Proximal methods for minimizing convex compositions

Courtney Paquette
Joint Work with D. Drusvyatskiy

Department of Mathematics
University of Washington (Seattle)

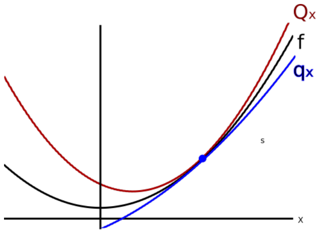
WCOM Fall 2016
October 1, 2016

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

$$q_x \leq f \leq Q_x$$

where

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

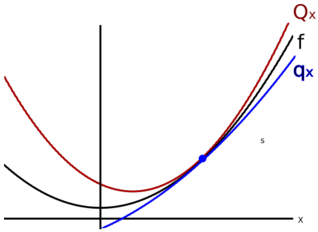


A function f is α-convex and β-smooth if

$$q_x \leq f \leq Q_x$$

where

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



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

Complexity of first-order methods

Gradient descent: $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$
Majorization view: $x_{k+1} = \text{argmin}_x\, Q_{x_k}(\cdot)$

	β-smooth	α-convex
Gradient descent	$\frac{\beta}{\varepsilon}$	$\kappa \cdot \log(\frac{1}{\varepsilon})$

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

(Nesterov '83, Yudin-Nemirovsky '83)

Complexity of first-order methods

Gradient descent: $x_{k+1} = x_k - \frac{1}{\beta} \nabla f(x_k)$
Majorization view: $x_{k+1} = \text{argmin}_x \, Q_{x_k}(\cdot)$

	β-smooth	α-convex
Gradient descent	$\frac{\beta}{\varepsilon}$	$\kappa \cdot \log(\frac{1}{\varepsilon})$
Optimal methods	$\sqrt{\frac{\beta}{\varepsilon}}$	$\sqrt{\kappa} \cdot \log(\frac{1}{\varepsilon})$

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

(Nesterov '83, Yudin-Nemirovsky '83)

General set-up

Convex-Composite Problem is

$$\min_x F(x) := h(c(x)) + g(x)$$

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

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

General set-up

Convex-Composite Problem is

$$\min_x F(x) := h(c(x)) + g(x)$$

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

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

Examples:

- Additive composite minimization

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

- Nonlinear least squares

$$\min_x \{\|c(x)\| \ : \ \ell_i \le x_i \le u_i, \quad i = 1, \ldots n\}$$

- Exact penalty subproblem:

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

Prox-Linear algorithm-Base Case

Seek points x which are *first-order stationary*: $F'(x; v) \geq 0 \quad \forall v$.
Equivalent to:

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

Prox-Linear algorithm-Base Case

Seek points x which are *first-order stationary*: $F'(x;v) \geq 0 \quad \forall v$.
Equivalent to:

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

Idea: Majorization

$$F(y) \leq h\left(c(x) + \nabla c(x)(y-x)\right) + \frac{\mu}{2}\|y-x\|^2 + g(y) \quad \forall\, y$$

Prox-linear mapping:

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

Prox-linear method:

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

(Burke '85, '91, Fletcher '82, Powell '84, Wright '90, Yuan '83)
Eg: proximal gradient, Levenberg-Marquardt

Prox-Linear algorithm-Base Case

Seek points x which are *first-order stationary*: $F'(x; v) \geq 0 \quad \forall v$.
Equivalent to:

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

Idea: Majorization

$$F(y) \leq h\left(c(x) + \nabla c(x)(y - x)\right) + \frac{\mu}{2}\|y - x\|^2 + g(y) \quad \forall y$$

Prox-linear mapping:

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

Prox-linear method:

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

(Burke '85, '91, Fletcher '82, Powell '84, Wright '90, Yuan '83)
Eg: proximal gradient, Levenberg-Marquardt
The **prox-gradient**

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

Prox-Linear algorithm

Convergence Rate:

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

What is $\|G(x_k)\|^2 < \varepsilon$?

$\text{dist}(0, \partial F(u_k)) \leq 5\,\|\mathcal{G}(x_k)\|$ with $\|u_k - x_k\| \approx \|\mathcal{G}(x_k)\|$
Pf: Ekeland's variational principle (Lewis-Drusvyatskiy '16)

Prox-Linear algorithm

Convergence Rate:

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

What is $\|G(x_k)\|^2 < \varepsilon$?

$\text{dist}(0, \partial F(u_k)) \leq 5\,\|\mathcal{G}(x_k)\|$ with $\|u_k - x_k\| \approx \|\mathcal{G}(x_k)\|$
Pf: Ekeland's variational principle (Lewis-Drusvyatskiy '16)

For nonconvex problems, the rate

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

is "essentially" the best

Prox-Linear algorithm

Convergence Rate:

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

What is $\|G(x_k)\|^2 < \varepsilon$?

$\text{dist}(0, \partial F(u_k)) \leq 5\,\|\mathcal{G}(x_k)\|$ with $\|u_k - x_k\| \approx \|\mathcal{G}(x_k)\|$
Pf: Ekeland's variational principle (Lewis-Drusvyatskiy '16)

For nonconvex problems, the rate

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

is "essentially" the best

Solving the Sub-problem: Inexact Prox-Linear

Prox-linear method requires solving:

$$\min_y \varphi(y,x) := h(c(x) + \nabla c(x)(y-x)) + \frac{\mu}{2}\|y-x\|^2 + g(y)$$

Suppose we can not solve exactly:

$$\varphi(x^+,x) \le \min_y \varphi(y,x) + \varepsilon$$

where x^+ is an ε-approximate optimal solution

Solving the Sub-problem: Inexact Prox-Linear

Prox-linear method requires solving:

$$\min_y \varphi(y,x) := h(c(x) + \nabla c(x)(y-x)) + \frac{\mu}{2}\|y-x\|^2 + g(y)$$

Suppose we can not solve exactly:

$$\varphi(x^+, x) \le \min_y \varphi(y,x) + \varepsilon$$

where x^+ is an ε-approximate optimal solution

Question

How accurately do we need to solve the subproblem to guarantee the same overall rate for the prox-linear?

Inexact Prox-Linear Algorithm

Want to bound

$\mathcal{G}(x_k) = \mu(x_k - x^*_{k+1})$, x^*_{k+1} is the **true** optimal point to the sub-problem

Inexact Prox-Linear Algorithm

Want to bound

$\mathcal{G}(x_k) = \mu(x_k - x^*_{k+1})$, x^*_{k+1} is the **true** optimal point to the sub-problem

Thm: (Drusvyatskiy-P '16)
Suppose x_{i+1} is an ε_{i+1}-approximate optimal solution. Then

$$\min_{i=1,\ldots,k} \|\mathcal{G}(x_i)\|^2 \leq \mathcal{O}\left(\frac{\mu + \sum_{i=1}^{k} \varepsilon_i}{k}\right).$$

- Generalizes (Schmidt-Le Roux-Bach '11)

Inexact Prox-Linear Algorithm

Want to bound

$\mathcal{G}(x_k) = \mu(x_k - x^*_{k+1})$, x^*_{k+1} is the **true** optimal point to the sub-problem

Thm: (Drusvyatskiy-P '16)
Suppose x_{i+1} is an ε_{i+1}-approximate optimal solution. Then

$$\min_{i=1,\ldots,k} \|\mathcal{G}(x_i)\|^2 \leq \mathcal{O}\left(\frac{\mu + \sum_{i=1}^k \varepsilon_i}{k}\right).$$

- Generalizes (Schmidt-Le Roux-Bach '11)

Question

Design an acceleration scheme

1. Optimal rate for convex problems
2. Rate no worse than prox-gradient for nonconvex problems
3. Detects convexity of the function

Acceleration

Measuring non-convexity,

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

Fact 1: $h \circ c(x)$ is convex if $x \mapsto \langle w, c(x)\rangle$ is convex for all $w \in \operatorname{dom} h^*$.

Fact 2: $x \mapsto \langle w, c(x)\rangle + \frac{\mu}{2}\|x\|^2$ is convex for all $w \in \operatorname{dom} h^*$

Acceleration

Measuring non-convexity,

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

Fact 1: $h \circ c(x)$ is convex if $x \mapsto \langle w, c(x) \rangle$ is convex for all $w \in \operatorname{dom} h^*$.

Fact 2: $x \mapsto \langle w, c(x) \rangle + \frac{\mu}{2} \|x\|^2$ is convex for all $w \in \operatorname{dom} h^*$

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

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

Acceleration

Algorithm 1: Accelerated prox-linear method

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

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

$\quad a_k \leftarrow \frac{2}{k+1}$

$\quad y_k \leftarrow a_k v_{k-1} + (1 - a_k)x_{k-1}$

$\quad x_k \leftarrow y_k^+$

$\quad v_k \leftarrow \operatorname{argmin}_z g(z) + \frac{1}{a_k} \cdot h\big(c(y_k) + a_k \nabla c(y_k)(z - v_{k-1})\big) + \frac{a_k}{2t} \|z - v_{k-1}\|^2$

$\quad k \leftarrow k + 1$

end

Thm: (Drusvyatskiy-P '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 additive composite

Inexact Accelerated Prox-Linear

Two sub-problems to solve:

- x_k is an ε_k-approximate optimal solution

$$\min_z g(z) + h\big(c(y_k) + \nabla c(y_k)(z - y_k)\big) + \frac{1}{2t}\|z - y_k\|^2$$

- v_k is an δ_k-approximate optimal solution

$$\min_z g(z) + \frac{1}{a_k} \cdot h\big(c(y_k) + a_k \nabla c(y_k)(z - v_{k-1})\big) + \frac{a_k}{2t}\|z - v_{k-1}\|^2$$

Inexact Accelerated Prox-Linear

Two sub-problems to solve:

- x_k is an ε_k-approximate optimal solution

$$\min_z g(z) + h\big(c(y_k) + \nabla c(y_k)(z - y_k)\big) + \tfrac{1}{2t}\|z - y_k\|^2$$

- v_k is an δ_k-approximate optimal solution

$$\min_z g(z) + \tfrac{1}{a_k} \cdot h\big(c(y_k) + a_k \nabla c(y_k)(z - v_{k-1})\big) + \tfrac{a_k}{2t}\|z - v_{k-1}\|^2$$

$$\operatorname{dist}(0, \partial F(u_k)) \leq C(\|x_k - y_k\| + \sqrt{\varepsilon_k}), \quad \|u_k - x_k\| \approx \|x_k - y_k\| + \sqrt{\varepsilon_k}$$

Inexact Accelerated Prox-Linear

Two sub-problems to solve:

- x_k is an ε_k-approximate optimal solution

$$\min_z \; g(z) + h\big(c(y_k) + \nabla c(y_k)(z - y_k)\big) + \tfrac{1}{2t}\|z - y_k\|^2$$

- v_k is an δ_k-approximate optimal solution

$$\min_z \; g(z) + \tfrac{1}{a_k} \cdot h\big(c(y_k) + a_k \nabla c(y_k)(z - v_{k-1})\big) + \tfrac{a_k}{2t}\|z - v_{k-1}\|^2$$

$\mathrm{dist}(0, \partial F(u_k)) \leq C(\|x_k - y_k\| + \sqrt{\varepsilon_k}), \quad \|u_k - x_k\| \approx \|x_k - y_k\| + \sqrt{\varepsilon_k}$

Thm: (Drusvyatskiy-P '16)

$$\min_{i=1,\ldots,k} \{\|x_k - y_k\|^2 + \varepsilon_i\} \leq \rho \cdot \mathcal{O}\left(\frac{\mu^2 R^2}{k}\right) + \mathcal{O}\left(\frac{\mu^2}{k^3}\right) + \frac{1}{k^3}\left(\sum_{i=1}^{k} \mathcal{O}(i^2 \varepsilon_i) + \mathcal{O}(i^2 \delta_i) + \mathcal{O}(i\sqrt{\delta_i})\right)$$

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

Need: $\varepsilon_i \sim \frac{1}{i^{3+r}}$ and $\delta_i \sim \frac{1}{i^{4+r}}$

Thank you!

References

Drusvyatskiy, D. and Kempton, C. (2016).
An accelerated algorithm for minimizing convex compositions.
Preprint arXiv: 1605.00125.

Drusvyatskiy, D. and Lewis, A. (2016).
Error bounds, quadratic growth, and linear convergence of proximal methods.
Preprint arXiv:1602.06661.

Ghadimi, S. and Lan, G. (2016).
Accelerated gradient methods for nonconvex nonlinear and stochastic programming.
Math. Program., 156(1-2, Ser. A):59–99.

Nesterov, Y. (2004).
Introductory lectures on convex optimization. A Basic Course.
Springer.

References (cont.)

Schmidt, M., Le Roux, N., and Bach, F. (2011).
Convergence rates of inexact proximal-gradient methods for convex optimization.
Advances in Neural Information Processing Systems.