Stochastic methods for nonsmooth nonconvex optimization

Dmitriy Drusvyatskiy
Mathematics, University of Washington

Joint work with V. Charisopoulos (Cornell), Y. Chen (Cornell), D. Davis (Cornell), D. Diaz (Cornell), K. MacPhee (Microsoft), C. Paquette (McGill)

EE and CSML seminar, Princeton University

Complexity of nonsmooth nonconvex stochastic optimization?

$$\min_x \ \mathbb{E}_{z\sim P}[f(x,z)]$$

Typical assumptions: **convexity** or **smoothness**

- different algorithms, analysis, guarantees

Nonsmooth and nonconvex losses arise often...

- structure (sparsity), robustness (outliers), stability (better conditioning)

Complexity of nonsmooth nonconvex stochastic optimization?

$$\min_x \; \mathbb{E}_{z\sim P}[f(x,z)]$$

Typical assumptions: **convexity** or **smoothness**

- different algorithms, analysis, guarantees

Nonsmooth and nonconvex losses arise often...

- structure (sparsity), robustness (outliers), stability (better conditioning)

Common problem class: $\text{(convex)} \circ \text{(smooth)}$

(Fletcher '80, Powell '83, Burke '85, Wright '90, Lewis-Wright '08, Cartis-Gould-Toint '11,. . .)

Outline

- Contemporary examples (low rank matrix recovery)
- **deterministic** rapid local search
- **stochastic** streaming and off-line algorithms

Example: Low-rank Matrix Recovery

Problem: Find rank r matrix $M_\sharp \succeq 0$ satisfying

$$\langle A_i, M_\sharp \rangle = b_i \qquad \forall i = 1, \ldots, m.$$

Example: Low-rank Matrix Recovery

Problem: Find rank r matrix $M_\sharp \succeq 0$ satisfying

$$\langle A_i, M_\sharp \rangle = b_i \qquad \forall i = 1, \ldots, m.$$

Measurement map:

$$\mathcal{A}(M) = (\langle A_1, M \rangle, \langle A_2, M \rangle, \ldots, \langle A_m, M \rangle)$$

Example: Low-rank Matrix Recovery

Problem: Find rank r matrix $M_\sharp \succeq 0$ satisfying

$$\langle A_i, M_\sharp \rangle = b_i \qquad \forall i = 1, \ldots, m.$$

Measurement map:

$$\mathcal{A}(M) = (\langle A_1, M \rangle, \langle A_2, M \rangle, \ldots, \langle A_m, M \rangle)$$

Restricted Isometry Property (RIP): Exist a norm $|||\cdot|||$ and constants $\kappa_1, \kappa_2 > 0$ satisfying

$$\kappa_1 \|M\|_F \leq |||\mathcal{A}(M)||| \leq \kappa_2 \|M\|_F$$

for all $M \in \mathbb{R}^{d \times d}$ of rank $2r$.

Example: Low-rank Matrix Recovery

Problem: Find rank r matrix $M_\sharp \succeq 0$ satisfying

$$\langle A_i, M_\sharp \rangle = b_i \qquad \forall i = 1, \ldots, m.$$

Measurement map:

$$\mathcal{A}(M) = (\langle A_1, M\rangle, \langle A_2, M\rangle, \ldots, \langle A_m, M\rangle)$$

Restricted Isometry Property (RIP): Exist a norm $|||\cdot|||$ and constants $\kappa_1, \kappa_2 > 0$ satisfying

$$\kappa_1 \|M\|_F \leq |||\mathcal{A}(M)||| \leq \kappa_2 \|M\|_F$$

for all $M \in \mathbb{R}^{d\times d}$ of rank $2r$.

Natural Penalty Formulation:

$$\min_{X \in \mathbb{R}^{d\times r}} \; |||\mathcal{A}(XX^\top) - b|||$$

Example: Low-rank Matrix Recovery

Problem: Find rank r matrix $M_\sharp \succeq 0$ satisfying

$$\langle A_i, M_\sharp \rangle = b_i \qquad \forall i = 1, \ldots, m.$$

Measurement map:

$$\mathcal{A}(M) = (\langle A_1, M \rangle, \langle A_2, M \rangle, \ldots, \langle A_m, M \rangle)$$

Restricted Isometry Property (RIP): Exist a norm $|||\cdot|||$ and constants $\kappa_1, \kappa_2 > 0$ satisfying

$$\kappa_1 \|M\|_F \leq |||\mathcal{A}(M)||| \leq \kappa_2 \|M\|_F$$

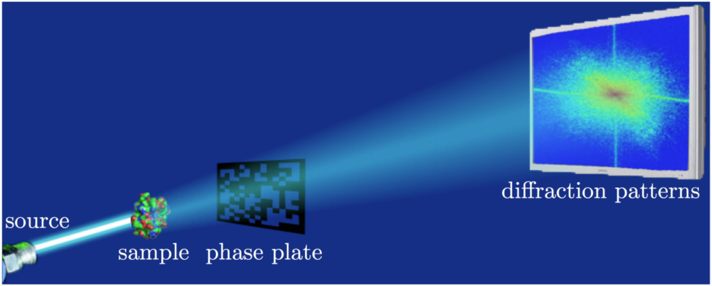
for all $M \in \mathbb{R}^{d \times d}$ of rank $2r$.

Natural Penalty Formulation:

$$\min_{X \in \mathbb{R}^{d \times r}} \; |||\mathcal{A}(XX^\top) - b|||$$

- Typical norms $|||\cdot||| = \frac{1}{\sqrt{m}} \| \cdot \|_2$ and $|||\cdot||| = \frac{1}{m} \| \cdot \|_1$
- ℓ_2-RIP valid for Gaussian A_i, leads to smooth problems
- ℓ_1-RIP valid for structured A_i, leads to nonsmooth problems

Example: phase retrieval[1]



[1] Candes, Li, Soltanolkotabi. Phase Retrieval from Coded Diffraction Patterns (2013)

Example: phase retrieval

Problem: Find $x_\sharp \in \mathbb{R}^d$ satisfying

$$(a_i^T x_\sharp)^2 = b_i$$

for $a_1, \ldots, a_m \in \mathbb{R}^d$ and $b_1, \ldots, b_m \in \mathbb{R}$.

[2] quadratic sensing (Chen-Chi-Goldsmith '15)

Example: phase retrieval

Problem: Find $x_\sharp \in \mathbb{R}^d$ satisfying

$$(a_i^T x_\sharp)^2 = b_i$$

for $a_1, \ldots, a_m \in \mathbb{R}^d$ and $b_1, \ldots, b_m \in \mathbb{R}$.

Measurement map: $$b_i = \langle a_i a_i^\top, x_\sharp x_\sharp^\top \rangle \qquad \Longrightarrow \qquad A_i := a_i a_i^\top$$

[2] quadratic sensing (Chen-Chi-Goldsmith '15)

Example: phase retrieval

Problem: Find $x_\sharp \in \mathbb{R}^d$ satisfying

$$(a_i^T x_\sharp)^2 = b_i$$

for $a_1, \ldots, a_m \in \mathbb{R}^d$ and $b_1, \ldots, b_m \in \mathbb{R}$.

Measurement map: $b_i = \langle a_i a_i^\top, x_\sharp x_\sharp^\top \rangle \implies A_i := a_i a_i^\top$

RIP:[2] Assume $m \geq 2d + 1$ and $a_i \sim N(0, I_d)$. Then w.p. $1 - e^{-cm}$ have

$$\kappa_1 \|M\|_F \leq \frac{1}{m} \|\mathcal{A}(M)\|_1 \leq \kappa_2 \|M\|_F \qquad \forall M \in \mathbb{R}_2^{d \times d}.$$

RIP fails with $|||\cdot||| = \frac{1}{\sqrt{m}} \|\cdot\|_2$

[2] quadratic sensing (Chen-Chi-Goldsmith '15)

Example: phase retrieval

Problem: Find $x_\sharp \in \mathbb{R}^d$ satisfying

$$(a_i^T x_\sharp)^2 = b_i$$

for $a_1, \ldots, a_m \in \mathbb{R}^d$ and $b_1, \ldots, b_m \in \mathbb{R}$.

Measurement map: $\qquad b_i = \langle a_i a_i^\top, x_\sharp x_\sharp^\top \rangle \qquad \Longrightarrow \qquad A_i := a_i a_i^\top$

RIP:[2] Assume $m \geq 2d + 1$ and $a_i \sim N(0, I_d)$. Then w.p. $1 - e^{-cm}$ have

$$\kappa_1 \|M\|_F \leq \frac{1}{m}\|\mathcal{A}(M)\|_1 \leq \kappa_2 \|M\|_F \qquad \forall M \in \mathbb{R}_2^{d\times d}.$$

RIP fails with $|||\cdot||| = \frac{1}{\sqrt{m}}\|\cdot\|_2$

Penalty Formulation: $\qquad \min_{x\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m |(a_i^\top x)^2 - b_i|.$

[2] quadratic sensing (Chen-Chi-Goldsmith '15)

Examples

- **Blind deconvolution/bi-convex sensing.** (Ling-Strohmer '15, Ahmed et al. '14)

$$\min_{x,y} \frac{1}{m}\sum_{i=1}^{m} |\langle u_i, x\rangle\langle v_i, y\rangle - b_i|$$

- **Robust PCA.** (Candès et al. '11, Chandrasekaran et al. '11, Netrapalli et al. '14)

$$\min_{L\in\mathbb{R}^{d\times r},\, V\in\mathbb{R}^{r\times m}} \|LV - M\|_1$$

- **Conditional Value-at-Risk.** (Rockafellar-Uryasev '10, Ben-Tal-Teboulle '86,'07)

$$\min_{x} \;\{\text{Expectation of } f(x,\cdot) \text{ on its } \alpha\text{-tail}\}\,.$$

  Equivalent formulation:

$$\min_{\gamma\in\mathbb{R},\, x\in\mathbb{R}^d} \;\gamma + \frac{1}{1-\alpha}\mathbb{E}_z[(f(x,z)-\gamma)_+]$$

- **covariance estimation**, **dictionary learning**, **group synchronization**, ...

Rapid local convergence

The two-part strategy

Typical approach.

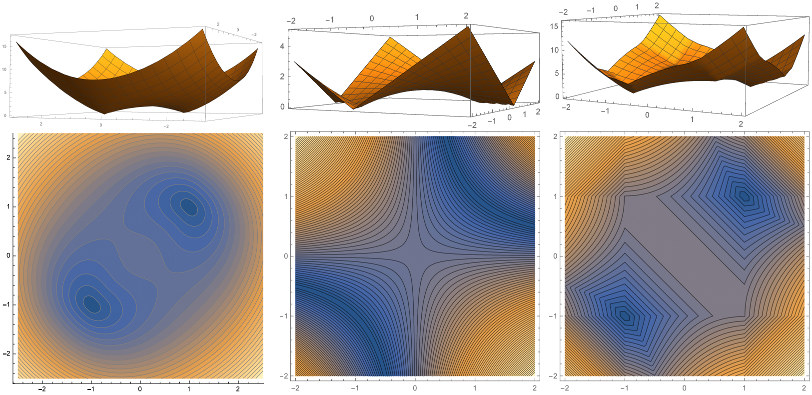
1. Find initial solution estimate $\hat{x}$.
   - Typically found via spectral method.

2. Run a "local search method."
   - Can be challenging to analyze.

Extensive literature in the smooth setting.

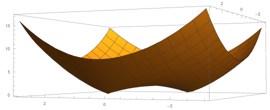
- http://sunju.org/research/nonconvex/
- Yuejie Chi, Yue M. Lu, and Yuxin Chen. "Nonconvex optimization meets low-rank matrix factorization: An overview." IEEE Transactions on Signal Processing 67.20 (2019): 5239-5269.

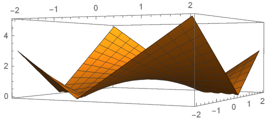
Conditioning in nonsmooth optimization

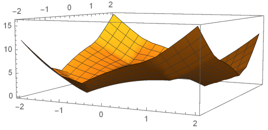


(a) (phase retrieval) (b) (blind deconvolution) (c) (robust PCA)

Conditioning in nonsmooth optimization

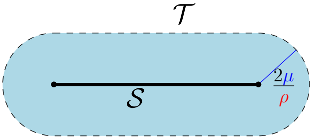




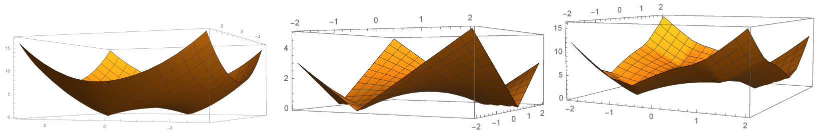


Three properties: Define $\mathcal{S} := \operatorname{argmin} F$.

- Weak convexity: $x \mapsto F(x) + \frac{\rho}{2}\|x\|^2$ is convex
- Sharpness: $F(x) - \min F \geq \mu \cdot \operatorname{dist}(x, \mathcal{S})$
- Lipschitz: F is L-Lipschitz on $\mathcal{T} := \left\{x \mid \operatorname{dist}(x, \mathcal{S}) < \frac{2\mu}{\rho}\right\}$



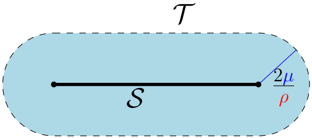
Conditioning in nonsmooth optimization



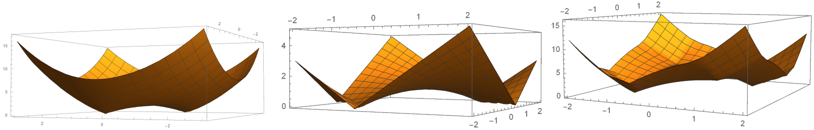
Three properties: Define $\mathcal{S} := \operatorname{argmin} F$.

- Weak convexity: $x \mapsto F(x) + \frac{\rho}{2}\|x\|^2 \quad \text{is convex}$
- Sharpness: $F(x) - \min F \geq \mu \cdot \operatorname{dist}(x, \mathcal{S})$
- Lipschitz: F is L-Lipschitz on $\mathcal{T} := \left\{x \mid \operatorname{dist}(x, \mathcal{S}) < \frac{2\mu}{\rho}\right\}$

Lemma: $\mathcal{T} \setminus \mathcal{S}$ contains no critical points.



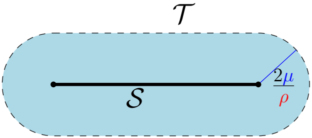
Conditioning in nonsmooth optimization



Three properties: Define $\mathcal{S} := \operatorname{argmin} F$.

- Weak convexity: $x \mapsto F(x) + \frac{\rho}{2}\|x\|^2$ is convex
- Sharpness: $F(x) - \min F \geq \mu \cdot \operatorname{dist}(x, \mathcal{S})$
- Lipschitz: F is L-Lipschitz on $\mathcal{T} := \left\{x \mid \operatorname{dist}(x, \mathcal{S}) < \frac{2\mu}{\rho}\right\}$

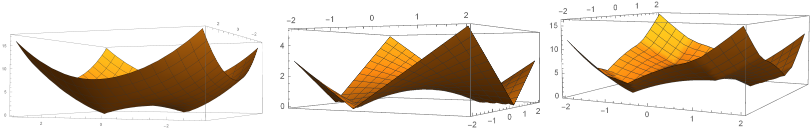
Lemma: $\mathcal{T} \setminus \mathcal{S}$ contains no critical points.



Summary:

- $\frac{\mu}{\rho}$ controls initialization
- $\frac{L}{\mu}$ controls speed

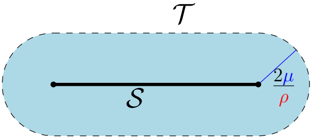
Conditioning in nonsmooth optimization



Three properties: Define $\mathcal{S} := \operatorname{argmin} F$.

- Weak convexity: $\quad x \mapsto F(x) + \frac{\rho}{2}\|x\|^2 \quad$ is convex
- Sharpness: $\quad F(x) - \min F \geq \mu \cdot \operatorname{dist}(x, \mathcal{S})$
- Lipschitz: $\quad F$ is L-Lipschitz on $\mathcal{T} := \left\{x \mid \operatorname{dist}(x, \mathcal{S}) < \frac{2\mu}{\rho}\right\}$

Lemma: $\mathcal{T} \setminus \mathcal{S}$ contains no critical points.



Summary:

- $\dfrac{\mu}{\rho}$ controls initialization $\qquad$ **RIP** $\Rightarrow \dfrac{\mu}{\rho} \asymp \dfrac{\kappa_1}{\kappa_2}\sqrt{\sigma_r(M_\sharp)}$
- $\dfrac{L}{\mu}$ controls speed $\qquad$ **RIP** $\Rightarrow \dfrac{L}{\mu} \asymp \dfrac{\kappa_2}{\kappa_1}\sqrt{\dfrac{\sigma_1(M_\sharp)}{\sigma_r(M_\sharp)}}$

Meta-Theorem:

Simple algorithms for sharp and weakly convex functions converge rapidly.

Meta-Theorem:

Simple algorithms for sharp and weakly convex functions converge rapidly.

Polyak subgradient method:

$$x^+ = x - \left(\frac{F(x) - \min F}{\|\nabla F(x)\|^2}\right) \nabla F(x)$$

Thm: (Polyak '67, Davis-D-MacPhee-Paquette '17)

Assuming $x_0 \in \mathcal{T}$, have

$$\frac{\operatorname{dist}(x_{t+1}; S)}{\operatorname{dist}(x_t; \mathcal{S})} \leq \sqrt{1 - \left(\frac{\mu}{L}\right)^2} \qquad \text{for all } t.$$

Meta-Theorem:

Simple algorithms for sharp and weakly convex functions converge rapidly.

Polyak subgradient method:

$$x^+ = x - \left(\frac{F(x) - \min F}{\|\nabla F(x)\|^2}\right) \nabla F(x)$$

Thm: (Polyak '67, Davis-D-MacPhee-Paquette '17)

Assuming $x_0 \in \mathcal{T}$, have

$$\frac{\operatorname{dist}(x_{t+1}; S)}{\operatorname{dist}(x_t; \mathcal{S})} \leq \sqrt{1 - \left(\frac{\mu}{L}\right)^2} \qquad \text{for all } t.$$

$\implies$ Off-the-shelf optimal sample and computational efficiency for phase retrieval (real/complex), blind deconvolution, and quadratic sensing.

Meta-Theorem:

Simple algorithms for sharp and weakly convex functions converge rapidly.

Polyak subgradient method:

$$x^+ = x - \left(\frac{F(x) - \min F}{\|\nabla F(x)\|^2}\right) \nabla F(x)$$

Thm: (Polyak '67, Davis-D-MacPhee-Paquette '17)

Assuming $x_0 \in \mathcal{T}$, have

$$\frac{\operatorname{dist}(x_{t+1}; S)}{\operatorname{dist}(x_t; \mathcal{S})} \leq \sqrt{1 - \left(\frac{\mu}{L}\right)^2} \qquad \text{for all } t.$$

$\implies$ Off-the-shelf optimal sample and computational efficiency for phase retrieval (real/complex), blind deconvolution, and quadratic sensing.

Remark:

- $\min F$ not known $\implies$ can update lower bounds (Hazan-Kakade '19)
- measurement errors $\implies$ linear convergence to a tolerance.

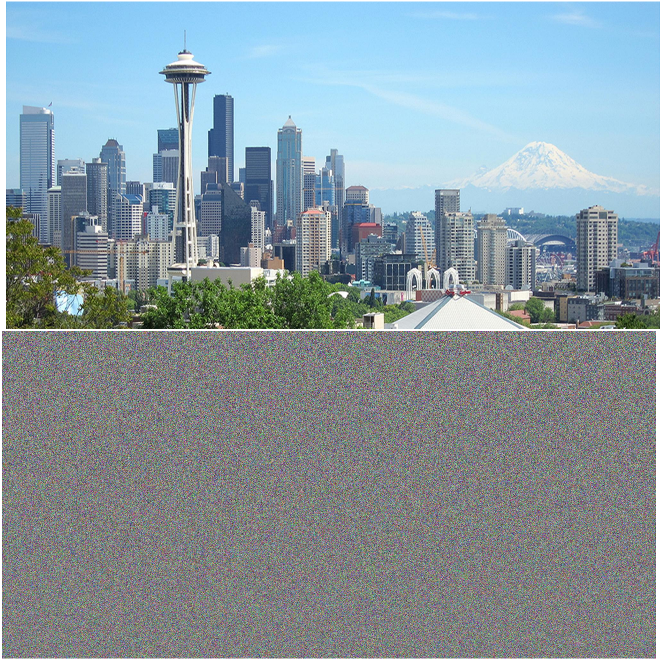


Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 1.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 2.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 3.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 4.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 5.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 6.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 7.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 8.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 9.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 10.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 11.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 12.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 13.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 14.

Figure: $(d, m) \approx (2^{23}, 2^{24})$. Iteration 15.

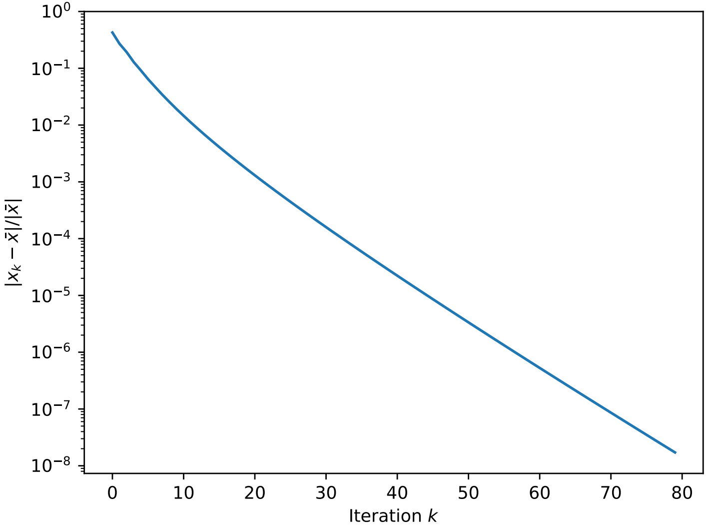


Figure: Convergence plot (iterates vs. $\|x_k - \bar{x}\|/\|\bar{x}\|$).

Stochastic weakly-convex minimization

Streaming & offline algorithms

$$\min_x \; F(x) = \mathbb{E}_z[f(x,z)]$$

Running assumption: weak convexity

$$f(\cdot, z) + \frac{\rho}{2}\|\cdot\|^2 \qquad \text{is convex.}$$

Streaming & offline algorithms

$$\min_{x} \; F(x) = \mathbb{E}_z[f(x, z)]$$

Running assumption: weak convexity

$$f(\cdot, z) + \frac{\rho}{2}\|\cdot\|^2 \qquad \text{is convex.}$$

Main example: convex compositions

$$x \mapsto h(c(x))$$

h is convex and L-Lipschitz; c is smooth with ℓ-Lipschitz Jacobian ($\rho = L\ell$)

Streaming & offline algorithms

$$\min_x \; F(x) = \mathbb{E}_z[f(x,z)]$$

Running assumption: weak convexity

$$f(\cdot,z) + \frac{\rho}{2}\|\cdot\|^2 \qquad \text{is convex.}$$

Main example: convex compositions

$$x \mapsto h(c(x))$$

h is convex and L-Lipschitz; c is smooth with ℓ-Lipschitz Jacobian ($\rho = L\ell$)

Two approaches:

- Streaming: Sample z_t and update x_t using $f(\cdot, z_t)$
- Offline: Sample $S = \{z_1, \ldots, z_n\}$ i.i.d. from P and approximate

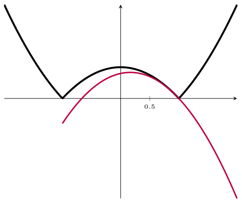
$$F(x) = \mathbb{E}_z[f(x,z)] \qquad \text{with} \qquad F^S(x) := \frac{1}{m}\sum_{i=1}^{m} f(x, z_i).$$

Interlude: subdifferential

Fact: For any $f\colon \mathbb{R}^d \to \mathbb{R}$, have equivalence:

- f is ρ-weakly convex
- Subgradient inequality: $\forall x \exists v_x$ satisfying

$$f(y) \geq f(x) + \langle v_x, y - x\rangle - \frac{\rho}{2}\|y - x\|^2$$

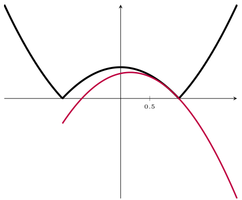


Interlude: subdifferential

Fact: For any $f\colon \mathbb{R}^d \to \mathbb{R}$, have equivalence:

- f is ρ-weakly convex
- Subgradient inequality: $\forall x \exists v_x$ satisfying

$$f(y) \geq f(x) + \langle v_x, y - x\rangle - \frac{\rho}{2}\|y - x\|^2$$



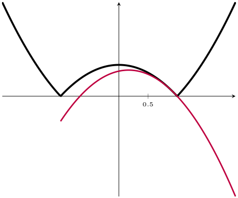
Subdifferential: $\partial f(x) := \{v_x\}$

Interlude: subdifferential

Fact: For any $f\colon \mathbb{R}^d \to \mathbb{R}$, have equivalence:

- f is ρ-weakly convex
- Subgradient inequality: $\forall x \exists v_x$ satisfying

$$f(y) \geq f(x) + \langle v_x, y - x\rangle - \frac{\rho}{2}\|y - x\|^2$$



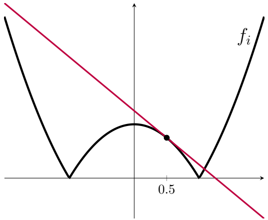
Subdifferential: $\partial f(x) := \{v_x\}$

Calculus: $\boxed{\partial(h \circ c)(x) := \nabla c(x)^T \partial h(c(x))}$

Four streaming algorithms

Problem:

$$\min_x \; F(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x))$$



Example: (Stochastic subgradient)[3] Choose $g \in \partial h_i(c_i(x))$ and

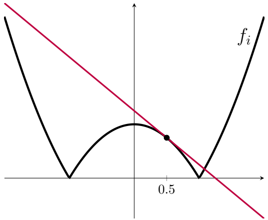
$$x^+ = x - \alpha \nabla c_i(x)^T g$$

[3] (Nemirovski-Juditsky-Lan-Shapiro '09, Ghadimi-Lan-Zhang '16. . .)

Four streaming algorithms

Problem:

$$\min_x \; F(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x))$$



Example: (Stochastic subgradient)[3] Choose $g \in \partial h_i(c_i(x))$ and

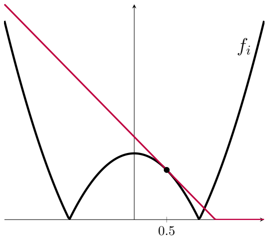
$$\begin{aligned} x^+ &= x - \alpha \nabla c_i(x)^T g \\ &= \operatorname*{argmin}_y \; \left\{ h_i(c_i(x)) + \langle \nabla c_i(x)^T g, y - x\rangle + \frac{1}{2\alpha}\|y - x\|^2 \right\} \end{aligned}$$

[3] (Nemirovski-Juditsky-Lan-Shapiro '09, Ghadimi-Lan-Zhang '16. . .)

Four streaming algorithms

Problem:

$$\min_x \; F(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x))$$



Example: (Stochastic clipped subgradient)[3] Choose $g \in \partial h_i(c_i(x))$ and

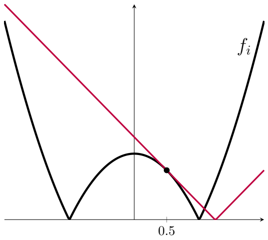
$$x^+ = \operatorname*{argmin}_y \; \left\{ \left[h_i(c_i(x)) + \langle \nabla c_i(x)^T g, y - x \rangle \right] \vee \texttt{lb} + \frac{1}{2\alpha}\|y - x\|^2 \right\}$$

[3] (Duchi-Ruan '17, Asi-Duchi '18 . . .)

Four streaming algorithms

Problem:

$$\min_{x} \; F(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x))$$



Example: (Stochastic prox-linear)[3]

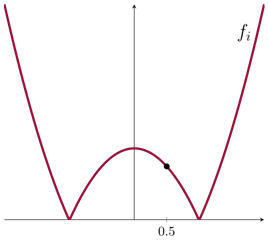
$$x^{+} = \operatorname*{argmin}_{y} \; \left\{ h_i\left(c_i(x) + \nabla c_i(x)(y-x)\right) + \frac{1}{2\alpha}\|y-x\|^2 \right\}$$

[3] (Burke '85, Lewis-Wright '15, Duchi-Ruan '17,. . .)

Four streaming algorithms

Problem:

$$\min_x \; F(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x))$$



Example: (Stochastic proximal point method)[3]

$$x^+ = \operatorname*{argmin}_y \; \left\{ h_i(c_i(y)) + \frac{1}{2\alpha}\|y - x\|^2 \right\}$$

[3] (Ryu-Boyd '16, Toulis-Tran-Airoldi '16, Bianchi '16...)

Model-Based streaming algorithm

$$\min_{x} \; F(x) = \mathbb{E}_z[f(x,z)].$$

Algorithm:

Sample: $z_t \sim P$

$$\text{Set: } x_{t+1} = \operatorname*{argmin}_{y} \; \left\{ f_{x_t}(y, z_t) + \frac{1}{2\alpha_t} \|y - x_t\|^2 \right\}$$

Model-Based streaming algorithm

$$\min_{x} \; F(x) = \mathbb{E}_z[f(x,z)].$$

Algorithm:

Sample: $z_t \sim P$

$$\text{Set: } x_{t+1} = \operatorname*{argmin}_{y} \; \left\{ f_{x_t}(y, z_t) + \frac{1}{2\alpha_t} \|y - x_t\|^2 \right\}$$

Assumption:

$$\boxed{f_x(x,z) = f(x,z) \quad \text{and} \quad f_x(y,z) \le f(y,z) + \frac{\tau}{2}\|y - x\|^2 \qquad \forall x, y}$$

Model-Based streaming algorithm

$$\min_x \; F(x) = \mathbb{E}_z[f(x,z)].$$

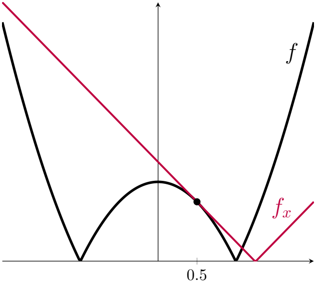
Algorithm:

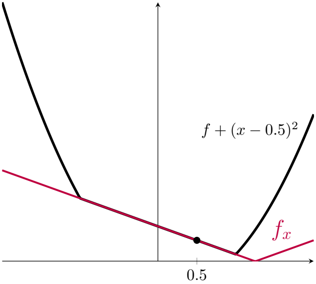
Sample: $z_t \sim P$

$$\text{Set: } x_{t+1} = \operatorname*{argmin}_y \; \left\{ f_{x_t}(y, z_t) + \frac{1}{2\alpha_t}\|y - x_t\|^2 \right\}$$

Assumption:

$$\boxed{f_x(x,z) = f(x,z) \quad \text{and} \quad f_x(y,z) \le f(y,z) + \frac{\tau}{2}\|y-x\|^2 \qquad \forall x, y}$$





Phase Retrieval Experiments

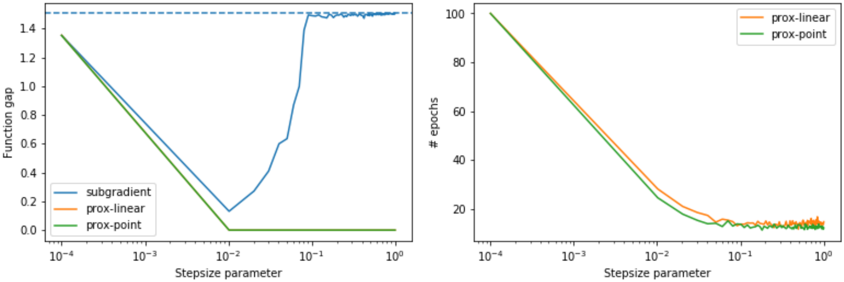


Figure: Target accuracy 10^{-4}.

Phase Retrieval Experiments

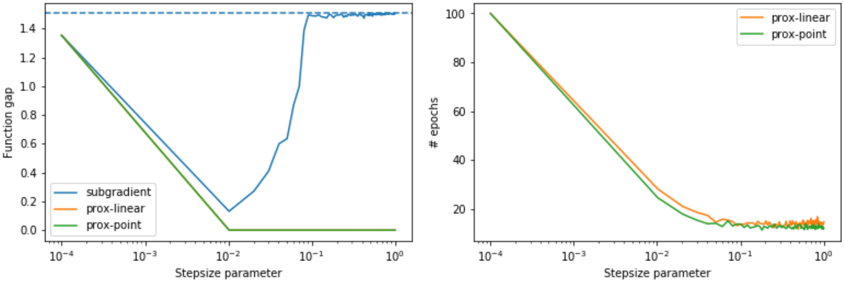


Figure: Target accuracy 10^{-4}.

Towards convergence guarantees. . .

Challenges

1. Biased search directions:

$$\text{SGD:} \quad \mathbb{E}\left[x_{t+1} - x_t\right] = -\alpha_t \nabla F(x_t)$$

$$\text{MODEL:} \quad \mathbb{E}\left[x_{t+1} - x_t\right] \qquad \text{no clear meaning!}$$

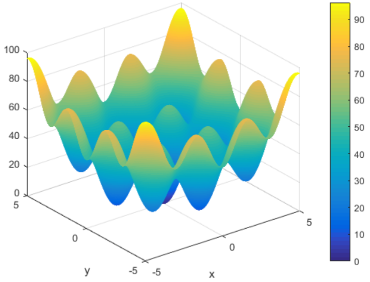
Challenges

1. Biased search directions:

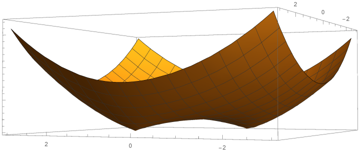
$$\text{SGD:} \quad \mathbb{E}\left[x_{t+1} - x_t\right] = -\alpha_t \nabla F(x_t)$$

$$\text{MODEL:} \quad \mathbb{E}\left[x_{t+1} - x_t\right] \qquad \textcolor{red}{\text{no clear meaning!}}$$

2. Unclear what to measure:



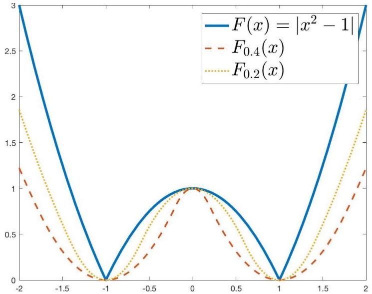
(a) $F(x) - \inf F \geq \Omega(1)$



(b) $\|\nabla F(x)\| \geq \Omega(1)$

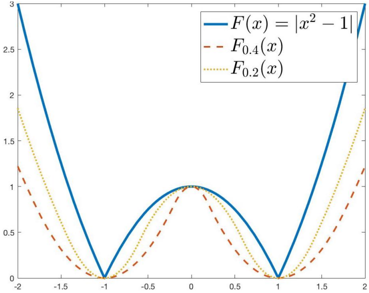
Moreau envelope

$$F_\lambda(x) = \inf_y \left\{ F(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}$$



Moreau envelope

$$F_\lambda(x) = \inf_y \left\{ F(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}$$



Implicit Smoothing. F_λ is C^1 for all $\lambda < \rho^{-1}$ with

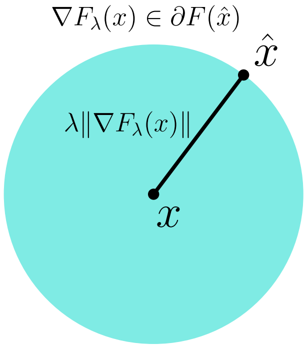
$$\boxed{\nabla F_\lambda(x) = \lambda^{-1}\left(x - \mathrm{prox}_{\lambda F}(x)\right)}$$

where

$$\mathrm{prox}_{\lambda F}(x) = \operatorname*{argmin}_y \left\{ F(y) + \tfrac{1}{2\lambda} \|y - x\|^2 \right\}$$

Moreau envelope

- Approximate stationarity: set $\hat{x} = \mathrm{prox}_{\lambda F}(x)$



Small $\|\nabla F_\lambda(x)\| \implies x$ is nearby a nearly stationary point of F.

Convergence guarantees

Assumptions: For all x, y, z, have

1. **(accuracy)** $\mathbb{E} f_x(x,z) = f(x)$ and $\mathbb{E} f_x(y,z) \le f(y) + \frac{\tau}{2}\|y-x\|^2$
2. **(convexity)** $f_x(\cdot, z)$ are ρ-weakly convex
3. **(Lipschitz)** $f_x(x,z) - f_x(y,z) \le L(z)\|y-x\|$ where $\mathbb{E}[L(z)^2] < \infty$

Moreau envelope is almost Lyapunov function for algorithm dynamics!

Theorem (Davis-D '18)

Setting $\lambda = 1/2(\rho+\tau)$, methods achieve approximate descent on envelope:

$$\mathbb{E}[F_\lambda(x_t) - F_\lambda(x_{t+1})] \ge \alpha_t \mathbb{E}\|\nabla F_\lambda(x_t)\|^2/\lambda - \alpha_t^2 \mathbb{E}\|L\|^2/\lambda$$

Hence for $\alpha_t \approx T^{-1/2}$ get complexity $\mathbb{E}\|\nabla F_\lambda(x_{t^*})\| = O(T^{-1/4})$.

[4] Duchi and Ruan. Stochastic methods for composite optimization problems. (2017)

[5] Nurminskii. The quasigradient method for the solving of the nonlinear programming problems (1973)

Convergence guarantees

Assumptions: For all x, y, z, have

1. **(accuracy)** $\mathbb{E} f_x(x,z) = f(x)$ and $\mathbb{E} f_x(y,z) \leq f(y) + \frac{\tau}{2}\|y-x\|^2$
2. **(convexity)** $f_x(\cdot, z)$ are ρ-weakly convex
3. **(Lipschitz)** $f_x(x,z) - f_x(y,z) \leq L(z)\|y-x\|$ where $\mathbb{E}[L(z)^2] < \infty$

Moreau envelope is almost Lyapunov function for algorithm dynamics!

Theorem (Davis-D '18)

Setting $\lambda = 1/2(\rho+\tau)$, methods achieve approximate descent on envelope:

$$\mathbb{E}[F_\lambda(x_t) - F_\lambda(x_{t+1})] \geq \alpha_t \mathbb{E}\|\nabla F_\lambda(x_t)\|^2/\lambda \qquad \textit{(gradient descent bound)}$$

Hence for $\alpha_t \approx T^{-1/2}$ get complexity $\mathbb{E}\|\nabla F_\lambda(x_{t^*})\| = O(T^{-1/4})$.

---

[4] Duchi and Ruan. Stochastic methods for composite optimization problems. (2017)

[5] Nurminskii. The quasigradient method for the solving of the nonlinear programming problems (1973)

Convergence guarantees

Assumptions: For all x, y, z, have

1. **(accuracy)** $\mathbb{E} f_x(x,z) = f(x)$ and $\mathbb{E} f_x(y,z) \leq f(y) + \frac{\tau}{2}\|y-x\|^2$
2. **(convexity)** $f_x(\cdot, z)$ are ρ-weakly convex
3. **(Lipschitz)** $f_x(x,z) - f_x(y,z) \leq L(z)\|y-x\|$ where $\mathbb{E}[L(z)^2] < \infty$

Moreau envelope is almost Lyapunov function for algorithm dynamics!

Theorem (Davis-D '18)

Setting $\lambda = 1/2(\rho+\tau)$, *methods achieve approximate descent on envelope:*

$$\mathbb{E}[F_\lambda(x_t) - F_\lambda(x_{t+1})] \geq \alpha_t \mathbb{E}\|\nabla F_\lambda(x_t)\|^2/\lambda - \alpha_t^2 \mathbb{E}\|L\|^2/\lambda$$

Hence for $\alpha_t \approx T^{-1/2}$ *get complexity* $\mathbb{E}\|\nabla F_\lambda(x_{t^*})\| = O(T^{-1/4})$.

[4] Duchi and Ruan. Stochastic methods for composite optimization problems. (2017)

[5] Nurminskii. The quasigradient method for the solving of the nonlinear programming problems (1973)

Convergence guarantees

Assumptions: For all x, y, z, have

1. **(accuracy)** $\mathbb{E} f_x(x, z) = f(x)$ and $\mathbb{E} f_x(y, z) \le f(y) + \frac{\tau}{2}\|y - x\|^2$
2. **(convexity)** $f_x(\cdot, z)$ are ρ-weakly convex
3. **(Lipschitz)** $f_x(x, z) - f_x(y, z) \le L(z)\|y - x\|$ where $\mathbb{E}[L(z)^2] < \infty$

Moreau envelope is almost Lyapunov function for algorithm dynamics!

Theorem (Davis-D '18)

Setting $\lambda = 1/2(\rho + \tau)$, methods achieve approximate descent on envelope:

$$\mathbb{E}[F_\lambda(x_t) - F_\lambda(x_{t+1})] \ge \alpha_t \mathbb{E}\|\nabla F_\lambda(x_t)\|^2/\lambda - \alpha_t^2 \mathbb{E}\|L\|^2/\lambda$$

Hence for $\alpha_t \approx T^{-1/2}$ get complexity $\mathbb{E}\|\nabla F_\lambda(x_{t^*})\| = O(T^{-1/4})$.

Almost sure convergence of stochastic prox-linear[4] and subgradient[5] previously known. Functional rates improve under convexity.

[4] Duchi and Ruan. Stochastic methods for composite optimization problems. (2017)

[5] Nurminskii. The quasigradient method for the solving of the nonlinear programming problems (1973)

Off-line Algorithms

Form i.i.d. sample $S = \{z_1, \ldots, z_n\} \subset \mathbb{R}^d$ from P and approximate

$$F(x) = \mathbb{E}_z[f(x,z)] \qquad \text{with} \qquad F^S(x) := \frac{1}{m}\sum_{i=1}^{m} f(x, z_i)$$

Theorem (Davis-D '18)

Setting $\lambda = 1/2\rho$, with probability $1-\gamma$, the estimate holds:

$$\sup_{\|x\|\le R} \|\nabla F^S_\lambda(x) - \nabla F_\lambda(x)\|_2 \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{L^2 d}{m} \cdot \ln\left(\frac{\rho R}{\gamma}\right)}\right)$$

- Estimate is tight even for smooth losses.

Off-line Algorithms

Uniform vs. Graphical Convergence:

$$\sup_{\|x\|\leq R} \|\nabla F_\lambda^S(x) - \nabla F_\lambda(x)\|_2 \approx \mathrm{dist}(\mathrm{gph}\ \partial F, \mathrm{gph}\ \partial F^S).$$

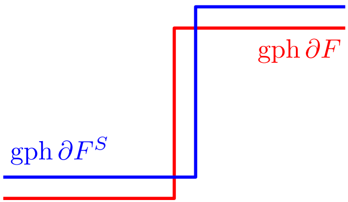


Figure: Graphical but not uniform

Off-line Algorithms

Uniform vs. Graphical Convergence:

$$\sup_{\|x\| \le R} \|\nabla F_\lambda^S(x) - \nabla F_\lambda(x)\|_2 \approx \mathrm{dist}(\mathrm{gph}\ \partial F, \mathrm{gph}\ \partial F^S).$$

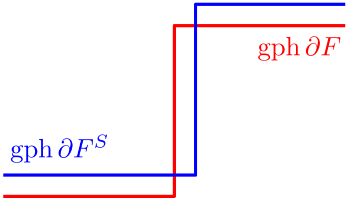


Figure: Graphical but not uniform

- Other results: d-independent rates for GLM, landscape analysis, regularity . . .

Proofs use

- nonsmooth analysis (Brøndsted-Rockafellar '65, Ekeland '79, Attouch '84)
- stability of ERM (Shalev-Shwartz et al. '09, Bousquet et al. '02)
- concentration (McDiarmid '89, Bartlett-Mendelson '02)

Fast stochastic algorithms

Back to phase retrieval:

$$\min_{x\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m}|(a_i^\top x)^2-(a_i^\top x_\sharp)^2| \qquad \approx \qquad \min_{x\in\mathbb{R}^d} \mathbb{E}_a|(a^\top x)^2-(a^\top x_\sharp)^2|.$$

Fast stochastic algorithms

Back to phase retrieval:

$$\min_{x \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} |(a_i^\top x)^2 - (a_i^\top x_\sharp)^2| \qquad \approx \qquad \min_{x \in \mathbb{R}^d} \mathbb{E}_a |(a^\top x)^2 - (a^\top x_\sharp)^2|.$$

When is the sample average well conditioned?

- sharpness μ is ubiquitous [small ball technique (Mendelson '14)]
- parameters ρ and L rely on **light tails**

Fast stochastic algorithms

Back to phase retrieval:

$$\min_{x\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} |(a_i^\top x)^2 - (a_i^\top x_\sharp)^2| \qquad \approx \qquad \min_{x\in\mathbb{R}^d} \mathbb{E}_a|(a^\top x)^2 - (a^\top x_\sharp)^2|.$$

When is the sample average well conditioned?

- sharpness μ is ubiquitous [small ball technique (Mendelson '14)]
- parameters ρ and L rely on **light tails**

Are there fast algorithms for the population objective?

$$\min_{x\in\mathbb{R}^d} \mathbb{E}_a f(x,a).$$

Fast stochastic algorithms

Back to phase retrieval:

$$\min_{x \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} |(a_i^\top x)^2 - (a_i^\top x_\sharp)^2| \qquad \approx \qquad \min_{x \in \mathbb{R}^d} \mathbb{E}_a |(a^\top x)^2 - (a^\top x_\sharp)^2|.$$

When is the sample average well conditioned?

- sharpness μ is ubiquitous [small ball technique (Mendelson '14)]
- parameters ρ and L rely on **light tails**

Are there fast algorithms for the population objective? **Yes!**

$$\min_{x \in \mathbb{R}^d} \mathbb{E}_a f(x, a).$$

Fast stochastic algorithms

Back to phase retrieval:

$$\min_{x\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m |(a_i^\top x)^2 - (a_i^\top x_\sharp)^2| \qquad \approx \qquad \min_{x\in\mathbb{R}^d} \mathbb{E}_a|(a^\top x)^2 - (a^\top x_\sharp)^2|.$$

When is the sample average well conditioned?

- sharpness μ is ubiquitous [small ball technique (Mendelson '14)]
- parameters ρ and L rely on **light tails**

Are there fast algorithms for the population objective? **Yes!**

$$\min_{x\in\mathbb{R}^d} \mathbb{E}_a f(x,a).$$

Theorem (Davis-D-Charisopoulos '19)

Stochastic algorithms on **weakly convex** *and* **sharp functions** *converge linearly in the tube $\mathcal{T}$ w.h.p.*

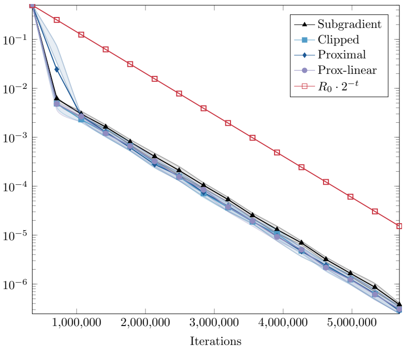
Surprising:

- Evaluating $\mathbb{E}_a[f(x,a)]$ to ε accuracy requires $O(\varepsilon^{-2})$ samples
- **This result:** to get ε close to minimizer, need $O\left(\frac{L^2}{\mu^2}\log(\varepsilon^{-1})\right)$ samples.

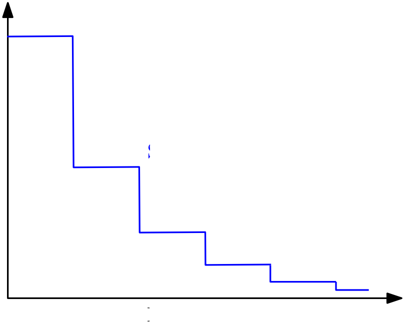
Fast stochastic algorithms

Algorithm:[6]

Model-based algorithms $+$ step-decay



(a) Phase retrieval



(b) Step decay

[6] related algorithm in convex setting (Xu-Lin-Yang '16)

Thank you

References

- Stochastic model-based minimization of weakly convex functions.
  Davis, D, SIAM J. Optim., 29, no. 1, 207-239, 2018.
- The nonsmooth landscape of phase retrieval.
  Davis, D, Paquette. To appear in IMA J. Numer. Anal. 2018
- Uniform graphical convergence of subgradients in nonconvex optimization
  Davis, D, To appear in Math. Oper. Res. arXiv:1810.07590.
- Subgradient methods for sharp weakly convex functions
  Davis, D, MacPhee, Paquette, J. Optim. Theory. Appl., 179, no. 3, 962-982, 2018.
- Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence
  Charisopoulos, Chen, Davis, D, Diaz, Ding, arXiv:1904.10020.
- Composite optimization for robust blind deconvolution
  Charisopoulos, Davis, Diaz, D, arXiv:1901.01624.
- Stochastic algorithms with step decay converge linearly on sharp functions
  Davis, D, Charisopoulos, arXiv:1907.09547.