Stochastic Gradient Descent-Ascent: Unified Theory and New Efficient Methods

Aleksandr Beznosikov[1] Eduard Gorbunov [2] Hugo Berard [3] Nicolas Loizou [4]

[1]Innopolis University [2]Mohamed bin Zayed University of Artificial Intelligence [3]Mila, Université de Montréal [4]Johns Hopkins University

The problem

Variational inequality problem (VIPs) :

$$\text{Find } x^* \in \mathbb{R}^d \text{ such that}$$
$$\langle F(x^*), x - x^*\rangle + R(x) - R(x^*) \geq 0 \quad \forall x \in \mathbb{R}^d.$$

- $F : \mathbb{R}^d \to \mathbb{R}^d$ is some operator,
- $R : \mathbb{R}^d \to \mathbb{R}$ is a regularization term.

Examples

- Minimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) + R(x),$$

for which $F(x) := \nabla f(x)$.

- Saddle point problem:

$$\min_{x_1\in\mathbb{R}^{d_1}} \max_{x_2\in\mathbb{R}^{d_2}} f(x_1,x_2) + R_1(x_1) - R_2(x_2),$$

for which $F(x) := (\nabla_{x_1} f(x_1,x_2), -\nabla_{x_2} f(x_1,x_2))$ and $R(x) := R_1(x_1) + R_2(x_2)$.

The method

SGDA (Stochastic Gradient Descent-Ascent):

$$x^{k+1} = \mathrm{prox}_{\gamma_k R}(x^k - \gamma_k g^k),$$

where g^k is an unbiased estimator of $F(x^k)$, $\gamma_k > 0$ is a stepsize at iteration k, and $\mathrm{prox}_{\gamma R}(x) := \arg\min_{y\in\mathbb{R}^d}\{R(y) + \|y-x\|^2/2\gamma\}$ is a proximal operator defined for any $\gamma > 0$ and $x \in \mathbb{R}^d$.

Setting

- Operator F is *μ-quasi-strongly monotone* and *ℓ-star-cocoercive*: there exist constants $\mu \geq 0$ and $\ell > 0$ such that for all $x \in \mathbb{R}^d$

$$\langle F(x) - F(x^*), x - x^*\rangle \geq \mu\|x - x^*\|^2,$$
$$\|F(x) - F(x^*)\|^2 \leq \ell\langle F(x) - F(x^*), x - x^*\rangle.$$

- g^k is an unbiased estimator of $F(x^k)$: $\mathbb{E}_k\left[g^k\right] = F(x^k)$, and

$$\mathbb{E}_k\left[\|g^k - g^{*,k}\|^2\right] \leq 2A\langle F(x^k) - g^{*,k}, x^k - x^{*,k}\rangle + B\sigma_k^2 + D_1,$$
$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq 2C\langle F(x^k) - g^{*,k}, x^k - x^{*,k}\rangle + (1-\rho)\sigma_k^2 + D_2,$$

where $x^{*,k} = \mathrm{proj}_{X^*}(x^k)$, $g^{*,k} = F(x^{*,k})$, $A, B, C, D_1, D_1 \geq 0$ and $\{\sigma_k\}_{k\geq 0}$ is a sequence of (possibly random) non-negative variables.

Main Contributions

◇ **Unified analysis of SGDA:** We propose a general assumption on the stochastic estimates and the VIP and show that several variants of SGDA satisfy this assumption. We derive general convergence results for (i) quasi-strongly monotone, (ii) monotone star-cocoercive, and (iii) cocoercive problems.

◇ **Extensions of known methods and analysis with sharp rates:** As a by-product of the generality of our theoretical framework, we derive new results for the proximal extensions of several known methods (e.g., proximal SGDA-AS, proximal SGDA with coordinate randomization).
For the known methods fitting our framework our general theorems either recover the best rates known for these methods (SGDA-AS) or tighten them (SGDA-SAGA, Coord. SGDA).

◇ **New methods:** The flexibility of our approach allows us to develop and analyze several new variants of SGDA.
For example, a new variance-reduced method (L-SVRGDA), and new distributed methods with compression (QSGDA, DIANA-SGDA, VR-DIANA-SGDA).

1. Unified analysis of SGDA

Theorem

Let F be μ-quasi-strongly monotone ($\mu > 0$) and let Assumption on g^k hold. Assume that $0 < \gamma \leq \min\{1/\mu, 1/2(A+CM)\}$ for some $M > B/\rho$. Then the iterates of **SGDA**, satisfy:

$$\mathbb{E}[V_k] \leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)^k V_0 + \frac{\gamma^2(D_1 + MD_2)}{\min\{\gamma\mu, \rho - B/M\}},$$

where the Lyapunov function V_k is defined by $V_k = \|x^k - x^{*,k}\|^2 + M\gamma^2\sigma_k^2$ for all $k \geq 0$.

Corollary

Let the assumptions of Theorem 1 hold. Then, for some choice of γ_k, the iterates of **SGDA** satisfy:

$$\mathbb{E}[V_K] \leq \frac{32hV_0}{\mu}\exp\left(-\frac{\mu}{h}K\right) + \frac{36(D_1 + 2BD_2/\rho)}{\mu^2 K},$$

where $h = \max\{2(A + 2BC/\rho), 2\mu/\rho\}$.

2. SGDA with Arbitrary Sampling

- Consider a random *sampling* vector $\xi = (\xi_1, \dots, \xi_n)^\top \in \mathbb{R}^n$. One can rewrite $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$ as

$$F(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\mathcal{D}[\xi_i F_i(x)] = \mathbb{E}_\mathcal{D}[F_\xi(x)],$$

where $F_\xi(x) = \frac{1}{n}\sum_{i=1}^n \xi_i F_i(x)$.

- Assume that stochastic operator $F_\xi(x), \xi \sim \mathcal{D}$ is $\ell_\mathcal{D}$-expected cocoercive, i.e. such that for all $x \in \mathbb{R}^d$,

$$\mathbb{E}_\mathcal{D}\left[\|F_\xi(x) - F_\xi(x^*)\|^2\right] \leq \ell_\mathcal{D}\langle F(x) - F(x^*), x - x^*\rangle,$$

where $x^* = \mathrm{proj}_{X^*}(x)$.

- **SGDA with Arbitrary Sampling** [2] : $g^k = F_{\xi^k}(x^k)$.

Corollary

Let F be μ-quasi-strongly monotone and $\ell_\mathcal{D}$-expected cocoercive. Then for all $K > 0$ there exists a choice of γ for which the iterates of **SGDA with Arbitrary Sampling**, satisfy:

$$\mathbb{E}[\|x^K - x^{*,K}\|^2] = \mathcal{O}\left(\frac{\ell_\mathcal{D}\Omega_0^2}{\mu}\exp\left(-\frac{\mu}{\ell_\mathcal{D}}K\right) + \frac{\sigma_*^2}{\mu^2 K}\right),$$

where $\Omega_0^2 = \|x^0 - x^{*,0}\|^2$.

3. SGDA with Variance Reduction

- Focus on the finite-sum problem:

$$F(x) = \tfrac{1}{n}\sum_{i=1}^n F_i(x)$$

- Assume that F is $\widehat{\ell}$-averaged star-cocoercive, i.e. there exists a constant $\widehat{\ell} > 0$ such that for all $x \in \mathbb{R}^d$

$$\frac{1}{n}\sum_{i=1}^n \|F_i(x) - F_i(x^*)\|^2 \leq \widehat{\ell}\langle F(x) - F(x^*), x - x^*\rangle,$$

where $x^* = \mathrm{proj}_{X^*}(x)$.

- **L-SVRGDA**:

$$g^k = F_{j_k}(x^k) - F_{j_k}(w^k) + F(w^k), \quad w^{k+1} = \begin{cases} x^k, & \text{with prob. } p, \\ w^k, & \text{with prob. } 1-p, \end{cases}$$

where in k^{th} iteration j_k is sampled uniformly at random from $[n]$.

Corollary

Let F be μ-quasi strongly monotone and $\widehat{\ell}$-averaged star-cocoercive. Then, for $p = n$, $\gamma = 1/6\widehat{\ell}$ and any $K \geq 0$ we have for **L-SVRGDA**

$$\mathbb{E}[\|x^k - x^*\|^2] \leq V_0 \exp\left(-\min\left\{\frac{\mu}{6\widehat{\ell}}, \frac{1}{2n}\right\}K\right).$$

Table: Summary of the complexity results for variance reduced methods. By default, operator F is assumed to be μ-**strongly monotone** and, as the result, the solution is unique. Our results rely on μ-**quasi strong monotonicity** of F. Methods supporting $R(x) \not\equiv 0$ are highlighted with *. Notation: $\overline{\ell}$, $\overline{L}$ = averaged cocoercivity/Lipschitz constants depending on the sampling strategy, e.g., for uniform sampling $\overline{\ell}^2 = \frac{1}{n}\sum_{i=1}^n \ell_i^2$, $\overline{L}^2 = \frac{1}{n}\sum_{i=1}^n L_i^2$ and for importance sampling $\overline{\ell} = \frac{1}{n}\sum_{i=1}^n \ell_i$, $\overline{L} = \frac{1}{n}\sum_{i=1}^n L_i$; $\widehat{\ell}$ = averaged star-cocoercivity constant.

Method	Citation	Assumptions	Complexity
SVRE	[3]	F_i is ℓ_i-cocoer.	$n + \frac{\overline{\ell}}{\mu}$
EG-VR*	[1]	F_i is L_i-Lip.	$n + \sqrt{n}\frac{\overline{L}}{\mu}$
SVRGDA*	[4]	F_i is L_i-Lip.	$n + \frac{\overline{L}^2}{\mu^2}$
SAGA-SGDA*	[4]	F_i is L_i-Lip.	$n + \frac{\overline{L}^2}{\mu^2}$
VR-AGDA	[5]	F_i is $L_{\max}$-Lip.	$\min\left\{n + \frac{L_{\max}^9}{\mu^9}, n^{2/3}\frac{L_{\max}^3}{\mu^3}\right\}$
L-SVRGDA*	**This paper**	$\widehat{\ell}$-av. st.-cocoer.	$n + \frac{\widehat{\ell}}{\mu}$
SAGA-SGDA*	**This paper**	$\widehat{\ell}$-av. st.-cocoer.	$n + \frac{\widehat{\ell}}{\mu}$

4. Distributed SGDA with Compression

- Assume that $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$, where $\{F_i\}_{i=1}^n$ are distributed across n devices connected with parameter-server in a centralized fashion.
- Operator $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}^d$ (possibly randomized) is called *unbiased compressor/quantization* if there exists a constant $\omega \geq 1$ such that for all $x \in \mathbb{R}^d$

$$\mathbb{E}[\mathcal{Q}(x)] = x, \quad \mathbb{E}[\|\mathcal{Q}(x) - x\|^2] \leq \omega\|x\|^2.$$

- **QSGDA**:

$$g^k = \tfrac{1}{n}\sum_{i=1}^n \mathcal{Q}(g_i^k)$$

for the setting, where all stochastic realizations g_i^k are unbiased and have bounded variance, i.e., for all $i \in [n]$ and $k \geq 0$ the following holds:

$$\mathbb{E}[g_i^k] = F_i(x^k), \quad \mathbb{E}[\|g_i^k - F_i(x^k)\|^2] \leq \sigma_i^2.$$

Corollary

Let F be μ-quasi strongly monotone, ℓ-star-cocoercive, and $\widehat{\ell}$-averaged star-cocoercive and the bounded variance assumption holds. Then, for some γ and any $K \geq 0$ we have for **QSGDA**

$$\mathbb{E}[\|x^K - x^{*,K}\|^2] \leq \frac{32(3\ell + 9\omega\widehat{\ell}/n)}{\mu}\Omega_0^2\exp\left(-\frac{\mu}{(3\ell + 9\omega\widehat{\ell}/n)}K\right) + \frac{36}{\mu^2 K}\cdot\frac{3(1+3\omega)\sigma^2 + 9\omega\zeta_*^2}{n},$$

where $\Omega_0^2 = \|x^0 - x^{*,0}\|^2$.

- **VR-DIANA-SGDA**:

$$\Delta_i^k = g_i^k - h_i^k, \quad h_i^{k+1} = h_i^k + \alpha\mathcal{Q}(\Delta_i^k),$$
$$g^k = h^k + \frac{1}{n}\sum_{i=1}^n \mathcal{Q}(\Delta_i^k), \quad h^{k+1} = \frac{1}{n}\sum_{i=1}^n h_i^{k+1} = h^k + \alpha\frac{1}{n}\sum_{i=1}^n \mathcal{Q}(\Delta_i^k),$$

for the finite-sum setting, where $F_i(x) = \frac{1}{m}\sum_{j=1}^m F_{ij}(x)$ under assumption that there exists a constant $\widetilde{\ell} > 0$ such that for all $x \in \mathbb{R}^d$

$$\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|F_{ij}(x) - F_{ij}(x^*)\|^2 \leq \widetilde{\ell}\langle F(x) - F(x^*), x - x^*\rangle,$$

where $x^* = \mathrm{proj}_{X^*}(x)$.

Corollary

Let F be μ-quasi strongly monotone, ℓ-star-cocoercive and $\widetilde{\ell}$-averaged star-cocoercive. Then, for $p = \frac{1}{m}$, $\alpha = \min\left\{\frac{1}{3m}, \frac{1}{1+\omega}\right\}$, some γ and any $K \geq 0$ we have for **VR-DIANA-SGDA**

$$\mathbb{E}[\|x^k - x^*\|^2] \leq V_0\exp\left(-\min\left\{\frac{\mu}{6\widehat{\ell}}, \frac{1}{2n}\right\}K\right).$$

5. Experiments

- We consider quadratic games:

$$F(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i x + b_i,$$

where each matrix $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ is non-symmetric with all eigenvalues having strictly positive real part. Enforcing all the eigenvalues to have strictly positive real part ensures that the operator is strongly monotone and cocoercive.
In our experiments we focus on two different settings:

(i) problems without constraints, and
(ii) problems with ℓ_1 regularization and constraints forcing the solution to lie in the ℓ_∞-ball of radius r.

- **Uniform sampling (US) vs Importance sampling (IS).**

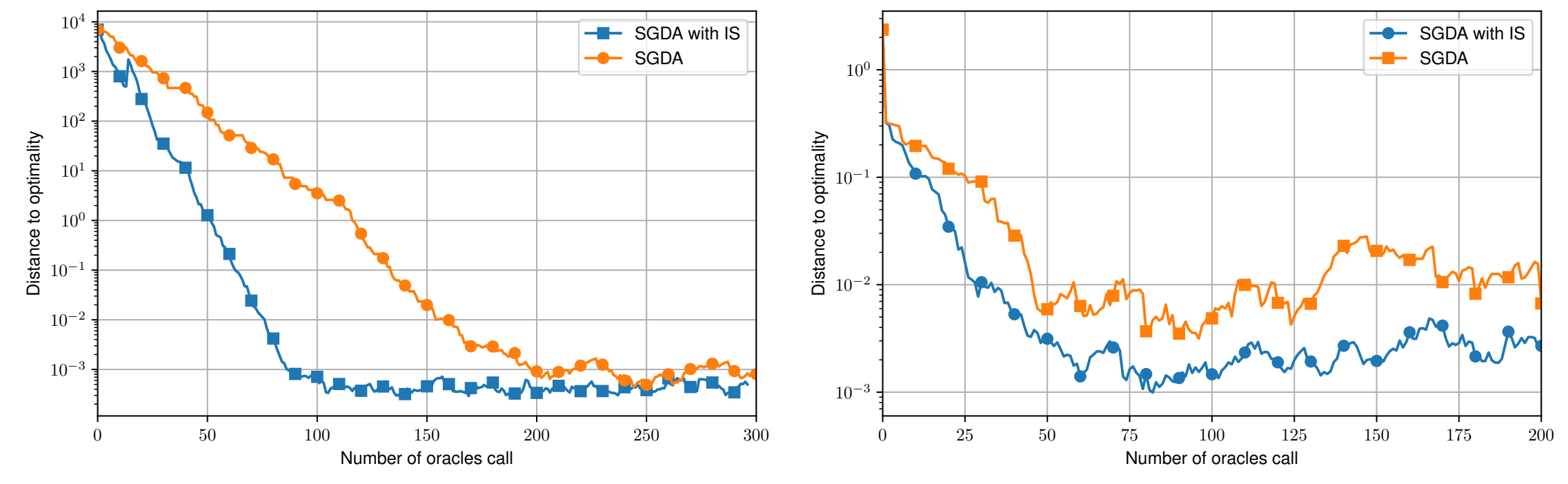

Figure: Comparison of Uniform Sampling (**US**) vs Importance Sampling (**IS**): the first plot shows the result for the problem without constraints, the second one – with constraints.

- **Comparison of variance reduced methods.**

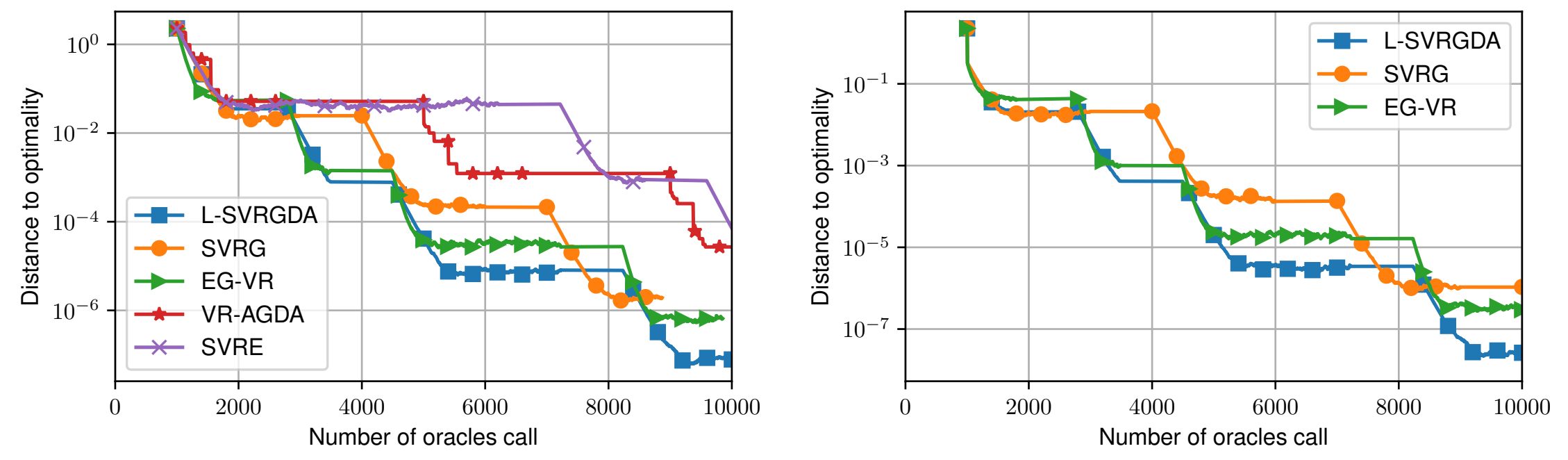

Figure: Comparison of variance reduced methods: the first plot shows the result for the problem without constraints, the second one – with constraints.

- **Comparison of distributed methods.**

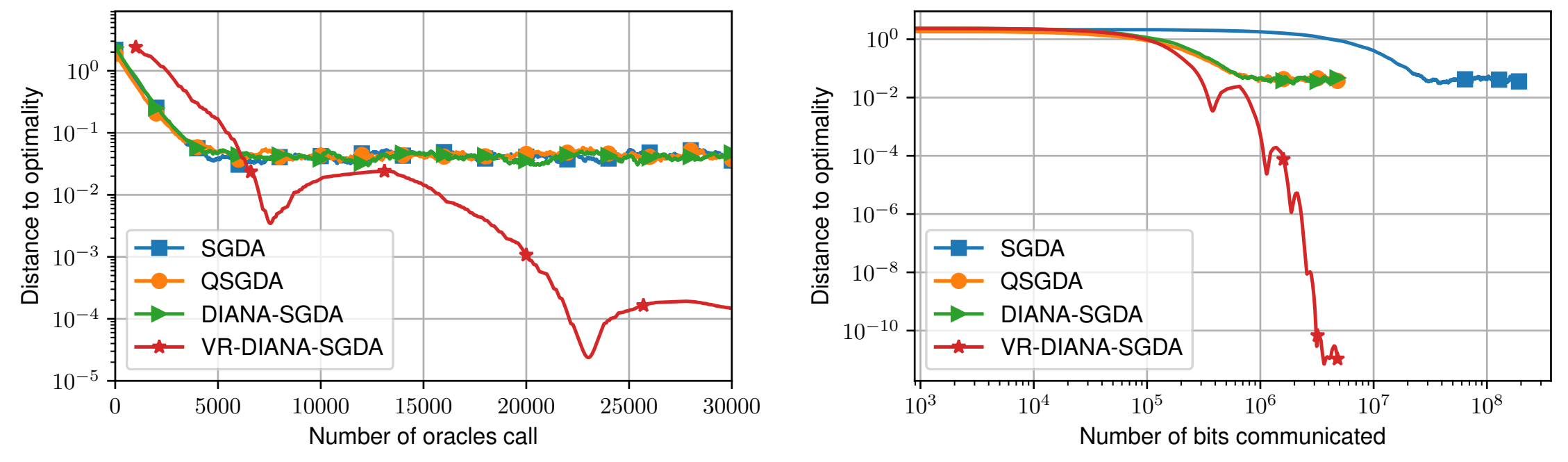

Figure: Comparison of algorithms in distributed setting: the first plot shows the results in terms of the number of oracle calls, the second – the number of bits communicated.

References

[1] A. Alacaoglu and Y. Malitsky. Stochastic variance reduction for variational inequality methods. COLT 2022.

[2] N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. NeurIPS 2021

[3] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. NeurIPS 2019.

[4] B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. NeurIPS 2016.

[5] J. Yang, N. Kiyavash, and N. He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. NeurIPS 2020.