signProx: 1-bit Proximal Algorithm for Nonconvex Stochastic Optimization

Xiaojian Xu[1], Ulugbek Kamilov[1,2]

[1] Computer Science and Engineering,
Washington University in St. Louis

[2] Electrical and Systems Engineering,
Washington University in St. Louis

A common inverse problem

- A common inverse problem model

True signal

$$\boldsymbol{x}^*$$

Forward model

$$\boldsymbol{H}$$

Observation

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}^*$$

A common inverse problem

- A common inverse problem model

True signal

$$\boldsymbol{x}^*$$

Forward model

$$\boldsymbol{H}$$

Observation

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}^*$$

The inverse problem could be formulated as an optimization problem

- A feasible optimization framework to solve the problem

$\boldsymbol{H}$ Forward model, $\boldsymbol{y}$ Observation ⇒ $\widehat{\boldsymbol{x}}$ Estimation of $\boldsymbol{x}^*$

- N. Parikh and S. Boyd, “Proximal algorithms,” Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123–231, 2014.

The inverse problem could be formulated as an optimization problem

- A feasible optimization framework to solve the problem

$\boldsymbol{H}$ Forward model, $\boldsymbol{y}$ Observation ⇒ $\widehat{\boldsymbol{x}}$ Estimation of $\boldsymbol{x}^*$

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \{ f(\boldsymbol{x}) \}$$

- N. Parikh and S. Boyd, “Proximal algorithms,” Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123–231, 2014.

The inverse problem could be formulated as an optimization problem

- A feasible optimization framework to solve the problem

$\boldsymbol{H}$ Forward model — $\boldsymbol{y}$ Observation ⇒ $\widehat{\boldsymbol{x}}$ Estimation of $\boldsymbol{x}^*$

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \{ f(\boldsymbol{x}) \}$$

↓

$$f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x})$$

- N. Parikh and S. Boyd, "Proximal algorithms," Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123–231, 2014.

The inverse problem could be formulated as an optimization problem

- A feasible optimization framework to solve the problem

$\boldsymbol{H}$ Forward model, $\boldsymbol{y}$ Observation → $\widehat{\boldsymbol{x}}$ Estimation of $\boldsymbol{x}^*$

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min}\,\{f(\boldsymbol{x})\}$$

Data-fidelity term

$$f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x}) \begin{cases} d(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|_2^2 \quad \cdots \\ \end{cases}$$

- N. Parikh and S. Boyd, "Proximal algorithms," Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123–231, 2014.

The inverse problem could be formulated as an optimization problem

- A feasible optimization framework to solve the problem

$\boldsymbol{H}$ Forward model, $\boldsymbol{y}$ Observation → $\widehat{\boldsymbol{x}}$ Estimation of $\boldsymbol{x}^*$

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x})\}$$

Data-fidelity term

$$f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x}) \begin{cases} d(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2 & \cdots \\ r(\boldsymbol{x}) = \end{cases}$$

Regularizer term

- N. Parikh and S. Boyd, "Proximal algorithms," Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123–231, 2014.

The inverse problem could be formulated as an optimization problem

- A feasible optimization framework to solve the problem

$\boldsymbol{H}$ Forward model, $\boldsymbol{y}$ Observation → $\widehat{\boldsymbol{x}}$ Estimation of $\boldsymbol{x}^*$

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x})\}$$

$$f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x})$$

Data-fidelity term: $d(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2 \quad \cdots$

Regularizer term: $r(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$

- N. Parikh and S. Boyd, "Proximal algorithms," Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123–231, 2014.

The inverse problem could be formulated as an optimization problem

- A feasible optimization framework to solve the problem

$\boldsymbol{H}$ Forward model, $\boldsymbol{y}$ Observation → $\widehat{\boldsymbol{x}}$ Estimation of $\boldsymbol{x}^*$

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x})\}$$

$$f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x})$$

Data-fidelity term

$$d(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2 \quad \cdots$$

Regularizer term

$$r(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 \quad \|\boldsymbol{x}\|_2$$

- N. Parikh and S. Boyd, "Proximal algorithms," Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123–231, 2014.

The inverse problem could be formulated as an optimization problem

- A feasible optimization framework to solve the problem

$\boldsymbol{H}$ Forward model, $\boldsymbol{y}$ Observation → $\widehat{\boldsymbol{x}}$ Estimation of $\boldsymbol{x}^*$

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x})\}$$

Data-fidelity term

$$f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x}) \begin{cases} d(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|_2^2 & \cdots \\ r(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 + \|\boldsymbol{x}\|_2 & \cdots \end{cases}$$

Regularizer term

- N. Parikh and S. Boyd, “Proximal algorithms,” Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123–231, 2014.

Guideline of the talk

- Large scale optimization problem

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min}\,\{f(\boldsymbol{x})\} \quad \text{with} \quad f(\boldsymbol{x}) = d(\boldsymbol{x}) + \frac{1}{K}\{r_1(\boldsymbol{x}) + \cdots + r_k(\boldsymbol{x}) + \cdots + r_K(\boldsymbol{x})\}$$

Inverse problem → Optimization problem → Alg: SPGM → **Alg: signProx** → Nonconvex task

Introduction

model

1. Baseline algorithm

2. Proposed algorithm

3. Simulation results

Proximal gradient method could minimize the non-differentiable functions

- Modeled optimization problem

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min}\{f(\boldsymbol{x})\} \quad \text{with} \quad f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x})$$

- J. J. Moreau, "Proximit´e et dualit´e dans un espace hilbertien," Bull. Soc. Math. France, vol. 93, pp. 273–299, 1965.

Proximal gradient method could minimize the non-differentiable functions

- Modeled optimization problem

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{R}^n}{\arg\min} \{ f(\boldsymbol{x}) \} \quad \text{with} \quad f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x})$$

- Proximal gradient method (PGM)

$$\boldsymbol{x}^t \leftarrow \mathrm{prox}_{\gamma r}(\boldsymbol{x}^{t-1} - \gamma \nabla d(\boldsymbol{x}^{t-1}))$$

Proximal gradient mapping

- J. J. Moreau, “Proximit´e et dualit´e dans un espace hilbertien,” Bull. Soc. Math. France, vol. 93, pp. 273–299, 1965.

Proximal gradient method could minimize the non-differentiable functions

- Modeled optimization problem

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{R}^n}{\arg\min} \{ f(\boldsymbol{x}) \} \quad \text{with} \quad f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x})$$

- Proximal gradient method (PGM)

$$\boldsymbol{x}^t \leftarrow \mathrm{prox}_{\gamma r}(\boldsymbol{x}^{t-1} - \gamma \nabla d(\boldsymbol{x}^{t-1})) \quad \left\{ \boldsymbol{s} = \boldsymbol{x}^{t-1} - \gamma \nabla d(\boldsymbol{x}^{t-1}) \right.$$

Proximal gradient mapping

- J. J. Moreau, "Proximit´e et dualit´e dans un espace hilbertien," Bull. Soc. Math. France, vol. 93, pp. 273–299, 1965.

Proximal gradient method could minimize the non-differentiable functions

- Modeled optimization problem

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \{ f(\boldsymbol{x}) \} \quad \text{with} \quad f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x})$$

- Proximal gradient method (PGM)

$$\boldsymbol{x}^t \leftarrow \mathrm{prox}_{\gamma r}(\boldsymbol{x}^{t-1} - \gamma \nabla d(\boldsymbol{x}^{t-1})) \quad \begin{cases} \boldsymbol{s} = \boldsymbol{x}^{t-1} - \gamma \nabla d(\boldsymbol{x}^{t-1}) \\ \boldsymbol{x}^t = \mathrm{prox}_{\gamma r}(\boldsymbol{s}) \triangleq \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{s}\|_2^2 + \gamma r(\boldsymbol{x}) \right\} \end{cases}$$

Proximal gradient mapping

Proximal operation

- J. J. Moreau, "Proximit´e et dualit´e dans un espace hilbertien," Bull. Soc. Math. France, vol. 93, pp. 273–299, 1965.

Proximal average is a good estimation of the true proximal gradient mapping

- Large scale optimization problem

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{R}^n}{\arg\min} \{ f(\boldsymbol{x}) \} \quad \text{with} \quad \begin{cases} f(\boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} f_k(\boldsymbol{x}) \\ \\ f_k(\boldsymbol{x}) = d(\boldsymbol{x}) + r_k(\boldsymbol{x}) \end{cases}$$

- Bauschke, Heinz H., et al. "The proximal average: basic theory." SIAM Journal on Optimization 19.2 (2008): 766-785.
- Yu, Yao-Liang. "Better approximation and faster algorithm using the proximal average." Advances in neural information processing systems. 2013.

Proximal average is a good estimation of the true proximal gradient mapping

- Large scale optimization problem

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min}\{f(\boldsymbol{x})\} \quad \text{with} \quad \begin{cases} f(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} f_k(\boldsymbol{x}) \\ f_k(\boldsymbol{x}) = d(\boldsymbol{x}) + r_k(\boldsymbol{x}) \end{cases}$$

- Proximal average

$$\mathsf{P}(\boldsymbol{x}) \triangleq \frac{1}{K}\sum_{k=1}^{K}\mathsf{P}_k(\boldsymbol{x}) \quad \text{with} \quad \mathsf{P}_k(\boldsymbol{x}) \triangleq \mathsf{prox}_{\gamma r_k}(\boldsymbol{x} - \gamma\nabla d(\boldsymbol{x})), \quad k \in [1, \dots, K]$$

- Bauschke, Heinz H., et al. "The proximal average: basic theory." SIAM Journal on Optimization 19.2 (2008): 766-785.
- Yu, Yao-Liang. "Better approximation and faster algorithm using the proximal average." Advances in neural information processing systems. 2013.

Stochastic proximal gradient method could approximate the proximal average

- Stochastic proximal gradient method

$$\mathsf{P}(\boldsymbol{x}) \triangleq \frac{1}{K}\sum_{k=1}^{K}\mathsf{P}_k(\boldsymbol{x})$$

Could be time consuming and computational resource demanding for K components

- H. Robbins and S. Monro, "A stochastic approximation method," The Annals of Mathematical Statistics, vol. 22, no. 3, pp. 400–407, September 1951.

Stochastic proximal gradient method could approximate the proximal average

- Stochastic proximal gradient method

$$\mathsf{P}(\boldsymbol{x}) \triangleq \frac{1}{K}\sum_{k=1}^{K}\mathsf{P}_k(\boldsymbol{x}) \quad \approx \quad \widehat{\mathsf{P}}(\boldsymbol{x}) \triangleq \frac{1}{B}\sum_{b=1}^{B}\mathsf{P}_{k_b}(\boldsymbol{x})$$

Could be time consuming and computational resource demanding for K components

$B \ll K$

- H. Robbins and S. Monro, "A stochastic approximation method," The Annals of Mathematical Statistics, vol. 22, no. 3, pp. 400–407, September 1951.

Stochastic proximal gradient method could approximate the proximal average

- Stochastic proximal gradient method

$$\mathsf{P}(\boldsymbol{x}) \triangleq \frac{1}{K}\sum_{k=1}^{K}\mathsf{P}_k(\boldsymbol{x}) \qquad \approx \qquad \widehat{\mathsf{P}}(\boldsymbol{x}) \triangleq \frac{1}{B}\sum_{b=1}^{B}\mathsf{P}_{k_b}(\boldsymbol{x})$$

Could be time consuming and computational resource demanding for K components

$B \ll K$

- H. Robbins and S. Monro, “A stochastic approximation method,” The Annals of Mathematical Statistics, vol. 22, no. 3, pp. 400–407, September 1951.

Stochastic proximal gradient method could approximate the proximal average

- Stochastic proximal gradient method

$$\mathsf{P}(\boldsymbol{x}) \triangleq \frac{1}{K}\sum_{k=1}^{K}\mathsf{P}_k(\boldsymbol{x}) \approx \widehat{\mathsf{P}}(\boldsymbol{x}) \triangleq \frac{1}{B}\sum_{b=1}^{B}\mathsf{P}_{k_b}(\boldsymbol{x})$$

Could be time consuming and computational resource demanding for K components

$B \ll K$

$$\boldsymbol{x}^t \leftarrow \widehat{\mathsf{P}}(\boldsymbol{x}^{t-1})$$

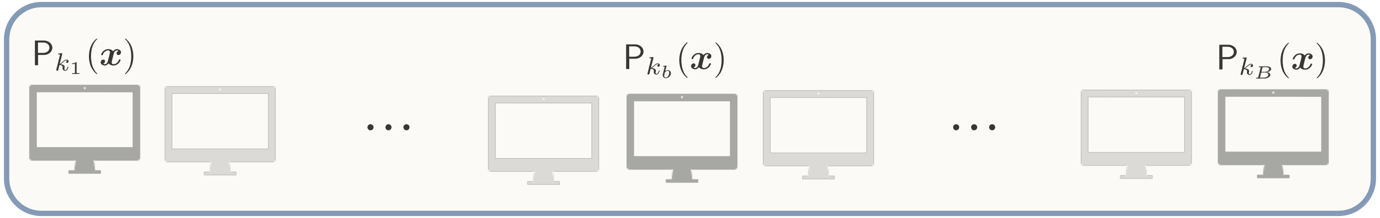
Stochastic proximal gradient method (SPGM)

- H. Robbins and S. Monro, “A stochastic approximation method,” The Annals of Mathematical Statistics, vol. 22, no. 3, pp. 400–407, September 1951.

SPGM could work parallelly

$$f(\boldsymbol{x}) = \frac{1}{K}\{\quad f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + \cdots + f_K(\boldsymbol{x}) \quad\}$$

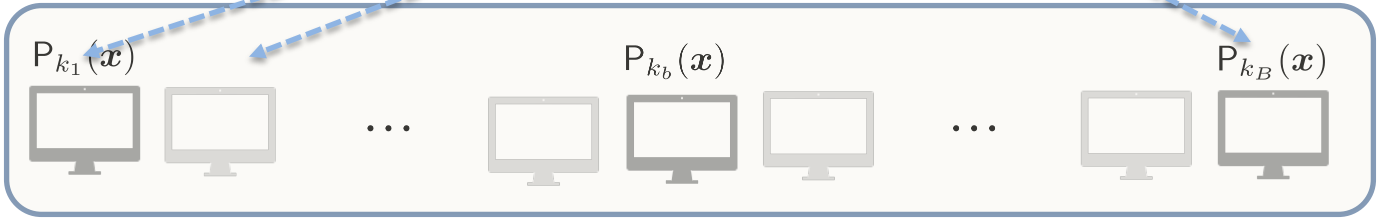
$\cdots$



SPGM could work parallelly

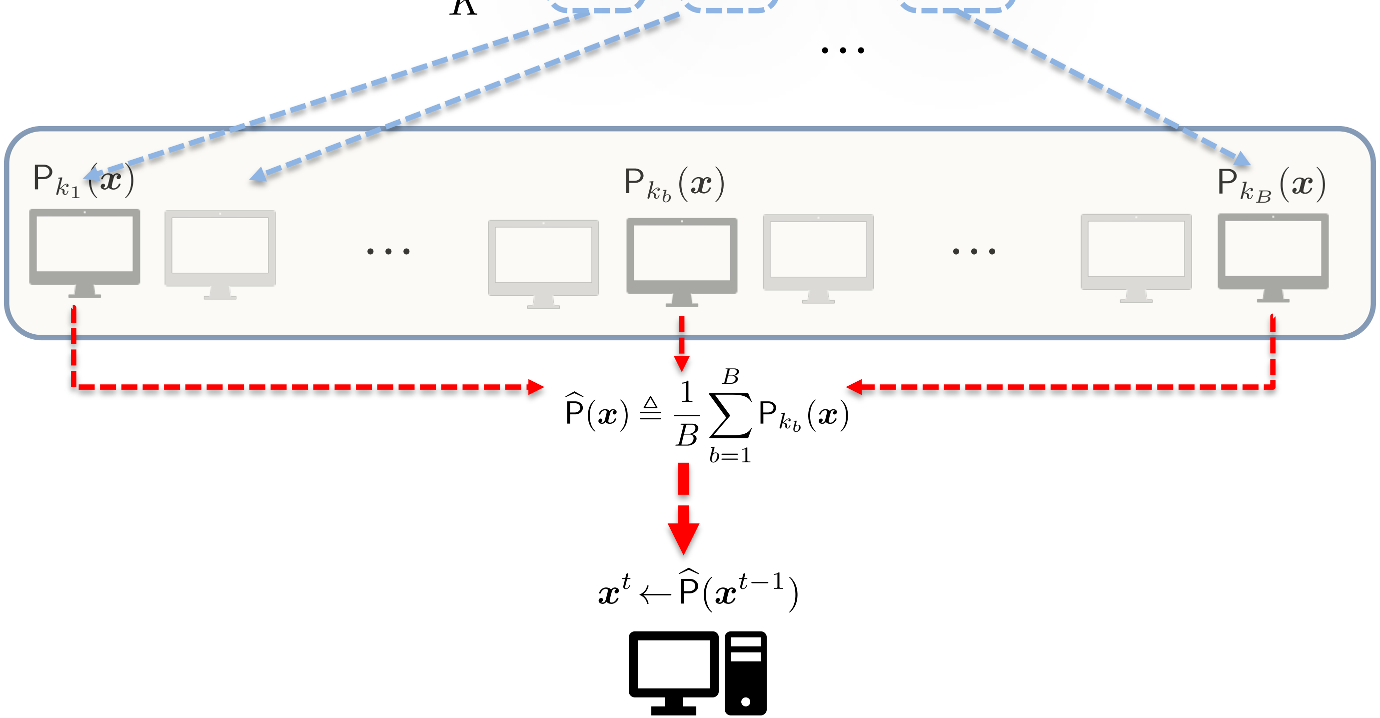
$$f(\boldsymbol{x}) = \frac{1}{K}\{\ f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + \cdots + f_K(\boldsymbol{x})\ \}$$

$\cdots$

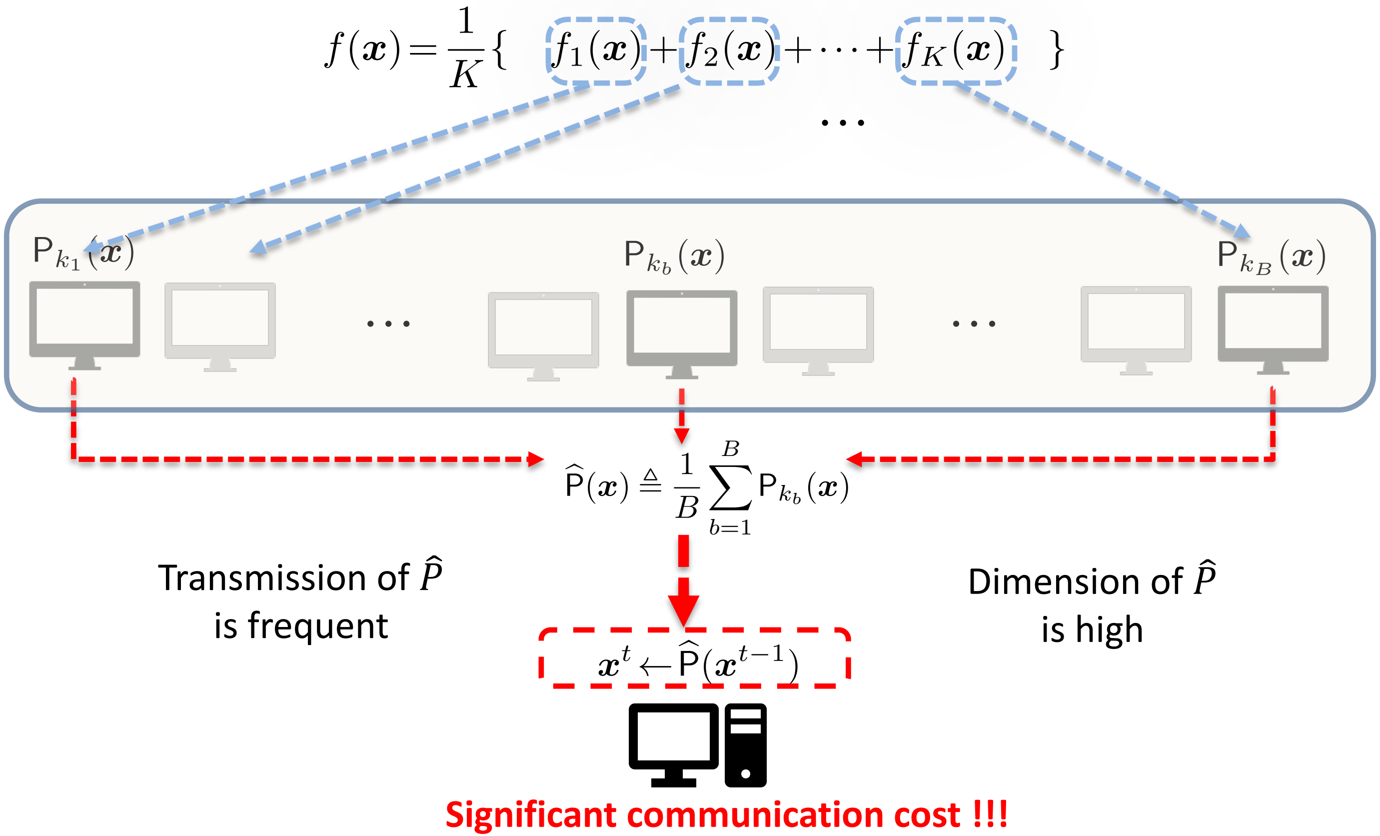


SPGM could work parallelly

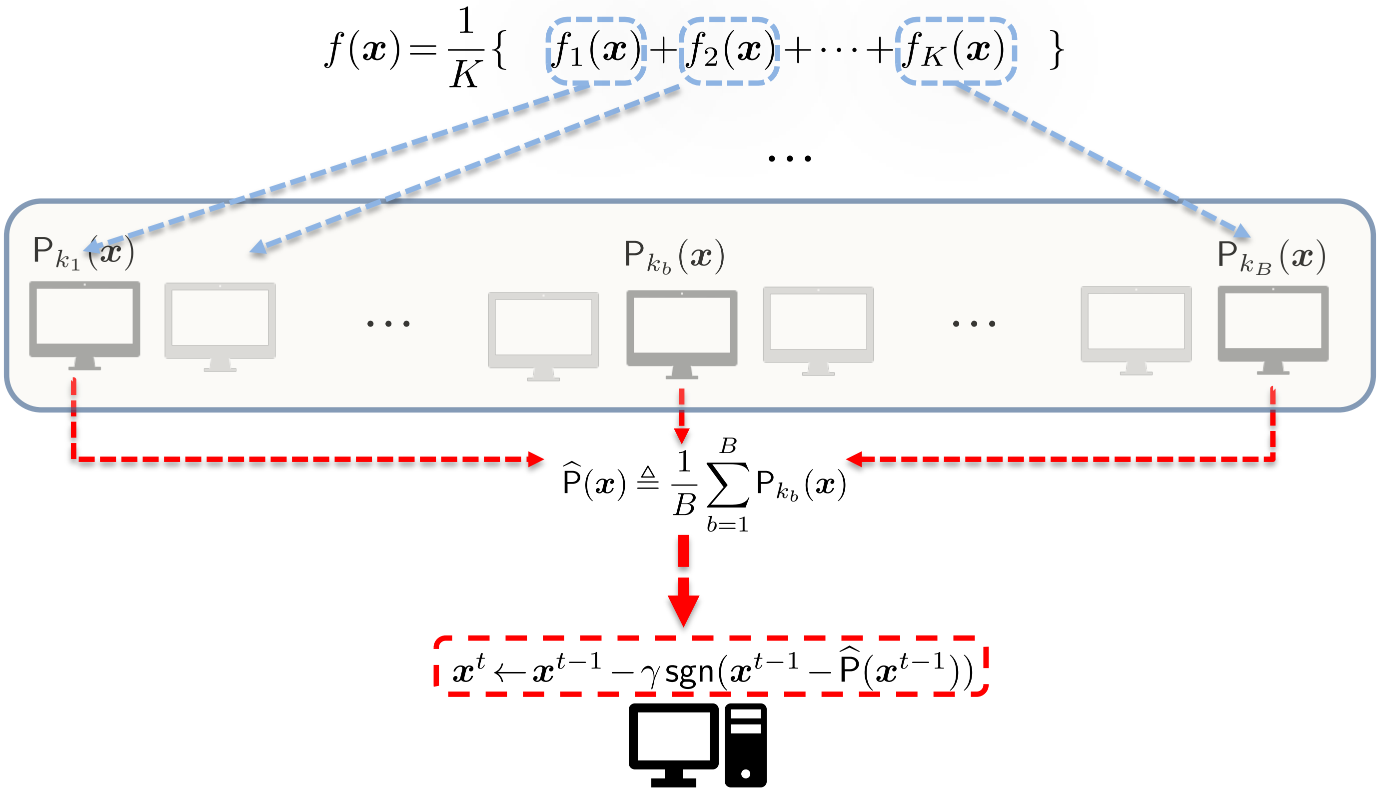
$$f(\boldsymbol{x}) = \frac{1}{K}\{ \; f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + \cdots + f_K(\boldsymbol{x}) \; \}$$



SPGM could work parallelly



signProx is more efficient than SPGM



- Seide, Frank, et al. "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns." Fifteenth Annual Conference of the International Speech Communication Association. 2014.
- J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, “signSGN: Compressed optimization for non-convex problems,” in Proc. 35th Int. Conf. Machine Learning (ICML), Stockholm, Sweden, July 2018.

SPGM uses the true direction to update while signProx only uses the sign

- Stochastic proximal gradient method (SPGM)
  - » Update rule: $\boldsymbol{x}^t \leftarrow \widehat{\mathsf{P}}(\boldsymbol{x}^{t-1})$

- 1-bit stochastic proximal gradient method (signProx)
  - » Update rule: $\boldsymbol{x}^t \leftarrow \boldsymbol{x}^{t-1} - \gamma\,\mathsf{sgn}(\boldsymbol{x}^{t-1} - \widehat{\mathsf{P}}(\boldsymbol{x}^{t-1}))$

signProx could achieve the comparable performance with SPGM

- Convergences rate tells how fast you can reduce you loss function

  - » Convergence conclusion for **signProx**:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{x}^{t-1}-\mathsf{P}(\boldsymbol{x}^{t-1})\|_1\right]\leq\frac{1}{\sqrt{T}}C_1=\mathcal{O}(\frac{1}{\sqrt{T}})$$

  - » Convergence conclusion for **SPGM**:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{x}^{t-1}-\mathsf{P}(\boldsymbol{x}^{t-1})\|_2^2\right]\leq\frac{1}{\sqrt{T}}C_2=\mathcal{O}(\frac{1}{\sqrt{T}})$$

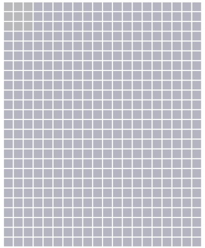
Check the performance of SPGM and signProx on a phase retrieval model

- **Nonconvex** phase retrieval task

$\boldsymbol{x}^* \in \mathbb{R}^{2500}$

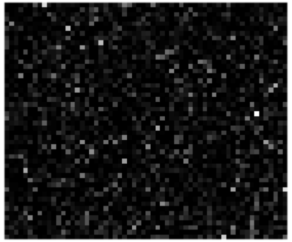
True image

$\boldsymbol{H} \in \mathbb{R}^{3000 \times 2500}$



Forward model

$\boldsymbol{y} = |\boldsymbol{H}\boldsymbol{x}^*|^2 \in \mathbb{R}^{3000}$



Observation

Check the performance of SPGM and signProx on a phase retrieval model

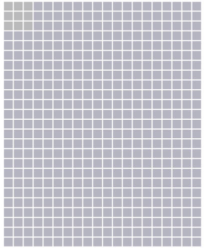
- **Nonconvex** phase retrieval task

Information about the sign is lost

$\boldsymbol{x}^* \in \mathbb{R}^{2500}$

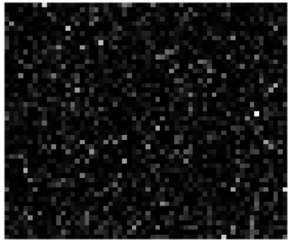
True image

$\boldsymbol{H} \in \mathbb{R}^{3000 \times 2500}$



Forward model

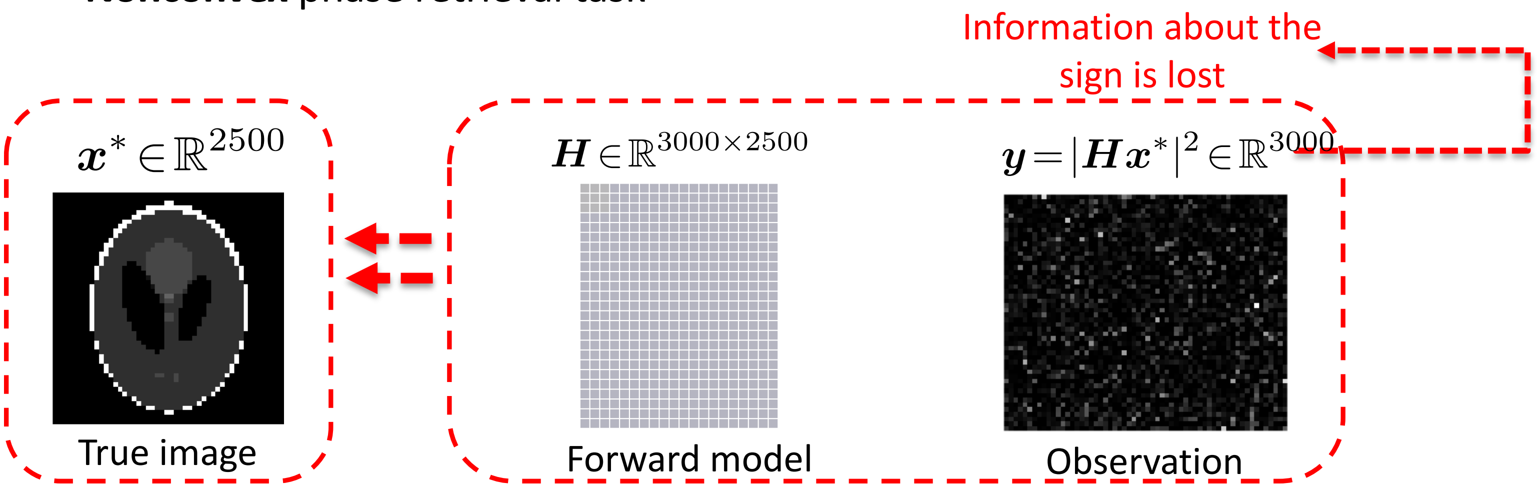
$\boldsymbol{y} = |\boldsymbol{H}\boldsymbol{x}^*|^2 \in \mathbb{R}^{3000}$



Observation

Check the performance of SPGM and signProx on a phase retrieval model

- **Nonconvex** phase retrieval task



Check the performance of SPGM and signProx on a phase retrieval model

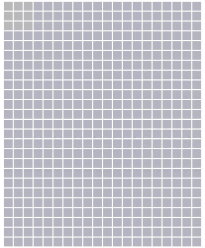
- **Nonconvex** phase retrieval task

Information about the sign is lost

$\boldsymbol{x}^* \in \mathbb{R}^{2500}$

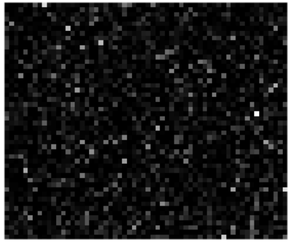
True image

$\boldsymbol{H} \in \mathbb{R}^{3000 \times 2500}$



Forward model

$\boldsymbol{y} = |\boldsymbol{H}\boldsymbol{x}^*|^2 \in \mathbb{R}^{3000}$



Observation

- Optimization objective function

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{R}^n}{\text{arg min}} \left\{ \frac{1}{2} \|\boldsymbol{y} - |\boldsymbol{H}\boldsymbol{x}|^2\|_2^2 + \text{TV}(\boldsymbol{x}) \right\}$$

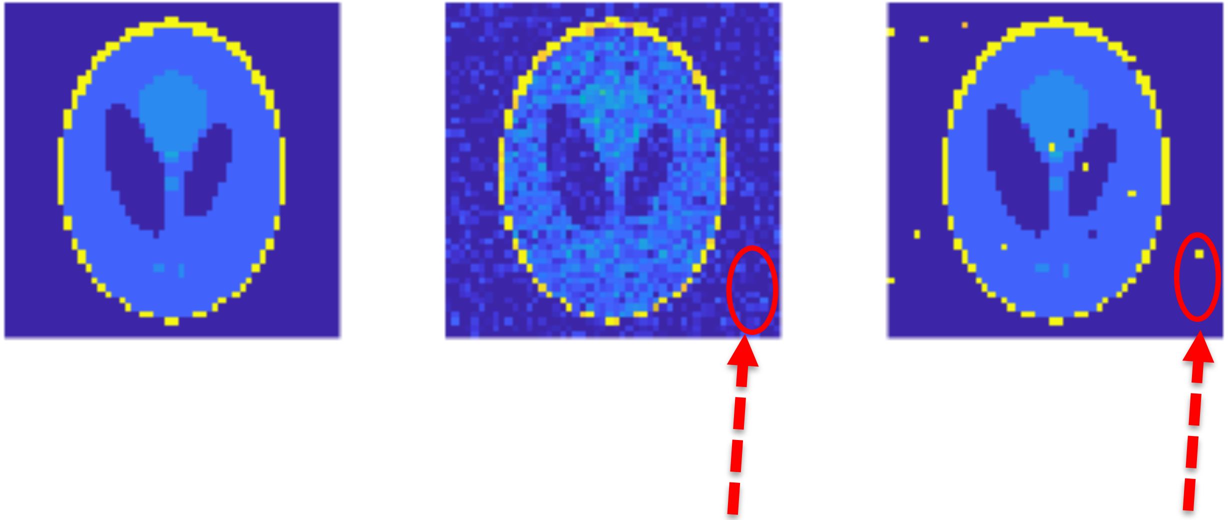
Simulate the stochasticity of the proximal gradient mapping by adding noise

- Stochastic simulation of proximal gradient mapping

Noise

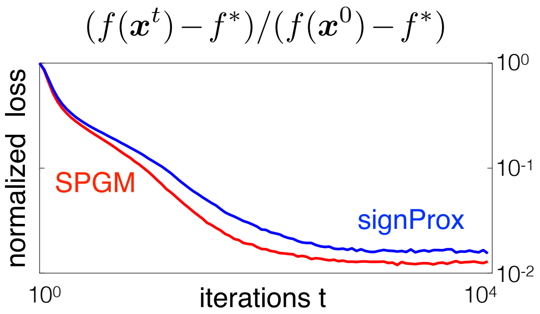
$$\widehat{\mathrm{P}} = \mathrm{P} + p(e)$$

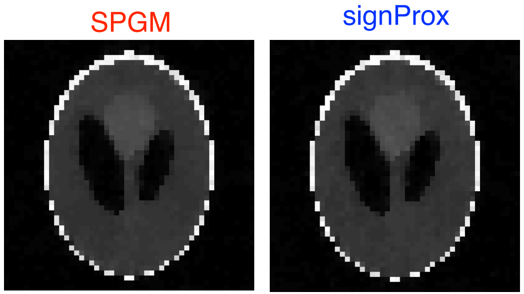
True proximal mapping	Dense stochasticity	Sparse stochasticity
P	$\hat{P}1$	$\hat{P}2$



signProx outperforms SPGM in the some sparse stochasticity scenario

Dense stochasticity scenario



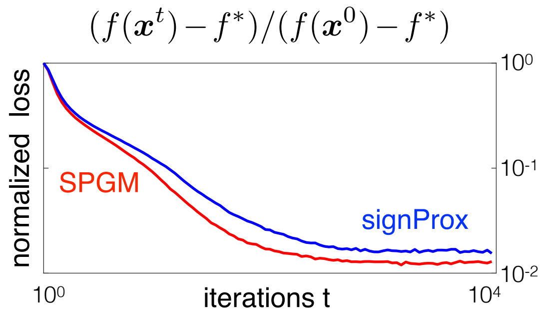


psnr = 19.12 dB

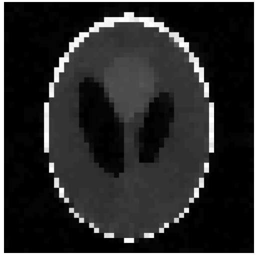
psnr = 18.60 dB

signProx outperforms SPGM in the some sparse stochasticity scenario

Dense stochasticity scenario

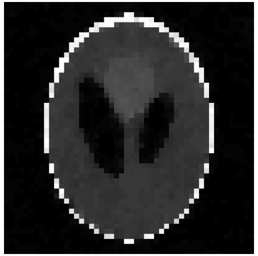


SPGM



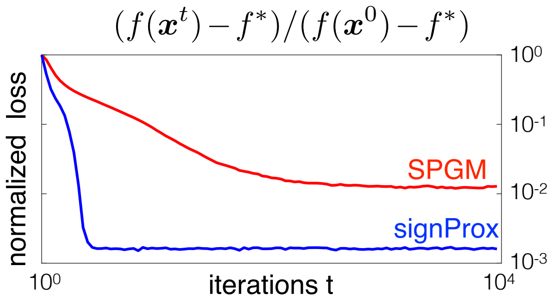
psnr = 19.12 dB

signProx

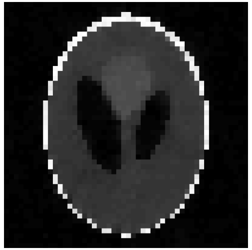


psnr = 18.60 dB

Sparse stochasticity scenario



SPGM



psnr = 19.01 dB

signProx

psnr = 21.15 dB

Conclusions

» Proposed a compressed proximal gradient method signProx to solve the low efficiency problem of SPGM in a large scale optimization scenario.

» Proved the convergence of the signProx under nonconvex assumption and showed it achieves the comparable theoretical performance with SPGM.

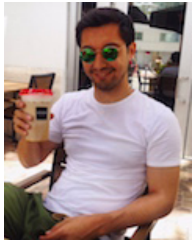
» Simulated a phase retrieval problem and showed signProx has a comparable performance with SPGM and in some scenario it even outperforms SPGM.

Acknowledgements

- Support
  - » National Science Foundation under Grant No. 1813910.

- Lab homepage
  - » https://cigroup.wustl.edu/

- Follow us
  - » https://twitter.com/wustlcig



Ulugbek Kamilov

Xiaojian Xu

Yu Sun

Jiaming Liu

Guangxiao Song

Thanks & questions?