ADUCA: Adaptive Delayed-Update Cyclic Algorithm for Variational Inequalities

Yi Wei, Xufeng Cai, Jelena Diakonikolas

UW-Madison
yeewei.math@gmail.com

January 22, 2025

Outline

Introduction

The generalized Minty variational inequality (GMVI) problem:

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

where $F : \mathbb{R}^d \to \mathbb{R}^d$, is a monotone, locally block-wise Lipschitz operator, and $g : \mathbb{R}^d \to (-\infty, +\infty]$ is an extended-valued, proper, convex, lower semicontinuous, block-separable function, with an efficiently computable proximal operator.

In particular, we focus on the setting where the vector $\mathbf{u} \in \mathbb{R}^d$ admits a block-wise structure, i.e., $\mathbf{u} = (\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^m)^\top$, where $\mathbf{u}^i \in \mathbb{R}^{d_i}$ is a subvector of $\mathbf{u}$ and $\sum_{i=1}^m d_i = d$.

Introduction

GMVI captures broad classes of optimization problems, such as convex-concave min-max optimization:

$$\min_{x^1 \in \mathbb{R}^{d^1}} \max_{x^2 \in \mathbb{R}^{d^2}} \Phi(x^1, x^2),$$

where

$$\Phi(x^1, x^2) := \phi(x^1, x^2) + g^1(x^1) - g^2(x^2),$$

$d^1 + d^2 = d$, ϕ is convex-concave and smooth, and g^1, g^2 are convex and "simple" (i.e., have efficiently computable proximal operators), and convex composite optimization:

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

where f is smooth and convex, and g is convex and "simple."

Introduction

To reduce convex-concave min-max optimization into GMVI, it suffices to stack vectors x^1, x^2 and define $x = (x^1, x^2)$,

$$F(x) = \begin{bmatrix} \nabla_{x^1}\phi(x^1, x^2) \\ -\nabla_{x^2}\phi(x^1, x^2) \end{bmatrix},$$

$g(x) = g^1(x^1) - g^2(x^2)$.
To reduce convex composite optimization into GMVI, it suffices to take $F(x) = \nabla f(x)$, while g is the same for both problems.

Assumptions

We make the following standard assumptions.

Assumption

There exists at least one $\mathbf{u}^* \in \mathbb{R}^d$ *that solves GMVI, and* $\mathcal{U}^*$ *denotes the solution set of GMVI.*

Assumption

The operator $F : \mathbb{R}^d \to \mathbb{R}^d$ *is monotone, i.e., for any* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$

$$\langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq 0.$$

Assumptions

Assumption (Block-wise Lipschitzness)

For each $i \in [m]$*, given positive definite diagonal matrix* $\mathbf{\Lambda}_i$*, there exists* $L^i \geq 0$ *such that* $F^i(\cdot)$ *is* L^i*-Lipschitz continuous with respect to the* $\mathbf{\Lambda}_i$*-norm, i.e., for any* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$*, we have*

$$\|F^i(\mathbf{u}) - F^i(\mathbf{v})\|_{\mathbf{\Lambda}_i^{-1}} \leq L^i \|\mathbf{u} - \mathbf{v}\|_{\mathbf{\Lambda}_i}. \tag{1}$$

Assumption (Local Block-wise Lipschitzness)

Given a positive integer m*, for every compact set* $\mathcal{C} \subseteq \mathbb{R}^d$ *and* $i \in [m]$*, there exist* $L^i_{\mathcal{C}} \geq 0$ *and positive definite diagonal matrix* $\mathbf{\Lambda}_i$ *such that* $F^i(\cdot)$ *is* $L^i_{\mathcal{C}}$*-Lipschitz continuous with respect to the norm* $\|\cdot\|_{\mathbf{\Lambda}_i}$*, i.e., for any* $\mathbf{u}, \mathbf{v} \in \mathcal{C}$

$$\|F^i(\mathbf{u}) - F^i(\mathbf{v})\|_{\mathbf{\Lambda}_i^{-1}} \leq L^i_{\mathcal{C}} \|\mathbf{u} - \mathbf{v}\|_{\mathbf{\Lambda}_i}. \tag{2}$$

Assumptions

Example (ℓ_1-regularized SVM)

The ℓ_1-regularized support vector machine (SVM) is to solve the following min-max optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in[-1,0]^n} \frac{1}{n}\sum_{i=1}^{n} y_i(-1 + b_i A_i \mathbf{x}) + \lambda\|\mathbf{x}\|_1 + \sum_{j=1}^{n} \mathbb{1}_{-1\le y_j\le 0},$$

where $A = [A_1, \ldots, A_n] \in \mathbb{R}^{n\times d}$, $\boldsymbol{b} \in \mathbb{R}^n$, and $\mathbb{1}_{-1\le y_j\le 0}$ is the convex indicator function of the interval $[-1, 0]$. Let $\bar{A} = [b_1A_1, \ldots, b_nA_n]$. This problem is an instance of GMVI with $F(\mathbf{x}, \mathbf{y}) = \frac{1}{n}[\bar{A}^\top \mathbf{y}, \mathbf{1} - \bar{A}\mathbf{x}] \in \mathbb{R}^{d+n}$, and $g(\mathbf{x}, \mathbf{y}) = \lambda\|\mathbf{x}\|_1 + \sum_{j=1}^{n} \mathbb{1}_{-1\le y_j\le 0}$. Suppose $m = 2$, $d_1 = d$ and $d_2 = n$. To normalize the non-uniform Lipschitz structure of different coordinates, we can set $\mathbf{\Lambda}_1 = \mathrm{diag}(\frac{1}{\|\bar{A}^\top[1]\|}, \ldots, \frac{1}{\|\bar{A}^\top[d]\|})$, and $\mathbf{\Lambda}_2 = \mathrm{diag}(\frac{1}{\|b_1A_1\|}, \ldots, \frac{1}{\|b_nA_n\|})$, where $\bar{A}^\top[i]$ is the i-th row of $\bar{A}^\top$ for any $i \in [m]$.

Assumptions

Assumption

The function $g : \mathbb{R}^d \to (-\infty, \infty]$ *is* μ*-strongly convex with respect to the norm* $\|\cdot\|_{\mathbf{\Lambda}}$ *for* $\mu \geq 0$ *and positive definite matrix* $\mathbf{\Lambda}$*, i.e., for any* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ *and* $g'(\mathbf{u}) \in \partial g(\mathbf{u})$*,*

$$g(\mathbf{v}) \geq g(\mathbf{u}) + \langle g'(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle + \frac{\mu}{2}\|\mathbf{u} - \mathbf{v}\|_{\mathbf{\Lambda}}^2, \tag{3}$$

where $\partial g(\mathbf{u})$ *denotes the subdifferential of* g *at* $\mathbf{u}$*. Further,* g *is block-separable over* $\{\mathcal{S}^j\}_{j=1}^m$*, i.e.,* $g(\mathbf{u}) = \sum_{j=1}^m g^j(\mathbf{u}^j)$*, and admits an efficiently computable proximal operator with respect to the norm* $\|\cdot\|_{\mathbf{\Lambda}}$*.*

Here without loss of generality and for notational brevity, we use the same diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}(\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_m)$ as the one in the previous assumption.

Algorithm

Algorithm 3.1 ADUCA: Adaptive Delayed-Update Cyclic Algorithm

Input: $a_0 > 0$; $\mathbf{u}_0 \in \mathrm{dom}(g)$; $\beta \in (\frac{-1+\sqrt{5}}{2}, 1)$; $\xi \in \left(\max\left\{\frac{1}{\beta(1+\beta)}, 1-\frac{4}{7\beta(1+\beta)}\right\}, 1\right)$; $\theta_0 > 0$; $\mu \geq 0$; $m > 0$, $\{\mathcal{S}^1, \ldots, \mathcal{S}^m\}$.

Initialization: $\mathbf{v}_0 = \mathbf{u}_0$; $a_{-1} = a_0$; $\phi_1 \in (1, \frac{1}{\beta})$, $\phi_2 = \xi\beta(1+\beta)$, $\phi_3 = \frac{4}{7\beta(1+\beta)(1-\xi)}$, $\phi_4 = \frac{1}{2}\sqrt{\frac{(1-\xi)(1+\beta)}{7\beta}}$, $\phi_5 = \frac{1}{7\beta}$.

$\mathbf{u}_1 = \mathrm{argmin}_{\mathbf{u}} \left\{a_0\langle \bar{F}_0, \mathbf{u}\rangle + a_0 g(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{v}_0\|^2_{\mathbf{\Lambda}}\right\}$

$\widetilde{F}_1^i = F^i(\mathbf{u}_1^1, \ldots, \mathbf{u}_1^{i-1}, \mathbf{u}_0^i, \ldots, \mathbf{u}_0^m)$ for $i \in [m]$

for $k = 1$ **to** K **do**

  $\theta_k = \frac{1+\mu a_{k-1}}{1+\mu\beta\phi_1 a_{k-1}} \cdot \theta_{k-1}$

  $\hat{L}_k = \frac{\|F(\mathbf{u}_k) - \widetilde{F}_k\|_{\mathbf{\Lambda}^{-1}}}{\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{\Lambda}}}$

  $L_k = \frac{\|F(\mathbf{u}_k) - F(\mathbf{u}_{k-1})\|_{\mathbf{\Lambda}^{-1}}}{\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{\Lambda}}}$.

  Find the stepsize:

  (3.1)

$$a_k = \min\left\{ \min\left\{\frac{\phi_1}{\mathbb{1}_{\mu>0}}, \frac{\phi_2}{[1-\xi\beta^2\mu a_{k-1}]_+}, \frac{\phi_3(1+\beta\mu a_{k-1})\theta_{k-1}}{\theta_k}\right\} \cdot a_{k-1}, \right.$$
$$\frac{\phi_4}{\hat{L}_k}\sqrt{\frac{(1+\beta\mu a_{k-2})a_{k-1}}{a_{k-2}}} \cdot \sqrt{\frac{\theta_{k-1}}{\theta_k}},$$
$$\left. \frac{\phi_5^2(1+\beta\mu a_{k-2})(1+\beta\mu a_{k-1})}{a_{k-2}L_k^2} \cdot \frac{\theta_{k-1}}{\theta_k} \right\}$$

  for $i = 1$ **to** m **do**

    $\bar{F}_k^i = \widetilde{F}_k^i + \frac{\theta_{k-1}a_{k-1}}{\theta_k a_k}(F^i(\mathbf{u}_{k-1}) - \widetilde{F}_{k-1}^i)$

    $\mathbf{v}_k^i = (1-\beta)\mathbf{u}_k^i + \beta\mathbf{v}_{k-1}^i$

    $\mathbf{u}_{k+1}^i = \mathrm{argmin}_{\mathbf{u}^i} \left\{a_k\langle \bar{F}_k^i, \mathbf{u}^i\rangle + a_k g^i(\mathbf{u}^i) + \frac{1}{2}\|\mathbf{u}^i - \mathbf{v}_k^i\|^2_{\mathbf{\Lambda}_i}\right\}$

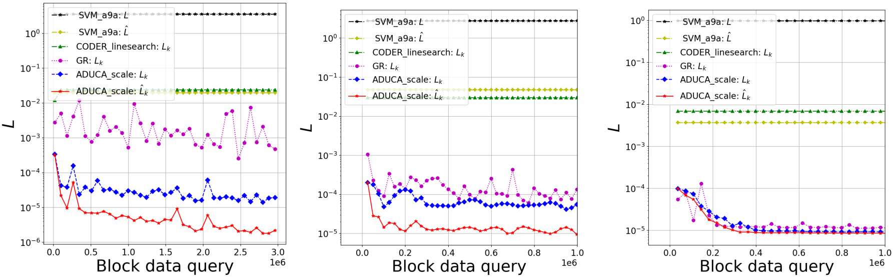
    $\widetilde{F}_{k+1}^i = F^i(\mathbf{u}_{k+1}^1, \ldots, \mathbf{u}_{k+1}^{i-1}, \mathbf{u}_k^i, \ldots, \mathbf{u}_k^m)$

  end for

end for

return $\mathbf{v}_K$, $\mathbf{u}_{K+1}$, $\hat{\mathbf{u}}_{K+1} = \frac{1}{A_{K+1}}\sum_{k=1}^{K+1}\theta_k a_k \mathbf{u}_k$

Algorithm



(a) a9a (b) w8a (c) real-sim

Figure: Comparisons of Lipschitzness estimates of different algorithms.

Convergence Analysis

We consider the following *restricted duality gap* function:

$$\text{Gap}(\mathbf{u}; \mathbf{v}) = \langle F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + g(\mathbf{u}) - g(\mathbf{v}), \ \mathbf{v} \in S$$

where S is a nonempty compact set to be determined. This is a valid suboptimality measure for VI if we can generate a point $\mathbf{u}_\epsilon^* \in S$ satisfying $\text{Gap}(\mathbf{u}_\epsilon^*; \mathbf{u}) \leq \epsilon$ for given $\epsilon > 0$ and any $\mathbf{u} \in S$, shown in Lemma 1 of [1].

Convergence Analysis

Consider $f(x, y) = xy$, $x \in \mathbb{R}, y \in \mathbb{R}$. Then

$$F\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} y \\ -x \end{bmatrix}.$$

Take any $(\bar{x}, \bar{y}) \neq (0, 0)$. Then:

$$\sup_{x\in\mathbb{R}, y\in\mathbb{R}} \left\langle F\left(\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}\right), \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} - \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle \geq \sup_{x\in\mathbb{R}, y\in\mathbb{R}} \left\langle F\left(\begin{bmatrix} x \\ y \end{bmatrix}\right), \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} - \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle.$$

$$= \sup_{x\in\mathbb{R}, y\in\mathbb{R}} \left\langle \begin{bmatrix} -y \\ x \end{bmatrix}, \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} - \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle$$

$$= \sup_{x\in\mathbb{R}, y\in\mathbb{R}} \left(-y\bar{x} + x\bar{y}\right).$$

$$= \infty$$

Convergence Analysis

Theorem

Let $\{\mathbf{u}_k\}_{k\geq 0}$, $\{a_k\}_{k\geq 0}$ and $\{\theta_k\}_{k\geq 0}$ be sequences generated by ADUCA. Suppose $\bar{F}_0 = \widetilde{F}_0 = F(\mathbf{u}_0)$ and $\mathbf{u}_0 = \mathbf{u}_{-1} = \mathbf{u}_{-2}$. Then $\forall\, \mathbf{u} \in dom(g)$, $K \geq 1$, we have:

$$
\begin{aligned}
&A_K \mathrm{Gap}(\hat{\mathbf{u}}_K; \mathbf{u}) + \frac{\theta_K(1+\mu a_K)}{4(1-\beta)}\|\mathbf{u} - \mathbf{v}_{K+1}\|_{\mathbf{\Lambda}}^2 + \sum_{k=1}^{K} \frac{\beta\theta_k a_k}{2a_{k-1}}\|\mathbf{u}_k - \mathbf{v}_{k-1}\|_{\mathbf{\Lambda}}^2 \\
&\leq \frac{\theta_1(1+\beta\mu a_1)}{2(1-\beta)}\|\mathbf{u} - \mathbf{v}_1\|_{\mathbf{\Lambda}}^2 + \Big(\frac{\beta^2\theta_1(1-\xi)(1+\beta)}{4} + \frac{2\theta_0(1+\beta\mu a_0)}{7\beta}\Big)\|\mathbf{u}_1 - \mathbf{u}_0\|_{\mathbf{\Lambda}}^2,
\end{aligned}
$$

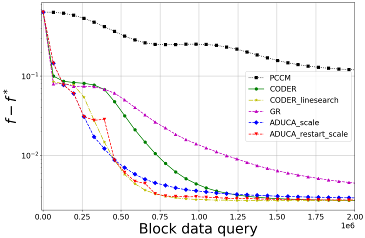
In particular, we have

$$
\mathrm{Gap}(\hat{\mathbf{u}}_K; \mathbf{u}) \leq \frac{1}{A_K}\left(\frac{\theta_1(1+\beta\mu a_1)}{2(1-\beta)}\|\mathbf{u} - \mathbf{v}_1\|_{\mathbf{\Lambda}}^2 + \Big(\frac{\beta^2\theta_1(1-\xi)(1+\beta)}{4} + \frac{2\theta_0(1+\beta\mu a_0)}{7\beta}\Big)\|\mathbf{u}_1 - \mathbf{u}_0\|_{\mathbf{\Lambda}}^2\right).
$$

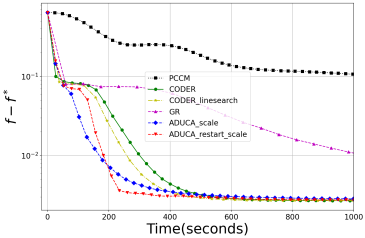
Furthermore, $\forall\, K \geq 1$, given $\mathbf{u}^* \in \mathcal{U}^*$, the iterates $\{\mathbf{u}_k\}_{k\geq 0}$ and $\{\mathbf{v}_k\}_{k\geq 0}$ are bounded. Given $\mathbf{u}^* \in \mathcal{U}^*$, we have

$$
\frac{\theta_K}{4(1-\beta)}\|\mathbf{u}^* - \mathbf{v}_{K+1}\|_{\mathbf{\Lambda}}^2 \leq \frac{\theta_1(1+\beta\mu a_1)}{2(1-\beta)}\|\mathbf{u}^* - \mathbf{v}_1\|_{\mathbf{\Lambda}}^2 + \Big(\frac{\beta^2\theta_1(1-\xi)(1+\beta)}{4} + \frac{2\theta_0(1+\beta\mu a_0)}{7\beta}\Big)\|\mathbf{u}_1 - \mathbf{u}_0\|_{\mathbf{\Lambda}}^2.
$$

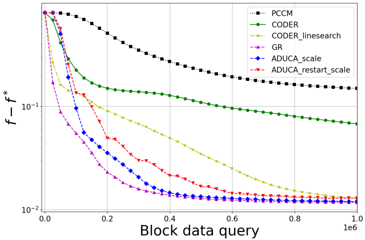
Experiments: SVM



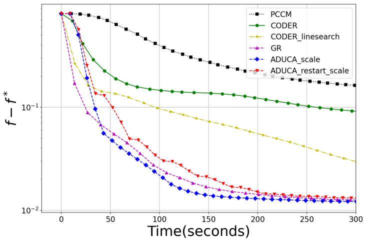
(a) a9a



(b) a9a

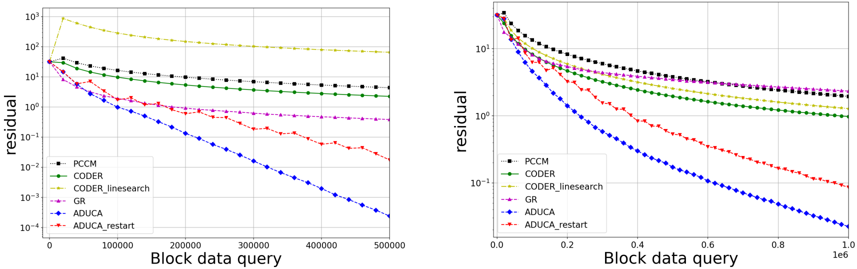


(c) w8a



(d) w8a

Experiments: Nash Equilibrium



(a) (b)

Figure: Results of Nash equilibrium problem. Scenario (a) on the left, (b) on the right.

Future Works

Distributed Block Coordinate methods.

Accelerated ADUCA in minimization problems.

Sharpness-awareness and local smoothness.

Parameter-free methods in neural networks with local smoothness.

Questions?

Thank you!

Any questions?

Yurii Nesterov.
Dual extrapolation and its applications to solving variational inequalities and related problems.
Math. Program., 109(2–3):319–344, mar 2007.