A Normal Map-Based Proximal Stochastic Gradient Method

Workshop: Optimization and Learning Theory and Applications, CRM

May 29th

Andre Milzarek

SDS / CUHK-SZ

Acknowledgements

- Joint work with Junwen Qiu (NUS) and Li Jiang (CUHK-SZ).
- Preprint: `arXiv:2305.05828v2` (May '25; recently updated).

Main Contents

- Background and Problem Formulation.
- The Proximal Stochastic Gradient Method.
- The Proposed Method: norm-SGD.
- Complexity, Iterate Convergence, and Identification.
- Numerical Illustrations.

Background and Problem Formulation

Problem Formulation

We consider the composite optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \ \psi(\mathbf{x}) := f(\mathbf{x}) + \varphi(\mathbf{x})$$

Basic Assumptions:

- $\varphi : \mathbb{R}^d \to (-\infty, \infty]$ is a lower semicontinuous, proper, and convex function (can be nonsmooth).
- $f : \mathbb{R}^d \to \mathbb{R}$ is smooth (can be nonconvex and large-scale).

Typical Situation:

- f measures the error between an iterate and given data.
- φ is a regularization term that promotes special structure.
- Evaluation of f / ∇f is too expensive $\rightsquigarrow$ use stochastic techniques.

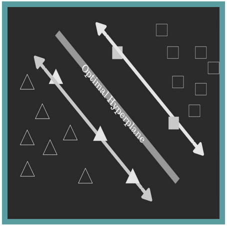
Examples and Applications

Examples:

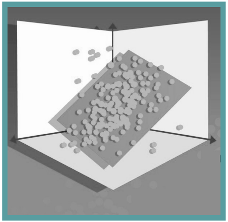
- Sparse / Low-rank optimization: $\varphi(\boldsymbol{x}) = \mu\|\boldsymbol{x}\|_1$, $\varphi(\boldsymbol{X}) = \mu\|\boldsymbol{X}\|_*$.
- Constrained optimization problems: $\varphi(\boldsymbol{x}) = \iota_{\mathcal{C}}(\boldsymbol{x})$.
- Expected / Empirical risk: $f(\boldsymbol{x}) = \mathbb{E}[F(\boldsymbol{x},\xi)]$, $f(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x};i)$.

⇝ Stochastic optimization techniques, nonsmoothness, and nonconvexity are prevalent in many large-scale and learning applications.

Neural Networks



Supervised Learning



Matrix Optimization

The Proximal Stochastic Gradient Method

Proximal Stochastic Gradient Descent

To solve $\min_{\boldsymbol{x}\in\mathbb{R}^d} \ \psi(\boldsymbol{x}) := f(\boldsymbol{x}) + \varphi(\boldsymbol{x})$, we (can) consider:

Proximal Stochastic Gradient Descent (prox-SGD):

$$\boldsymbol{x}^{k+1} = \mathrm{prox}_{\alpha_k\varphi}(\boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k)$$

- $\boldsymbol{g}^k \approx \nabla f(\boldsymbol{x}^k)$ is a stochastic approximation of $\nabla f(\boldsymbol{x}^k)$.
- $\{\alpha_k\}_k$ are suitable step sizes.
- $\mathrm{prox}_{\alpha\varphi}(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathbb{R}^d} \ \varphi(\boldsymbol{y}) + \frac{1}{2\alpha}\|\boldsymbol{x} - \boldsymbol{y}\|^2$ is the well-known proximity operator of φ.

Literature:

- Duchi and Singer '11, Xiao and Zhang '14, Nitanda '14, Ghadimi et al. '16, Atchadé et al. '17, Davis and Drusvyatskiy '19, ...

prox-SGD: What Do We Know?

Discussion:

⇝ Theory and convergence guarantees seem well-developed.

- If f (or ψ) is convex or strongly convex: analysis is close to **SGD** and the deterministic case.

  Ghadimi and Lan '13, Rosasco et al. '20, Khaled et al. '20, Patrascu and Irofti '21, Garrigos and Gower '24, ...

- Understanding convergence of **prox-SGD** if f is nonconvex was a long open problem.
- Finally addressed by Davis and Drusvyatskiy:

Complexity Bound for **prox-SGD:** (Davis and Drusvyatskiy '19)

$$\min_{k\in\{0,1,\dots,T-1\}} \mathbb{E}[\|F^{\lambda}_{\mathrm{nat}}(\boldsymbol{x}^k)\|^2] = \mathcal{O}(T^{-1/2})$$

prox-SGD: What Do We Know?

Discussion:

⇝ Theory and convergence guarantees seem well-developed.

- If f (or ψ) is convex or strongly convex: analysis is close to **SGD** and the deterministic case.

  Ghadimi and Lan '13, Rosasco et al. '20, Khaled et al. '20, Patrascu and Irofti '21, Garrigos and Gower '24, . . .

- Understanding convergence of **prox-SGD** if f is nonconvex was a long open problem.
- Finally addressed by Davis and Drusvyatskiy:

Complexity Bound for **prox-SGD:** (Davis and Drusvyatskiy '19)

$$\min_{k\in\{0,1,\dots,T-1\}} \mathbb{E}[\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\|^2] = \mathcal{O}(T^{-1/2})$$

prox-SGD: What Do We Know?

Discussion:

⇝ Theory and convergence guarantees seem well-developed.

- If f (or ψ) is convex or strongly convex: analysis is close to **SGD** and the deterministic case.

  Ghadimi and Lan '13, Rosasco et al. '20, Khaled et al. '20, Patrascu and Irofti '21, Garrigos and Gower '24, ...

- Understanding convergence of **prox-SGD** if f is nonconvex was a long open problem.
- Finally addressed by Davis and Drusvyatskiy:

Complexity Bound for **prox-SGD:** (Davis and Drusvyatskiy '19)

$$\min_{k \in \{0,1,\ldots,T-1\}} \mathbb{E}[\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\|^2] = \mathcal{O}(T^{-1/2})$$

prox-SGD: What Do We Know?

Natural Residual:

$$F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}) := \frac{1}{\lambda}(\boldsymbol{x} - \mathrm{prox}_{\lambda\varphi}(\boldsymbol{x} - \lambda\nabla f(\boldsymbol{x}))), \quad \lambda > 0,$$

is a popular stationarity measure for proximal methods:

$$\mathbf{0} \in \partial\psi(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) + \partial\varphi(\boldsymbol{x}) \quad \iff \quad F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}) = \mathbf{0}.$$

Stochastic Conditions:

- Complexity and convergence is based on the standard assumptions:

$$\mathbb{E}[\boldsymbol{g}^k \mid \mathcal{F}_k] = \nabla f(\boldsymbol{x}^k) \qquad \text{(unbiased)}$$

$$\mathbb{E}[\|\boldsymbol{g}^k - \nabla f(\boldsymbol{x}^k)\|^2 \mid \mathcal{F}_k] \leq \sigma^2 \qquad \text{(bounded variance)}$$

  (on some suitable underlying probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_k, \mathbb{P})$).
- Earlier results under variance reduction $\sigma \to 0$, (Ghadimi et al. '16, Xiao and Zhang '14, Reddi et al. '16).

prox-SGD: What Do We Know?

Natural Residual:

$$F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}) := \frac{1}{\lambda}(\boldsymbol{x} - \mathrm{prox}_{\lambda\varphi}(\boldsymbol{x} - \lambda \nabla f(\boldsymbol{x}))), \quad \lambda > 0,$$

is a popular stationarity measure for proximal methods:

$$\mathbf{0} \in \partial\psi(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) + \partial\varphi(\boldsymbol{x}) \quad \iff \quad F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}) = \mathbf{0}.$$

Stochastic Conditions:

- Complexity and convergence is based on the standard assumptions:

$$\mathbb{E}[\boldsymbol{g}^k \mid \mathcal{F}_k] = \nabla f(\boldsymbol{x}^k) \qquad \text{(unbiased)}$$
$$\mathbb{E}[\|\boldsymbol{g}^k - \nabla f(\boldsymbol{x}^k)\|^2 \mid \mathcal{F}_k] \leq \sigma^2 \qquad \text{(bounded variance)}$$

  (on some suitable underlying probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_k, \mathbb{P})$).

- Earlier results under variance reduction $\sigma \to 0$, (Ghadimi et al. '16, Xiao and Zhang '14, Reddi et al. '16).

prox-SGD: What Do We Know?

Asymptotic Convergence of **prox-SGD:** (Li and Milzarek '22)

$$\lim_{k\to\infty} \|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\| = 0 \quad \text{almost surely}$$

- Requires diminishing step sizes $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.
- φ needs to be Lipschitz on $\mathrm{dom}(\varphi)$.
- Stronger asymptotic guarantees in the convex case ($\rightsquigarrow$ folklore).

... seems pretty comprehensive

... anything open / missing?

... any major drawbacks of **prox-SGD**?

(... which might require / motivate some new research ☺)

prox-SGD: What Do We Know?

Asymptotic Convergence of **prox-SGD:** (Li and Milzarek '22)

$$\lim_{k\to\infty} \|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\| = 0 \quad \text{almost surely}$$

- Requires diminishing step sizes $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.
- φ needs to be Lipschitz on $\mathrm{dom}(\varphi)$.
- Stronger asymptotic guarantees in the convex case ($\rightsquigarrow$ folklore).

... seems pretty comprehensive

... anything open / missing?

... any major drawbacks of **prox-SGD**?

(... which might require / motivate some new research ☺)

Research Questions

The following questions have not been (re-)solved for **prox-SGD**:

- Can we guarantee asymptotic convergence without requiring global Lipschitz continuity of φ?
  - Is a full theory "**SGD** $\rightsquigarrow$ **prox-SGD**" possible?

    (*a bit boring*)
- Can we show $\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k)) \to 0$ (a.s.)?

  (*open*)
- Can we say more? Can we ensure $\boldsymbol{x}^k \to \boldsymbol{x}^*$ in the stochastic, nonconvex, nonsmooth case?

  (*open*)
- **prox-SGD** is known to not have a manifold identification property.

  (*limitation*)

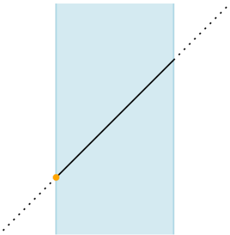
Manifold Identification

Failure of Identification: Illustration

Toy Example: (Duchi and Ruan '21)

$$\min_{x\in[-1,1]} \; f(x) := x.$$

- Global solution $x^* = -1$.
- Active set $\mathcal{M}_{x^*} = \{x \in [-1,1] : x = -1\} = \{x^*\}$.



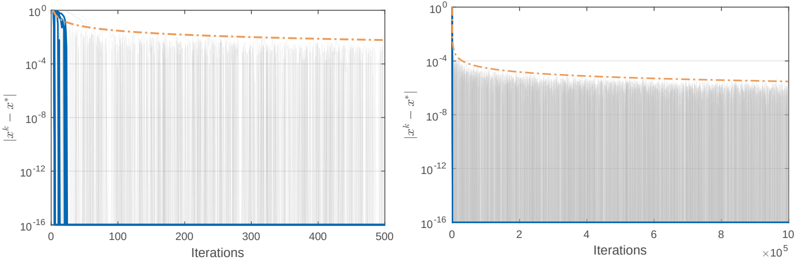
- We run **prox-SGD**,

$$x^{k+1} = \mathrm{proj}_{[-1,1]}(x^k - \alpha_k g^k),$$

with $g^k = f'(x^k) + e^k$, $e^k \sim \mathcal{N}(0,1)$, $\alpha_k = \frac{1}{k}$, and $x^0 = 100$.

- Comparison with **prox-GD** ($e^k = 0$, $\alpha_k \equiv \alpha = 1$).

Failure of Identification: Toy Example



- **Fig.: prox-GD** (■), **prox-SGD** (■), $k \mapsto \frac{3}{k}$ (■ ■)

> **Fact:** The iterates $\{x^k\}_k$ generated by **prox-SGD** satisfy
>
> $$\mathbb{P}(x^k \notin \mathcal{M}_{x^*}) \geq \eta \quad \text{for some } \eta > 0.$$

Active Manifold Identification

- (Active) manifolds $\mathcal{M}_{\boldsymbol{x}^*}$ can capture the smooth local sub-structure of the objective function ψ at a point $\boldsymbol{x}^*$.

Manifold Identification: There is $K \in \mathbb{N}$ such that

$$\boldsymbol{x}^k \in \mathcal{M}_{\boldsymbol{x}^*} \quad \forall\, k \geq K \quad \text{(almost surely)}.$$

Low-rank. Let $\varphi(\boldsymbol{X}) = \|\boldsymbol{X}\|_*$ and $\boldsymbol{X}^* \in \mathbb{R}^{m \times n}$ be given and set:

$$\mathcal{M}_{\boldsymbol{X}^*} = \{\boldsymbol{X} \in \mathbb{R}^{m \times n} : \text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{X}^*)\}.$$

The nuclear norm is smooth on $\mathcal{M}_{\boldsymbol{X}^*}$ (Vaiter et al. '17).

Remark: Once the (low-rank) sub-structure has been identified, more efficient algorithmic strategies can be used.

Active Manifold Identification

- (Active) manifolds $\mathcal{M}_{\boldsymbol{x}^*}$ can capture the smooth local sub-structure of the objective function ψ at a point $\boldsymbol{x}^*$.

Manifold Identification: There is $K \in \mathbb{N}$ such that

$$\boldsymbol{x}^k \in \mathcal{M}_{\boldsymbol{x}^*} \quad \forall\ k \geq K \quad \text{(almost surely)}.$$

Low-rank. Let $\varphi(\boldsymbol{X}) = \|\boldsymbol{X}\|_*$ and $\boldsymbol{X}^* \in \mathbb{R}^{m \times n}$ be given and set:

$$\mathcal{M}_{\boldsymbol{X}^*} = \{\boldsymbol{X} \in \mathbb{R}^{m \times n} : \text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{X}^*)\}.$$

The nuclear norm is smooth on $\mathcal{M}_{\boldsymbol{X}^*}$ (Vaiter et al. '17).

Remark: Once the (low-rank) sub-structure has been identified, more efficient algorithmic strategies can be used.

Active Manifold Identification

- Identification typically relies on the concept of partial smoothness and on the strict complementarity condition.

Partial Smoothness (for $\psi = f + \varphi$): (Lewis '03)

ψ is partly smooth at $\boldsymbol{x}^* \in \mathrm{dom}(\partial\varphi)$ relative to $\mathcal{M}_{\boldsymbol{x}^*}$ if:

- (Smoothness) $\mathcal{M}_{\boldsymbol{x}^*}$ is a C^2-manifold and $\psi|_{\mathcal{M}_{\boldsymbol{x}^*}}$ is C^2 near $\boldsymbol{x}^*$;
- (Sharpness) affine span of $\partial\psi(\boldsymbol{x}^*)$ is parallel to $N_{\mathcal{M}_{\boldsymbol{x}^*}}(\boldsymbol{x}^*)$;
- (Continuity) $\partial\psi$ restricted to $\mathcal{M}_{\boldsymbol{x}^*}$ is continuous at $\boldsymbol{x}^*$.

Theorem (Informal):

prox-GD has a manifold identification property.

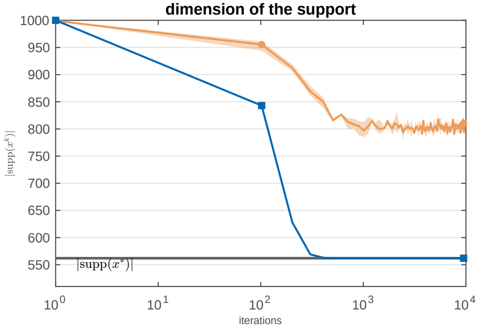
References: Lewis '03, Hare and Lewis '04, Lewis and Wright '08, Lee and Wright '12, Liang et al. '17, Poon et al. '18, ...

Failure of Identification: LASSO

Least-Squares with ℓ_1-**Regularizer:**

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \; f(\boldsymbol{x}) + \varphi(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \mu\|\boldsymbol{x}\|_1.$$

- **Fig.: prox-GD** (blue), **prox-SGD** (orange).
- Observed in (Xiao '10, Lee and Wright '12, Poon et al. '18, ...).



Solutions and Motivation

Enabling Identification of prox-SGD

Can stochastic proximal-type methods achieve identification?

Current Solutions and **Limitations:**

- Incorporate variance reduction or use averaging techniques: **RDA**, **SAGA**, **prox-SVRG**, **prox-STORM**.

⇝ Advantage: can work with fixed step size $\alpha_k \equiv \alpha$, variance vanishes.

- Most results limited to the (strongly) convex case; a.s. convergence $\boldsymbol{x}^k \to \boldsymbol{x}^*$ is often assumed as prerequisite.

References: Xiao '10, Lee and Wright '12, Poon et al. '18, Sun et al. '19, Duchi and Ruan '21, Huang and Lee '22, Dai et al. '23.

Enabling Identification of prox-SGD

Can stochastic proximal-type methods achieve identification?

Current Solutions and **Limitations:**

- Incorporate variance reduction or use averaging techniques: **RDA**, **SAGA**, **prox-SVRG**, **prox-STORM**.

$\rightsquigarrow$ Advantage: can work with fixed step size $\alpha_k \equiv \alpha$, variance vanishes.

- Most results limited to the (strongly) convex case; a.s. convergence $\boldsymbol{x}^k \to \boldsymbol{x}^*$ is often assumed as prerequisite.

References: Xiao '10, Lee and Wright '12, Poon et al. '18, Sun et al. '19, Duchi and Ruan '21, Huang and Lee '22, Dai et al. '23.

Observation

Can stochastic proximal-type methods achieve identification **without** variance reduction techniques?

Prox-SGD:

$$\boldsymbol{x}^{k+1} = \mathrm{prox}_{\alpha_k \varphi}(\boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k) \quad \text{with} \quad \boldsymbol{g}^k \approx \nabla f(\boldsymbol{x}^k).$$

- Diminishing step sizes, $\alpha_k \to 0$, are required to ensure convergence.
- Small α_k can harm identification properties.

How about keeping the proximal parameter constant?

$$\boldsymbol{x}^{k+1} = (1 - \alpha_k)\boldsymbol{x}^k + \alpha_k \mathrm{prox}_{\lambda\varphi}(\boldsymbol{x}^k - \lambda \boldsymbol{g}^k).$$

- **No**, this does not work (variance scaled with "α_k"). ☹

Observation

Can stochastic proximal-type methods achieve identification **without** variance reduction techniques?

Prox-SGD:

$$x^{k+1} = \mathrm{prox}_{\alpha_k \varphi}(x^k - \alpha_k g^k) \quad \text{with} \quad g^k \approx \nabla f(x^k).$$

- Diminishing step sizes, $\alpha_k \to 0$, are required to ensure convergence.
- Small α_k can harm identification properties.

How about keeping the proximal parameter constant?

$$x^{k+1} = (1 - \alpha_k)x^k + \alpha_k \mathrm{prox}_{\lambda\varphi}(x^k - \lambda g^k).$$

- **No**, this does not work (variance scaled with "α_k"). ☹

Observation

Can stochastic proximal-type methods achieve identification
without variance reduction techniques?

Prox-SGD:

$$\boldsymbol{x}^{k+1} = \mathrm{prox}_{\alpha_k \varphi}(\boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k) \quad \text{with} \quad \boldsymbol{g}^k \approx \nabla f(\boldsymbol{x}^k).$$

- Diminishing step sizes, $\alpha_k \to 0$, are required to ensure convergence.
- Small α_k can harm identification properties.

How about keeping the proximal parameter constant?

$$\boldsymbol{x}^{k+1} = (1 - \alpha_k)\boldsymbol{x}^k + \alpha_k \mathrm{prox}_{\lambda\varphi}(\boldsymbol{x}^k - \lambda \boldsymbol{g}^k).$$

- **No**, this does not work (variance scaled with "α_k"). ☹

Revisiting prox-GD

prox-GD:

$$x^{k+1} = \mathrm{prox}_{\lambda\varphi}(x^k - \lambda\nabla f(x^k)).$$

Introduce an auxiliary iterate z^k:

$$\begin{bmatrix} z^{k+1} = x^k - \lambda\nabla f(x^k), \\ x^{k+1} = \mathrm{prox}_{\lambda\varphi}(z^{k+1}). \end{bmatrix}$$

We rearrange

$$\begin{bmatrix} z^{k+1} = z^k - \alpha\cdot[\nabla f(x^k) + \lambda^{-1}(z^k - x^k)] \\ x^{k+1} = \mathrm{prox}_{\lambda\varphi}(z^{k+1}) \quad \text{with} \quad \alpha = \lambda. \end{bmatrix}$$

⇝ We also have $\lambda^{-1}(z^k - x^k) = \nabla \mathrm{env}_{\lambda\varphi}(z^k) \in \partial\varphi(x^k)$.

- **Idea:** Keep λ fixed and vary the parameter $\alpha \rightsquigarrow \alpha_k$.

Revisiting prox-GD

prox-GD:

$$\boldsymbol{x}^{k+1} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{x}^k - \lambda \nabla f(\boldsymbol{x}^k)).$$

Introduce an auxiliary iterate $\boldsymbol{z}^k$:

$$\begin{cases} \boldsymbol{z}^{k+1} = \boldsymbol{x}^k - \lambda \nabla f(\boldsymbol{x}^k), \\ \boldsymbol{x}^{k+1} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}^{k+1}). \end{cases}$$

We rearrange

$$\begin{cases} \boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \alpha \cdot [\nabla f(\boldsymbol{x}^k) + \lambda^{-1}(\boldsymbol{z}^k - \boldsymbol{x}^k)] \\ \boldsymbol{x}^{k+1} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}^{k+1}) \quad \text{with} \quad \alpha = \lambda. \end{cases}$$

$\rightsquigarrow$ We also have $\lambda^{-1}(\boldsymbol{z}^k - \boldsymbol{x}^k) = \nabla \mathrm{env}_{\lambda\varphi}(\boldsymbol{z}^k) \in \partial\varphi(\boldsymbol{x}^k)$.

- **Idea:** Keep λ fixed and vary the parameter $\alpha \rightsquigarrow \alpha_k$.

Revisiting prox-GD

prox-GD:

$$x^{k+1} = \mathrm{prox}_{\lambda\varphi}(x^k - \lambda\nabla f(x^k)).$$

Introduce an auxiliary iterate z^k:

$$\begin{cases} z^{k+1} = x^k - \lambda\nabla f(x^k), \\ x^{k+1} = \mathrm{prox}_{\lambda\varphi}(z^{k+1}). \end{cases}$$

We rearrange

$$\begin{cases} z^{k+1} = z^k - \alpha \cdot [\nabla f(x^k) + \lambda^{-1}(z^k - x^k)] \\ x^{k+1} = \mathrm{prox}_{\lambda\varphi}(z^{k+1}) \quad \text{with} \quad \alpha = \lambda. \end{cases}$$

$\rightsquigarrow$ We also have $\lambda^{-1}(z^k - x^k) = \nabla \mathrm{env}_{\lambda\varphi}(z^k) \in \partial\varphi(x^k)$.

- **Idea:** Keep λ fixed and vary the parameter $\alpha \rightsquigarrow \alpha_k$.

The Proposed Method: norm-SGD

Proposed Method

Normal Map-based Proximal SGD (norm-SGD):

- $\boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \alpha_k \cdot [\boldsymbol{g}^k + \lambda^{-1}(\boldsymbol{z}^k - \boldsymbol{x}^k)]$ (Normal map step)
- $\boldsymbol{x}^{k+1} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}^{k+1})$ (Proximal step)

- The normal map (Robinson '92) is defined as

$$F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) := \nabla f(\boldsymbol{x}) + \underbrace{\lambda^{-1}(\boldsymbol{z} - \boldsymbol{x})}_{\in \partial\varphi(\boldsymbol{x})} \quad \text{where} \quad \boldsymbol{x} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}).$$

Since $\psi = f + \varphi$, it holds that $F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) \in \partial\psi(\boldsymbol{x})$.

⇝ The $\boldsymbol{z}$-update can be seen as a special stochastic subgradient step!

- The normal map has been primarily used in variational inequalities and generalized equations (Facchinei and Pang '03).

Proposed Method

Normal Map-based Proximal SGD (norm-SGD):

- $\boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \alpha_k \cdot [\boldsymbol{g}^k + \lambda^{-1}(\boldsymbol{z}^k - \boldsymbol{x}^k)]$ (Normal map step)
- $\boldsymbol{x}^{k+1} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}^{k+1})$ (Proximal step)

- The normal map (Robinson '92) is defined as

$$F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) := \nabla f(\boldsymbol{x}) + \underbrace{\lambda^{-1}(\boldsymbol{z} - \boldsymbol{x})}_{\in \partial\varphi(\boldsymbol{x})} \quad \text{where} \quad \boldsymbol{x} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}).$$

  Since $\psi = f + \varphi$, it holds that $F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) \in \partial\psi(\boldsymbol{x})$.

$\rightsquigarrow$ The $\boldsymbol{z}$-update can be seen as a special stochastic subgradient step!

- The normal map has been primarily used in variational inequalities and generalized equations (Facchinei and Pang '03).

Proposed Method

Normal Map-based Proximal SGD (norm-SGD):

- $\boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \alpha_k \cdot [\boldsymbol{g}^k + \lambda^{-1}(\boldsymbol{z}^k - \boldsymbol{x}^k)]$ (Normal map step)
- $\boldsymbol{x}^{k+1} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}^{k+1})$ (Proximal step)

- The normal map (Robinson '92) is defined as

$$F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) := \nabla f(\boldsymbol{x}) + \underbrace{\lambda^{-1}(\boldsymbol{z} - \boldsymbol{x})}_{\in \partial\varphi(\boldsymbol{x})} \quad \text{where} \quad \boldsymbol{x} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}).$$

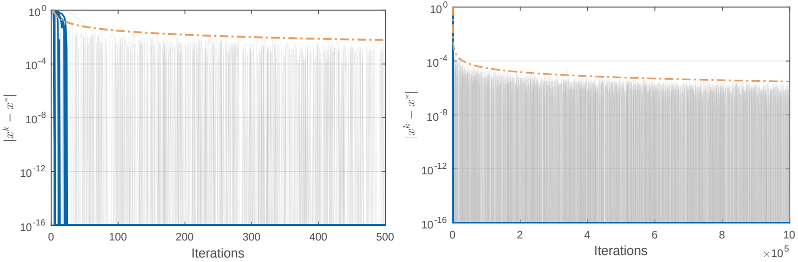
Since $\psi = f + \varphi$, it holds that $F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) \in \partial\psi(\boldsymbol{x})$.

$\rightsquigarrow$ The $\boldsymbol{z}$-update can be seen as a special stochastic subgradient step!

- The normal map has been primarily used in variational inequalities and generalized equations (Facchinei and Pang '03).

Does It Work?

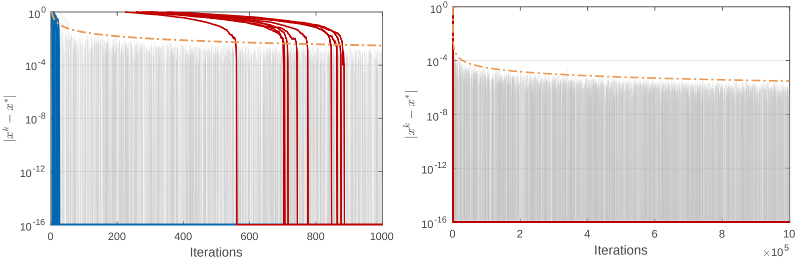
Let us revisit the earlier toy example:



- **Fig.: prox-GD** (▬), **prox-SGD** (▬), $k \mapsto \frac{3}{k}$ (▬ ▬)

Does It Work?

Let us revisit the earlier toy example:



- **Fig.: prox-GD** (■), **prox-SGD** (■), **norm-SGD** (■), and $k \mapsto \frac{3}{k}$ (■ ■) ☺

Why Does It Work? (Theory)

The Normal Map as Stationarity Measure

Normal Map (Robinson '92, Pang '93, ...)**:**

$$F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) = \nabla f(\boldsymbol{x}) + \lambda^{-1}(\boldsymbol{z} - \boldsymbol{x}) \quad \text{where} \quad \boldsymbol{x} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}), \quad \lambda > 0$$
$$= \nabla f(\mathrm{prox}_{\lambda\varphi}(\boldsymbol{z})) + \lambda^{-1}(\boldsymbol{z} - \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z})).$$

Comparison with $F_{\mathrm{nat}}^{\lambda}$:

$$F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) \in \nabla f(\boldsymbol{x}) + \partial\varphi(\boldsymbol{x}) = \partial\psi(\boldsymbol{x})$$

$$F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}) \in \nabla f(\boldsymbol{x}) + \partial\varphi(\boldsymbol{x}^{+}) \neq \partial\psi(\boldsymbol{x})$$

where $\boldsymbol{x}^{+} := \mathrm{prox}_{\lambda\varphi}(\boldsymbol{x} - \lambda\nabla f(\boldsymbol{x}))$.

The Normal Map as Stationarity Measure

Normal Map (Robinson '92, Pang '93, ...)**:**

$$F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) = \nabla f(\boldsymbol{x}) + \lambda^{-1}(\boldsymbol{z} - \boldsymbol{x}) \quad \text{where} \quad \boldsymbol{x} = \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}), \quad \lambda > 0$$

$$= \nabla f(\mathrm{prox}_{\lambda\varphi}(\boldsymbol{z})) + \lambda^{-1}(\boldsymbol{z} - \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z})).$$

Comparison with $F_{\mathrm{nat}}^{\lambda}$:

$$F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) \in \nabla f(\boldsymbol{x}) + \partial\varphi(\boldsymbol{x}) = \partial\psi(\boldsymbol{x})$$

$$F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}) \in \nabla f(\boldsymbol{x}) + \partial\varphi(\boldsymbol{x}^{+}) \neq \partial\psi(\boldsymbol{x})$$

where $\boldsymbol{x}^{+} := \mathrm{prox}_{\lambda\varphi}(\boldsymbol{x} - \lambda\nabla f(\boldsymbol{x}))$.

The Normal Map as Stationarity Measure

Stationarity:

- If $F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) = \mathbf{0}$, then $\boldsymbol{x} := \mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}) \in \mathrm{crit}(\psi) := \{\boldsymbol{x} : \mathbf{0} \in \partial\psi(\boldsymbol{x})\}$.
- If $F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}) = \mathbf{0}$, then $\boldsymbol{z} := \boldsymbol{x} - \lambda\nabla f(\boldsymbol{x})$ satisfies $F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z}) = \mathbf{0}$.

Relationship: For all $\boldsymbol{x} \in \mathrm{dom}(\partial\varphi)$ and $\boldsymbol{z} \in \mathbb{R}^d$, we have

$$\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x})\| \leq \mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x})), \quad \mathrm{dist}(\mathbf{0}, \partial\psi(\mathrm{prox}_{\lambda\varphi}(\boldsymbol{z}))) \leq \|F_{\mathrm{nor}}^{\lambda}(\boldsymbol{z})\|$$

(see, e.g., Drusvyatskiy and Lewis '18)

- **Remark:** In general: $\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x})\| \leq \varepsilon \quad \not\Longrightarrow \quad \mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x})) \leq \varepsilon$.

Complexity of norm-SGD

Basic Assumptions

(**A.1**) f is C^1; φ is convex, lsc., proper; $\inf_{\boldsymbol{x}\in\mathbb{R}^d}\ \psi(\boldsymbol{x}) > -\infty$.

(**A.2**) The gradient mapping ∇f is L-continuous on $\mathrm{dom}(\varphi)$.

(**A.3**) (Variance Bound). We assume $\mathbb{E}[\boldsymbol{g}^k \mid \mathcal{F}_k] = \nabla f(\boldsymbol{x}^k)$ and $\mathbb{E}[\|\boldsymbol{g}^k - \nabla f(\boldsymbol{x}^k)\|^2 \mid \mathcal{F}_k] \leq \sigma^2$ (for all k, a.s.).

Theorem: Iteration Complexity of **norm-SGD** (QJM '25)

Under (A.1)–(A.3), and if $\alpha_k \equiv \alpha \sim T^{-1/2}$, then:

$$\min_{k\in\{0,\dots,T-1\}} \mathbb{E}[\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k))^2] = \mathcal{O}(T^{-1/2}).$$

$\rightsquigarrow$ **prox-SGD**: $\min_{k\in\{0,\dots,T-1\}}\ \mathbb{E}[\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\|^2] = \mathcal{O}(T^{-1/2})$.

Complexity of norm-SGD

Basic Assumptions

- (**A.1**) f is C^1; φ is convex, lsc., proper; $\inf_{\boldsymbol{x}\in\mathbb{R}^d}\ \psi(\boldsymbol{x}) > -\infty$.
- (**A.2**) The gradient mapping ∇f is L-continuous on $\mathrm{dom}(\varphi)$.
- (**A.3**) (Variance Bound). We assume $\mathbb{E}[\boldsymbol{g}^k \mid \mathcal{F}_k] = \nabla f(\boldsymbol{x}^k)$ and $\mathbb{E}[\|\boldsymbol{g}^k - \nabla f(\boldsymbol{x}^k)\|^2 \mid \mathcal{F}_k] \leq \sigma^2$ (for all k, a.s.).

Theorem: Iteration Complexity of **norm-SGD** (QJM '25)

Under (A.1)–(A.3), and if $\alpha_k \equiv \alpha \sim T^{-1/2}$, then:

$$\min_{k\in\{0,\dots,T-1\}} \ \mathbb{E}[\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k))^2] = \mathcal{O}(T^{-1/2}).$$

$\rightsquigarrow$ **prox-SGD**: $\min_{k\in\{0,\dots,T-1\}}\ \mathbb{E}[\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\|^2] = \mathcal{O}(T^{-1/2})$.

Asymptotic Convergence

Theorem: Asymptotic Convergence of **norm-SGD** (QJM '25)

Under (A.1)–(A.3), and if $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2$, then:

$$\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k)) \to 0 \quad \text{and} \quad \psi(\boldsymbol{x}^k) \to \boldsymbol{\psi}^* \quad \text{almost surely;}$$

and we have $\mathbb{E}[\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k))^2] \to 0$ and $\mathbb{E}[\psi(\boldsymbol{x}^k)] \to \mathbb{E}[\boldsymbol{\psi}^*]$.

⇝ **prox-SGD**: $\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\| \to 0$ and $\psi(\boldsymbol{x}^k) \to \boldsymbol{\psi}^*$ almost surely.

- (Hare and Lewis '04): Let ψ be C^2-partly smooth at $\boldsymbol{x}^*$ rel. to $\mathcal{M}_{\boldsymbol{x}^*}$ with $\mathbf{0} \in \mathrm{ri}(\partial\psi(\boldsymbol{x}^*))$. Suppose $\boldsymbol{x}^k \to \boldsymbol{x}^*$ and $\psi(\boldsymbol{x}^k) \to \psi(\boldsymbol{x}^*)$. Then:

$$\boldsymbol{x}^k \in \mathcal{M}_{\boldsymbol{x}^*} \quad \forall\, k \text{ sufficiently large} \quad \Longleftrightarrow \quad \mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k)) \to 0.$$

⇝ **Manifold Identification:** We only need to show $\boldsymbol{x}^k \to \boldsymbol{x}^*$ (a.s.)!

Asymptotic Convergence

Theorem: Asymptotic Convergence of **norm-SGD** (QJM '25)

Under (A.1)–(A.3), and if $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2$, then:

$$\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k)) \to 0 \quad \text{and} \quad \psi(\boldsymbol{x}^k) \to \boldsymbol{\psi}^* \quad \text{almost surely;}$$

and we have $\mathbb{E}[\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k))^2] \to 0$ and $\mathbb{E}[\psi(\boldsymbol{x}^k)] \to \mathbb{E}[\boldsymbol{\psi}^*]$.

⇝ **prox-SGD**: $\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\| \to 0$ and $\psi(\boldsymbol{x}^k) \to \boldsymbol{\psi}^*$ almost surely.

- (Hare and Lewis '04): Let ψ be C^2-partly smooth at $\boldsymbol{x}^*$ rel. to $\mathcal{M}_{\boldsymbol{x}^*}$ with $\mathbf{0} \in \mathrm{ri}(\partial\psi(\boldsymbol{x}^*))$. Suppose $\boldsymbol{x}^k \to \boldsymbol{x}^*$ and $\psi(\boldsymbol{x}^k) \to \psi(\boldsymbol{x}^*)$. Then:

$$\boldsymbol{x}^k \in \mathcal{M}_{\boldsymbol{x}^*} \quad \forall\, k \text{ sufficiently large} \quad \iff \quad \mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k)) \to 0.$$

⇝ **Manifold Identification:** We only need to show $\boldsymbol{x}^k \to \boldsymbol{x}^*$ (a.s.)!

Asymptotic Convergence

Theorem: Asymptotic Convergence of **norm-SGD** (QJM '25)

Under (A.1)–(A.3), and if $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2$, then:

$$\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k)) \to 0 \quad \text{and} \quad \psi(\boldsymbol{x}^k) \to \boldsymbol{\psi}^* \quad \text{almost surely;}$$

and we have $\mathbb{E}[\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k))^2] \to 0$ and $\mathbb{E}[\psi(\boldsymbol{x}^k)] \to \mathbb{E}[\boldsymbol{\psi}^*]$.

$\rightsquigarrow$ **prox-SGD**: $\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\| \to 0$ and $\psi(\boldsymbol{x}^k) \to \boldsymbol{\psi}^*$ almost surely.

- (Hare and Lewis '04): Let ψ be C^2-partly smooth at $\boldsymbol{x}^*$ rel. to $\mathcal{M}_{\boldsymbol{x}^*}$ with $\mathbf{0} \in \mathrm{ri}(\partial\psi(\boldsymbol{x}^*))$. Suppose $\boldsymbol{x}^k \to \boldsymbol{x}^*$ and $\psi(\boldsymbol{x}^k) \to \psi(\boldsymbol{x}^*)$. Then:

$$\boldsymbol{x}^k \in \mathcal{M}_{\boldsymbol{x}^*} \quad \forall\, k \text{ sufficiently large} \quad \Longleftrightarrow \quad \mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k)) \to 0.$$

$\rightsquigarrow$ **Manifold Identification:** We only need to show $\boldsymbol{x}^k \to \boldsymbol{x}^*$ (a.s.)!

Iterate Convergence and Identification

Iterate Convergence

- Can we guarantee $\boldsymbol{x}^k \to \boldsymbol{x}^*$ (almost surely)?
- Can we show manifold identification of **norm-SGD**? (✓)

Core Idea:

- We apply extended Kurdyka-Łojasiewicz analysis techniques.

Assumptions for Iterate Convergence

(**B.1**) The function ψ is definable in an *o*-minimal structure.

(**B.2**) We assume $\mathbb{P}(\{\omega : \liminf_{k\to\infty} \|\boldsymbol{x}^k(\omega)\| < \infty\}) = 1$.

- Semialgebraic and globally subanalytic functions and functions in log-exp structures are definable.
- **Literature:** Łojasiewicz '65, '93, Kurdyka '98, van den Dries '97, Attouch and Bolte '09, . . .

Iterate Convergence

- Can we guarantee $\boldsymbol{x}^k \to \boldsymbol{x}^*$ (almost surely)?
- Can we show manifold identification of **norm-SGD**? ($\checkmark$)

Core Idea:

- We apply extended Kurdyka-Łojasiewicz analysis techniques.

Assumptions for Iterate Convergence

(**B.1**) The function ψ is definable in an *o*-minimal structure.

(**B.2**) We assume $\mathbb{P}(\{\omega : \liminf_{k\to\infty} \|\boldsymbol{x}^k(\omega)\| < \infty\}) = 1$.

- Semialgebraic and globally subanalytic functions and functions in log-exp structures are definable.
- **Literature:** Łojasiewicz '65, '93, Kurdyka '98, van den Dries '97, Attouch and Bolte '09, ...

Iterate Convergence and Manifold Identification

Theorem: Iterate Convergence (QJM '25)

Let (A.1)–(A.3), (B.1)–(B.2) hold and assume

$$\sum_{k=0}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k^2 \Big(\sum_{i=0}^{k} \alpha_i\Big)^p < \infty$$

for some $p > 1$. Then, $\lim_{k\to\infty} \boldsymbol{x}^k = \boldsymbol{x}^* \in \mathrm{crit}(\psi)$ almost surely.

- Holds for step sizes $\alpha_k \sim k^{-\gamma}$ with $\gamma \in (\frac{2}{3}, 1]$.

Theorem: Manifold Identification (QJM '25)

... in addition, if ψ is C^2-partly smooth at $\boldsymbol{x}^*$ and if $\mathbf{0} \in \mathrm{ri}(\partial\psi(\boldsymbol{x}^*))$ for almost every ω, then:

$$\boldsymbol{x}^k \in \mathcal{M}_{\boldsymbol{x}^*} \quad \text{for all } \boldsymbol{k} \text{ large, almost surely.}$$

Proof Snippets — I

- Measure descent via a merit function (Ouyang and Milzarek '21):

$$H_\xi(\boldsymbol{z}) := \psi(\mathrm{prox}_{\lambda\varphi}(\boldsymbol{z})) + \xi\|F^\lambda_{\mathrm{nor}}(\boldsymbol{z})\|^2, \quad \lambda, \xi > 0.$$

⇝ This allows us to leverage the unbiasedness in the $\boldsymbol{z}$-updates!

- Analysis of H_ξ along natural time scales $\sum_{i=k}^{n-1} \alpha_i$. For $\tau > 0$, we define the time indices $\{t_k\}_k$ via $t_0 = 0$:

$$t_{k+1} := \varsigma(t_k, \tau) \quad \text{where} \quad \varsigma(k, \tau) := \sup\{n \geq k : \textstyle\sum_{i=k}^{n-1} \alpha_i \leq \tau\}.$$

(Ljung '77, Benveniste et al. '92, Kushner, Yin '03, Tadić '15, ...).

Proof Snippets — II

- Time window-based approximate descent:

$$H_\xi(\boldsymbol{z}^{t_{k+1}}) - H_\xi(\boldsymbol{z}^{t_k}) \le -C_1\|F^\lambda_{\mathrm{nor}}(\boldsymbol{z}^{t_k})\|^2 + C_2\boldsymbol{s}_k^2.$$

- Aggregated error $\boldsymbol{s}_k := \max_{t_k < j \le t_{k+1}} \|\sum_{i=t_k}^{j-1} \alpha_i[\boldsymbol{g}^i - \nabla f(\boldsymbol{x}^i)]\|$ are controllable via the Burkholder-Davis-Gundy inequality.

- Convergence behavior of iterates $j \in (t_k, t_{k+1})$ can be recovered via Gronwall's inequality.

- Use a specialized KL inequality to handle the additional error terms in the descent condition.

Related Work

Traditional KL-Framework:

- Absil et al. '05, Attouch and Bolte '09, Attouch et al. '10, '13, Bolte et al. '10, '14, Frankel et al. '15, ...

KL-Results for SGD and RR:

- Tadić '09, '15: step sizes $\{\alpha_k\}_k$ with $\alpha_k = \frac{\alpha}{(k+\beta)^\gamma}$, $\gamma \in (\frac{3}{4}, 1)$,

  $$|f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)| = \mathcal{O}(k^{-p}), \quad \|\boldsymbol{x}^k - \boldsymbol{x}^*\| = \mathcal{O}(k^{-q}), \quad k \to \infty$$

  a.s. on $\{\omega : \sup_k \|\boldsymbol{x}^k(\omega)\| < \infty\}$ where $p \in (0, 1]$, $q \in (0, \frac{1}{2}]$.
- **Related:** Benaïm '18, Dereich and Kassing '21, Chouzenoux et al. '23; **RR:** Li et al. '21; **SGD** with **momentum:** Qiu et al. '24.

Global KL / PL: Karimi et al. '16, Gadat, Panloup '17, Wojtowytsch '23, Fatkhullin et al. '22, ...

In Expectation: Driggs et al. '21 (Stochastic PALM with VR), ...

Related Work

Traditional KL-Framework:

- Absil et al. '05, Attouch and Bolte '09, Attouch et al. '10, '13, Bolte et al. '10, '14, Frankel et al. '15, ...

KL-Results for SGD and RR:

- Tadić '09, '15: step sizes $\{\alpha_k\}_k$ with $\alpha_k = \frac{\alpha}{(k+\beta)^\gamma}$, $\gamma \in (\frac{3}{4}, 1)$,

  $$|f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)| = \mathcal{O}(k^{-p}), \quad \|\boldsymbol{x}^k - \boldsymbol{x}^*\| = \mathcal{O}(k^{-q}), \quad k \to \infty$$

  a.s. on $\{\omega : \sup_k \|\boldsymbol{x}^k(\omega)\| < \infty\}$ where $p \in (0, 1]$, $q \in (0, \frac{1}{2}]$.
- **Related:** Benaïm '18, Dereich and Kassing '21, Chouzenoux et al. '23; **RR:** Li et al. '21; **SGD** with **momentum:** Qiu et al. '24.

Global KL / PL: Karimi et al. '16, Gadat, Panloup '17, Wojtowytsch '23, Fatkhullin et al. '22, ...

In Expectation: Driggs et al. '21 (Stochastic PALM with VR), ...

Numerical Illustrations

Experiment: Sparse + Low-rank Recovery

We consider the application:

$$\min_{X,Y\in\mathbb{R}^{m\times n}} \ \frac{1}{2}\|X+Y-M\|_F^2+\nu_1\|X\|_*+\nu_2\|Y\|_1, \quad \nu_1,\nu_2>0.$$

Background and Remarks:

- M is a video clip; each column M_i of M is a vectorized frame.
- The model aims to decompose M into a low-rank background X and a sparse component Y (for movements).
- We test **prox-SGD** and **norm-SGD** on a video $M \in [0,1]^{230\,400\times 351}$.
- We use the stochastic gradients $g^k = \frac{1}{|S_k|}\sum_{i\in S_k}\nabla f_i(X^k,Y^k)$ where $f_i(X,Y)=\frac{n}{2}\|X_i+Y_i-M_i\|^2$ and $|S_k|=8$.

$\rightsquigarrow$ Each stochastic gradient only accesses 8 frames of M each iteration.

- We use $\alpha_k \sim 1/k^{3/4}$, $\lambda\in\{1,2\}$, and $\nu_1=150$, $\nu_2=0.25$.

Experiment: Sparse + Low-rank Recovery

We consider the application:

$$\min_{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{m\times n}} \ \frac{1}{2}\|\boldsymbol{X}+\boldsymbol{Y}-\boldsymbol{M}\|_F^2+\nu_1\|\boldsymbol{X}\|_*+\nu_2\|\boldsymbol{Y}\|_1, \quad \nu_1,\nu_2>0.$$

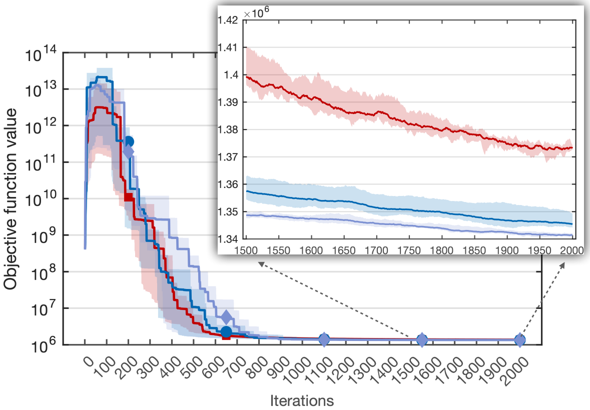
Background and Remarks:

- $\boldsymbol{M}$ is a video clip; each column $\boldsymbol{M}_i$ of $\boldsymbol{M}$ is a vectorized frame.
- The model aims to decompose $\boldsymbol{M}$ into a low-rank background $\boldsymbol{X}$ and a sparse component $\boldsymbol{Y}$ (for movements).
- We test **prox-SGD** and **norm-SGD** on a video $\boldsymbol{M}\in[0,1]^{230\,400\times 351}$.
- We use the stochastic gradients $\boldsymbol{g}^k=\frac{1}{|S_k|}\sum_{i\in S_k}\nabla f_i(\boldsymbol{X}^k,\boldsymbol{Y}^k)$ where $f_i(\boldsymbol{X},\boldsymbol{Y})=\frac{n}{2}\|\boldsymbol{X}_i+\boldsymbol{Y}_i-\boldsymbol{M}_i\|^2$ and $|S_k|=8$.

$\rightsquigarrow$ Each stochastic gradient only accesses 8 frames of $\boldsymbol{M}$ each iteration.

- We use $\alpha_k\sim 1/k^{3/4}$, $\lambda\in\{1,2\}$, and $\nu_1=150$, $\nu_2=0.25$.

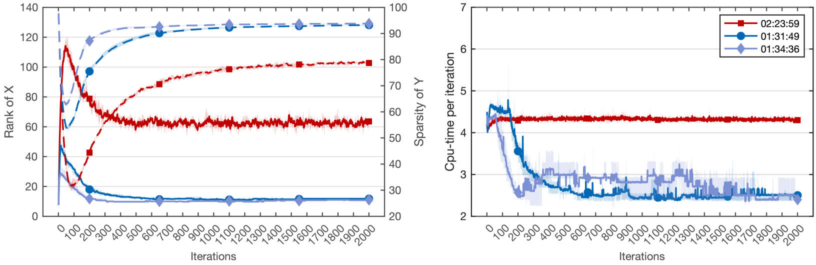
Experiment: Sparse + Low-rank Recovery



- **Fig.:** Plot of objective function values.

norm-SGD, $\lambda = 1$ (■), **norm-SGD**, $\lambda = 2$ (■), **prox-SGD** (■).

Experiment: Sparse + Low-rank Recovery

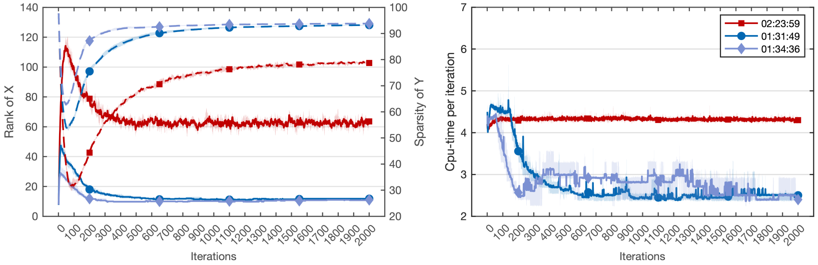


- **Fig.:** Left: rank (solid line) & sparsity (dashed line); Right: cpu-time per iteration.
- **prox-SGD** spent 57% more time than **norm-SGD**.

norm-SGD, $\lambda = 1$ (■), **norm-SGD**, $\lambda = 2$ (■), **prox-SGD** (■).

⇝ More results and experiments in the paper.

Experiment: Sparse + Low-rank Recovery



- **Fig.:** Left: rank (solid line) & sparsity (dashed line); Right: cpu-time per iteration.
- **prox-SGD** spent 57% more time than **norm-SGD**.

norm-SGD, $\lambda = 1$ (■), **norm-SGD**, $\lambda = 2$ (■), **prox-SGD** (■).

⇝ More results and experiments in the paper.

Summary

Conclusions

Take-Away: New normal map-based perspective & KL-analysis techniques for stochastic methods.

$\rightsquigarrow$ Can be applied to many other contexts and problems: random reshuffling, distributed algorithms, ...

Theory	**prox-SGD**	**norm-SGD**
Complexity	$\mathcal{O}(T^{-1/2})$	$\mathcal{O}(T^{-1/2})$
Asymp. Conv.	$\|F_{\mathrm{nat}}^{\lambda}(\boldsymbol{x}^k)\| \to 0$	$\mathrm{dist}(\mathbf{0}, \partial\psi(\boldsymbol{x}^k)) \to 0$
Iter. Conv.	✗ (?)	$\boldsymbol{x}^k \to \boldsymbol{x}^*$ a.s.
Identification	✗	$\boldsymbol{x}^k \in \mathcal{M}_{\boldsymbol{x}^*}$ a.s.

Joint work with **Junwen Qiu** and **Li Jiang**:

"A Normal Map-Based Proximal Stochastic Gradient Method: Convergence and Identification Properties"

Thank you very much! ☺