# Variance Reduced Halpern Iteration for Finite-Sum Monotone Inclusions

Xufeng Cai*[1], Ahmet Alacaoglu*[2], Jelena Diakonikolas[1]

[1]Department of Computer Sciences, University of Wisconsin-Madison

[2] Wisconsin Institute for Discovery, University of Wisconsin–Madison

* Equal contribution.

Computer Sciences
SCHOOL OF COMPUTER, DATA & INFORMATION SCIENCES
UNIVERSITY OF WISCONSIN–MADISON

WISCONSIN INSTITUTE FOR DISCOVERY
AT THE UNIVERSITY OF WISCONSIN-MADISON

## Problem Setting

**Find $\mathbf{u}_* \in \mathbb{R}^d$ such that $\mathbf{0} \in F(\mathbf{u}_*) + G(\mathbf{u}_*)$, where $F = \frac{1}{n}\sum_{i=1}^n F_i$**

- **Assumptions:** For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, the operator $F(\mathbf{u})\colon \mathbb{R}^d \to \mathbb{R}^d$ is
  1. *monotone and $L_F$-Lipschitz*: $\langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle \ge 0$, $\|F(\mathbf{u}) - F(\mathbf{v})\| \le L_F\|\mathbf{u} - \mathbf{v}\|$,
  2. *$L_Q$-Lipschitz in expectation*: $\mathbb{E}_{\xi\sim Q}\|F_\xi(\mathbf{u}) - F_\xi(\mathbf{v})\|^2 \le L_Q^2\|\mathbf{u} - \mathbf{v}\|^2$, given an oracle $F_\xi$ and distribution such that $\mathbb{E}[F_\xi(\mathbf{u})] = F(\mathbf{u})$;
  3. *$1/L$-cocoercive on average*: $\langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle \ge \frac{1}{nL}\sum_{i=1}^n \|F_i(\mathbf{u}) - F_i(\mathbf{v})\|^2$;

  and the operator $G(\mathbf{u})\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is possibly multi-valued and *maximally monotone* with *access to the resolvent* $J_{\eta G}$ of $\eta G$ for $\eta > 0$ (generalizing the proximal operator).
- **Applications:** constrained finite-sum minimization, variational inequality (VI) problems, robust machine learning, adversarial training, multi-agent RL.
- **Optimality measure:**
$$\mathrm{Res}_{F+G}(\mathbf{u}) = \|F(\mathbf{u}) + g(\mathbf{u})\|,$$
for some $g(\mathbf{u}) \in G(\mathbf{u})$ and hence $\mathrm{dist}(F(\mathbf{u}) + G(\mathbf{u}), \mathbf{0}) = \min_{g(\mathbf{u})\in G(\mathbf{u})} \|F(\mathbf{u}) + g(\mathbf{u})\| \le \mathrm{Res}_{F+G}(\mathbf{u})$.
  - computable in most cases as the algorithms have access to $F(\mathbf{u}) + g(\mathbf{u})$,
  - implies other optimality measures (such as duality gap for VI problems),
  - meaningful for some classes of structured non-monotone operators.
- **Oracle complexity:** the number of calls to to make an optimality measure small (the number of calls to the resolvent if of the same order).

## Contributions

★The first $\tilde{\mathcal{O}}(n + \sqrt{n}L\varepsilon^{-1})$ variance-reduced complexity result for the **residual guarantee** when $F$ is either average $1/L$-cocoercive or monotone and $L$-Lipschitz in expectation, that could lead to a $\sqrt{n}$ improvement compared to the methods without variance reduction. This guarantee is also on the **last-iterate.**

Table: Comparison of our results with state of the art in the monotone Lipschitz settings.

| Reference | Complexity for $\mathrm{Res}_{F+G}$ | Complexity for Gap | Assumption | High Probability Result |
|---|---|---|---|---|
| Kovalev & Gasnikov (2022) | $\mathcal{O}(nL_F\varepsilon^{-1})$ | $\mathcal{O}(nL_F\varepsilon^{-1})$ | Assumption 1 | N/A |
| Nemirovski (2004) | $\mathcal{O}(nL_F^2\varepsilon^{-2})$ | $\mathcal{O}(nL_F\varepsilon^{-1})$ | Assumption 1 | N/A |
| Cai et al. (2022a) | $\mathcal{O}((\sigma^2L + L^3)\varepsilon^{-3})$ | $\mathcal{O}((\sigma^2L + L^3)\varepsilon^{-3})$ | Assumption 1, 2, $G \equiv \mathbf{0}$ $\mathbb{E}_i\|F_i(\mathbf{x}) - F(\mathbf{x})\|^2 \le \sigma^2$ | – |
| Luo et al. (2021) | $\tilde{\mathcal{O}}(\sigma^2\varepsilon^{-2} + L_F\varepsilon^{-1})$ | $\tilde{\mathcal{O}}(\sigma^2\varepsilon^{-2} + L_F\varepsilon^{-1})$ | Assumption 1, $G \equiv \mathbf{0}$ $F = \binom{\nabla_{\mathbf{x}}\Phi(\mathbf{x},\mathbf{y})}{-\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})}$ $\mathbb{E}_i\|F_i(\mathbf{x}) - F(\mathbf{x})\|^2 \le \sigma^2$ | – |
| Carmon et al. (2019) | – | $\tilde{\mathcal{O}}(n + \sqrt{n}L\varepsilon^{-1})$ | Assumption 1, 2 bounded domain (cf. Sec 5.4 in (Carmon et al., 2019)) | – |
| Palaniappan & Bach (2016) | – | $\tilde{\mathcal{O}}(n + \sqrt{n}L\varepsilon^{-1})$ | Assumption 1, 2 bounded domain (cf. (C) in Sec. 2 in (Palaniappan & Bach, 2016)) | – |
| Alacaoglu & Malitsky (2022) | – | $\mathcal{O}(n + \sqrt{n}L\varepsilon^{-1})$ | Assumption 1, 2 (cf. Assumption 1(iv) in (Alacaoglu & Malitsky, 2022)) | – |
| **[Our results, Theorem 4.2]** | $\tilde{\mathcal{O}}(n + \sqrt{n}L\varepsilon^{-1})$ | $\tilde{\mathcal{O}}(n + \sqrt{n}L\varepsilon^{-1})$ | Assumption 1, 2 | ✓ |

## Cocoercive Case

**Recipe:** stochastic constrained Halpern Iteration + PAGE (Li et al., 2021)

**Input:** $\mathbf{u}_0 \in \mathbb{R}^d$, step size $\eta = \frac{1}{4L}$, batch size $b = \lceil\sqrt{n}\rceil$, $\lambda_1 = \frac{2}{5}$

$\mathbf{u}_1 = J_{\frac{\eta}{2\lambda_1}G}\left(\mathbf{u}_0 - \frac{\eta}{2\lambda_1}F(\mathbf{u}_0)\right)$, $\quad \widetilde{F}(\mathbf{u}_1) = F(\mathbf{u}_1)$

**for** $k = 1, 2, \ldots$ **do**

$$\lambda_k = \frac{2}{k+4}, \quad p_{k+1} = \begin{cases} \frac{4}{k+5} & \forall k \le \sqrt{n} \\ \frac{4}{\sqrt{n}+5} & \forall k \ge \sqrt{n}\end{cases}$$

$$\mathbf{u}_{k+1} = J_{\eta G}(\lambda_k\mathbf{u}_0 + (1-\lambda_k)\mathbf{u}_k - \eta\widetilde{F}(\mathbf{u}_k))$$

Sample $\mathcal{S}_{k+1} \subseteq \{1, \ldots, n\}$ without replacement and uniformly at random with $|\mathcal{S}_{k+1}| = b$

$$\widetilde{F}(\mathbf{u}_{k+1}) = \begin{cases} F(\mathbf{u}_{k+1}) & \text{w.p. } p_{k+1}, \\ \widetilde{F}(\mathbf{u}_k) + \frac{1}{b}\sum_{i\in\mathcal{S}_{k+1}}(F_i(\mathbf{u}_{k+1}) - F_i(\mathbf{u}_k)) & \text{w.p. } 1 - p_{k+1}.\end{cases}$$

### Convergence Analysis

Potential function:

$$\mathscr{C}_k = \frac{\eta}{2\lambda_k}\|F(\mathbf{u}_k) + \mathbf{g}_k\|^2 + \langle F(\mathbf{u}_k) + \mathbf{g}_k, \mathbf{u}_k - \mathbf{u}_0\rangle + c_k\|F(\mathbf{u}_k) - \tilde{F}(\mathbf{u}_k)\|^2,$$

where $\mathbf{g}_k = \frac{1}{\eta}(\lambda_{k-1}\mathbf{u}_0 + (1-\lambda_{k-1})\mathbf{u}_{k-1} - \eta\tilde{F}(\mathbf{u}_{k-1}) - \mathbf{u}_k) \in G(\mathbf{u}_k)$ and $c_k = \frac{(\sqrt{n}+2)(k+4)}{4L}$.

- Use average cocoercivity of $F$ and recursive variance bound of PAGE to show:
$$\mathbb{E}[\mathscr{C}_{k+1}] \le (1-\lambda_k)\mathbb{E}[\mathscr{C}_k].$$

### Improved Complexity

Under Assumptions 1 and 3:

$$\mathbb{E}[\mathrm{Res}_{F+G}(\mathbf{u}_k)] \le \left(\mathbb{E}[\mathrm{Res}^2_{F+G}(\mathbf{u}_k)]\right)^{1/2} \le \frac{16L\|\mathbf{u}_0 - \mathbf{u}_*\|}{k+4},$$

$\Rightarrow \tilde{\mathcal{O}}(n + \sqrt{n}L\varepsilon^{-1})$ stochastic oracle complexity.

- Up to $\sqrt{n}$ improvement compared to complexity results $\tilde{\mathcal{O}}(nL_F\varepsilon^{-1})$ of deterministic algorithms (Diakonikolas, 2020).
- Improve in the regime $\varepsilon = o(1/\sqrt{n})$ compared to complexity results for infinite-sum stochastic settings in Cai et al. (2022a); Chen & Luo (2022).
- Provide the best-known guarantees (among direct approaches) with a single-loop algorithm for finite-sum minimization.

## Monotone and Lipschitz Case

### Recipe

- For any $\eta > 0$, finding a point $\mathbf{u}$ with $\|P^\eta(\mathbf{u})\| \le \eta\varepsilon$ is sufficient to guarantee $\mathrm{Res}_{F+G}(J_{\eta(F+G)}(\mathbf{u})) \le \varepsilon$, where $P^\eta(\mathbf{u}) := \mathbf{u} - J_{\eta(F+G)}(\mathbf{u})$ is 1/2-cocoercive. We can replace $J_{\eta(F+G)}(\mathbf{u})$ in the guarantee by a computable output.
- (Stochastic) inexact Halpern iteration converges at an optimal rate with appropriate level of inexactness:
$$\mathbf{u}_{k+1} = \lambda_k\mathbf{u}_0 + (1-\lambda)\tilde{J}_{\eta(F+G)}(\mathbf{u}_k) = \lambda_k\mathbf{u}_0 + (1-\lambda)(\mathbf{u}_k - P^\eta(\mathbf{u}_k)) - (1-\lambda)\mathbf{e}_k,$$
where $\tilde{J}_{\eta(F+G)}$ is an approximation of $J_{\eta(F+G)}$ and $\mathbf{e}_k = J_{\eta(F+G)}(\mathbf{u}_k) - \tilde{J}_{\eta(F+G)}(\mathbf{u}_k)$.
- Approximating $J_{\eta(F+G)}$ corresponds to solving the strongly monotone and expected Lipschitz MI with finite-sum structure, which can be computed fast (Alacaoglu & Malitsky, 2022).

## Monotone and Lipschitz Case (Cont.)

**Recipe:** inexact Halpern Iteration + VR-FoRB (subsolver)

**Input:** $\mathbf{u}_0 \in \mathbb{R}^d$, $L = L_Q$ with the distribution $Q = \{q_i\}_{i=1}^n, n, \eta = \frac{\sqrt{n}}{L}$

**for** $k = 0, 1, 2, \ldots$ **do**

$$\lambda_k = \frac{1}{k+2},\ M_k = \lceil 56(n + \sqrt{n})\log(2k+4)\rceil$$

$$\tilde{J}_{\eta(F+G)}(\mathbf{u}_k) = \mathtt{VR\text{-}FoRB}(\mathbf{u}_k, M_k, \mathrm{Id} + \eta(F+G) - \mathbf{u}_k, Q)$$

$$\mathbf{u}_{k+1} = \lambda_k\mathbf{u}_0 + (1-\lambda_k)\tilde{J}_{\eta(F+G)}(\mathbf{u}_k)$$

**Input:** $\mathbf{v}_0 = \mathbf{w}_0 = \mathbf{w}_{-1} = \mathbf{u}$, $p = \frac{1}{n}$, $\alpha = 1 - p$, $\tau = \frac{\sqrt{p(1-p)}}{2L_A}$, distribution $Q = \{q_i\}_{i=1}^n$

**for** $k = 0, 1, \ldots, M-1$ **do**

$\hat{\mathbf{v}}_k = \alpha\mathbf{v}_k + (1-\alpha)\mathbf{w}_k$

Sample $i \in \{1, \ldots, n\}$ according to $Q$

$\mathbf{v}_{k+1} = J_{\tau B}(\hat{\mathbf{v}}_k - \tau[A(\mathbf{w}_k) - (nq_i)^{-1}A_i(\mathbf{w}_{k-1}) + (nq_i)^{-1}A_i(\mathbf{v}_k)])$

$$\mathbf{w}_{k+1} = \begin{cases}\mathbf{v}_{k+1} & \text{w.p. } p \\ \mathbf{w}_k & \text{w.p. } 1-p\end{cases}$$

### Convergence Analysis

- **Resolvent approximation (VR-FoRB):** Let $A$ be monotone and $L_A$-Lipschitz in expectation with $A = \sum_{i=1}^n A_i$. Let $B$ be maximally monotone, and $A + B$ be $\mu$-strongly monotone. Given $\bar{\varepsilon} > 0$, VR-FoRB returns $\mathbf{v}_M$ such that $\mathbb{E}[\|\mathbf{v}_M - \mathbf{v}_*\|^2] \le \bar{\varepsilon}^2$ in
$$O((n + \sqrt{n}L_A/\mu)\log(\|\mathbf{v}_0 - \mathbf{v}_*\|/\bar{\varepsilon}))$$
iterations and oracle queries.
- **Inexact Halpern iteration:** Let $F$ be $L$-Lipschitz in expectation, then we have
  - $\mathbb{E}_k[\|\mathbf{e}_k\|^2] \le \|P^\eta(\mathbf{u}_k)\|/(k+2)^8$,
  - $\mathbb{E}[\|P^\eta(\mathbf{u}_k)\|] \le (\mathbb{E}[\|P^\eta(\mathbf{u}_k)\|^2]) \le 7\|\mathbf{u}_0 - \mathbf{u}_*\|/k$,

### Improved Complexity

Under Assumptions 1 and 2, given accuracy $\varepsilon > 0$, to return a point $\mathbf{u}_k$ such that $\mathbb{E}[\|P^\eta(\mathbf{u}_k)\|] \le \eta\varepsilon$ with $\eta = \sqrt{n}/L$, the stochastic oracle complexity is

$$\tilde{\mathcal{O}}(n + \sqrt{n}L\varepsilon^{-1}).$$

- Leads to high probability guarantees using a confidence boosting mechanism.
- Up to $\sqrt{n}$ improvement compared to complexity results $\tilde{\mathcal{O}}(nL_F\varepsilon^{-1})$ of deterministic algorithms for the residual (Diakonikolas, 2020; Yoon & Ryu, 2021).
- Implies prior gap guarantee results (Alacaoglu & Malitsky, 2022; Carmon et al., 2019) which are suboptimal for the residual. The implication also ensures the near-optimality of our results.
- Extends to $\rho$-cohypomonotone settings with $G \equiv \mathbf{0}$ for any $\eta > 0$ such that $\rho < \min\{\eta/2, 1/\eta L_F^2\}$.

## Numerical Results

Compare with extragradient (**EG**) (Korpelevich, 1977), constrained anchored extragradient (**EAG**) (Cai et al., 2022b), and variance-reduced extragradient (**VR-EG**) (Alacaoglu & Malitsky, 2022). Use uniform sampling for all algorithms and tune the step size for each method individually.

- **Matrix game:** $\min_{\mathbf{x}\in\Delta^{m_1}}\max_{\mathbf{y}\in\Delta^{m_2}}\langle\mathbf{A}\mathbf{x}, \mathbf{y}\rangle$ with simplex constraints ($m_1 = m_2 = 500$) and the policeman and burglar matrix (Nemirovski, 2013).
- **Quadratic program:** $\min_{\mathbf{x}\in\mathbb{R}^{m_1}}\max_{\mathbf{y}\in\mathbb{R}^{m_2}}\frac{1}{2}\mathbf{x}^\top\mathbf{H}\mathbf{x} - \mathbf{h}^\top\mathbf{x} - \langle\mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle$ with $m_1 = m_2 = 200$ and the difficult instance for establishing lower bounds for min-max optimization (Ouyang & Xu, 2021).

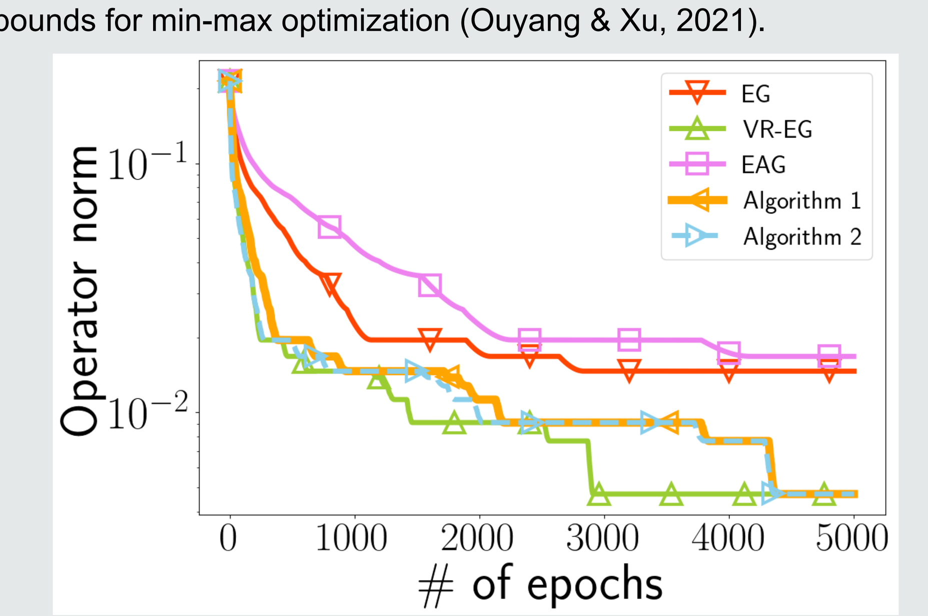


Fig. 1: Matrix game

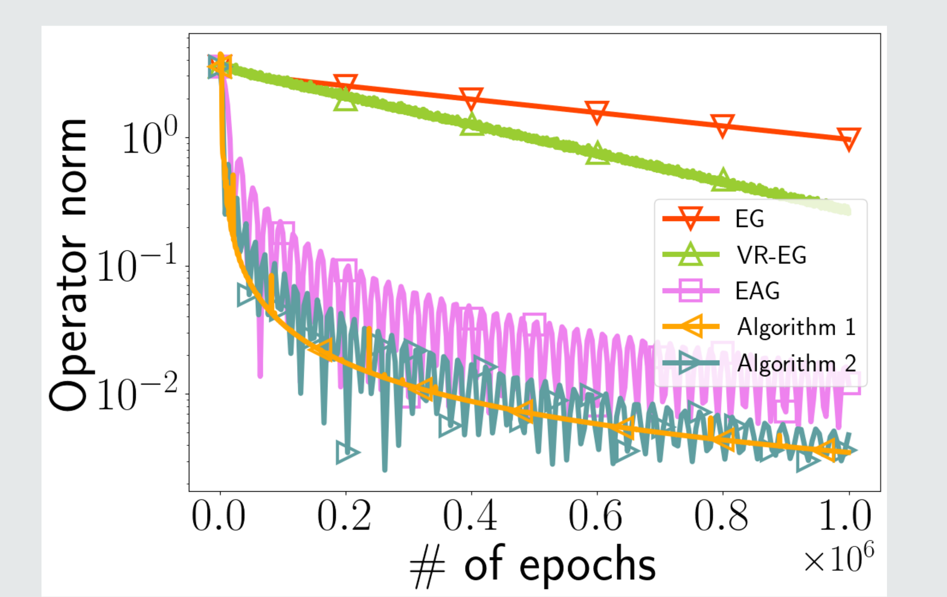


Fig. 2: Quadratic program

References

A. Alacaoglu and Y. Malitsky. Stochastic variance reduction for variational inequality methods. In Proc. COLT'22, 2022.
X. Cai, C. Song, C. Guzman, and J. Diakonikolas. Stochastic Halpern iteration with variance reduction for stochastic monotone inclusions. In Proc. NeurIPS'22, 2022a.
Y. Cai, A. Oikonomou, and W. Zheng. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. arXiv preprint arXiv:2206.05248, 2022b.
Y. Carmon, Y. Jin, A. Sidford, and K. Tian. Variance reduction for matrix games. In Proc. NeurIPS'19, 2019.
L. Chen and L. Luo. Near-optimal algorithms for making the gradient small in stochastic minimax optimization. arXiv preprint arXiv:2208.05925, 2022.
J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In Proc. COLT'20, 2020.
GM Korpelevich. Extragradient method for finding saddle points and other problems. Matekon, 13 (4):35–49, 1977.
D. Kovalev and A. Gasnikov. The first optimal algorithm for smooth and stronglyconvex-strongly-concave minimax optimization. In Proc. NeurIPS'22, 2022.
A. Nemirovski. Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM Journal on Optimization, 15(1):229–251, 2004.
A. Nemirovski. Mini-course on convex programming algorithms. Lecture Notes, 2013.
Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convexconcave bilinear saddle-point problems. Mathematical Programming, 185(1-2):1–35, 2021.
B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In Proc. NeurIPS'16, 2016.
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 Proc. ICML'21, 2021.

Acknowledgements

This research was supported in part by the NSF grant 2023239, the NSF grant 2007757, the NSF grant 2224213, the AFOSR award FA9550-21-1-0084, the Office of Naval Research under contract number N00014-22-1-2348.