Error bounds and convergence of proximal methods for composite minimization

Adrian Lewis
(with D. Drusvyatskiy, A. Ioffe, and S. Wright)

ORIE Cornell

CMO-BIRS Workshop on Splitting Algorithms September 2017

Outline

Prox-linear algorithms for composite minimization:

$$g + c \quad \text{or} \quad h\big(c(\cdot)\big)$$

for simple nonsmooth g and h, and smooth c.

- Background: Fletcher '82, . . .
  . . . , sparse estimation via nonconvex regularization.
- Global convergence: limit points and sublinear rate.
- Prox-gradient steps as a stopping criterion.
- Error bounds, quadratic growth, linear convergence.
- Partial smoothness and second-order acceleration.

ProxDescent algorithm for min $h(c(\cdot))$

Nonsmooth but **"simple"** $h\colon \mathbf{R}^m \to \mathbf{R}$ (initially finite convex).
Smooth $c\colon \mathbf{R}^n \to \mathbf{R}^m$. Around **iterate** x,

$$\tilde{c}(d) \;=\; c(x) + \nabla c(x) d \;\approx\; c(x+d).$$

Unique **step** d solves **easy** subproblem

$$\min_d \; h\big(\tilde{c}(d)\big) + \mu \|d\|^2.$$

Update **prox parameter** μ: **if**

$$\text{actual decrease} \;=\; h\big(c(x)\big) - h\big(c(x+d)\big)$$

less than half

$$\text{predicted decrease} \;=\; h\big(c(x)\big) - h\big(\tilde{c}(d)\big),$$

reject: $\mu \leftarrow 2\mu$; otherwise,
accept: $x \leftarrow x + d, \quad \mu \leftarrow \frac{\mu}{2}$.
Repeat. (L-Wright '15)

Examples: exact penalties, compressive sensing

$$\min_x p(x) + \nu \sum_i q_i^+(x).$$

Easy subproblems:

$$\min_d \; s^T d + \sum_i (a_i^T d + b_i)^+ + \mu \|d\|^2.$$

Follows Fletcher '82, Powell '84, Yuan '85, Burke '85, Wright '90, Byrd et al. '05 (KNITRO), Friedlander et al. 07...

Sparse solve $Ax = b$ (Candès, Donoho, Tao et al. '06...) via

$$\min_x \; \|Ax - b\|^2 + \tau \|x\|_1.$$

Separable subproblems:

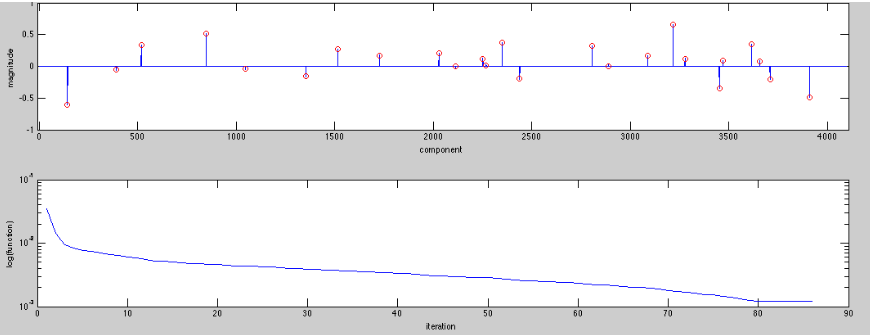
$$\min_{d \in \mathbf{R}^n} \; s^T d + \tau \|x + d\|_1 + \mu \|d\|^2.$$

Just $O(n)$ operations: SpaRSA (Wright et al. '09).

Example: nonconvex regularizers for sparse estimation

$$\min_{\mathbf{x}} \|A\mathbf{x} - b\|^2 + \tau \sum_i \phi(\mathbf{x}_i) \qquad \text{(Zhao et al. '10)}.$$

Random 256-by-4096 A, sparse $\hat{\mathbf{x}}$, and $b = A\hat{\mathbf{x}} + \text{noise}$.



Eventual slow linear convergence.

Global convergence of prox-linear methods

Theorem (L-Wright '15)

For arbitrary h (nonsmooth or extended-valued), limit points $\bar{x}$ of iterates are **stationary** for objective $f = h\big(c(\cdot)\big)$:

$$f(x) - f(\bar{x}) \geq o(\|x - \bar{x}\|).$$

Rate? More generally, if g, h convex (for now),

$$\underset{x}{\text{minimize}} \;\; g(x) + h\big(c(x)\big)$$

via iteration $x \leftarrow x +$ **prox-gradient step**

$$d(x, \mu) \;=\; \underset{d}{\text{argmin}} \;\; g(x+d) + h\big(\tilde{c}(d)\big) + \mu\|d\|^2.$$

If $h, \nabla c$ are β, γ-Lipschitz, the steps $d_1, d_2, \ldots$ become small:

$$d_k = O(k^{-\frac{1}{2}})$$

providing $\mu \geq \beta\gamma$. (Drusvyatskiy-L '16).

Small prox-gradient steps $\Rightarrow$ near-stationarity

When should we stop the prox-linear method for minimizing

$$f(x) = g(x) + h\big(c(x)\big)?$$

Theorem (Drusvyatskiy-L '16)

If the step d is small, then the iterate x is "nearly" stationary.

Precisely: corresponding to the step $d = d(x, \mu)$
is a point $\hat{x}$ and a vector v satisfying

$$\|x + d - \hat{x}\| \le \|d\| \quad \text{and} \quad \|v\| \le 5\mu\|d\|,$$

such that

$$f(\cdot) + \langle v, \cdot \rangle$$

is stationary at $\hat{x}$. (Proof via Ekeland principle).

Linear convergence and prox-gradient error bounds

Minimizing $f(\cdot) = g(\cdot) + h\big(c(\cdot)\big)$ gives iterates x_k. Around any limit point $\bar{x}$, suppose stepsize bounds **distance** to a minimizer:

$$\mathrm{dist}(x) \;=\; \min_{y \in \operatorname{argmin} f} \|x - y\| \;\leq\; \frac{1}{\alpha}\|d(x,\mu)\|. \qquad (*)$$

Then (Luo-Tseng '93) the **excess** $e(\cdot) = f(\cdot) - \min f$ shrinks:

$$\frac{e(x_{k+1})}{e(x_k)} \;\leq\; 1 - \alpha^2.$$

Theorem (Drusvyatskiy-L '16)

The error bound $(*)$ is equivalent to local **quadratic growth**:

$$e(x) \;\geq\; \frac{\mu\alpha}{2}\mathrm{dist}^2(x)$$

(and to "metric subregularity" of the subdifferential ∂f).

General Taylor-like models

Stationary points for closed objective f on complete metric (X, d) have nearby points (with nearby value) and small **slope**:

$$|\nabla f|(x) \;=\; \limsup_{y \to x} \frac{\big(f(x) - f(y)\big)^{+}}{d(x, y)}.$$

Algorithms iteratively minimize closed **model** m around current x:

$$|m(y) - f(y)| \;\leq\; \eta d^2(x, y) \qquad (y \in X).$$

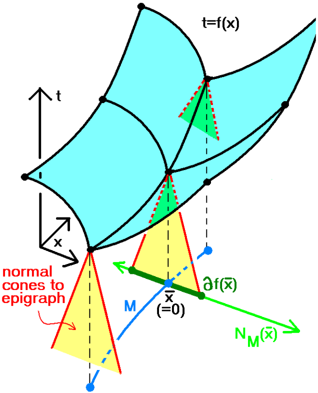
Model minimizer x^+ gives **step size** $\epsilon = d(x, x^+)$.
Then (Drusvyatskiy-Ioffe-L '17), some $\hat{x}$ satisfies

$$\begin{aligned} d(\hat{x}, x^+) &\leq \epsilon \\ f(\hat{x}) - f(x^+) &\leq \eta\epsilon^2 \\ |\nabla f|(\hat{x}) &\leq 10\eta\epsilon. \end{aligned}$$

Small steps $\Rightarrow$ nearly stationary.

Partial smoothness: the easiest nonconvex case

- Well-behaved on "active manifold" $\mathcal{M}$: $f|_{\mathcal{M}}$ smooth and critical at $\bar{x}$, fixed-directional derivatives $f'(\cdot;y)$ continuous.
- Prox-regularity: points near $\big(\bar{x}, f(\bar{x})\big)$ have unique nearest points in the epigraph $\big\{(x,t) : t \geq f(x)\big\}$.
- Sharp growth: $f'(\bar{x};y) > 0$ for unit normals y to $\mathcal{M}$.



The "active set" philosophy

Quadratic growth, and hence linear convergence, simplifies for partly smooth f:

$$f \text{ grows at least quadratically} \quad \Leftrightarrow \quad f|_{\mathcal{M}} \text{ grows quadratically}$$

(verifiable simply via a Hessian.)

Furthermore $\mathcal{M}$ is **identifiable** (Wright '93): $y_k \to 0$ and $f + \langle y_k, \cdot \rangle$ stationary at $x_k \to \bar{x} \quad \Rightarrow \quad x_k \in \mathcal{M}$ eventually.

Hence high-dimensional nonsmooth optimization

$$\min f$$

reduces locally to low-dimensional smooth equality-constrained

$$\min f|_{\mathcal{M}}.$$

Now **accelerate** using a second-order model.

Acceleration

The prox-linear method for minimizing $f(\cdot) = h\big(c(\cdot)\big)$ generates steps d_k, and corresponding iterates x_k having a limit point $\bar{x}$.

Suppose h is partly smooth at $c(\bar{x})$ relative to a manifold $\mathcal{N}$, and assume objective quadratic growth. Then $x_k \to \bar{x}$ (linearly).

Identifiability $\Rightarrow$ $c(x_k) + \nabla c(x_k) d_k \in \mathcal{N}$ eventually.

Classical algorithms

- use d_k to predict the active set.
- accelerate using a second-order model.

Generalize for simple h (L-Wright '15):

- "Track" $\mathcal{N}$.
- Build a second-order model from c and $h|_{\mathcal{N}}$.

(See also Mifflin-Sagastizábal '05).

Composite prox-linear methods: highlights

- Simple and intuitive.
- Unifying classical and modern algorithmic frameworks.
- Robust and scaleable in practice.
- Comprehensive convergence theory:
  - limit points are stationary
  - basic sublinear rate
  - stopping criterion
  - linear convergence and quadratic growth
  - partial smoothness and second-order acceleration.
- The lens of variational analysis.