Mirror Descent Methods with Weighting Scheme for Outputs for Constrained Variational Inequality Problems

Mohammad S. Alkousa

Innopolis University

Jointly work with: B. Alashqar, F. Stonyakin, T. Nabhani, S. Ablaev.

March 24, 2025

1. Problem statement
2. Preliminaries
3. Mirror-Descent method for constrained variational inequality problem
4. Mirror-Descent method for variational inequality problem with functional constraints
5. Numerical experiments

Problem statement

We consider two classes of problems. The first one is the following variational inequality problem

$$\text{Find} \quad x^* \in Q: \quad \langle F(x), x^* - x\rangle \leq 0 \quad \forall x \in Q, \tag{1}$$

where Q compact convex set and F an operator is given on Q satisfies some conditions such as continuity, monotonicity, boundedness, ...

Problem statement

We consider two classes of problems. The first one is the following variational inequality problem

$$\text{Find} \quad x^* \in Q : \quad \langle F(x), x^* - x \rangle \leq 0 \quad \forall x \in Q, \tag{1}$$

where Q compact convex set and F an operator is given on Q satisfies some conditions such as continuity, monotonicity, boundedness, ...

For the second class of problems, we consider a set of convex subdifferentiable Lipschitz-continuous functionals $g_i : Q \longrightarrow \mathbb{R}$, $i = 1, 2, \ldots, p$. The problem is the following

$$\text{Find} \quad x^* \in Q : \quad \langle F(x), x^* - x \rangle \leq 0 \quad \forall x \in Q,$$
$$\text{and} \quad g_i(x) \leq 0 \quad \forall i = 1, 2, \ldots, p.$$

Variational inequalities play a key role in machine learning research

- generative models[1] and generative adversarial networks[2],
- supervised/unsupervised learning[3],
- reinforcement learning[4],
- adversarial training[5].

[1] Gauthier Gidel, et.al: A variational inequality perspective on generative adversarial networks. 2018.

[2] Ian Goodfellow, et.al: Generative adversarial networks. Communications of the ACM, 2020.

[3] Thorsten Joachims: A support vector method for multivariate performance measures. International Conference on Machine Learning, 2005.

[4] Shayegan Omidshafiei, et.al: Deep decentralized multi-task multi-agent reinforcement learning under partial observability. 2017.

[5] Aleksander Madry, et.al.: Towards deep learning models resistant to adversarial attacks. 2017.

Next, we mention three common special cases for VIs.

Example (Minimization problem)

Let us consider the minimization problem

$$\min_{x \in Q} f(x), \tag{2}$$

and assume that $F(x) = \nabla f(x)$, where $\nabla f(x)$ denotes the (sub)gradient of f at x. Then, if f is convex, it can be proved that $x^* \in Q$ is a solution to (1) if and only if x^* is a solution to (2).

Example (Saddle point problem)

Let us consider the saddle point problem

$$\min_{u \in Q_u} \max_{v \in Q_v} f(u, v), \tag{3}$$

and assume that $F(x) := F(u, v) = [\nabla_u f(u, v) \quad -\nabla_v f(u, v)]^\top$, where $Q = Q_u \times Q_v$ with $Q_u \subseteq \mathbb{R}^{n_u}, Q_v \subseteq \mathbb{R}^{n_v}$. Then if f is convex in u and concave in v, it can be proved that $x^* \in Q$ is a solution to (1) if and only if $x^* = (u^*, v^*) \in Q$ is a solution to (3).

Example (Fixed point problem)

Let us consider the fixed point problem

$$\text{find } \; x^* \in Q \; \text{ such that } \quad T(x^*) = x^*, \tag{4}$$

where $T : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is an operator. Taking $F(x) = x - T(x)$, it can be proved that $x^* \in Q = \mathbb{R}^n$ is a solution to (1) if $F(x^*) = \mathbf{0} \in \mathbb{R}^n$, that is, x^* is a solution to (4).

Preliminaries

Let $(\mathbf{E}, \|\cdot\|)$ be a normed finite-dimensional vector space, with an arbitrary norm $\|\cdot\|$, and $\mathbf{E}^*$ be the conjugate space of $\mathbf{E}$ with $\|y\|_* = \max\limits_{x \in \mathbf{E}}\{\langle y, x\rangle : \|x\| \leq 1\}$, where $\langle y, x\rangle$ is the value of the continuous linear functional $y \in \mathbf{E}^*$ at $x \in \mathbf{E}$.

Let $Q \subset \mathbf{E}^n$ be a compact convex set with diameter $D > 0$, and $\psi : Q \longrightarrow \mathbb{R}$ be a proper closed differentiable and σ-strongly convex (called prox function). The corresponding Bregman divergence is defined as $V_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle \; \forall x, y \in Q$.

Preliminaries

Let $(\mathbf{E}, \|\cdot\|)$ be a normed finite-dimensional vector space, with an arbitrary norm $\|\cdot\|$, and $\mathbf{E}^*$ be the conjugate space of $\mathbf{E}$ with $\|y\|_* = \max\limits_{x \in \mathbf{E}}\{\langle y, x\rangle : \|x\| \leq 1\}$, where $\langle y, x\rangle$ is the value of the continuous linear functional $y \in \mathbf{E}^*$ at $x \in \mathbf{E}$.

Let $Q \subset \mathbf{E}^n$ be a compact convex set with diameter $D > 0$, and $\psi : Q \longrightarrow \mathbb{R}$ be a proper closed differentiable and σ-strongly convex (called prox function). The corresponding Bregman divergence is defined as $V_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle \; \forall x, y \in Q$.

Definition (δ-monotone operator)

Let $\delta > 0$. The operator $F : Q \longrightarrow \mathbf{E}^*$ is called δ-monotone, if it holds

$$\langle F(y) - F(x), y - x\rangle \geq -\delta \quad \forall x, y \in Q.$$

Preliminaries

For example[6], we can consider $F = \nabla_\delta f$ for δ-subgradient $\nabla_\delta f(x)$ of convex function f at a point $x \in Q$:

$$f(y) - f(x) \geq \langle \nabla_\delta f(x), y - x \rangle - \delta \;\; \forall y \in Q.$$

When $\delta = 0$, then the operator F is called monotone, i.e.,

$$\langle F(x) - F(y), x - y \rangle \geq 0 \quad \forall x, y \in Q.$$

[6] B. T. Polyak. Introduction to optimization. Optimization Software, Inc, New York, 1987.

Preliminaries

For example[6], we can consider $F = \nabla_\delta f$ for δ-subgradient $\nabla_\delta f(x)$ of convex function f at a point $x \in Q$:

$$f(y) - f(x) \geq \langle \nabla_\delta f(x), y - x \rangle - \delta \ \ \forall y \in Q.$$

When $\delta = 0$, then the operator F is called monotone, i.e.,

$$\langle F(x) - F(y), x - y \rangle \geq 0 \quad \forall x, y \in Q.$$

Definition (bounded operator)

We say that the operator F is bounded on Q, if there exist $L_F > 0$ such that

$$\|F(x)\|_* \leq L_F \quad \forall x \in Q.$$

[6] B. T. Polyak. Introduction to optimization. Optimization Software, Inc, New York, 1987.

Constrained variational inequality problem

Let $F : Q \longrightarrow \mathbf{E}^*$ be a continuous, bounded, and δ-monotone operator. We consider the following variational inequality problem

$$\text{Find} \quad x^* \in Q : \quad \langle F(x), x^* - x \rangle \leq 0 \quad \forall x \in Q. \tag{5}$$

For problem (5), we propose Algorithm 1, under consideration

$$V_\psi(x, y) \leq V_\psi(x, x^1) < \infty \quad \forall x, y \in Q, \tag{6}$$

where $x^1 \in Q$ is a chosen (dependently on Q) initial point.

Algorithm 1

Mirror descent method for constrained variational inequality problem.

Require: step sizes $\{\gamma_k\}_{k\geq 1}$, initial point $x^1 \in Q$ s.t. (6) holds, number of iterations N.

1. **for** $k = 1, 2, \ldots, N$ **do**
2. $\quad x^{k+1} = \arg\min_{x \in Q} \left\{ \langle x, F(x^k) \rangle + \frac{1}{\gamma_k} V_\psi(x, x^k) \right\}.$
3. **endfor**

For the quality of a candidate solution $\widehat{x}$, we use the following restricted gap (or merit) function[7]

$$\mathrm{Gap}(\widehat{x}) = \max_{u \in Q} \langle F(u), \widehat{x} - u \rangle.$$

[7] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. Mathematical Programming, 109(2):319–344, 2007.

For Algorithm 1, we have the following result.

Theorem (1)

Let $F : Q \longrightarrow \mathbf{E}^*$ be a continuous, bounded, and δ-monotone operator. Then for problem (5)*, by Algorithm 1, with a positive non-increasing sequence of step sizes $\{\gamma_k\}_{k\geq 1}$, for any fixed $m \geq -1$, it satisfies the following inequality*

$$\operatorname{Gap}(\widehat{x}) \leq \frac{1}{\sum_{k=1}^{N} \gamma_k^{-m}} \left(\frac{R^2}{\gamma_N^{m+1}} + \frac{1}{2\sigma} \sum_{k=1}^{N} \frac{\|F(x^k)\|_*^2}{\gamma_k^{m-1}} \right) + \delta,$$

where $R > 0$, such that $\max_{x \in Q} V_\psi(x, x^1) \leq R^2$, and

$$\widehat{x} = \frac{1}{\sum_{k=1}^{N} \gamma_k^{-m}} \sum_{k=1}^{N} \gamma_k^{-m} x^k.$$

Now, let us take

$$\gamma_k = \frac{\sqrt{2\sigma}}{L_F\sqrt{k}}, \quad \text{or} \quad \gamma_k = \frac{\sqrt{2\sigma}}{\|F(x^k)\|_*\sqrt{k}}, \qquad k = 1, 2, \ldots, N, \tag{7}$$

When $m = -1$, we have the following result.

Corollary (1)

Let $F : Q \longrightarrow \mathbf{E}^*$ be a continuous, bounded, and δ-monotone operator. Then for problem (5), by Algorithm 1, with $m = -1$, and the time-varying step sizes given in (7), it satisfies the following

$$\operatorname{Gap}(\widetilde{x}) \leq \frac{L_F\left(R^2 + 1 + \log(N)\right)}{\sqrt{\sigma}} \cdot \frac{1}{\sqrt{N}} + \delta = O\left(\frac{\log(N)}{\sqrt{N}}\right) + \delta,$$

where $\widetilde{x} = \frac{1}{\sum_{k=1}^N \gamma_k} \sum_{k=1}^N \gamma_k x^k$.

When $m \geq 0$, we have the following result.

Corollary (2)

Let $F : Q \longrightarrow \mathbf{E}^*$ be a continuous, bounded, and δ-monotone operator. Then for problem (5), by Algorithm 1, and the time-varying step sizes given in (7),

- *with $m = 0$, and $\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x^k$ it satisfies*

$$\operatorname{Gap}(\overline{x}) \leq \frac{L_F\left(2+R^2\right)}{\sqrt{2\sigma}} \cdot \frac{1}{\sqrt{N}} + \delta = O\left(\frac{1}{\sqrt{N}}\right) + \delta,$$

- *with any $m \geq 1$, and $\widehat{x} = \frac{1}{\sum_{k=1}^{N}\gamma_k^{-m}}\sum_{k=1}^{N}\gamma_k^{-m}x^k$ it satisfies*

$$\operatorname{Gap}(\widehat{x}) \leq \frac{L_F(m+2)(1+R^2)}{2\sqrt{2\sigma}} \cdot \frac{1}{\sqrt{N}} + \delta = O\left(\frac{1}{\sqrt{N}}\right) + \delta.$$

Variational inequality problem with functional constraints

Consider a set of convex subdifferentiable functionals $g_i : Q \longrightarrow \mathbb{R}$, $i = 1, 2, \ldots, p$. Assume that all functionals g_i are Lipschitz-continuous with some constants $M_{g_i} > 0$, i.e.,

$$|g_i(x) - g_i(y)| \leq M_{g_i} \|x - y\| \quad \forall\, x, y \in Q \;\text{ and }\; i = 1, \ldots, p.$$

Let $F : Q \longrightarrow \mathbf{E}^*$ be a continuous, bounded, and δ-monotone operator. We consider the following variational inequality problem

$$\begin{aligned} &\text{Find} \quad x^* \in Q : \quad \langle F(x), x^* - x \rangle \leq 0 \quad \forall x \in Q, \\ &\qquad\qquad \text{and} \quad g_i(x) \leq 0 \quad \forall i = 1, 2, \ldots, p. \end{aligned} \tag{8}$$

Let us set $g(x) := \max_{1 \leq i \leq p} g_i(x)$, and $M_g := \max_{1 \leq i \leq p} M_{g_i}$.

Algorithm 2

We propose the following algorithm for the problem (8).

Algorithm 2 Mirror descent algorithm for VIs with functional constraints.

Require: $\varepsilon > 0$, initial point $x^1 \in Q$, step sizes $\{\gamma_k^F\}_{k \geq 1}, \{\gamma_k^g\}_{k \geq 1}$, number of iterations N.

$I \longrightarrow \emptyset, J \longrightarrow \emptyset$.

for $k = 1, 2, \ldots, N$ **do**

  if $g(x^k) \leq \varepsilon$ **then**

    $x^{k+1} = \arg\min_{x \in Q} \left\{ \langle x, F(x^k) \rangle + \frac{1}{\gamma_k^F} V_\psi(x, x^k) \right\}.$

    $k \longrightarrow I$ "productive step"

  else

    Calculate $\nabla g(x^k) \in \partial g(x^k)$,

    $x^{k+1} = \arg\min_{x \in Q} \left\{ \langle x, \nabla g(x^k) \rangle + \frac{1}{\gamma_k^g} V_\psi(x, x^k) \right\}.$

    $k \longrightarrow J$ "non-productive step"

  end if

end for

Let $\gamma_k := \gamma_k^F$ if $k \in I$ and $\gamma_k := \gamma_k^g$ if $k \in J$. For Algorithm 2, we have the following result.

Theorem (2)

Let $F : Q \longrightarrow \mathbf{E}^*$ be a continuous, bounded, and δ-monotone operator. Then for problem (8), by Algorithm 2, with a positive non-increasing sequence of step sizes $\{\gamma_k\}_{k \geq 1}$, for any fixed $m \geq -1$, after $N \geq 1$ iterations, with $\widehat{x} = \frac{1}{\sum_{k \in I} (\gamma_k^F)^{-m}} \sum_{k \in I} (\gamma_k^F)^{-m} x^k$ it satisfies the following inequality

$$\mathrm{Gap}(\widehat{x}) < \frac{1}{\sum_{k \in I} (\gamma_k^F)^{-m}} \left(\frac{R^2}{\gamma_N^{m+1}} + \frac{1}{2\sigma} \sum_{k \in I} \frac{\|F(x^k)\|_*^2}{(\gamma_k^F)^{m-1}} + \frac{1}{2\sigma} \sum_{k \in J} \frac{\|\nabla g(x^k)\|_*^2}{(\gamma_k^g)^{m-1}} - (\varepsilon - M_g D) \sum_{k \in J} (\gamma_k^g)^{-m} \right) + \delta.$$

For Algorithm 2, let us take the following time-varying step size rules

$$\gamma_k = \begin{cases} \gamma_k^F := \dfrac{\sqrt{2\sigma}}{L_F\sqrt{k}}, \quad \text{or} \quad \gamma_k^F := \dfrac{\sqrt{2\sigma}}{\|F(x^k)\|_*\sqrt{k}}; & \text{if } k \in I, \\ \gamma_k^g := \dfrac{\sqrt{2\sigma}}{M_g\sqrt{k}}, \quad \text{or} \quad \gamma_k^g := \dfrac{\sqrt{2\sigma}}{\|\nabla g(x^k)\|_*\sqrt{k}}; & \text{if } k \in J. \end{cases} \tag{9}$$

Corollary (3)

Let $F : Q \longrightarrow \mathbf{E}^*$ be a continuous, bounded, and δ-monotone operator. Let $g(x) = \max_{1 \leq i \leq p}\{g_i(x)\}$ be an M_g-Lipschitz convex function, where $g_i : Q \longrightarrow \mathbb{R}$, $\forall i = 1, 2, \ldots, p$ are M_{g_i}-Lipschitz, and $M_g = \max_{1 \leq i \leq p}\{M_{g_i}\}$. Then after $N \geq 1$ iterations of Algorithm 2, such that

$$\frac{M(1+R^2)}{\sqrt{2\sigma}\sqrt{N}} + \frac{MD|J|}{N} \leq \varepsilon,$$

for any fixed $m > 0$, with step size rules given in (9), it satisfies

$$\mathrm{Gap}(\widehat{x}) = \max_{x \in Q} \langle F(x), \widehat{x} - x \rangle < \varepsilon + \delta, \quad \text{and} \quad g(\widehat{x}) \leq \varepsilon,$$

where $\widehat{x} = \frac{1}{\sum_{k \in I}\left(\gamma_k^f\right)^{-m}} \sum_{k \in I} \left(\gamma_k^f\right)^{-m} x^k.$

For $m = 0$, we can formulate the following result.

Corollary (4)

Let $F : Q \longrightarrow \mathbf{E}^*$ be a continuous, bounded, and δ-monotone operator. Let $g(x) = \max_{1 \leq i \leq p}\{g_i(x)\}$ be an M_g-Lipschitz convex function, where $g_i : Q \longrightarrow \mathbb{R}$, $\forall i = 1, 2, \ldots, p$ are M_{g_i}-Lipschitz, and $M_g = \max_{1 \leq i \leq p}\{M_{g_i}\}$. Then, after $N \geq 1$ iterations of Algorithm 2, such that

$$\frac{M\left(2 + R^2\right)}{\sqrt{2\sigma}\sqrt{N}} + \frac{MD|J|}{N} \leq \varepsilon,$$

with $m = 0$ and step size rules given in (9), it satisfies

$$\operatorname{Gap}(\overline{x}) = \max_{x \in Q} \langle F(x), \overline{x} - x \rangle < \varepsilon + \delta, \quad \text{and} \quad g(\overline{x}) \leq \varepsilon,$$

where $\overline{x} = \frac{1}{|I|} \sum_{k \in I} x^k$.

Numerical experiments

We compare the performance of Algorithm 1 with the Modified Projection Method (MPM). We take the Euclidean setup, Q as a unit ball in $\mathbb{R}^n$ and the initial point $x^1 = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) \in \mathbb{R}^n$.

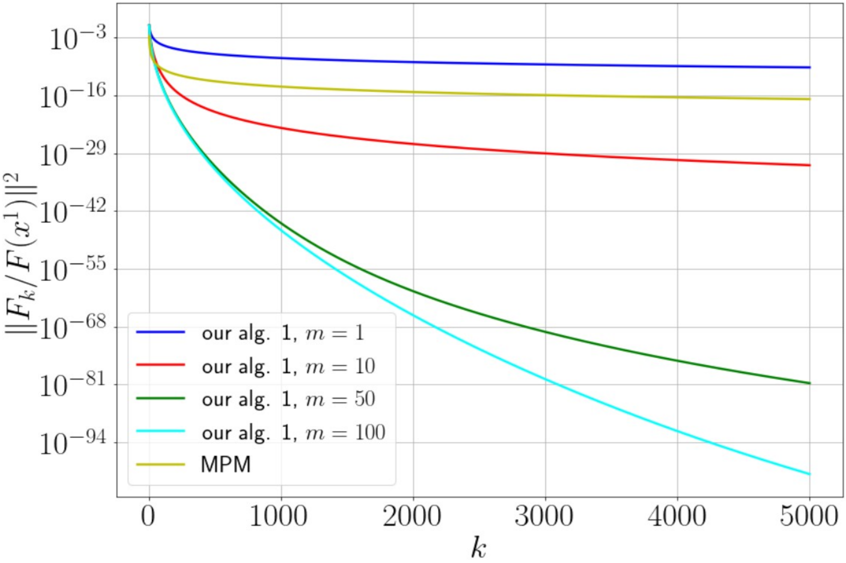
Example (1)

Let $F : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ be a monotone and bounded operator in the unit ball, defined as follows

$$F(x_1, x_2, x_3) = \left(F_1(x_1, x_2, x_3), F_2(x_1, x_2, x_3), F_3(x_1, x_2, x_3)\right),$$

where $F_1 = x_1 - x_2 + x_3 + \sin(x_1)$, $F_2 = x_2 - x_3 + x_1 + \sin(x_2)$, and $F_3 = x_3 - x_1 + x_2 + \sin(x_3)$.

The result for Example (1).



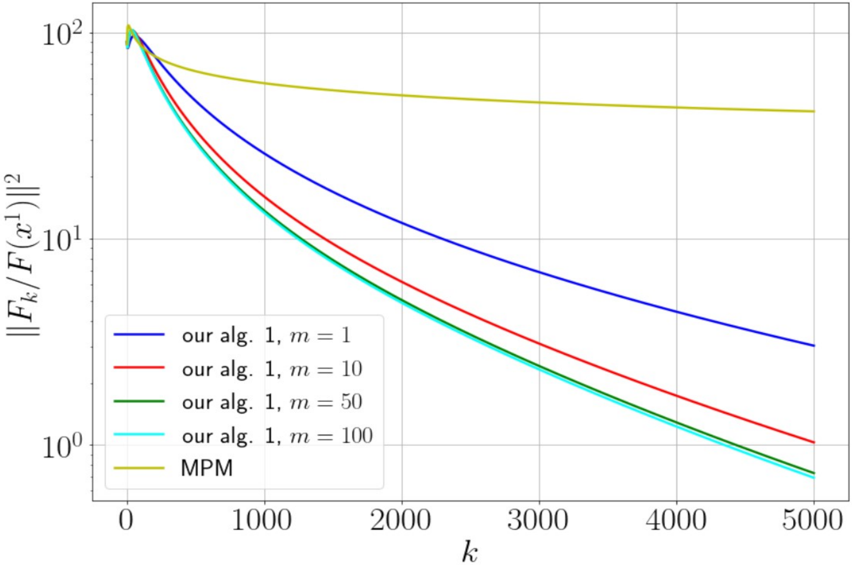
Example (HpHard (or Harker-Pang) problem)

This problem is a well-known issue in nonnegative matrix factorization, appearing in several practical applications. Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be an operator defined by

$$F(x) = Kx + q, \quad K = AA^\top + B + C, q \in \mathbb{R}^n,$$

where $A \in \mathbb{R}^{n \times n}$ is a matrix, $B \in \mathbb{R}^{n \times n}$ is a skew-symmetric matrix (A and B are randomly generated from a normal (Gaussian) distribution with mean equals 0 and scale equals 0.01) and $C \in \mathbb{R}^{n \times n}$ is a diagonal matrix with non-negative diagonal entries (randomly generated from the continuous uniform distribution over the interval $[0, 1)$). Therefore, it follows that K is positive semidefinite. The operator F is monotone and bounded in the unit ball with constant $L_F = \|K\|_2 + \|q\|_2$. For $q = \mathbf{0} \in \mathbb{R}^n$, the solution of problem (5), is $x^* = \mathbf{0} \in \mathbb{R}^n$.

The result for the Haker-Pang problem with $n = 100$.



Thank You for Your Attention!!!