SARAH-based Variance-reduced Algorithm for Stochastic Finite-sum Cocoercive Variational Inequalities

Aleksandr Beznosikov

MIPT

31st January 2023

A. Beznosikov, A. Gasnikov. SARAH-based Variance-reduced Algorithm for Stochastic Finite-sum Cocoercive Variational Inequalities

Variational Inequality

Definition

We consider the unconstrained variational inequality (VI) problem:

$$\text{Find } z^* \in \mathbb{R}^d \text{ such that } F(z^*) = 0,$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is some operator.

Variational Inequality

- Minimization problem:

$$\min_{z \in \mathbb{R}^d} f(z).$$

With $F(z) \stackrel{\text{def}}{=} \nabla f(z)$. And try to find $\nabla f(z^*) = 0$.

- Saddle point problem:

$$\min_{x \in \mathbb{R}^{d_x}} \min_{y \in \mathbb{R}^{d_y}} g(x, y).$$

With $F(z) \stackrel{\text{def}}{=} F(x, y) = [\nabla_x g(x, y), -\nabla_y g(x, y)]$. And try to find $\nabla_x g(x^*, y^*) = 0$.

- Fixed point problem:

$$\text{Найти } z^* \in \mathbb{R}^d \text{ такую, что } T(z^*) = z^*,$$

where $T : \mathbb{R}^d \to \mathbb{R}^d$ is an operator. We need to take $F(z) = z - T(z)$.

Stochastic Variational Inequality

Definition

We consider the unconstrained variational inequality (VI) problem:

$$F(z) = \frac{1}{n}\sum_{i=1}^{n} F_i(z).$$

where $F_i : \mathbb{R}^d \to \mathbb{R}^d$ are some operators.

Stochastic Variational Inequality: examples

- Emperical risk minimization:

$$\min_{z \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i, z), y_i),$$

  where $\{x_i, y_i\}_{i=1}^n$ – данные, f – модель в параметрами z, l – функция потерь.

- Adversarial training:

$$\min_{z \in \mathbb{R}^d} \max_{\|\delta_i\| \le e} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i + \delta_i, z), y_i),$$

  where δ_i – the so-called adversarial noise.

It is very expensive to calculate the full gradient for such problems, therefore stochastic approaches are used.

Assumptions

Definition (Cocoercivity)

Each operator F_i is ℓ-cocoercive, i.e. for all $u, v \in \mathbb{R}^d$ we have

$$\|F_i(u) - F_i(v)\|^2 \leq \ell \langle F_i(u) - F_i(v), u - v \rangle.$$

This assumption is somehow a more restricted analogue of the Lipschetzness of F_i. For convex minimization problems, ℓ-Lipschitzness and ℓ-cocoercivity are equivalent.

Definition (Strong monotonicity)

The operator F is μ-strongly monotone, i.e. for all $u, v \in \mathbb{R}^d$ we have

$$\langle F(u) - F(v); u - v \rangle \geq \mu \|u - v\|^2.$$

For minimization problems this property means strong convexity, and for saddle point problems strong convexity–strong concavity.

Stochastic Methods

$$z^{k+1} = z^k - \eta v^k,$$

where $\eta > 0$ is a predefined step-size.

- SGD or SGDA:
$$v^k = F_i(z^k),$$
where $i \in [n]$ is chosen randomly.
- SVRG or SVRGA:
$$v^k = F_i(z^k) - F_i(w^k) + F(w^k),$$
where $i \in [n]$ is chosen randomly, w^k is a reference point, which is rarely updated.
- SARAH:
$$v^k = F_i(z^k) - F_i(z^{k-1}) + v^{k-1},$$
where $i \in [n]$ is chosen randomly.

Algorithm

Algorithm SARAH for Stochastic Cocoercive Variational Inequalities

Parameters: Stepsize $\gamma > 0$, number of iterations K, S.
Initialization: Choose $\tilde{z}^0 \in \mathbb{R}^d$.
for $s = 1, 2, \ldots, S$ **do**
    $z^0 = \tilde{z}^{s-1}$
    $v^0 = F(z^0)$
    $z^1 = z^0 - \gamma v^0$
    for $k = 1, 2, \ldots, K-1$ **do**
        Sample i_k independently and uniformly from $[n]$
        $v^k = F_{i_k}(z^k) - F_{i_k}(z^{k-1}) + v^{k-1}$
        $z^{k+1} = z^k - \gamma v^k$
    end for
    $\tilde{z}^s = z^K$
end for

Convergence

Theorem

Suppose that Assumptions on cocoercivity and strong monotonicity hold. Consider SARAH Algorithm with $\gamma = \frac{2}{9\ell}$ *and* $K = \frac{10\ell}{\mu}$. *Then, we have*

$$\mathrm{E}[\|F(\tilde{z}^s)\|^2] \leq \frac{1}{2}\mathrm{E}[\|F(\tilde{z}^{s-1})\|^2].$$

Corollary

Suppose that Assumptions on cocoercivity and strong monotonicity hold. Consider SARAH Algorithm with $\gamma = \frac{2}{9\ell}$ *and* $K = \frac{10\ell}{\mu}$. *Then, to achieve* ε*-solution* ($\mathrm{E}\|F(\tilde{z}^S)\|^2 \sim \varepsilon^2$)*, we need*

$$\mathcal{O}\left(\left[n + \frac{\ell}{\mu}\right] \log_2 \frac{\|F(z^0)\|^2}{\varepsilon^2}\right) \quad \textit{oracle calls.}$$

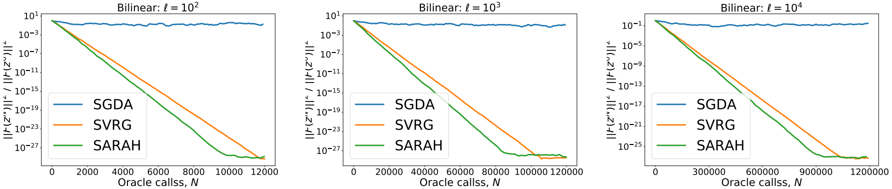
Experiments

- We conduct our experiments on a finite-sum bilinear saddle point problem:

$$g(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left[g_i(x,y) = x^\top A_i y + a_i^\top x + b_i^\top y + \frac{\lambda}{2}\|x\|^2 - \frac{\lambda}{2}\|y\|^2\right],$$

  where $A_i \in \mathbb{R}^{d\times d}$, $a_i, b_i \in \mathbb{R}^d$. This problem is λ-strongly convex–strongly concave and, moreover, L-smooth with $L = \|A\|_2$ for $A = \frac{1}{n}\sum_{i=1}^{n} A_i$. We take $n = 10$, $d = 100$ and generate matrix A and vectors a_i, b_i randomly, $\lambda = 1$. For this problem the cocoercivity constant $\ell = \frac{\|A\|_2^2}{\lambda}$. We run three experiment setups: with small $\ell \approx 10^2$, medium $\ell \approx 10^3$ and big $\ell \approx 10^4$.
- We use SGD, SVRG for comparison with SARAH. The steps of the methods are selected for best convergence. For SVRG and SARAH the number of iterations for the inner loops is taken as $\frac{\ell}{\lambda}$.

Experiments



(a) small ℓ (b) medium ℓ (c) big ℓ

Figure: Bilinear problem: Comparison of state-of-the-art SGD-based methods for stochastic cocoercive VIs.