When ExtraGradient meets VR: bridjing two giants to boost VIs

Y. Kabikov, G. Chirkov
D. Medyakov, G. Molodtsov
MIPT, 2025

INTRODUCTION

variational inequality:

$$\forall z \in \mathcal{Z} \hookrightarrow \langle F(z^*), z - z^* \rangle \geqslant 0$$

finite-sum approach:

$$F(z) = \frac{1}{n} \sum_{i=1}^{n} F_i(z) \qquad \min_{x \in \mathbb{R}^d} \left[F(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\xi_i \sim \mathcal{D}_i} \left[F_{\xi_i}(x) \right] \right]$$

Stochastic EG and variance problem:

$$z^{t+\frac{1}{2}} = z^t - \gamma F_{i_t}(z^t),$$
$$z^{t+1} = z^t - \gamma F_{j_t}(z^{t+\frac{1}{2}})$$

$$\mathbf{E}\{\|F_{i_t}(z^t) - F(z^t)\|^2\} \not\to 0, \quad z^t \to z^*$$

Related works

Felix E Browder. Nonexpansive nonlinear operators in a banach space. Proceedings of the National Academy of Sciences, 54(4):1041–1044, 1965.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror prox algorithm. Stochastic Systems, 1(1):17–58, 2011.

Aaron Defazio, Justin Domke, et al. Finito: A faster, per mutable incremental gradient method for big data prob lems. In International Conference on Machine Learning, pages 1125–1133. PMLR, 2014b.

$$
\begin{aligned}
z^{t+1/2} &= z^t - \gamma G^{t-1} \\
G^t &= F_{i_t}(z^{t+1/2}) - y_{i_t}^t + \tfrac{1}{n}\textstyle\sum_{j=1}^n y_j^t \\
y_i^{t+1} &= F_{i_t}(z^{t+1/2}) \text{ for } i = i_t \\
z^{t+1} &= z^t - \gamma G^t
\end{aligned}
$$

ExtraSAGA method

(final)

$$
\begin{aligned}
z^{t+1/2} &= z^t - \tfrac{1}{n}\textstyle\sum_{j=1}^n y_j^t \\
G^t &= F_{i_t}(z^{t+1/2}) - y_{i_t}^t + \tfrac{1}{n}\textstyle\sum_{j=1}^n y_j^t \\
y_i^{t+1} &= F_{i_t}(z^{t+1/2}) \text{ for } i = i_t \\
z^{t+1} &= z^t - \gamma G^t
\end{aligned}
$$

ExtraSAGA method

SETUP

-Lipschitzness

$$\mathbb{E}\left[\|F_\xi(z_1) - F_\xi(z_2)\|^2\right] \leqslant L^2\|z_1 - z_2\|^2$$

-Strong monotonicity

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geqslant \mu\|z_1 - z_2\|^2$$

CONVERGENCE

$$\mathbb{E}\left[\sigma_{t+1}^2 | z^{t+\frac{1}{2}}\right] \quad = \quad \mathbb{E}\left[\frac{1}{n}\sum_j \left\|y_j^{t+1} - F_j(z^*)\right\|^2\right]$$

$$\gamma \leq \frac{1}{6L}$$

linear rate

$$\mathbb{E}\left[\left\|z^{t+1} - z^*\right\| + M\sigma_{t+1}^2 | z^{t+\frac{1}{2}}\right] \quad \leq$$

$$\leq \quad \left(1 - \frac{\gamma\mu}{2}\right)\left\|z^t - z^*\right\|^2 + \left(1 - \frac{1}{2n}\right)\sigma_t^2$$

CONVERGENCE

Co-coercivity

Stochastic Gradient Descent-Ascent and Consensus Optimization for Smooth Games: Convergence Analysis under Expected Co-coercivity.
N. Loizou et al., 2021

Definition 3.1 (Co-coercivity / Co-coercive *around* w^*). We say that an operator ξ is ℓ–co-coercive if there exist $\ell > 0$ such that, $\|\xi(x) - \xi(y)\|^2 \leq \ell \langle \xi(x) - \xi(y), x - y \rangle \quad \forall x, y \in \mathbb{R}^d$.
If there exist $w^* \in \mathbb{R}^d$ and $\ell > 0$ such that $\|\xi(x) - \xi(w^*)\|^2 \leq \ell \langle \xi(x) - \xi(w^*), x - w^* \rangle \quad \forall x \in \mathbb{R}^d$. then we say that the operator ξ is ℓ–co-coercive *around* w^*. Note that in the last definition the point w^* is not necessarily a point where $\xi(w^*) = 0$.

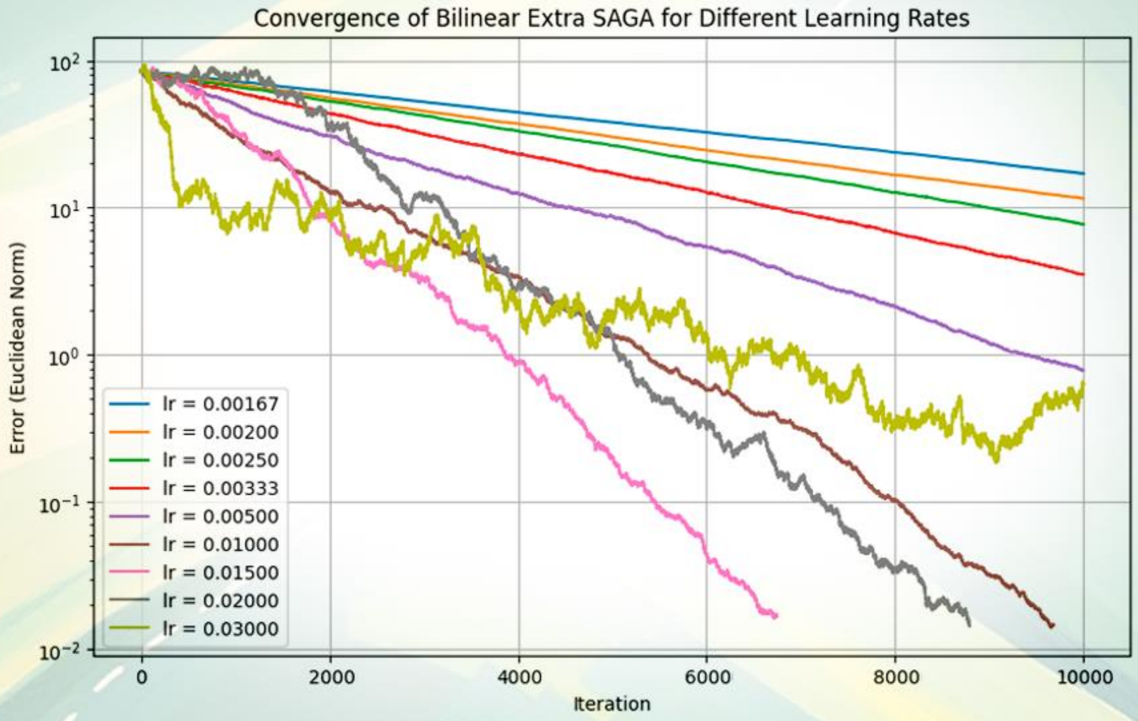
EXPERIMENTS

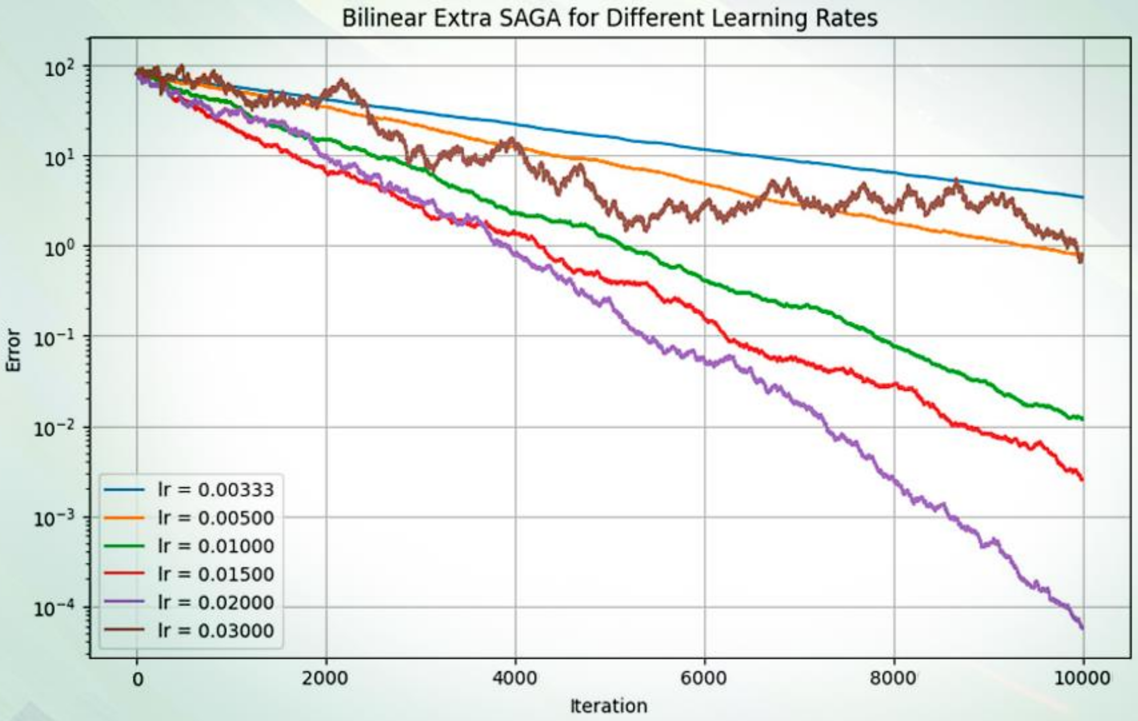
Experiments:
Bilinear convex-concave min-max (saddle points)

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} [f(x, y)]$$

$$f(x, y) = (x - b_x)^T A(y - b_y) + \frac{\lambda}{2}\|x - b_x\|^2 - \frac{\lambda}{2}\|y - b_y\|^2$$

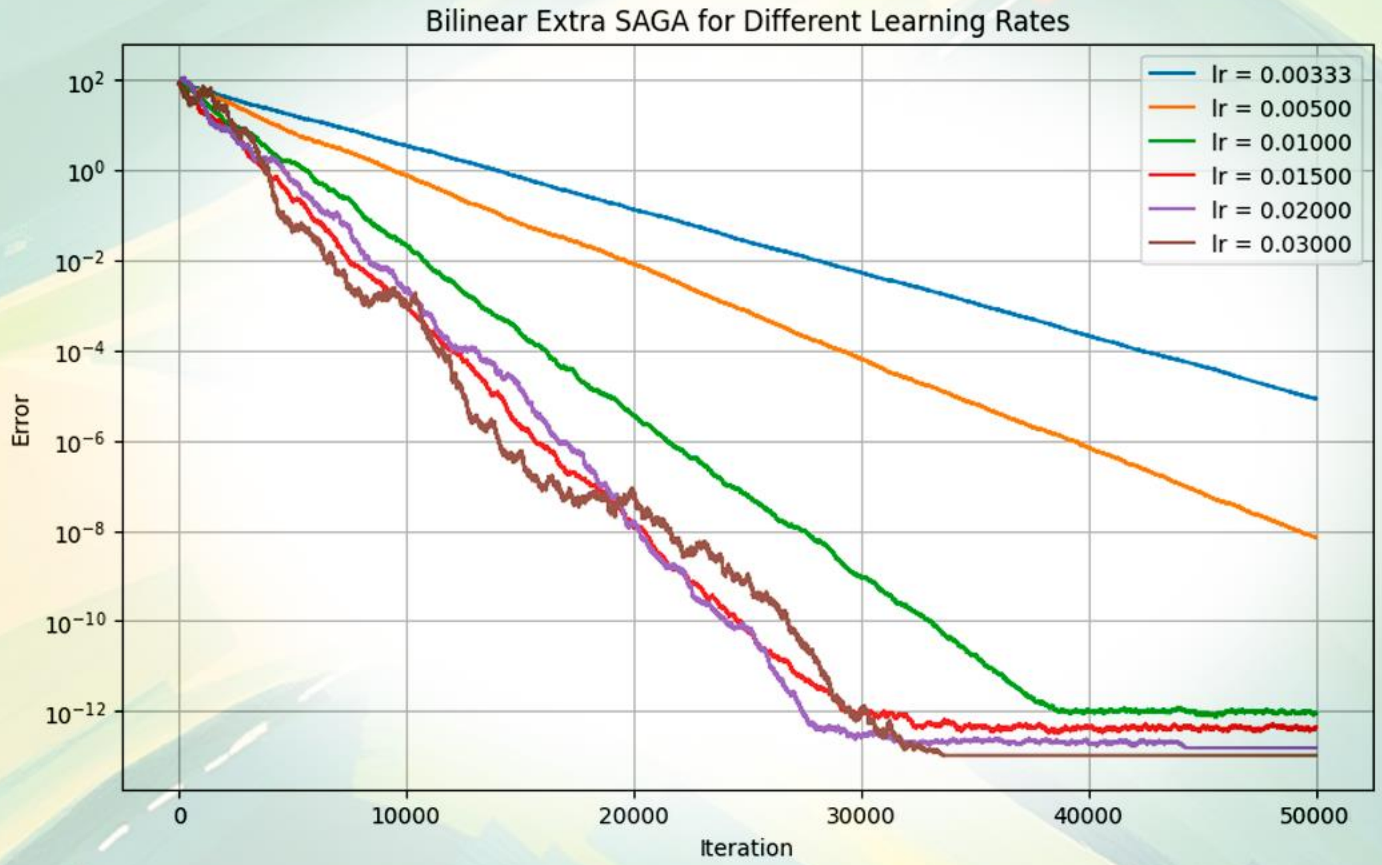
Experiments:
Bilinear convex-concave min-max (saddle points)





L = 100, mu = 0.1, n = 100, d = 100

Experiments:
Bilinear convex-concave min-max (saddle points)



L = 100, mu = 0.1, n = 100, d = 100

Denoising

$$\{(ih, jh) : 1 \le i \le M, 1 \le j \le N\}$$

picture = cartesian grid

$$(\nabla u)_{i,j} = \begin{pmatrix} (\nabla u)^1_{i,j} \\ (\nabla u)^2_{i,j} \end{pmatrix} \quad (\nabla u)^1_{i,j} = \begin{cases} \dfrac{u_{i+1,j} - u_{i,j}}{h} & \text{if } i < M \\ 0 & \text{if } i = M \end{cases}, \quad (\nabla u)^2_{i,j} = \begin{cases} \dfrac{u_{i,j+1} - u_{i,j}}{h} & \text{if } j < N \\ 0 & \text{if } j = N \end{cases}$$

$$\min_x \int_\Omega |Du| + \frac{\lambda}{2} \|u - g\|_2^2$$

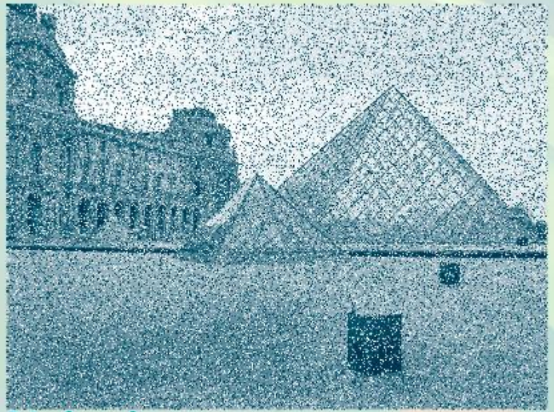
the ROF model

$$\min_{u \in X} \|\nabla u\|_1 + \lambda \|u - g\|_1$$

TV-L 1 model

Denoising

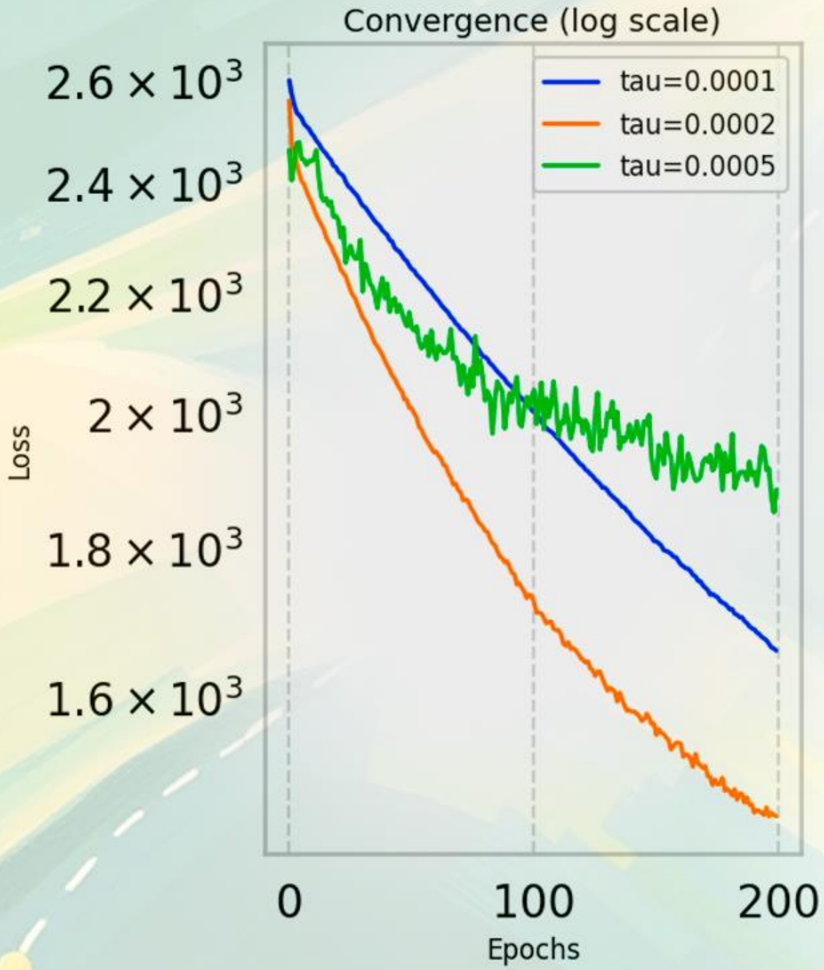

Clean image


Noisy image


ROF (lambda = 8)


TV-L 1 (lambda = 1.5)

DENOISING



Reconstructed (tau=0.0002)

Reconstructed (tau=0.0005)

Results

Convergence is proved for strongly monotonic and monotonic cases.

Under the assumption of co-coercivity, linear convergence is obtained

Various modifications have been investigated

Experiments have been conducted (minmax, denoising, models learning)

THANKS FOR YOUR ATTENTION!