Proximal-Gradient Homotopy Methods for Sparse Optimization

Lin Xiao (Microsoft Research)

Joint work with
Tong Zhang (Rutgers University)
Qihang Lin (Carnegie Mellon University)

IPAM Workshop on
Structure and Randomness in System Identification and Learning
UCLA, January 16, 2013

The sparse least-squares problem

- ℓ_1-regularized least-squares (ℓ_1-LS) problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $\lambda > 0$

The sparse least-squares problem

- ℓ_1-regularized least-squares (ℓ_1-LS) problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

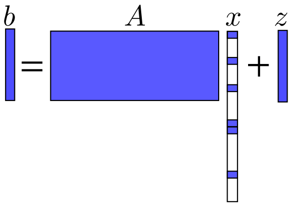
where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $\lambda > 0$

- example: sparse recovery

$$b = Ax + z$$

x: sparse signal

z: observation noise



The sparse least-squares problem

- ℓ_1-regularized least-squares (ℓ_1-LS) problem

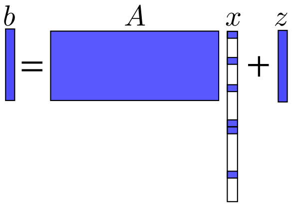
$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $\lambda > 0$

- example: sparse recovery

$$b = Ax + z$$

x: sparse signal
z: observation noise



- many applications
  - machine learning, signal/image processing and statistics
  - recent revival through *compressed sensing* theory

Efficiency of optimization methods

- assumptions:
  - $m < n$ ($Ax = b$ is underdetermined, "high-dimensional")
  - solution is sparse (λ sufficiently large, $s = \|x^\star(\lambda)\|_0$ small)

Efficiency of optimization methods

- assumptions:
  - $m < n$ ($Ax = b$ is underdetermined, "high-dimensional")
  - solution is sparse (λ sufficiently large, $s = \|x^\star(\lambda)\|_0$ small)

- complexities for finding ϵ-optimal solution

numerical methods	cost per iteration	iteration complexity
interior-point methods	$O\left(m^2 n\right)$	$O\left(\sqrt{n} \log\left(\frac{1}{\epsilon}\right)\right)$
proximal-gradient (PG)	$O\left(mn\right)$	$O\left(\frac{1}{\epsilon}\right)$
accelerated PG	$O\left(mn\right)$	$O\left(\frac{1}{\sqrt{\epsilon}}\right)$

Efficiency of optimization methods

- assumptions:
  - $m < n$ ($Ax = b$ is underdetermined, "high-dimensional")
  - solution is sparse (λ sufficiently large, $s = \|x^\star(\lambda)\|_0$ small)

- complexities for finding ϵ-optimal solution

numerical methods	cost per iteration	iteration complexity
interior-point methods	$O\left(m^2 n\right)$	$O\left(\sqrt{n}\log\left(\frac{1}{\epsilon}\right)\right)$
proximal-gradient (PG)	$O\left(mn\right)$	$O\left(\frac{1}{\epsilon}\right)$
accelerated PG	$O\left(mn\right)$	$O\left(\frac{1}{\sqrt{\epsilon}}\right)$

- **this talk:** with an additional RIP-like condition on A

(accelerated) PG + homotopy continuation $\qquad O\left(mn \cdot \log\left(\frac{1}{\epsilon}\right)\right)$

Outline

- **background: first-order methods and their complexities**
- proximal-gradient (PG) method + homotopy
- accelerated proximal gradient (APG) method + homotopy
- numerical experiments and summary

Classes of convex functions

- **convex**

$$f(\alpha x+(1-\alpha)y) \le \alpha f(x)+(1-\alpha)f(y), \quad \forall\, \alpha \in [0,1]$$

- **smooth** with parameter L

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2, \quad \forall\, x, y$$

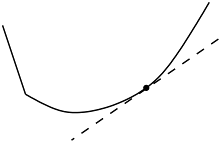
- **strongly convex** with parameter μ

$$f(\alpha x+(1-\alpha)y) \le \alpha f(x)+(1-\alpha)f(y)-\alpha(1-\alpha)\frac{\mu}{2}\|x-y\|_2^2$$

Classes of convex functions

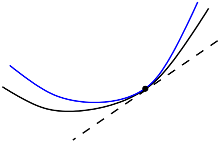
- **convex**

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$



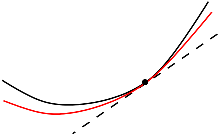
- **smooth** with parameter L

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2$$



- **strongly convex** with parameter μ

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2$$



Proximal mapping

proximal mapping (prox-operator) of a convex function Ψ is

$$\mathbf{prox}_{\Psi}(x) = \operatorname*{argmin}_{u} \left\{ \Psi(u) + \frac{1}{2}\|u - x\|_2^2 \right\}$$

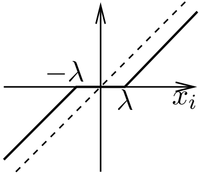
examples:

- projection: $\Psi(x) = I_C(x)$ (indicator function of a convex set C)

$$\mathbf{prox}_{\Psi}(x) = \operatorname*{argmin}_{u \in C} \|u - x\|_2^2$$

- soft thresholding (shrinkage): $\Psi(x) = \lambda\|x\|_1$

$$\mathbf{prox}_{\lambda\|\cdot\|_1}(x)_i = \begin{cases} x_i - \lambda & x_i > \lambda \\ 0 & |x_i| \le \lambda \\ x_i + \lambda & x_i < -\lambda \end{cases}$$



Proximal gradient (PG) method

minimizing composite objective

$$\underset{x}{\text{minimize}} \quad \left\{ \phi(x) \triangleq f(x) + \Psi(x) \right\}$$

- f convex and smooth with parameter L (and $\mu \geq 0$)
- Ψ convex and simple (can easily compute $\mathbf{prox}_{\Psi}$)

Proximal gradient (PG) method

minimizing composite objective

$$\underset{x}{\text{minimize}} \quad \left\{\phi(x) \triangleq f(x) + \Psi(x)\right\}$$

- f convex and smooth with parameter L (and $\mu \geq 0$)
- Ψ convex and simple (can easily compute $\mathbf{prox}_\Psi$)

PG method:

$$x^{(k+1)} = \mathbf{prox}_{\frac{1}{L}\Psi}\left(x^{(k)} - \frac{1}{L}\nabla f(x^{(k)})\right)$$

interpretation:

$$x^{(k+1)} = \underset{y}{\text{argmin}}\left\{f(x^{(k)}) + \nabla f(x^{(k)})^T(y - x^{(k)}) + \frac{L}{2}\|y - x^{(k)}\|_2^2 + \Psi(y)\right\}$$

Structure and complexity

- problem: composite convex optimization

$$\underset{x}{\text{minimize}} \quad \left\{ \phi(x) \triangleq f(x) + \Psi(x) \right\}$$

- iteration complexity to reach $\phi(x^{(k)}) - \phi^\star \leq \epsilon$

class of f	smooth	smooth + strongly convex
PG	$O\left(\frac{L}{\epsilon}\right)$	$O\left(\frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$ (not require μ)

Structure and complexity

- problem: composite convex optimization

$$\underset{x}{\text{minimize}} \quad \left\{\phi(x) \triangleq f(x) + \Psi(x)\right\}$$

- iteration complexity to reach $\phi(x^{(k)}) - \phi^\star \le \epsilon$

class of f	smooth	smooth + strongly convex
PG	$O\left(\frac{L}{\epsilon}\right)$	$O\left(\frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$ (not require μ)
accelerated PG	$O\left(\sqrt{\frac{L}{\epsilon}}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$ (require μ)

Nesterov (04, 07), Beck & Teboulle (08), Tseng (08)

Two simple APG methods

- a simple variant of FISTA (Beck & Teboulle 2008): $O\left(\sqrt{\frac{L}{\epsilon}}\right)$

$$
\begin{aligned}
y^{(k)} &= x^{(k)} + \frac{k-1}{k+2}(x^{(k)} - x^{(k-1)}) \\
x^{(k+1)} &= \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right)
\end{aligned}
$$

Two simple APG methods

- a simple variant of FISTA (Beck & Teboulle 2008): $O\left(\sqrt{\frac{L}{\epsilon}}\right)$

$$\begin{aligned} y^{(k)} &= x^{(k)} + \frac{k-1}{k+2}(x^{(k)} - x^{(k-1)}) \\ x^{(k+1)} &= \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right) \end{aligned}$$

- Nesterov's constant step scheme III (2004): $O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$

$$\begin{aligned} y^{(k)} &= x^{(k)} + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(x^{(k)} - x^{(k-1)}) \\ x^{(k+1)} &= \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right) \end{aligned}$$

(does not work for $\mu = 0$, but there are more general variants)

Outline

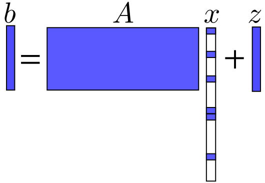
- background: first-order methods and their complexities
- **proximal-gradient (PG) method + homotopy**
- accelerated proximal gradient (APG) method + homotopy
- numerical experiments and summary

Complexity of solving ℓ_1-LS problem

given $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $\lambda > 0$

$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1$$

(focus on the case $m < n$)

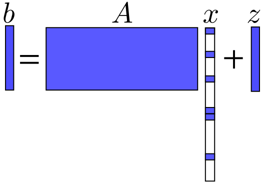


Complexity of solving ℓ_1-LS problem

given $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $\lambda > 0$

$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

(focus on the case $m < n$)



check $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

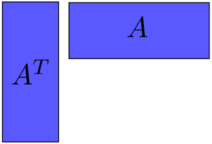
- smooth:

$$L_f = \lambda_{\max}(A^TA)$$

- **not** strongly convex:

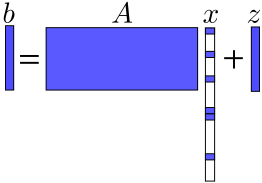
$$\mu_f = \lambda_{\min}(A^TA) = 0$$

$\nabla^2 f(x) =$



Complexity of solving ℓ_1-LS problem

given $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $\lambda > 0$



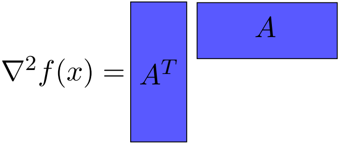
$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

(focus on the case $m < n$)

check $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

- smooth:

$$L_f = \lambda_{\max}(A^T A)$$



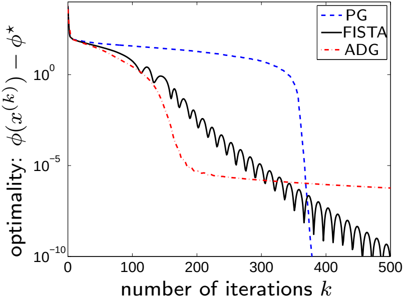
- **not** strongly convex:

$$\mu_f = \lambda_{\min}(A^T A) = 0$$

so we only expect sublinear convergence: $O\left(\frac{L_f}{\epsilon}\right)$ or $O\left(\sqrt{\frac{L_f}{\epsilon}}\right)$

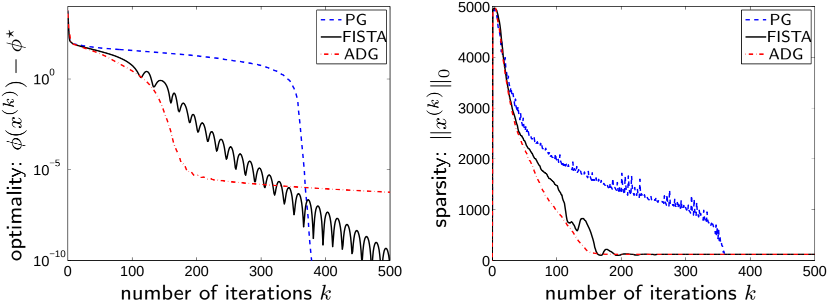
Experiments: two phases of PG method

- $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ generated randomly ($m = 1000$, $n = 5000$)
- algorithms: PG, ADG (Nesterov 07), FISTA (Beck & Teboulle 08)



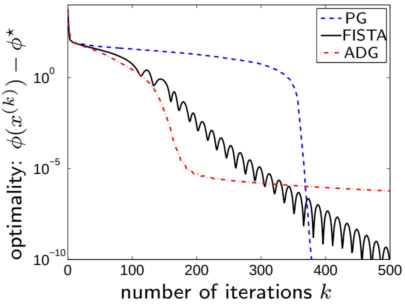
Experiments: two phases of PG method

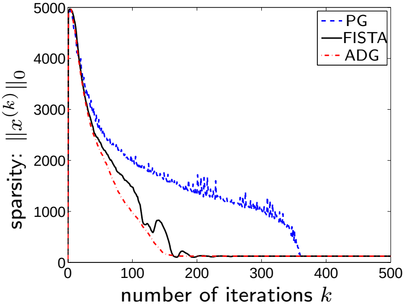
- $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ generated randomly ($m = 1000$, $n = 5000$)
- algorithms: PG, ADG (Nesterov 07), FISTA (Beck & Teboulle 08)



Experiments: two phases of PG method

- $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ generated randomly ($m = 1000$, $n = 5000$)
- algorithms: PG, ADG (Nesterov 07), FISTA (Beck & Teboulle 08)



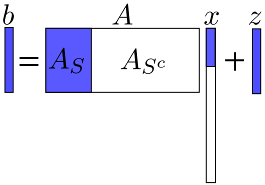


key observations:

- slow global convergence (sublinear rate $O(1/\epsilon)$)
- fast linear rate when iterates become sparse *and* close to optimal

Structure: restricted strong convexity

suppose optimal solution is sparse



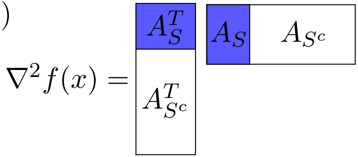
$$x^\star = \begin{bmatrix} x^\star_S \\ x^\star_{S^c} \end{bmatrix} = \begin{bmatrix} x^\star_S \\ 0 \end{bmatrix}$$

- restricted smoothness:

$$L_S = \lambda_{\max}(A_S^T A_S) < \lambda_{\max}(A^T A)$$

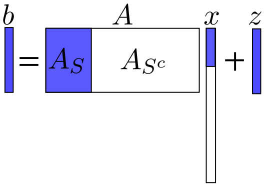
- restricted strong convexity:

$$\mu_S = \lambda_{\min}(A_S^T A_S) > 0$$



Structure: restricted strong convexity

suppose optimal solution is sparse



$$x^\star = \begin{bmatrix} x_S^\star \\ x_{S^c}^\star \end{bmatrix} = \begin{bmatrix} x_S^\star \\ 0 \end{bmatrix}$$

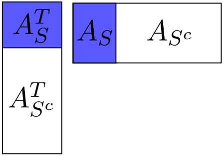
- restricted smoothness:

$$L_S = \lambda_{\max}(A_S^T A_S) < \lambda_{\max}(A^T A)$$

- restricted strong convexity:

$$\mu_S = \lambda_{\min}(A_S^T A_S) > 0$$

$\nabla^2 f(x) =$



conclusion: if we can identify the sparse subspace S, then minimizing $\frac{1}{2}\|A_S x_S - b\|_2^2$ gives fast linear convergence rat

The homotopy continuation idea

key observations:

- PG has slow global convergence (sublinear rate $O(1/\epsilon)$)
- fast linear rate when iterates become sparse and near optimal

The homotopy continuation idea

key observations:

- PG has slow global convergence (sublinear rate $O(1/\epsilon)$)
- fast linear rate when iterates become sparse and near optimal

idea: always engage in sparse mode (fast local convergence)

homotopy continuation: solve ℓ_1-LS for decreasing values of λ

$$\|A^T b\|_\infty = \lambda_0 > \lambda_1 > \lambda_2 > \cdots > \lambda_{\text{tgt}}$$

for each λ_K, solve with PG and use previous solution to warm start

The homotopy continuation idea

key observations:

- PG has slow global convergence (sublinear rate $O(1/\epsilon)$)
- fast linear rate when iterates become sparse and near optimal

idea: always engage in sparse mode (fast local convergence)

homotopy continuation: solve ℓ_1-LS for decreasing values of λ

$$\|A^T b\|_\infty = \lambda_0 > \lambda_1 > \lambda_2 > \cdots > \lambda_{\mathrm{tgt}}$$

for each λ_K, solve with PG and use previous solution to warm start

- not a new idea (e.g., Hale et al. 2008, Wright et al. 2009)
- superior performance reported, but no global complexity analysis
- **questions:** how to decrease λ? how accurate for each λ_K?

Proximal-gradient homotopy (PGH) method

parameters: $\eta \in (0,1)$, $\delta \in (0,1)$

Algorithm: $\hat{x}^{(\mathsf{tgt})} \leftarrow \texttt{Homotopy}(A, b, \lambda_{\mathsf{tgt}}, \epsilon)$

---

initialize: $\lambda_0 \leftarrow \|A^T b\|_\infty, \;\; \hat{x}^{(0)} \leftarrow 0$

$N \leftarrow \lfloor \ln(\lambda_0/\lambda_{\mathsf{tgt}}) \,/\, \ln(1/\eta) \rfloor$

repeat: for $K = 0, 1, \ldots, N-1$

$\lambda_{K+1} \leftarrow \eta \lambda_K$ (geometric decrease $\lambda_K = \eta^K \lambda_0$)
$\hat{\epsilon}_{K+1} \leftarrow \delta \lambda_{K+1}$ (low accuracy proportional to λ_K)

$x^{(K+1)} \leftarrow \texttt{ProxGrad}\big(\lambda_{K+1}, \hat{\epsilon}_{K+1}, \hat{x}^{(K)}\big)$

end

$\hat{x}^{(\mathsf{tgt})} \leftarrow \texttt{ProxGrad}\big(\lambda_{\mathsf{tgt}}, \epsilon, \hat{x}^{(N)}\big)$ (final stage for high accuracy)

Stopping criterion and line search

$$\underset{x}{\text{minimize}} \quad \left\{ \phi_\lambda(x) \triangleq f(x) + \lambda \|x\|_1 \right\}$$

- stopping criterion

$$\omega_\lambda(x) \triangleq \min_{\xi \in \partial \|x\|_1} \|\nabla f(x) + \lambda \xi\|_\infty \;\leq\; \epsilon$$

(x optimal iff there exists $\xi \in \partial \|x\|_1$ such that $\nabla f(x) + \lambda \xi = 0$)

Stopping criterion and line search

$$\underset{x}{\text{minimize}} \quad \left\{ \phi_\lambda(x) \triangleq f(x) + \lambda \|x\|_1 \right\}$$

- stopping criterion

$$\omega_\lambda(x) \triangleq \min_{\xi \in \partial \|x\|_1} \|\nabla f(x) + \lambda \xi\|_\infty \ \leq \ \epsilon$$

(x optimal iff there exists $\xi \in \partial \|x\|_1$ such that $\nabla f(x) + \lambda \xi = 0$)

- line search to use $L \approx \lambda_{\max}(A_S^T A_S)$ in PG method

$$x^+ = \mathbf{prox}_{\lambda \|\cdot\|_1} \left(x - \frac{1}{L} \nabla f(x) \right) = \underset{y}{\text{argmin}} \ \psi_{\lambda,L}(x; y)$$

where $\psi_{\lambda,L}(x; y) = f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2 + \lambda \|y\|_1$

PG with adaptive line search (Nesterov 07)

parameters: $\gamma_{\mathsf{inc}} \geq 1$

Algorithm: $\{x^+, L\} \leftarrow \texttt{LineSearch}(\lambda, x, L)$

---

repeat:

$L \leftarrow L\gamma_{\mathsf{inc}}$

$x^+ = \mathbf{prox}_{\lambda\|\cdot\|_1}\left(x - \frac{1}{L}\nabla f(x)\right)$

until: $\phi_\lambda(x^+) <= \psi_{\lambda,L}(x, x^+)$

PG with adaptive line search (Nesterov 07)

parameters: $\gamma_{\text{inc}} \geq 1$, $\gamma_{\text{dec}} \geq 1$, $L_{\text{min}} > 0$

Algorithm: $\{x^+, L\} \leftarrow \texttt{LineSearch}(\lambda, x, L)$

---

repeat:

$L \leftarrow L\gamma_{\text{inc}}$

$x^+ = \mathbf{prox}_{\lambda\|\cdot\|_1}\left(x - \frac{1}{L}\nabla f(x)\right)$

until: $\phi_\lambda(x^+) <= \psi_{\lambda,L}(x, x^+)$

Algorithm: $\{\hat{x}, \hat{M}\} \leftarrow \texttt{ProxGrad}(\lambda, \hat{\epsilon}, x^{(0)}, L_0)$

---

repeat: for $k = 0, 1, 2, \ldots$

$\{x^{(k+1)}, M_k\} \leftarrow \texttt{LineSearch}(\lambda, x^{(k)}, L_k)$

$L_{k+1} \leftarrow \max\{L_{\text{min}}, M_k/\gamma_{\text{dec}}\}$

until: $\omega_\lambda(x^{(k+1)}) \leq \hat{\epsilon}$

$\hat{x} \leftarrow x^{(k+1)}$, $\hat{M} \leftarrow M_k$

Numerical experiments: setup

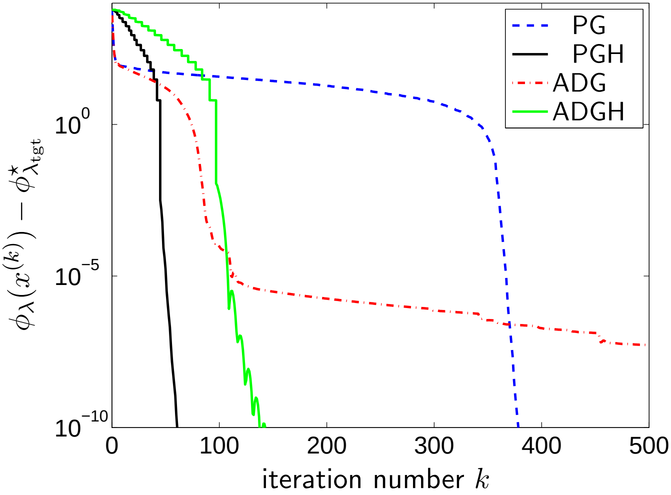
randomly generated data using the model

$$b = A\bar{x} + z$$

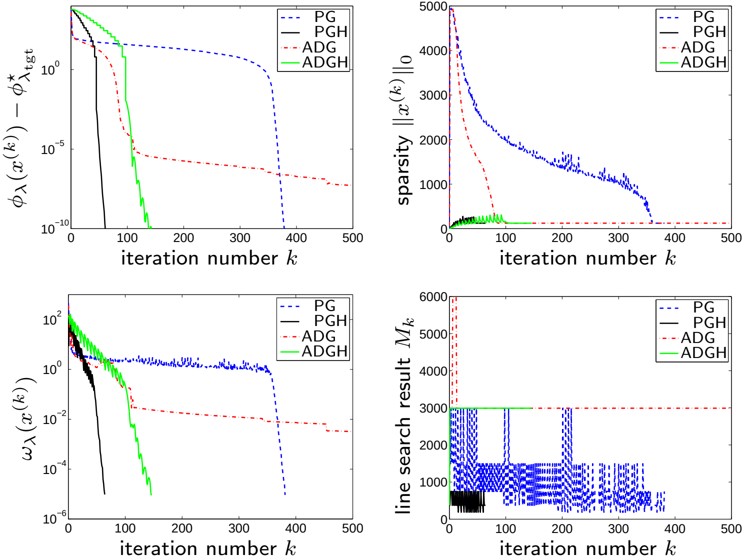
- $A \in \mathbb{R}^{m \times n}$, with $m = 1000$ and $n = 5000$, $A_{ij} \sim \mathsf{U}[-1, 1]$
- $\|\bar{x}\|_0 = 100$, nonzeros entries i.i.d. uniform on $[-1, 1]$
- z entries i.i.d. uniform on $[-0.01, 0.01]$, and $\|A^T z\|_\infty \approx 0.4$
- regularization parameter $\lambda_{\mathsf{tgt}} = 1$ (note $\lambda_0 = \|A^T b\|_\infty = 480$)

PGH method parameters: $\eta = 0.8$, $\delta = 0.2$

Numerical experiments



Numerical experiments



Restricted eigenvalue (RE) conditions

for some $s < m$, there exist $\rho_+(A,s) \geq \rho_-(A,s) > 0$ such that

$$
\begin{aligned}
\rho_+(A,s) &= \sup\left\{ \frac{x^T A^T A x}{x^T x} : x \neq 0,\ \|x\|_0 \leq s \right\} \\
\rho_-(A,s) &= \inf\left\{ \frac{x^T A^T A x}{x^T x} : x \neq 0,\ \|x\|_0 \leq s \right\}
\end{aligned}
$$

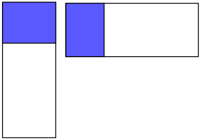
Restricted eigenvalue (RE) conditions

for some $s < m$, there exist $\rho_+(A,s) \geq \rho_-(A,s) > 0$ such that

$$\begin{aligned} \rho_+(A,s) &= \sup\left\{\frac{x^TA^TAx}{x^Tx} : x \neq 0,\ \|x\|_0 \leq s\right\} \\ \rho_-(A,s) &= \inf\left\{\frac{x^TA^TAx}{x^Tx} : x \neq 0,\ \|x\|_0 \leq s\right\} \end{aligned}$$

by definition

$$0 = \lambda_{\mathsf{min}}(A^TA) \ \leq\ \rho_-(A,s) \ \leq\ \rho_+(A,s) \ \leq\ \lambda_{\mathsf{max}}(A^TA)$$

recall the picture: $\nabla^2 f(x) = A^TA =$ 

Restricted smoothness & strong convexity

Suppose for some integer $s < m$, two sparse vectors x and y satisfy

$$\left|\mathsf{supp}(x) \cup \mathsf{supp}(y)\right| \leq s$$

- restricted smoothness

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\rho_+(A, s)}{2} \|y - x\|_2^2$$

- restricted strong convexity

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\rho_-(A, s)}{2} \|y - x\|_2^2$$

- restricted condition number: $\quad \kappa(A, s) \triangleq \dfrac{\rho_+(A, s)}{\rho_-(A, s)}$

Convergence analysis: assumptions

suppose $b = A\bar{x} + z$; let $\bar{S} = \mathsf{supp}(\bar{x})$ and $\bar{s} = |\bar{S}|$

- there exist $\gamma > 0$ and $\delta' \in (0, 0.2)$ such that $\gamma > \frac{1+\delta'}{1-\delta'}$ and

$$\lambda_{\mathsf{tgt}} \geq 4 \max\left\{2,\ \frac{\gamma+1}{(1-\delta')\gamma - (1+\delta')}\right\} \|A^T z\|_\infty$$

Convergence analysis: assumptions

suppose $b = A\bar{x} + z$; let $\bar{S} = \text{supp}(\bar{x})$ and $\bar{s} = |\bar{S}|$

- there exist $\gamma > 0$ and $\delta' \in (0, 0.2)$ such that $\gamma > \frac{1+\delta'}{1-\delta'}$ and

$$\lambda_{\text{tgt}} \geq 4 \max\left\{2,\ \frac{\gamma + 1}{(1-\delta')\gamma - (1+\delta')}\right\} \|A^T z\|_\infty$$

- there exists an integer $\tilde{s}$ such that $\rho_-(A, \bar{s} + 2\tilde{s}) > 0$ and

$$\tilde{s} \ > \ \frac{8\big(\gamma_{\text{inc}}\rho_+(A, \bar{s} + 2\tilde{s}) + \rho_+(A, \tilde{s})\big)}{\rho_-(A, \bar{s} + \tilde{s})}(1+\gamma)\bar{s}.$$

Convergence analysis: assumptions

suppose $b = A\bar{x} + z$; let $\bar{S} = \mathrm{supp}(\bar{x})$ and $\bar{s} = |\bar{S}|$

- there exist $\gamma > 0$ and $\delta' \in (0, 0.2)$ such that $\gamma > \frac{1+\delta'}{1-\delta'}$ and

$$\lambda_{\mathsf{tgt}} \geq 4 \max\left\{2,\ \frac{\gamma + 1}{(1-\delta')\gamma - (1+\delta')}\right\} \|A^T z\|_\infty$$

- there exists an integer $\tilde{s}$ such that $\rho_-(A, \bar{s} + 2\tilde{s}) > 0$ and

$$\tilde{s} \ > \ \frac{8\big(\gamma_{\mathsf{inc}}\rho_+(A, \bar{s} + 2\tilde{s}) + \rho_+(A, \tilde{s})\big)}{\rho_-(A, \bar{s} + \tilde{s})}(1+\gamma)\bar{s}.$$

for example, if RIP is satisfied for $\nu = 0.1$ at $s = 45(1+\gamma)\bar{s}$,

$$(1-\nu)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\nu)\|x\|_2^2, \quad \forall x : \|x\|_0 \leq s$$

then can take $\gamma_{\mathsf{inc}} = 1.2$ and $\tilde{s} = 22(1+\gamma)\bar{s}$

Convergence analysis: local results

theorem: suppose previous assumptions hold and $x^{(0)}$ satisfies

$$
\begin{aligned}
\left\|x^{(0)}_{\bar{S}^c}\right\|_0 &\leq \tilde{s} && \text{(sparse)} && (*)\\
\omega_\lambda(x^{(0)}) &\leq \delta'\lambda && \text{(close to optimal)} && (**)
\end{aligned}
$$

then for all $k > 0$,

$$
\begin{aligned}
\left\|x^{(k)}_{\bar{S}^c}\right\|_0 &\leq \tilde{s}\\
\phi_\lambda(x^{(k)}) - \phi^\star_\lambda &\leq \left(1 - \frac{1}{4\gamma_{\mathsf{inc}}\,\kappa(A, \bar{s} + 2\tilde{s})}\right)^k \left(\phi_\lambda(x^{(0)}) - \phi^\star_\lambda\right)
\end{aligned}
$$

Convergence analysis: global rate

theorem: suppose previous assumptions hold, and η and δ satisfy

$$\frac{1+\delta}{1+\delta'} \leq \eta < 1,$$

then

- $(*)$ holds for the starting points of each λ_K, and number of iterations for each intermediate λ_K is no more than

$$\ln\left(\frac{C}{\delta^2}\right) \Big/ \ln\left(1 - \frac{1}{4\gamma_{\mathsf{inc}}\kappa}\right)^{-1} \qquad \text{(independent of } \lambda_K\text{)}$$

- for $K = 1, \ldots, N-1$,

$$\phi_{\lambda_{\mathrm{tgt}}}(\hat{x}^{(K)}) - \phi^\star_{\lambda_{\mathrm{tgt}}} \leq \eta^{2(K+1)} \frac{5(1+\gamma)\lambda_0^2 \bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}$$

- total number of iterations is $O\left(\kappa(A, \bar{s}+2\tilde{s}) \ln\left(\frac{\lambda_0}{\epsilon}\right)\right)$

Theory versus practice

RIP-like condition

- more restrictive than ("static") conditions on sparse recovery
- depends on algorithmic parameters γ_{inc}, δ and η
- hard to estimate for choosing parameters in practice

Theory versus practice

RIP-like condition

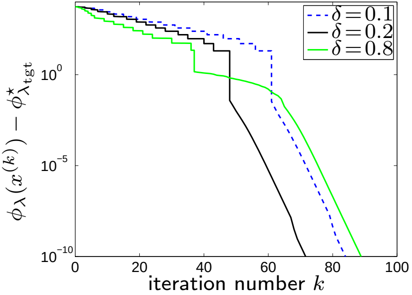
- more restrictive than (“static”) conditions on sparse recovery
- depends on algorithmic parameters γ_{inc}, δ and η
- hard to estimate for choosing parameters in practice

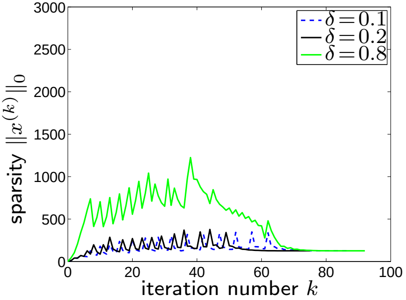
in practice

- two most effective rules supported by theory:
  - geometric decrease of $\lambda_{K+1} = \eta\lambda_K$
  - proportional precision $\hat{\epsilon}_K = \delta\lambda_K$
- best choices of δ and η may not satisfy our conditions
  - geometric convergence at each stages may not be necessary
  - important to start final stage within fast convergence zone

Effects of varying δ

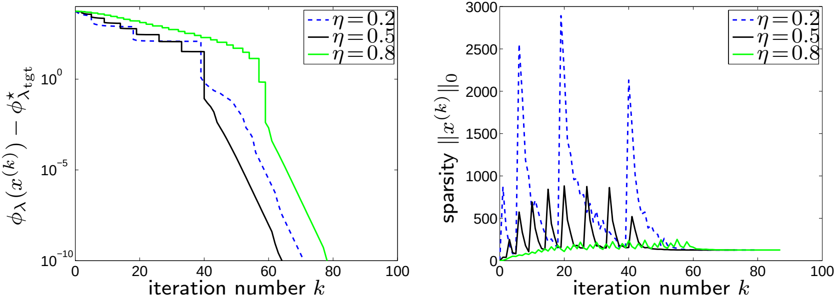
fixed $\eta = 0.7$



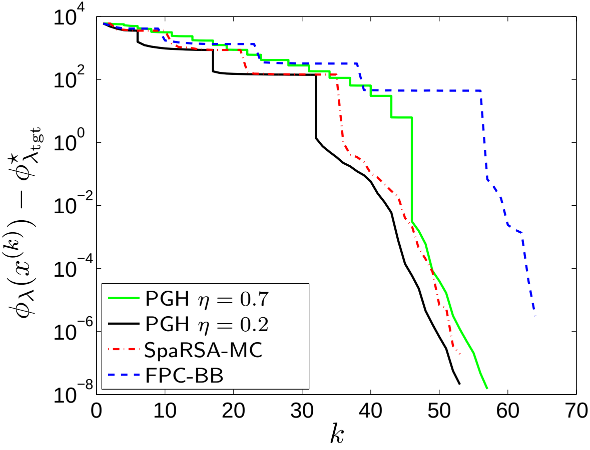


Effects of varying η

fixed $\delta = 0.2$



Comparison with SpaRSA and FPC



SpaRSA: Wright, Nowak and Figueiredo (09)

FPC: Hale, Yin and Zhang (08)

Outline

- background: first-order methods and their complexities
- proximal-gradient (PG) method + homotopy
- **accelerated proximal gradient (APG) method + homotopy**
- numerical experiments and summary

What can we expect?

iteration complexity for: $\underset{x}{\text{minimize}} \quad \left\{ \phi(x) \triangleq f(x) + \Psi(x) \right\}$

class of f	smooth	smooth + strongly convex
PG	$O\left(\frac{L}{\epsilon}\right)$	$O\left(\frac{L}{\mu} \log\left(\frac{1}{\epsilon}\right)\right)$ (not require μ)
accelerated PG	$O\left(\sqrt{\frac{L}{\epsilon}}\right)$	$O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{1}{\epsilon}\right)\right)$ (require μ)

by exploiting **restricted strong convexity** of ℓ_1-LS problem

- PG + homotopy can achieve: $O\left(\kappa(A, s) \log\left(\frac{1}{\epsilon}\right)\right)$

What can we expect?

iteration complexity for: $\underset{x}{\text{minimize}} \quad \left\{\phi(x) \triangleq f(x) + \Psi(x)\right\}$

class of f	smooth	smooth + strongly convex
PG	$O\left(\frac{L}{\epsilon}\right)$	$O\left(\frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$ (not require μ)
accelerated PG	$O\left(\sqrt{\frac{L}{\epsilon}}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$ (require μ)

by exploiting **restricted strong convexity** of ℓ_1-LS problem

- PG + homotopy can achieve: $O\left(\kappa(A,s)\log\left(\frac{1}{\epsilon}\right)\right)$
- accelerated PG + homotopy: $O\left(\sqrt{\kappa(A,s')}\log\left(\frac{1}{\epsilon}\right)\right)$?

APG methods without knowledge of μ_f

basic strategy: **restart**

- simple schemes that do not estimate μ_f explicitly
  - restart FISTA periodically (not very sensitive to period)
  - restart FISTA whenever objective increases

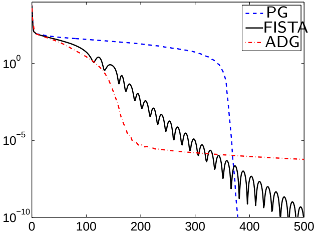
- restart based on adaptive estimation of μ_f
  - admits rigorous complexity analysis
  - involves some extra overheads

Restart FISTA

a simple variant of FISTA (Beck & Teboulle 2008):

$$y^{(k)} = x^{(k)} + \frac{k-1}{k+2}(x^{(k)} - x^{(k-1)})$$

$$x^{(k+1)} = \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right)$$



two simple schemes that work very well in practice:

- restart whenever $f(x^{(k)}) > f(x^{(k-1)})$
- restart whenever $\nabla f(y^{(k-1)})^T(x^{(k)} - x^{(k-1)}) > 0$

(recent analysis by O'Donoghue & Candès 2012)

Restart based on adaptive estimation of μ_f

estimate μ_f by measuring reduction of norm of *gradient mapping*

- first proposed by Nesterov (2007)
  - in the context of his accelerated dual gradient (ADG) method
  - complexity $\sqrt{\kappa_f}\log(\kappa_f)\log(1/\epsilon)$, where $\kappa_f = L_f/\mu_f$
- we will focus on Nesterov's constant step scheme III (2004):

$$
\begin{aligned}
y^{(k)} &= x^{(k)} + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(x^{(k)} - x^{(k-1)}) \\
x^{(k+1)} &= \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right)
\end{aligned}
$$

our scheme also includes a line search procedure for tuning L

Accelerated line search (on L)

Algorithm:

$\{x^{(k+1)}, M_k, \alpha_k\} \leftarrow \texttt{AccelLineSearch}(x^{(k)}, x^{(k-1)}, L_k, \mu, \alpha_{k-1})$

---

repeat

$$L \leftarrow L\gamma_{\mathsf{inc}}$$

$$\alpha_k \leftarrow \sqrt{\frac{\mu}{L}}$$

$$y^{(k)} \leftarrow x^{(k)} + \frac{\alpha^{(k)}(1-\alpha^{(k-1)})}{\alpha^{(k-1)}(1+\alpha^{(k)})}(x^{(k)} - x^{(k-1)})$$

$$x^{(k+1)} \leftarrow \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right)$$

until: $\phi(x^{(k+1)}) \le \psi_L(y^{(k)}; x^{(k+1)})$

$M_k \leftarrow L$

(here μ is an estimate of μ_f, not necessarily less than μ_f)

APG method with line search (on L)

Algorithm: $\{\hat{x}, \hat{M}\} \leftarrow \texttt{scAPG}(x^{(0)}, L_0, \hat{\epsilon})$

parameters: $L_{\mathsf{min}} \geq \mu > 0$, $\gamma_{\mathsf{dec}} \geq 1$, $\gamma_{\mathsf{inc}} \geq 1$

$x^{(-1)} \leftarrow x^{(0)}$, $\alpha_{-1} = 1$

repeat for $k = 0, 1, 2, \ldots$

$\quad \{x^{(k+1)}, M_k, \alpha_k\} \leftarrow \texttt{AccelLineSearch}(x^{(k)}, x^{(k-1)}, L_k, \mu, \alpha_{k-1})$

$\quad L_{k+1} \leftarrow \max\{L_{\mathsf{min}}, M_k/\gamma_{\mathsf{dec}}\}$

until $\omega(x^{(k+1)}) \leq \hat{\epsilon}$

$\hat{x} \leftarrow x^{(k+1)}$, $\hat{M} \leftarrow M_k$

(important to keep $L_{\mathsf{min}} \geq \mu$, especially in the case of $\mu \geq \mu_f$)

Convergence results for $0 < \mu \le \mu_f$

theorem: let $x^\star$ is the minimizer of $\phi(x) \triangleq f(x) + \Psi(x)$, and suppose $0 < \mu \le \mu_f$, then Algorithm `scAPG` guarantees that

$$\begin{aligned} \phi(x^{(k)}) - \phi(x^\star) &\le \tau_k \left[\phi(x^{(0)}) - \phi(x^\star) + \frac{\mu}{2}\|x^{(0)} - x^\star\|^2 \right] \\ \frac{\mu}{2}\|y^{(k)} - x^\star\|_2^2 &\le \tau_k \left[\phi(x^{(0)}) - \phi(x^\star) + \frac{\mu}{2}\|x^{(0)} - x^\star\|^2 \right] \end{aligned}$$

where

$$\tau_k = \begin{cases} 1 & k = 0 \\ \prod_{i=0}^{k-1} (1 - \alpha_i) & k \ge 1 \end{cases}$$

moreover

$$\tau_k \le \left(1 - \sqrt{\frac{\mu}{L_f \gamma_{\mathsf{inc}}}} \right)^k$$

The non-blowout property

lemma: suppose $0 < \mu \leq L_{\mathsf{min}}$, then Algorithm `scAPG` guarantees

$$\phi(x^{(k+1)}) \leq \phi(x^{(k)}) + \frac{M_{k-1}}{2}\left\|x^{(k)} - x^{(k-1)}\right\|^2 - \frac{M_k}{2}\left\|x^{(k+1)} - x^{(k)}\right\|^2$$

corollary: suppose $0 < \mu \leq L_{\mathsf{min}}$, then we have

$$\phi(x^{(k+1)}) \leq \phi(x^{(0)}) - \frac{M_k}{2}\left\|x^{(k+1)} - x^{(k)}\right\|^2$$

- holds for both situations: $0 < \mu \leq \mu_f$ and $\mu > \mu_f$
- useful for adaptive estimation of μ_f
- critical in analysis of homotopy method: maintain sparsity

Composite gradient mapping

analogue of gradient for composite objective $\phi(x) = f(x) + \Psi(x)$

$$g_L(x) = L\left(x - \mathbf{prox}_{\frac{1}{L}\Psi}\left(x - \frac{1}{L}\nabla f(x)\right)\right)$$

proximal gradient method can be written as

$$x^+ = \mathbf{prox}_{\frac{1}{L}\Psi}\left(x - \frac{1}{L}\nabla f(x)\right) = x - \frac{1}{L}g_L(x)$$

- if $\Psi \equiv 0$, then $g_L(x) = \nabla\phi(x) = \nabla f(x)$ for any $L > 0$
- in general, $g_L(x) \in \nabla f(x) + \partial\Psi\left(x - \frac{1}{L}g_L(x)\right)$
- $g_L(x) = 0$ if and only if x minimizes $\phi(x) = f(x) + \Psi(x)$

Reduction of gradient mapping

lemma: suppose $0 < \mu \leq \mu_f$ and $x^{(0)}$ in `scAPG` is computed by

$$\{x^{(0)}, M_{-1}, g^{(-1)}, S_{-1}\} \leftarrow \texttt{LineSearch}\,(x^{(-1)}, L_{-1})$$

with arbitrary $x^{(-1)} \in \mathbb{R}^n$ and $L_{-1} \geq L_{\text{min}}$, then for any $k \geq 0$

$$\left\| g_{M_k}(y^{(k)}) \right\|_2 \leq 2\sqrt{2\tau_k}\,\frac{M_k}{\mu}\left(1 + \frac{S_{-1}}{M_{-1}}\right)\left\| g^{(-1)} \right\|_2$$

Reduction of gradient mapping

lemma: suppose $0 < \mu \le \mu_f$ and $x^{(0)}$ in `scAPG` is computed by

$$\{x^{(0)}, M_{-1}, g^{(-1)}, S_{-1}\} \leftarrow \texttt{LineSearch}\,(x^{(-1)}, L_{-1})$$

with arbitrary $x^{(-1)} \in \mathbb{R}^n$ and $L_{-1} \ge L_{\min}$, then for any $k \ge 0$

$$\left\| g_{M_k}(y^{(k)}) \right\|_2 \le 2\sqrt{2\tau_k}\,\frac{M_k}{\mu}\left(1 + \frac{S_{-1}}{M_{-1}}\right)\left\|g^{(-1)}\right\|_2$$

where S_{-1} is a local estimate of Lipschitz constant

$$S_{-1} = \frac{\|\nabla f(x^{(0)}) - \nabla f(x^{(-1)})\|_2}{\|x^{(0)} - x^{(-1)}\|_2}$$

can easily compute $S(y^{(k)})$ based on line search result

AdapAPG: restart based on estimation of μ_f

two conditions to test:

A: $\left\|g_{M_k}(y^{(k)})\right\| \leq \theta \left\|g^{(-1)}\right\|$ (reduction factor: $0 < \theta < 1$)

B: $2\sqrt{2\tau_k}\,\frac{M_k}{\mu}\left(1 + \frac{S_{-1}}{M_{-1}}\right) \leq \theta$

- if A is satisfied first, then restart with

$$x^{(0)} \leftarrow x^{(k+1)},\ g^{(-1)} \leftarrow g_{M_k}(y^{(k)}),\ M_{-1} \leftarrow M_k,\ S_{-1} \leftarrow S(y^{(k)})$$

- if B is satisfied first (indicating $\mu > \mu_f$), then restart with

$$\mu \leftarrow \mu/10$$

AdapAPG: restart based on estimation of μ_f

two conditions to test:

A: $\left\| g_{M_k}(y^{(k)}) \right\| \leq \theta \left\| g^{(-1)} \right\|$ (reduction factor: $0 < \theta < 1$)

B: $2\sqrt{2\tau_k}\,\frac{M_k}{\mu}\left(1 + \frac{S_{-1}}{M_{-1}}\right) \leq \theta$

- if A is satisfied first, then restart with

$$x^{(0)} \leftarrow x^{(k+1)},\ g^{(-1)} \leftarrow g_{M_k}(y^{(k)}),\ M_{-1} \leftarrow M_k,\ S_{-1} \leftarrow S(y^{(k)})$$

- if B is satisfied first (indicating $\mu > \mu_f$), then restart with

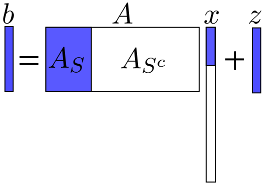
$$\mu \leftarrow \mu/10$$

overall complexity: $O\left(\sqrt{\kappa_f}\log(\kappa_f)\log\left(\frac{1}{\hat{\epsilon}}\right)\right) + O\left(\sqrt{\kappa_f}\log(\kappa_f)\right)$

Review: structure of ℓ_1-LS problem

suppose optimal solution is sparse

$$x^\star = \begin{bmatrix} x^\star_S \\ x^\star_{S^c} \end{bmatrix} = \begin{bmatrix} x^\star_S \\ 0 \end{bmatrix}$$



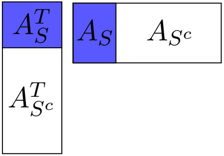
- restricted smoothness:

$$L_S = \lambda_{\max}(A_S^T A_S) < \lambda_{\max}(A^T A)$$

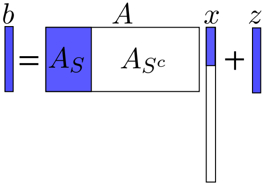
- restricted strong convexity:

$$\mu_S = \lambda_{\min}(A_S^T A_S) > 0$$

$\nabla^2 f(x) =$



Review: structure of ℓ_1-LS problem



suppose optimal solution is sparse

$$x^\star = \begin{bmatrix} x^\star_S \\ x^\star_{S^c} \end{bmatrix} = \begin{bmatrix} x^\star_S \\ 0 \end{bmatrix}$$

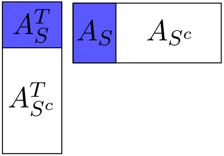
- restricted smoothness:

$$L_S = \lambda_{\max}(A_S^T A_S) < \lambda_{\max}(A^T A)$$

- restricted strong convexity:

$$\mu_S = \lambda_{\min}(A_S^T A_S) > 0$$

$\nabla^2 f(x) =$



solution: `AdapAPG` $+$ homotopy continuation

AdapAPG + homotopy continuation

parameters: $\eta \in (0,1)$, $\delta \in (0,1)$

Algorithm: $\hat{x}^{(\text{tgt})} \leftarrow \texttt{Homotopy}(A, b, \lambda_{\text{tgt}}, \epsilon)$

initialize: $\lambda_0 \leftarrow \|A^T b\|_\infty$, $\hat{x}^{(0)} \leftarrow 0$

$N \leftarrow \lfloor \ln(\lambda_0/\lambda_{\text{tgt}}) \,/\, \ln(1/\eta) \rfloor$

repeat: for $K = 0, 1, \ldots, N-1$

$\lambda_{K+1} \leftarrow \eta \lambda_K$ (geometric decrease $\lambda_K = \eta^K \lambda_0$)

$\hat{\epsilon}_{K+1} \leftarrow \delta \lambda_{K+1}$ (low accuracy proportional to λ_K)

$x^{(K+1)} \leftarrow \texttt{AdapAPG}\big(\lambda_{K+1}, \hat{\epsilon}_{K+1}, \hat{x}^{(K)}\big)$

end

$\hat{x}^{(\text{tgt})} \leftarrow \texttt{AdapAGP}\big(\lambda_{\text{tgt}}, \epsilon, \hat{x}^{(N)}\big)$ (final stage for high accuracy)

Convergence analysis: assumptions

suppose $b = A\bar{x} + z$; let $\bar{S} = \text{supp}(\bar{x})$ and $\bar{s} = |\bar{S}|$

- there exist $\gamma > 0$ and $\delta' \in (0, 0.2)$ such that $\gamma > \frac{1+\delta'}{1-\delta'}$ and

$$\lambda_{\text{tgt}} \geq 4 \max\left\{2,\ \frac{\gamma+1}{(1-\delta')\gamma - (1+\delta')}\right\} \|A^T z\|_\infty$$

- there exists an integer $\tilde{s}$ such that $\rho_-(A, \bar{s} + 3\tilde{s}) > 0$ and

$$\tilde{s} \ > \ \frac{24\big(\gamma_{\text{inc}}\rho_+(A, \bar{s} + 3\tilde{s}) + \rho_+(A, \tilde{s})\big)}{\rho_-(A, \bar{s} + \tilde{s})}(1+\gamma)\bar{s}.$$

(similar to conditions for PGH, but with slightly larger constants)

Convergence results

local: suppose previous assumptions hold and $x^{(0)}$ satisfies

$$\left\|x_{\bar{S}^c}^{(0)}\right\|_0 \le \tilde{s}, \qquad \omega_\lambda(x^{(0)}) \le \delta'\lambda$$

then for all $k > 0$,

$$\begin{aligned} \left\|x_{\bar{S}^c}^{(k)}\right\|_0 &\le \tilde{s} \\ \phi_\lambda(x^{(k)}) - \phi_\lambda^\star &\le \left(1 - \frac{1}{4\gamma_{\mathsf{inc}}\,\kappa(A, \bar{s} + 3\tilde{s})}\right)^k \left(\phi_\lambda(x^{(0)}) - \phi_\lambda^\star\right) \end{aligned}$$

Convergence results

local: suppose previous assumptions hold and $x^{(0)}$ satisfies

$$\left\|x^{(0)}_{\bar{S}^c}\right\|_0 \;\le\; \tilde{s}, \qquad \omega_\lambda(x^{(0)}) \;\le\; \delta'\lambda$$

then for all $k > 0$,

$$\begin{aligned} \left\|x^{(k)}_{\bar{S}^c}\right\|_0 \;&\le\; \tilde{s} \\ \phi_\lambda(x^{(k)}) - \phi_\lambda^\star \;&\le\; \left(1 - \frac{1}{4\gamma_{\mathsf{inc}}\,\kappa(A, \bar{s} + 3\tilde{s})}\right)^k \left(\phi_\lambda(x^{(0)}) - \phi_\lambda^\star\right) \end{aligned}$$

global: if δ and η are chosen such that $\dfrac{1+\delta}{1+\delta'} \le \eta < 1$, then the total number of iterations is $O\left(\sqrt{\kappa}\log(\kappa)\log\left(\frac{\lambda_0}{\epsilon}\right)\right)$

Numerical experiments: setup

$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$$

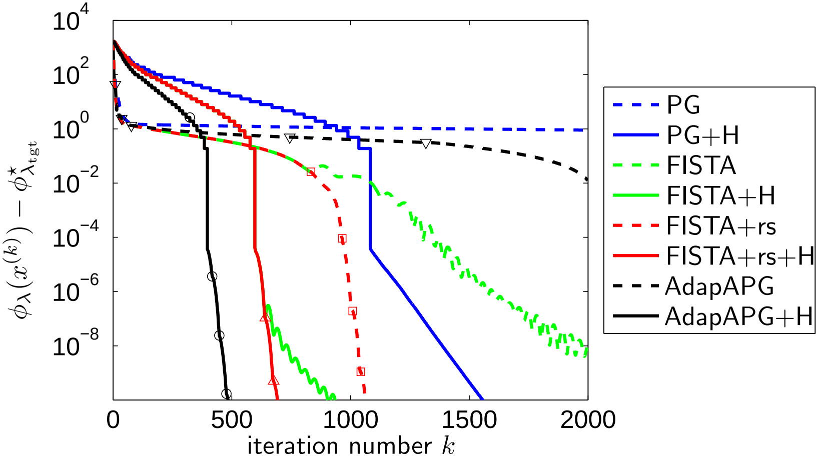
- generate A following experiments in Agarwal et al. (2011)
  - first generate $B \in \mathbb{R}^{m,n}$ with $B_{ij} \sim$ i.i.d. standard Gaussian
  - choose $\omega \in [0,1)$ and for $i = 1,\ldots,m$, generate row $A_{i,:}$ as

$$\begin{array}{rcl} A_{i,1} & = & B_{i,1}/\sqrt{1-\omega^2} \\ A_{i,j+1} & = & \omega A_{i,j} + B_{i,j}, \quad j = 2,\ldots,n \end{array}$$

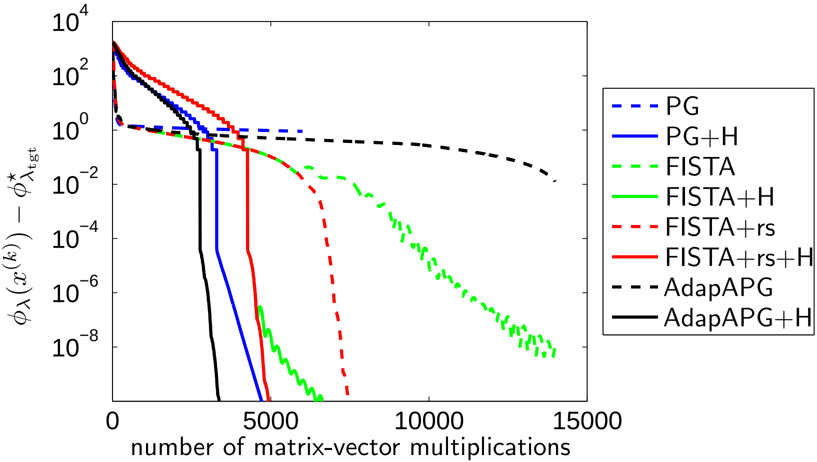
  eigenvalues of $\mathbf{E}[A^TA]$ lie within $\left[\frac{1}{(1+\omega)^2}, \frac{2}{(1-\omega)^2(1+\omega)}\right]$

- parameters: $\eta = 0.8$, $\delta = 0.2$, and initialize $\mu = L_0/100$

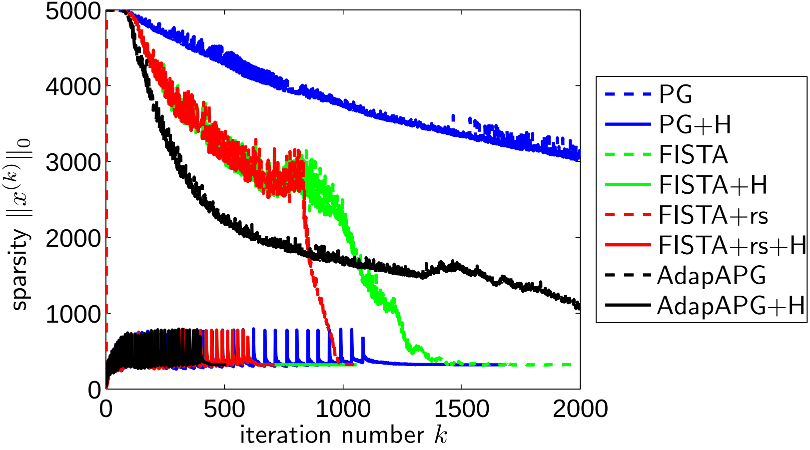
Numerical experiments ($\omega = 0.9$)



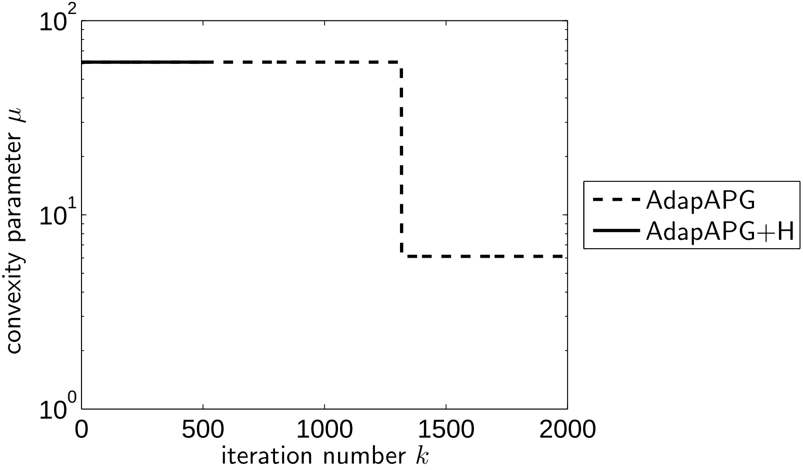
Numerical experiments



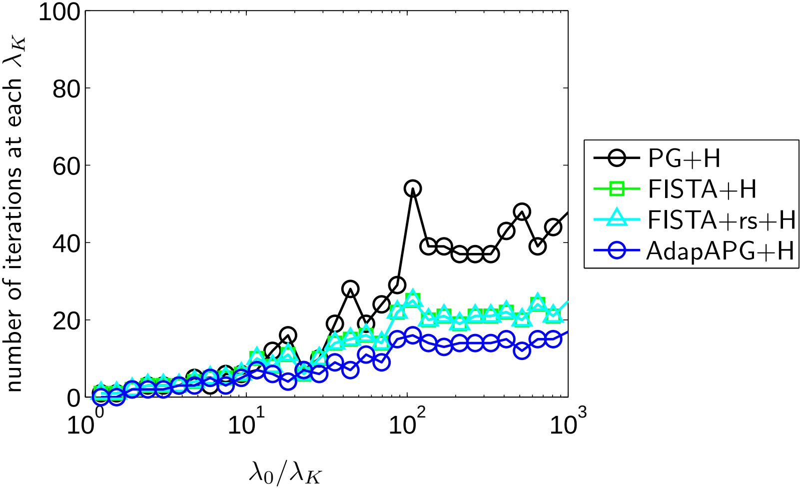
Numerical experiments



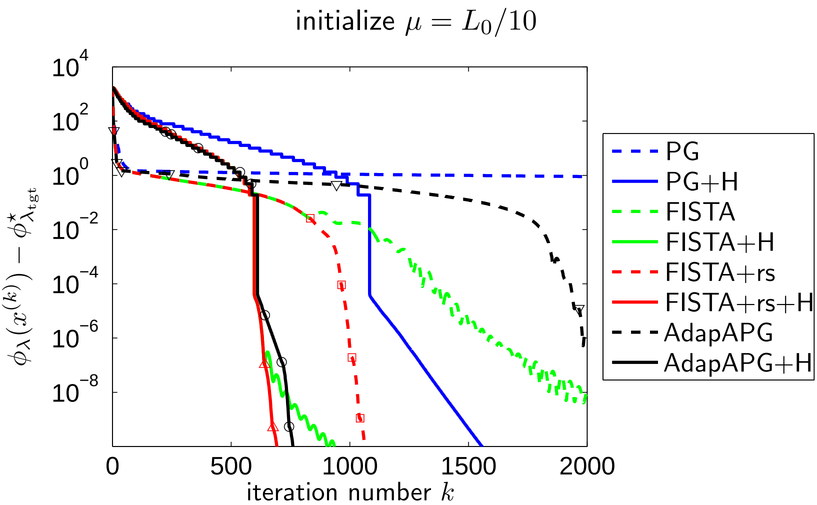
Numerical experiments



Numerical experiments

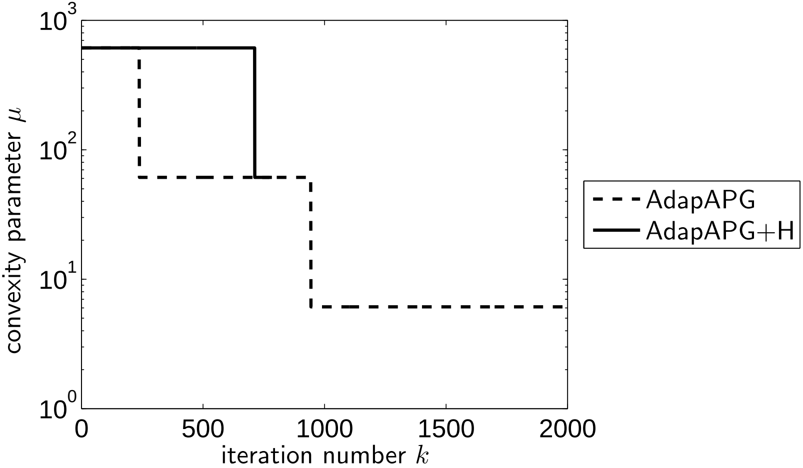


Numerical experiments



Numerical experiments

initialize $\mu = L_0/10$



Outline

- background: first-order methods and their complexities
- proximal-gradient (PG) method + homotopy
- accelerated proximal gradient (APG) method + homotopy
- **summary**

Summary

computational complexities for the sparse least-squares problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

numerical methods	cost per iteration	iteration complexity
interior-point methods	$O\left(m^2 n\right)$	$O\left(\sqrt{n}\log\left(\frac{1}{\epsilon}\right)\right)$
proximal-gradient (PG)	$O\left(mn\right)$	$O\left(\frac{1}{\epsilon}\right)$
accelerated PG	$O\left(mn\right)$	$O\left(\frac{1}{\sqrt{\epsilon}}\right)$
PG + homotopy	$O\left(mn\right)$	$O\left(\kappa(A,s)\log\left(\frac{1}{\epsilon}\right)\right)$
APG + homotopy	$O\left(mn\right)$	$O\left(\sqrt{\kappa}\log(\kappa)\log\left(\frac{1}{\epsilon}\right)\right)$