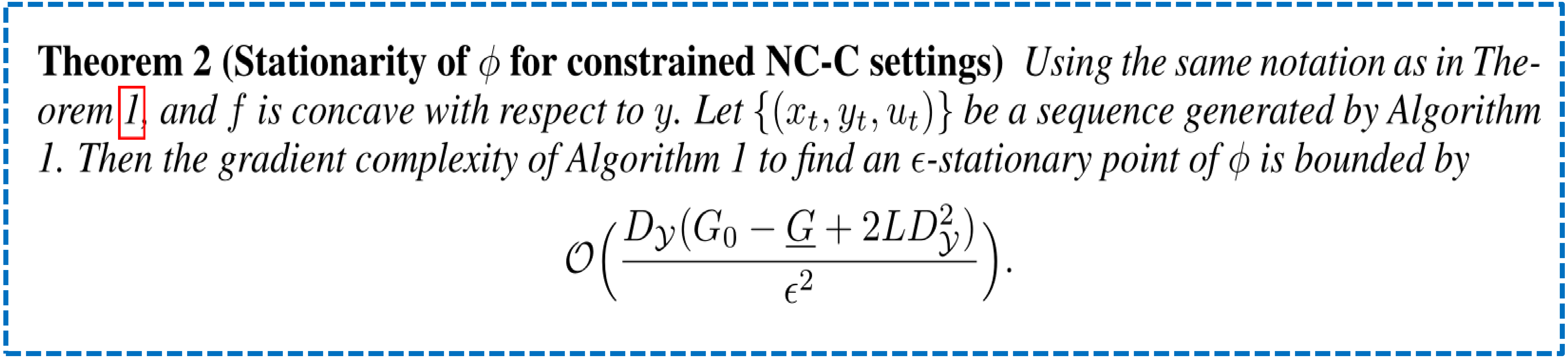
Theorem 1 (Stationarity of f in constrained NC-NC settings) *Suppose Assumptions 1 and 2 hold. Then the complexity of Algorithm 1 to find an ϵ-stationary point of f with $\eta_x \leq \frac{1}{4L}$, $\beta \leq \frac{1}{60L}$ and $\tau \geq 1/2$ is bounded by*

$$\mathcal{O}\Big(\frac{G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2}{\epsilon^2}\Big),$$

where $G_0 := G(x_0, y_0)$, and $\underline{G} := \min_{x \in \mathcal{X}} \phi(x)$.

For constrained NC-NC problems, the gradient complexity of our EGDA algorithm to find an ϵ-stationary point of f is $\mathcal{O}(\epsilon^{-2})$. That is, our EGDA algorithm is first to obtain the gradient complexity in constrained NC-NC setting.

Convergence Guarantees



Theorem 2 (Stationarity of ϕ for constrained NC-C settings) *Using the same notation as in Theorem 1, and f is concave with respect to y. Let $\{(x_t, y_t, u_t)\}$ be a sequence generated by Algorithm 1. Then the gradient complexity of Algorithm 1 to find an ϵ-stationary point of ϕ is bounded by*

$$\mathcal{O}\Big(\frac{D_{\mathcal{Y}}(G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2)}{\epsilon^2}\Big).$$

That is, the proposed algorithm can improve the best-known gradient complexity from $\widetilde{\mathcal{O}}(\epsilon^{-3})$ to $\mathcal{O}(\epsilon^{-2})$.

Experimental Results

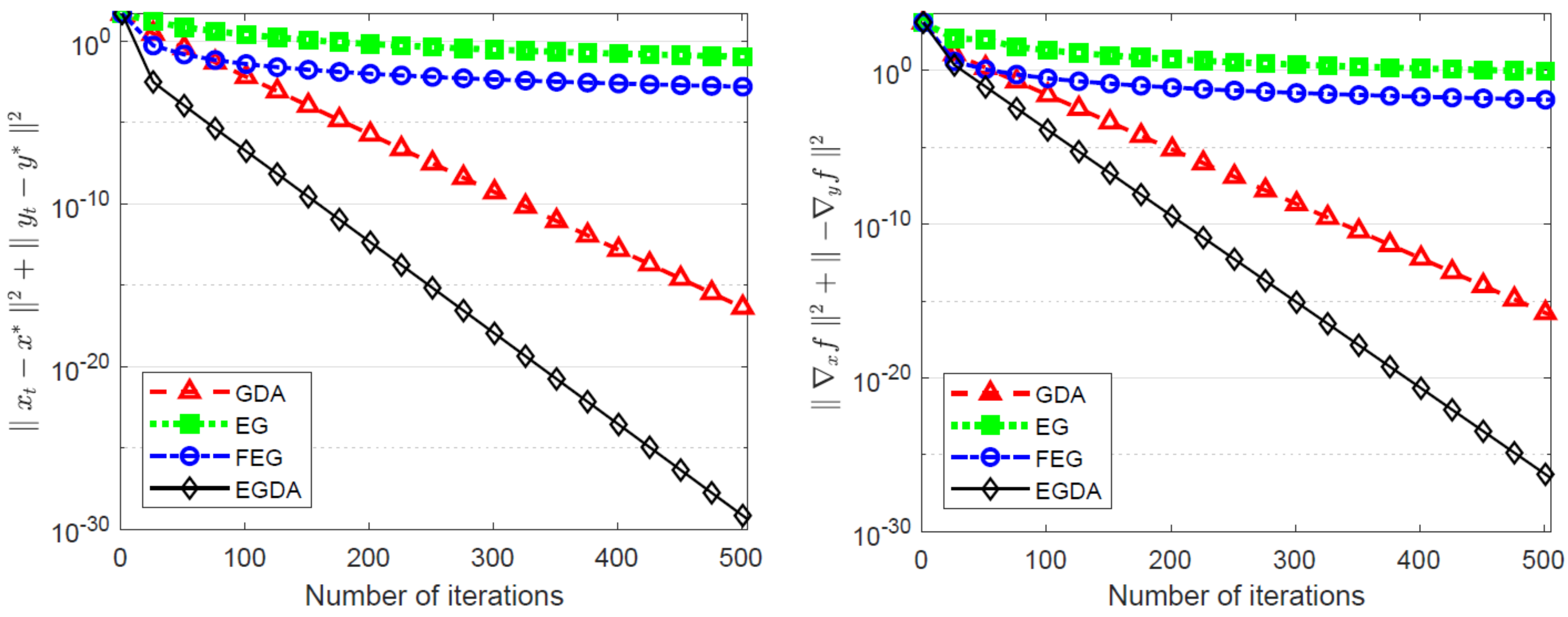


Figure 1: Comparison of all the methods for solving the NC-NC problem, $f(x, y) = x^2 + 3\sin^2 x \sin^2 y - 4y^2 - 10\sin^2 y$. Left: Convergence in terms of $\|x_t - x^*\|^2 + \|y_t - y^*\|^2$, where (x^*, y^*) is the global saddle point; Right: Convergence in terms of $\|\nabla_x f\|^2 + \|\nabla_y f\|^2$.

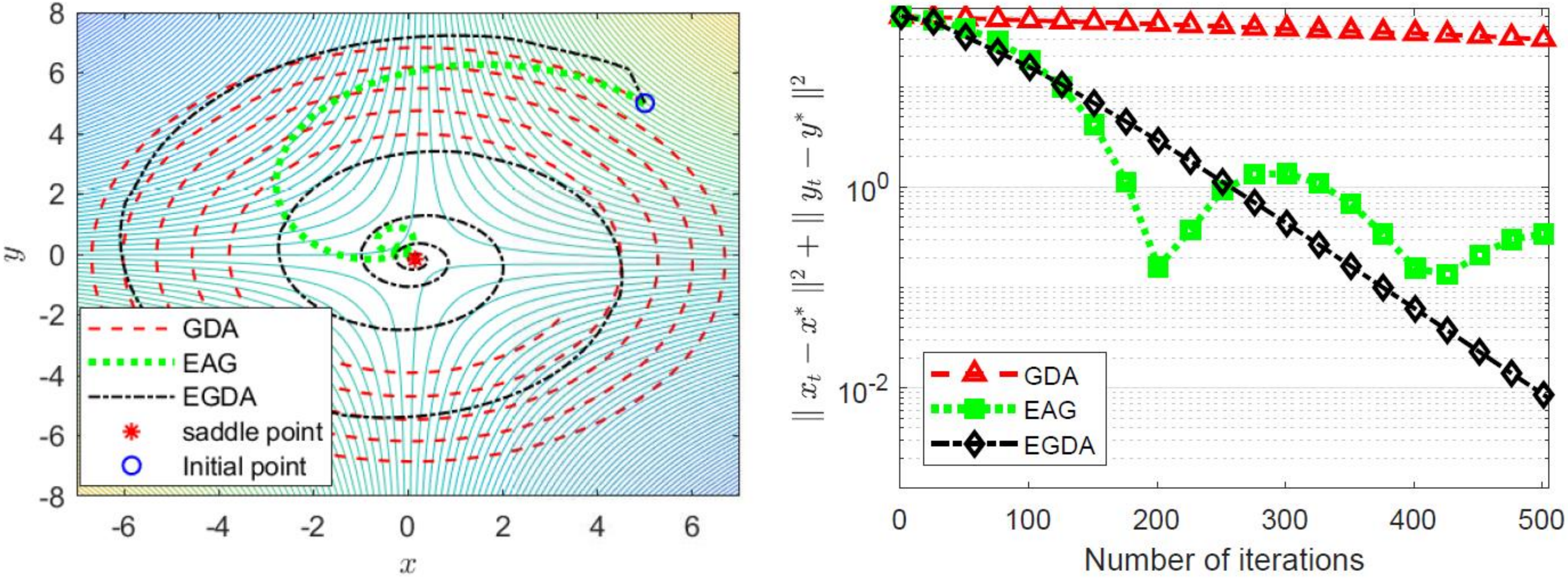


Figure 2: Comparison of all the methods for solving the convex-concave problem, $f(x, y) = \log(1 + e^x) + 3xy - \log(1 + e^y)$. Left: Trajectories of the three algorithms; Right: Convergence in terms of $\|x_t - x^*\|^2 + \|y_t - y^*\|^2$.

Experimental Results

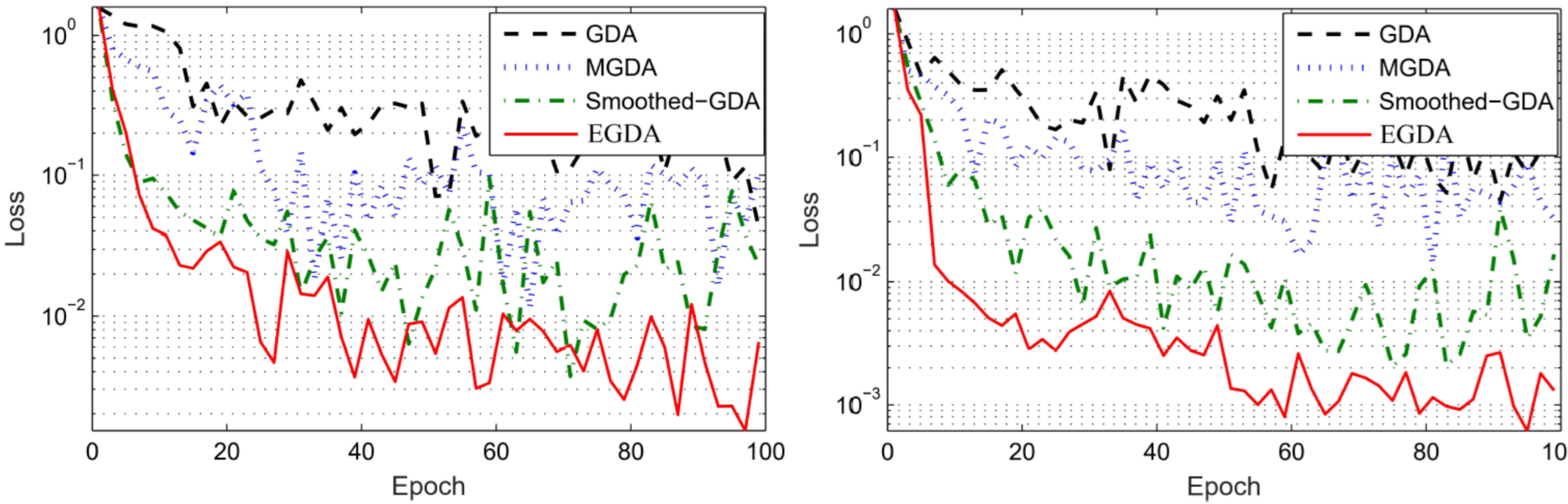


Figure 3: Convergence speed of all the algorithms on Fashion MNIST (left) and MNIST (right).

Thank you!