Advances in first-order methods: constraints, non-smoothness and faster convergence

Stephen Becker

Laboratoire Jacques-Louis Lions, UPMC/CNRS;
Fondation Sciences Mathématiques de Paris

May 20 2012

Joint work with:

Michael Grant (Caltech, CVX Research)
Emmanuel Candès (Stanford)
Jalal Fadili (GREYC-ENSICAEN)

Current state of first-order methods

Smooth unconstrained convex minimization

$$\min_x f(x) \qquad f \in \Gamma_0(\mathbb{R}^n) \cap C^1(\mathbb{R}^n) \quad \text{and} \quad \nabla f \text{ is } L-\text{Lipschitz}$$

Current tricks: limited memory quasi-Newton (e.g. L-BFGS), non-linear conjugate gradients. Results are very good.

Current state of first-order methods

Smooth unconstrained convex minimization

$$\min_x f(x) \qquad f \in \Gamma_0(\mathbb{R}^n) \cap C^1(\mathbb{R}^n) \quad \text{and} \quad \nabla f \text{ is } L-\text{Lipschitz}$$

Current tricks: limited memory quasi-Newton (e.g. L-BFGS), non-linear conjugate gradients. Results are very good.

Now consider $\min_x f(x) + \psi(x)$.

- $\psi(x) = \|x\|_1$ is non-smooth
- $\psi(x) = \iota_C(x) \equiv \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$ is indicator of closed convex set

e.g. $C = \{x : x \geq 0\}$ or $C = \{x : \|x\|_\infty \leq 1\}$

Results are not good. Worst-case behavior well-understood (Nemirovskii, Nesterov) and algorithms rarely exceed worst-case behavior.

Proximity operator: generalization of projection

Let $\psi \in \Gamma_0(\mathbb{R}^n)$ (proper, closed/lsc, convex functions)

Definition (Proximity operator)

$$\operatorname{prox}_f(y) = \operatorname*{argmin}_x \psi(x) + \frac{1}{2}\|x-y\|^2$$

Example

$\psi(x) = \iota_C(x) \equiv \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$ where C is closed, non-empty and convex. Then $\operatorname{prox}_C = \mathcal{P}_C$.

Efficient for many functions

Problem classes

Our interest is in "reasonable" non-smooth and constrained convex problems.
Assume still f is nice, $\psi_i \in \Gamma_0(\mathbb{R}^n)$, and prox_{ψ_i} is easy.

Definition (primal-dual class)

$$\min_x f(x) + \sum_i \psi_i(A_i x + b_i), \quad \psi_i \in \Gamma_0(\mathbb{R}^n)$$

Definition (primal class)

$$\min_x f(x) + \psi(x), \quad \psi \text{ separable}$$

Definition (generalized primal)

$$\min_x f(x) + \sum_i \psi_i(x), \quad \psi_i \text{ separable}$$

For the primal-dual class, there were no reasonable first-order methods *at all* until 2010. We show one such method.

For the primal class, we can solve, but *slowly*. Cannot apply quasi-Newton or non-linear CG. We show a quasi-Newton method.

Outline

Solving the primal-dual class

$$\min_x F(x) = f(x) + \sum_i \psi_i(A_i x + b_i)$$

Problems with primal approach

- prox_ψ is easy but $\operatorname{prox}_{\psi \circ A}$ is not
- prox_ψ is easy but $\operatorname{prox}_{\psi_1 + \psi_2}$ is not

Solving the primal-dual class

$$\min_x F(x) = f(x) + \sum_i \psi_i(A_i x + b_i)$$

Problems with primal approach

- prox_ψ is easy but $\mathrm{prox}_{\psi \circ A}$ is not
- prox_ψ is easy but $\mathrm{prox}_{\psi_1 + \psi_2}$ is not

Our approach (in "TFOCS" package: `tfocs.stanford.edu`)

- Solve via proximal point
  - $y_{k+1} = \mathrm{argmin}_x\, F(x) + \mu/2\|x - y_k\|^2$
- Solve sub-problem via a dual method
  - The strongly convex term makes the dual nice
- The dual problem is separable in ψ_i^*
- The A_i terms are now innocuous

Solving the primal-dual class: other approaches

Approach is simple, but... all methods are recent

- TFOCS (Becker, Candès, Grant 2010)
- relaxed Arrow-Hurwicz (Chambolle, Pock 2010)
  - extension (He, Yuan 2010; Condat 2011; Vũ 2011)
- monotone+skew (Briceño-Arias, Combettes 2011)
  - forward-backward-forward (Tseng 1998)
- product-space (Combettes, Pesquet 2011)

Chen, Teboulle 1994 is similar

Outline

1. Primal-dual class
2. Primal class

Solving the primal class

Goal is to speed up the problem $\min_x f(x) + \psi(x)$

Standard algorithm: proximal/projected gradient descent.

$$x_{k+1} = \operatorname*{argmin}_x \; Q_f(x; x_k) + \psi(x)$$

where $Q_f(\cdot; x_k)$ is a quadratic approximation to f at x_k.

Solving the primal class

Goal is to speed up the problem $\min_x f(x) + \psi(x)$

Standard algorithm: proximal/projected gradient descent.

$$x_{k+1} = \operatorname*{argmin}_x \; Q_f(x; x_k) + \psi(x)$$

where $Q_f(\cdot; x_k)$ is a quadratic approximation to f at x_k.

To apply fancy algorithms (CG, BFGS), three common strategies

- If ψ is non-smooth, pretend it is smooth and try BFGS anyhow
- Active-set methods
- Use a non-trivial quadratic model Q_f and solve subproblem approximately

We take the last approach, but provide an algorithm that *exactly* solves the subproblem in $\mathcal{O}(n \log n)$ time for a specific type of $Q_f(\cdot; x_k)$

Quadratic approximation

Key step in algorithm:

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \; Q_f(x; x_k) + \psi(x) \tag{1}$$

Typically

$$Q_f(x; x_k) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \langle x - x_k, B_k(x - x_k) \rangle / 2$$

where B_k is diagonal (e.g. $B_k = LI$).

Quasi-Newton idea: want $B \simeq \nabla^2 f(x_k)$ and do this via $B_{k+1} = B_k + \Delta_k$ where Δ_k is a rank-1 or rank-2 term update.

We can apply this, but must solve (1) iteratively.
e.g. PQN (Schmidt, van den Berg, Friedlander, Murphy 2009)

Consider the special class formed by 0-memory SR1 quasi-Newton:

$$B_k = \tilde{D}_k + \tilde{u}\tilde{u}^T, \quad \tilde{D}_k = \operatorname{diag}(\tilde{d}_k) \succ 0$$

Proximity operator in scaled norm

$$B_k = \tilde{D}_k + \tilde{u}\tilde{u}^T, \quad H_k = B_k^{-1} = D_k + uu^T$$

Key computation is equivalent to

$$x_{k+1} = \operatorname*{argmin}_x \psi(x) + \frac{1}{2}\|x_k - H_k^{-1}\nabla f(x_k)\|_{B_k}^2$$

Can we solve this (efficiently)? Yes! If ψ is separable and piecewise linear

- $\psi(x) = \|x\|_1$
- $\psi(x) = \max(0, 1 - x)$ hinge-loss
- $\psi(x) = \iota_{\{x:x\geq 0\}}$ or $\psi(x) = \iota_{\{x:a_1\leq x\leq a_2\}}$
- $\psi(x) = \iota_{\{x:\|x\|_\infty\leq 1\}}$

Inspired by fast projections onto ℓ_1 ball

Example: prox of non-negativity constraint

Nonnegative least-squares (NNLS) with non-diagonal norm

$$\min_{x \geq 0} \frac{1}{2} \|x - y\|^2_{H^{-1}}, \quad H = D + uv^T$$

$D = \text{diag}(d) \succ 0$, $u = v \in \mathbb{R}^n$.

Introduce Lagrange multiplier $\lambda \in \mathbb{R}^n$. KKT conditions are:

$$x \geq 0, \, \lambda \geq 0, \quad \langle x, \lambda \rangle = 0, \quad x = y + (D + uv^T)\lambda$$

Define the scalar $s = \langle v, \lambda \rangle$. If s is known, problem is solved:

$$x_i = \lfloor y_i + s u_i \rfloor_+, \quad \lambda_i = \lfloor -(y_i + s u_i)/D_{ii} \rfloor_+$$

In other words, we solve the simple diagonal weighted problem with $y \leftarrow y + su$.

Example: prox of non-negativity constraint. Finding s

Define

$$\lambda_i^{(s)} = \lfloor -(y_i + s u_i)/D_{ii} \rfloor_+$$

We need a value of s that satisfies $s = \langle v, \lambda^{(s)} \rangle$.

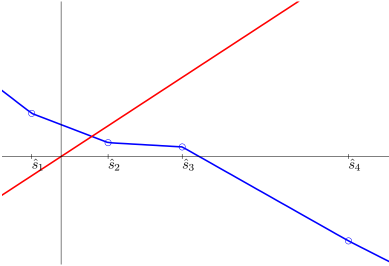
Let $\hat{s}_i = \text{sort}(-y_i/u_i)$ and $p(s) = \langle v, \lambda^{(s)} \rangle$.

Then p is linear in $[\hat{s}_i, \hat{s}_{i+1}]$ and thus trivial to find $p(s) = s$ there.

How to find i? Bisection search.

$p(s) = \sum_i \left(-(v_i y_i + s v_i u_i)/D_{ii}\right) \chi_i(s)$ where $\chi(s)$ is 0 or 1.

Slope composed of $v_i u_i / D_{ii}$ terms, so negative due to assumptions on u, v, D.
Thus $s - p(s)$ is monotonic.



red is s
blue is $p(s)$

Zero-memory SR1 method

$$
\begin{aligned}
x_{k+1} &= \underset{x}{\operatorname{argmin}}\, Q_f(x; x_k) + \psi(x) \\
Q(x; x_k) &= f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \|x - x_k\|^2_{B_k} \\
H_k &= B_k^{-1} = D + uu^T \\
D &= .8\tau_{\text{BB}} I
\end{aligned}
$$

Choose u to satisfy the secant equation

$$
H_k y_k = s_k, \quad \text{where} \quad s_k = x_k - x_{k-1},\; y_k = \nabla f(x_k - \nabla f(x_{k-1}) \tag{SR1}
$$

Here, τ_{BB} is a Barzilai-Borwein updates

$$
\tau_{\text{BB}} = \langle s_k, y_k \rangle / \|y_k\|^2
$$

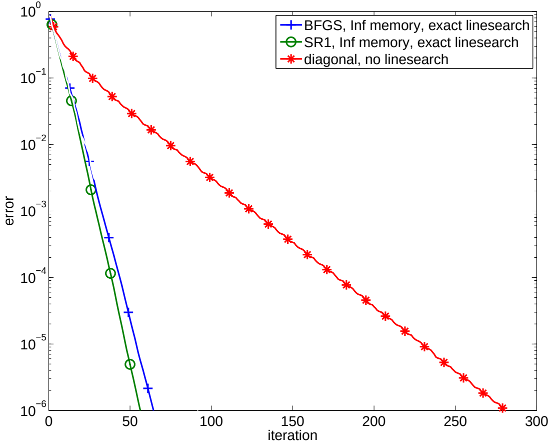
For convergence results, add (non-monotonic) line search

Question: does the extra rank-1 term really matter? N.B.: rank-2 term = zero-memory BFGS = conjugate gradients (for quadratics)

Quasi-Newton vs plain first-order

$$\min f(x) + \psi(x)$$

Both f and ψ quadratic. Approximate f, keep ψ unchanged. $n = 1000$

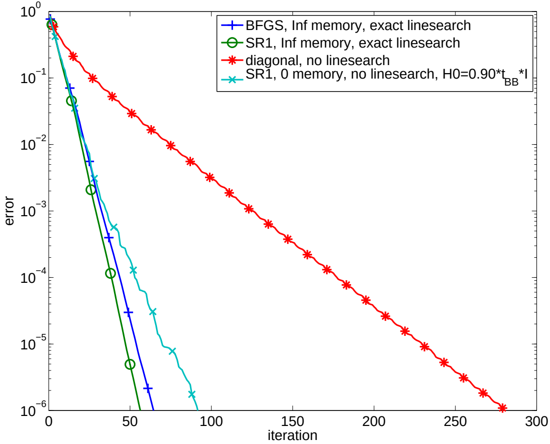


"diagonal" = Spectral Projected Gradient (i.e. Barzilai-Borwein)

Quasi-Newton vs limited-memory quasi-Newton

$$\min f(x) + \psi(x)$$

Both f and ψ quadratic. Approximate f, keep ψ unchanged. $n = 1000$



"diagonal" = Spectral Projected Gradient (i.e. Barzilai-Borwein)

Numerical comparisons

First-order methods

- Spectral projected gradient (SPG) (Birgin, Martínez, Raydan 2000)
  - Uses BB stepsize (Barzilai, Borwein 1988)
  - Extend to non-smooth case
- FISTA (Nesterov 1983; Beck, Teboulle 2009)
  - Use BB stepsize and line search
  - Restart every 1000 iterations

"1.5"-order methods. Most use active-set strategy

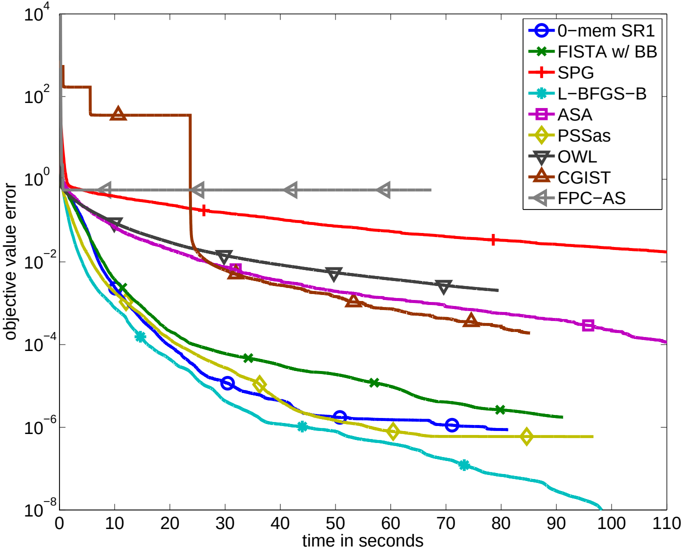
- L-BFGS-B† (Byrd, Lu, Nocedal, Zhu 1995)
- ASA† "Active Set Algorithm" (Hager, Zhang 2006)
- CGIST "CG + IST" (Goldstein, Setzer 2011)
- FPC-AS "FPC + Active Set" (Wen, Yin, Goldfarb, Zhang 2010)
- PSSas "Projected Scaled Sub-gradient + Active Set" (Schmidt, Fung, Rosales 2007)
- OWL "Orthant-wise Learning" (Andrew, Gao 2007)

† require splitting $x = x_+ - x_-$ with $x_+, x_- \geq 0$

all in MATLAB except L-BFGS-B in Fortran and ASA in C

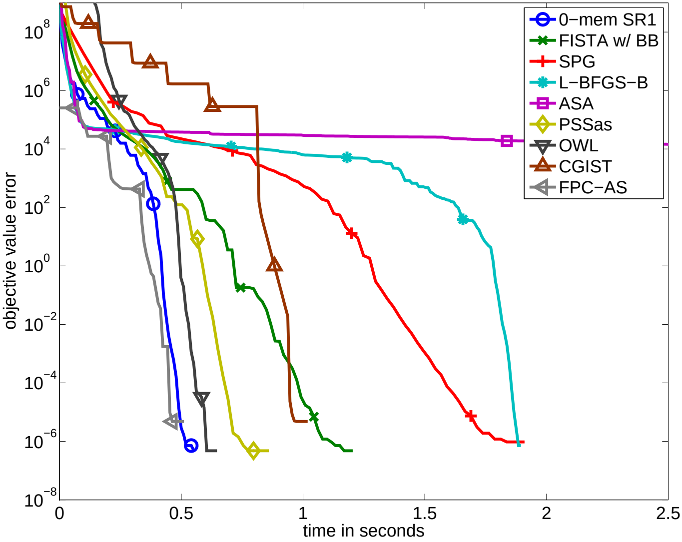
Numerical comparisons: test 1

Solve LASSO: $\min_x \ \lambda\|x\|_1 + \frac{1}{2}\|Ax - b\|^2$
A is 1500×3000 and $\mathcal{N}(0,1)$ iid, $\lambda = 0.1$



Numerical comparisons: test 2

Also LASSO, but pick A and b according to 3D discrete differential operator used by Fletcher, $N = 13^3 = 2197$, $\lambda = 1$



Numerical comparisons: summary

Subjective rankings in order

faster	L-BFGS-B
	0-mem SR1
	PSSas
	FISTA
slower	ASA
	CGIST
	OWL
	SPG
	FPC-AS

Table: Test 1

faster	FPC-AS
	0-mem SR1
	PSSas
	OWL
	FISTA
	CGIST
slower	SPG
	L-BFGS-B
	ASA

Table: Test 2

Generalized primal algorithm

Recall

Definition (generalized primal)

$$\min_x f(x) + \sum_i \psi_i(x), \;\; \psi_i \text{ separable}$$

Do not assume that $\operatorname{prox}_{\sum \psi_i}$ is easy. How to solve?

Solution:
Generalized forward-backward algorithm (Raguet, Fadili, Peyré 2011).

Extensions

- Apply quasi-Newton (or static preconditioner) to saddle-point (i.e. primal-dual) problems
- Rank-1 proximity operators for larger class of functions
- Rank-2 proximity operators
- Block-diagonal preconditioner

$$B = \left[\begin{array}{c|c|c} D + uu^t & 0 & 0 \\ \hline 0 & D + uu^t & 0 \\ \hline 0 & 0 & D + uu^t \end{array}\right]$$