On Distributed Methods for Variational Inequalities and Beyond

Aleksandr Beznosikov

MIPT

17 January 2024

Aleksandr Beznosikov

On Distributed Variational Inequalities

17 January 2024

Definition (Stampacchia VIs)

 $\text{Find } z^* \in \mathcal{Z} \text{ such that } \langle F(z^*), z-z^* \rangle + g(z) - g(z^*) \geq 0, \ \forall z \in \mathcal{Z},$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is some operator and g is a proper convex lower semicontinuous function.

Definition (Minty VIs)

 $\text{Find } z^* \in \mathcal{Z} \text{ such that } \langle F(z), z-z^* \rangle + g(z) - g(z^*) \geq 0, \ \forall z \in \mathcal{Z},$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is some operator and g is a proper convex lower semicontinuous function.

- Two formulations are equivariant for smooth monotone operators.
- In the case when $g \equiv 0$ and $\mathcal{Z} = \mathbb{R}^d$, then VI is equal to

Find $z^* \in \mathcal{Z}$ such that F(z) = 0.

Variational Inequality: examples

• Minimization:

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

We take
$$F(z) \stackrel{\text{def}}{=} \nabla f(z)$$
.

• Saddle point problem:

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

Here
$$F(z) \stackrel{\text{def}}{=} F(x, y) = [\nabla_x g(x, y), -\nabla_y g(x, y)].$$

• Fixed point problem:

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

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

• Game theory and economy (comes from von Neumann). Simple example – matrix game (bilinear sadddle point problem on simplexes):

$$\min_{x\in\Delta^{d_x}}\max_{y\in\Delta^{d_y}}x^TAy,$$

where A - cost matrix, x u y - probability of actions.

• Constrained optimization and Lagrange multipliers.

Variational Inequality: ML example

• From classical minimization problem:

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

where $\{x_i, y_i\}_{i=1}^n$ - data, f - model z, l - loss.

• To robust formulation via saddle point problem:

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

where δ_i – adversarial noise.

- GAN represents two models generator G and discriminator D.
- *D* takes an element *x* as input and determines whether this element is real (from a data sample) or artificially generated by the generator.
- The generator is given some random vector *z* as input, from which the generator constructs a "fake" instance similar to the real sample.
- Formally, the GAN training problem is formulated as a saddle point problem:

$$\min_{G} \max_{D} V(D,G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

• We know what to do with:

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

Gradient descent:

$$z^{k+1} = z^k - \gamma \nabla f(z^k).$$

What to do with VIs and saddle point problems? The same idea – descent-ascent:

$$z^{k+1} = z^k - \gamma F(z^k).$$

• The idea of descent-ascent isn't bad and often works, but physical intuition tells that it has some not-so-pleasant aspects.

- The idea of descent-ascent isn't bad and often works, but physical intuition tells that it has some not-so-pleasant aspects.
- Consider min_{x∈R} max_{y∈R} xy. With starting point (1,1). Where is the solution? Point (0,0).
- Vector: \$\begin{pmatrix} \nabla_{xg}(x^k,y^k) \\ -\nabla_{yg}(x^k,y^k) \end{pmatrix}\$ is always orthogonal to \$\begin{pmatrix} x^k x^* \\ y^k y^* \end{pmatrix}\$. What does it means? Method diverges.
- Intuition is not strict, but it can tell us to try something a little different.

Algorithm Extragradient method

Bxoq: stepsize $\gamma > 0$, staring $z^0 \in \mathbb{R}^d$, number of iterations K1: for $k = 0, 1, \dots, K - 1$ do 2: $z^{k+1/2} = z^k - \gamma F(z^k)$ 3: $z^{k+1} = z^k - \gamma F(z^{k+1/2})$ 4: end for

It is easy to check that for this method on the problem $\min_{x \in R} \max_{y \in R} xy$, the directions of the final step in the scalar product with the direction to the solution gives a number greater than 0, hence an acute angle.

Contemporary challenges

Exponential growth in model sizes and data volumes.



Figure: Dynamics of growth of modern language models



Figure: Dynamics of dataset growth

- Cluster learning (large players): we train within one large and powerful computing cluster
- Collaborative learning (all players): pooling computing resources over the Internet
- Federated learning (another paradigm): learn on users' local data using their computational power



Figure: Federated learning

• Formulation (horizontal):

$$F(z) := rac{1}{M}\sum_{m=1}^M F_m(z) := rac{1}{M}\sum_{m=1}^M \mathbb{E}_{\xi\sim\mathcal{D}_m}[F_m(z,\xi)].$$

• The problem is shared among M computing devices, each device m has access only to its own operator F_m or its stochastic realization.

13/47

Communication setups



Figure: Centralized and decentralized connections

э

Communicating through the server

• Let us look at an example of how Extragradient becomes centralized.

Algorithm Centralized Extragradient

Bxc	pg: Stepsize $\gamma > 0$, starting point $z_0 \in \mathbb{R}^d$,	number of iterations <i>K</i>
1:	for $k = 0, 1,, K - 1$ do	
2:	Send z^k to all workers	⊳ by server
3:	for $m = 1, \ldots, M$ in parallel do	
4:	Recieve z^k from server	⊳ by workers
5:	Compute $F_m(z^k)$ in z^k	⊳ by workers
6:	Send $F_m(z^k)$ to server	⊳ by workers
7:	end for	
8:	Recieve $F_m(z^k)$ from all workers	⊳ by server
9:	Compute $F(z^k) = \frac{1}{M} \sum_{m=1}^{M} F_m(z^k)$	⊳ by server
10:	$z^{k+1/2} = z^k - \gamma F(z^k)$	⊳ by server
11:	Similarly for z^{k+1}	
12:	end for	

• In the decentralized setting, it does not work, there is no server.

• Assumption 1. F_m is Lipschitz with constant L, i.e. for all $z_1, z_2 \in \mathcal{Z}$

$$\|F_m(z_1) - F_m(z_2)\| \le L\|z_1 - z_2\|.$$

(smoothness)

• Assumption 2. F_m is strongly monotone with constant μ , i.e. for all $z_1, z_2 \in \mathcal{Z}$

$$\langle F_m(z_1) - F_m(z_2), z_1 - z_2 \rangle \ge \mu ||z_1 - z_2||^2.$$

(strong convexity and strong-convexity-strong-concavity)

• Assumption 3. (for decentralized setting) F_m is stored locally on its own device. All devices are connected in a network (undirected may be time-varying graph $G_k(\mathcal{V}_k, \mathcal{E}_k)$ with max diameter Δ and max condition number χ of Laplace matrix).

- Deterministic case
- Stochastic: bounded variance
- Stochastic: finite sum
- Compression
- Similarity
- Similarity + compression

э

• We can compute full F_m on each device:

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

3

Lower bounds

Lower bounds for distributed algorithms with K communications.

centralized				
VIs	$\Omega\left(R_0^2\exp\left(-\frac{32\mu\kappa}{L} ight) ight)$			
Minimization (exists)	$\Omega\left(R_0^2\exp\left(-rac{\sqrt{\mu}\kappa}{\sqrt{L}} ight) ight)$			
decentralized (fixed network)				
VIs	$\Omega\left(R_0^2\exp\left(-\frac{128\mu K}{L\sqrt{\chi}}\right)\right)$			
Minimization (exists)	$\Omega\left(R_0^2\exp\left(-\frac{\sqrt{\mu}K}{\sqrt{L}\sqrt{\chi}}\right)\right)$			
decentralized (time-varying network)				
VIs	$\Omega\left(R_0^2\exp\left(-\frac{128\mu K}{L\chi}\right)\right)$			
Minimization (exists)	$\Omega\left(R_0^2\exp\left(-\frac{\sqrt{\mu}K}{\sqrt{L_{\chi}}}\right)\right)$			

Table: Lower bounds for distributed VIsp

(4) E > (4) E >

æ

- No "problem" acceleration (unlike minimization) since VIs is a broader class of problems
- No "network" acceleration in the time-varying setting

э

Lower bounds: idea

• Problem:

$$f_m(x,y) = \begin{cases} f_1(x,y) = \frac{M}{2|B_d|} \cdot \frac{L}{2} x^T A_1 y + \frac{\mu}{2} ||x||^2 - \frac{\mu}{2} ||y||^2 + \frac{M}{2|B_d|} \cdot \frac{L^2}{2\mu} e_1^T y, & m \in B_d \\ f_2(x,y) = \frac{M}{2|B|} \cdot \frac{L}{2} x^T A_2 y + \frac{\mu}{2} ||x||^2 - \frac{\mu}{2} ||y||^2, & m \in B \\ f_3(x,y) = \frac{\mu}{2} ||x||^2 - \frac{\mu}{2} ||y||^2, & \text{otherwise} \end{cases}$$

where $e_1 = (1, 0 \dots, 0)$ and

$$A_1 = \begin{pmatrix} 1 & 0 & & & & \\ & 1 & -2 & & & & \\ & & 1 & 0 & & & \\ & & & 1 & -2 & & \\ & & & & 1 & -2 & \\ & & & & & 1 & -2 \\ & & & & & & 1 & 0 \\ & & & & & & & 1 \end{pmatrix}, \ A_2 = \text{analogically}$$

• Network - chain

표 문 문

Optimal algorithms

- In the centralized case, just Centralized Extragradient
- In the decentralized case, we can simulate server communication via gossip procedures.

Algorithm FastMix

Parameters: Vectors $z_1, ..., z_M$, communic. rounds *P*.

Initialization: Construct matrix \mathbf{z} with rows $z_1^T, ..., z_M^T$, choose $\mathbf{z}^{-1} = \mathbf{z}$, $\mathbf{z}^0 = \mathbf{z}$, $\eta = \frac{1 - \sqrt{1 - \lambda_2^2(\tilde{W})}}{1 + \sqrt{1 - \lambda_2^2(\tilde{W})}}$. for h = 0, 1, 2, ..., P - 1 do $\mathbf{z}^{h+1} = (1 + \eta)\tilde{W}\mathbf{z}^h - \eta\mathbf{z}^{h-1}$, end for Output: rows $z_1, ..., z_M$ of \mathbf{z}^P . • We can compute only stochastic realizations $F_m(z,\xi)$ for each device:

$$F(z) := rac{1}{M}\sum_{m=1}^M F_m(z) = rac{1}{M}\sum_{m=1}^M \mathbb{E}_{\xi\sim\mathcal{D}_m}[F_m(z,\xi)].$$

• Assumption. $F_m(z,\xi)$ is unbiased and has bounded variance, i.e. for all $z \in \mathcal{Z}$

$$\mathbb{E}[F_m(z,\xi)] = F_m(z), \ \mathbb{E}[\|F_m(z,\xi) - F_m(z)\|^2] \leq \sigma^2.$$

Lower bounds

Lower bounds for distributed algorithms with K communications and T local computations (T > K).

centralized				
VIs		$\Omega\left(R_0^2\exp\left(-rac{32\mu K}{L} ight)+rac{\sigma^2}{\mu^2 MT} ight)$		
Minimization ((exists)	$\Omega\left(R_0^2\exp\left(-\frac{\sqrt{\mu}K}{\sqrt{L}}\right) + \frac{\sigma^2}{\mu^2 MT}\right)$		
decentralized (fixed network)				
VIs		$\Omega\left(R_0^2\exp\left(-\frac{128\mu K}{L\sqrt{\chi}}\right) + \frac{\sigma^2}{\mu^2 MT}\right)$		
Minimization ((exists)	$\Omega\left(R_0^2\exp\left(-\frac{\sqrt{\mu}K}{\sqrt{L}\sqrt{\chi}}+\frac{\sigma^2}{\mu^2 MT}\right)\right)$		
decentralized (time-varying network)				
VIs		$\Omega\left(R_0^2\exp\left(-\frac{128\mu K}{L\chi}\right) + \frac{\sigma^2}{\mu^2 MT}\right)$		
Minimization ((exists)	$\Omega\left(R_0^2 \exp\left(-\frac{\sqrt{\mu}K}{\sqrt{L_{\chi}}}\right) + \frac{\sigma^2}{\mu^2 MT}\right)$		
Aleksandr Beznosikov	On Distributed Va	riational Inequalities 17 January 2024		

24 / 47

• Consider

$$\min_{x\in\mathbb{R}}f(x)=\frac{\mu}{2}(x-x_0)^2,$$

where we do not know the constant $x_0 \neq 0$.

• Using stochastic first order oracle

$$abla f(x,\xi) = \mu(x+\xi-x_0), ext{ where } \xi \in \mathcal{N}\left(0,rac{\sigma^2}{\mu^2}
ight).$$

э

Batching as additional idea to the deterministic algorithm

Algorithm 1 Centralized Extra Step Method

Parameters: Stepsize $\gamma \leq \frac{1}{4L}$; Communication rounds K, number of local steps T. **Initialization:** Choose $(x^0, y^0) = z^0 \in \mathbb{Z}, k = \left\lfloor \frac{K}{r} \right\rfloor$ and batch size $b = \left\lfloor \frac{T}{2k} \right\rfloor$. for $t = 0, 1, 2, \dots, k$ do for each machine m do $g_m^t = \frac{1}{b}\sum\limits_{i=1}^{s}F_m(z^t,\xi_m^{t,i}), \ \text{send} \ g_m^t,$ on server: $z^{t+1/2} = \operatorname{proj}_{\mathcal{Z}}(z^t - \tfrac{\gamma}{M} \sum_{=}^{M} g_m^t), \text{ send } z^{t+1/2},$ for each machine m do $g_m^{t+1/2} = \frac{1}{b} \sum_{i=1}^{o} F_m(z^{t+1/2}, \xi_m^{t+1/2,i})$, send $g_m^{t+1/2}$, on server: $z^{t+1} = \operatorname{proj}_{\mathcal{Z}}(z^t - \frac{\gamma}{M} \sum_{m=-1}^{M} g_m^{t+1/2}), \text{ send } z^{t+1},$ end for **Output:** z^{k+1} or z^{k+1}_{ava} .

Optimal algorithms: decentralized

Algorithm 2 Decentralized Extra Step Method

Parameters: Stepsize $\gamma \leq \frac{1}{4L}$; Communication rounds K, number of local calls T. **Initialization:** Choose $(x^0, y^0) = z^0 \in \mathcal{Z}, z_m^0 = z^0, k = \left|\frac{K}{H}\right|$ and batch size $b = \left|\frac{T}{2^k}\right|$. for $t = 0, 1, 2, \dots, k$ do for each machine m do $g_m^t = \frac{1}{b} \sum_{i=1}^{b} F_m(z_m^t, \xi_m^{t,i}), \quad \hat{z}_m^{t+1/2} = z_m^t - \gamma g_m^t,$ **communication** $\tilde{z}_{1}^{t+1/2}, ..., \tilde{z}_{M}^{t+1/2} = \text{FastMix}(\hat{z}_{1}^{t+1/2}, ..., \hat{z}_{M}^{t+1/2}, H),$ for each machine m do $z_m^{t+1/2} = \operatorname{proj}_{\mathcal{Z}}(\tilde{z}_m^{t+1/2}),$ $g_m^{t+1/2} = \frac{1}{b} \sum_{i=1}^{b} F_m(z_m^{t+1/2}, \xi_m^{t+1/2,i}),$ $\hat{z}_m^{t+1} = z_m^t - \gamma g_m^{t+1/2},$ communication $\tilde{z}_{1}^{t+1}, \dots, \tilde{z}_{M}^{t+1} = \text{FastMix}(\hat{z}_{1}^{t+1}, \dots, \hat{z}_{M}^{t+1}, H),$ for each machine m do $z_m^{t+1} = \operatorname{proj}_{\mathcal{Z}}(\tilde{z}_m^{t+1}),$ end for **Output:** \bar{z}^{k+1} or \bar{z}^{k+1}_{ava} .

• □ ▶ • • □ ▶ • • □ ▶

3

• We can compute full F_m on each device:

$$F(z) := rac{1}{M} \sum_{m=1}^{M} F_m(z) = rac{1}{M} \sum_{m=1}^{M} rac{1}{n} \sum_{i=1}^{n} F_{m,i}(z).$$

• But we don't want to do it since expensive, we compute only random part $F_{m,i}$.

Lower bounds

Lower bounds for distributed algorithms with ${\cal K}$ communications and ${\cal T}$ local computations.

centralized						
VIs	$\Omega\left(R_0^2\exp\left(-\frac{32\mu K}{L}\right) + R_0^2\exp\left(-\frac{16\mu K}{\sqrt{nL}}\right)\right)$					
Minimization (exists)	$\Omega\left(R_0^2 \exp\left(-\frac{\sqrt{\mu}K}{\sqrt{L}}\right) + R_0^2 \exp\left(-\frac{\sqrt{\mu}K}{\sqrt{n}\sqrt{L}}\right)\right)$					
decentralized (fixed network)						
VIs	$\Omega\left(R_0^2 \exp\left(-\frac{128\mu K}{L\sqrt{\chi}}\right) + R_0^2 \exp\left(-\frac{16\mu K}{\sqrt{nL}}\right)\right)$					
Minimization (exists)	$\Omega\left(R_0^2\exp\left(-\frac{\sqrt{\mu}\kappa}{\sqrt{L}\sqrt{\chi}}\right) + R_0^2\exp\left(-\frac{\sqrt{\mu}\kappa}{\sqrt{n}\sqrt{L}}\right)\right)$					
decentralized (time-varying network)						
VIs	$\Omega\left(R_0^2 \exp\left(-\frac{128\mu K}{L\chi}\right) + R_0^2 \exp\left(-\frac{16\mu K}{\sqrt{nL}}\right)\right)$					
Minimization (exists)	$\Omega\left(R_0^2\exp\left(-\frac{\sqrt{\mu}K}{\sqrt{L_{\chi}}}\right) + R_0^2\exp\left(-\frac{\sqrt{\mu}K}{\sqrt{n_{\chi}}L}\right)\right) < \mathbb{C}$					
Aleksandr Beznosikov On Dist	ributed Variational Inequalities 17 January 2024 29 / 47					

- Double separation
- Random choice of batch

3

Optimal algorithms: non-distributed with bathching

New variance reduction algorithm

Algorithm 4

- Parameters: Stepsizes η > 0, momentums α, γ, batchsize b ∈ {1,..., n}, probability p ∈ (0, 1)
 Initialization: Choose z⁰ = w⁰ ∈ dom g. Put z⁻¹ = z⁰, w⁻¹ = w⁰
- 3: for k = 0, 1, 2... do
- Sample j_1^k, \ldots, j_h^k independently from $\{1, \ldots, m\}$ uniformly at random 4:

$$\begin{aligned} & 5: \quad S^{k} = \{j_{1}^{k}, \dots, j_{b}^{k}\} \\ & 6: \quad \Delta^{k} = \frac{1}{b} \sum_{j \in S^{k}} \left(F_{j}(x^{k}) - F_{j}(w^{k-1}) + \alpha(F_{j}(x^{k}) - F_{j}(x^{k-1}))\right) + F(w^{k-1}) \\ & 7: \quad x^{k+1} = \operatorname{prox}_{\eta g}(x^{k} + \gamma(w^{k} - x^{k}) - \eta \Delta^{k}) \\ & 8: \quad w^{k+1} = \begin{cases} x^{k+1}, & \text{with probability } p \\ w^{k}, & \text{with probability } 1 - p \end{cases} \end{aligned}$$

9: end for

Optimal algorithms: fixed network

Algorithm 1

1: **Parameters:** Stepsizes $\eta, \theta > 0$, momentums α, β, γ , batchsize $b \in \{1, ..., n\}$, probability $p \in (0, 1)$ 2: Initialization: Choose $\mathbf{z}^0 = \mathbf{w}^0 \in (\operatorname{dom} g)^M$, $\mathbf{y}^0 \in$ \mathcal{L}^{\perp} . Put $\mathbf{z}^{-1} = \mathbf{z}^{0}, \mathbf{w}^{-1} = \mathbf{w}^{0}, \mathbf{y}^{-1} = \mathbf{y}^{0}$ 3: for k = 0, 1, 2... do Sample $j_{m,1}^k, \ldots, j_{m,b}^k$ independently from [n]4: $S^{k} = \{j_{m 1}^{k}, \dots, j_{m k}^{k}\}$ 5: Sample $j_{m,1}^{k+1/2}, \ldots, j_{m,b}^{k+1/2}$ independently from [n]6: 7: $S^{k+1/2} = \{j_{m,1}^{k+1/2}, \dots, j_{m,k}^{k+1/2}\}$ $\delta^{k} = \frac{1}{b} \sum_{j \in S^{k}} \left(\mathbf{F}_{j}(\mathbf{z}^{k}) - \mathbf{F}_{j}(\mathbf{w}^{k-1}) \right)$ 8: $+\alpha [\mathbf{F}_{i}(\mathbf{z}^{k}) - \mathbf{F}_{i}(\mathbf{z}^{k-1})] + \mathbf{F}(\mathbf{w}^{k-1})$ $\Delta^k = \delta^k - (\mathbf{v}^k + \alpha(\mathbf{v}^k - \mathbf{v}^{k-1}))$ 9: $\mathbf{z}^{k+1} = \operatorname{prox}_{n\mathbf{g}}(\mathbf{z}^k + \gamma(\mathbf{w}^k - \mathbf{z}^k) - \eta \Delta^k)$ 10: $\Delta^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} \left(\mathbf{F}_j(\mathbf{z}^{k+1}) - \mathbf{F}_j(\mathbf{w}^k) \right)$ 11: $+\mathbf{F}(\mathbf{w}^k)$ $\mathbf{y}^{k+1} = \mathbf{y}^k - \theta(\mathbf{W} \otimes \mathbf{I}_d)(\mathbf{z}^{k+1} - \beta(\Delta^{k+1/2} - \mathbf{v}^k))$ 12: $\mathbf{w}^{k+1} = \begin{cases} \mathbf{z}^k, & \text{with probability } p \\ \mathbf{w}^k, & \text{with probability } 1 - p \end{cases}$ 13: 14: end for

3

Optimal algorithms: fixed network

Algorithm 2

$\begin{array}{l} \alpha, \gamma, \omega, \tau, \text{ parameters } \nu, \beta, \text{ batchsize } b \in \{1, \ldots, n\},\\ \text{ probability } p \in (0, 1) \\ 2: \text{ Initialization: Choose } x^0 = w^0 \in (\operatorname{dom} g)^M, y^0 \in (\mathbb{R}^d)^M, x^0 \in \mathcal{L}^1. \text{ Put } x^{-1} = x^0, w^{-1} = w^0, y_f = y^{-1} = y^0, x_f = x^{-1} = x^0, w_0 = 0^{dM} \\ 3: \text{ for } k = 0, 1, 2, \ldots, d_b \\ 4: \text{ Sample } j_{m,1}, \ldots, j_{m,b}^{k+1/2} \text{ independently from } [n] \\ 5: S^k \in \{j_{m,1}^k, \ldots, j_{m,b}^{k+1/2} \mid \text{independently from } [n] \\ 6: \text{ Sample } j_{m,1}, \ldots, j_{m,b}^{k+1/2} \text{ independently from } [n] \\ 7: S^{k+1/2} = \{j_{m+1/2}^{k+1/2}, \ldots, j_{m+1/2}^{k+1/2} \} \\ 8: \delta^k = \frac{1}{b} \sum_{j \in S^k} \left(\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{z}^{k-1})\right) + \mathbf{F}(\mathbf{w}^{k-1}) \\ \qquad $	1:	Parameters: Stepsizes $\eta_z, \eta_y, \eta_x, \theta > 0$, momentums	
$ \begin{array}{l} \mbox{probability} p \in (0,1) \\ \mbox{2: Initialization: Choose } z^0 = w^0 \in (\operatorname{dom} g)^M, y^0 \in (\mathbb{R}^d)^M, x^0 \in \mathcal{L}^\perp, \operatorname{Put} z^{-1} = z^0, w^{-1} = w^0, y_f = \\ & (\mathbb{R}^d)^M, x^0 \in \mathcal{L}^\perp, \operatorname{Put} z^{-1} = z^0, m_0 = 0^{dM} \\ \mbox{3: for } k = 0, 1, 2, \ldots, do \\ \mbox{3: for } k = 0, 1, 2, \ldots, do \\ \mbox{4: Sample } j_{m,1}^{k,1}, \ldots, j_{m,b}^{k,1} (\operatorname{integraphical} product of [n] \\ \mbox{5: } S^k = \{j_{m,1}^{k,1}, \ldots, j_{m,b}^{k,1/2} (\operatorname{integraphical} product of [n] \\ \mbox{5: } S^{k+1/2} = \{j_{m,1}^{k+1/2}, \ldots, j_{m,b}^{k+1/2}\} \\ \mbox{6: } Sample \; j_{m,1}^{k+1/2}, \ldots, j_{m,b}^{k+1/2} (\operatorname{integraphical} product of [n] \\ \mbox{7: } S^{k+1/2} = \{j_{m,1}^{k+1/2}, \ldots, j_{m,b}^{k+1/2}\} \\ \mbox{8: } \delta^k = \frac{1}{b} \sum_{j \in S^k} \left(\mathbb{F}_j(z^k) - \mathbb{F}_j(z^{k-1}) \right) \\ & + \alpha(\mathbb{F}_j(z^k) - \mathbb{F}_j(z^{k-1})) \\ & + \alpha(\mathbb{F}_j(z^k) - \mathbb{F}_j(z^{k-1})) \\ \mbox{7: } y_j^{k+1/2} = 1 \\ \mbox{7: } x_j^{k+1/2} = (\mathbb{F}_j(z^k) - \mathbb{F}_j(z^k)) \\ \mbox{7: } z_j^{k+1/2} = (\mathbb{F}_j(z^k) - \mathbb{F}_j(z^k)) \\ \mbox{7: } z_j^{k+1} = z^k - (1 - \gamma) x_j^k \\ \mbox{7: } z_j^{k+1/2} = \mathbb{F}_j(z_j \in \mathbb{F}_j(z^k) - \mathbb{F}_j(z^k)) \\ \mbox{7: } z_j^{k+1/2} = \mathbb{F}_j(z_j \in \mathbb{F}_j(z^k) - \mathbb{F}_j(z^k)) \\ \mbox{7: } z_j^{k+1} = z^k - (1 - \gamma) x_j^k \\ \mbox{7: } z_j^{k+1} = z^k - (1 - \gamma) x_j^k \\ \mbox{7: } z_j^{k+1} = z^k - (1 - \gamma) x_j^k \\ \mbox{7: } z_j^{k+1/2} = \mathbb{F}_j(z^k) + \beta(x^k + \delta^{k+1/2}) \\ \mbox{7: } z_j^{k+1} = \mathbb{F}_j^k - \eta_j z_j^k \\ \mbox{7: } z_j^{k+1} = \mathbb{F}_j^k - \eta_j z_j^k \\ \mbox{7: } z_j^{k+1} = \mathbb{F}_j^k - (\mathbb{F}_j(x^k) \otimes \mathbb{I}_j)(\eta_z \Delta_k^k + m^k) \\ \mbox{7: } z_j^{k+1} = x_j^k - (\mathbb{W}_T(Tk) \otimes \mathbb{I}_j)(y_z \Delta_k^k + m^k) \\ \mbox{7: } z_j^{k+1} = x_j^k - (\mathbb{W}_T(Tk) \otimes \mathbb{I}_j)(y_z \Delta_k^k + m^k) \\ \mbox{7: } z_j^{k+1} = \mathbb{F}_j^k - (\mathbb{W}_T(Tk) \otimes \mathbb{I}_j)(y_z \Delta_k^k + m^k) \\ \mbox{7: } z_j^{k+1} = \mathbb{F}_j^k + (\mathbb{F}_j^k) \\ \mbox{7: } z_j^{k+1} = \mathbb{F}_j^k + (\mathbb{F}_j^k) \\ \mbox{7: } z_j^k = \mathbb{F}_j^k + \mathbb{F}_$		$\alpha, \gamma, \omega, \tau$, parameters ν, β , batchsize $b \in \{1, \ldots, n\}$,	
2: Initialization: Choose $\vartheta = \mathbf{w}^0 \in (\operatorname{dom} g)^M, \mathbf{y}^0 \in (\mathbb{R}^d)^M, \mathbf{x}^0 \in \mathcal{L}^\perp, \operatorname{Put} \mathbf{z}^{-1} = \boldsymbol{v}^0, \mathbf{w}^{-1} = \mathbf{w}^0, \mathbf{y}_f = \mathbf{y}^{-1} = \boldsymbol{v}^0, \mathbf{z}_f = \mathbf{x}^{-1} = \mathbf{z}^0, m_0 = \boldsymbol{g}^{dM}$ 3: for $k = 0, 1, 2, \dots, d\theta$ 4: Sample $j_{m,1}^k, \dots, j_{m,b}^k$ independently from $[n]$ 5: $S^k = \{j_{m,1}^k, \dots, j_{m,b}^{k+1/2} \text{ independently from } [n]$ 6: Sample $j_{m,1}^{k+1/2}, \dots, j_{m,b}^{k+1/2}$ independently from $[n]$ 7: $S^{k+1/2} = \{j_{k+1}^{k+1/2}, \dots, j_{k-1}^{k+1/2}\}$ 8: $\delta^k = \frac{1}{b} \sum_{j \in S^k} \left(\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{w}^{k-1}) + \mathbf{F}(\mathbf{w}^{k-1})\right)$ 9: $\Delta_s^k = \delta^k - v\mathbf{z}^k - \mathbf{y}^k - \alpha(\mathbf{y}^k - \mathbf{y}^{k-1})$ 10: $\mathbf{z}^{k+1} = \operatorname{prov}_{n,sg}(\mathbf{z}^k + \omega(\mathbf{w}^k - \mathbf{z}^k) - \eta_z \Delta_z^k)$ 11: $y_c^k = \tau \mathbf{y}^k + (1 - \tau)\mathbf{y}_f^k$ 12: $\mathbf{x}_b^k = \tau \mathbf{y}^k + (1 - \tau)\mathbf{x}_f^k$ 13: $\Delta_b^k = v^{-1}(\mathbf{y}_c^k + \mathbf{x}_c^k) + \mathbf{z}^{k+1} + \gamma(\mathbf{y}^k + \mathbf{x}^k + v\mathbf{z}^k)$ 14: $\delta^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} (\mathbf{F}_j(\mathbf{z}^{k+1}) - \mathbf{F}_j(\mathbf{w}^k))$ 15: $\Delta_s^k = v^{-1}(\mathbf{y}_c^k + \mathbf{x}_c^k) + \beta(\mathbf{x}^k + \delta^{k+1/2})$ 16: $\mathbf{y}^{k+1} = \mathbf{y}^k - \eta_D \Delta_b^k$ 17: $\mathbf{x}^{k+1} = \mathbf{x}^k - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + m^k)$ 18: $m^{k+1} = \mathbf{y}_a \Delta_b^k + m^k$ $-(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_a^k + m^k)$ 19: $\mathbf{y}_f^{k+1} = \mathbf{x}_c^k - d(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(y_c + \mathbf{x}_c^k)$ 21: $\mathbf{w}^{k+1} = \left\{\mathbf{z}^k, \text{ with probability } p$ 22: end for		probability $p \in (0, 1)$	
$ \begin{aligned} & (\mathbb{R}^d)^M, \ & \mathbb{Q} \in \mathcal{L}^4, \ \text{Put } \mathbf{z}^{-1} = \mathbf{z}^0, \ \mathbf{w}^{-1} = \mathbf{w}^0, \ \mathbf{y}_f = \\ & \mathbf{y}^{-1} = \mathbf{y}^0, \ \mathbf{x}_f = \mathbf{x}^{-1} = \mathbf{z}^0, \ \mathbf{m}_0 = 0^{dM} \\ & 3 \ \text{ for } k = 0, 1, 2, \dots, d_0 \\ & 4 \ & \text{ Sample } \frac{1}{\beta_{h-1}}, \dots, \frac{1}{\beta_{h,h}} \text{ independently from } [n] \\ & 5 \ & \mathbf{s}^k = \{j_{h-1}^k, \dots, j_{h,h}^k\} \\ & 6 \ & \text{ Sample } \frac{1}{\beta_{h-1}}, \dots, \frac{1}{\beta_{h-1}}^k \text{ independently from } [n] \\ & 7 \ & \mathbf{s}^{k+1/2} = \{j_{m-1}^k, \dots, j_{h-1}^{k+1/2} \ \text{ independently from } [n] \\ & 7 \ & \mathbf{s}^{k+1/2} = \{j_{m-1}^k, \dots, j_{h-1}^{k+1/2} \} \\ & 8 \ & \delta^k = \frac{1}{b} \sum_{j \in S^k} \left(\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{w}^{k-1}) \\ & + \alpha[\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{z}^{k-1})] \right) + \mathbf{F}(\mathbf{w}^{k-1}) \\ & + \alpha[\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{w}^{k-1})] \\ & 9 \ & \mathbf{z}^{k+1} = p \operatorname{rox}_{n,\mathbf{g}}(\mathbf{z}^k + \mathbf{w}_{h}(\mathbf{w}^k - \mathbf{z}^k) - \eta_z \Delta_z^k) \\ & 10 \ & \mathbf{z}^{k+1} = \operatorname{prox}_{n,\mathbf{g}}(\mathbf{z}^k + \mathbf{w}_{h}(\mathbf{w}^k - \mathbf{z}^k) - \eta_z \Delta_z^k) \\ & 11 \ & \mathbf{y}^k_b = \tau \mathbf{y}^k + (1 - \tau) \mathbf{y}^k_f \\ & 12 \ & \mathbf{x}^k_c = \tau \mathbf{x}^k + (1 - \tau) \mathbf{x}^k_f \\ & 13 \ & \Delta_b^k = \nu^{-1}(\mathbf{y}^k_c + \mathbf{x}^k) + \mathbf{z}^{k+1} + \gamma(\mathbf{y}^k + \mathbf{x}^k + \nu \mathbf{z}^k) \\ & 14 \ & \delta^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} \left(\mathbf{F}_j(\mathbf{z}^{k+1}) - \mathbf{F}(\mathbf{w}^k) \right) \\ & 15 \ & \Delta_b^k = \nu^{-1}(\mathbf{y}^k_c + \mathbf{x}^k) + \beta(\mathbf{x}^k + \delta^{k+1/2}) \\ & 16 \ & \mathbf{y}^{k+1} = \mathbf{y}^k - \eta_y \Delta_b^k \\ & 17 \ & \mathbf{x}^{k+1} = \mathbf{x}^k - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + m^k) \\ & 18 \ & m^{k+1} = \eta_x \Delta_x^k + m^k \\ & - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + m^k) \\ & 19 \ & \mathbf{y}^{k+1}_h = \mathbf{y}^k_h + \tau(\mathbf{y}^{k+1} \to \mathbf{y}^k) \\ & 20 \ & \mathbf{x}^{k+1}_h = \mathbf{x}^k_c - 0(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(y_c + \mathbf{x}^k_c) \\ & 21 \ & \mathbf{w}^{k+1} = \left\{ \mathbf{z}^k, \text{ with probability } p \\ & \mathbf{w}^k, \text{ with probability } 1 - p \\ & 22 \ & \text{ end for } \end{array} \right\}$	2:	Initialization: Choose $\mathbf{z}^0 = \mathbf{w}^0 \in (\operatorname{dom} g)^M$, $\mathbf{y}^0 \in$	
$ \begin{aligned} \mathbf{y}^{-1} &= \mathbf{y}^{0} \ x_{f} = \mathbf{x}^{-1} = \mathbf{x}^{0}, \ m_{0} = 0^{dM} \\ 3: \ \text{for } k = 0, \ 1, 2, \dots, 0 \\ 4: \ \text{Sample } \mathcal{J}_{m,1}^{(1)}, \dots, \mathcal{J}_{m,k}^{(m)} \text{ independently from } [n] \\ 5: \ S^{k} &= \{\mathcal{J}_{k,1}^{k}, \dots, \mathcal{J}_{m,k}^{k+1/2}\} \\ 6: \ \text{Sample } \mathcal{J}_{m,1}^{(1)}, \dots, \mathcal{J}_{m,k}^{k+1/2} \text{ logendently from } [n] \\ 7: \ S^{k+1/2} &= \{\mathcal{J}_{m,1}^{k+1/2}, \dots, \mathcal{J}_{m,k}^{k+1/2}\} \\ 8: \ \delta^{k} &= \frac{1}{b} \sum_{j \in S^{k}} \left(\mathbf{F}_{j}(\mathbf{z}^{k}) - \mathbf{F}_{j}(\mathbf{w}^{k-1}) \\ &\qquad + \alpha \{\mathbf{F}_{j}(\mathbf{z}^{k}) - \mathbf{F}_{j}(\mathbf{w}^{k-1})\} \\ 9: \ \Delta_{k}^{k} &= \delta^{k} - \nu \mathbf{z}^{k} - \mathbf{y}^{k} - \alpha (\mathbf{y}^{k} - \mathbf{y}^{k-1}) \\ 10: \ \mathbf{z}^{k+1} = \operatorname{prox}_{n;\mathbf{g}}(\mathbf{z}^{k} + \mathbf{u}(\mathbf{w}^{k} - \mathbf{z}^{k}) - \eta_{z} \Delta_{z}^{k}) \\ 11: \ \mathbf{y}_{c}^{k} &= \tau \mathbf{y}^{k} + (1 - \tau) \mathbf{y}_{f}^{k} \\ 12: \ \mathbf{x}_{c}^{k} &= \tau \mathbf{x}^{k} + (1 - \tau) \mathbf{y}_{f}^{k} \\ 13: \ \Delta_{y}^{k} &= \nu^{-1} (\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) + \mathbf{z}^{k+1} + \gamma (\mathbf{y}^{k} + \mathbf{x}^{k} + \nu \mathbf{z}^{k}) \\ 14: \ \delta^{k+1/2} &= \frac{1}{b} \sum_{j \in S^{k+1/2}} (\mathbf{F}_{j}(\mathbf{z}^{k+1}) - \mathbf{F}_{j}(\mathbf{w}^{k})) \\ 15: \ \Delta_{x}^{k} &= \nu^{-1} (\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) + \beta (\mathbf{x}^{k} + \delta^{k+1/2}) \\ 16: \ \mathbf{y}^{k+1} &= \mathbf{y}^{k} - \eta_{\mu} \Delta_{y}^{k} \\ 17: \ \mathbf{x}_{s}^{k+1} &= \mathbf{x}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\eta_{z} \Delta_{x}^{k} + m^{k}) \\ 18: \ m^{k+1} &= \eta_{x} - \Delta_{x}^{k} + m^{k} \\ 0: \ \mathbf{x}_{j}^{k+1} &= \mathbf{x}_{c}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\eta_{z} \Delta_{x}^{k} + m^{k}) \\ 19: \ \mathbf{y}_{s}^{k+1} &= \mathbf{x}_{c}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (y_{c}^{k} + \mathbf{x}_{c}^{k}) \\ 21: \ \mathbf{w}^{k+1} &= \begin{cases} \mathbf{z}^{k}, & \text{with probability } p \\ \mathbf{w}^{k}, & \text{with probability } 1 - p \end{cases} \\ 22: \ \text{end for} \end{cases} $		$(\mathbb{R}^d)^M$, $\mathbf{x}^0 \in \mathcal{L}^{\perp}$. Put $\mathbf{z}^{-1} = \mathbf{z}^0$, $\mathbf{w}^{-1} = \mathbf{w}^0$, $\mathbf{y}_f =$	
3: for $k = 0, 1, 2,, d_0$ 4: Sample $j_{m,1},, j_{m,k}^{k+1/2}$ independently from $[n]$ 5: $S^k = \{j_{m,1}^k,, j_{m,k}^{k+1/2} independently from [n]7: S^{k+1/2} = \{j_{m+1/2}^{k+1/2},, j_{m,k}^{k+1/2}\}8: \delta^k = \frac{1}{b} \sum_{j \in S^k} \left(\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{w}^{k-1}) + \alpha [\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{z}^{k-1})] \right) + \mathbf{F}(\mathbf{w}^{k-1})9: \Delta_k^k = \delta^k - \nu \mathbf{z}^k - \mathbf{y}^k - \alpha (\mathbf{y}^k - \mathbf{y}^{k-1})10: \mathbf{z}^{k+1} = \operatorname{prox}_{n,g}(\mathbf{z}^k) + \alpha (\mathbf{w}^k - \mathbf{z}^k) - \eta_z \Delta_z^k)11: \mathbf{y}_c^k = \tau \mathbf{y}^k + (1 - \tau) \mathbf{y}_f^k12: \mathbf{x}_c^k = \tau \mathbf{x}^k + (1 - \tau) \mathbf{x}_f^k13: \Delta_b^k = \nu^{-1} (\mathbf{y}_c^k + \mathbf{x}_c^k) + \mathbf{z}^{k+1} + \gamma (\mathbf{y}^k + \mathbf{x}^k + \nu \mathbf{z}^k)14: \delta^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} (\mathbf{F}_j(\mathbf{z}^{k+1}) - \mathbf{F}_j(\mathbf{w}^k))15: \Delta_k^k = \nu^{-1} (\mathbf{y}_c^k + \mathbf{x}_c^k) + \beta (\mathbf{x}^k + \delta^{k+1/2})16: \mathbf{y}^{k+1} = \mathbf{y}^k - \eta_y \Delta_y^k17: \mathbf{x}^{k+1} = \mathbf{x}^k - (W_T(Tk) \otimes \mathbf{I}_d) (\eta_z \Delta_x^k + m^k)18: m^{k+1} = \eta_z \Delta_k^k + m^k-(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d) (\mathbf{y}_c \Delta_x^k + m^k)19: \mathbf{y}_f^{k+1} = \mathbf{x}_c^k - \theta(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d) (\mathbf{y}_c + \mathbf{x}_c^k)21: \mathbf{w}^{k+1} = \left\{ \mathbf{z}^k, \text{ with probability } p \right\}22: end for$		$\mathbf{v}^{-1} = \mathbf{v}^0, \mathbf{x}_\ell = \mathbf{x}^{-1} = \mathbf{x}^0, m_0 = 0^{dM}$	
$ \begin{array}{ll} \begin{array}{l} 4: & {\rm Sample}\;j_{m,1}^{k}, \ldots, j_{m,b}^{k}\; {\rm independently}\; {\rm from}\; [n] \\ 5: & S^{k}=\{j_{m,1}^{k}, \ldots, j_{m,b}^{k+1/2}\; {\rm independently}\; {\rm from}\; [n] \\ 6: & {\rm Sample}\; j_{m,1}^{k+1/2}, \ldots, j_{m,b}^{k+1/2}\; {\rm independently}\; {\rm from}\; [n] \\ 7: & S^{k+1/2}=\{j_{m,1}^{k+1/2}, \ldots, j_{m,b}^{k+1/2}\} \\ 8: & \delta^{k}=\frac{1}{b}\sum_{j\in S^{k}}\left({\rm F}_{j}({\bf z}^{k})-{\rm F}_{j}({\bf w}^{k-1})\right) \\ & \qquad \qquad$	3:	for $k = 0, 1, 2, \dots$ do	
5: $S^{k} = \{j_{k,1}^{k}, \dots, j_{m,b}^{k}\}$ 6: Sample $j_{m,1}^{k+1/2}, \dots, j_{m,b}^{k+1/2}$ independently from $[n]$ 7: $S^{k+1/2} = \{j_{k+1}^{k+1/2}, \dots, j_{m,b}^{k+1/2}\}$ 8: $\delta^{k} = \frac{1}{b} \sum_{j \in S^{k}} \left(\mathbf{F}_{j}(\mathbf{z}^{k}) - \mathbf{F}_{j}(\mathbf{w}^{k-1}) + \mathbf{f}(\mathbf{w}^{k-1}) + \alpha(\mathbf{F}_{j}(\mathbf{z}^{k}) - \mathbf{F}_{j}(\mathbf{z}^{k-1})]\right) + \mathbf{F}(\mathbf{w}^{k-1})$ 9: $\Delta_{\mathbf{z}}^{k} = \delta^{k} - \nu \mathbf{z}^{k} - \nu^{k} - \alpha(\mathbf{y}^{k} - \mathbf{y}^{k-1})$ 10: $\mathbf{z}^{k+1} = \operatorname{prox}_{n_{1}\mathbf{z}}(\mathbf{z}^{k} + \alpha(\mathbf{w}^{k} - \mathbf{z}^{k}) - \eta_{z}\Delta_{z}^{k})$ 11: $\mathbf{y}_{c}^{k} = \tau \mathbf{y}^{k} + (1 - \tau)\mathbf{y}_{t}^{k}$ 12: $\mathbf{x}_{c}^{k} = \pi \mathbf{z}^{k} + (1 - \tau)\mathbf{x}_{t}^{k}$ 13: $\Delta_{y}^{k} = \nu^{-1}(\mathbf{y}_{c}^{k} + \mathbf{z}^{k}) + \mathbf{z}^{k+1} + \gamma(\mathbf{y}^{k} + \mathbf{x}^{k} + \nu \mathbf{z}^{k})$ 14: $\delta^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} (\mathbf{F}_{j}(\mathbf{z}^{k+1}) - \mathbf{F}_{j}(\mathbf{w}^{k}))$ 15: $\Delta_{x}^{k} = \nu^{-1}(\mathbf{y}_{c}^{k} + \mathbf{z}^{k}) + \beta(\mathbf{x}^{k} + \delta^{k+1/2})$ 16: $\mathbf{y}^{k+1} = \mathbf{y}^{k} - \eta \Delta_{y}^{k}$ 17: $\mathbf{x}^{k+1} = \mathbf{x}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\eta_{z}\Delta_{x}^{k} + m^{k})$ 18: $m^{k+1} = \eta_{z}\Delta_{x}^{k} + m^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\eta_{z}\Delta_{x}^{k} + m^{k})$ 19: $\mathbf{y}_{t}^{k+1} = \mathbf{x}_{c}^{k} - \theta(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k})$ 21: $\mathbf{w}^{k+1} = \left\{\mathbf{z}_{c}^{k}, \text{ with probability } p - \mathbf{y}\right\}$ 22: end for	4:	Sample $j_{m,1}^k, \ldots, j_{m,k}^k$ independently from [n]	
$ \begin{aligned} & \delta & = 0 & \sum_{k=1}^{k+1/2} \sum_{j \in \mathbb{N}^{k+1/2}} \left[independently from [n] \\ & f & Sample f_{m,1}^{j}, \dots, f_{m,k}^{j+1/2} \\ & f & starting \\ & \delta^k = \frac{1}{b} \sum_{j \in \mathbb{N}^k} \left(\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{w}^{k-1}) \right) \\ & + \alpha[\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{z}^{k-1})] \right) + \mathbf{F}(\mathbf{w}^{k-1}) \\ & + \alpha[\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{z}^{k-1})] \\ & 0 & z^{k+1} = \operatorname{prox}_{n,g}(\mathbf{z}^k + \omega(\mathbf{w}^k - \mathbf{z}^k) - \eta_z \Delta_z^k) \\ & 10 & z^{k+1} = \operatorname{prox}_{n,g}(\mathbf{z}^k + \omega(\mathbf{w}^k - \mathbf{z}^k) - \eta_z \Delta_z^k) \\ & 11 & y_c^k = \tau y^k + (1 - \tau) y_f^k \\ & 12 & \mathbf{x}_c^k = \tau \mathbf{x}^k + (1 - \tau) \mathbf{x}_f^k \\ & 13 & \Delta_b^k = \nu^{-1}(\mathbf{y}_c^k + \mathbf{x}_c^k) + \mathbf{z}^{k+1} + \gamma(\mathbf{y}^k + \mathbf{x}^k + \nu \mathbf{z}^k) \\ & \mathbf{z}^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} \left[\mathbf{F}_j(\mathbf{z}^{k+1}) - \mathbf{F}_j(\mathbf{w}^k) \right] \\ & + \mathbf{F}(\mathbf{w}^k) \\ & 15 & \Delta_b^k = \nu^{-1}(\mathbf{y}_c^k + \mathbf{x}_c^k) + \beta(\mathbf{x}^k + \delta^{k+1/2}) \\ & 16 & \mathbf{y}^{k+1} = \mathbf{y}^k - \eta_y \Delta_b^k \\ & 17 & \mathbf{x}^{k+1} = \mathbf{x}^k - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + m^k) \\ & 18 & m^{k+1} = \eta_z \Delta_x^k + m^k \\ & - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + m^k) \\ & 19 & \mathbf{y}_j^{k+1} = \mathbf{x}_c^k - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(y_c^k + \mathbf{x}_c^k) \\ & 21 & \mathbf{w}_j^{k+1} = \left\{ \mathbf{z}^k, \text{with probability } p \\ & \mathbf{w}^k, \text{with probability } 1 - p \\ & 22 & \text{end for} \end{aligned} \right$	5:	$S^k = \{j_{k_1}^k, \dots, j_{k_{n-1}}^k\}$	
$\begin{array}{ll} \begin{array}{l} \text{6.} & \text{Sample } y_{m,1}, \dots, y_{m,k}, & \text{molecularly from } [n] \\ \text{7.} & S^{k+1/2} = \{j_{m,1}^{k+1/2}, \dots, j_{m,k}^{k+1/2}\} \\ \text{8.} & \delta^k = \frac{1}{b} \sum_{j \in S^k} \left(\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{w}^{k-1}) \\ & + \alpha[\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{z}^{k-1})] \right) + \mathbf{F}(\mathbf{w}^{k-1}) \\ \text{9.} & \Delta_k^k = \delta^k - \nu \mathbf{z}^k - \mathbf{y}^k - \alpha(\mathbf{y}^k - \mathbf{y}^{k-1}) \\ \text{10.} & \mathbf{z}^{k+1} = \operatorname{prox}_{n,g}(\mathbf{z}^k + \mathbf{u}(\mathbf{w}^k - \mathbf{z}^k) - \eta_z \Delta_k^k) \\ \text{11.} & \mathbf{y}_k^k = \tau \mathbf{y}^k + (1 - \tau) \mathbf{y}_j^k \\ \text{12.} & \mathbf{x}_k^k = \tau \mathbf{x}^k + (1 - \tau) \mathbf{y}_j^k \\ \text{13.} & \Delta_y^k = \nu^{-1}(\mathbf{y}_k^k + \mathbf{x}_k^k) + \mathbf{z}^{k+1} + \gamma(\mathbf{y}^k + \mathbf{x}^k + \nu \mathbf{z}^k) \\ \text{14.} & \delta^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} (\mathbf{F}_j(\mathbf{z}^{k+1}) - \mathbf{F}_j(\mathbf{w}^k)) \\ \text{15.} & \Delta_k^k = \nu^{-1}(\mathbf{y}_k^k + \mathbf{x}_k^k) + \beta(\mathbf{x}^k + \delta^{k+1/2}) \\ \text{16.} & \mathbf{y}^{k+1} = \mathbf{y}^k - \eta_j \Delta_k^k \\ \text{17.} & \mathbf{x}^{k+1} = \mathbf{x}^k - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + \mathbf{m}^k) \\ \text{18.} & \mathbf{m}^{k+1} = \eta_z \Delta_k^k + \mathbf{m}^k \\ & -(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + \mathbf{m}^k) \\ \text{19.} & \mathbf{y}_j^{k+1} = \mathbf{x}_k^k - \theta(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(y_k - \mathbf{x}_k^k) \\ \text{20.} & \mathbf{x}_j^{k+1} = \mathbf{x}_k^k - \theta(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(y_k - \mathbf{x}_k^k) \\ \text{21.} & \mathbf{w}^{k+1} = \left\{ \mathbf{z}_j^k, \text{ with probability } p \\ \mathbf{w}^k, \text{ with probability } 1 - p \\ \text{22.} & \text{end for} \end{array} \right\} \xrightarrow{\mathbf{P}} $	2	$\mathbf{S}_{\text{constant}} = \frac{(k+1/2)}{(k+1/2)} \cdot \frac{(k+1/2)}{(k+1/2)} \cdot$	
7: $S^{k+1/2} = \{j_{w_1}^{k+1/2}, \dots, j_{w_k}^{k+1/2}\}\$ 8: $\delta^k = \frac{1}{b} \sum_{j \in S^k} \left(\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{w}^{k-1}) + \mathbf{h}(\mathbf{F}_i(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{z}^{k-1})]\right) + \mathbf{F}(\mathbf{w}^{k-1})\$ 9: $\Delta_{\mathbf{z}}^k = \delta^k - \nu \mathbf{z}^k - \mathbf{y}^k - \alpha(\mathbf{y}^k - \mathbf{y}^{k-1})\$ 10: $\mathbf{z}^{k+1} = \operatorname{prox}_{n_{12}} \mathbf{g}(\mathbf{z}^k + \omega(\mathbf{w}^k - \mathbf{z}^k) - \eta_z \Delta_z^k)\$ 11: $\mathbf{y}_c^k = \tau \mathbf{y}^k + (1 - \tau) \mathbf{y}_f^k\$ 12: $\mathbf{x}_c^k = \tau \mathbf{x}^k + (1 - \tau) \mathbf{x}_f^k\$ 13: $\Delta_y^k = \nu^{-1}(\mathbf{y}_c^k + \mathbf{x}_c^k) + \mathbf{z}^{k+1} + \gamma(\mathbf{y}^k + \mathbf{x}^k + \nu \mathbf{z}^k)\)\$ 14: $\delta^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} \left(\mathbf{F}_j(\mathbf{z}^{k+1}) - \mathbf{F}_j(\mathbf{w}^k)\right)\$ 15: $\Delta_x^k = \nu^{-1}(\mathbf{y}_c^k + \mathbf{x}_c^k) + \beta(\mathbf{x}^k + \delta^{k+1/2})\$ 16: $\mathbf{y}^{k+1} = \mathbf{y}^k - \eta_\lambda \Delta_y^k\$ 17: $\mathbf{x}^{k+1} = \mathbf{x}^k - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + m^k)\$ 18: $m^{k+1} = \mathbf{y}_c^k \pm \tau (\mathbf{y}^{k+1} - \mathbf{y})\$ 20: $\mathbf{x}_f^{k+1} = \mathbf{x}_c^k - \theta(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\mathbf{y}_c + \mathbf{x}_c^k)\$ 21: $\mathbf{w}^{k+1} = \left\{\mathbf{z}_c^k, \text{ with probability } p\$ 22: end for	0:	Sample $j_{m,1}$,, $j_{m,b}$ independently from $[n]$	
8: $ \delta^{k} = \frac{1}{b} \sum_{j \in S^{k}} \left(\mathbf{F}_{j}(\mathbf{z}^{k}) - \mathbf{F}_{j}(\mathbf{w}^{k-1}) + \mathbf{a}[\mathbf{F}_{j}(\mathbf{z}^{k}) - \mathbf{F}_{j}(\mathbf{z}^{k-1})] \right) + \mathbf{F}(\mathbf{w}^{k-1}) $ 9: $ \Delta_{k}^{k} = \delta^{k} - t\mathbf{z}^{k} - \mathbf{y}^{k} - \alpha(\mathbf{y}^{k} - \mathbf{y}^{k-1}) $ 10: $ \mathbf{z}^{k-1} = \operatorname{prox}_{n,\mathbf{g}}(\mathbf{z}^{k} + \omega(\mathbf{w}^{k} - \mathbf{z}^{k}) - \eta_{z}\Delta_{z}^{k}) $ 11: $ \mathbf{y}_{c}^{k} = \tau\mathbf{y}^{k} + (1 - \tau)\mathbf{y}_{f}^{k} $ 12: $ \mathbf{x}_{c}^{k} = \tau\mathbf{z}^{k} + (1 - \tau)\mathbf{x}_{f}^{k} $ 13: $ \Delta_{b}^{k} = \nu^{-1}(\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) + \mathbf{z}^{k+1} + \gamma(\mathbf{y}^{k} + \mathbf{x}^{k} + \nu\mathbf{z}^{k}) $ 14: $ \delta^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} \left(\mathbf{F}_{j}(\mathbf{z}^{k+1}) - \mathbf{F}_{j}(\mathbf{w}^{k}) \right) $ 15: $ \Delta_{b}^{k} = \nu^{-1}(\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) + \beta(\mathbf{x}^{k} + \delta^{k+1/2}) $ 16: $ \mathbf{y}^{k+1} = \mathbf{y}^{k} - \eta_{y}\Delta_{y}^{k} $ 17: $ \mathbf{x}^{k+1} = \mathbf{x}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\eta_{z}\Delta_{x}^{k} + m^{k}) $ 18: $ m^{k+1} = \eta_{z}\Delta_{x}^{k} + m^{k} $ 19: $ \mathbf{y}_{f}^{k+1} = \mathbf{x}_{c}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\eta_{z}\Delta_{x}^{k} + m^{k}) $ 20: $ \mathbf{x}_{f}^{k+1} = \mathbf{x}_{c}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\mathbf{y}_{c} + \mathbf{x}_{c}^{k}) $ 21: $ \mathbf{w}^{k+1} = \begin{cases} \mathbf{z}^{k}, & \text{ with probability } p \\ \mathbf{w}^{k}, & \text{ with probability } 1 - p \end{cases} $ 22: end for	7:	$S^{k+1/2} = \{j_{m,1}^{k+1/2}, \dots, j_{m,b}^{k+1/2}\}$	
$\begin{aligned} &+\alpha[\mathbf{F}_{j}(\mathbf{z}^{k})-\mathbf{F}_{j}(\mathbf{z}^{k-1})]\Big)+\mathbf{F}(\mathbf{w}^{k-1})\\ 9: \Delta_{x}^{k}=\delta^{k}-\nu \mathbf{z}^{k}-\mathbf{y}^{k}-\alpha(\mathbf{y}^{k}-\mathbf{y}^{k-1})\\ 10: \mathbf{z}^{k+1}=\operatorname{prox}_{n,\mathbf{g}}(\mathbf{z}^{k}+\omega(\mathbf{w}^{k}-\mathbf{z}^{k})-\eta_{z}\Delta_{z}^{k})\\ 11: \mathbf{y}_{z}^{k}=\tau \mathbf{y}^{k}+(1-\tau)\mathbf{y}_{j}^{k}\\ 12: \mathbf{x}_{z}^{k}=r\mathbf{z}^{k}+(1-\tau)\mathbf{y}_{j}^{k}\\ 13: \Delta_{y}^{k}=\nu^{-1}(\mathbf{y}_{z}^{k}+\mathbf{x}_{z}^{k})+\mathbf{z}^{k+1}+\gamma(\mathbf{y}^{k}+\mathbf{x}^{k}+\nu \mathbf{z}^{k})\\ 14: \delta^{k+1/2}=\frac{1}{k}\sum_{j\in S^{k+1/2}}(\mathbf{F}_{j}(\mathbf{z}^{k+1})-\mathbf{F}_{j}(\mathbf{w}^{k}))\\ 15: \Delta_{z}^{k}=\nu^{-1}(\mathbf{y}_{z}^{k}+\mathbf{x}_{z}^{k})+\beta(\mathbf{x}^{k}+\delta^{k+1/2})\\ 16: \mathbf{y}^{k+1}=\mathbf{x}^{k}-\eta_{z}\Delta_{z}^{k}\\ 17: \mathbf{x}^{k+1}=\mathbf{x}^{k}-(\mathbf{W}_{T}(Tk)\otimes \mathbf{I}_{d})(\eta_{z}\Delta_{x}^{k}+m^{k})\\ 18: m^{k+1}=\eta_{z}\Delta_{x}^{k}+m^{k}\\ &-(\mathbf{W}_{T}(Tk)\otimes \mathbf{I}_{d})(\eta_{z}\Delta_{x}^{k}+m^{k})\\ 19: \mathbf{y}_{z}^{k+1}=\mathbf{x}_{z}^{k}-\theta(\mathbf{W}_{T}(Tk)\otimes \mathbf{I}_{d})(\mathbf{y}_{z}^{k}+\mathbf{x}_{z}^{k})\\ 20: \mathbf{x}_{j}^{k+1}=\mathbf{x}_{z}^{k}-\theta(\mathbf{W}_{T}(Tk)\otimes \mathbf{I}_{d})(\mathbf{y}_{z}^{k}+\mathbf{x}_{z}^{k})\\ 21: \mathbf{w}^{k+1}=\left\{\mathbf{z}_{z}^{k}, \text{with probability }p\\ \mathbf{w}^{k}, \text{with probability }1-p\\ 22: \text{end for} \end{aligned}$	8:	$\delta^k = \frac{1}{h} \sum_{j \in S^k} \left(\mathbf{F}_j(\mathbf{z}^k) - \mathbf{F}_j(\mathbf{w}^{k-1}) \right)$	
$\begin{array}{l} & (\mathbf{x}_{1}) = \mathbf{x}_{1}(\mathbf{z}_{1}) + \mathbf{x}_{1}(\mathbf{w}_{1}) \\ & (\mathbf{x}_{2}) = \mathbf{x}_{1}(\mathbf{z}_{1}) + \mathbf{x}_{1}(\mathbf{w}_{1}) \\ & (\mathbf{x}_{2}) = \mathbf{x}_{1}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{x}_{2}) \\ & (\mathbf{z}_{2}) = \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) \\ & (\mathbf{z}_{2}) = \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) \\ & (\mathbf{z}_{2}) = \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) \\ & (\mathbf{z}_{2}) = \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) \\ & (\mathbf{z}_{2}) = \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) + \mathbf{x}_{2}(\mathbf{z}_{2}) \\ & (\mathbf{z}_{2}) = \mathbf{z}_{2} \\ & (\mathbf{z}_{2}) = \mathbf{z}_{2$		$+\alpha[\mathbf{F}_{k}(\mathbf{z}^{k}) - \mathbf{F}_{k}(\mathbf{z}^{k-1})] + \mathbf{F}(\mathbf{w}^{k-1})$	
9: $\Delta_{k}^{k} = \delta^{k} - \nu z^{k} - y^{k} - \alpha(y^{k} - y^{k-1})$ 10: $z^{k+1} = \operatorname{prov}_{\eta_{z}z}(z^{k} + \omega(w^{k} - z^{k}) - \eta_{z}\Delta_{z}^{k})$ 11: $y_{c}^{k} = \tau y^{k} + (1 - \tau)y_{f}^{k}$ 12: $x_{c}^{k} = \tau x^{k} + (1 - \tau)x_{f}^{k}$ 13: $\Delta_{y}^{k} = \nu^{-1}(y_{c}^{k} + x_{c}^{k}) + z^{k+1} + \gamma(y^{k} + x^{k} + \nu z^{k})$ 14: $\delta^{k+1/2} = \frac{1}{k}\sum_{j \in S^{k+1/2}} (F_{j}(z^{k+1}) - F_{j}(w^{k}))$ 15: $\Delta_{x}^{k} = \nu^{-1}(y_{c}^{k} + x_{c}^{k}) + \beta(x^{k} + \delta^{k+1/2})$ 16: $y^{k+1} = y^{k} - \eta_{a}\Delta_{y}^{k}$ 17: $x^{k+1} = x^{k} - (W_{T}(T^{k}) \otimes I_{d})(\eta_{z}\Delta_{x}^{k} + m^{k})$ 18: $m^{k+1} = \eta_{z}\Delta_{x}^{k} + m^{k}$ $-(W_{T}(T^{k}) \otimes I_{d})(\eta_{z}\Delta_{x}^{k} + m^{k})$ 19: $y_{f}^{k+1} = x_{c}^{k} - \theta(W_{T}(T^{k}) \otimes I_{d})(y_{c}^{k} + x_{c}^{k})$ 21: $w^{k+1} = \begin{cases} z^{k}, & \text{with probability } p \\ w^{k}, & \text{with probability } 1 - p \end{cases}$ 22: end for		$+\alpha [\mathbf{r}_j(\mathbf{z}_j) - \mathbf{r}_j(\mathbf{z}_j)] + \mathbf{r}(\mathbf{w}_j)$	
$ \begin{array}{llllllllllllllllllllllllllllllllllll$	9:	$\Delta_{\mathbf{z}}^{\kappa} = \delta^{\kappa} - \nu \mathbf{z}^{\kappa} - \mathbf{y}^{\kappa} - \alpha (\mathbf{y}^{\kappa} - \mathbf{y}^{\kappa-1})$	
$ \begin{array}{llllllllllllllllllllllllllllllllllll$	10:	$\mathbf{z}^{\kappa+1} = \operatorname{prox}_{\eta_z \mathbf{g}}(\mathbf{z}^{\kappa} + \omega(\mathbf{w}^{\kappa} - \mathbf{z}^{\kappa}) - \eta_z \Delta_z^{\kappa})$	
$ \begin{array}{ll} 12: & \mathbf{x}_{k}^{k} = \tau \mathbf{x}^{k} + (1 - \tau) \mathbf{x}_{k}^{k} \\ 13: & \Delta_{y}^{k} = \nu^{-1} (\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) + \mathbf{z}^{k+1} + \gamma (\mathbf{y}^{k} + \mathbf{x}^{k} + \nu \mathbf{z}^{k}) \\ 14: & \delta^{k+1/2} = \frac{1}{k} \sum_{j \in S^{k+1/2}} \left(\mathbf{F}_{j} (\mathbf{z}^{k+1}) - \mathbf{F}_{j} (\mathbf{w}^{k}) \right) \\ 15: & \Delta_{k}^{k} = \nu^{-1} (\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) + \beta (\mathbf{x}^{k} + \delta^{k+1/2}) \\ 16: & \mathbf{y}^{k+1} = \mathbf{y}^{k} - \eta_{\mu} \Delta_{y}^{k} \\ 17: & \mathbf{x}^{k+1} = \mathbf{x}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\eta_{z} \Delta_{x}^{k} + m^{k}) \\ 18: & m^{k+1} = \mathbf{y}_{c}^{k} + \pi (\mathbf{y}^{k+1} - \mathbf{y}^{k}) \\ 19: & \mathbf{y}_{f}^{k+1} = \mathbf{x}_{c}^{k} - \theta (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\eta_{z} \Delta_{x}^{k} + m^{k}) \\ 19: & \mathbf{x}_{f}^{k+1} = \mathbf{x}_{c}^{k} - \theta (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\mathbf{y}_{c} + \mathbf{x}_{c}^{k}) \\ 21: & \mathbf{w}^{k+1} = \begin{cases} \mathbf{z}^{k}, & \text{with probability } p \\ \mathbf{w}^{k}, & \text{with probability } 1 - p \end{cases} \end{array} $	11:	$\mathbf{y}_c^k = au \mathbf{y}^k + (1- au) \mathbf{y}_f^k$	
$ \begin{array}{ll} 13: & \Delta_y^k = \nu^{-1}(\mathbf{y}_c^k + \mathbf{x}_c^k) + \mathbf{z}^{k+1} + \gamma(\mathbf{y}^k + \mathbf{x}^k + \nu \mathbf{z}^k) \\ 14: & \delta^{k+1/2} = \frac{1}{b} \sum_{j \in S^{k+1/2}} (\mathbf{F}_j(\mathbf{z}^{k+1}) - \mathbf{F}_j(\mathbf{w}^k)) \\ & + \mathbf{F}(\mathbf{w}^k) \\ 15: & \Delta_x^k = \nu^{-1}(\mathbf{y}_c^k + \mathbf{x}_c^k) + \beta(\mathbf{x}^k + \delta^{k+1/2}) \\ 16: & \mathbf{y}^{k+1} = \mathbf{y}^k - \eta_y \Delta_y^k \\ 17: & \mathbf{x}^{k+1} = \mathbf{x}^k - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + m^k) \\ 18: & m^{k+1} = \eta_z \Delta_x^k + m^k \\ & -(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + m^k) \\ 19: & \mathbf{y}_y^{k+1} = \mathbf{y}_c^k - \mathbf{f}(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\mathbf{y}_c + \mathbf{x}_c^k) \\ 20: & \mathbf{x}_j^{k+1} = \mathbf{x}_c^k - \theta(\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\mathbf{y}_c^k + \mathbf{x}_c^k) \\ 21: & \mathbf{w}^{k+1} = \begin{cases} \mathbf{z}^k, & \text{with probability } p \\ \mathbf{w}^k, & \text{with probability } 1 - p \end{cases} \\ 22: & \text{end for} \end{array} $	12:	$\mathbf{x}_{c}^{k} = \tau \mathbf{x}^{k} + (1 - \tau) \mathbf{x}_{f}^{k}$	
$ \begin{array}{ll} 14: & \delta^{k+1/2} = \int_{0}^{\infty} \sum_{j \in S^{k+1/2}} \left(\mathbf{F}_{j}(\mathbf{z}^{k+1}) - \mathbf{F}_{j}(\mathbf{w}^{k}) \right) & + \mathbf{F}(\mathbf{w}^{k}) \\ & + \mathbf{F}(\mathbf{w}^{k}) \\ 15: & \Delta_{\mathbf{x}}^{k} = \nu^{-1}(\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) + \beta(\mathbf{x}^{k} + \delta^{k+1/2}) \\ 16: & \mathbf{y}^{k+1} = \mathbf{y}^{k} - \eta_{\mu} \Delta_{\mathbf{y}}^{k} \\ 17: & \mathbf{x}^{k+1} = \mathbf{x}^{k} - (\mathbf{W}_{T}(\mathbf{T}k) \otimes \mathbf{I}_{d})(\eta_{x} \Delta_{\mathbf{x}}^{k} + m^{k}) \\ 18: & m^{k+1} = \eta_{x} \Delta_{\mathbf{x}}^{k} + m^{k} \\ & - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\eta_{x} \Delta_{\mathbf{x}}^{k} + m^{k}) \\ 19: & \mathbf{y}_{f}^{k+1} = \mathbf{x}_{c}^{k} - \theta(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) \\ 20: & \mathbf{x}_{f}^{k+1} = \mathbf{x}_{c}^{k} - \theta(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) \\ 21: & \mathbf{w}^{k+1} = \left\{ \mathbf{z}_{c}^{k}, & \text{with probability } p \\ \mathbf{w}^{k}, & \text{with probability } 1 - p \\ 22: & \text{end for} \end{array} \right\} \leqslant \mathcal{O} \diamond \mathcal{O} $	13:	$\Delta_{\mu}^{k} = \nu^{-1} (\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) + \mathbf{z}^{k+1} + \gamma (\mathbf{y}^{k} + \mathbf{x}^{k} + \nu \mathbf{z}^{k})$	
$ \begin{array}{l} \mathbf{b} = \sum_{k=0}^{k-1} (\mathbf{y}_{k}^{k} + \mathbf{x}_{k}^{k}) + \beta(\mathbf{x}^{k} + \delta^{k+1/2}) \\ \mathbf{15:} \mathbf{b}_{k}^{k} = \mathbf{y}^{k} - \eta_{\mu} \Delta_{y}^{k} \\ \mathbf{17:} \mathbf{x}^{k+1} = \mathbf{x}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\eta_{z} \Delta_{x}^{k} + m^{k}) \\ \mathbf{18:} m^{k+1} = \mathbf{x}_{z}^{k} \Delta_{x}^{k} + m^{k} \\ - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\eta_{z} \Delta_{x}^{k} + m^{k}) \\ \mathbf{19:} \mathbf{y}_{f}^{k+1} = \mathbf{x}_{c}^{k} - \sigma(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\eta_{z} \Delta_{x}^{k} + m^{k}) \\ \mathbf{20:} \mathbf{x}_{f}^{k+1} = \mathbf{x}_{c}^{k} - \sigma(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\mathbf{y}_{c}^{c} + \mathbf{x}_{c}^{k}) \\ \mathbf{21:} \mathbf{w}^{k+1} = \left\{ \mathbf{z}_{c}^{k}, \text{with probability } p \\ \mathbf{w}^{k}, \text{with probability } 1 - p \\ \mathbf{22:} \text{end for} \end{array} \right\} \boldsymbol{\in} \boldsymbol{\mathbb{R}} $	14:	$\delta^{k+1/2} = \frac{1}{k} \sum_{i \in S^{k+1/2}} \left(\mathbf{F}_i(\mathbf{z}^{k+1}) - \mathbf{F}_i(\mathbf{w}^k) \right)$	
$ \begin{array}{ll} 15: & \Delta_{x}^{k} = \nu^{-1}(\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) + \beta(\mathbf{x}^{k} + \delta^{k+1/2}) \\ 16: & \mathbf{y}^{k+1} = \mathbf{y}^{k} - \eta_{p} \Delta_{y}^{k} \\ 17: & \mathbf{x}^{k+1} = \mathbf{x}^{k} - (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\eta_{z} \Delta_{x}^{k} + m^{k}) \\ 18: & m^{k+1} = \eta_{z} \Delta_{x}^{k} + m^{k} \\ -(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\eta_{z} \Delta_{x}^{k} + m^{k}) \\ 19: & \mathbf{y}_{f}^{k+1} = \mathbf{y}_{c}^{k} + r(\mathbf{y}^{k+1} - \mathbf{y}^{k}) \\ 20: & \mathbf{x}_{f}^{k+1} = \mathbf{x}_{c}^{k} - \theta(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\mathbf{y}_{c} \Delta_{x}^{k} + m^{k}) \\ 21: & \mathbf{w}^{k+1} = \left\{ \mathbf{z}_{c}^{k}, & \text{with probability } p \\ \mathbf{w}^{k}, & \text{with probability } 1 - p \\ 22: & \text{end for} \end{array} \right. $		$+\mathbf{F}(\mathbf{w}^k)$	
$ \begin{array}{ll} 16: & \mathbf{y}^{k+1}_{k+1} = \mathbf{y}^{k-c}_{k-1} - \eta_{y} \boldsymbol{\xi}^{k}_{y} \\ 16: & \mathbf{y}^{k+1}_{k+1} = \mathbf{x}^{k}_{k-1} (\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\eta_{x} \Delta^{k}_{x} + m^{k}) \\ 17: & \mathbf{x}^{k+1} = \eta_{x} \Delta^{k}_{x} + m^{k} \\ & -(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\eta_{x} \Delta^{k}_{x} + m^{k}) \\ 19: & \mathbf{y}^{k+1}_{f} = \mathbf{y}^{k}_{c} + \tau(\mathbf{y}^{k+1} - \mathbf{y}^{k}) \\ 20: & \mathbf{x}^{k+1}_{f} = \mathbf{x}^{k}_{c} - \theta(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}) (\mathbf{y}^{k}_{c} + \mathbf{x}^{k}_{c}) \\ 21: & \mathbf{w}^{k+1}_{t} = \left\{ \mathbf{z}^{k}, & \text{with probability } p \\ \mathbf{w}^{k}, & \text{with probability } 1 - p \end{array} \right. $	15:	$\Delta^k = \nu^{-1} (\mathbf{v}^k + \mathbf{x}^k) + \beta (\mathbf{x}^k + \delta^{k+1/2})$	
$\begin{array}{l} 17: \mathbf{x}^{k+1} = \mathbf{x}^k - (\mathbf{W}_T(Tk) \otimes \mathbf{I}_d)(\eta_z \Delta_x^k + m^k) \\ 18: m^{k+1} = \eta_z \Delta_z^k + m^k \\ \qquad $	16:	$\mathbf{v}^{k+1} = \mathbf{v}^k - n \Delta^k$	
$\begin{array}{llllllllllllllllllllllllllllllllllll$	17.	$\mathbf{y}^{k+1} = \mathbf{y}^k = (\mathbf{W}_m(Tk) \otimes \mathbf{I}_n)(n \ \Delta^k + m^k)$	
10. $ \mathbf{m}^{k} = \mathbf{y}_{c}\mathbf{x}_{s}\mathbf{m}^{k} - \mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d}(\eta_{x}\Delta_{x}^{k} + m^{k}) $ 19. $ \mathbf{y}_{f}^{k+1} = \mathbf{y}_{c}^{k} + \tau(\mathbf{y}^{k+1} - \mathbf{y}^{k}) $ 20. $ \mathbf{x}_{f}^{k+1} = \mathbf{x}_{c}^{k} - \theta(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k}) $ 21. $ \mathbf{w}^{k+1} = \begin{cases} \mathbf{z}^{k}, & \text{with probability } p \\ \mathbf{w}^{k}, & \text{with probability } 1 - p \end{cases} $ 22. end for	18.	$\mathbf{x} = \mathbf{x} - (\mathbf{v} T(1 k) \otimes \mathbf{I}_d)(\eta_x \Delta_x + m)$ $m^{k+1} - n \Delta^k + m^k$	
19: $\mathbf{y}_{l}^{k+1} = \mathbf{y}_{c}^{k} + \tau(\mathbf{y}^{k+1} - \mathbf{y}^{k})$ 20: $\mathbf{x}_{l}^{k+1} = \mathbf{x}_{c}^{k} - \theta(\mathbf{W}_{T}(Tk) \otimes \mathbf{I}_{d})(\mathbf{y}_{c}^{k} + \mathbf{x}_{c}^{k})$ 21: $\mathbf{w}_{c}^{k+1} = \begin{cases} \mathbf{z}^{k}, & \text{with probability } p \\ \mathbf{w}^{k}, & \text{with probability } 1 - p \end{cases}$ 22: end for	10.	$-(\mathbf{W}_{x}(Tk)\otimes\mathbf{L})(n\;\Delta^{k}+m^{k})$	
$\begin{array}{llllllllllllllllllllllllllllllllllll$	10.	$\mathbf{v}^{k+1} - \mathbf{v}^k + \tau(\mathbf{v}^{k+1} - \mathbf{v}^k)$	
20: $\mathbf{x}_{f} = \mathbf{x}_{c}^{-b} \operatorname{GW}(\mathbf{x}) \otimes 4_{d} (\mathbf{y}_{c} + \mathbf{x}_{c})$ 21: $\mathbf{w}^{k+1} = \begin{cases} \mathbf{z}^{k}, & \text{with probability } p \\ \mathbf{w}^{k}, & \text{with probability } 1 - p \end{cases}$ 22: end for	no.	$\mathbf{y}_f = \mathbf{y}_c + \mathbf{y}(\mathbf{y} + \mathbf{y})$ $\mathbf{x}_c^{k+1} = \mathbf{x}_c^k + \mathbf{h}(\mathbf{y}_c + \mathbf{y}_c^k) \otimes \mathbf{I}(\mathbf{x}_c^k + \mathbf{x}_c^k)$	
21: $\mathbf{w}^{k+1} = \begin{cases} \mathbf{z}^n, & \text{with probability } p \\ \mathbf{w}^k, & \text{with probability } 1-p \end{cases}$ 22: end for	20:	$\mathbf{x}_f = \mathbf{x}_c - \mathbf{v} (\mathbf{v} \mathbf{v}_T (1 \ \kappa) \otimes \mathbf{I}_d) (\mathbf{y}_c + \mathbf{x}_c)$	
$[\mathbf{w}^{\kappa}, \text{ with probability } 1-p$ 22: end for $ \langle \mathbb{P} \rangle \land \mathbb{P} \land $	21:	$\mathbf{w}^{k+1} = \begin{cases} \mathbf{z}^n, & \text{with probability } p \end{cases}$	
22: end for		$(\mathbf{w}^{\kappa}, \text{ with probability } 1-p$	-
	22:	end for	

Aleksandr Beznosikov

On Distributed Variational Inequalities

17 January 2024

୬ < ୍ୟ 33 / 47

Quantization and compression

Definition (Quantization)

A stochastic operator $Q: \mathbb{R}^d \to \mathbb{R}^d$ is called *quantization* if there exists a constant $q \ge 1$ such that

Q(z) = z, $\mathbb{E} \|Q(z)\|^2 \leq q \|z\|^2$, $\forall z \in \mathbb{R}^d$.

Expected/average compression (how much less the compressed vector takes up in memory): $\beta^{-1} \stackrel{\text{def}}{=} \frac{\mathbb{E} ||Q(z)||_{\text{bits}}}{||z|||_{\text{bits}}}$. Note that $\beta \ge 1$. Examples: random selection of coordinates. Definition (Compression)

(Stochastic) operator $C : \mathbb{R}^d \to \mathbb{R}^d$ is called *compression* if there exists $\delta \ge 1$ such that

$$\mathbb{E} \|C(z)-z\|^2 \leq (1-1/\delta)\|z\|^2, \quad \forall z\in \mathbb{R}^d.$$

Expected/average compression (how much less the compressed vector occupies in memory): $\beta^{-1} \stackrel{\text{def}}{=} \frac{\mathbb{E} \|C(z)\|_{\text{bits}}}{\|z\|_{\text{bits}}}$. Отметим, что $\beta \ge 1$. Examples: Greedy choice of coordinates, low-rank decompositions, $z = 200^{\circ}$ • For example, quantized extragradient method

$$z^{k+1/2} = z^k - \gamma \cdot \frac{1}{M} \sum_{m=1}^M Q_1(F_m(z^k))$$

$$z^{k+1} = z^k - \gamma \cdot \frac{1}{M} \sum_{m=1}^M Q_2(F_m(z^{k+1/2}))$$

• Different Q are taken here. In fact it can be the same operator in terms of physics, but with different or the same randomness.

- Good idea: variance reduction.
- Non-distributed problem:

$$\min_{z\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^n f_i(z)$$

And the next method:

$$z^{k+1} = z^k - \gamma \cdot (\nabla f_{i_k}(z^k) - \nabla f_{i_k}(w^k) + \nabla f(w^k))$$
$$w^{k+1} = \begin{cases} w^k & \text{with prob.} \quad \tau\\ z^k & \text{with prob} \quad 1 - \tau \end{cases}$$

3

MASHA1

Algorithm 1 MASHA1

Parameters: Stepsize $\gamma > 0$, parameter $\tau \in (0; 1)$, number of iterations K. Initialization: Choose $z^0 = w^0 \in \mathbb{Z}$. Devices send $F_m(w^0)$ to server and get $F(w^0)$ for $k = 0, 1, 2, \dots, K - 1$ 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 $z^{k+1} = z^{k+1/2} - \gamma q^k$ 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

Image: A marked and A mar A marked and A 3

• Convergence of MASHA1 in transmitted information:

$$\mathcal{O}([1+\sqrt{rac{1}{M}+rac{1}{eta}}\cdotrac{L}{\mu}]\lograc{1}{arepsilon});$$

• Extragradient without quantization:

$$\mathcal{O}(\frac{L}{\mu}\log\frac{1}{\varepsilon});$$

• Quantization gives boost.

• Consider the following distributed problem with M = 3, d = 3 and local functions:

 $f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|$

where a = (-3, 2, 2), b = (2, -3, 2) is c = (2, 2, -3).

• Consider the following distributed problem with M = 3, d = 3 and local functions:

 $f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|$

where a = (-3, 2, 2), b = (2, -3, 2) v c = (2, 2, -3).

• Question: where is her optimum?

• Consider the following distributed problem with M = 3, d = 3 and local functions:

 $f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|$

where a = (-3, 2, 2), b = (2, -3, 2) is c = (2, 2, -3).

• Question: where is her optimum? (0, 0, 0).

• Consider the following distributed problem with M = 3, d = 3 and local functions:

 $f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|$

where a = (-3, 2, 2), b = (2, -3, 2) is c = (2, 2, -3).

- Question: where is her optimum? (0,0,0).
- Let the starting point $w_0 = (t, t, t)$ for some t > 0. Then the local gradients are:

 $\nabla f_1(w_0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(w_0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(w_0) = \frac{t}{2}(9, 9, -11).$

• Question: what will the QGD (gradient descent with compressions) step look like if we use *Top1* compression?

3

• Consider the following distributed problem with M = 3, d = 3 and local functions:

 $f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \ f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|$

where a = (-3, 2, 2), b = (2, -3, 2) is c = (2, 2, -3).

- Question: where is her optimum? (0,0,0).
- Let the starting point $w_0 = (t, t, t)$ for some t > 0. Then the local gradients are:

 $\nabla f_1(w_0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(w_0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(w_0) = \frac{t}{2}(9, 9, -11).$

• Question: what will the QGD (gradient descent with compressions) step look like if we use *Top1* compression?

$$w_1 = (t,t,t) + \gamma \cdot rac{11}{6}(t,t,t) = \left(1 + rac{11\gamma}{6}\right)w_0.$$

• We move away from the solution geometrically for any $\gamma > 0$.

• Let's try to remember what we didn't pass on in the communication process:

$$e_{1,m} = 0 + \gamma F_m(z_0) - C(0 + \gamma F_m(z_0)).$$

• And add this to future parcels:

$$C(e_{1,m} + \gamma F_m(z_1))$$

• In an arbitrary iteration, it is written as follows:

Parcel:
$$C(e_{k,m} + \gamma F_m(w_k))$$
,
 $e_{k+1,m} = e_{k,m} + \gamma F_m(z_k) - C(e_{k,m} + \gamma F_m(z_k))$

• This technique is called error compensation (error feedback).

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 $\ddot{F}(w^0)$ for $k = 0, 1, 2, \dots, K - 1$ do for each device m in parallel do $z^{k+1/2} = \tau z^k + (1-\tau)w^k - \gamma F(w^k)$ Sends $q_m^k = C_m^{\text{dev}}(\gamma F_m(z^{k+1/2}) - \gamma F_m(w^k) + e_m^k)$ to server $e_m^{k+1} = e_m^k + \gamma F_m(z^{k+1/2}) - \gamma F_m(w^k) - q_m^k$ 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 $z^{k+1} = z^{k+1/2} - \gamma q^k$ 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

 ▶ < ≧ ▶ < ≧ ▶ </td>
 ≧
 ✓

 17 January 2024
 41

• Convergence of MASHA2 in transmitted information:

$$\mathcal{O}([1+\frac{L}{\mu}]\log\frac{1}{\varepsilon});$$

• Convergence of MASHA1 in transmitted information:

$$\mathcal{O}([1+\sqrt{\frac{1}{M}+\frac{1}{\beta}}\cdot\frac{L}{\mu}]\log\frac{1}{\varepsilon});$$

• Extragradient without quantization:

$$\mathcal{O}(\frac{L}{\mu}\log\frac{1}{\varepsilon});$$

• Compression does not give boost in theory.

• Convergence of MASHA2 in transmitted information:

$$\mathcal{O}([1+\frac{L}{\mu}]\log\frac{1}{\varepsilon});$$

• Convergence of MASHA1 in transmitted information:

$$\mathcal{O}([1+\sqrt{\frac{1}{M}+\frac{1}{\beta}}\cdot\frac{L}{\mu}]\log\frac{1}{\varepsilon});$$

• Extragradient without quantization:

$$\mathcal{O}(\frac{L}{\mu}\log\frac{1}{\varepsilon});$$

• Compression does not give boost in theory.

• The operators $\{F_m\}$ is δ -related in mean. It means that for any j operators $\{F_m - F_j\}$ is δ -Lipschitz continuous in mean, i.e. for all u, v we have

$$\frac{1}{M}\sum_{m=1}^{M}\|F_m(u)-F_j(u)-F_m(v)+F_j(v)\|^2 \leq \delta^2\|u-v\|^2.$$

• Comes from hessian (second derivatives similarity):

$$\|\nabla^2 f_m(z) - \nabla^2 f_j(z)\| \leq \delta$$

 Natural assumption since from Hoeffding: δ = Õ(L/√b) or even δ = Õ(L/b), where b is the number of local data points on each of the devices. • Mirror descent for minimization problems:

$$z^{k+1} = \arg\min_{w \in \mathbb{R}^d} \left(\gamma \langle \nabla f(z^k), w \rangle + V(w, z^k) \right),$$

where V(x, y) is the Bregman divergence generated by the function $\varphi(x)$ (here we need to require that f_1 is convex):

$$\varphi(x) = f_1(x) + \frac{\delta}{2} \|x\|^2.$$

The function f_1 is stored on the server.

• Mirror descent for minimization problems:

$$z^{k+1} = \arg\min_{w \in \mathbb{R}^d} \left(\gamma \langle
abla f(z^k), w
angle + V(w, z^k)
ight),$$

where V(x, y) is the Bregman divergence generated by the function $\varphi(x)$ (here we need to require that f_1 is convex):

$$\varphi(x) = f_1(x) + \frac{\delta}{2} \|x\|^2.$$

The function f_1 is stored on the server.

 What is the number of communications that occur in K iterations of such a mirror descent? K of communications (the number of gradient counts ∇f), computing arg min requires only computations on the server.

Theorem (convergence for data similarity)

Let f be strongly convex, f_i be convex, and ℓ be smooth, and $\varphi(w) = f_1(w) + \delta ||w||^2$, then mirror descent with step $\gamma = 1$ converges and is satisfied:

$$V(w^*, w_K) \leq \left(1 - rac{\mu}{\mu + 2\delta}
ight)^K V(w^*, w_0).$$

Theorem (convergence for data similarity)

Let f be strongly convex, f_i be convex, and ℓ be smooth, and $\varphi(w) = f_1(w) + \delta ||w||^2$, then mirror descent with step $\gamma = 1$ converges and is satisfied:

$$V(w^*, w_K) \leq \left(1 - rac{\mu}{\mu + 2\delta}
ight)^K V(w^*, w_0).$$

 It means that if we need to achieve an accuracy ε (V(w^{*}, w_K) ~ ε), then we need to

$$\mathcal{K} = \left(\left[1 + \frac{\delta}{\mu} \right] \log \frac{V(w^*, w_0)}{\varepsilon} \right)$$
 communications.

Similarity + compression

Double kill of two ideas

Algorithm 1 Three Pillars Algorithm **Parameters:** stepsizes γ and η , momentum τ , probability $p \in (0, 1]$, number of local steps H: Initialization: Choose $z^0 = m^0 = (x^0, y^0) \in \mathbb{Z}$; 1: for $k = 0, 1, \dots, K - 1$ do Server takes $u_0^k = z^k$: 2: 3: for $t = 0, 1, \dots, H - 1$ do Server computes $u_{t+1/2}^k = \text{proj}_{\mathcal{Z}}[u_t^k - \eta(F_1(u_t^k) - F_1(m^k) + F(m^k) + \frac{1}{2}(u_t^k - z^k - \tau(m^k - z^k)))];$ 4: Server updates $u_{t+1}^k = \operatorname{proj}_{\mathcal{Z}}[u_t^k - \eta(F_1(u_{t+1/2}^k) - F_1(m^k) + F(m^k) + \frac{1}{2}(u_{t+1/2}^k - z^k - \tau(m^k - z^k)))];$ 5: 6. end for Server broadcasts u_H^k and $F_1(u_H^k)$ to devices; 7: 8: Devices in parallel compute $Q_i(F_i(m^k) - F_1(m^k) - F_i(u_H^k) + F_1(u_H^k))$ and send to server; Server updates $z^{k+1} = \text{proj}_{z} [u_{H}^{k} + \gamma \cdot \frac{1}{n} \sum_{i=1}^{n} Q_{i} (F_{i}(m^{k}) - F_{1}(m^{k}) - F_{i}(u_{H}^{k}) + F_{1}(u_{H}^{k}))];$ 9. Server updates $m^{k+1} = \begin{cases} z^k, & \text{with probability } p, \\ m^k, & \text{with probability } 1-p, \end{cases}$ 10: if $m^{k+1} = z^k$ then 11: Server broadcasts m^{k+1} to devices: 12. 13: Devices in parallel compute $F_i(m^k)$ and send to server; Server computes $F(m^{k+1}) = \frac{1}{n} \sum_{i=1}^{n} F_i(m^{k+1});$ 14: end if 15. 16: end for

э

47 / 47

イロト イポト イヨト イヨト