Stochastic and Variance-Reduced Monotone Operator Splitting

Patrick Johnstone
Brookhaven National Laboratory, NY, USA
OWOS, February 7th, 2022

Collaborators: Jonathan Eckstein (Rutgers), Thomas Flynn (BNL), Shinjae Yoo (BNL)

Talk Overview

- Focus: projective splitting method for solving monotone inclusions
- Develop new stochastic version
  - almost sure iterate convergence + convergence rate
- Develop variance-reduced version
  - same rates as deterministic methods but with improved constants

Convex Optimization I

Consider

$$\min_{x\in\mathbb{R}^d}\left\{\sum_{i=1}^{n} f_i(G_i x) + h(x)\right\}$$

where

- $f_i : \mathbb{R}^{d_i} \to \mathbb{R} \cup \{+\infty\}$ are convex, closed, proper
- $G_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$ are linear
- $h : \mathbb{R}^d \to \mathbb{R}^d$ is convex and smooth
- multiple regularizers and constraints
  - $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$, else $+\infty$
- certain regularizers such as total variation (TV)

Convex Optimization II

$$\min_{x\in\mathbb{R}^d}\left\{\sum_{i=1}^{n} f_i(G_i x) + h(x)\right\}$$

- (Fermat's Rule): first order sufficient[1] condition

$$\exists\left\{\begin{array}{c} w_1 \in \partial f_1(G_1 x)\\ w_2 \in \partial f_2(G_2 x)\\ \vdots \\ w_n \in \partial f_n(G_n x)\end{array}\right\} \quad : \quad 0 = \sum_{i=1}^{n} G_i{}^\top w_i + \nabla h(x)$$

- With Minkowski summation $\left(A + B = \{a+b : a \in A, b \in B\}\right)$, write as

$$0 \in \sum_{i=1}^{n} G_i{}^\top \partial f_i(G_i x) + \nabla h(x)$$

[1]and necessary under additional Slater-like conditions

Monotone Inclusions I

Instead of

$$0 \in \sum_{i=1}^{n} G_i^\top \partial f_i(G_i x) + \nabla h(x)$$

solve

$$\text{Find } z \in \mathbb{R}^d \quad \text{s.t.} \quad 0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$

where

- $A_i : \mathbb{R}^{d_i} \to 2^{\mathbb{R}^{d_i}}$ are maximal-monotone

$$\forall x_1, x_2 \in \mathbb{R}^d, y_1 \in Ax_1, y_2 \in Ax_2 : \\ \langle y_1 - y_2, x_1 - x_2 \rangle \geq 0$$

- $G_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$ are linear
- $B : \mathbb{R}^d \to \mathbb{R}^d$ is monotone and single-valued and *continuous* to some degree

Monotone Inclusions II

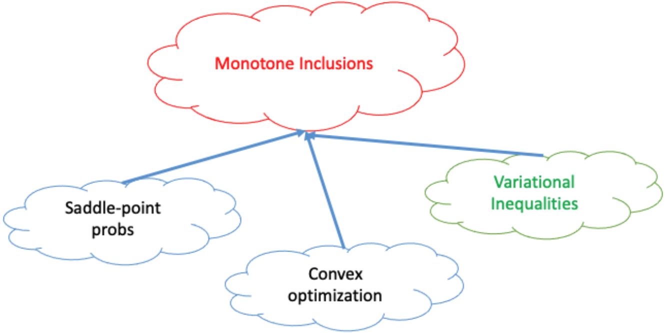
$$\text{Find } z \in \mathbb{R}^d \quad \text{s.t.} \quad 0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$

where

- $A_i : \mathbb{R}^{d_i} \to 2^{\mathbb{R}^{d_i}}$ are maximal-monotone
- $G_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$ are linear
- B is monotone and single-valued

$$\text{Find } (z, w_1, \ldots, w_n) : \left\{ \begin{array}{c} w_1 \in A_1(G_1 z) \\ w_2 \in A_2(G_2 z) \\ \vdots \\ w_n \in A_n(G_n z) \end{array} \right\} \quad : \quad 0 = \sum_{i=1}^{n} G_i^\top w_i + Bz$$

Why Care About Monotone Inclusions?



- Umbrella problem
- Same algorithms/analysis for all problems

Saddle-point Problems

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^m}\left\{\sum_{i=1}^{n}(f_i(R_i x) - g_i(H_i y)) + F(x,y)\right\}$$

- f_i, g_i are convex, F is convex-concave and smooth
- first-order sufficient conditions

$$0 \in \sum_{i=1}^{n}\begin{bmatrix} R_i^\top \partial f_i(R_i x) \\ H_i^\top \partial g_i(H_i y) \end{bmatrix} + \begin{bmatrix} \nabla_x F(x,y) \\ -\nabla_y F(x,y) \end{bmatrix}$$

Set

$$z = \begin{bmatrix} x \\ y \end{bmatrix} \quad G_i = \begin{bmatrix} R_i & 0 \\ 0 & H_i \end{bmatrix} \quad A_i = \partial f_i \times \partial g_i \quad Bz = \begin{bmatrix} \nabla_x F(x,y) \\ -\nabla_y F(x,y) \end{bmatrix}$$

- B is monotone (Rockafellar 1970), A_i is maximal-monotone

$$\text{Find } z \in \mathbb{R}^{d+m} \quad \text{s.t.} \quad 0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$

Operator Splitting Algorithms

$$\text{Find } z \in \mathbb{R}^d \quad \text{s.t.} \quad 0 \in \sum_{i=1}^{n} G_i{}^\top A_i(G_i z) + Bz$$

Solve problem (i.e. converge to a solution) using

- direct evaluation for single-valued B (a.k.a. forward step)
- resolvents for set-valued A_i (a.k.a. backward step)
- direct and transpose application for linear G_i
- Basic vector operations (norms, inner products, vector addition, scalar multiplication)

Resolvents

$$J_A \triangleq (I + A)^{-1}$$

- (Minty's theorem): For maximal-monotone A, resolvent is single-valued and defined everywhere
- $A = \partial f$ reduces to *proximal operator*

$$J_{\partial f} x_0 = \text{prox}_f(x_0) \triangleq \arg\min_x \left\{ f(x) + \frac{1}{2}\|x - x_0\|^2 \right\}$$

- Example: ℓ_1-norm

$$\text{prox}_{\|\cdot\|_1}(x)_i = \begin{cases} x_i - 1 & : x_i \geq 1 \\ x_i + 1 & : x_i \leq -1 \\ 0 & : \text{ else} \end{cases}$$

- Constraints: $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$, else $+\infty$, then $\text{prox}_{\iota_{\mathcal{C}}} = \text{proj}_{\mathcal{C}}$

Handling Linear Composition: ${G_i}^\top A_i(G_i z)$ and $f_i(G_i x)$

- Example: Vector TV norm

$$x = (x^1, \dots, x^d), \quad f(x) = \sum_{i=1}^{d-1} |x^{i+1} - x^i|$$

- Prox no closed form $\text{prox}_f(y) = \arg\min_x \left\{ f(x) + \frac{1}{2}\|x - y\|^2 \right\}$ how can operator splitting algorithms process?

Handling Linear Composition: $G_i^\top A_i(G_i z)$ and $f_i(G_i x)$

- Example: Vector TV norm

$$x = (x^1, \ldots, x^d), \quad f(x) = \sum_{i=1}^{d-1} |x^{i+1} - x^i|$$

- Prox no closed form $\text{prox}_f(y) = \arg\min_x \left\{ f(x) + \frac{1}{2}\|x - y\|^2 \right\}$ how can operator splitting algorithms process?
- Rewrite as

$$\tilde{f}(x) = \|Gx\|_1, \text{ where } G = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1 \end{bmatrix}$$

- Operator splitting algorithms process $f(Gx)$ via $prox_f$, G and $G^\top$. For example projective splitting (Alotaibi 2013)

$$x^k = \text{prox}_{\tau f}(Gz^k + \tau w^k) \text{ and } G^\top y^k$$

- other applications: overlapping group lasso, graph-guided fused lasso, linear constraints $Gx \geq 0 \implies \iota_{\mathcal{C}}(Gx)$ where $\mathcal{C} = \{v : v \geq 0\}$

The Issue of B's Continuity

- Lipschitz: $\|Bx - By\| \leq L\|x - y\|$
- Cocoercive: $\langle Bx - By, x - y\rangle \geq (1/L)\|Bx - By\|^2$
- Cocoercive $\Longrightarrow$ Lipschitz
- but not the opposite direction
  - Example: skew-symmetric linear operators

$$Bz = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} z$$

  - Example: saddle-point games

$$Bz = \begin{bmatrix} \nabla_x F(x,y) \\ -\nabla_y F(x,y) \end{bmatrix}$$

The Issue of B's Continuity

- Lipschitz: $\|Bx - By\| \leq L\|x - y\|$
- Cocoercive: $\langle Bx - By, x - y\rangle \geq (1/L)\|Bx - By\|^2$
- Cocoercive $\Longrightarrow$ Lipschitz
- but not the opposite direction
  - Example: skew-symmetric linear operators

$$Bz = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} z$$

  - Example: saddle-point games

$$Bz = \begin{bmatrix} \nabla_x F(x, y) \\ -\nabla_y F(x, y) \end{bmatrix}$$

- However (Baillon-Haddad): for $Bz = \nabla f$ Lipschitz $\iff$ cocoercive

Product Space Reformulation

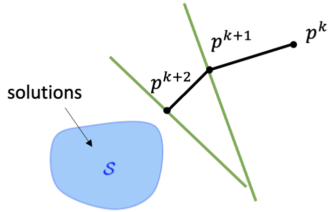
$$0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz \tag{1}$$

- Many splitting algorithms exist for $n = 1, 2$. For example Forward-Backward, three-operator splitting, Douglas-Rachford (ADMM), Tseng's method (FBF), forward-reflected-backward method, Chambolle-Pock splitting etc.
- To extend to arbitrary n, usually use a *product space reformulation* to reduce $(n + 1)$-operator problem to 2 or 3-operator problem in enlargened space.
- *Projective splitting* (PS) solves (1) but is not based on product space
- However Gisselson (2021) rewrites it this way in some special cases

Projective Splitting in a Nutshell

$$0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$

$$\mathcal{S} = \left\{ \underbrace{(z, w_1, \ldots, w_n)}_{p} : \quad w_i \in A_i(G_i z), \quad 0 = \sum_{i=1}^{n} G_i^\top w_i + Bz \right\}$$



$$p^k = (z^k, w_1^k, \ldots, w_n^k)$$

A Brief History of Projective Splitting

- 1999 Origins with projection-type methods by Solodov, Svaiter, Iusem, others
- 2008 Eckstein and Svaiter invent method called "projective splitting" (PS) to solve: $0 \in \sum_{i=1}^{n} A_i z$
- 2013 Alotaibi et al. allow for linear compositions: $0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z)$
- 2015 Combettes et al. extension to *asynchronous* and *block-iterative* operation
- 2018 PJ and Eckstein *2-forward step* version for Lipschitz operators: $0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$
- 2019 PJ and Eckstein *1-forward step* version for *cocoercive* B: $0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$
- 2020 M. Marques Alves et al. *inertial (momentum) version* of PS

Benefits/Quirks of PS

- *Not* based on a fixed-point analysis

$$p^{k+1} = \mathcal{M}(p^k),\ \text{study } p^* = \mathcal{M}(p^*) \text{ and firmly nonexpansive}$$

- Explicitly perform projection gives nice properties directly (eg: Fejér monotonicity)
- Need to prove "good" separating hyperplanes
- Decomposition (full splitting)
- Flexibility
  - Async, block iterative
  - permissive stepsize constraints
  - inexact resolvents (relative error)
  - mix-and-match updates (resolvent, 2-forward-step, 1-forward-step, a different 1-forward step[2], Newton-step[3])

[2]Due to Maicon Marques Alves
[3]Also due to Maicon

Limitation: No Stochastic Oracle

$$\sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$

- Can only access B through noisy oracle $\tilde{B}z = Bz + \epsilon$
- Eg: $Bz = \frac{1}{N}\sum_{j=1}^{N} B_j z$, sample $\tilde{B}z = B_J z$ where $J \sim \text{uniform}\{1, \ldots, N\}$
- Projective splitting cannot handle that

Contributions of this Work

$$\sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$

- Assume can access $Bz + \epsilon$ where B is Lipschitz
  1. $\mathbb{E}_k[\epsilon] = 0$
  2. $\mathbb{E}_k[\|\epsilon\|^2] \leq N_1 + N_2\|Bz\|^2$
- Extend Projective Splitting (PS) using correct *decaying stepsizes* and prove
  1. almost-sure convergence of iterates to a solution
  2. $\mathcal{O}(1/\sqrt{k})$ rate for the expected solution residual

Contributions of this Work

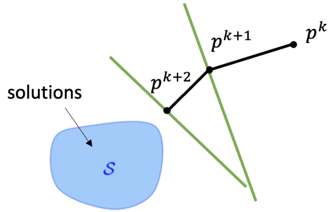
$$\sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$

- Assume can access $Bz + \epsilon$ where B is Lipschitz
  1. $\mathbb{E}_k[\epsilon] = 0$
  2. $\mathbb{E}_k[\|\epsilon\|^2] \leq N_1 + N_2\|Bz\|^2$
- Extend Projective Splitting (PS) using correct *decaying stepsizes* and prove
  1. almost-sure convergence of iterates to a solution
  2. $\mathcal{O}(1/\sqrt{k})$ rate for the expected solution residual
- When B is cocoercive, for several *variance reduced estimators* extend PS
  1. $\mathcal{O}(1/k)$ rate of expected solution residual
  2. Linear rate of iterates under additional strong monotonicy + cocoercivity
  3. Better computational complexities than deterministic PS

Deterministic Projective Splitting Background

$$0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$

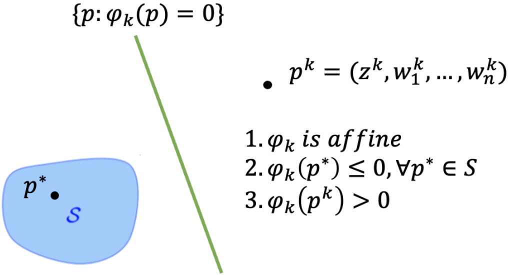
$$\mathcal{S} = \left\{ \underbrace{(z, w_1, \ldots, w_n)}_{p} : \quad w_i \in A_i(G_i z), \quad 0 = \sum_{i=1}^{n} G_i^\top w_i + Bz \right\}$$



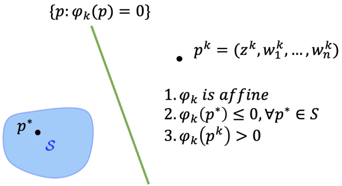
$$p^k = (z^k, w_1^k, \ldots, w_n^k)$$

Constructing a Separating Hyperplane I

- Construct scalar function $\varphi_k : \mathbb{R}^d \times \mathbb{R}^{d_1} \times \ldots \mathbb{R}^{d_n} \to \mathbb{R}$
- $\varphi_k(p) = \varphi_k(z, w_1, \ldots, w_n)$



Constructing a Separating Hyperplane II[4]



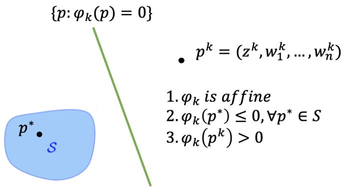
$$\mathcal{S} = \{(z^*, w_1^*, \dots, w_n^*) : w_i^* \in A_i(G_i z^*), -\sum_{i=1}^n G_i^\top w_i^* = Bz^*\}$$

$$\varphi_k(z, w_1, \dots, w_n) = \sum_{i=1}^n \langle G_i z - x_i^k, y_i^k - w_i \rangle + \langle z - x_{n+1}^k, y_{n+1}^k + \sum_{i=1}^n G_i^\top w_i \rangle$$

- (x_i^k, y_i^k) parameterize hyperplane.
- Can be shown that
  1. φ_k is affine
  2. Choosing $y_i^k \in A_i x_i^k$ and $y_{n+1}^k = B x_{n+1}^k \implies \varphi_k(p^*) \leq 0$

[4] Alotaibi et al. 2013

Meta-Algorithm



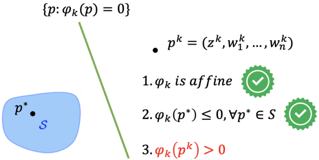
$$\varphi_k(z, w_1, \ldots, w_n) = \sum_{i=1}^{n} \langle G_i z - x_i^k, y_i^k - w_i \rangle + \langle z - x_{n+1}^k, y_{n+1}^k + \sum_{i=1}^{n} G_i^\top w_i \rangle$$

1. Select (x_i^k, y_i^k) for $i = 1, \ldots, n+1$ such that $y_i^k \in A_i x_i^k$ and $y_{n+1}^k = B x_{n+1}^k$ and $\varphi_k(p^k) \gg 0$
2. Project p^k onto hyperplane (easy) to get p^{k+1}
   - $\nabla_z \varphi_k = \sum_{i=1}^n G_i^\top y_i^k + y_{n+1}^k, \quad \nabla_{w_i} \varphi_k = x_i^k - G_i x_{n+1}^k$
   - $\alpha_k = \varphi_k(p^k)/\|\nabla \varphi_k\|^2$

$$p^{k+1} = p^k - \alpha_k \nabla \varphi_k$$

Good Separators I

$$0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$



$$\varphi_k(p^k) = \sum_{i=1}^{n} \langle G_i z^k - x_i^k, y_i^k - w_i^k \rangle + \langle z^k - x_{n+1}^k, y_{n+1}^k - w_{n+1}^k \rangle$$

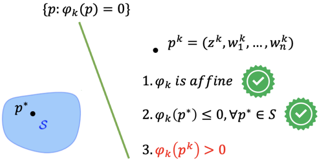
- Treat each i separately (splitting)
- Choose (Eckstein et al. 2008, Alotaibi et al. 2013)

$$x_i^k = J_{\tau A_i}(G_i z^k + \tau w_i^k), \quad y_i^k = \tau^{-1}(G_i z^k + w_i^k - x_i^k)$$

- Simple properties of resolvent:
  1. $y_i^k \in A_i x_i^k$
  2. $\langle G_i z - x_i^k, y_i^k - w_i^k \rangle = (1/\tau)\|G_i z^k - x_i^k\|^2$

Good Separators II

$$0 \in \sum_{i=1}^{n} G_i^\top A_i(G_i z) + Bz$$



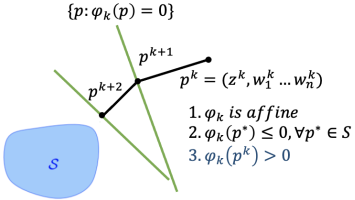
$$\varphi_k(p^k) = \sum_{i=1}^{n} \langle G_i z^k - x_i^k, y_i^k - w_i^k \rangle + \langle z^k - x_{n+1}^k, y_{n+1}^k + \sum_{i=1}^{n} G_i^\top w_i^k \rangle$$

- Choose (PJ and Eckstein 2018)

$$x_{n+1}^k = z^k - \rho(Bz^k + \sum_{i=1}^{n} G_i^\top w_i^k), \quad y_{n+1}^k = Bx_{n+1}^k$$

$$\langle z^k - x_{n+1}^k, y^k + \sum_{i=1}^{n} G_i^\top w_i^k \rangle \geq \left(\rho^{-1} - L\right) \|z^k - x_{n+1}^k\|^2.$$

Algorithm Summary[5]



$$\varphi_k(p^k) = \sum_{i=1}^{n} \langle G_i z^k - x_i^k, y_i^k - w_i^k \rangle + \langle z^k - x_{n+1}^k, y_{n+1}^k + \sum_{i=1}^{n} G_i^\top w_i^k \rangle$$

1. Find good separator (i.e. choose (x_i^k, y_i^k))
   1. For $i = 1, \dots, n$ $\quad x_i^k = J_{\tau A_i}(G_i z^k + \tau w_i^k), \quad y_i^k = \tau^{-1}(G_i z^k + w_i^k - x_i^k)$
   2. For $i = n+1$: $w_{n+1}^k = -\sum_{i=1}^n G_i^\top w_i^k$,
      $x_{n+1}^k = z^k - \rho(Bz^k - w_{n+1}^k), \quad y_{n+1}^k = Bx_{n+1}^k$
2. Project $p^k = (z^k, w_1^k, \dots, w_n^k)$ onto hyperplane
   1. $\nabla_z \varphi_k = \sum_{i=1}^n G_i^\top y_i^k + y_{n+1}^k, \quad \nabla_{w_i} \varphi_k = x_i^k - G_i x_{n+1}^k$
   2. $\alpha_k = \varphi_k(p^k)/\|\nabla \varphi_k\|^2$

$$p^{k+1} = p^k - \alpha_k \nabla \varphi_k$$

[5] Projective Splitting with Forward Steps, PJ and Eckstein 2018

Making things Stochastic

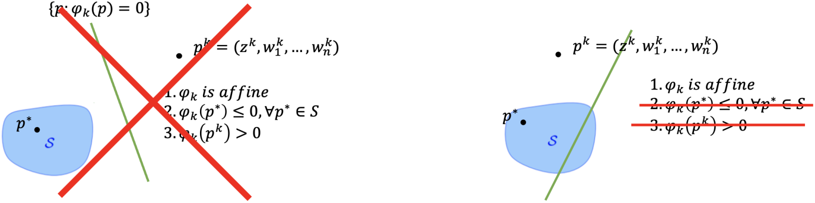
1. Find good separator (i.e. choose (x_i^k, y_i^k))
    1. For $i = 1, \ldots, n$ $\quad x_i^k = J_{\tau A_i}(G_i z^k + \tau w_i^k), \quad y_i^k = \tau^{-1}(G_i z^k + w_i^k - x_i^k)$
    2. For $i = n+1$: $w_{n+1}^k = -\sum_{i=1}^n G_i^\top w_i^k$,

$$\boxed{x_{n+1}^k = z^k - \rho(Bz^k + \epsilon^k - w_{n+1}^k), \quad y_{n+1}^k = Bx_{n+1}^k + e^k}$$

2. Project p^k onto hyperplane
    1. $\nabla_z \varphi_k = \sum_{i=1}^n G_i^\top y_i^k + y_{n+1}^k, \quad \nabla_{w_i}\varphi_k = x_i^k - G_i x_{n+1}^k$
    2. $\alpha_k = \varphi_k(p^k)/\|\nabla\varphi_k\|^2$

$$p^{k+1} = p^k - \alpha_k \nabla\varphi_k$$

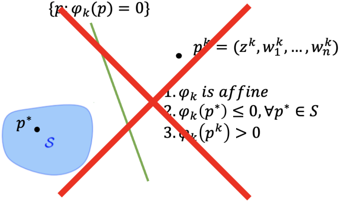
Effect of Noise

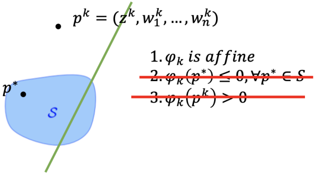


1. Find good separator (i.e. choose (x_i^k, y_i^k))
   1. For $i = 1, \ldots, n$ $\quad x_i^k = J_{\tau A_i}(G_i z^k + \tau w_i^k), \quad y_i^k = \tau^{-1}(G_i z^k + w_i^k - x_i^k)$
   2. For $i = n+1$: $w_{n+1}^k = -\sum_{i=1}^n G_i^\top w_i^k$,
      $x_{n+1}^k = z^k - \rho(Bz^k + \epsilon^k - w_{n+1}^k), \quad y_{n+1}^k = Bx_{n+1}^k + e^k$
2. Project p^k onto hyperplane
   1. $\nabla_z \varphi_k = \sum_{i=1}^n G_i^\top y_i^k + y_{n+1}^k, \quad \nabla_{w_i} \varphi_k = x_i^k - G_i x_{n+1}^k$
   2. $\alpha_k = \varphi_k(p^k)/\|\nabla \varphi_k\|^2$

$$p^{k+1} = p^k - \alpha_k \nabla \varphi_k$$

The Way Out

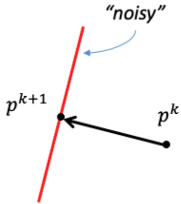


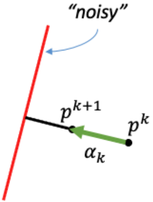


1. Find good separator (i.e. choose (x_i^k, y_i^k))
   1. For $i = 1, \ldots, n$ $\quad x_i^k = J_{\tau A_i}(G_i z^k + \tau w_i^k), \quad y_i^k = \tau^{-1}(G_i z^k + w_i^k - x_i^k)$
   2. For $i = n+1$: $w_{n+1}^k = -\sum_{i=1}^n G_i^\top w_i^k$,
      $x_{n+1}^k = z^k - \underline{\rho_k}(Bz^k + \epsilon^k - w_{n+1}^k), \quad y_{n+1}^k = Bx_{n+1}^k + e^k$
2. Project p^k onto hyperplane
   1. $\nabla_z \varphi_k = \sum_{i=1}^n G_i^\top y_i^k + y_{n+1}^k, \quad \nabla_{w_i} \varphi_k = x_i^k - G_i x_{n+1}^k$
   2. ~~$\alpha_k = \varphi_k(p^k)/\|\nabla \varphi_k\|^2$~~

$$p^{k+1} = p^k - \underline{\alpha_k} \nabla \varphi_k$$

Change of Perspective





1. Find good separator (i.e. choose (x_i^k, y_i^k))
    1. For $i = 1, \ldots, n \quad x_i^k = J_{\tau A_i}(G_i z^k + \tau w_i^k), \quad y_i^k = \tau^{-1}(G_i z^k + w_i^k - x_i^k)$
    2. For $i = n+1$: $w_{n+1}^k = -\sum_{i=1}^n G_i^\top w_i^k$,
       $x_{n+1}^k = z^k - \rho_k(Bz^k + \epsilon^k - w_{n+1}^k), \quad y_{n+1}^k = Bx_{n+1}^k + e^k$
2. Project p^k onto hyperplane
    1. $\nabla_z \varphi_k = \sum_{i=1}^n G_i^\top y_i^k + y_{n+1}^k, \quad \nabla_{w_i}\varphi_k = x_i^k - G_i x_{n+1}^k$
    2. ~~$\alpha_k = \varphi_k(p^k)/\|\nabla\varphi_k\|^2$~~

$$p^{k+1} = p^k - \alpha_k \nabla\varphi_k$$

Analysis I: Stochastic Quasi-Fejér Monotonicity (SQFM)

Theorem

(Combettes et al. 2015), For any $p^* \in \mathcal{S}$, *if*

$$\mathbb{E}_k[\|p^{k+1} - p^*\|^2] \le (1 + \chi^k)\|p^k - p^*\|^2 - \nu_k + \eta^k$$

where $\nu_k \ge 0$, $\sum_{k=1}^{\infty} \eta^k < \infty$, *and* $\sum_{k=1}^{\infty} \chi^k < \infty$. *Then (a.s.):*

1. p^k *is bounded,*
2. $\|p^k - p^*\|$ *has a limit, and*
3. $\sum_k \nu_k < \infty$

Intuition:

- Without noise and summable terms becomes (Fejér monotonicity)

$$\|p^{k+1} - p^*\|^2 \le \|p^k - p^*\|^2 - \nu_k$$

Analysis II: A Simple Recursion

Using $p^{k+1} = p^k - \alpha_k \nabla \varphi_k$ for any $p^* \in \mathcal{S}$

$$\begin{aligned}\|p^{k+1} - p^*\|^2 &= \|p^k - \alpha_k \nabla \varphi_k - p^*\|^2 \\ &= \|p^k - p^*\|^2 - 2\alpha_k \langle \nabla \varphi_k, p^k - p^* \rangle + \alpha_k^2 \|\nabla \varphi_k\|^2 \\ &= \|p^k - p^*\|^2 - 2\alpha_k \left(\varphi_k(p^k) - \varphi_k(p^*) \right) + \alpha_k^2 \|\nabla \varphi_k\|^2\end{aligned}$$

General strategy:

Analysis II: A Simple Recursion

Using $p^{k+1} = p^k - \alpha_k \nabla \varphi_k$ for any $p^* \in \mathcal{S}$

$$\begin{aligned}
\|p^{k+1} - p^*\|^2 &= \|p^k - \alpha_k \nabla \varphi_k - p^*\|^2 \\
&= \|p^k - p^*\|^2 - 2\alpha_k \langle \nabla \varphi_k, p^k - p^* \rangle + \alpha_k^2 \|\nabla \varphi_k\|^2 \\
&= \|p^k - p^*\|^2 - 2\alpha_k \left(\varphi_k(p^k) - \varphi_k(p^*)\right) + \alpha_k^2 \|\nabla \varphi_k\|^2
\end{aligned}$$

General strategy:

1. Upper bound $\mathbb{E}_k \|\nabla \varphi_k\|^2$
2. Optimality condition lower bound for $\mathbb{E}_k(\varphi_k(p^k) - \varphi_k(p^*)) \geq \mathbb{E}_k[\mathcal{G}_k]$

Analysis II: A Simple Recursion

Using $p^{k+1} = p^k - \alpha_k \nabla\varphi_k$ for any $p^* \in \mathcal{S}$

$$\begin{aligned}\|p^{k+1} - p^*\|^2 &= \|p^k - \alpha_k \nabla\varphi_k - p^*\|^2 \\ &= \|p^k - p^*\|^2 - 2\alpha_k \langle \nabla\varphi_k, p^k - p^* \rangle + \alpha_k^2 \|\nabla\varphi_k\|^2 \\ &= \|p^k - p^*\|^2 - 2\alpha_k \left(\varphi_k(p^k) - \varphi_k(p^*)\right) + \alpha_k^2 \|\nabla\varphi_k\|^2\end{aligned}$$

General strategy:

1. Upper bound $\mathbb{E}_k \|\nabla\varphi_k\|^2$
2. Optimality condition lower bound for $\mathbb{E}_k(\varphi_k(p^k) - \varphi_k(p^*)) \geq \mathbb{E}_k[\mathcal{G}_k]$
3. Use SQFM: If

$$\mathbb{E}_k[\|p^{k+1} - p\|^2] \leq (1 + \chi^k)\|p^k - p\|^2 - \nu_k + \eta^k$$

   Then p^k bounded, $\|p^k - p\|$ has a limit, and ν_k is summable (a.s.)
4. Pick stepsizes α_k and ρ_k to make it all work out!

Analysis III: Putting it together

Bounds we found:

1. $\mathbb{E}_k\|\nabla\varphi_k\|^2 \leq C_1\|p^k - p^*\|^2 + C_2$
2. $\mathbb{E}_k[\varphi_k(p^k) - \varphi_k(p^*)] \geq \rho_k\mathcal{G}_k - \rho_k^2 N$ where

$$\mathcal{G}_k = \sum_{i=1}^{n}\|y_i^k - w_i^k\|^2 + \sum_{i=1}^{n}\|G_i z^k - x_i^k\|^2 + \|Bz^k - w_{n+1}^k\|^2$$

Analysis III: Putting it together

Bounds we found:

1. $\mathbb{E}_k\|\nabla\varphi_k\|^2 \leq C_1\|p^k - p^*\|^2 + C_2$
2. $\mathbb{E}_k[\varphi_k(p^k) - \varphi_k(p^*)] \geq \rho_k\mathcal{G}_k - \rho_k^2 N$ where

$$\mathcal{G}_k = \sum_{i=1}^{n}\|y_i^k - w_i^k\|^2 + \sum_{i=1}^{n}\|G_i z^k - x_i^k\|^2 + \|Bz^k - w_{n+1}^k\|^2$$

- Put in

$$\|p^{k+1} - p^*\|^2 = \|p^k - p^*\|^2 - 2\alpha_k\left(\varphi_k(p^k) - \varphi_k(p^*)\right) + \alpha_k^2\|\nabla\varphi_k\|^2$$

to get

$$\mathbb{E}_k\|p^{k+1} - p^*\|^2 \leq (1 + C_1\alpha_k^2)\|p^k - p^*\|^2 - 2\alpha_k\rho_k\mathcal{G}_k + \alpha_k^2 C_2 + \alpha_k\rho_k^2 C_3$$

Solution Certificate / Optimality Condition

$$\mathcal{G}_k = \sum_{i=1}^{n} \|y_i^k - w_i^k\|^2 + \sum_{i=1}^{n} \|G_i z^k - x_i^k\|^2 + \|Bz^k - w_{n+1}^k\|^2$$

$$(\mathcal{G}_k = 0) \iff w_i^k = y_i^k \in A_i x_i^k = A_i G_i z^k \quad \text{and} \ \underbrace{-\sum_{i=1}^{n} G_i^\top w_i^k}_{w_{n+1}^k} = Bz^k$$

$$\iff p^k = (z^k, w_1^k \ldots, w_n^k) \in \mathcal{S} \text{ (is a solution)}$$

$$\mathcal{S} = \left\{ (z, w_1, \ldots, w_n) : \quad w_i \in A_i(G_i z), \quad 0 = \sum_{i=1}^{n} G_i^\top w_i + Bz \right\}$$

Analysis IV: Exploiting Stochastic Quasi-Fejér Monotonicity (SQFM)

$$\mathbb{E}_k\|p^{k+1} - p^*\|^2 \leq (1 + C_1\alpha_k^2)\|p^k - p^*\|^2 - 2\alpha_k\rho_k\mathcal{G}_k + \alpha_k^2 C_2 + \alpha_k\rho_k^2 C_3$$
$$\mathbb{E}_k[\|p^{k+1} - p\|^2] \leq (1 + \chi^k)\|p^k - p\|^2 - \nu_k + \eta^k \quad (SQFM)$$

Need summable χ^k and η^k:

$$\sum_k \alpha_k^2 < \infty, \quad \sum_k \alpha_k\rho_k^2 < \infty,$$

- Conclude (via SQFM) $\sum_k \nu_k = 2\sum_k \alpha_k\rho_k\mathcal{G}_k < \infty$.
- If $\sum_k \alpha_k\rho_k = \infty$, then $\liminf \mathcal{G}_k = 0$
- with standard arguments can derive:

$$p^k \to \hat{p} \in \mathcal{S} \quad (a.s.)$$

Analysis IV: Exploiting Stochastic Quasi-Fejér Monotonicity (SQFM)

$$\mathbb{E}_k \|p^{k+1} - p^*\|^2 \leq (1 + C_1 \alpha_k^2)\|p^k - p^*\|^2 - 2\alpha_k \rho_k \mathcal{G}_k + \alpha_k^2 C_2 + \alpha_k \rho_k^2 C_3$$

$$\mathbb{E}_k[\|p^{k+1} - p\|^2] \leq (1 + \chi^k)\|p^k - p\|^2 - \nu_k + \eta^k \quad (SQFM)$$

Need summable χ^k and η^k:

$$\sum_k \alpha_k^2 < \infty, \quad \sum_k \alpha_k \rho_k^2 < \infty,$$

- Conclude (via SQFM) $\sum_k \nu_k = 2\sum_k \alpha_k \rho_k \mathcal{G}_k < \infty$.
- If $\sum_k \alpha_k \rho_k = \infty$, then $\liminf \mathcal{G}_k = 0$
- with standard arguments can derive:

$$p^k \to \hat{p} \in \mathcal{S} \quad (a.s.)$$

- Ex. $\alpha_k = k^{-1/2-\varepsilon} \quad \rho_k = k^{-1/4}$
- Double Stepsize Extragradient (Hsieh et al. 2020) is special case when solving $0 = Bz$ (no $A_1, \ldots, A_n$ i.e. no regularizers/constraints).

Convergence Rate

$$\mathbb{E}_k\|p^{k+1}-p^*\|^2 \leq (1+C_1\alpha_k^2)\|p^k-p^*\|^2 - 2\alpha_k\rho_k\mathcal{G}_k + \alpha_k^2 C_2 + \alpha_k\rho_k^2 C_3$$

Fix iterations K set

$$\rho_k = K^{-1/4} \quad \alpha_k = K^{-1/2}$$

can derive

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\mathcal{G}_k] = \mathcal{O}(1/\sqrt{K}).$$

Example Application: Distributionally Robust Sparse Logistic Regression

- Wasserstein Robust logistic regression. Finite dimensional representation, (Yu 2021).
- Add an L_1 regularizer to promote sparsity.
- Training data: $(\hat{x}_i, \hat{y}_i)$ $i = 1, \ldots, m$

$$\min_{\substack{\beta \in \mathbb{R}^d \\ \lambda \in \mathbb{R}}} \max_{\gamma \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{i=1}^{m} \Psi(\langle \hat{x}_i, \beta \rangle) + \frac{1}{m} \sum_{i=1}^{m} \gamma_i (\hat{y}_i \langle \hat{x}_i, \beta \rangle - \lambda) + c \|\beta\|_1 \right\}$$
$$\text{s.t.} \quad \|\beta\|_2 \leq \lambda/2 \qquad \|\gamma\|_\infty \leq 1.$$

- Ψ is the logistic loss

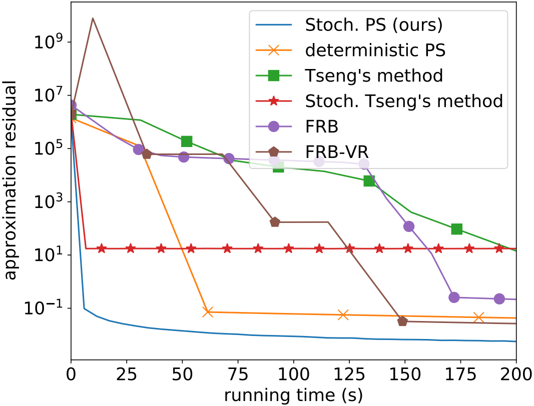
Splitting Up the Problem

$$\min_{\substack{\beta\in\mathbb{R}^d\\ \lambda\in\mathbb{R}}} \max_{\gamma\in\mathbb{R}^m} \quad \left\{\lambda(\delta-\kappa)+\frac{1}{m}\sum_{i=1}^m \Psi(\langle \hat{x}_i,\beta\rangle)+\frac{1}{m}\sum_{i=1}^m \gamma_i(\hat{y}_i\langle \hat{x}_i,\beta\rangle-\lambda\kappa)+c\|\beta\|_1\right\}$$
$$\text{s.t.} \qquad \|\beta\|_2\leq\lambda/2 \qquad \|\gamma\|_\infty\leq 1.$$

- The constraints $\|\beta\|_2 \leq \lambda/2$ and $\|\gamma\|_\infty \leq 1$ are handled by a single set-valued operator A_1
- The $c\|\beta\|_1$ regularization penalty is handled by second set-valued operator A_2
- Everything else is Lipschitz and absorbed into B

$$0 \in A_1 z + A_2 z + Bz \quad \text{where} \quad z = (\beta, \lambda, \gamma)$$

Experimental Results



S-Tseng: *Two steps at a time – taking GAN training in stride with Tseng's method*, Böhm et al., **FRB-VR:** *Forward-reflected-backward method with variance reduction*, Alacaoglu et al., **FRB:** *A Forward-Backward Splitting Method for Monotone Inclusions Without Cocoercivity*, Malitsky et al.

Variance Reduced Projective Splitting

- Assume $Bz^k = (1/N)\sum_{j=1}^{N} B_j z^k$ and B_j are cocoercive
- Assume *some* variance-reduced estimator: $y^k \approx Bz^k = (1/N)\sum_{j=1}^{N} B_j z^k$
  - Must satisfy some simple recursions
  - Holds for estimators based on SVRG, loopess SVRG, SAGA, SEGA...
- Extension to monotone inclusions of condition in "A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent" Gorbunov et al.

Variance Reduction - Convergence Rates

- Under these conditions obtain

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\mathcal{G}_k] \leq \mathcal{O}(K^{-1}) \quad \text{(up from } \mathcal{O}(K^{-1/2})\text{)}$$

$$\|z^k - z^*\| \leq \mathcal{O}(q^{-k}) \quad \text{under strong monotonicity, cocoercivity}$$

- Same rates as deterministic PS

Variance Reduction - Computational Complexities

Method	Comp. Complexity
deterministic PS[6]	$\mathcal{O}\left(\frac{NL}{\epsilon}\right)$
VR-PS (this work)	$\mathcal{O}\left(\frac{N+L}{\epsilon}\right)$
stochastic PS (this work)	$\mathcal{O}\left(\frac{L^2}{\epsilon^2}\right)$

where L=max Lipschitz constant of $B_j, j = 1, \ldots, N$
Complexity measures the # of stoch. gradient (equiv.) oracles

[6] *convergence rates for projective splitting*, PJ and Eckstein 2019

Conclusion

Paper:

- *Stochastic Projective Splitting: Solving Saddle-Point Problems with Multiple Regularizers.* arXiv:2106.13067, PJ, Jonathan Eckstein, Thomas Flynn, Shinjae Yoo

Thank you!

Variance Reduction

$$\mathbb{E}_k[\|y^k - Bz^*\|^2] \leq \frac{\gamma_1}{N}\sum_{j=1}^{N}\|B_j z^k - B_j z^*\|^2 + \gamma_2\sigma_k^2$$

$$\mathbb{E}_k[\sigma_{k+1}^2] \leq (1-\gamma_3)\sigma_k^2 + \frac{\gamma_3\gamma_1}{N}\sum_{j=1}^{N}\|B_j z^k - B_j z^*\|^2$$