Journées SMAI MODE 2018

A primal-dual algorithm for finding zeros of random non-monotone operators in Hilbert spaces

Kimon Antonakopoulos
Univ. Grenoble Alpes, CNRS, Inria, LIG, F-38000 Grenoble, France

Panayotis Mertikopoulos
Univ. Grenoble Alpes, CNRS, Inria, LIG, F-38000 Grenoble, France

Abstract. Motivated by Nesterov's *dual averaging* method for solving (stochastic) convex programs and monotone variational inequalities, we propose a primal-dual algorithm for finding zeros of random, non-monotone operators. Given the close connection between the zero set of a maximal monotone operator and the set of solutions of a (Minty-type) variational inequality, we focus on a class of *non-monotone* operators for which the associated Minty variational inequality admits a solution. This property, which we call *variational coherence*, is wide enough to properly include all maximal monotone, pseudomonotone ("+" or "∗"), and other relevant classes of operators. Under mild assumptions for the randomness of the problem at hand (bounded second moments), we show that the algorithm converges to a zero point with probability 1, and we estimate its rate of convergence.

Keywords: Dual averaging; variational coherence; non-monotone operators.

Let $\mathcal{X}$ be a convex and compact of some (possibly infinite-dimensional) Hilbert space $\mathcal{H}$, and let $A\colon \mathcal{X} \to \mathcal{P}(\mathcal{H})$ be an $\mathcal{H}$-valued mapping which is upper-hemicontinuous in the strong-weak sense.[1] Our goal in this paper is to solve the *zero-point problem:*

$$\text{Find } x^* \in \mathcal{X} \text{ such that } 0 \in A(x^*). \tag{Z}$$

In the applications we have in mind (dictionary learning, generative adversarial networks, learning in games, etc.), A is typically derived from a random operator of the general form

$$\mathcal{A}\colon \mathcal{X} \times \Omega \to \mathcal{P}(\mathcal{H}), \tag{1}$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is a complete probability space and

$$A(x) = \mathbb{E}[A(x;\omega)] \quad \text{for all } x \in \mathcal{X}, \tag{2}$$

with integrals and expectations defined in the set-valued sense of Aumann [1]. In this way, (Z) becomes a stochastic problem, for which we make the following standard assumptions:

(A1) There is a mechanism (a *stochastic oracle*) which, for a given input point $(x;\omega) \in \mathcal{X} \times \Omega$, returns a vector $v(x;\omega) \in \mathcal{A}(x;\omega)$ such that $\mathbb{E}[v(x;\omega)] \in A(x)$.

[1] That is, if $x_n \in \mathcal{X}$ converges strongly to x and $v_n \in A(x_n)$ converges weakly to v, then $v \in A(x)$, or, equivalently, the graph of A is sequentially closed in the $\mathcal{H}^{\text{strong}} \times \mathcal{H}^{\text{weak}}$ topology.

(A2) The oracle is bounded in L^2, i.e., $\mathbb{E}[\|v(x;\omega)\|^2] \leq \sigma^2$ for some $\sigma \geq 0$ and all $x \in \mathcal{X}$.

The above problem plays a key role in optimization theory and its applications. A fundamental example is that of minimizing a function $f\colon \mathcal{X} \to \mathbb{R}$, i.e.

$$\begin{aligned} \text{minimize} \quad & f(x), \\ \text{subject to} \quad & x \in \mathcal{X}. \end{aligned} \tag{Opt}$$

When f is convex, the *Fenchel-Moreau subdifferential* ∂f of f is maximal monotone [2], so minimizing f is equivalent to finding a zero of ∂f.

Another key example is that of solving variational inequalities [5]. More concretely, a *Minty-type variational inequality* can be formulated as

$$\text{Find } x^* \in \mathcal{X} \text{ such that } \langle A(x), x - x^* \rangle \geq 0 \text{ for all } x \in \mathcal{X}, \tag{MVI}$$

and, when A is maximal monotone, it is well-known that x^* is a solution of (MVI) if and only if it is a solution of (Z).

The above links have triggered a vast literature for finding zeros of maximal monotone operators. The most widely studied algorithm for this task is the seminal *forward-backward algorithm* (we refer the reader to [2, 3] and [4] for the deterministic and stochastic case respectively). However, monotonicity (and, to a lesser degree, maximality) is crucial for the convergence analysis of forward-backward schemes: beyond this setting, few (if any) results are known, especially for stochastic problems.

Motivated by the above, we focus on a class of operators which we call *variationally coherent* and which are defined as follows:

Definition 1. We say that A is **variationally coherent** if

1. The solution set $\mathcal{X}^*$ of (Z) is nonempty and, for all $x \in \mathcal{X}$, $y \in A(x)$ and $x^* \in \mathcal{X}^*$, we have $\langle y, x - x^* \rangle \geq 0$.
2. $0 \in \langle A(x), x - x^* \rangle$ for all $x^* \in \mathcal{X}^*$ if and only if $x \in \mathcal{X}^*$.

It is straightforward to show that the class of variationally coherent operators strictly includes all maximal monotone operators and several proper relaxations thereof – such as pseudomonotone ("+" or "∗") operators, etc. [5, 6].

Owing to its success in solving monotone variational inequalities, we will focus below on a variant of Nesterov's *dual averaging* algorithm [7], suitably adapted for the setting at hand. The algorithm itself can be described by the recursion:

$$\begin{aligned} y_{n+1} &= y_n + \gamma_n v(x_n; \omega_{n+1}) \\ x_{n+1} &= \arg\max\nolimits_{x \in \mathcal{X}} \{ \langle y_{n+1}, x \rangle - h(y_{n+1}) \}, \end{aligned} \tag{DA}$$

where $h\colon \mathcal{X} \to \mathbb{R}$ is a strongly convex "regularizer function", γ_n is a (positive) step-size sequence, and ω_n is an independent sequence of events drawn from Ω according to $\mathbb{P}$.

Heuristically, the main idea of the method is as follows: At each iteration $n = 1, 2, \ldots,$ the algorithm takes as input a random sample generated by the oracle at the algorithm's current state. Subsequently, the method takes a dual step along the input provided by the oracle in

$\mathcal{H}^* \cong \mathcal{H}$, the outcome is mirrored back to the problem's domain $\mathcal{X}$ to obtain a new solution candidate x_{n+1}, and the process repeats. As a standard example, if $h(x) = 1/2\|x\|^2$, the mirror step is simply a (lazy) projection with respect to the inner product $\langle \cdot, \cdot \rangle$ of $\mathcal{H}$.

In this context, our main result may be stated as follows:

Main Theorem. *Assume that* (DA) *is run with oracle input satisfying* (A1) *and* (A2) *and a step-size sequence* γ_n *such that* $\sum_{k=1}^n \gamma_k^2 / \sum_{k=1}^n \gamma_k \to 0$ *as* $n \to \infty$. *If* A *is variationally coherent,* x_n *converges to a solution of* (Z) *with probability* 1.

As an immediate corollary of this result, it follows that the so-called "ergodic average"

$$\bar{x}_n = \frac{\sum_{k=1}^n \gamma_k x_k}{\sum_{k=1}^n \gamma_k}$$

also converges with probability 1. In particular, under a stronger assumption for A, that of *strong variational coherence* (i.e. that there exists some $L > 0$ such that $\langle A(x), x - x^* \rangle \geq \frac{1}{2} L \|x - x^*\|^2$ for some x^* and all x in $\mathcal{X}$), we also show that the parameter choice $\gamma_n \propto 1/\sqrt{n}$ leads to the convergence rate estimate $\|\bar{x}_n - x^*\|^2 = \mathcal{O}(1/\sqrt{n})$, valid for any initialization of (DA).

References

[1] R. J. Aumann, *Intergrals of set-valued functions*, J. Math. Anal. 12,1-12, 1965.

[2] H. H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2010.

[3] H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, H. Wolkowicz, *Fixed Point Algorithms for inverse problems in Science and Engineering*, Springer, 2010.

[4] P. Bianchi, W. Hachem, *Dynamical Behavior of a Stochastic Forward-Backward Algorithm using monotone operators*,arxiv, 1508.02845v3.

[5] F. Facchinei, J.S. Pang, *Finite- Dimensional Variational Inequalities and Complementarity Problems*, Springer Series in Operations Research, Springer.

[6] N. Hadjisavvas, S. Schaible, N.C. Wong, *Pseudomonotone operators: A survey of the theory and its applications*, J. Optim. Theory Appl. (2012), 152:1-20.

[7] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., Ser B (2009) 120:221-259, Springer.