Efficiency of inner-outer algorithms at scale

Dmitriy Drusvyatskiy
Mathematics, University of Washington

Joint work with
Aravkin (UW), Harchaoui (UW), Lewis (Cornell),
Lin (Inria), Mairal (Inria), Paquette (UW), and van Leeuwen (Utrecht)

Optimization and Statistical Learning 2017

AFOSR YIP

Outline

1. Partial minimization

2. Proximal point acceleration (Catalyst)

3. Complexity of minimizing $h \circ c$

Outline

1. Partial minimization

$$\boxed{\min_{x,y} \varphi(x,y) = \min_{x} \left\{ F(x) := \min_{y} \varphi(x,y) \right\}}$$

2. Proximal point acceleration (Catalyst)

$$\boxed{x_{t+1} = \operatorname*{argmin}_{x} \; f(x) + \tfrac{\kappa}{2}\|x - x_t\|^2}$$

3. Complexity of minimizing $h \circ c$

$$\boxed{x_{t+1} = \operatorname*{argmin}_{x} \; h\Big(c(x_t) + \nabla c(x_t)(x - x_t)\Big) + \tfrac{\kappa}{2}\|x - x_t\|^2}$$

Structured partial minimization

Structured partial minimization

Problem class:

$$\min_{y,x} \; f(x) + g(y) \quad \text{subject to} \quad A(y)x = q.$$

- f and g are convex and simple.
- $A(\cdot)$ smooth and $A(y)$ invertible $\forall y$.

Structured partial minimization

Problem class:

$$\min_{y,x} \; f(x) + g(y) \quad \text{subject to} \quad A(y)x = q.$$

- f and g are convex and simple.
- $A(\cdot)$ smooth and $A(y)$ invertible $\forall y$.

Applications: Kalman filtering, PDE-constrained optimization, boundary control, optimal transport, etc.

Structured partial minimization

Problem class:

$$\min_{y,x} \; f(x) + \tfrac{\lambda}{2}\|A(y)x - q\|^2 + g(y).$$

- f and g are convex and simple.
- $A(\cdot)$ smooth and $A(y)$ invertible $\forall y$.
- $\lambda > 0$ is a relaxation parameter.

Applications: Kalman filtering, PDE-constrained optimization, boundary control, optimal transport, etc.

Structured partial minimization

Problem class:

$$\min_{y,x} \; f(x) + \tfrac{\lambda}{2}\|A(y)x - q\|^2 + g(y).$$

- f and g are convex and simple.
- $A(\cdot)$ smooth and $A(y)$ invertible $\forall y$.
- $\lambda > 0$ is a relaxation parameter.

Applications: Kalman filtering, PDE-constrained optimization, boundary control, optimal transport, etc.

Difficulty: Relaxation $\lambda \uparrow \infty \implies$ poor conditioning!

Structured partial minimization

Problem class:

$$\boxed{\min_{y,x} \; f(x) + \tfrac{\lambda}{2}\|A(y)x - q\|^2 + g(y).}$$

- f and g are convex and simple.
- $A(\cdot)$ smooth and $A(y)$ invertible $\forall y$.
- $\lambda > 0$ is a relaxation parameter.

Applications: Kalman filtering, PDE-constrained optimization, boundary control, optimal transport, etc.

Difficulty: Relaxation $\lambda \uparrow \infty \implies$ poor conditioning!

Partial minimization:

$$\min_{y} \; \widetilde{f}(y) + g(y)$$

where

$$\widetilde{f}(y) = \min_{x} \; f(x) + \tfrac{\lambda}{2}\|A(y)x - q\|^2.$$

Structured partial minimization

Problem class:

$$\boxed{\min_{y,x} \; f(x) + \tfrac{\lambda}{2}\|A(y)x - q\|^2 + g(y).}$$

- f and g are convex and simple.
- $A(\cdot)$ smooth and $A(y)$ invertible $\forall y$.
- $\lambda > 0$ is a relaxation parameter.

Applications: Kalman filtering, PDE-constrained optimization, boundary control, optimal transport, etc.

Difficulty: Relaxation $\lambda \uparrow \infty \implies$ poor conditioning!

Partial minimization:

$$\min_{y} \; \widetilde{f}(y) + g(y)$$

where

$$\widetilde{f}(y) = \min_{x} \; f(x) + \tfrac{\lambda}{2}\|A(y)x - q\|^2.$$

Thm: (Aravkin-van Leeuwen-D '16)
$\mathrm{lip}(\nabla \widetilde{f})$ is independent of λ

Structured partial minimization

Problem class:

$$\boxed{\min_{y,x} \; f(x) + \tfrac{\lambda}{2}\|A(y)x - q\|^2 + g(y).}$$

- f and g are convex and simple.
- $A(\cdot)$ smooth and $A(y)$ invertible $\forall y$.
- $\lambda > 0$ is a relaxation parameter.

Applications: Kalman filtering, PDE-constrained optimization, boundary control, optimal transport, etc.

Difficulty: Relaxation $\lambda \uparrow \infty \implies$ poor conditioning!

Partial minimization:

$$\min_{y} \; \widetilde{f}(y) + g(y)$$

where

$$\widetilde{f}(y) = \min_{x} \; f(x) + \tfrac{\lambda}{2}\|A(y)x - q\|^2.$$

Thm: (Aravkin-van Leeuwen-D '16)
$\mathrm{lip}(\nabla \widetilde{f})$ is independent of $\lambda \implies$ inexact prox-gradient $\sim \log(\lambda)$.

Example: optimal transport

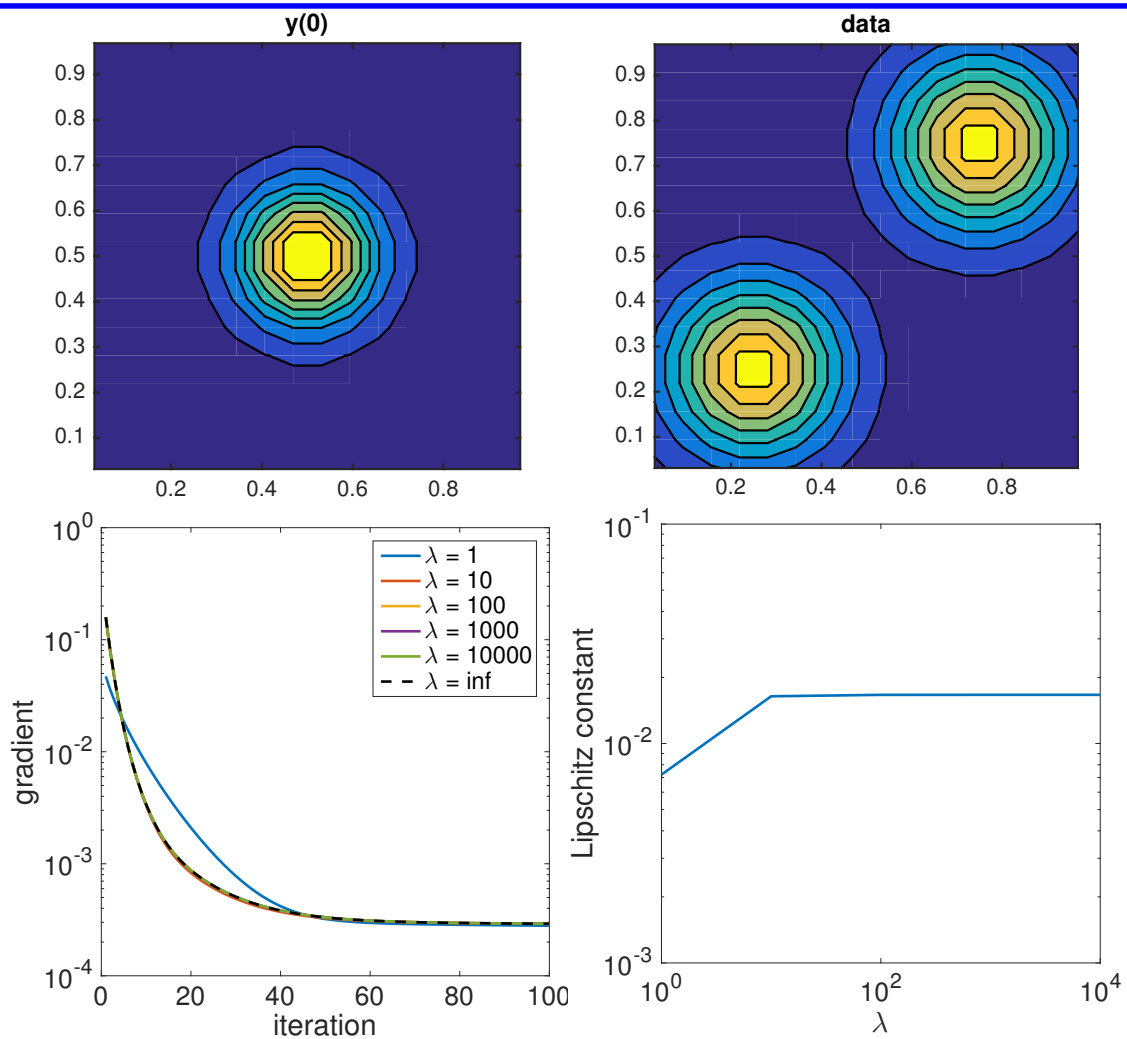

Figure: Initial density, target density, convergence rate, $\mathrm{lip}(\nabla \tilde{f})$.

Accelerated proximal-point method (Catalyst)

Proximal-point algorithm

Suppose f is μ-strongly convex.

Proximal-point algorithm

Suppose f is μ-strongly convex.

Proximal-point method (Martinet '70,72, Rockafellar '76):

$$x_{t+1} = \operatorname*{argmin}_x \; f(x) + \tfrac{\kappa}{2}\|x - x_t\|^2$$

Proximal-point algorithm

Suppose f is μ-strongly convex.

Proximal-point method (Martinet '70,72, Rockafellar '76):

$$x_{t+1} = \operatorname*{argmin}_x \; f(x) + \tfrac{\kappa}{2}\|x - x_t\|^2$$

Complexity: $f(x_t) - f^* \sim \min\left\{\frac{\kappa}{t}, \left(1 - \frac{\mu}{\kappa}\right)^t\right\}$

Proximal-point algorithm

Suppose f is μ-strongly convex.

Proximal-point method (Martinet '70,72, Rockafellar '76):

$$x_{t+1} = \operatorname*{argmin}_x \; f(x) + \tfrac{\kappa}{2}\|x - x_t\|^2$$

Complexity: $f(x_t) - f^* \sim \min\left\{\frac{\kappa}{t}, (1 - \frac{\mu}{\kappa})^t\right\}$

Accelerated proximal-point method:
(Nesterov '83, Güler '92, Beck-Teboulle '09)

$$\left\{\begin{array}{c} x_t = \operatorname*{argmin}_x \; \left\{f(x) + \frac{\kappa}{2}\|x - y_{t-1}\|^2\right\} \\ \text{Solve } \alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2 + \dfrac{\mu}{\mu+\kappa}\alpha_t \\ y_t = x_t + \beta_k(x_t - x_{t-1}) \\ \beta_t = \dfrac{\alpha_{t-1}(1-\alpha_{t-1})}{\alpha_{t-1}^2 + \alpha_t} \end{array}\right\}$$

Proximal-point algorithm

Suppose f is μ-strongly convex.

Proximal-point method (Martinet '70,72, Rockafellar '76):

$$x_{t+1} = \operatorname*{argmin}_x \; f(x) + \tfrac{\kappa}{2}\|x - x_t\|^2$$

Complexity: $f(x_t) - f^* \sim \min\left\{\frac{\kappa}{t}, (1 - \frac{\mu}{\kappa})^t\right\}$

Accelerated proximal-point method:
(Nesterov '83, Güler '92, Beck-Teboulle '09)

$$\left\{\begin{array}{c} x_t = \operatorname*{argmin}_x \; \left\{f(x) + \frac{\kappa}{2}\|x - y_{t-1}\|^2\right\} \\ \text{Solve } \alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2 + \dfrac{\mu}{\mu+\kappa}\alpha_t \\ y_t = x_t + \beta_k(x_t - x_{t-1}) \\ \beta_t = \dfrac{\alpha_{t-1}(1-\alpha_{t-1})}{\alpha_{t-1}^2 + \alpha_t} \end{array}\right\}$$

Complexity: $f(x_t) - f^* \sim \min\left\{\frac{\kappa}{t^2}, \left(1 - \sqrt{\frac{\mu}{\kappa}}\right)^t\right\}$

Renewed interest from finite sums

Renewed interest from finite sums

Finite-sums:

$$\min_{x \in \mathbf{R}^d} \ \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x).$$

- f_i convex, ∇f_i is β-Lipschitz, and sum is μ-strongly convex.

Renewed interest from finite sums

Finite-sums:

$$\min_{x \in \mathbf{R}^d} \ \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x).$$

- f_i convex, ∇f_i is β-Lipschitz, and sum is μ-strongly convex.

	Progress	Efficiency

Renewed interest from finite sums

Finite-sums:

$$\min_{x \in \mathbf{R}^d} \ \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x).$$

- f_i convex, ∇f_i is β-Lipschitz, and sum is μ-strongly convex.

	Progress	Efficiency
Incremental methods	$\mathbb{E}[f(x_t)] - f^* \leq \epsilon$	$\left(n + \frac{\beta}{\mu}\right) \cdot \ln \frac{1}{\epsilon}$

SAG (Schmidt, Le Roux-Bach '13), SVRG (Johnson, Zhang '13), SAGA (Defazio, Bach, Lacoste-Julien '13), etc.

Renewed interest from finite sums

Finite-sums:

$$\min_{x \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x).$$

- f_i convex, ∇f_i is β-Lipschitz, and sum is μ-strongly convex.

	Progress	Efficiency
Incremental methods	$\mathbb{E}[f(x_t)] - f^* \leq \epsilon$	$\left(n + \frac{\beta}{\mu}\right) \cdot \ln \frac{1}{\epsilon}$

SAG (Schmidt, Le Roux-Bach '13), SVRG (Johnson, Zhang '13), SAGA (Defazio, Bach, Lacoste-Julien '13), etc.

Accelerated incremental methods?

Answer: Catalyst (Lin, Mairal, Harchaoui '15),
(aka inexact accelerated proximal-point algorithm!)

Renewed interest from finite sums

Finite-sums:

$$\min_{x \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x).$$

- f_i convex, ∇f_i is β-Lipschitz, and sum is μ-strongly convex.

	Progress	Efficiency
Incremental methods	$\mathbb{E}[f(x_t)] - f^* \leq \epsilon$	$\left(n + \frac{\beta}{\mu}\right) \cdot \ln \frac{1}{\epsilon}$

SAG (Schmidt, Le Roux-Bach '13), SVRG (Johnson, Zhang '13), SAGA (Defazio, Bach, Lacoste-Julien '13), etc.

Accelerated incremental methods?

Answer: Catalyst (Lin, Mairal, Harchaoui '15),
(aka inexact accelerated proximal-point algorithm!)

Other PPA approaches: (Frostig et al. '15),
(Shalev-Schwartz, Zhang '14)

Renewed interest from finite sums

Finite-sums:

$$\min_{x \in \mathbf{R}^d} \ \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x).$$

- f_i convex, ∇f_i is β-Lipschitz, and sum is μ-strongly convex.

	Progress	Efficiency
Incremental methods	$\mathbb{E}[f(x_t)] - f^* \leq \epsilon$	$\left(n + \frac{\beta}{\mu}\right) \cdot \ln \frac{1}{\epsilon}$

SAG (Schmidt, Le Roux-Bach '13), SVRG (Johnson, Zhang '13), SAGA (Defazio, Bach, Lacoste-Julien '13), etc.

Accelerated incremental methods?

Answer: Catalyst (Lin, Mairal, Harchaoui '15),
(aka inexact accelerated proximal-point algorithm!)

Other PPA approaches: (Frostig et al. '15), (Shalev-Schwartz, Zhang '14)

Direct approaches: (Lan, Zhou '15), (Allen-Zhu '16)

Non-convex catalyst?

Can catalyst "work" for nonconvex problems?

Non-convex catalyst?

Can catalyst "work" for nonconvex problems?

Baseline: For nonconvex f, prox-points x_t satisfy

$$\operatorname{dist}(0; \partial f(x_t)) \leq \|\kappa(x_t - x_{t-1})\| \sim \frac{\kappa}{\sqrt{t}}.$$

Non-convex catalyst?

Can catalyst "work" for nonconvex problems?

Baseline: For nonconvex f, prox-points x_t satisfy

$$\operatorname{dist}(0; \partial f(x_t)) \leq \|\kappa(x_t - x_{t-1})\| \sim \frac{\kappa}{\sqrt{t}}.$$

Modest goal: Catalyst for $\min_x f(x)$ with

- f non-convex $\Rightarrow$ rate no worse than PPM
- f convex $\Rightarrow$ accelerated rate

Non-convex catalyst?

Can catalyst "work" for nonconvex problems?

Baseline: For nonconvex f, prox-points x_t satisfy

$$\operatorname{dist}(0; \partial f(x_t)) \leq \|\kappa(x_t - x_{t-1})\| \sim \frac{\kappa}{\sqrt{t}}.$$

Modest goal: Catalyst for $\min_x f(x)$ with

- f non-convex $\Rightarrow$ rate no worse than PPM
- f convex $\Rightarrow$ accelerated rate

Problem class: f is ρ-weakly convex if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex.

Non-convex catalyst?

Can catalyst "work" for nonconvex problems?

Baseline: For nonconvex f, prox-points x_t satisfy

$$\operatorname{dist}(0; \partial f(x_t)) \leq \|\kappa(x_t - x_{t-1})\| \sim \tfrac{\kappa}{\sqrt{t}}.$$

Modest goal: Catalyst for $\min_x f(x)$ with

- f non-convex $\Rightarrow$ rate no worse than PPM
- f convex $\Rightarrow$ accelerated rate

Problem class: f is ρ-weakly convex if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex.

- C^2-smooth f is ρ-weakly convex $\iff$ $\nabla^2 f \succeq -\rho I$.

Non-convex catalyst?

Can catalyst "work" for nonconvex problems?

Baseline: For nonconvex f, prox-points x_t satisfy

$$\operatorname{dist}(0; \partial f(x_t)) \leq \|\kappa(x_t - x_{t-1})\| \sim \tfrac{\kappa}{\sqrt{t}}.$$

Modest goal: Catalyst for $\min_x f(x)$ with

- f non-convex $\Rightarrow$ rate no worse than PPM
- f convex $\Rightarrow$ accelerated rate

Problem class: f is ρ-weakly convex if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex.

- C^2-smooth f is ρ-weakly convex $\iff$ $\nabla^2 f \succeq -\rho I$.

Nonconvex catalyst (Paquette, Lin, D., Mairal, Harchaoui '17):
"Generic acceleration schema beyond convexity" (poster session)

Non-convex catalyst?

Can catalyst "work" for nonconvex problems?

Baseline: For nonconvex f, prox-points x_t satisfy

$$\operatorname{dist}(0; \partial f(x_t)) \leq \|\kappa(x_t - x_{t-1})\| \sim \tfrac{\kappa}{\sqrt{t}}.$$

Modest goal: Catalyst for $\min_x f(x)$ with

- f non-convex $\Rightarrow$ rate no worse than PPM
- f convex $\Rightarrow$ accelerated rate

Problem class: f is ρ-weakly convex if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex.

- C^2-smooth f is ρ-weakly convex $\iff$ $\nabla^2 f \succeq -\rho I$.

Nonconvex catalyst (Paquette, Lin, D., Mairal, Harchaoui '17): "Generic acceleration schema beyond convexity" (poster session)
Algorithmic idea: Interlace prox-point steps and accelerated steps (Ghadimi-Lan-Zhang '15).

Illustration of 4WD-Catalyst

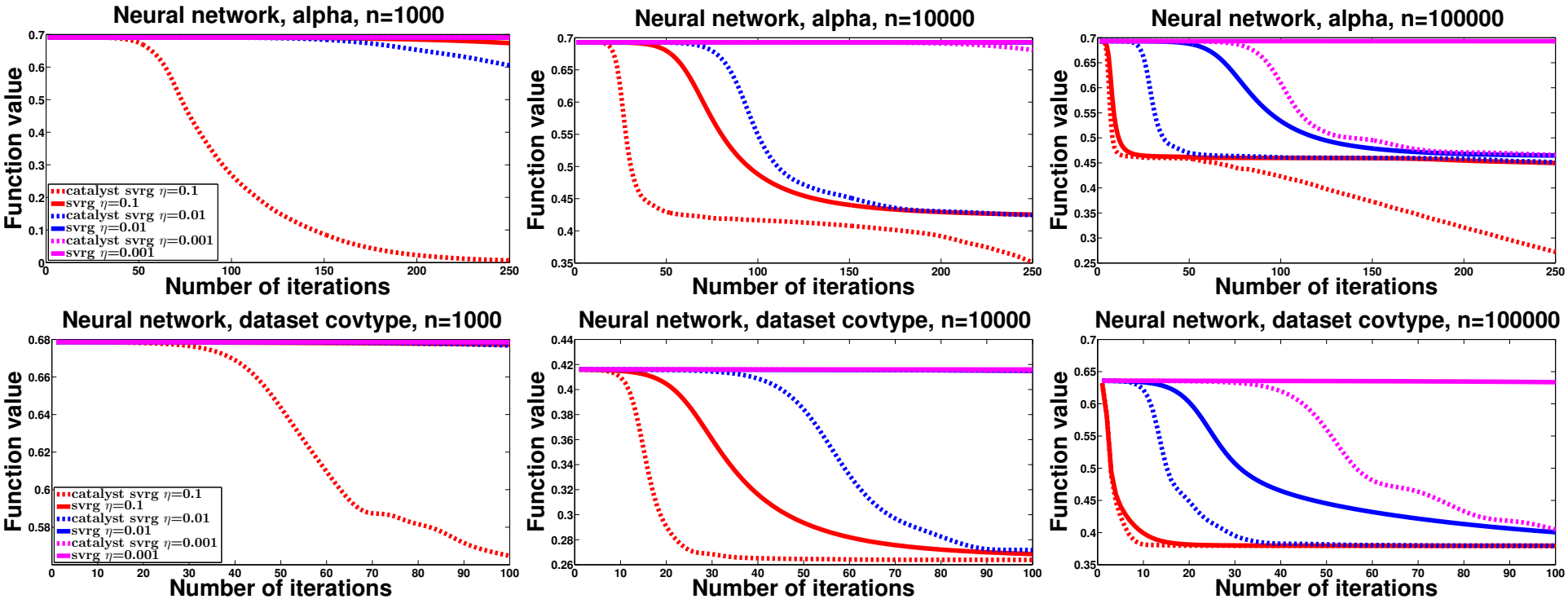

Figure: Two-layer neural network on subsets of two datasets alpha (top) and covtype (bottom).

Complexity for minimizing $h \circ c$

Nonsmooth & Nonconvex minimization

Convex composition

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

where

- $g\colon \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is closed, convex.
- $h\colon \mathbf{R}^m \to \mathbf{R}$ is convex and L-Lipschitz.
- $c\colon \mathbf{R}^d \to \mathbf{R}^m$ is C^1-smooth and ∇c is β-Lipschitz.

Nonsmooth & Nonconvex minimization

Convex composition

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

where

- $g\colon \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is closed, convex.
- $h\colon \mathbf{R}^m \to \mathbf{R}$ is convex and L-Lipschitz.
- $c\colon \mathbf{R}^d \to \mathbf{R}^m$ is C^1-smooth and ∇c is β-Lipschitz.

For convenience, set $\mu = L\beta$.

Nonsmooth & Nonconvex minimization

Convex composition

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

where

- $g\colon \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is closed, convex.
- $h\colon \mathbf{R}^m \to \mathbf{R}$ is convex and L-Lipschitz.
- $c\colon \mathbf{R}^d \to \mathbf{R}^m$ is C^1-smooth and ∇c is β-Lipschitz.

For convenience, set $\mu = L\beta$.

Main examples:

- Additive composite minimization:

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

- Nonlinear least squares:

$$\min_x \; \{\|c(x)\| : l_i \le x_i \le u_i \qquad \text{for } i = 1, \ldots, m\}$$

- Exact penalty subproblem:

$$\min_x \; g(x) + \operatorname{dist}_K(c(x))$$

(Burke '85,'91, Fletcher '82, Powell '84, Wright '90, Yuan '83)

Roadmap

What is the complexity of this problem class?

Roadmap:

- Idealized method: "prox-linear algorithm",
- smoothing,
- fast-gradient subproblem solves.

Prox-linear algorithm

Prox-linear mapping

$$x^+ = \operatorname*{argmin}_y \; g(y) + h\Big(c(x) + \nabla c(x)(y - x)\Big) + \frac{\mu}{2}\|y - x\|^2$$

Prox-linear algorithm

Prox-linear mapping

$$x^+ = \operatorname*{argmin}_y \; g(y) + h\Big(c(x) + \nabla c(x)(y - x)\Big) + \frac{\mu}{2}\|y - x\|^2$$

Prox-linear method (Burke, Fletcher, Osborne, Powell,. . . '80s):

$$x_{t+1} = x_t^+.$$

Prox-linear algorithm

Prox-linear mapping

$$x^+ = \operatorname*{argmin}_y \; g(y) + h\Big(c(x) + \nabla c(x)(y-x)\Big) + \frac{\mu}{2}\|y-x\|^2$$

Prox-linear method (Burke, Fletcher, Osborne, Powell,. . . '80s):

$$x_{t+1} = x_t^+.$$

Stationarity measure:

The prox-gradient $\quad \mathcal{G}(x) := \mu(x - x^+).$

Prox-linear algorithm

Prox-linear mapping

$$x^+ = \operatorname*{argmin}_y \; g(y) + h\Big(c(x) + \nabla c(x)(y-x)\Big) + \frac{\mu}{2}\|y-x\|^2$$

Prox-linear method (Burke, Fletcher, Osborne, Powell,. . . '80s):

$$x_{t+1} = x_t^+.$$

Stationarity measure:

The prox-gradient $\quad \mathcal{G}(x) := \mu(x - x^+).$

Convergence rate:

$$\|\mathcal{G}(x_t)\| < \epsilon \quad \text{after} \quad \mathcal{O}\left(\frac{\mu}{\epsilon^2}\right) \quad \text{iterations}$$

Prox-linear algorithm

Prox-linear mapping

$$x^+ = \operatorname*{argmin}_y \; g(y) + h\Big(c(x) + \nabla c(x)(y-x)\Big) + \frac{\mu}{2}\|y-x\|^2$$

Prox-linear method (Burke, Fletcher, Osborne, Powell,. . . '80s):

$$x_{t+1} = x_t^+.$$

Stationarity measure:

The prox-gradient $\quad \mathcal{G}(x) := \mu(x - x^+).$

Convergence rate:

$$\|\mathcal{G}(x_t)\| < \epsilon \quad \text{after} \quad \mathcal{O}\left(\frac{\mu}{\epsilon^2}\right) \quad \text{iterations}$$

Why is "$\|\mathcal{G}(x_t)\| < \epsilon$" meaningful?

Prox-linear algorithm

Prox-linear mapping

$$x^+ = \operatorname*{argmin}_y \; g(y) + h\Big(c(x) + \nabla c(x)(y-x)\Big) + \frac{\mu}{2}\|y-x\|^2$$

Prox-linear method (Burke, Fletcher, Osborne, Powell,. . . '80s):

$$x_{t+1} = x_t^+.$$

Stationarity measure:

$$\text{The prox-gradient} \quad \mathcal{G}(x) := \mu(x - x^+).$$

Convergence rate:

$$\|\mathcal{G}(x_t)\| < \epsilon \quad \text{after} \quad \mathcal{O}\left(\frac{\mu}{\epsilon^2}\right) \quad \text{iterations}$$

Why is "$\|\mathcal{G}(x_t)\| < \epsilon$" meaningful?

Lemma:

$$4 \cdot \|\mathcal{G}(x)\| \geq \|2\mu(x - \operatorname{prox}_{F/2\mu}(x))\|.$$

Smoothing

Lemma: For $\nu > 0$, define

$$F_\nu := g + h_\nu \circ c.$$

with the Moreau envelope

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

Smoothing

Lemma: For $\nu > 0$, define

$$F_\nu := g + h_\nu \circ c.$$

with the Moreau envelope

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

Then

$$\|\mathcal{G}(x)\| \leq \|\mathcal{G}^\nu(x)\| + \sqrt{L^3\beta\nu}.$$

Smoothing

Lemma: For $\nu > 0$, define

$$F_\nu := g + h_\nu \circ c.$$

with the Moreau envelope

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

Then

$$\|\mathcal{G}(x)\| \leq \|\mathcal{G}^\nu(x)\| + \sqrt{L^3\beta\nu}.$$

Thm: (D-Paquette '16)
Smoothing $\nu \sim \frac{\epsilon^2}{L^3\beta}$, prox-linear, fast-grad subsolves $\Rightarrow$

$$\|\mathcal{G}(x)\| < \epsilon \quad \text{after} \quad \widetilde{\mathcal{O}}\left(\frac{L^2\beta \cdot \mathrm{lip}(c)}{\epsilon^3}\right) \quad \text{prox-grad steps.}$$

Extension to finite sums

Finite sum:

$$\min_x \ \frac{1}{n}\sum_{i=1}^{n} h_i(c_i(x)) + g(x)$$

Thm:(D-Paquette '16)
Smoothing, prox-linear, fast-incremental subsolves $\Rightarrow$

$$\mathbb{E}\|\mathcal{G}(x)\| < \epsilon \quad \text{after} \quad \widetilde{\mathcal{O}}\left(\frac{L\beta n^{1/2}}{\epsilon^2} + \frac{L^2\beta\cdot\Omega}{n^{1/2}\epsilon^3}\right)$$

evaluation of $\nabla c_i/\text{prox}_h/\text{prox}_g$, with

$$\Omega = \max_{i,\ldots,m} \sup_{x\in\text{dom}\, g} \|\nabla c_i\|.$$

Alternate approach: (Duchi-Ruan '17)
"Composite optimization at Scale: Structures and Solvability of some non-smooth non-convex problems" (poster session)

Conclusions

Inner-outer algorithms $\implies$ improved complexity

Three examples:

- Structured variable projection
- Proximal point acceleration
- Complexity of minimizing $h \circ c$

Thank you!

References

• "Quadratic penalization through the variable projection technique", Aravkin-van Leeuwen-D, 2016, arXiv:1606.02395.

• "Catalyst acceleration for gradient-based non-convex optimization", Paquette-Lin-D-Mairal-Harchaoui, 2017, arXiv:1703.10993.

• "Efficiency of minimizing compositions of convex functions and smooth maps", D-Paquette, 2016, arXiv:1605.00125.