Advanced Topics in Data Sciences

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 5: Stochastic gradient methods

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-731 (Spring 2016)

lions@epfl

MARIE CURIE ACTIONS
FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION
erc
EPFL

License Information for Mathematics of Data Slides

- This work is released under a Creative Commons License with the following terms:
- **Attribution**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- **Non-Commercial**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- **Share Alike**
  - The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- Full Text of the License

Outline

- This lecture
  1. Stochastic projected gradient method
  2. Stochastic projected gradient method with averaging
  3. Stochastic proximal gradient method
  4. Stochastic proximal gradient method with averaging
  5. Accelerated stochastic proximal gradient methods

Recommended reading materials

1. V. Cevher; S. Becker, and M. Schmidt. Convex optimization for big data. *IEEE Signal Process. Mag.*, vol. 31, pp. 32–43, 2014.
2. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, vol. 19, pp. 1574–1609, 2008.

What is this class about?

Gradient method

Choose a starting point $\mathbf{x}^0$ and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k)$$

where γ_k is a step-size to be chosen so that $\mathbf{x}^k$ converges to $\mathbf{x}^\star$.

What is this class about?

Gradient method

Choose a starting point $\mathbf{x}^0$ and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k)$$

where γ_k is a step-size to be chosen so that $\mathbf{x}^k$ converges to $\mathbf{x}^\star$.

Stochastic gradient method

Let $G(\mathbf{x}^k, \theta_k)$ be an unbiased estimate of the gradient $\nabla f(\mathbf{x}^k)$, i.e.,

$$\mathbb{E}_\theta[G(\mathbf{x}^k, \theta_k)] = \nabla f(\mathbf{x}^k),$$

where θ_k is a random vector whose probability distribution is supported on set Θ.

Choose a starting point $\mathbf{x}^0$ and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k)$$

where γ_k is a step-size to be chosen so that $\mathbf{x}^k$ converges to $\mathbf{x}^\star$.

Motivation: Big n

Problem (Least squares)

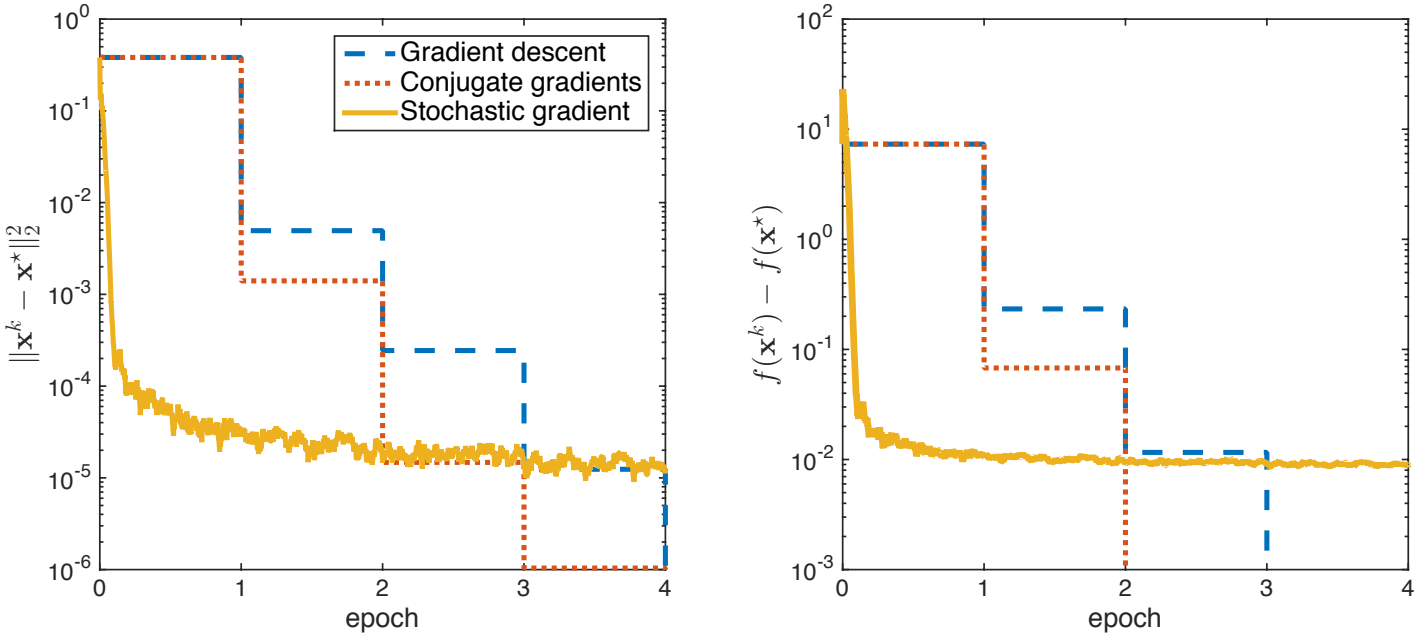
Given $\mathbf{A} \in \mathbb{R}^{n \times p}$ *and* $\mathbf{b} \in \mathbb{R}^n$ *where* $n \gg p$*, solve:*

$$f^\star := \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \right\}. \tag{1}$$

Complexity per iteration for gradient method

- Evaluating $\nabla f(\mathbf{x}^k) = \mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{b})$ requires one $\mathbf{A}\mathbf{x}$ and one $\mathbf{A}^T\mathbf{y}$.
- **Optional**: Evaluating $L = \|\mathbf{A}^T\mathbf{A}\|$ (spectral norm) - via **power iterations** (e.g., 50 iterations, each iteration requires one $\mathbf{A}\mathbf{x}$ and one $\mathbf{A}^T\mathbf{y}$).

Example: Least squares



Synthetic data generation

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$.
- $n = 10^4$, $p = 10^2$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$ where $\mathbf{w}$ is Gaussian white noise. SNR is 30dB.

Motivation: Statistical learning

A basic statistical learning model [1]

A statistical learning model consists of the following three elements.

1. A sample of i.i.d. random variables $(\mathbf{a}_i, b_i) \in \mathcal{A} \times \mathcal{B}$, $i = 1, \ldots, n$, following an *unknown* probability distribution $\mathbb{P}$.
2. A class (set) $\mathcal{F}$ of functions $f : \mathcal{A} \to \mathcal{B}$.
3. A loss function $L : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$.

Definition

Let $(\mathbf{a}, b)$ follow the probability distribution $\mathbb{P}$ and be independent of $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_n, b_n)$. Then, the risk corresponding to any $f \in \mathcal{F}$ is its expected loss:

$$R(f) := \mathbb{E}_{(\mathbf{a},b)} \left[L(f(\mathbf{a}), b) \right].$$

Statistical learning seeks to find a $f^\star \in \mathcal{F}$ that minimizes the risk, i.e., it solves

$$f^\star \in \arg\min_f \left\{ R(f) : f \in \mathcal{F} \right\}.$$

Many problems in machine learning cast into this formulation!

Example: Learning from a training set

- Data can be **decentralized**, or even **streaming**.
- By the law of large numbers, we can expect that for each $f \in \mathcal{F}$,

$$R(f) := \mathbb{E}\left[L(\mathbf{a}, b)\right] \approx \frac{1}{n} \sum_{i=1}^{n} L(f(\mathbf{a}_i), b_i)$$

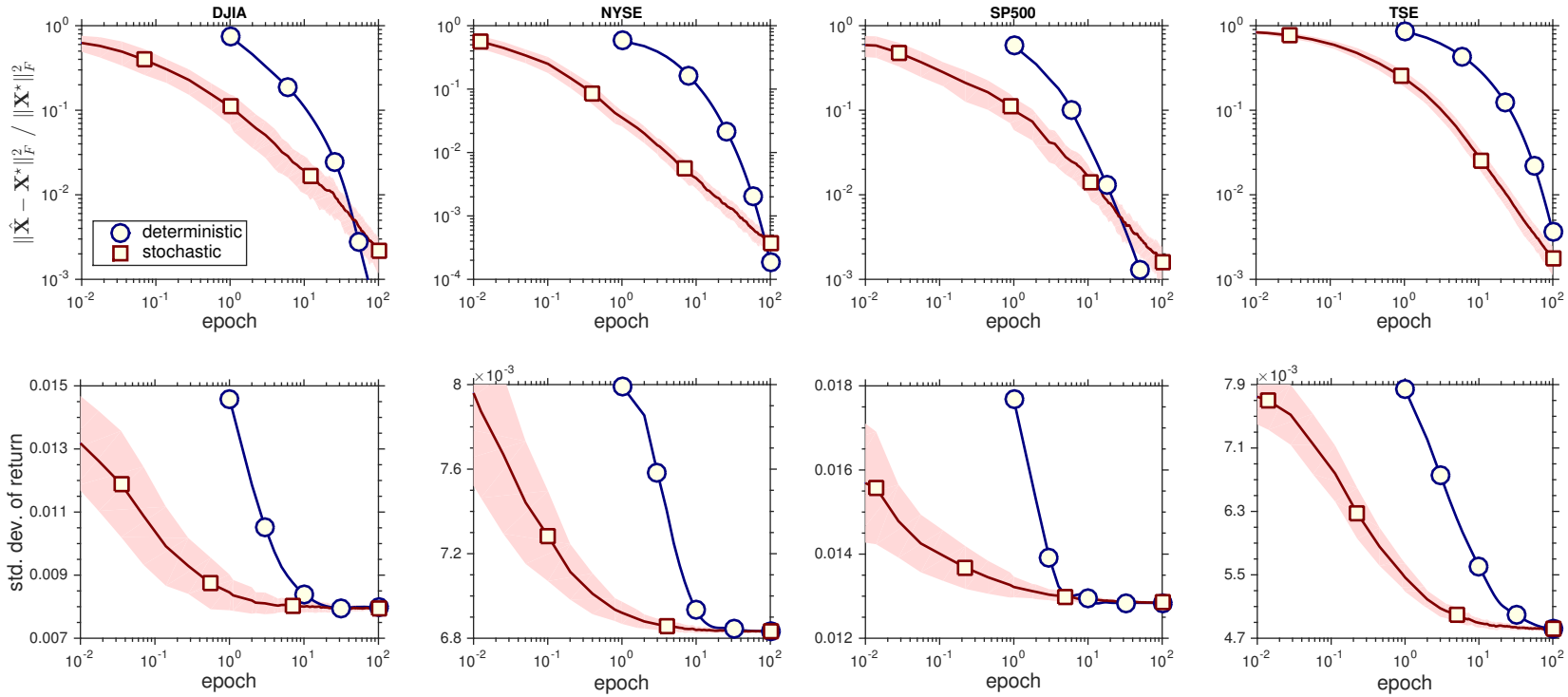
when n is large enough, with high probability.

Empirical risk minimization (ERM) [1]

We approximate $f^\star$ by minimizing the *empirical average of the loss* instead of the risk. That is, we consider the optimization problem

$$\hat{f}_n \in \arg\min_{f} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(f(\mathbf{a}_i), b_i) : f \in \mathcal{F} \right\}.$$

Example: Markowitz portfolio optimization [2]



Problem (Markowitz portfolio optimization)

$$F^\star := \min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathbb{E}\left[|\rho - \langle \mathbf{x}, \theta_t \rangle|^2 \right] \right\}$$

- $\rho \in \mathbb{R}$ *is the desired return.*
- $\mathcal{X}$ *is intersection of the standard simplex and the constraint* $\langle \mathbf{x}, \mathbb{E}[\theta_t] \rangle \geq \rho$.

*Datasets are available from `http://www.cs.technion.ac.il/~rani/portfolios`

Revisiting common loss functions: Least squares

Recall that the LS estimator is given by

$$\hat{\mathbf{x}}_{\mathsf{LS}} \in \arg\min_{\mathbf{x}} \left\{ \frac{1}{2} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\} = \arg\min_{\mathbf{x}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (b_i - \langle \mathbf{a}_i, \mathbf{x} \rangle)^2 : \mathbf{x} \in \mathbb{R}^p \right\},$$

where we define $\mathbf{b} := (b_1, \ldots, b_n)$ and $\mathbf{a}_i$ to be the i-th row of $\mathbf{A}$.

A statistical learning view of least squares

This corresponds to a statistical learning model, for which

- the sample is given by $(\mathbf{a}_i, b_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$,
- the function class $\mathcal{F}$ is given by $\mathcal{F} := \{f_{\mathbf{x}}(\cdot) := \langle \cdot, \mathbf{x} \rangle : \mathbf{x} \in \mathbb{R}^p\}$, and
- the loss function is given by $L(f_{\mathbf{x}}(\mathbf{a}), b) := (b - f_{\mathbf{x}}(\mathbf{a}))^2$.

The corresponding ERM solution is

$$\hat{f}_n(\cdot) := \langle \cdot, \hat{\mathbf{x}}_{\mathsf{LS}} \rangle .$$

• Given $\mathbf{a}$, LS estimator seeks to minimize the error of predicting the corresponding b by a linear function, in terms of the squared error.

Revisiting common loss functions: Hinge loss

Recall the unconstrained SVM formulation

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \left\{ \frac{1}{n} \sum_{j=1}^{n} \max\left\{1 - b_j \langle \mathbf{a}_j, \mathbf{x} \rangle, 0\right\} + \lambda \|\mathbf{x}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$

where $\mathbf{b} := (b_1, \ldots, b_n) \in \{-1, 1\}^n$.

A statistical learning view of SVM

This corresponds to a statistical learning model, for which

- the sample is given by $(\mathbf{a}_i, b_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$,
- the function class $\mathcal{F}$ is given by $\mathcal{F} := \{f_{\mathbf{x}}(\cdot) := \langle \cdot, \mathbf{x} \rangle : \mathbf{x} \in \mathbb{R}^p\}$, and
- the loss function is given by $L(f_{\mathbf{x}}(\mathbf{a}), b) := \max\{0, 1 - b f_{\mathbf{x}}(\mathbf{a})\}$.

The corresponding ERM solution is

$$\hat{f}_n(\cdot) := \langle \cdot, \hat{\mathbf{x}} \rangle .$$

• Given $\mathbf{a}$, SVM aims to minimize the error of predicting the corresponding b by a linear function, in terms of the hinge loss.

Revisiting common loss functions: Logistic loss (Log-loss)

Recall the logistic regression formulation

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x},\delta} \left\{ \frac{1}{n} \sum_{j=1}^{n} \log\left(1 + e^{-b_j(\langle \mathbf{x}, \mathbf{a}_j \rangle + \delta)}\right) : \mathbf{x} \in \mathbb{R}^p, \delta \in \mathbb{R} \right\}$$

where $\mathbf{b} := (b_1, \ldots, b_n) \in \{-1, 1\}^n$.

A statistical learning view of logistic regression

This corresponds to a statistical learning model, for which

- the sample is given by $(\mathbf{a}_i, b_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$,
- the function class $\mathcal{F}$ is given by $\mathcal{F} := \{f_{\mathbf{x}}(\cdot) := \langle \cdot, \mathbf{x} \rangle : \mathbf{x} \in \mathbb{R}^p\}$, and
- the loss function is given by $L(f_{\mathbf{x}}(\mathbf{a}), b) := \log\left(1 + e^{-b f_{\mathbf{x}}(\mathbf{a})}\right)$.

The corresponding ERM solution is

$$\hat{f}_n(\cdot) := \langle \cdot, \hat{\mathbf{x}} \rangle .$$

• Given $\mathbf{a}$, logistic regression aims to minimize the error of predicting the corresponding b by a linear function, in terms of the log-loss.

Constrained convex minimization

Problem (**Mathematical formulation**)

Consider the following constrained convex minimization problem:

$$f^\star = \min_{\mathbf{x}\in\mathcal{X}} \Big\{ f(\mathbf{x}) := \mathbb{E}[h(\mathbf{x},\theta)] \Big\}$$

- $\mathcal{X} \subset \mathbb{R}^p$ *is a non-empty bounded closed convex set.*
- θ *is a random vector whose probability distribution is supported on set* Θ.
- f *is continuous and convex on* $\mathcal{X}$.
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in dom(f) : f(\mathbf{x}^\star) = f^\star\}$ *is nonempty.*

Stochastic projected gradient method (SG)

Stochastic projected gradient method (SG)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = P_{\mathcal{X}}\left(\mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k)\right).$$

Remark

SG shares the same structure as the projected gradient descent method, but the gradient is replaced by an unbiased estimate in the 2nd step. The cost of computing this estimate is typically much cheaper than that of $\nabla f(\mathbf{x}^k)$.

Stochastic projected gradient method (SG)

Stochastic projected gradient method (SG)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = P_{\mathcal{X}}\left(\mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k)\right).$$

Remark

SG shares the same structure as the projected gradient descent method, but the gradient is replaced by an unbiased estimate in the 2nd step. The cost of computing this estimate is typically much cheaper than that of $\nabla f(\mathbf{x}^k)$.

Least squares

$$\hat{\mathbf{x}}_{\mathsf{LS}} \in \arg\min\left\{\frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x} \in \mathcal{X}\right\} = \arg\min\left\{\frac{1}{2n}\sum_{i=1}^{n}(b_i - \langle \mathbf{a}_i, \mathbf{x}\rangle)^2 : \mathbf{x} \in \mathcal{X}\right\}$$

$$\nabla f(\mathbf{x}^k) = \mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{b}), \qquad G(\mathbf{x}^k, (\mathbf{a}_i, b_i)) = \mathbf{a}_i(\langle \mathbf{a}_i, \mathbf{x}^k\rangle - b_i).$$

Different notions of convergence: Convergence in expectation

Theorem (Mean convergence of SG [3])

Suppose that:

1. f *is* μ*-strongly convex,*
2. $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2] \leq M^2$,
3. $\gamma_k = \gamma_0/(k+1)$ *with* $\gamma_0 > \frac{\mu}{2}$.

Then,

$$\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|^2] \leq \frac{1}{k} \max\left\{ \frac{\gamma_0^2 M^2}{2\gamma_0 \mu - 1}, \|\mathbf{x}^0 - \mathbf{x}^\star\|^2 \right\}.$$

If, in addition,

1. $\mathbf{x}^\star \in \operatorname{int}(\mathcal{X})$,
2. ∇f *is* L*-Lipschitz continuous.*

Then,

$$\mathbb{E}[f(\mathbf{x}^k) - f(\mathbf{x}^\star)] \leq \frac{L}{2k} \max\left\{ \frac{\gamma_0^2 M^2}{2\gamma_0 \mu - 1}, \|\mathbf{x}^0 - \mathbf{x}^\star\|^2 \right\}.$$

• $\mathcal{O}(1/k)$ rate in the objective residual is optimal for stochastic gradient methods under strong convexity assumption.

Different notions of convergence: Almost sure convergence

Theorem (Almost sure convergence of SG [3])

Denote $\mathcal{F}_k = \sigma(\mathbf{x}^0, \theta_0, \ldots, \theta_{k-1})$. *Suppose that:*

1. ∇f *is* L*-Lipschitz continuous,*
2. $\sum_{k=0}^{\infty} \gamma_k = \infty$ *and* $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$.
3. $\sum_{k=0}^{\infty} \gamma_k^2 \, \mathbb{E}[\| G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k) \|^2 | \mathcal{F}_k] < +\infty$ *almost surely.*

Then,

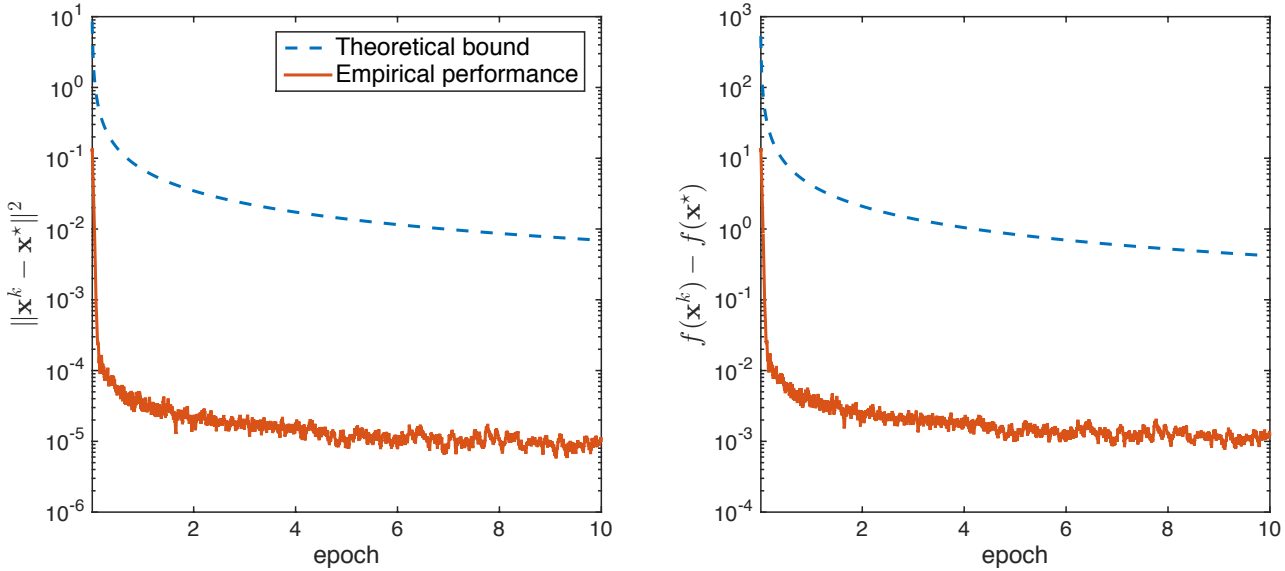
$$\mathbf{x}^k \to \mathbf{x}^\star \text{ almost surely.}$$

Remarks

- (2) is satisfied: $\gamma_k = \gamma_0/(k+1)$.
- (3) is satisfied: $\sup_{k \in \mathbb{N}} \mathbb{E}[\| G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k) \|^2 | \mathcal{F}_k] < +\infty$.

Example: Constrained least squares problem

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_1 \leq 1 \right\}$$

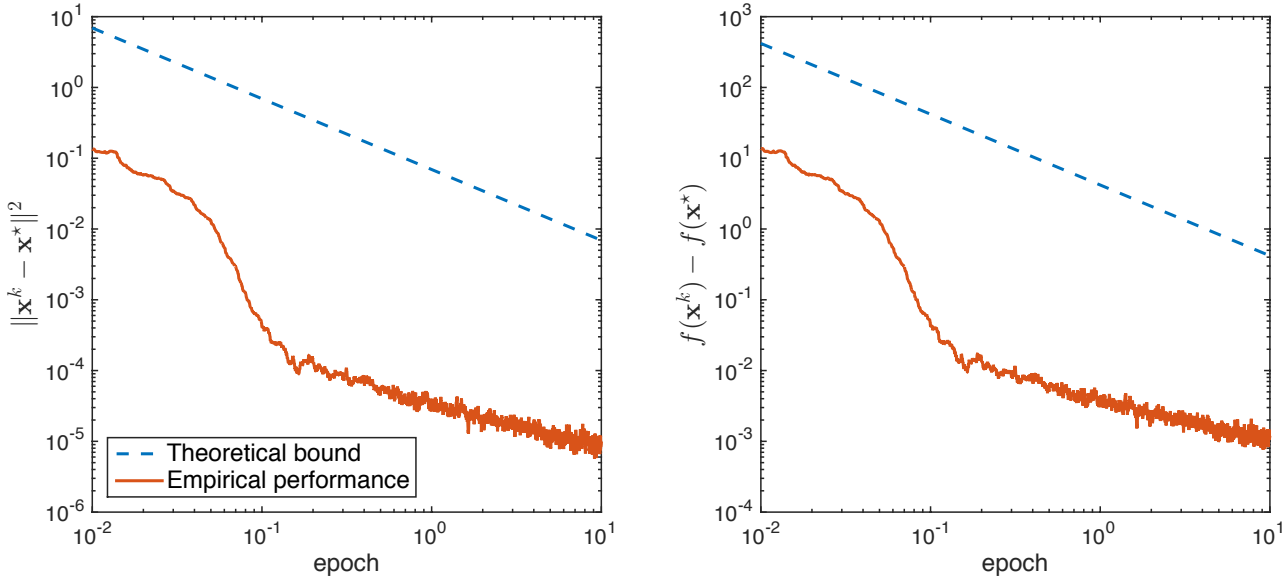


Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 10 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_1 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise. SNR is 30dB.
- $\gamma_0 = \mu/2$.

Example: Constrained least squares problem

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_1 \leq 1 \right\}$$

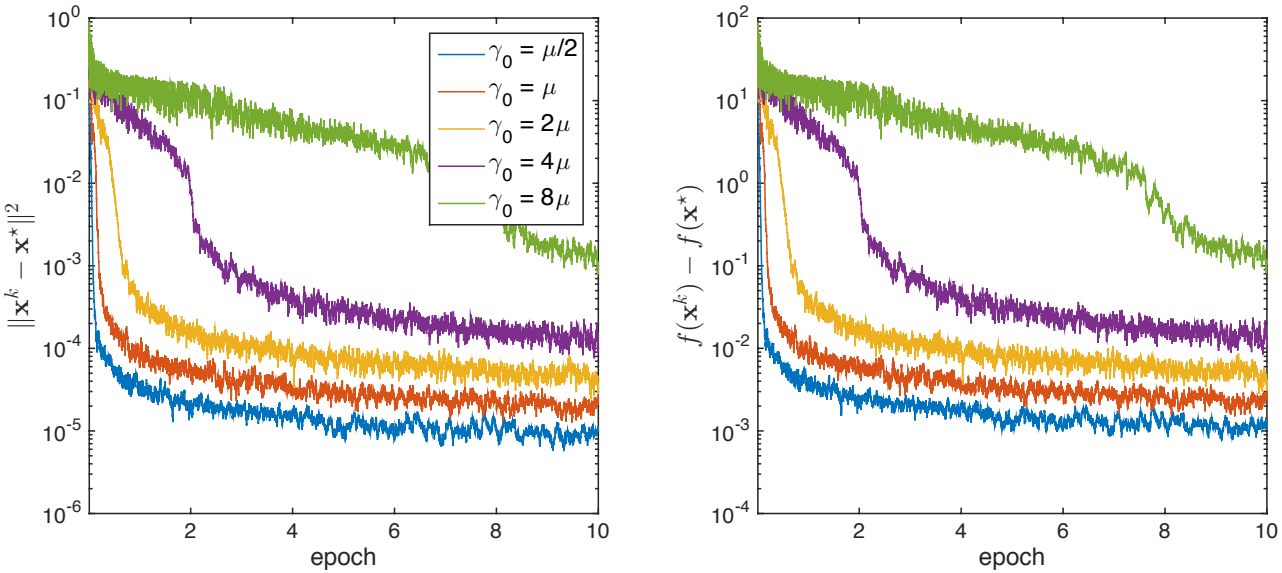


Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 10 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_1 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise. SNR is 30dB.
- $\gamma_0 = \mu/2$.

Example: SG method with different step sizes

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_1 \leq 1 \right\}$$



Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 10 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_1 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise. SNR is 30dB.

Stochastic projected gradient with averaging

Stochastic gradient method with averaging (ASG)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.

2a. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = P_{\mathcal{X}}(\mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k)).$$

2b. $\bar{\mathbf{x}}^k = (\sum_{j=0}^k \gamma_j)^{-1} \sum_{j=0}^k \gamma_j \mathbf{x}^j$.

Theorem (Mean convergence of ASG [3])

Denote $D_{\mathcal{X}} = \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}^0 - \mathbf{x}\|^2$ *and assume that* $\mathbb{E}[\| G(\mathbf{x}^k, \theta_k)\|^2] \leq M^2$ *for some* $M \in]0, +\infty[$. *Then,*

$$\mathbb{E}[f(\bar{\mathbf{x}}^k) - f(\mathbf{x}^\star)] \leq \frac{D_{\mathcal{X}}^2 + M^2 \sum_{j=0}^k \gamma_j^2}{2\sum_{j=0}^k \gamma_j},$$

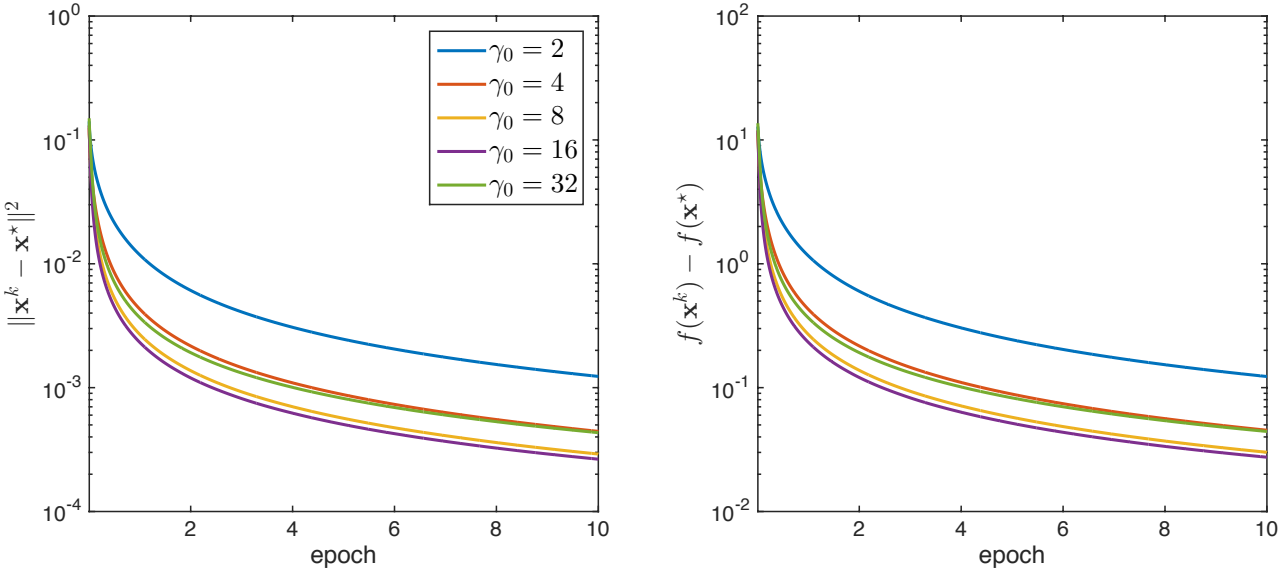
In addition, choosing $\gamma_k = D_{\mathcal{X}}/(M\sqrt{k})$, *we get,*

$$\mathbb{E}[f(\bar{\mathbf{x}}^k) - f(\mathbf{x}^\star)] \leq \frac{M D_{\mathcal{X}}}{\sqrt{k}}.$$

- $\mathcal{O}(1/\sqrt{k})$ rate is optimal for stochastic methods unless we assume strong convexity.

Example: ASG method with different step sizes

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_1 \leq 1 \right\}$$

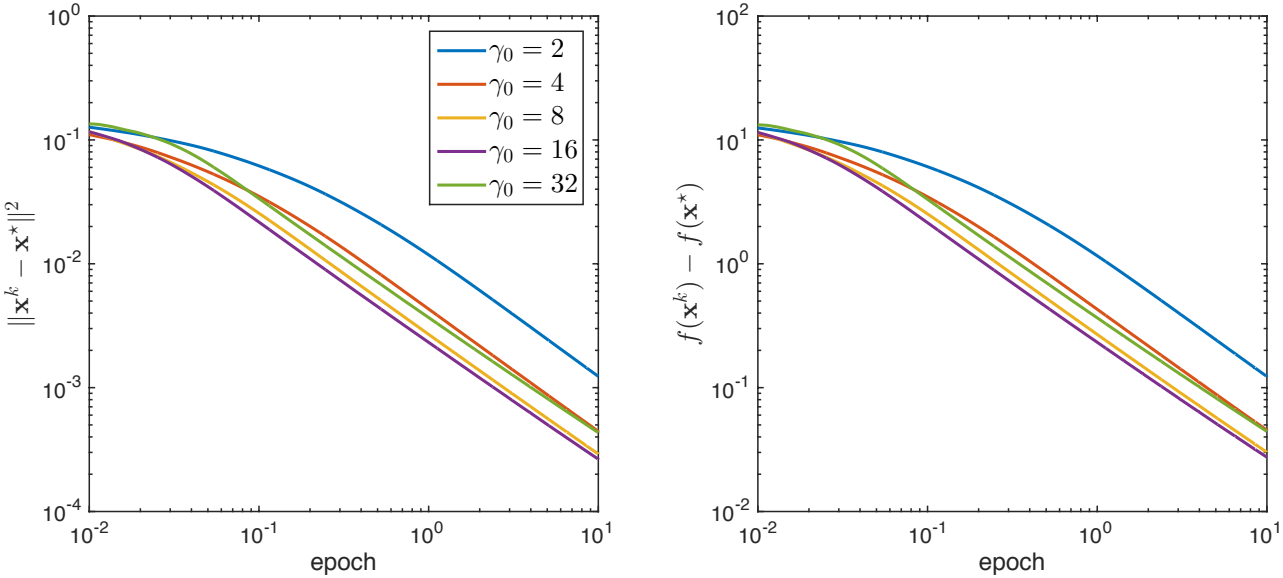


Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 10 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_1 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise. SNR is 30dB.

Example: ASG method with different step sizes

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_1 \leq 1 \right\}$$



Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 10 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_1 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise. SNR is 30dB.

Stochastic approximation for unconstrained convex optimization

For the special case $\mathcal{X} = \mathbb{R}^p$ [4]

Assumptions (with probability 1):

H1. $\|G(\mathbf{x}, \theta_k) - G(\mathbf{y}, \theta_k)\|_2 \leq L_k \|\mathbf{x} - \mathbf{y}\|_2$.

H2. $\mathbb{E}[\|G(\mathbf{x}^\star, \theta_k)\|^2 | \mathcal{F}_k] \leq M^2$ for some $M \in]0, \infty[$.

H3. $\sup_{D>0} \sup_{\{\|\mathbf{x}\| \leq D\}} \|G(\mathbf{x}, \theta_k)\| < +\infty$.

Convergence's rates without averaging:

- $\gamma_k = \gamma_0 k^{-2/3}$, + H1.+ H2. $\implies \mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|_2^2] = \mathcal{O}(k^{-1/3})$.
- $\gamma_k = \gamma_0 k^{-\alpha}$, $\alpha \in [1/3, 1/2]$, + H1.+ H3. $\implies \mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|_2^2] = \mathcal{O}(k^{-3\alpha/2+1/2})$.

Convergence's rates with averaging:

- $\gamma_k = \gamma_0 k^{-\alpha}$, $\alpha \in [1/2, 1]$, + H1.+ H2. $\implies \mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|_2^2] = \mathcal{O}(k^{-\alpha})$.
- $\gamma_k = \gamma_0 k^{-\alpha}$, $\alpha \in [0, 1]$, + H3. $\implies \mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|_2^2] = \mathcal{O}(k^{\alpha-1})$.

- These bounds are non-asymptotic, see [4] for the exact expressions of these bounds.

Important remark!

All the results we have shown so far can be generalized for the non-smooth objectives, simply by replacing the gradient with a subgradient.

Recall: Subdifferentials and (sub)gradients in convex functions

- Subdifferential: generalizes ∇ to *nondifferentiable functions*

Definition

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ is defined by the set:

$$\partial f(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^p \ : \ f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \ \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathcal{Q}\}\,.$$

Each element $\mathbf{v}$ of $\partial f(\mathbf{x})$ is called *subgradient* of f at $\mathbf{x}$.

Definition

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a differentiable convex function. Then, the subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ contains only the gradient, i.e., $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

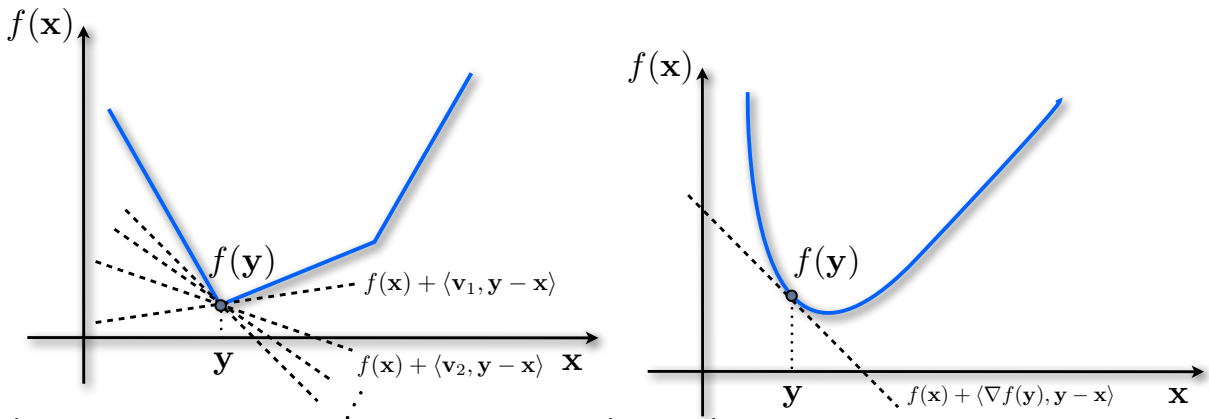


Figure: (**Left**) Non-differentiability at point $\mathbf{y}$. (**Right**) Gradient as a subdifferential with a singleton entry.

Composite convex minimization

Problem (Unconstrained composite convex minimization)

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

- f *and* g *are both proper, closed, and convex.*
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in dom(F) : F(\mathbf{x}^\star) = F^\star\}$ *is nonempty.*

Composite convex minimization

Problem (Unconstrained composite convex minimization)

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

- *f and g are both proper, closed, and convex.*
- *The solution set $\mathcal{S}^\star := \{\mathbf{x}^\star \in dom(F) : F(\mathbf{x}^\star) = F^\star\}$ is nonempty.*

Two remarks

- Nonsmoothness: At least one of the two functions f and g is **nonsmooth**
  - General nonsmooth convex optimization methods are inefficient.
- Generality: it covers a wider range of problems than smooth unconstrained problems.
  - f is a loss function, a data fidelity, or negative log-likelihood function.
  - g is a regularizer, encouraging structure and/or constraints in the solution.

A short detour: Proximal-point operators

Definition (Proximal operator [5])

Let $g : \mathbb{R}^p \to \mathbb{R}$ be a closed proper convex function. Then, the proximal operator (or prox-operator) of g is defined as:

$$\boxed{\operatorname{prox}_g(\mathbf{x}) \equiv \arg\min_{\mathbf{y}\in\mathbb{R}^p} \left\{ g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \right\}.}$$

A short detour: Proximal-point operators

Definition (Proximal operator [5])

Let $g : \mathbb{R}^p \to \mathbb{R}$ be a closed proper convex function. Then, the proximal operator (or prox-operator) of g is defined as:

$$\boxed{\operatorname{prox}_g(\mathbf{x}) \equiv \arg\min_{\mathbf{y}\in\mathbb{R}^p} \left\{ g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \right\}.}$$

Numerical efficiency

For problem (2):

- Many well-known convex functions g, we can compute $\operatorname{prox}_g(\mathbf{x})$ analytically or very efficiently.
- If ∇f is Lipschitz continuous and $\operatorname{prox}_g(\mathbf{x})$ is **cheap** to compute, then solving (2) is as **efficient** as solving $\boxed{\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})}$ in terms of complexity.

A non-exhaustive list of proximal tractability functions

Name	Function	Proximal operator	Complexity
ℓ_1-norm	$f(\mathbf{x}) := \|\mathbf{x}\|_1$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \mathrm{sign}(\mathbf{x}) \otimes [\|\mathbf{x}\| - \lambda]_+$	$\mathcal{O}(p)$
ℓ_2-norm	$f(\mathbf{x}) := \|\mathbf{x}\|_2$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = [1 - \lambda/\|\mathbf{x}\|_2]_+\mathbf{x}$	$\mathcal{O}(p)$
Support function	$f(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{C}} \mathbf{x}^T\mathbf{y}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \mathbf{x} - \lambda\pi_{\mathcal{C}}(\mathbf{x})$	
Box indicator	$f(\mathbf{x}) := \delta_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \pi_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$	$\mathcal{O}(p)$
Positive semidefinite cone indicator	$f(\mathbf{X}) := \delta_{\mathbb{S}^p_+}(\mathbf{X})$	$\mathrm{prox}_{\lambda f}(\mathbf{X}) = \mathbf{U}[\Sigma]_+\mathbf{U}^T$, where $\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^T$	$\mathcal{O}(p^3)$
Hyperplane indicator	$f(\mathbf{x}) := \delta_{\mathcal{X}}(\mathbf{x})$, $\mathcal{X} := \{\mathbf{x} : \mathbf{a}^T\mathbf{x} = b\}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \pi_{\mathcal{X}}(\mathbf{x}) = \mathbf{x} + \left(\frac{b-\mathbf{a}^T\mathbf{x}}{\|\mathbf{a}\|_2}\right)\mathbf{a}$	$\mathcal{O}(p)$
Simplex indicator	$f(\mathbf{x}) = \delta_{\mathcal{X}}(\mathbf{x}), \mathcal{X} := \{\mathbf{x} : \mathbf{x} \geq 0,\ \mathbf{1}^T\mathbf{x} = 1\}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (\mathbf{x} - \nu\mathbf{1})$ for some $\nu \in \mathbb{R}$, which can be efficiently calculated	$\tilde{\mathcal{O}}(p)$
Convex quadratic	$f(\mathbf{x}) := (1/2)\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{q}^T\mathbf{x}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (\lambda\mathbb{I} + \mathbf{Q})^{-1}\mathbf{x}$	$\mathcal{O}(p\log p) \rightarrow \mathcal{O}(p^3)$
Square ℓ_2-norm	$f(\mathbf{x}) := (1/2)\|\mathbf{x}\|_2^2$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (1/(1+\lambda))\mathbf{x}$	$\mathcal{O}(p)$
log-function	$f(\mathbf{x}) := -\log(x)$	$\mathrm{prox}_{\lambda f}(x) = ((x^2 + 4\lambda)^{1/2} + x)/2$	$\mathcal{O}(1)$
log det-function	$f(\mathbf{x}) := -\log\det(\mathbf{X})$	$\mathrm{prox}_{\lambda f}(\mathbf{X})$ is the log-function prox applied to the individual eigenvalues of $\mathbf{X}$	$\mathcal{O}(p^3)$

Here: $[\mathbf{x}]_+ := \max\{0, \mathbf{x}\}$ and $\delta_{\mathcal{X}}$ is the indicator function of the convex set $\mathcal{X}$, sign is the sign function, $\mathbb{S}^p_+$ is the cone of symmetric positive semidefinite matrices.

For more functions, see [6, 7].

Unconstrained composite convex minimization

Problem (**Mathematical formulation**)

Consider the following unconstrained composite convex minimization problem:

$$\boxed{F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x}) := \mathbb{E}[h(\mathbf{x},\theta)] + g(\mathbf{x})\}} \tag{2}$$

- $f := \mathbb{E}[h(\mathbf{x},\theta)]$ *and* g *are both proper, closed, and convex.*
- ∇f *is* L*-Lipschitz continuous.*
- θ *is a random vector whose distribution is supported on* Θ
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in dom(F) : F(\mathbf{x}^\star) = F^\star\}$ *is nonempty.*

We assume that

- It is possible to generate a i.i.d sample $(\theta_k)_{k\in\mathbb{N}}$ of realizations of θ.
- Given $(\mathbf{x},\theta) \in \mathbb{R}^p \times \Omega$, one can find a vector $G(\mathbf{x},\theta)$ such that $\mathbb{E}[G(\mathbf{x},\theta)] = \nabla f(\mathbf{x})$.
- prox_g is tractable.

Stochastic proximal gradient method

Stochastic proximal gradient method (SPGM)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma_k g}(\mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k)).$$

Remark

SPGM shares the same structure as the classical proximal gradient method, but the gradient is replaced by an unbiased estimate in the 2nd step.

Two special cases:

- $g = \mathbf{1}_{\mathcal{X}}$, i.e., the indicator function of $\mathcal{X}$: SPGM reduces to SG.
- $G(\mathbf{x}, \theta) = \nabla f(\mathbf{x})$: SPGM reduces to the classical proximal gradient method.

Stochastic proximal gradient method

Stochastic proximal gradient method (SPGM)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma_k g}(\mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k)).$$

Theorem (Mean convergence of the iterates of SPGM [8])

Suppose that:

1. f *and* g *are (strongly) convex with* $\mu_f \geq 0$ *and* $\mu_g \geq 0$ *such that* $\mu := \mu_f + \mu_g > 0$.
2. $\sup_{k\in\mathbb{N}} \mathbb{E}[\|G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k)\|^2 | \mathcal{F}_k] \leq M^2$.
3. $\gamma_k = \gamma_0/k^\alpha$ *with* $0 < \alpha \leq 1$.

Then, for k *large enough,*

$$\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|_2^2] = \begin{cases} \mathcal{O}(k^\alpha) & \text{if} \quad 0 < \alpha < 1, \\ \mathcal{O}(1/k^\beta) + \mathcal{O}(1/k) & \text{if} \quad \alpha = 1, \end{cases}$$

where $\beta = 2\gamma_0(\mu_g + \mu_f\varepsilon)/(1 + \mu_g)^2$ *for some fixed* $0 < \varepsilon < 1$.

Remark: If γ_0 is large enough then $\beta > 1$.

Stochastic proximal gradient method

Stochastic proximal gradient method (SPGM)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \mathrm{prox}_{\gamma_k g}(\mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k)).$$

Theorem (Almost sure convergence of the iterates of SPGM [8])

Suppose that:

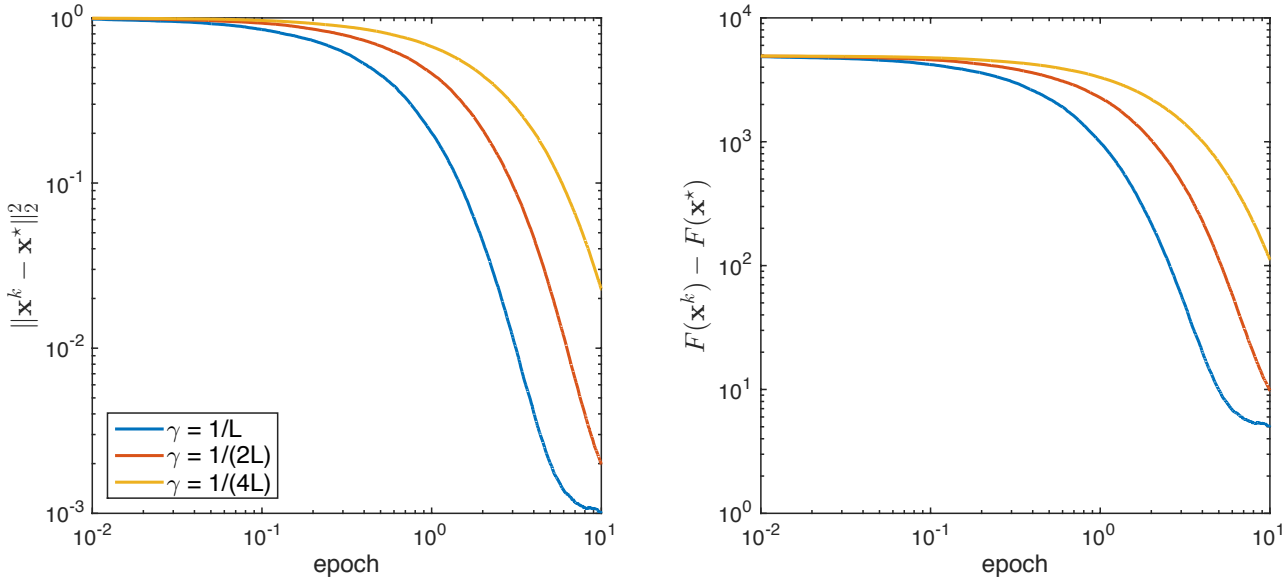
1. $0 < \gamma_k \leq 1/L$ *and* $\sum_{k=0}^{\infty} \gamma_k = \infty$.
2. $\limsup_k \|\mathbf{x}^k\| < +\infty$ *almost surely.*
3. $\sum_{k\geq 0} \gamma_k^2 \, \mathbb{E}[\| G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k)\|^2 | \mathcal{F}_k] < +\infty$.

Then,

$$\mathbf{x}^k \to \mathbf{x}^\star \text{ almost surely.}$$

Example: ASPGM I with different number of iterations

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \rho\|\mathbf{x}\|_1 : \mathbf{x} \in \mathbb{R}^p \right\}$$



LASSO: Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 10 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_2 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise. SNR is 30dB.
- $\rho = 10^{-2}$.

Stochastic proximal gradient method with averaging

SPGM with averaging

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}}$, $(\chi_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.

2a. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma_k h}(\mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k)).$$

2b. $\bar{\mathbf{x}}^k = (\sum_{j=0}^k \chi_j)^{-1} \sum_{j=0}^k \chi_j \mathbf{x}^j$.

Theorem (Mean convergence of SPGM with averaging [11])

Denote $A_k = \sum_{j=0}^k \chi_j$. *Suppose that:*

1. $(\chi_k/\gamma_k)_{k\in\mathbb{N}}$ *is non decreasing.*
2. $(\exists c > 0)\ \sup_{k\in\mathbb{N}} \|\mathbf{x}^k\| \leq c$ *almost surely.*

Then,

$$\mathbb{E}[F(\bar{\mathbf{x}}^k) - F(\mathbf{x}^\star)] \leq \frac{(c + \|\mathbf{x}^\star\|)^2 \chi_k}{2\gamma_k A_k} + \frac{1}{A_k} \sum_{j=0}^k \chi_j \gamma_j \mathbb{E}[\| G(\mathbf{x}^j, \theta_j) - \nabla f(\mathbf{x}^j)\|^2].$$

Accelerated stochastic proximal gradient method I

Accelerated stochastic proximal gradient method I (ASPGM I)

1. Choose $\mathbf{x}_0 = \mathbf{z}_0 = \mathbf{0}$. Define $\alpha_k := 2/(k+2)$ and $\gamma_k := \alpha_k(2 + N^{3/2}/L)$
2. For $k = 0, 1, \dots N$ perform:
2a. $\mathbf{y}^k = (1-\alpha_k)\mathbf{x}^k + \alpha_k \mathbf{z}^k$
2b. $\mathbf{z}^{k+1} = \operatorname{prox}_{\frac{1}{\gamma_k L} g}(\mathbf{z}^k - \frac{1}{\gamma_k L} G(\mathbf{y}^k, \theta_k))$
2c. $\mathbf{x}^{k+1} = (1-\alpha_k)\mathbf{x}^k + \alpha_k \mathbf{z}^{k+1}$.

Theorem (Mean convergence of ASPGM I [9])

Suppose that $\mathbb{E}[\| G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k)\|] \leq M^2$ *for some* $M \in]0, \infty[$. *Then,*

$$\mathbb{E}[F(\mathbf{x}^{N+1}) - F(\mathbf{x}^\star)] \leq \frac{2\|\mathbf{x}^\star - \mathbf{z}_0\|^2 + M^2}{\sqrt{N+2}} + L\frac{4\|\mathbf{x}^\star - \mathbf{z}_0\|^2 + 2M^2}{N+2}.$$

Accelerated stochastic proximal gradient method I

Accelerated stochastic proximal gradient method I (ASPGM I)

1. Choose $\mathbf{x}_0 = \mathbf{z}_0 = \mathbf{0}$. Define $\alpha_k := 2/(k+2)$ and $\gamma_k := \alpha_k(2 + N^{3/2}/L)$
2. For $k = 0, 1, \dots N$ perform:
2a. $\mathbf{y}^k = (1-\alpha_k)\mathbf{x}^k + \alpha_k \mathbf{z}^k$
2b. $\mathbf{z}^{k+1} = \operatorname{prox}_{\frac{1}{\gamma_k L} g}(\mathbf{z}^k - \frac{1}{\gamma_k L} G(\mathbf{y}^k, \theta_k))$
2c. $\mathbf{x}^{k+1} = (1-\alpha_k)\mathbf{x}^k + \alpha_k \mathbf{z}^{k+1}$.

Theorem (Mean convergence of ASPGM I [9])

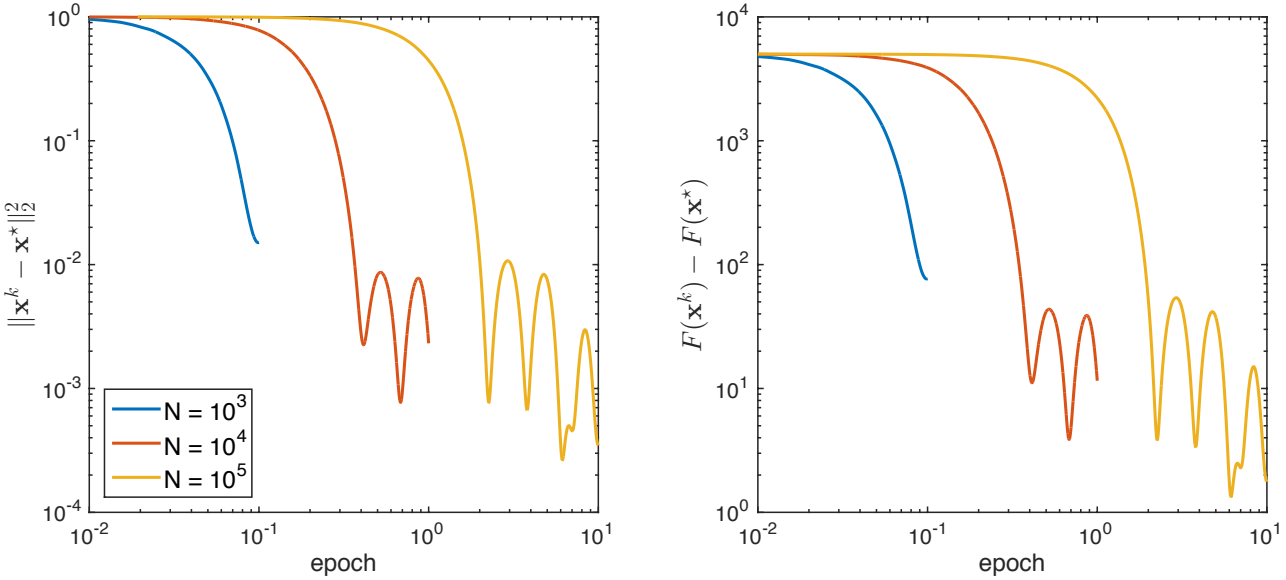
Suppose that $\mathbb{E}[\| G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k)\|] \leq M^2$ *for some* $M \in]0, \infty[$. *Then,*

$$\mathbb{E}[F(\mathbf{x}^{N+1}) - F(\mathbf{x}^\star)] \leq \frac{2\|\mathbf{x}^\star - \mathbf{z}_0\|^2 + M^2}{\sqrt{N+2}} + L\frac{4\|\mathbf{x}^\star - \mathbf{z}_0\|^2 + 2M^2}{N+2}.$$

Number of iterations N needs to be fixed in advance!

Example: ASPGM I with different number of iterations

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \rho\|\mathbf{x}\|_1 : \mathbf{x} \in \mathbb{R}^p \right\}$$



LASSO: Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 10 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_2 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise. SNR is 30dB.
- $\rho = 10^{-2}$.

Accelerated stochastic proximal gradient method II

Accelerated stochastic proximal gradient method II (ASPGM II)

1. Choose $\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{0}$, $(\gamma_k)_{k\in\mathbb{N}}$, $(\alpha_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$, $\alpha_0 = 1, \gamma_0 = L + \mu$.
2. For $k = 0, 1, \ldots$ perform:
2a. $\mathbf{x}^{k+1} = (1 - \alpha_k)\mathbf{y}^k + \alpha_k \mathbf{z}^k$.
2b. $\mathbf{y}^{k+1} = \operatorname{prox}_{\frac{1}{\gamma_k} g}(\mathbf{x}^{k+1} - \frac{1}{\gamma_k} G(\mathbf{x}^{k+1}, \theta_k))$.
2c. $\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{1}{\gamma_k \alpha_k + \mu} \left(\gamma_k(\mathbf{x}^{k+1} - \mathbf{y}^{k+1}) + \mu(\mathbf{z}^k - \mathbf{x}^{k+1})\right)$.

Theorem (Mean convergence of ASPGM II [10])

Suppose that:

1. $\mathbb{E}[\|\mathbf{z}^k - \mathbf{x}^\star\|^2] \leq D^2$ *for some* $D \in\ ]0, \infty[$.
2. $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k)\|^2] \leq M^2$ *for some* $M \in]0, \infty[$.
3. $\gamma_k = c(k+1)^{3/2} + L$ *for a fixed* $c > 0$, *and* $\alpha_k = 2/(k+2)$.

Then,

$$\mathbb{E}[F(\mathbf{y}^{k+1}) - F(\mathbf{x}^\star)] \leq \frac{3D^2 L}{k^2} + \left(3D^2 c + \frac{5M^2}{3c}\right)\frac{1}{\sqrt{k}}.$$

Accelerated stochastic proximal gradient method II

Accelerated stochastic proximal gradient method II (ASPGM II)

1. Choose $\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{0}$, $(\gamma_k)_{k\in\mathbb{N}}$, $(\alpha_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$, $\alpha_0 = 1, \gamma_0 = L + \mu$.

2. For $k = 0, 1, \ldots$ perform:

2a. $\mathbf{x}^{k+1} = (1 - \alpha_k)\mathbf{y}^k + \alpha_k \mathbf{z}^k$.

2b. $\mathbf{y}^{k+1} = \operatorname{prox}_{\frac{1}{\gamma_k} g}(\mathbf{x}^{k+1} - \frac{1}{\gamma_k} G(\mathbf{x}^{k+1}, \theta_k))$.

2c. $\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{1}{\gamma_k \alpha_k + \mu} \left(\gamma_k(\mathbf{x}^{k+1} - \mathbf{y}^{k+1}) + \mu(\mathbf{z}^k - \mathbf{x}^{k+1})\right)$.

Theorem (Mean convergence of ASPGM II with strong convexity [10])

Define $\lambda_k = \prod_{j=1}^{k}(1 - \alpha_j)$ *and* $\lambda_0 = 1$. *Suppose that:*

1. f *is* μ*-strongly convex.*
2. $\mathbb{E}[\|\mathbf{z}^k - \mathbf{x}^\star\|^2] \leq D^2$ *for some* $D \in\ ]0, \infty[$.
3. $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k)\|^2] \leq M^2$ *for some* $M \in]0, \infty[$.
4. $\gamma_k = L + \frac{\mu}{\lambda_{k-1}}$ *and* $\alpha_k = \sqrt{\lambda_{k-1} + \frac{\lambda_{k-1}^2}{4}} - \frac{\lambda_{k-1}}{2}$.

Then,

$$\mathbb{E}[F(\mathbf{y}^{k+1}) - F(\mathbf{x}^\star)] \leq \frac{2(L+\mu)D^2}{k^2} + \frac{6M^2}{\mu k}.$$

Stochastic FISTA

Stochastic FISTA

1. Choose $\mathbf{y}^0 \in \mathrm{dom} g$, $(\gamma_k)_{k\in\mathbb{N}}$, $(\alpha_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$, and

$$\lambda_0 = 1, \quad (\forall k \geq 1) \quad \gamma_{k+1}\lambda_k(\lambda_k - 1) \leq \gamma_k \lambda_{k-1}^2.$$

2. $\mathbf{y}^0 \in \mathrm{dom} g$ and $\mathbf{y}^1 = \mathrm{prox}_{\gamma_1 g}(\mathbf{y}^0 - \gamma_1 G(\mathbf{y}^0, \theta_0))$.

3. For $k = 1, 2 \ldots$ perform:

3a. $\mathbf{x}^k = \mathbf{y}^k + \frac{\lambda_{k-1}-1}{\lambda_k}(\mathbf{y}^k - \mathbf{y}^{k-1})$.

3b. $\mathbf{y}^{k+1} = \mathrm{prox}_{\gamma_{k+1} g}(\mathbf{x}^k - \gamma_{k+1} G(\mathbf{x}^k, \theta_k))$.

Remark. Some rules to update $(\lambda_k)_{k\in\mathbb{N}}$ when $(\gamma_k)_{k\in\mathbb{N}}$ is non-increasing:

- $\lambda_k \propto (k + k_0)^\alpha$ with $0 < \alpha < 1$, for some $k_0 > 0$.
- $\lambda_k = 1 + \frac{k}{2}$.
- $\lambda_{k+1} = \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\lambda_k^2}$.

Stochastic FISTA: Almost sure convergence

Theorem (Almost sure convergence of stochastic FISTA [11])

Suppose that:

1. $0 < \gamma_k < 1/L$ *and* $\lim \gamma_k \lambda_k^2 = +\infty$.
2. $\limsup_{k\to\infty} \|\mathbf{x}^k\| < +\infty$.
3. *For any* $B > 0$, *there exist constants* $(K_k)_{k\in\mathbb{N}}$ *such that*

$$\sum_{k\in\mathbb{N}} \gamma_{k+1}\lambda_k(1+\gamma_{k+1}\lambda_k)K_k < +\infty,$$

$$\mathbb{E}[\|G(\mathbf{x}^k,\theta_k) - \nabla f(\mathbf{x}^k)\|^2 | \mathcal{F}_k]\mathbf{1}_{\left\{\cap_{j\leq k}\{\|\mathbf{x}^j\|\leq B\}\right\}} \leq K_k.$$

Then,

$$F(\mathbf{y}^k) \to F(\mathbf{x}^\star) \text{ almost surely.}$$

Stochastic FISTA - Convergence in expectation

Theorem (Mean convergence of Stochastic FISTA [11])

Suppose that:

1. $0 < \gamma_k \leq 1/L$.
2. $\sup_k \|\mathbf{x}^k\| \leq B$ *almost surely.*
3. $B_k = D + 2\sum_{j=1}^{k} \gamma_{j+1}^2 \lambda_j^2 \mathbb{E}[\|\Delta_j\|^2] < +\infty$ *where*

$$\Delta_k = G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k),$$

$$D = \gamma_1 \mathbb{E}[F(\mathbf{x}^0) - F(\mathbf{x}^\star)] + \frac{1}{2}(B + \|\mathbf{x}^\star\|)^2.$$

Then,

$$\mathbb{E}[F(\mathbf{y}^{k+1}) - F(\mathbf{x}^\star)] \leq \frac{B_k}{\gamma_{k+1}\lambda_k^2}.$$

References I

[1] V. N. Vapnik.
An overview of statistical learning theory.
IEEE Trans. Inf. Theory, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[2] V. Cevher, B. C. Vũ, and A. Yurtsever.
Stochastic forward Douglas-Rachford splitting for monotone inclusions.
infoscience.epfl.ch/record/215759.

[3] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro.
Robust stochastic approximation approach to stochastic programming.
SIAM J. Optim., vol. 19, pp. 1574–1609, 2008.

[4] F. Bach and E. Moulines.
Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning.
Advances in Neural Information Processing Systems (NIPS), 2011.

[5] R.T. Rockafellar.
Monotone operators and the proximal point algorithm.
SIAM Journal on Control and Optimization, 14:877–898, 1976.

References II

[6] P. Combettes and Pesquet J.-C.
Signal recovery by proximal forward-backward splitting.
In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer-Verlag, 2011.

[7] N. Parikh and S. Boyd.
Proximal algorithms.
Foundations and Trends in Optimization, 1(3):123–231, 2013.

[8] L. Rosasco, S. Villa, and B. C. Vũ.
Convergence of stochastic proximal gradient algorithm.
http://arxiv.org/abs/1403.5074.

[9] Q. Lin, X. Chen and J. Peña.
A smoothing stochastic gradient method for composite optimization.
Optimization Methods and Software, vol. 29, pp. 1281–1301, 2014.

[10] J. T. Kwok, C. Hu and W. Pan.
Accelerated gradient methods for stochastic optimization and online learning.
Advances in Neural Information Processing Systems, vol. 22, pp. 781–789, 2009.

References III

[11] Y. F. Atchade, G. Fort and E. Moulines.
On stochastic proximal gradient algorithms.
arXiv:1402.2365, 2014.

[12] Y. Nesterov.
Introductory Lectures on Convex Optimization: A Basic Course.
Kluwer, Boston, 2004.

[13] A. Nitanda.
Stochastic proximal gradient descent with acceleration techniques.
Advances in Neural Information Processing Systems, pp. 1574–1582, 2014.