A hybrid quasi-Newton projected-gradient method with application to Lasso and basis-pursuit denoise

Ewout van den Berg

Human Language Technologies Group
IBM T.J. Watson Research Center

Work done at the Department of Statistics
Stanford University

October 10, 2014

This work was partially supported by National Science Foundation Grant DMS 0906812
(American Reinvestment and Recovery Act).

Background

Basis pursuit denoise

$$\underset{x}{\text{minimize}} \quad \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 \leq \sigma$$

SPGL1 reduces this by to a series of Lasso problems [B, Friedlander, 2008]

$$\underset{x}{\text{minimize}} \quad \tfrac{1}{2}\|Ax - b\|_2^2 \quad \text{subject to} \quad \|x\|_1 \leq \tau$$

Root finding with $\tau^+ = \tau + (\|r\|_2^2 - \sigma\|r\|)/\|A^T r\|_\infty$

Background

Basis pursuit denoise

$$\underset{x}{\text{minimize}} \quad \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 \leq \sigma$$

SPGL1 reduces this by to a series of Lasso problems [B, Friedlander, 2008]

$$\underset{x}{\text{minimize}} \quad \tfrac{1}{2}\|Ax - b\|_2^2 \quad \text{subject to} \quad \|x\|_1 \leq \tau$$

Root finding with $\tau^+ = \tau + (\|r\|_2^2 - \sigma\|r\|)/\|A^T r\|_\infty$

Background

$$\underset{x}{\text{minimize}} \quad \tfrac{1}{2}\|Ax - b\|_2^2 \quad \text{subject to} \quad \|x\|_1 \leq \tau$$

General form

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad x \in \mathcal{C}$$

Solved using spectral projected-gradient (SPG) method:

$$\begin{aligned} d &= -\nabla f(x) \cdot \beta \\ x^+ &= \mathcal{P}(x + \alpha d) \end{aligned} \qquad \textit{or} \qquad \begin{aligned} d &= \mathcal{P}(x - \nabla f(x) \cdot \beta) - x \\ x^+ &= x + \alpha d \end{aligned}$$

With

- β: Barzilai-Borwein scaling parameter [Barzilai,Borwein,1988]
- α: Step length from non-monotone line search [Birgin et al., 2000]
- $\mathcal{P}$: Orthogonal projection onto $\mathcal{C}$

$$\mathcal{P}(x) := \underset{v}{\text{argmin}} \quad \|x - v\|_2 \quad \text{subject to} \quad v \in \mathcal{C}$$

Motivation

Observation

- (Sometimes) difficult to get a highly accurate solution
- Iterates remain on the same face of $\mathcal{C}$ (same sign pattern)
- Very little progress

Motivation

Observation

- (Sometimes) difficult to get a highly accurate solution
- Iterates remain on the same face of $\mathcal{C}$ (same sign pattern)
- Very little progress

Typical solution

- Detect stagnation on a fixed face
- Solve problem constrained to the given face
- Check optimality for global problem
- Resume if not optimal

Motivation

Observation

- (Sometimes) difficult to get a highly accurate solution
- Iterates remain on the same face of $\mathcal{C}$ (same sign pattern)
- Very little progress

Typical solution

- Detect stagnation on a fixed face
- Solve problem constrained to the given face
- Check optimality for global problem
- Resume if not optimal

Difficulties

- When to initiate this procedure?
- Solving subproblem on incorrect face is wasteful
- Waiting too long defeats the purpose

Outline

1. Propose a new hybrid method for polyhedral $\mathcal{C}$
   (Practical only for simple $\mathcal{C}$: ℓ_1, bound constrained, simplex)
2. Convergence of the method
3. Application to Lasso and basis pursuit

Hybrid method

Basic idea

- Take regular SPG steps by default
- After each iteration, check whether $\mathcal{F}(x^+) = \mathcal{F}(x)$ $(\neq \mathcal{C})$
- Initialize or update L-BFGS model
- Use quasi-Newton search direction in next iteration

Hybrid method

Basic idea

- Take regular SPG steps by default
- After each iteration, check whether $\mathcal{F}(x^+) = \mathcal{F}(x)$ $(\neq \mathcal{C})$
- Initialize or update L-BFGS model
- Use quasi-Newton search direction in next iteration

Some issues

- Quasi-Newton direction cannot simply be projected onto $\mathcal{C}$
- Naive implementation ignores problem structure

Hybrid method

Basic idea

- Take regular SPG steps by default
- After each iteration, check whether $\mathcal{F}(x^+) = \mathcal{F}(x)$ $(\neq \mathcal{C})$
- Initialize or update L-BFGS model
- Use quasi-Newton search direction in next iteration

Some issues

- Quasi-Newton direction cannot simply be projected onto $\mathcal{C}$
- Naive implementation ignores problem structure

Solution

- Form an L-BFGS model restricted to the current face
- Capture only relevant information

Reduced L-BFGS model

Local function

- We only want to model $f(x)$ over the current d-face $\mathcal{F}$
- Find an orthonormal basis $B \in \mathbb{R}^{n \times d}$ for $\mathrm{lin}(\mathcal{F} - \mathcal{F})$
- Define $\bar{f}(c) : \mathbb{R}^d \to \mathbb{R}$ for some fixed $x_0 \in \mathcal{F}$

$$\bar{f}(c) = f(x_0 + Bc)$$

- Choosing $c = B^T(x - x_0)$ gives $\bar{f}(c) = f(x)$ for $x \in \mathcal{F}$

Reduced L-BFGS model

Local function

- We only want to model $f(x)$ over the current d-face $\mathcal{F}$
- Find an orthonormal basis $B \in \mathbb{R}^{n \times d}$ for $\text{lin}(\mathcal{F} - \mathcal{F})$
- Define $\bar{f}(c) : \mathbb{R}^d \to \mathbb{R}$ for some fixed $x_0 \in \mathcal{F}$

$$\bar{f}(c) = f(x_0 + Bc)$$

- Choosing $c = B^T(x - x_0)$ gives $\bar{f}(c) = f(x)$ for $x \in \mathcal{F}$

Model updates

- Standard L-BFGS uses $s = x^+ - x$ and $y = \nabla f(x^+) - \nabla f(x)$
- We use $s = c^+ - c$, and $y = \nabla \bar{f}(c^+) - \nabla \bar{f}(c)$:

$$s = B^T(x^+ - x), \quad y = B^T(\nabla f(x^+) - \nabla f(x))$$

- Never need to choose x_0

Quasi-Newton search direction

Computing the search direction

- Want to compute search direction at current x
- Denote by H^{-1} the inverse approximate Hessian ($\mathbb{R}^{d \times d}$)
- In the reduced space we compute the search direction

$$\bar{d} = -H^{-1} \nabla \bar{f}(c) = -H^{-1} B^T \nabla f(x)$$

- Project back to ambient space using $B\bar{d}$:

$$d = -BH^{-1} B^T \nabla f(x)$$

Properties

- Search direction along the face: $(x + \alpha d) \in \mathcal{F}$ for $0 \leq \alpha \leq \alpha_{\text{max}}$
- Guaranteed descent direction

Self-projection cone

Remaining issues

- Quasi-Newton step must be restricted to the face ($\alpha \leq \alpha_{\mathsf{max}}$)
- Fall back to SPG step if line search fails (reset Hessian, history)
- Misses mechanism to avoid local minimum on relint($\mathcal{F}$)

Remaining issues

- Quasi-Newton step must be restricted to the face ($\alpha \leq \alpha_{\text{max}}$)
- Fall back to SPG step if line search fails (reset Hessian, history)
- Misses mechanism to avoid local minimum on relint($\mathcal{F}$)

Self-projection cone

- Update and use L-BFGS model only if $-\nabla f(x^+) \in \mathcal{S}(\mathcal{F}(x))$
- Where $\mathcal{S}(\mathcal{F}(x))$ is the self-projection cone of $\mathcal{F}(x)$:

$$
\begin{aligned}
\mathcal{S}(\mathcal{F}(x)) &:= \{d \in \mathbb{R}^n \mid \exists\, \alpha > 0 : \mathcal{F}[\mathcal{P}(x + \alpha d)] = \mathcal{F}(x)\} \\
&= \mathcal{N}(x) + \operatorname{lin}(\mathcal{F}(x) - \mathcal{F}(x))
\end{aligned}
$$

Convergence

Theorem

Let $f(x)$ be a twice continuously differentiable convex function that is bounded below and for which there exist constants $0 < \mu_1 \leq \mu_2 < \infty$ such that for all $x, v \in \mathbb{R}^n$

$$\mu_1 \|v\|_2^2 \leq v^T \nabla^2 f(x) v \leq \mu_2 \|v\|_2^2.$$

Then for any starting point $x_0 \in \mathcal{C}$, the sequence $\{x_k\}$ generated by the hybrid algorithm converges to the minimizer of $f(x)$ over $\mathcal{C}$.

Proof sketch:

- Finitely many quasi-Newton steps: done or SPG converges
- Infinitely many quasi-Newton steps:
- Successful quasi-Newton (L-BFGS) step (Liu and Nocedal):

$$f(x^+) - f(x^*) \leq (1 - c)(f(x) - f(x^*))$$

- Finite number of quasi-Newton steps on incorrect faces

Application to general problems

Challenges for general problems

- Projection in SPG is difficult for general $\mathcal{C}$
- Facial structure is often unknown
- Finding orthonormal basis for face may be expensive
- Even true for weighted ℓ_1 ball

Well suited for simple problems

- Cross polytope (ℓ_1-norm)
- Box constrained problems
- Simplex

Additional conditions

- Typically $A \in \mathbb{R}^{m \times n}$ with $m < n$
- Hessian not full rank for d-faces with $d > m$
- Use quasi-Newton steps only when $d \leq m$

Application to Lasso

Additional conditions

- Typically $A \in \mathbb{R}^{m \times n}$ with $m < n$
- Hessian not full rank for d-faces with $d > m$
- Use quasi-Newton steps only when $d \leq m$

Orthogonal projection

- Reduces to soft-thresholding, $\mathcal{O}(n \log n)$ complexity

Application to Lasso

Additional conditions

- Typically $A \in \mathbb{R}^{m \times n}$ with $m < n$
- Hessian not full rank for d-faces with $d > m$
- Use quasi-Newton steps only when $d \leq m$

Orthogonal projection

- Reduces to soft-thresholding, $\mathcal{O}(n \log n)$ complexity

Orthonormal basis

- Normalize signs and permute indices: $\mathcal{F} = \mathrm{conv}\{e_1, \ldots, e_{d+1}\}$
- Compute QR factorization of $[e_2 - e_1, \ldots, e_{d+1} - e_1]$:

$$Q_{i,j} = \begin{cases} -\sqrt{1/(j^2 + j)} & i \leq j \\ \sqrt{j/(j+1)} & i = j + 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Implicit B and B^T, can apply in $\mathcal{O}(n)$ time

Application to Lasso

Self-projection cone

- Let $d = -\nabla f(x)$ and define

$$\begin{aligned} \mathcal{I}_1 &= \{i \in [n] \mid (x_i > 0 \text{ and } d_i < 0) \text{ or } (x_i < 0 \text{ and } d_i > 0)\}, \\ \mathcal{I}_2 &= \{i \in [n] \mid (x_i > 0 \text{ and } d_i \geq 0) \text{ or } (x_i < 0 \text{ and } d_i \leq 0)\}, \\ \mathcal{I}_3 &= (\mathcal{I}_1 \cup \mathcal{I}_2)^c, \end{aligned}$$

- Set $s_j := \sum_{i \in \mathcal{I}_j} |d_i|$ and assume that $x \notin \operatorname{relint}(\mathcal{C})$, then

$$d \in \mathcal{S}(\mathcal{F}(x)) \quad \text{iff} \quad \begin{cases} s_1 = s_2 + s_3 \text{ and } s_3 = 0, \ \textit{or} \\ s_1 < s_2 + s_3 \text{ and } \max_{i \in \mathcal{I}_3} |d_i| \leq \dfrac{s_2 - s_1}{|\mathcal{I}_1 \cup \mathcal{I}_2|} \end{cases}$$

Application to Lasso

Self-projection cone

- Let $d = -\nabla f(x)$ and define

$$\begin{aligned}\mathcal{I}_1 &= \{i \in [n] \mid (x_i > 0 \text{ and } d_i < 0) \text{ or } (x_i < 0 \text{ and } d_i > 0)\},\\ \mathcal{I}_2 &= \{i \in [n] \mid (x_i > 0 \text{ and } d_i \geq 0) \text{ or } (x_i < 0 \text{ and } d_i \leq 0)\},\\ \mathcal{I}_3 &= (\mathcal{I}_1 \cup \mathcal{I}_2)^c,\end{aligned}$$

- Set $s_j := \sum_{i \in \mathcal{I}_j} |d_i|$ and assume that $x \notin \mathrm{relint}(\mathcal{C})$, then

$$d \in \mathcal{S}(\mathcal{F}(x)) \quad \text{iff} \quad \begin{cases} s_1 = s_2 + s_3 \text{ and } s_3 = 0, \ \textit{or} \\ s_1 < s_2 + s_3 \text{ and } \max\limits_{i \in \mathcal{I}_3} |d_i| \leq \dfrac{s_2 - s_1}{|\mathcal{I}_1 \cup \mathcal{I}_2|} \end{cases}$$

Line search

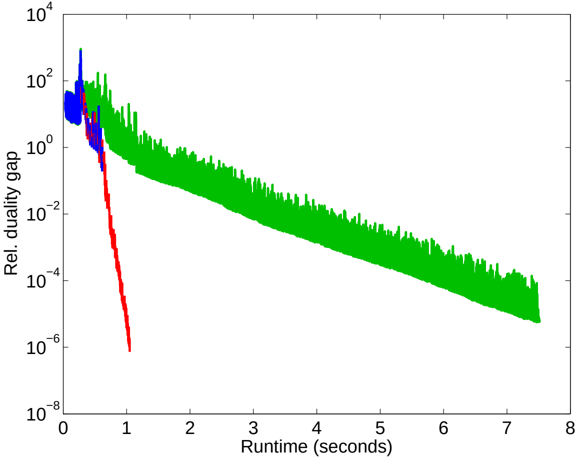
- Can compute maximum step length α_{max} to stay on face
- Objective is quadratic, can find minimum along search direction
- Can compute interval $[\alpha_{wmin}, \alpha_{wmax}]$ satisfying Wolfe conditions

Numerical experiments

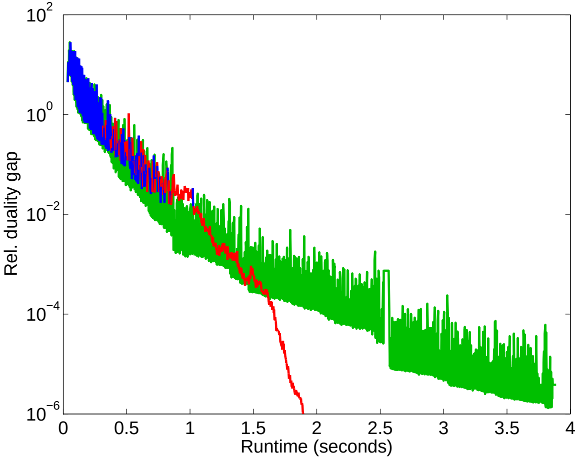
- 10 Sparco problems, each with three τ values [B et al., 2009]
- Random problems: A, $A + c$, $b = Ax$, b
- Heaviside matrix, random b

Numerical experiments

Heaviside matrix, random b

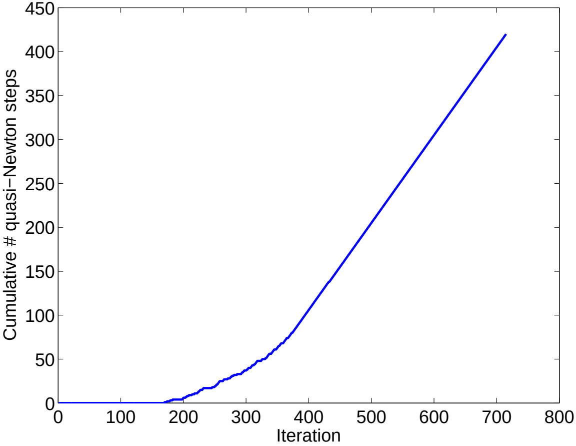


Random 300×800 A, random b



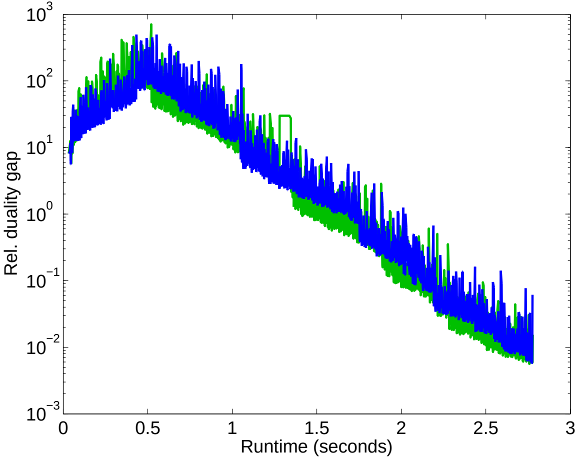
Numerical experiments

Random 300×800 A, random b



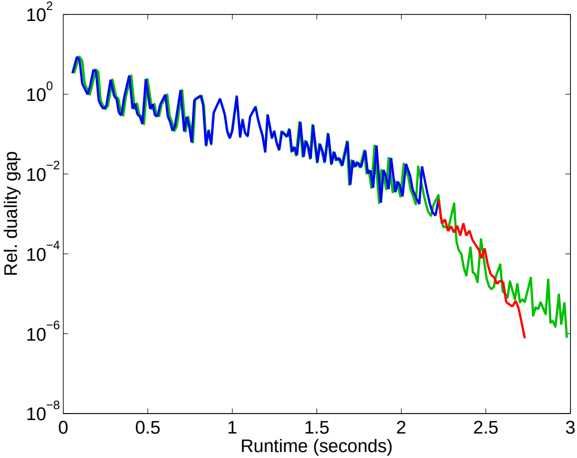
Numerical experiments

300×800 random + offset A, $b = Ax_0$, 50-sparse x_0



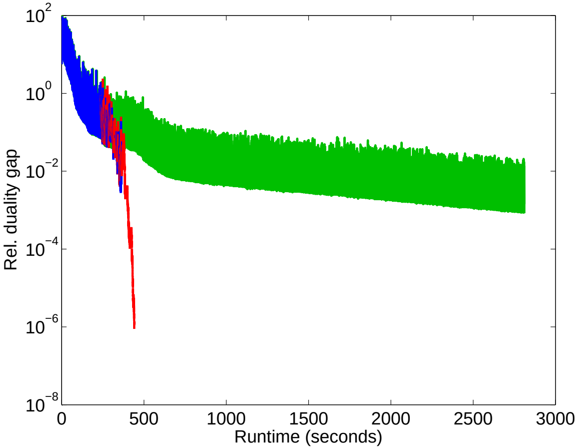
Numerical experiments

Sparco `blurspike`, $\tau \to \sigma \approx 0.1\|b\|_2$

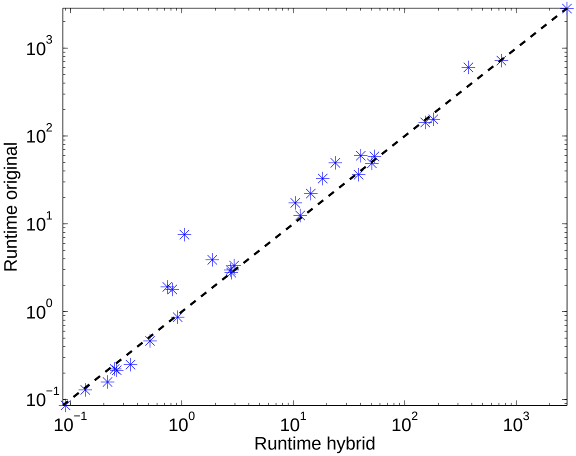


Numerical experiments

Sparco `p3poly`, $\tau \to \sigma \approx 10^{-3}\|b\|_2$



Numerical experiments



Numerical experiments

Lasso

- Sometimes the procedure is never used, small overhead
- Does well on problems that take longer to solve

Basis pursuit denoise

- SPGL1 has enthusiastic (aggressive) update strategy
- Subproblem terminated before quasi-Newton steps are taken
- Update strategy can lead to run-away behavior
- In those cases accurate solves with hybrid method can help

Conclusions

- Hybrid method shows encouraging results
- Apply to box-contrained problems

Reference

- J. Barzilai and J.M. Borwein, Two-point step size gradient methods, *IMA Journal of Numerical Analysis*, 8 (1988), pp. 141–148
- E.G. Birgin, J.M. Martínez, and M. Raydan, Nonmonotone spectral projected gradient methods on convex sets, *SIAM Journal on Optimization*, 10 (2000), pp. 1196–1211
- E. v.d. Berg and M.P. Friedlander, *Probing the Pareto frontier for basis pursuit solutions*, SIAM Journal on Scientific Computing, 2 (2008), pp. 890–912
- E. v.d. Berg, M.P. Friedlander, G. Hennenfent, F. Herrmann, R. Saab, and Ö. Yılmaz, Algorithm 890: Sparco: A testing framework for sparse reconstruction, *ACM Transactions on Mathematical Software*, 35 (2009), pp. 1–16