Minimization of convex composites

Courtney Paquette
Joint work with D. Drusvyatskiy (UW)

Department of Mathematics
Ohio State University (Columbus)

Lehigh University
September 22, 2017

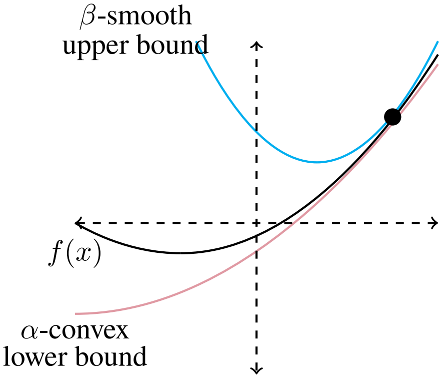
A function f is α-convex and β-smooth if

$$q_x \leq f \leq Q_x$$

where

$$q_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2$$

$$Q_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2$$



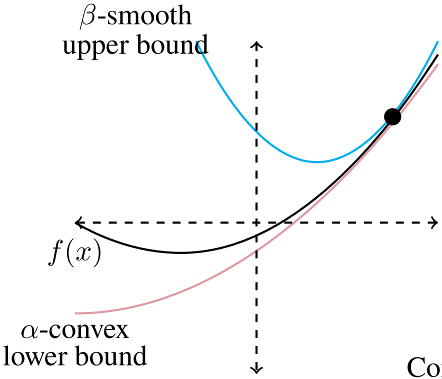
A function f is α-convex and β-smooth if

$$q_x \leq f \leq Q_x$$

where

$$q_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2$$

$$Q_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2$$



Condition number: $\kappa = \frac{\beta}{\alpha}$

A brief history of 1st order methods...

Black-box model: objective function accessed via *oracles*:

- zeroth-order: $f(x)$
- $\boxed{\text{1st-order: } f(x) \text{ and } \nabla f(x)}$
- 2nd-order: $f(x)$, $\nabla f(x)$, and $\nabla^2 f(x)$

(Yudin-Nemirovsky '83)

An interlude into 1st-order lower bounds

Lower complexity: the num. of calls to the oracle that **any** algorithm needs to obtain an ε-approx. minima

Measuring complexity:

- **(Gradients):** $\|\nabla f(x)\| < \varepsilon$
- **(Function values):** $f(x) - f^* < \varepsilon$
- **(Iterates):** $\|x - x^*\| < \varepsilon$

Complexity of first-order methods

Gradient descent: $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$

	$\|\nabla f(x)\| < \varepsilon$		
	β-smooth	β-smooth/convex	α-convex
Gradient descent	$\left(\frac{\beta}{\varepsilon}\right)^2$	$\frac{\beta}{\varepsilon}$	$\kappa \cdot \log(\frac{1}{\varepsilon})$

(Nesterov '83, Yudin-Nemirovsky '83)

Complexity of first-order methods

Gradient descent: $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$

	$\|\nabla f(x)\| < \varepsilon$		
	β-smooth	β-smooth/convex	α-convex
Gradient descent	$\left(\frac{\beta}{\varepsilon}\right)^2$	$\frac{\beta}{\varepsilon}$	$\kappa \cdot \log(\frac{1}{\varepsilon})$
Accelerated gradient	?	$\left(\frac{\beta}{\varepsilon}\right)^{2/3}$	$\sqrt{\kappa} \cdot \log(\frac{1}{\varepsilon})$

(Nesterov '83, Yudin-Nemirovsky '83)

Nonsmooth & Nonconvex minimization

Convex composition

$$\min_{x} \; F(x) := h(c(x)) + g(x)$$

- $c : \mathbf{R}^n \to \mathbf{R}^m$ is C^1-smooth with β-Lipschitz Jacobian
- $h : \mathbf{R}^m \to \mathbf{R}$ is closed, convex, and 1-Lipschitz
- $g : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ is convex

(Burke '85, '91, Fletcher '82, Powell '84, Wright '90, Yuan '83, Cartis-Gould-Toint '11)

Nonsmooth & Nonconvex minimization

Examples:

- Additive composite minimization

$$\min_{x}\ c(x) + g(x)$$

- Nonlinear least squares

$$\min_{x \in Q}\ \|c(x)\|$$

  - Robust Phase Retrieval: $\min_x \sum_{i=1}^m |(a_i^T x)^2 - b_i|$
  - Non-neg. Factorization: $\min_{X,Y \geq 0} \left\|XY^T - D\right\|$

- Exact penalty subproblem:

$$\min_{x}\ g(x) + \mathrm{dist}_K(c(x))$$

Prox-Linear algorithm

$$\boxed{\min_x \; F(x) = g(x) + h(c(x))}$$

Local Model:

$$F_x(y) := g(y) + h\big(c(x) + \nabla c(x)(y - x)\big)$$

Accuracy: $|F_x(y) - F(x)| \leq \frac{\beta}{2} \|y - x\|^2$

Prox-Linear algorithm

$$\boxed{\min_x \; F(x) = g(x) + h(c(x))}$$

Local Model:

$$F_x(y) := g(y) + h\big(c(x) + \nabla c(x)(y - x)\big)$$

Accuracy: $|F_x(y) - F(x)| \leq \frac{\beta}{2} \|y - x\|^2$

$$\Rightarrow \quad F(x) \leq F_x(y) + \frac{\beta}{2}\|y - x\|^2 \quad \forall\, y$$

Prox-linear method:

$$x^+ = \operatorname*{argmin}_x \left\{ F_x(y) + \frac{\beta}{2}\|y - x\|^2 \right\}$$

Big assumption: x^+ is computable (for now)

Example of Prox-linear method

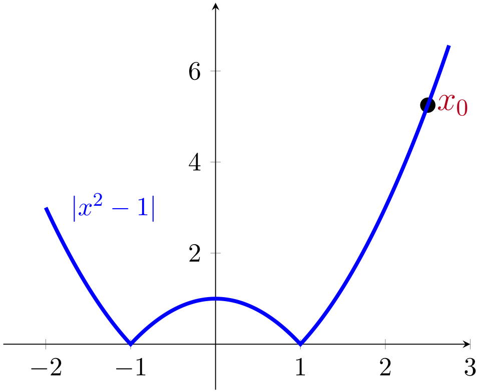
Looking for *first-order stationary*: $0 \in \partial F(x) = \partial g(x) + \nabla c(x)^* \partial h(c(x))$

Example of Prox-linear method

Looking for *first-order stationary*: $0 \in \partial F(x) = \partial g(x) + \nabla c(x)^* \partial h(c(x))$

$$\boxed{\min_{x \in \mathbb{R}} F(x) := |x^2 - 1|}$$

Iterate	$F(x)$	$\partial F(x)$
$x_0 = 2.50$	5.25	5.00

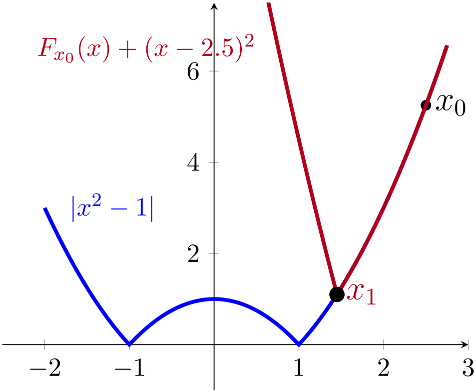


Example of Prox-linear method

Looking for *first-order stationary*: $0 \in \partial F(x) = \partial g(x) + \nabla c(x)^* \partial h(c(x))$

$$\boxed{\min_{x\in\mathbb{R}} F(x) := |x^2 - 1|}$$

Iterate	$F(x)$	$\partial F(x)$
$x_0 = 2.50$	5.25	5.00
$x_1 = 1.45$	1.10	2.90

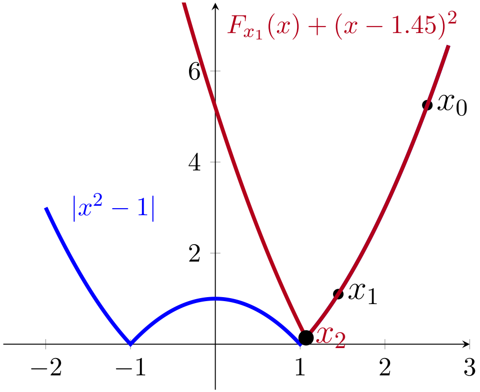


Example of Prox-linear method

Looking for *first-order stationary*: $\quad 0 \in \partial F(x) = \partial g(x) + \nabla c(x)^* \partial h(c(x))$

$$\boxed{\min_{x \in \mathbb{R}} F(x) := |x^2 - 1|}$$

Iterate	$F(x)$	$\partial F(x)$
$x_0 = 2.50$	5.25	5.00
$x_1 = 1.45$	1.10	2.90
$x_2 = 1.07$	0.14	2.13

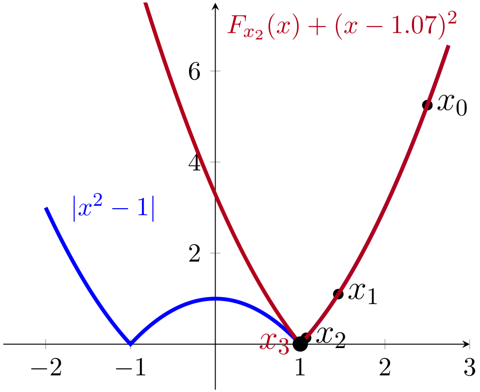


Example of Prox-linear method

Looking for *first-order stationary*: $\quad 0 \in \partial F(x) = \partial g(x) + \nabla c(x)^* \partial h(c(x))$

$$\boxed{\min_{x\in\mathbb{R}} F(x) := |x^2 - 1|}$$

Iterate	$F(x)$	$\partial F(x)$
$x_0 = 2.50$	5.25	5.00
$x_1 = 1.45$	1.10	2.90
$x_2 = 1.07$	0.14	2.13
$x_3 = 1.002$	0.0046	2.005

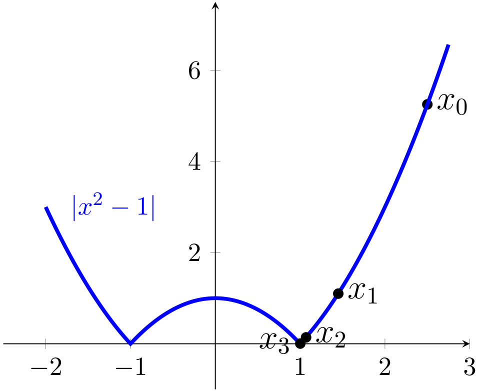


Example of Prox-linear method

Looking for *first-order stationary*: $0 \in \partial F(x) = \partial g(x) + \nabla c(x)^* \partial h(c(x))$

$$\boxed{\min_{x \in \mathbb{R}} F(x) := |x^2 - 1|}$$

Iterate	$F(x)$	$\partial F(x)$
$x_0 = 2.50$	5.25	5.00
$x_1 = 1.45$	1.10	2.90
$x_2 = 1.07$	0.14	2.13
$x_3 = 1.002$	0.0046	2.005



No finite termination!

Sublinear rate

The **prox-gradient**

$$\mathcal{G}(x) = \beta(x - x^+)$$

Convergence Rate: If x^+ can be computed exactly,

$$\|\mathcal{G}(x_t)\| < \varepsilon \quad \text{after} \quad \mathcal{O}\left(\frac{\beta}{\varepsilon^2}\left(F(x_0) - F^*\right)\right) \text{ iterations}$$

$$\boxed{\text{What does } \|\mathcal{G}(x_t)\| < \varepsilon \text{ mean?}}$$

Sublinear rate

The **prox-gradient**

$$\mathcal{G}(x) = \beta(x - x^+)$$

Convergence Rate: If x^+ can be computed exactly,

$$\|\mathcal{G}(x_t)\| < \varepsilon \quad \text{after} \quad \mathcal{O}\left(\frac{\beta}{\varepsilon^2}\big(F(x_0) - F^*\big)\right) \text{ iterations}$$

What does $\|\mathcal{G}(x_t)\| < \varepsilon$ mean?

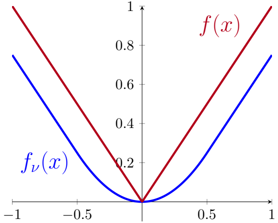
Additive composite $F(x) = c(x) + g(x)$ setting:

$$\operatorname{dist}(0, \partial F(x_{t+1})) \leq 2 \|\mathcal{G}(x_t)\|.$$

False in general!

What does $\|\mathcal{G}(x)\| < \varepsilon$ mean?

Moreau envelope: $\quad f_\nu(x) := \min_y \left\{ f(y) + \frac{1}{2\nu}\|y-x\|^2 \right\}$



Different Motivation: $F + \beta\|\cdot\|^2$ is convex

$$\Rightarrow \qquad \boxed{\nabla F_s(x) = s^{-1}(x - \mathrm{prox}_{sF}(x))} \qquad \forall s < \beta^{-1}$$

Thm: (Drusvyatskiy-P ’16)

$$\frac{1}{4}\left\|\nabla F_{\frac{1}{2\beta}}(x)\right\| \le \|\mathcal{G}(x)\| \le \frac{3}{2}\left(1 + \frac{1}{\sqrt{2}}\right)\left\|\nabla F_{\frac{1}{2\beta}}(x)\right\|$$

$$\Rightarrow \quad \mathrm{dist}((x,0), \mathrm{gph}\, \partial F) \le \|\mathcal{G}(x)\|.$$

Complexity Theory Question

Convergence Rate: If x^+ can be computed exactly,

$$\|\mathcal{G}(x_t)\| < \varepsilon \quad \text{after} \quad \mathcal{O}\left(\tfrac{\beta}{\varepsilon^2}\left(F(x_0) - F^*\right)\right) \text{ iterations}$$

What is the complexity of this problem class if x^+ is computed **inexactly**?

$$x^+ := \underset{y}{\operatorname{argmin}} \left\{ h\left(c(x) + \nabla c(x)(y - x)\right) + \frac{\beta}{2}\|y - x\|^2 + g(y) \right\}$$

Inexact Prox-Linear

$$\min_x \; h(c(x)) + g(x)$$

$$x^+ := \operatorname*{argmin}_y \left\{ h\left(c(x) + \nabla c(x)(y-x)\right) + \frac{\beta}{2}\|y-x\|^2 + g(y) \right\}$$

Thm: (Drusvyatskiy-P '16) Define

$$\|\nabla c\| := \max_{x \in \text{dom} g} \|\nabla c(x)\|.$$

There exists a method such that

$$\|\mathcal{G}(x)\| < \varepsilon \quad \text{after} \quad \tilde{\mathcal{O}}\left(\frac{\beta \|\nabla c\|}{\varepsilon^{2+1}} (F(x_0) - F^*) \right) \quad \text{iterations.}$$

Inexact Prox-Linear

$$\min_x \; h(c(x)) + g(x)$$

$$x^+ := \operatorname*{argmin}_y \left\{h\left(c(x) + \nabla c(x)(y-x)\right) + \frac{\beta}{2}\|y-x\|^2 + g(y)\right\}$$

Thm: (Drusvyatskiy-P '16) Define

$$\|\nabla c\| := \max_{x \in \mathrm{dom} g} \|\nabla c(x)\|.$$

There exists a method such that

$$\|\mathcal{G}(x)\| < \varepsilon \quad \text{after} \quad \tilde{\mathcal{O}}\left(\frac{\beta\|\nabla c\|}{\varepsilon^{2+1}}(F(x_0) - F^*)\right) \quad \text{iterations.}$$

Strategies:

- (inexact) prox-linear
- Smooth+(inexact) prox-linear + fast grad. subsolves
  - Smooth the function h
  - Run prox-linear method
  - Solve sub-problems with a fast gradient method
- Smooth + prox-gradient

Inexact Prox-Linear

$$\min_x \ h(c(x)) + g(x)$$

$$x^+ := \operatorname*{argmin}_y \left\{h\left(c(x) + \nabla c(x)(y-x)\right) + \frac{\beta}{2}\|y-x\|^2 + g(y)\right\}$$

Thm: (Drusvyatskiy-P '16) Define

$$\|\nabla c\| := \max_{x \in \mathrm{dom} g} \|\nabla c(x)\|.$$

There exists a method such that

$$\|\mathcal{G}(x)\| < \varepsilon \quad \text{after} \quad \tilde{\mathcal{O}}\left(\frac{\beta\|\nabla c\|}{\varepsilon^{2+1}}(F(x_0) - F^*)\right) \quad \text{iterations.}$$

Strategies:

- (inexact) prox-linear
- Smooth+(inexact) prox-linear + fast grad. subsolves
  - Smooth the function h
  - Run prox-linear method
  - Solve sub-problems with a fast gradient method
- Smooth + prox-gradient

Smoothing

We "smooth" the function h with the Moreau envelope

$$h_\nu(x) := \min_y \left\{h(y) + \tfrac{1}{2\nu}\|y - x\|^2\right\}.$$

Lemma I: (Optimality conditions) (Drusvyatskiy-P '16)

For any $\nu > 0$ and h_ν a smoothed h, then

$$\|\mathcal{G}(x)\| \leq \|\mathcal{G}^\nu(x)\| + \sqrt{\frac{\beta\nu}{2}}.$$

Smoothing parameter: $\frac{\varepsilon}{2} = \sqrt{\frac{\beta\nu}{2}} \quad \Rightarrow \quad \boxed{\nu = \frac{\varepsilon^2}{2\beta}}$

Smoothing + prox-linear + fast grad. subsolves

Step t:
Run a linearly convergent method $\mathcal{M}$ starting at z_0 to solve

$$x_{t+1}^{+} \approx \operatorname*{argmin}_{y} \Big\{ h_\nu(c(x_t) + \nabla c(x_t)(y - x_t)) + \frac{\beta}{2}\|y - x_t\|^2 + g(y) \Big\}.$$

A method $\mathcal{M}$ is linearly convergent if

$$f(z_i) - f^* \leq A\,(1-\tau)^i\,(f(z_0) - f^*)$$

Accuracy of the subproblem?

The subproblem:

$$x_{t+1}^{+} \approx \underset{y}{\operatorname{argmin}} \left\{ h_{\nu}(c(x_t) + \nabla c(x_t)(y - x_t)) + \frac{\beta}{2} \|y - x_t\|^2 + g(y) \right\}.$$

- Solve with $\mathcal{M}$: num. of inner iterations to get an ε-approx. iterate

$$\mathcal{O}\left(\frac{1}{\tau} \cdot \ln\left(\frac{A\|z_0 - z^*\|^2}{\varepsilon}\right)\right)$$

Difficulty in bounding $\|z_0 - z^*\|$

Inner complexity: adaptive approach

$$\varphi(y; x_t) := h_\nu(c(x_t) + \nabla c(x_t)(y - x_t)) + \tfrac{\beta}{2} \|y - x_t\|^2 + g(y)$$

Relative decrease condition

$$\varphi(x^+; x_t) - \varphi^*(x_t) \leq \frac{\beta}{4} \left\|x^+ - x_t^*\right\|^2$$

Inner complexity: adaptive approach

$$\varphi(y; x_t) := h_\nu(c(x_t) + \nabla c(x_t)(y - x_t)) + \tfrac{\beta}{2}\|y - x_t\|^2 + g(y)$$

Relative decrease condition

$$\varphi(x^+; x_t) - \varphi^*(x_t) \leq \frac{\beta}{4}\left\|x^+ - x_t^*\right\|^2$$

Lemma II (Controlling inner complexity) (Drusvyatskiy-P '16):
Run the subproblem with a method $\mathcal{M}$ for

$$T := \left\lceil \frac{1}{\tau} \log\left(2\beta^{-1} A \cdot \mathrm{lip}(\nabla\varphi)\right) \right\rceil$$

Then

$$\varphi(x_{t+1}; x_t) - \varphi^*(x_t) \leq \frac{\beta}{4}\left\|x_{t+1} - x_t^*\right\|^2$$

Complexity result

Lemma III: Outer complexity (Drusvyatskiy-P '16)
Suppose

$$\varphi(x_{t+1}; x_t) - \varphi^*(x_t) \le \frac{\beta}{4} \left\| x_{t+1} - x_t^* \right\|^2, \quad \forall t$$

Then

$$\|\mathcal{G}^\nu(x)\| < \varepsilon \quad \text{after} \quad \mathcal{O}\left(\frac{\beta}{\varepsilon^2} \big(F(x_0) - F^* \big) \right)$$

Complexity result

Lemma III: Outer complexity (Drusvyatskiy-P '16)
Suppose

$$\varphi(x_{t+1}; x_t) - \varphi^*(x_t) \leq \frac{\beta}{4} \|x_{t+1} - x_t^*\|^2, \quad \forall t$$

Then

$$\|\mathcal{G}^\nu(x)\| < \varepsilon \quad \text{after} \quad \mathcal{O}\left(\frac{\beta}{\varepsilon^2}\big(F(x_0) - F^*\big)\right)$$

Global Complexity (Inner · Outer):

$$\|\mathcal{G}^\nu(x)\| < \varepsilon \quad \text{after} \quad \tilde{\mathcal{O}}\left(\frac{1}{\tau} \cdot \frac{\beta}{\varepsilon^2}\big(F(x_0) - F^*\big)\right)$$

Smoothing parameter

Lemma I: (Optimality conditions) (Drusvyatskiy-P '16)
For any $\nu > 0$ and h_ν a smoothed h, then

$$\|\mathcal{G}(x)\| \leq \|\mathcal{G}^\nu(x)\| + \sqrt{\frac{\beta\nu}{2}}.$$

Smoothing parameter: $$\frac{\varepsilon}{2} = \sqrt{\frac{\beta\nu}{2}} \quad \Rightarrow \quad \boxed{\nu = \frac{\varepsilon^2}{2\beta}}$$

Thm: (Drusvyatskiy-P '16)
With $\nu \sim \frac{\varepsilon^2}{\beta}$, (inexact) prox-linear + smoothing + fast gradient subsolves $\Rightarrow$

$$\|\mathcal{G}(x)\| < \varepsilon \quad \text{after} \quad \tilde{O}\left(\frac{\beta \cdot \|\nabla c\|}{\varepsilon^3}\big(F(x_0) - F^*\big)\right)$$

Acceleration

Goal

Design a convexity adapting acceleration scheme.

Acceleration

Goal

Design a convexity adapting acceleration scheme.

Measuring non-convexity,

$$h \circ c(x) = \sup_w \{\langle w, c(x)\rangle - h^*(w)\}$$

Defn: Parameter $\rho \in [0,1]$ such that

$$x \mapsto \langle w, c(x)\rangle + \rho \cdot \frac{\beta}{2} \|x\|^2 \quad \text{is convex for all } w \in \operatorname{dom} h^*$$

Acceleration

Algorithm 1: Accelerated prox-linear method

Initialize : Fix two points $x_0, v_0 \in \operatorname{dom} g$.

while $\|\mathcal{G}(y_{t-1})\| > \varepsilon$ **do**

$a_t \leftarrow \frac{2}{t+1}$

$y_t \leftarrow a_t v_{t-1} + (1 - a_t) x_{t-1}$

$x_t \leftarrow y_t^+$

$v_t \leftarrow \operatorname{argmin}_z \; g(z) + \frac{1}{a_t} \cdot h\big(c(y_t) + a_t \nabla c(y_t)(z - v_{t-1})\big) + \frac{a_t}{2s} \|z - v_{t-1}\|^2$

$t \leftarrow t + 1$

end

Thm: (Drusvyatskiy-P '16)

$$\min_{i=1,\ldots,t} \|\mathcal{G}(x_i)\|^2 \leq \mathcal{O}\left(\frac{\beta^2}{t^3} \|x_0 - x^*\|^2\right) + \rho \cdot \mathcal{O}\left(\frac{\beta^2 R^2}{t}\right)$$

where $R = \operatorname{diam}(\operatorname{dom} g)$

- Generalizes (Ghadimi-Lan '16) for additive composite

Recent related work

- Local rates (Drusvyatskiy-Lewis '15):
  Error-bound property: $\alpha \cdot \text{dist}(x; \mathcal{S}) \leq \text{dist}(0; \partial F(x)), \forall x \in \text{nbhd}(\mathcal{S})$

$$F(x_{t+1}) - F^* \leq \left(1 - \left(\frac{\alpha}{\beta}\right)^2\right) F(x_t) - F^*$$

- Robust Phase Retrieval (Duchi-Ruan '17):

  Quadratic convergence w.h.p. on $\min_x \frac{1}{m} \sum_{i=1}^{m} |(a_i^T x)^2 - b_i|$

  (uses Eldar-Mendelson '12)
- Sampling methods (Duchi-Ruan '16): Almost sure conv. on

$$\min_x g(x) + \int_{\mathcal{S}} h(c(x, s), s)\, dP(s).$$

  Inner-outer subgradient methods w/ rates (Davis-Grimmer '17)
- Large-finite sums (P-Lin, Drusvyatskiy, Mairal, Harchaoui '17):

  "Accelerate" non-convex $\frac{1}{n} \sum_{i=1}^{n} h_i(x) + g(x)$

Nonconvex catalyst: design a generic scheme

Our problem:

$$\min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x)$$

- $f_i : \mathbb{R}^p \to \mathbb{R}$ is smooth, $\boldsymbol{L}$-Lipschitz continuous gradient
- $\psi : \mathbb{R}^p \to \overline{\mathbb{R}}$ may be nonsmooth
- f is $\boldsymbol{\rho}$-weakly convex

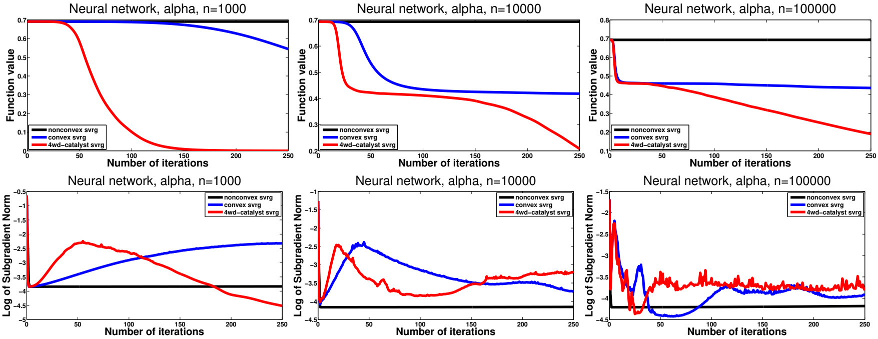
Examples

1. Robust penalties (e.g. MCP, SCAD)
2. Neural networks

Experiments: neural networks

Given data $\{(a_i, b_i)\}_{i=1}^n$

$$\min_{W_1 \in \mathbb{R}^{n \times d}, W_2 \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-b_i\left(W_2^T \sigma(W_1^T a_i)\right)\right)\right)$$



References

Drusvyatskiy, D. and Paquette, C. (2016).
Efficiency of minimizing compositions of convex functions and smooth maps.
Preprint arXiv: 1605.00125.

Paquette, C., Lin, H., Drusvyatskiy, D., Mairal, J., and Harchaoui, Z. (2017).
Catalyst for gradient-based nonconvex optimization.
Preprint arXiv: 1703.10993.