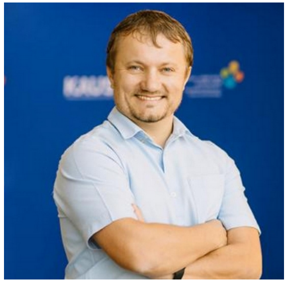
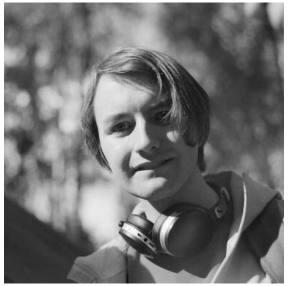
Distributed Methods with Compressed Communication for Solving Variational Inequalities, with Theoretical Guarantees

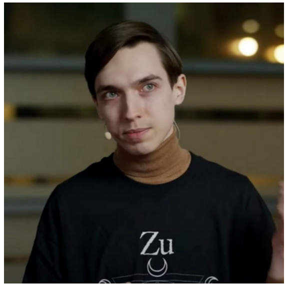
Aleksandr Beznosikov
Innopolis, MIPT, HSE and Yandex



Peter Richtarik
KAUST



Michael Diskin
HSE and Yandex



Max Ryabinin
Yandex and HSE

Alexander Gasnikov
MIPT, HSE and IITP

INNOPOLIS UNIVERSITY · MIPT MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY · HIGHER SCHOOL OF ECONOMICS · Yandex · KAUST

NeurIPS 2022

Variational Inequality Problem

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

Variational Inequality Problem

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

- $\min_{z \in \mathbb{R}^d} f(z) \longrightarrow F(z) := \nabla f(z)$

Variational Inequality Problem

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

- $\min_{z \in \mathbb{R}^d} f(z) \longrightarrow F(z) := \nabla f(z)$
- $\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} g(x, y) \longrightarrow F(z) := [\nabla_x g(x, y), -\nabla_y g(x, y)]$

Variational Inequality Problem

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

- $\min_{z \in \mathbb{R}^d} f(z) \longrightarrow F(z) := \nabla f(z)$
- $\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} g(x, y) \longrightarrow F(z) := [\nabla_x g(x,y), -\nabla_y g(x,y)]$
- $\text{Find } z^* \in \mathbb{R}^d \text{ such that } T(z^*) = z^*$

$$\downarrow$$

$$F(z) := z - T(z)$$

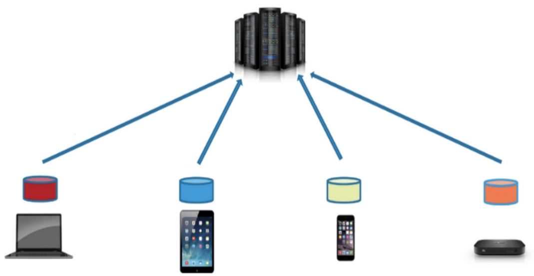
Distributed Variational Inequality

$$F(z) := \frac{1}{M}\sum_{m=1}^{M} F_m(z)$$

Distributed Variational Inequality

$$F(z) := \frac{1}{M} \sum_{m=1}^{M} F_m(z)$$

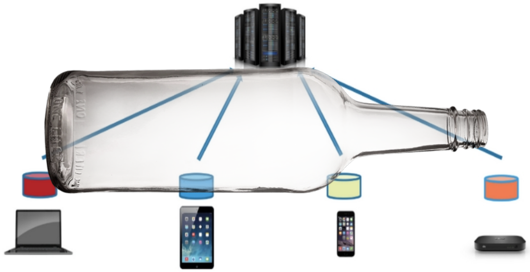
F_m on local devices



Distributed Variational Inequality

$$F(z) := \frac{1}{M}\sum_{m=1}^{M} F_m(z)$$

F_m on local devices → **Communication bottleneck!**

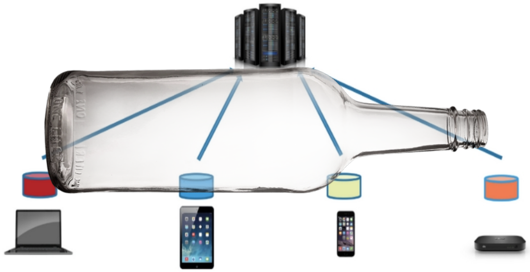


Distributed Variational Inequality

$$F(z) := \frac{1}{M} \sum_{m=1}^{M} F_m(z)$$

F_m on local devices → **Communication bottleneck!**

→ **Use compression in communications!**



Compression

- Unbiased: $\mathbb{E}Q(z) = z, \quad \mathbb{E}\|Q(z)\|^2 \leq q\|z\|^2$

  Examples: random choice of coordinates

Compression

- Unbiased: $\mathbb{E}Q(z) = z, \quad \mathbb{E}\|Q(z)\|^2 \leq q\|z\|^2$

  Examples: random choice of coordinates

- Contractive: $\mathbb{E}\|C(z) - z\|^2 \leq (1 - 1/\delta)\|z\|^2$

  Examples: choice of top coordinates, rounding

Algorithms

Algorithm 1 MASHA1

Parameters: Stepsize $\gamma > 0$, parameter $\tau \in (0; 1)$, number of iterations K.
Initialization: Choose $z^0 = w^0 \in \mathcal{Z}$.
Devices send $F_m(w^0)$ to server and get $F(w^0)$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  for each device m in parallel **do**
    $z^{k+1/2} = \tau z^k + (1-\tau)w^k - \gamma F(w^k)$
    Sends $g_m^k = Q_m^{\text{dev}}(F_m(z^{k+1/2}) - F_m(w^k))$ to server
  end for
  for server **do**
    Sends to devices $g^k = Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^M g_m^k\right]$
    Sends to devices one bit b_k : 1 with probability $1-\tau$, 0 with with probability τ
  end for
  for each device m in parallel **do**
    $z^{k+1} = z^{k+1/2} - \gamma g^k$
    If $b_k = 1$ then $w^{k+1} = z^k$, sends $F_m(w^{k+1})$ to server and gets $F(w^{k+1})$
    else $w^{k+1} = w^k$
  end for
end for

Algorithms

Algorithm 1 MASHA1

Parameters: Stepsize $\gamma > 0$, parameter $\tau \in (0; 1)$, number of iterations K.
Initialization: Choose $z^0 = w^0 \in \mathcal{Z}$.
Devices send $F_m(w^0)$ to server and get $F(w^0)$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  for each device m in parallel **do**
    $z^{k+1/2} = \tau z^k + (1-\tau) w^k - \gamma F(w^k)$
    Sends $g_m^k = Q_m^{\text{dev}}(F_m(z^{k+1/2}) - F_m(w^k))$ to server
  end for
  for server **do**
    Sends to devices $g^k = Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^{M} g_m^k\right]$
    Sends to devices one bit b_k : 1 with probability $1-\tau$, 0 with with probability τ
  end for
  for each device m in parallel **do**
    $z^{k+1} = z^{k+1/2} - \gamma g^k$
    If $b_k = 1$ then $w^{k+1} = z^k$, sends $F_m(w^{k+1})$ to server and gets $F(w^{k+1})$
    else $w^{k+1} = w^k$
  end for
end for

Ideas:

- Extragradient

Algorithms

Algorithm 1 MASHA1

Parameters: Stepsize $\gamma > 0$, parameter $\tau \in (0;1)$, number of iterations K.
Initialization: Choose $z^0 = w^0 \in \mathcal{Z}$.
Devices send $F_m(w^0)$ to server and get $F(w^0)$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  for each device m in parallel **do**
    $z^{k+1/2} = \tau z^k + (1-\tau) w^k - \gamma F(w^k)$
    Sends $g_m^k = Q_m^{\text{dev}}(F_m(z^{k+1/2}) - F_m(w^k))$ to server
  end for
  for server **do**
    Sends to devices $g^k = Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^{M} g_m^k\right]$
    Sends to devices one bit b_k : 1 with probability $1-\tau$, 0 with with probability τ
  end for
  for each device m in parallel **do**
    $z^{k+1} = z^{k+1/2} - \gamma g^k$
    If $b_k = 1$ then $w^{k+1} = z^k$, sends $F_m(w^{k+1})$ to server and gets $F(w^{k+1})$
    else $w^{k+1} = w^k$
  end for
end for

Ideas:

- Extragradient
- Negative momentum
+
- VR teqnique

Algorithms

Algorithm 1 MASHA1

Parameters: Stepsize $\gamma > 0$, parameter $\tau \in (0;1)$, number of iterations K.
Initialization: Choose $z^0 = w^0 \in \mathcal{Z}$.
Devices send $F_m(w^0)$ to server and get $F(w^0)$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  for each device m in parallel **do**
    $z^{k+1/2} = \tau z^k + (1-\tau)w^k - \gamma F(w^k)$
    Sends $g_m^k = Q_m^{\text{dev}}(F_m(z^{k+1/2}) - F_m(w^k))$ to server
  end for
  for server **do**
    Sends to devices $g^k = Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^{M} g_m^k\right]$
    Sends to devices one bit b_k : 1 with probability $1-\tau$, 0 with with probability τ
  end for
  for each device m in parallel **do**
    $z^{k+1} = z^{k+1/2} - \gamma g^k$
    If $b_k = 1$ then $w^{k+1} = z^k$, sends $F_m(w^{k+1})$ to server and gets $F(w^{k+1})$
    else $w^{k+1} = w^k$
  end for
end for

Ideas:

- Extragradient
- Negative momentum
+
- VR teqnique
- Compression of difference

Algorithms

Algorithm 2 MASHA2

Parameters: Stepsize $\gamma > 0$, parameter τ, number of iterations K.
Initialization: Choose $z^0 = w^0 \in \mathcal{Z}$, $e^0_m = 0$, $e^0 = 0$.
Devices send $F_m(w^0)$ to server and get $F(w^0)$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  for each device m in parallel **do**
    $z^{k+1/2} = \tau z^k + (1-\tau)w^k - \gamma F(w^k)$
    Sends $g^k_m = C^{\text{dev}}_m(\gamma F_m(z^{k+1/2}) - \gamma F_m(w^k) + e^k_m)$ to server
    $e^{k+1}_m = e^k_m + \gamma F_m(z^{k+1/2}) - \gamma F_m(w^k) - g^k_m$
  end for
  for server **do**
    Sends to devices $g^k = C^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^M g^k_m + e^k\right]$
    $e^{k+1} = e^k + \frac{1}{M}\sum_{m=1}^M g^k_m - g^k$
    Sends to devices one bit b_k : 1 with probability $1-\tau$, 0 with with probability τ
  end for
  for each device m in parallel **do**
    $z^{k+1} = z^{k+1/2} - \gamma g^k$
    If $b_k = 1$ then $w^{k+1} = z^k$, sends $F_m(w^{k+1})$ to server and gets $F(w^{k+1})$
    else $w^{k+1} = w^k$
  end for
end for

Algorithms

Algorithm 2 MASHA2

Parameters: Stepsize $\gamma > 0$, parameter τ, number of iterations K.
Initialization: Choose $z^0 = w^0 \in \mathcal{Z}$, $e_m^0 = 0$, $e^0 = 0$.
Devices send $F_m(w^0)$ to server and get $F(w^0)$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  for each device m in parallel **do**
    $z^{k+1/2} = \tau z^k + (1-\tau)w^k - \gamma F(w^k)$
    Sends $g_m^k = C_m^{\text{dev}}(\gamma F_m(z^{k+1/2}) - \gamma F_m(w^k) + e_m^k)$ to server
    $e_m^{k+1} = e_m^k + \gamma F_m(z^{k+1/2}) - \gamma F_m(w^k) - g_m^k$
  end for
  for server **do**
    Sends to devices $g^k = C^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^M g_m^k + e^k\right]$
    $e^{k+1} = e^k + \frac{1}{M}\sum_{m=1}^M g_m^k - g^k$
    Sends to devices one bit b_k : 1 with probability $1-\tau$, 0 with with probability τ
  end for
  for each device m in parallel **do**
    $z^{k+1} = z^{k+1/2} - \gamma g^k$
    If $b_k = 1$ then $w^{k+1} = z^k$, sends $F_m(w^{k+1})$ to server and gets $F(w^{k+1})$
    else $w^{k+1} = w^k$
  end for
end for

Main difference:

- Error feedback

Thank you!