Expanding the reach of optimal methods

Dmitriy Drusvyatskiy
Mathematics, University of Washington

Joint work with
C. Kempton (UW), M. Fazel (UW), A.S. Lewis (Cornell), and S. Roy (UW)

BURKAPALOOZA!

WCOM SPRING 2016

AFOSR YIP

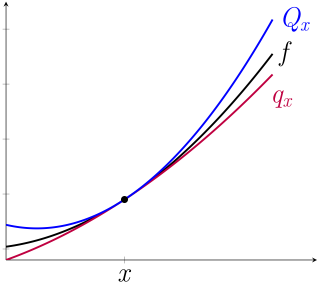
Notation

Function $f\colon \mathbf{R}^n \to \mathbf{R}$ is α-convex and β-smooth if

$$q_x \le f \le Q_x$$

where

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



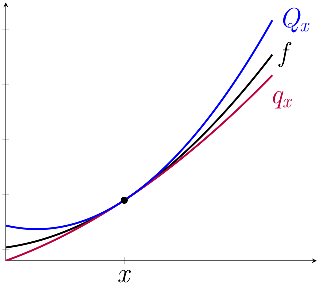
Notation

Function $f\colon \mathbf{R}^n \to \mathbf{R}$ is α-convex and β-smooth if

$$q_x \le f \le Q_x$$

where

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



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

Complexity of first-order methods

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

Complexity of first-order methods

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

Majorization view: $x_{k+1} = \text{argmin } Q_{x_k}(\cdot)$

Complexity of first-order methods

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

Majorization view: $x_{k+1} = \text{argmin } Q_{x_k}(\cdot)$

	β-smooth	β-smooth & α-convex
Gradient Descent	$\frac{\beta}{\epsilon}$	$\kappa \cdot \ln \frac{1}{\epsilon}$

Table: Iterations until $f(x_k) - f^* < \epsilon$

Complexity of first-order methods

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

Majorization view: $x_{k+1} = \text{argmin } Q_{x_k}(\cdot)$

	β-smooth	β-smooth & α-convex
Gradient Descent	$\frac{\beta}{\epsilon}$	$\kappa \cdot \ln \frac{1}{\epsilon}$
Optimal Methods	$\sqrt{\frac{\beta}{\epsilon}}$	$\sqrt{\kappa} \cdot \ln \frac{1}{\epsilon}$

Table: Iterations until $f(x_k) - f^* < \epsilon$

Complexity of first-order methods

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

Majorization view: $x_{k+1} = \text{argmin } Q_{x_k}(\cdot)$

	β-smooth	β-smooth & α-convex
Gradient Descent	$\frac{\beta}{\epsilon}$	$\kappa \cdot \ln \frac{1}{\epsilon}$
Optimal Methods	$\sqrt{\frac{\beta}{\epsilon}}$	$\sqrt{\kappa} \cdot \ln \frac{1}{\epsilon}$

Table: Iterations until $f(x_k) - f^* < \epsilon$

(Nesterov '83, Yudin-Nemirovsky '83)

Complexity of first-order methods

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

Majorization view: $x_{k+1} = \text{argmin } Q_{x_k}(\cdot)$

	β-smooth	β-smooth & α-convex
Gradient Descent	$\frac{\beta}{\epsilon}$	$\kappa \cdot \ln \frac{1}{\epsilon}$
Optimal Methods	$\sqrt{\frac{\beta}{\epsilon}}$	$\sqrt{\kappa} \cdot \ln \frac{1}{\epsilon}$

Table: Iterations until $f(x_k) - f^* < \epsilon$

(Nesterov '83, Yudin-Nemirovsky '83)

Optimal methods have downsides:

- Not intuitive
- Non-monotone
- Difficult to augment with "memory"

Optimal method by optimal averaging

Idea: Maximize lower models of f.

Optimal method by optimal averaging

Idea: Maximize lower models of f.

Notation:

$$x^+ = x - \frac{1}{\beta}\nabla f(x) \qquad \text{and} \qquad x^{++} = x - \frac{1}{\alpha}\nabla f(x)$$

Optimal method by optimal averaging

Idea: Maximize lower models of f.

Notation:

$$x^+ = x - \frac{1}{\beta}\nabla f(x) \qquad \text{and} \qquad x^{++} = x - \frac{1}{\alpha}\nabla f(x)$$

Convexity bound $f \geq q_x$ in canonical form:

$$f(y) \geq \left(f(x) - \frac{|\nabla f(x)|^2}{2\alpha} \right) + \frac{\alpha}{2}|y - x^{++}|^2$$

Optimal method by optimal averaging

Idea: Maximize lower models of f.

Notation:

$$x^+ = x - \frac{1}{\beta}\nabla f(x) \qquad \text{and} \qquad x^{++} = x - \frac{1}{\alpha}\nabla f(x)$$

Convexity bound $f \geq q_x$ in canonical form:

$$f(y) \geq \left(f(x) - \frac{|\nabla f(x)|^2}{2\alpha}\right) + \frac{\alpha}{2}|y - x^{++}|^2$$

Lower models:

$$Q_A(x) = v_A + \frac{\alpha}{2}|x - x_A|^2 \qquad Q_B(x) = v_B + \frac{\alpha}{2}|x - x_B|^2$$

Optimal method by optimal averaging

Idea: Maximize lower models of f.

Notation:

$$x^{+} = x - \frac{1}{\beta}\nabla f(x) \qquad \text{and} \qquad x^{++} = x - \frac{1}{\alpha}\nabla f(x)$$

Convexity bound $f \geq q_x$ in canonical form:

$$f(y) \geq \left(f(x) - \frac{|\nabla f(x)|^2}{2\alpha} \right) + \frac{\alpha}{2}|y - x^{++}|^2$$

Lower models:

$$Q_A(x) = v_A + \frac{\alpha}{2}|x - x_A|^2 \qquad Q_B(x) = v_B + \frac{\alpha}{2}|x - x_B|^2$$

$\Longrightarrow$ for any $\lambda \in [0,1]$ new lower-model

$$Q_\lambda := \lambda Q_A + (1-\lambda) Q_B = v_\lambda + \frac{\alpha}{2}|\cdot - x_\lambda|^2$$

Optimal method by optimal averaging

Idea: Maximize lower models of f.

Notation:

$$x^{+} = x - \frac{1}{\beta}\nabla f(x) \qquad \text{and} \qquad x^{++} = x - \frac{1}{\alpha}\nabla f(x)$$

Convexity bound $f \geq q_x$ in canonical form:

$$f(y) \geq \left(f(x) - \frac{|\nabla f(x)|^2}{2\alpha} \right) + \frac{\alpha}{2}|y - x^{++}|^2$$

Lower models:

$$Q_A(x) = v_A + \frac{\alpha}{2}|x - x_A|^2 \qquad Q_B(x) = v_B + \frac{\alpha}{2}|x - x_B|^2$$

$\Longrightarrow$ for any $\lambda \in [0,1]$ new lower-model

$$Q_\lambda := \lambda Q_A + (1-\lambda) Q_B = v_\lambda + \frac{\alpha}{2}|\cdot - x_\lambda|^2$$

Key observation: $v_\lambda \leq f^*$

Optimal method by optimal averaging

The minimum v_λ is maximized when

$$\bar{\lambda} = \mathrm{proj}_{[0,1]}\left(\frac{1}{2} + \frac{v_A - v_B}{\alpha|x_A - x_B|^2}\right).$$

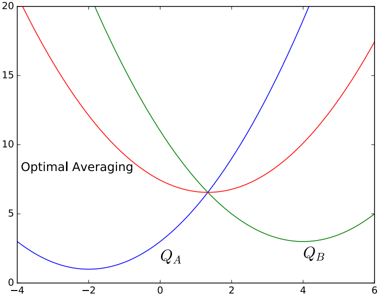
The quadratic $Q_{\bar{\lambda}}$ is the optimal averaging of (Q_A, Q_B).

Optimal method by optimal averaging

The minimum v_λ is maximized when

$$\bar{\lambda} = \mathrm{proj}_{[0,1]}\left(\frac{1}{2} + \frac{v_A - v_B}{\alpha|x_A - x_B|^2}\right).$$

The quadratic $Q_{\bar{\lambda}}$ is the optimal averaging of (Q_A, Q_B).

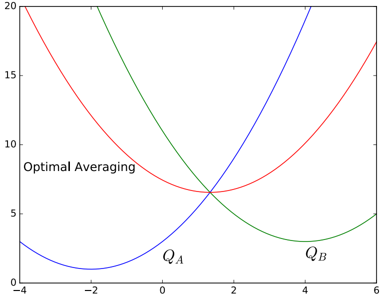


Optimal method by optimal averaging

The minimum v_λ is maximized when

$$\bar{\lambda} = \mathrm{proj}_{[0,1]}\left(\frac{1}{2} + \frac{v_A - v_B}{\alpha|x_A - x_B|^2}\right).$$

The quadratic $Q_{\bar{\lambda}}$ is the optimal averaging of (Q_A, Q_B).



Related to cutting plane, bundle methods, geometric descent (Bubeck-Lee-Singh '15)

Optimal method by optimal averaging

for $k = 1, \ldots, K$ **do**

  Set $Q(x) = \left(f(x_k) - \frac{|\nabla f(x_k)|^2}{2\alpha}\right) + \frac{\alpha}{2}\left|x - x_k^{++}\right|^2$

  Let $Q_k(x) = v_k + \frac{\alpha}{2}|x - c_k|^2$ be optim. average of (Q, Q_{k-1}).

  Set $x_{k+1} =$ `line_search` $\left(c_k, x_k^+\right)$

end

Algorithm: Optimal averaging

The line-search is due to (Bubeck-Lee-Singh '15)

Optimal method by optimal averaging

for $k = 1, \ldots, K$ **do**

Set $Q(x) = \left(f(x_k) - \frac{|\nabla f(x_k)|^2}{2\alpha}\right) + \frac{\alpha}{2}\left|x - x_k^{++}\right|^2$

Let $Q_k(x) = v_k + \frac{\alpha}{2}|x - c_k|^2$ be optim. average of (Q, Q_{k-1}).

Set $x_{k+1} = \texttt{line\_search}\left(c_k, x_k^+\right)$

end

Algorithm: Optimal averaging

The line-search is due to (Bubeck-Lee-Singh '15)

Optimal Linear Rate (D-Fazel-Roy '16):

$$f(x_k^+) - v_k \leq \epsilon \quad \text{after} \quad \mathcal{O}\left(\sqrt{\frac{\beta}{\alpha}} \cdot \ln \frac{1}{\epsilon}\right) \quad \text{iterations.}$$

Optimal method by optimal averaging

for $k = 1, \ldots, K$ **do**

  Set $Q(x) = \left(f(x_k) - \frac{|\nabla f(x_k)|^2}{2\alpha}\right) + \frac{\alpha}{2}\left|x - x_k^{++}\right|^2$

  Let $Q_k(x) = v_k + \frac{\alpha}{2}|x - c_k|^2$ be optim. average of (Q, Q_{k-1}).

  Set $x_{k+1} =$ `line_search` $\left(c_k, x_k^{+}\right)$

end

Algorithm: Optimal averaging

The line-search is due to (Bubeck-Lee-Singh '15)

Optimal Linear Rate (D-Fazel-Roy '16):

$$f(x_k^+) - v_k \leq \epsilon \quad \text{after} \quad \mathcal{O}\left(\sqrt{\frac{\beta}{\alpha}} \cdot \ln \frac{1}{\epsilon}\right) \quad \text{iterations.}$$

- Intuitive
- Monotone in $f(x_k^+)$ and in v_k.
- "Memory" by optimally averaging $(Q, Q_{k-1}, \ldots, Q_{k-t})$.

Optimal method by optimal averaging

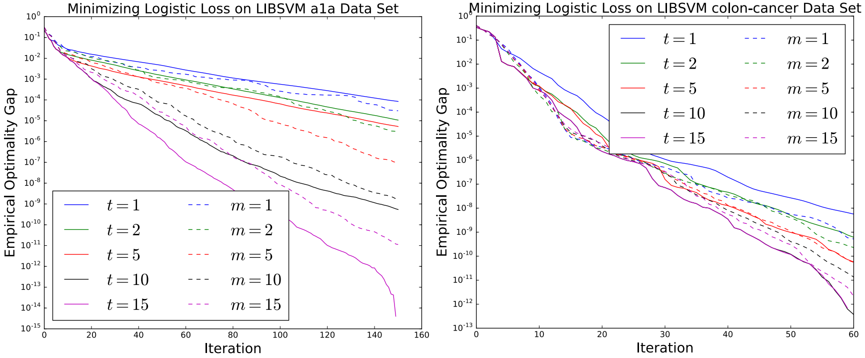


Figure: Logistic regression with regularization $\alpha = 10^{-4}$.

Nonsmooth & Nonconvex minimization

Convex composition

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

where

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

Nonsmooth & Nonconvex minimization

Convex composition

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

where

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

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

Convex composition

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

where

- $g \colon \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is closed, convex.
- $h \colon \mathbf{R}^m \to \mathbf{R}$ is convex and L-Lipschitz.
- $c \colon \mathbf{R}^n \to \mathbf{R}^m$ 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:

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

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

Prox-linear algorithm

Prox-linear mapping

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

and the prox-gradient

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

Prox-linear algorithm

Prox-linear mapping

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

and the prox-gradient

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

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

$$x_{k+1} = x_k^+.$$

Prox-linear algorithm

Prox-linear mapping

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

and the prox-gradient

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

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

$$x_{k+1} = x_k^+.$$

Eg: proximal gradient, Levenberg-Marquardt methods

Prox-linear algorithm

Prox-linear mapping

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

and the prox-gradient

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

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

$$x_{k+1} = x_k^{+}.$$

Eg: proximal gradient, Levenberg-Marquardt methods

Convergence rate:

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

Stopping criterion

$$\boxed{\text{What does } \|\mathcal{G}(x)\|^2 < \epsilon \text{ actually mean?}}$$

Stopping criterion

$$\boxed{\text{What does } \|\mathcal{G}(x)\|^2 < \epsilon \text{ actually mean?}}$$

Stationarity for target problem:

$$0 \in \partial g(x) + \nabla c(x)^* \partial h(c(x))$$

Stationarity for prox-subproblem:

$$\mathcal{G}(x) \in \partial g(x^+) + \nabla c(x)^* \partial h\big(c(x) + \nabla c(x)(x^+ - x)\big)$$

Stopping criterion

$$\boxed{\text{What does } \|\mathcal{G}(x)\|^2 < \epsilon \text{ actually mean?}}$$

Stationarity for target problem:

$$0 \in \partial g(x) + \nabla c(x)^* \partial h(c(x))$$

Stationarity for prox-subproblem:

$$\mathcal{G}(x) \in \partial g(x^+) + \nabla c(x)^* \partial h(c(x) + \nabla c(x)(x^+ - x))$$

Thm: (D-Lewis '16)
x^+ is nearly stationary because $\exists\ (\hat{x}, \hat{v})$ with

$$\hat{v} \in \partial g(\hat{x}) + \nabla c(\hat{x})^* \partial h(c(\hat{x}))$$

where

$$\|\hat{x} - x^+\| \leq \mu \|\mathcal{G}(x)\| \qquad \text{and} \qquad \|\hat{v}\| \leq 5 \|\mathcal{G}(x)\|$$

Stopping criterion

$$\boxed{\text{What does } \|\mathcal{G}(x)\|^2 < \epsilon \text{ actually mean?}}$$

Stationarity for target problem:

$$0 \in \partial g(x) + \nabla c(x)^* \partial h(c(x))$$

Stationarity for prox-subproblem:

$$\mathcal{G}(x) \in \partial g(x^+) + \nabla c(x)^* \partial h\big(c(x) + \nabla c(x)(x^+ - x)\big)$$

Thm: (D-Lewis '16)
x^+ is nearly stationary because $\exists\ (\hat{x}, \hat{v})$ with

$$\hat{v} \in \partial g(\hat{x}) + \nabla c(\hat{x})^* \partial h(c(\hat{x}))$$

where

$$\|\hat{x} - x^+\| \leq \mu \|\mathcal{G}(x)\| \qquad \text{and} \qquad \|\hat{v}\| \leq 5\|\mathcal{G}(x)\|$$

(*pf:* Ekeland's variational principle)

Local linear convergence

Error bound property (Luo-Tseng '93)

$$\boxed{\operatorname{dist}(x, \{\text{soln. set}\}) \le \frac{1}{\alpha}\|\mathcal{G}(x)\| \qquad \text{for } x \text{ near } \bar{x}}$$

Local linear convergence

Error bound property (Luo-Tseng '93)

$$\boxed{\operatorname{dist}(x, \{\text{soln. set}\}) \le \frac{1}{\alpha}\|\mathcal{G}(x)\| \qquad \text{for } x \text{ near } \bar{x}}$$

$\Longrightarrow$ local linear convergence

$$F(x_{k+1}) - F^* \le \left(1 - \frac{\alpha^2}{\mu^2}\right)(F(x_k) - F^*)$$

Local linear convergence

Error bound property (Luo-Tseng '93)

$$\boxed{\operatorname{dist}(x, \{\text{soln. set}\}) \le \frac{1}{\alpha}\|\mathcal{G}(x)\| \qquad \text{for } x \text{ near } \bar{x}}$$

$\Longrightarrow$ local linear convergence

$$F(x_{k+1}) - F^* \le \left(1 - \frac{\alpha^2}{\mu^2}\right)(F(x_k) - F^*)$$

The following are "essentially" equivalent (D-Lewis '16):

- **EB property**
- **Subregularity**:

$$\operatorname{dist}(x; \{\text{soln. set}\}) \le \frac{1}{\alpha} \cdot \operatorname{dist}(0; \partial F(x)) \qquad \text{for } x \text{ near } \bar{x}$$

- **Quadratic growth**:

$$F(x) \ge F(\bar{x}) + \frac{\alpha}{2} \cdot \operatorname{dist}^2(x, \{\text{soln. set}\}) \qquad \text{for } x \text{ near } \bar{x}$$

Local linear convergence

Error bound property (Luo-Tseng '93)

$$\boxed{\operatorname{dist}(x, \{\text{soln. set}\}) \leq \frac{1}{\alpha}\|\mathcal{G}(x)\| \qquad \text{for } x \text{ near } \bar{x}}$$

$\Longrightarrow$ local linear convergence

$$F(x_{k+1}) - F^* \leq \left(1 - \frac{\alpha^2}{\mu^2}\right)(F(x_k) - F^*)$$

The following are "essentially" equivalent (D-Lewis '16):

- **EB property**
- **Subregularity**:

$$\operatorname{dist}(x; \{\text{soln. set}\}) \leq \frac{1}{\alpha} \cdot \operatorname{dist}(0; \partial F(x)) \qquad \text{for } x \text{ near } \bar{x}$$

- **Quadratic growth**:

$$F(x) \geq F(\bar{x}) + \frac{\alpha}{2} \cdot \operatorname{dist}^2(x, \{\text{soln. set}\}) \qquad \text{for } x \text{ near } \bar{x}$$

Rate becomes $\frac{\alpha}{\mu}$ under tilt-stability (Poliquin-Rockafellar '98)

Acceleration

For nonconvex problems, the rate

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

is “essentially” best possible.

Acceleration

For nonconvex problems, the rate

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

is "essentially" best possible.

Measuring non-convexity:

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

Acceleration

For nonconvex problems, the rate

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

is "essentially" best possible.

Measuring non-convexity:

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

Fact: $x \mapsto \langle w, c(x)\rangle + \frac{\mu}{2}|x|^2 \quad \text{convex} \quad \forall w \in \operatorname{dom} h^\star$

Acceleration

For nonconvex problems, the rate

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

is "essentially" best possible.

Measuring non-convexity:

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

Fact: $\quad x \mapsto \langle w, c(x)\rangle + \frac{\mu}{2}|x|^2 \quad$ convex $\quad \forall w \in \operatorname{dom} h^\star$

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

$$x \mapsto \langle w, c(x)\rangle + \rho \cdot \frac{\mu}{2}|x|^2 \qquad \text{is convex} \quad \forall w \in \operatorname{dom} h^\star$$

Acceleration

$\boxed{\text{Accelerated prox-linear method!}}$

Accelerated prox-linear method!

Thm: (D-Kempton '16)

$$\min_{i=1,\ldots,k} \|\mathcal{G}(x_i)\|^2 \leq \mathcal{O}\left(\frac{\mu^2}{k^3}\right) + \rho \cdot \mathcal{O}\left(\frac{\mu^2 R^2}{k}\right)$$

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

Accelerated prox-linear method!

Thm: (D-Kempton '16)

$$\min_{i=1,\ldots,k} \|\mathcal{G}(x_i)\|^2 \leq \mathcal{O}\left(\frac{\mu^2}{k^3}\right) + \rho \cdot \mathcal{O}\left(\frac{\mu^2 R^2}{k}\right)$$

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

- Generalizes (Ghadimi, Lan '16) for additively composite.

Conclusions

Balanced approach:

- Computational complexity
- Acceleration
- Variational analysis

Happy Birthday, Jim!