Deterministic and Stochastic Gradient Methods for Nonsmooth Nonconvex Regularized Optimization

Akiko Takeda

The University of Tokyo / RIKEN

Joint work with
1: Tianxiang Liu (RIKEN)
2: Ting Kei Pong (Hong Kong Polytechnic University)
3: Michael R. Metel (RIKEN)

Nonsmooth Nonconvex Optimization

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})$$

f: L-smooth func. (∇f is Lipschitz continuous with L)

$g_0, g_1, \ldots, g_m$: prox-friendly

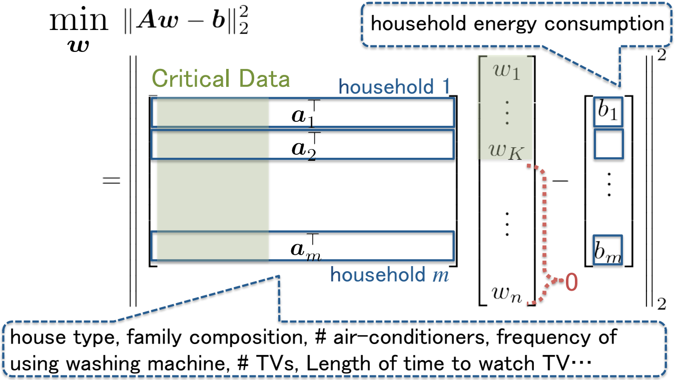
($\mathbf{prox}_{\lambda g_i}(\boldsymbol{w}) := \operatorname*{argmin}_{\boldsymbol{x}} g_i(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x}-\boldsymbol{w}\|_2^2$ is easily computed)

$f + g_0$: level-bounded

- f, g_i: can be nonconvex
- g_i: can be nonsmooth
- $\mathcal{A}_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$ are linear mappings

Constrained Sparse Regression Problem

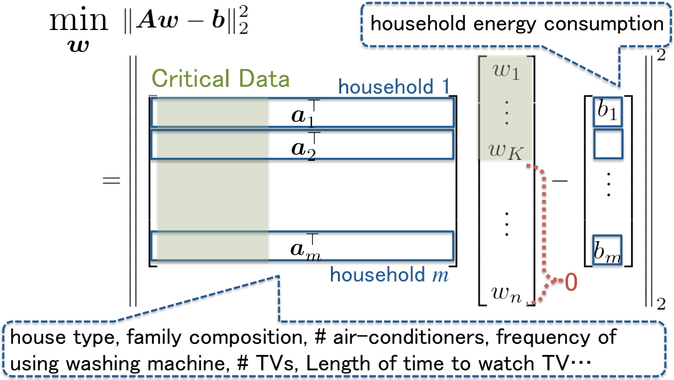
$$\min_{\boldsymbol{w}} \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2$$



under the ℓ_0-norm const. $\|\boldsymbol{w}\|_0 \leq K$ and another one $\boldsymbol{w} \in C$.

$$\Longrightarrow \min_{\boldsymbol{w}} \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2 + \delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \leq K\}}(\boldsymbol{w}) + \delta_C(\boldsymbol{w}).$$

Constrained Sparse Regression Problem

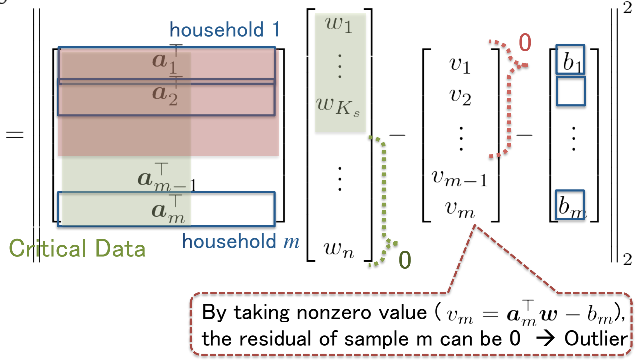


under the ℓ_0-norm const. $\|\boldsymbol{w}\|_0 \leq K$ and another one $\boldsymbol{w} \in C$.

$$\Longrightarrow \min_{\boldsymbol{w}} \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2 + \delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \leq K\}}(\boldsymbol{w}) + \delta_C(\boldsymbol{w}).$$

Simultaneous Sparse Recovery and Outlier Detection

$$\min_{\boldsymbol{v},\boldsymbol{w}} \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{v} - \boldsymbol{b}\|_2^2$$



under two ℓ_0-norm constraints: $\|\boldsymbol{v}\|_0 \le K_o$, $\|\boldsymbol{w}\|_0 \le K_s$

Nonzero elements in $\boldsymbol{v}$ are regarded as outliers (the residual $=$0)

Unconstrained Sparse Regularized Optimization

$$\min_{\boldsymbol{w}} \quad f(\boldsymbol{w}) + \delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}}(\boldsymbol{w})$$

Assump.: f is L-smooth func. (∇f is Lipschitz continuous with L)

- Proximal Gradient Method (PGM) iteratively solves

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \mathbf{prox}_{\delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}}}\left(\boldsymbol{w}_t - \tfrac{1}{L}\nabla f(\boldsymbol{w}_t)\right) \\ &= \mathbf{proj}_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}}\left(\boldsymbol{w}_t - \tfrac{1}{L}\nabla f(\boldsymbol{w}_t)\right), \end{aligned}$$

  where the projection is easy.

  Blumensath, Davies ('09), Bertimas, King & Mazumder ('16)

- PGM finds a stationary point of the problem.
- Weakness: When other constraints exist, projection onto the intersection of $\{w : \|w\|_0 \le K\}$ and C is difficult in many cases.

Unconstrained Sparse Regularized Optimization

$$\min_{\boldsymbol{w}} \quad f(\boldsymbol{w}) + \delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}}(\boldsymbol{w})$$

Assump.: f is L-smooth func. (∇f is Lipschitz continuous with L)

- Proximal Gradient Method (PGM) iteratively solves

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \mathbf{prox}_{\delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}}} \left(\boldsymbol{w}_t - \tfrac{1}{L}\nabla f(\boldsymbol{w}_t)\right) \\ &= \mathbf{proj}_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}} \left(\boldsymbol{w}_t - \tfrac{1}{L}\nabla f(\boldsymbol{w}_t)\right), \end{aligned}$$

  where the projection is easy.

  Blumensath, Davies ('09), Bertimas, King & Mazumder ('16)

- PGM finds a stationary point of the problem.
- Weakness: When other constraints exist, projection onto the intersection of $\{w : \|w\|_0 \le K\}$ and C is difficult in many cases.

Unconstrained Sparse Regularized Optimization

$$\min_{\boldsymbol{w}} \quad f(\boldsymbol{w}) + \delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}}(\boldsymbol{w})$$

Assump.: f is L-smooth func. (∇f is Lipschitz continuous with L)

- Proximal Gradient Method (PGM) iteratively solves

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \mathbf{prox}_{\delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}}} \left(\boldsymbol{w}_t - \tfrac{1}{L}\nabla f(\boldsymbol{w}_t)\right) \\ &= \mathbf{proj}_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}} \left(\boldsymbol{w}_t - \tfrac{1}{L}\nabla f(\boldsymbol{w}_t)\right), \end{aligned}$$

  where the projection is easy.

  Blumensath, Davies ('09), Bertimas, King & Mazumder ('16)

- PGM finds a stationary point of the problem.
- Weakness: When other constraints exist, projection onto the intersection of $\{w : \|w\|_0 \le K\}$ and C is difficult in many cases.

Unconstrained Sparse Regularized Optimization

$$\min_{\boldsymbol{w} \in C} \quad f(\boldsymbol{w}) + \delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \leq K\}}(\boldsymbol{w})$$

Assump.: f is L-smooth func. (∇f is Lipschitz continuous with L)

- Proximal Gradient Method (PGM) iteratively solves

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \mathbf{prox}_{\delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \leq K\}}}\left(\boldsymbol{w}_t - \tfrac{1}{L}\nabla f(\boldsymbol{w}_t)\right) \\ &= \mathbf{proj}_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \leq K\}}\left(\boldsymbol{w}_t - \tfrac{1}{L}\nabla f(\boldsymbol{w}_t)\right), \end{aligned}$$

  where the projection is easy.

  Blumensath, Davies ('09), Bertimas, King & Mazumder ('16)

- PGM finds a stationary point of the problem.
- Weakness: When other constraints exist, projection onto the intersection of $\{\boldsymbol{w} : \|\boldsymbol{w}\|_0 \leq K\}$ and C is difficult in many cases.

Two Difficulties in General Form

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i\boldsymbol{w})$$

$g_0,\ g_1, \ldots,\ g_m$: prox-friendly

- In general,

$$\mathbf{prox}_{g_i+g_j}(\boldsymbol{w}) \neq \mathbf{prox}_{g_i}(\mathbf{prox}_{g_j}(\boldsymbol{w})) \neq \mathbf{prox}_{g_j}(\mathbf{prox}_{g_i}(\boldsymbol{w})),$$

  though some sufficinet condition is shown in Yu (NIPS, 2013).

- $\min_{\boldsymbol{w}\in\mathbb{R}^n} f(\boldsymbol{w}) + g_i(\mathcal{A}_i\boldsymbol{w})$ becomes difficult than $\min_{\boldsymbol{w}\in\mathbb{R}^n} f(\boldsymbol{w}) + g_i(\boldsymbol{w})$, because the proximal operator of $\tilde{g}_i(\boldsymbol{w}) := g_i(\mathcal{A}_i\boldsymbol{w})$ is complicated by the presence of $\mathcal{A}_i$.

Two Difficulties in General Form

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i\boldsymbol{w})$$

$g_0,\ g_1, \ldots,\ g_m$: prox-friendly

- In general,

$$\mathbf{prox}_{g_i+g_j}(\boldsymbol{w}) \neq \mathbf{prox}_{g_i}(\mathbf{prox}_{g_j}(\boldsymbol{w})) \neq \mathbf{prox}_{g_j}(\mathbf{prox}_{g_i}(\boldsymbol{w})),$$

  though some sufficinet condition is shown in Yu (NIPS, 2013).
- $\min_{\boldsymbol{w}\in\mathbb{R}^n} f(\boldsymbol{w}) + g_i(\mathcal{A}_i\boldsymbol{w})$ becomes difficult than $\min_{\boldsymbol{w}\in\mathbb{R}^n} f(\boldsymbol{w}) + g_i(\boldsymbol{w})$, because the proximal operator of $\tilde{g}_i(\boldsymbol{w}) := g_i(\mathcal{A}_i\boldsymbol{w})$ is complicated by the presence of $\mathcal{A}_i$.

Examples of Nonconvex Nonsmooth Problems

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \quad f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})$$

Assumption: $\mathbf{prox}_{\lambda g_i}(\boldsymbol{w}) := \operatorname*{argmin}_{\boldsymbol{x}} g_i(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x}-\boldsymbol{w}\|^2$ is easy

- Sparse regularizer, e.g., $g_i(\boldsymbol{w}) := \|\boldsymbol{w}\|_1$, $\|\boldsymbol{w}\|_0$, MCP, SCAD ...
  $\mathcal{A}_i$ can express certain structural sparsity such as $\sum_{i=2}^{n} |w_i - w_{i-1}|$
  $\Longrightarrow$ used in image processing for making small the horizontal or/and vertical differences between pixels

- Simple constraint $g_i(\boldsymbol{w}) := \delta_{\{\boldsymbol{w}:\, h_i(\boldsymbol{w})\le 0\}}(\boldsymbol{w})$ such as
  - $\mathrm{rank}\,(\boldsymbol{W}) \le K_r$ $\quad$ ($\mathrm{rank}\,(\mathcal{H}(\boldsymbol{W})) \le K_r$ is also possible)
  - $W_{ij} = M_{ij}, \quad (i,j) \in \mathcal{I}$
  - $\|\mathrm{vec}(\boldsymbol{W})\|_0 \le K_0$
  - $\boldsymbol{W} \succeq \boldsymbol{O}, \cdots$

Examples of Nonconvex Nonsmooth Problems

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \quad f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i\boldsymbol{w})$$

Assumption: $\mathbf{prox}_{\lambda g_i}(\boldsymbol{w}) := \underset{\boldsymbol{x}}{\operatorname{argmin}}\, g_i(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x}-\boldsymbol{w}\|^2$ is easy

- Sparse regularizer, e.g., $g_i(\boldsymbol{w}) := \|\boldsymbol{w}\|_1,\ \|\boldsymbol{w}\|_0$, MCP, SCAD ...
  $\mathcal{A}_i$ can express certain structural sparsity such as $\sum_{i=2}^{n}|w_i - w_{i-1}|$
  $\Longrightarrow$ used in image processing for making small the horizontal or/and vertical differences between pixels

- Simple constraint $g_i(\boldsymbol{w}) := \delta_{\{\boldsymbol{w}:\ h_i(\boldsymbol{w})\le 0\}}(\boldsymbol{w})$ such as
  - $\operatorname{rank}(\boldsymbol{W}) \le K_r$ $\quad$ ($\operatorname{rank}(\mathcal{H}(\boldsymbol{W})) \le K_r$ is also possible)
  - $W_{ij} = M_{ij},\quad (i,j)\in\mathcal{I}$
  - $\|\mathrm{vec}(\boldsymbol{W})\|_0 \le K_0$
  - $\boldsymbol{W} \succeq \boldsymbol{O}, \cdots$

Our Recent Works

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})$$

f: L-smooth func. (∇f is Lipschitz continuous with L)

g_i: prox-friendly ($\min_{\boldsymbol{x}} g_i(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{w}\|_2^2$ is easy)

$f + g_0$: level-bounded

- DC approach for special cases with $m = 1$ (constrained sparse opt.) [Tono, T. & Gotoh, arXiv '17], [Gotoh, T. & Tono, MathProg '18]
- DC approach for the general problem [Liu, Pong & T., MathProg '19]
- Applications to system identification, [Liu, Markovsky, Pong & T., SIMAX '20]
  sparse recovery and outlier detection [Liu, Pong & T., COAP '19]
- Stochastic DC approach [Metel & T., ICML '19]

Our Recent Works

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})$$

f: L-smooth func. (∇f is Lipschitz continuous with L)

g_i: prox-friendly ($\min_{\boldsymbol{x}} g_i(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{w}\|_2^2$ is easy)

$f + g_0$: level-bounded

- DC approach for special cases with $m = 1$ (constrained sparse opt.)
  [Tono, T. & Gotoh, arXiv '17], [Gotoh, T. & Tono, MathProg '18]
- DC approach for the general problem [Liu, Pong & T., MathProg '19]
- Applications to system identification, [Liu, Markovsky, Pong & T., SIMAX '20]
  sparse recovery and outlier detection [Liu, Pong & T., COAP '19]
- Stochastic DC approach [Metel & T., ICML '19]

Our Recent Works

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})$$

f: L-smooth func. (∇f is Lipschitz continuous with L)

g_i: prox-friendly ($\min_{\boldsymbol{x}} g_i(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{w}\|_2^2$ is easy)

$f + g_0$: level-bounded

- DC approach for special cases with $m = 1$ (constrained sparse opt.)
  [Tono, T. & Gotoh, arXiv '17], [Gotoh, T. & Tono, MathProg '18]
- DC approach for the general problem [Liu, Pong & T., MathProg '19]
- Applications to system identification,
  [Liu, Markovsky, Pong & T., SIMAX '20]
  sparse recovery and outlier detection [Liu, Pong & T., COAP '19]
- Stochastic DC approach [Metel & T., ICML '19]

Related Works

Assumption: f: L-smooth func., g_0: prox-friendly

Deterministic approach

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + g_0(\boldsymbol{w})$$

Global convergence to a stationary point when $g_0(\boldsymbol{w})$ is nonconvex

Wright, Nowak & Figueiredo ('09), Gong et al. ('13)

Stochastic approach

$$\min_{\boldsymbol{w}} \; h(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w}, \boldsymbol{\xi})] +$$

Non-asymptotic convergence to the *gradient mapping* (an approximation of the gradient of h) when $g_0(\boldsymbol{w})$ is convex

Ghadimi, Lan & Zhang, ('16), Li & Li ('18)

Related Works

Assumption: f: L-smooth func., g_0: prox-friendly

Deterministic approach

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + g_0(\boldsymbol{w})$$

Global convergence to a stationary point when $g_0(\boldsymbol{w})$ is nonconvex

Wright, Nowak & Figueiredo ('09), Gong et al. ('13)

Stochastic approach

$$\min_{\boldsymbol{w}} \ h(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w}, \boldsymbol{\xi})] + g_0(\boldsymbol{w})$$

Non-asymptotic convergence to the *gradient mapping* (an approximation of the gradient of h) when $g_0(\boldsymbol{w})$ is convex

Ghadimi, Lan & Zhang, ('16), Li & Li ('18)

Related Works

Assumption: f: L-smooth func., g_0: prox-friendly

Deterministic approach

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} \underbrace{g_i(\mathcal{A}_i \boldsymbol{w})}_{\text{nonconvex}}$$

Global convergence to a stationary point when $g_0(\boldsymbol{w})$ is nonconvex

Wright, Nowak & Figueiredo ('09), Gong et al. ('13)

Stochastic approach

$$\min_{\boldsymbol{w}} \ h(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w}, \boldsymbol{\xi})] + \underbrace{g_0(\boldsymbol{w})}_{\text{nonconvex}}$$

Non-asymptotic convergence to an ϵ-stationary point $\overline{\boldsymbol{w}}$

$$\mathbb{E}\left[\operatorname{dist}(0, \partial h(\overline{\boldsymbol{w}}))\right] \leq \epsilon$$

Key Idea of Our Algorithm (SDCAM)

$$\min_{\boldsymbol{w}} \quad f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \underbrace{\sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})}$$

Key Idea of Our Algorithm (SDCAM)

$$\min_{\boldsymbol{w}} \quad f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \underbrace{\sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})}$$

$$\geq \sum_{i=1}^{m} \text{Moreau envelope func. of } g_i(\mathcal{A}_i \boldsymbol{w})$$

Moreau env.: $e_{\lambda_i} g_i(\boldsymbol{w}) := \min_{\boldsymbol{x}} \; g_i(\boldsymbol{x}) + \frac{1}{2\lambda_i}\|\boldsymbol{x} - \boldsymbol{w}\|_2^2$

$$\mathbf{prox}_{\lambda_i g_i}(\boldsymbol{w}) := \operatorname*{argmin}_{\boldsymbol{x}} g_i(\boldsymbol{x}) + \frac{1}{2\lambda_i}\|\boldsymbol{x} - \boldsymbol{w}\|^2$$

Key Idea of Our Algorithm (SDCAM)

$$\min_{\boldsymbol{w}} \quad f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \underbrace{\sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})}$$

$$\geq \sum_{i=1}^{m} \text{Moreau envelope func. of } g_i(\mathcal{A}_i \boldsymbol{w})$$

$$= \sum_{i=1}^{m} \left(\frac{1}{2\lambda_i} \|\mathcal{A}_i \boldsymbol{w}\|_2^2 - g_i^{\lambda_i}(\mathcal{A}_i \boldsymbol{w}) \right)$$

$g_i^{\lambda_i}$: convex functions

$\frac{1}{\lambda_i}\mathbf{prox}_{\lambda_i g_i}(\mathcal{A}_i \boldsymbol{w}) \subseteq$ the subdifferential of the conv. func $g_i^{\lambda_i}(\mathcal{A}_i \boldsymbol{w})$.

Moreau env.: $e_{\lambda_i} g_i(\boldsymbol{w}) := \min_{\boldsymbol{x}} \ g_i(\boldsymbol{x}) + \frac{1}{2\lambda_i}\|\boldsymbol{x} - \boldsymbol{w}\|_2^2$

$$\mathbf{prox}_{\lambda_i g_i}(\boldsymbol{w}) := \operatorname*{argmin}_{\boldsymbol{x}} g_i(\boldsymbol{x}) + \frac{1}{2\lambda_i}\|\boldsymbol{x} - \boldsymbol{w}\|^2$$

Key Idea of Our Algorithm (SDCAM)

$$\min_{\boldsymbol{w}} \quad f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \underbrace{\sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})}$$

$$\geq \sum_{i=1}^{m} \text{Moreau envelope func. of } g_i(\mathcal{A}_i \boldsymbol{w})$$

$$= \sum_{i=1}^{m} \left(\frac{1}{2\lambda_i} \|\mathcal{A}_i \boldsymbol{w}\|_2^2 - g_i^{\lambda_i}(\mathcal{A}_i \boldsymbol{w}) \right)$$

$g_i^{\lambda_i}$: convex functions

$\frac{1}{\lambda_i} \mathbf{prox}_{\lambda_i g_i}(\mathcal{A}_i \boldsymbol{w}) \subseteq$ the subdifferential of the conv. func $g_i^{\lambda_i}(\mathcal{A}_i \boldsymbol{w})$.

Moreau env.: $e_{\lambda_i} g_i(\boldsymbol{w}) := \min_{\boldsymbol{x}} \; g_i(\boldsymbol{x}) + \frac{1}{2\lambda_i} \|\boldsymbol{x} - \boldsymbol{w}\|_2^2 \xrightarrow{\lambda_i \to 0} g_i(\boldsymbol{w})$

$$\mathbf{prox}_{\lambda_i g_i}(\boldsymbol{w}) := \operatorname*{argmin}_{\boldsymbol{x}} g_i(\boldsymbol{x}) + \frac{1}{2\lambda_i} \|\boldsymbol{x} - \boldsymbol{w}\|^2$$

Example of Moreau Envelope

$$\text{Moreau envelope of } g_i(\mathcal{A}_i \boldsymbol{w}) = \frac{1}{2\lambda_i}\|\mathcal{A}_i \boldsymbol{w}\|_2^2 - \underbrace{g_i^{\lambda_i}(\mathcal{A}_i \boldsymbol{w})}_{\text{convex}}$$

$$\text{Moreau envelope of } \delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0 \le K\}}(\boldsymbol{w}) = \frac{1}{2\lambda}\left(\|\boldsymbol{w}\|_2^2 - |||\boldsymbol{w}|||_{2,K}^2\right)$$

$$\|\boldsymbol{w}\|_0 \le K \quad \Leftrightarrow \quad \|\boldsymbol{w}\|_2^2 - |||\boldsymbol{w}|||_{2,K}^2 = 0$$

$|||\boldsymbol{w}|||_{2,K}^2$: Sum of large K elements among w_i^2, $i \in \{1, 2, \ldots, n\}$

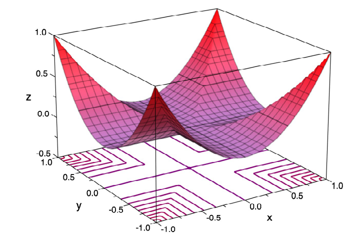
$$\overbrace{\underbrace{w_{(1)}^2 \ge w_{(2)}^2 \ge \ldots \ge w_{(K)}^2}_{\text{Sum: } |||\boldsymbol{w}|||_{2,K}^2} \ge \underbrace{w_{(K+1)}^2 \cdots \ge w_{(n)}^2}_{\text{all } 0}}^{\text{Sum: } \|\boldsymbol{w}\|_2^2}$$

Figs of DC representations

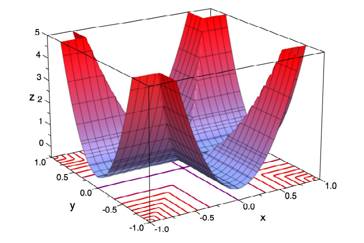
Moreau env. of $\delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0\le K\}}(\boldsymbol{w}) = \frac{1}{2\lambda}(\|\boldsymbol{w}\|_2^2 - |\!|\!|\boldsymbol{w}|\!|\!|_{2,K}^2)$: Penalty term

$(\boldsymbol{w} \in \mathbb{R}^2$ and $K = 1)$

$\frac{1}{2\lambda} = 1$:



$\frac{1}{2\lambda} = 10$:



Moreau env. of $\delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0\le K\}}(\cdot) \xrightarrow{\lambda\to 0} \delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0\le K\}}(\cdot)$

SDCAM: Successive Difference-of-Convex Approximation Method

[Liu, Pong & T., MathProg '19]

$$f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^m g_i(\mathcal{A}_i\boldsymbol{w})$$

$$\geq \left(f(\boldsymbol{w}) + \sum_{i=1}^m \frac{1}{2\lambda_i}\|\mathcal{A}_i\boldsymbol{w}\|_2^2\right) + g_0(\boldsymbol{w}) - \sum_{i=1}^m g_i^{\lambda_i}(\mathcal{A}_i\boldsymbol{w}) =: F^{\boldsymbol{\lambda}}(\boldsymbol{w})$$

1. Decrease $\boldsymbol{\lambda}$ and ϵ. Terminate when they become almost 0.
2. Find an ϵ-stationary point for $\min_{\boldsymbol{w}} F^{\boldsymbol{\lambda}}(\boldsymbol{w})$ with fixed $\boldsymbol{\lambda}$.
   - Linearize $-\sum_{i=1}^m g_i^{\lambda_i}(\mathcal{A}_i\boldsymbol{w})$ at $\boldsymbol{w}_t$ and construct a convex subproblem
   - Solve the subproblem $\Rightarrow$ Opt. sol: $\boldsymbol{w}_{t+1}$
   - Repeat until some termination criteria with ϵ are satisfied.

Theorem $\boldsymbol{w}^*$: accumulation point of $\{\boldsymbol{w}_t\}$. Under some constraint qualification, $\boldsymbol{w}^*$ satisfies the first-order optimality condition:

$$\mathbf{0} \in \nabla f(\boldsymbol{w}^*) + \partial g_0(\boldsymbol{w}^*) + \sum^m \mathcal{A}_i^*\partial g_i(\mathcal{A}_i\boldsymbol{w}^*)$$

NPG$_{\text{major}}$ for $\min_{\boldsymbol{w}} F^{\boldsymbol{\lambda}}(\boldsymbol{w})$

We can use NPG$_{\text{major}}$ for

Wright, Nowak & Figueiredo ('09)

$$\min_{\boldsymbol{w}} \underbrace{\left(f(\boldsymbol{w}) + \sum_{i=1}^{m} \frac{1}{2\lambda_i}\|\mathcal{A}_i\boldsymbol{w}\|_2^2\right)}_{\text{smooth}} + \underbrace{g_0(\boldsymbol{w})}_{\text{prox-friendly}} - \underbrace{\sum_{i=1}^{m} g_i^{\lambda_i}(\mathcal{A}_i\boldsymbol{w})}_{\text{convex} \to \text{linearize}}$$

In NPG$_{\text{major}}$, solve the t-th subproblem $(t = 0, 1, \ldots)$:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \frac{L}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2 + \left(\nabla f(\boldsymbol{w}_t) + \sum_{i=1}^{m} \frac{1}{\lambda_i}\mathcal{A}_i^*[\mathcal{A}_i\boldsymbol{w}_t - \boldsymbol{s}_t]\right)^\top \boldsymbol{w} + g_0(\boldsymbol{w})$$

until the convergence.

$$\frac{1}{\lambda_i}\mathcal{A}_i^*\boldsymbol{s}_t \in \frac{1}{\lambda_i}\mathcal{A}_i^*\mathbf{prox}_{\lambda_i g_i}(\mathcal{A}_i\boldsymbol{w}_t) \subseteq \mathcal{A}_i^*\partial g_i^{\lambda_i}(\mathcal{A}_i\boldsymbol{w}_t)$$

The t-th Subproblem in NPG$_{\text{major}}$

In NPG$_{\text{major}}$, solve the t-th subproblem ($t = 0, 1, \ldots$):

$$\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \frac{L}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2 + \left(\nabla f(\boldsymbol{w}_t) + \sum_{i=1}^{m} \frac{1}{\lambda_i}\mathcal{A}_i^*[\mathcal{A}_i\boldsymbol{w}_t - \boldsymbol{s}_t]\right)^\top \boldsymbol{w} + g_0(\boldsymbol{w})$$

$$= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \frac{L}{2}\left\|\boldsymbol{w} - \left(\boldsymbol{w}_t - \frac{1}{L}\left(\nabla f(\boldsymbol{w}_t) + \sum_{i=1}^{m} \frac{1}{\lambda_i}\mathcal{A}_i^*[\mathcal{A}_i\boldsymbol{w}_t - \boldsymbol{s}_t]\right)\right)\right\| + g_0(\boldsymbol{w})$$

$$\Longrightarrow \boldsymbol{w}_{t+1} = \mathbf{prox}_{\frac{1}{L}g_0}\left(\boldsymbol{w}_t - \frac{1}{L}\left(\nabla f(\boldsymbol{w}_t) + \sum_{i=1}^{m} \frac{1}{\lambda_i}\mathcal{A}_i^*[\mathcal{A}_i\boldsymbol{w}_t - \boldsymbol{s}_t]\right)\right)$$

$$\frac{1}{\lambda_i}\mathcal{A}_i^*\boldsymbol{s}_t \in \frac{1}{\lambda_i}\mathcal{A}_i^*\mathbf{prox}_{\lambda_i g_i}(\mathcal{A}_i\boldsymbol{w}_t) \subseteq \mathcal{A}_i^*\partial g_i^{\lambda_i}(\mathcal{A}_i\boldsymbol{w}_t)$$

Which Constraint Should be Approximated by Moreau Env.

Low-rank + sparse matrix completion

$$\begin{array}{cl} \min_{\boldsymbol{W}\in\mathbb{R}^{m\times n}} & \frac{1}{2}\|\boldsymbol{W}-\boldsymbol{M}\|_F^2 \\ \text{s.t.} & \|\mathrm{vec}(\boldsymbol{W})\|_0 \le s, \quad \mathrm{rank}(\boldsymbol{W}) \le K \end{array}$$

- $\boldsymbol{M} = \boldsymbol{M}_1\boldsymbol{M}_2 + \sigma\boldsymbol{\Delta}$, where $\sigma > 0$ is a noise factor. $\boldsymbol{M}_1 \in \mathbb{R}^{m\times K}$, $\boldsymbol{M}_2 \in \mathbb{R}^{K\times n}$ and $\boldsymbol{\Delta}$ have i.i.d. standard Gaussian entries but $m/10$ random rows of $\boldsymbol{M}_1$ fixed to zero.
- Fix $n = 500$, $K = 10$ and $s = mn/10$ and change $\sigma \in \{.005, .010, .020\}$ and $m \in \{1000, 2000, 3000\}$.
- Decrease the param. of Moreau envelope λ as $\lambda_t = \frac{1}{10^{t+1}}$.
- Stopping criteria: (dist. of $\boldsymbol{X}^t$ to each constraint) $\le 10^{-6} \cdot \|\boldsymbol{X}^t\|_F$.

Which Constraint Should be Approximated by Moreau Env.

Low-rank + sparse matrix completion

$$\min_{\boldsymbol{W}\in\mathbb{R}^{m\times n}} \quad \frac{1}{2}\|\boldsymbol{W}-\boldsymbol{M}\|_F^2$$
$$\text{s.t.} \quad \|\mathrm{vec}(\boldsymbol{W})\|_0 \le s, \quad \mathrm{rank}(\boldsymbol{W}) \le K$$

noise	m	iter		CPU		feas.vio	
		ME.ℓ_0	ME.rank	ME.ℓ_0	ME.rank	ME.ℓ_0	ME.rank
.005	1000	41	5597	4.7	378.1	4.76e-04	1.05e-04
	2000	12	5298	4.0	647.0	6.71e-04	1.52e-04
	3000	12	4618	6.0	862.8	8.20e-04	1.89e-04
.010	1000	4508	7900	379.3	529.2	9.43e-05	2.10e-04
	2000	4453	7526	653.6	912.6	1.34e-04	3.06e-04
	3000	4428	5721	969.5	1080.6	1.64e-04	3.77e-04
.020	1000	4922	11631	413.7	769.2	1.90e-04	4.22e-04
	2000	4634	10267	675.5	1251.3	2.68e-04	6.11e-04
	3000	4580	10859	1003.5	2043.0	3.28e-04	7.55e-04

ME.ℓ_0 (or ME.rank): ℓ_0-norm (or rank) const. is approximated

Stochastic Setting for Nonsmooth Nonconvex Optimization

Assumption: $F(\cdot, \boldsymbol{\xi})$: L-smooth func., g_0: nonsmooth, prox-friendly
$\boldsymbol{\xi}$: random vector following a probability distr. P

Stochastic problem

$$\min_{\boldsymbol{w}} \; h(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w}, \boldsymbol{\xi})] + g_0(\boldsymbol{w})$$

or the finite-sum problem where the expectation is taken over an empirical distribution func.:

$$\min_{\boldsymbol{w}} \; h(\boldsymbol{w}) := \frac{1}{m} \sum_{i=1}^{m} F(\boldsymbol{w}, \boldsymbol{\xi}_i) + g_0(\boldsymbol{w})$$

- Regression: $F(\boldsymbol{w}, \boldsymbol{\xi}_i) = (\boldsymbol{a}_i^\top \boldsymbol{w} - b_i)^2$ for each data point $\boldsymbol{\xi}_i := (\boldsymbol{a}_i, b_i)$
- Assume that data size m is huge and $g_0(\boldsymbol{w})$ is a regularizer, e.g., $g_0(\boldsymbol{w}) = \|\boldsymbol{w}\|_p^p \quad (p = 1, 2)$, MCP, SCAD ...

SGD for Nonsmooth Nonconvex Optimization

Assumption: $F(\cdot, \boldsymbol{\xi})$: L-smooth func., g_0: nonsmooth, prox-friendly
$\boldsymbol{\xi}$: random vector following a probability distr. P

Stochastic approach

$$\min_{\boldsymbol{w}} \; h(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w}, \boldsymbol{\xi})] + \underbrace{g_0(\boldsymbol{w})}_{\text{nonconvex}}$$

Non-asymptotic convergence to an ϵ-stationary point $\overline{\boldsymbol{w}}$

$$\mathbb{E}\left[\mathrm{dist}(0, \partial h(\overline{\boldsymbol{w}}))\right] \leq \epsilon$$

- $g_0(\boldsymbol{w})$ is convex in Ghadimi, Lan & Zhang ('16), Li & Li ('18), etc.
- $g_0(\boldsymbol{w})$ can be nonconvex in Metel & Takeda ('19), Xu et al. ('19). A key ingredient is "Moreau envelope".

Mini-batch Stochastic Gradient Algo. (MBSGA)

Assumption: $F(\cdot, \boldsymbol{\xi})$: L-smooth func., g_0: nonsmooth, prox-friendly
$\boldsymbol{\xi}$: random vector following a probability distr. P

$$\min_{\boldsymbol{w}} \; h(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w}, \boldsymbol{\xi})] + g_0(\boldsymbol{w})$$

Input: $M := \lceil N^{1/4} \rceil, \lambda = \frac{1}{N^{1/4}}, \; \gamma = \min\{\frac{1}{L+\frac{1}{\lambda}}, \; \frac{1}{\sigma\sqrt{N}}\}$

$R \sim$ uniform$\{1, \cdots, N\}$

for $k = 1, \ldots, R-1$ do

- randomly generate $\{\boldsymbol{\xi}_1^k, ..., \boldsymbol{\xi}_M^k\}$
- compute the approximate gradient
  $\boldsymbol{d}^k := \frac{1}{M} \sum_{j=1}^{M} \nabla F(\boldsymbol{w}^k, \boldsymbol{\xi}_j^k) + \frac{1}{\lambda}(\boldsymbol{w}^k - \boldsymbol{s}^k)$ with $\boldsymbol{s}^k \in \mathbf{prox}_{\lambda g_0}(\boldsymbol{w}^k)$
- compute the next iterate $\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \gamma \boldsymbol{d}^k$

Return $\boldsymbol{w}^R$

Convergence Results

$$\mathbb{E}\left[\mathrm{dist}(0, \partial(h(\boldsymbol{w}^R)))\right] \le O(N^{-1/4})$$

MBSGA gives an ϵ-stationary point $\bar{w}$ in expectation, i.e.,

$$\mathbb{E}\left[\mathrm{dist}(0, \partial(h(\bar{w})))\right] \le \epsilon$$

less than $N = O(\epsilon^{-4})$ iterations.

Table: Comparison of Mini-batch algorithm (MBSGA) and Variance reduction algorithm (VRSGA) obtained in (b: Xu et al. ('19), a: it's arXiv)

Algorithm	Finite-sum Assumption	Gradient Call Complexity	Proximal Operator Complexity
SSDC-SPG[a]	×	$O(\epsilon^{-8})$	$O(\epsilon^{-8})$
SSDC-SVRG[a]	√	$O(n\epsilon^{-4})$	$O(\epsilon^{-4})$
MBSGA	×	$O(\epsilon^{-5})$	$O(\epsilon^{-4})$
VRSGA	√	$O(n^{2/3}\epsilon^{-3})$	$O(\epsilon^{-3})$
SSDC-SPG[b]	×	$O(\epsilon^{-5})$	$O(\epsilon^{-5})$
SSDC-SVRG[b]	√	$\tilde{O}(n\epsilon^{-3})$	$\tilde{O}(\epsilon^{-3})$

Experimental Results

Application: Binary classification with smooth non-convex loss function and log-sum penalty as regularizer.

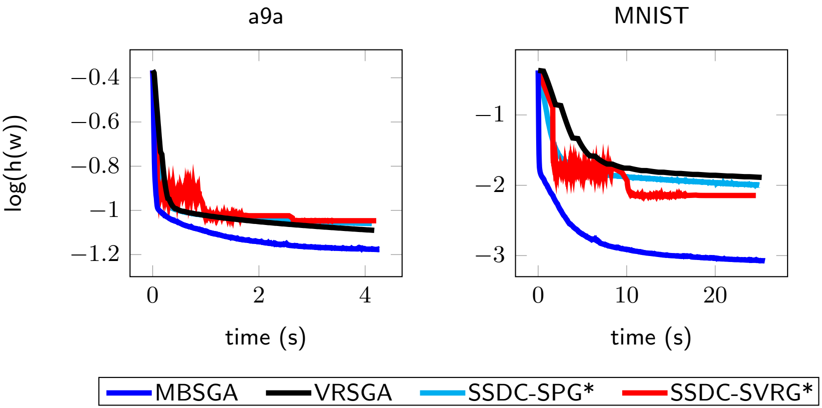


Figure: Comparison to algorithms of Xu et al. (arXiv)

Summary

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})$$

Proposed an approach using proximal operators for
general nonconvex nonsmooth optimization problems
(e.g. nonconvex sparse problems)

Application

- System identification,

  [Liu, Markovsky, Pong & T., SIMAX '20]

  $\Longrightarrow$ Three rank constraints for Henkel matrices

- Sparse recovery and outlier detection [Liu, Pong & T., COAP '19]

  $\Longrightarrow$ Sparse regularization term and ℓ_0-norm const

Summary

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + g_0(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\mathcal{A}_i \boldsymbol{w})$$

Proposed an approach using proximal operators for
general nonconvex nonsmooth optimization problems
(e.g. nonconvex sparse problems)

Application

- System identification,

  [Liu, Markovsky, Pong & T., SIMAX '20]

  $\Longrightarrow$ Three rank constraints for Henkel matrices
- Sparse recovery and outlier detection [Liu, Pong & T., COAP '19]

  $\Longrightarrow$ Sparse regularization term and ℓ_0-norm const