Variable Metric Monotone Operator Splitting

Jalal Fadili

GREYC, CNRS-ENSICAEN-Université de Caen,

http://www.greyc.ensicaen.fr/~jfadili

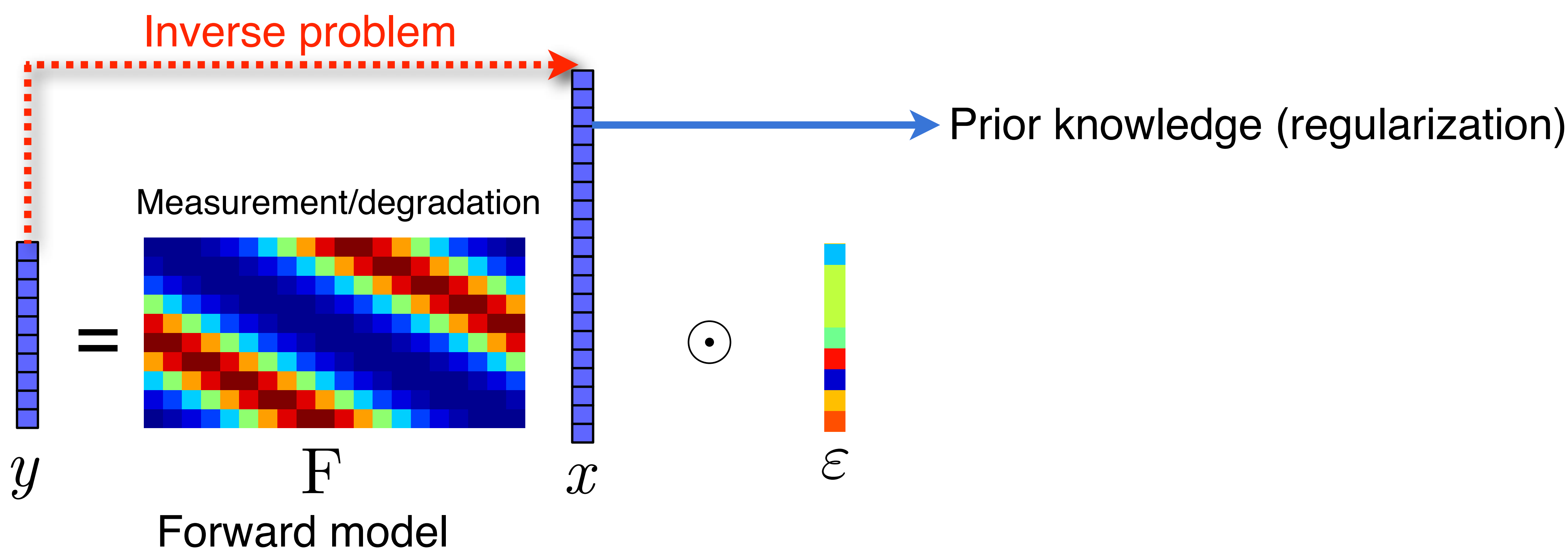
Joint work with Stephen Becker (LJLL UPMC)

Journées Bordelaises d'Analyse Mathématique des Images 2012

Motivations

- Optimization problems in regularized inverse problems :

$$\min_{x\in\mathbb{R}^N} \underbrace{f(x)}_{\text{Data fidelity}} + \underbrace{g_1(x) + \cdots + g_K(x)}_{\text{Regularization, constraints}}$$



Typical models

Smooth, piecewise-smooth, sparse, cartoon, etc..

Motivations

- Optimization problems in regularized inverse problems :

$$\min_{x\in\mathbb{R}^N} \underbrace{f(x)}_{\text{Data fidelity}} + \underbrace{g_1(x)+\cdots+g_K(x)}_{\text{Regularization, constraints}}$$

- Assumptions :
  - $f, g_i : \mathbb{R}^N \to \mathbb{R}\cup\{+\infty\}$ are proper, lsc and convex, $f, g_i \in \Gamma_0(\mathbb{R}^N)$;
  - Domain qualification condition(s) ;
  - Set of minimizers $\mathcal{M}^\star \neq \emptyset$.
- Example : sparsity regularization, e.g. :

$$\min_{x\in\mathbb{R}^N} \underbrace{f(x)}_{\text{Data fidelity}} + \underbrace{g_1(x)+\cdots+g_K(x)}_{\text{e.g. } \ell_1,\, \ell_p-\ell_q \text{ norm on overlapping blocks, etc.}}$$

- A variety of potential applications : signal and image processing, machine learning, classification, statistical estimation, etc..

Warm-up: the smooth case

$$\min_{x \in \mathbb{R}^N} f(x), \quad \nabla f \in C^0(\mathbb{R}^N) \cap \beta\text{-Lip}.$$

$$x^{(n+1)} = x^{(n)} - \gamma_n \nabla f(x^{(n)}), \quad 0 < \underline{\gamma} \leq \gamma_n \leq \overline{\gamma}$$

- Idea of gradient descent : let $\gamma > 0$,

$$f(x^{(n)} + \gamma d) - f(x^{(n)}) = \gamma \left\langle \nabla f(x^{(n)}), d \right\rangle + o(\gamma \|d\|).$$

The decrease in the function when moving from $x^{(n)}$ is bounded as

$$0 < f(x^{(n)}) - f(x^{(n)} + \gamma d) \leq \gamma \|\nabla f(x^{(n)})\| \, \|d\|$$

with equality in the upper bound if and only if $\nabla f(x^{(n)}) \propto -d$.

- The direction of the steepest descent is $-\nabla f(x^{(n)})$, and the largest decrease is $-\left\|\nabla f(x^{(n)})\right\|^2$.
- Convergence bound $1/n$ (on the objective) and $1/n^2$ for multi-step accelerations (weighted gradient memory).

Pros and cons of first-order methods

- Strong points:
  - Broad family of problems with a provably convergent algorithm.
  - Rates (objective, iterates in some circumstances).
  - Simplicity: at each step, a single evaluation of ∇f and a single (fixed or variable step size), or a small number of evaluations of f (line search).
- Weak points:
  - Its relatively low rate of convergence: even with multi-step acceleration (complexity bounds).
  - Bad in applications with ill-conditioned problems and where high-accuracy solution is required.
- Can we hope for better ? Yes, variable metric.

Variable Metric methods: the gist

- The gradient and the Hessian of a nonlinear function f are specific representations of the first and second-order derivatives tied to the standard Euclidian inner product on $\mathbb{R}^N$.
- Let us change a new inner product and metric : $V \in \mathbb{R}^{N\times N}$ sdp matrix,

$$\langle x, y\rangle_V = \langle Vx, y\rangle = x^T V y, \quad \|x\|_V = \langle Vx, x\rangle^{1/2}.$$

- The gradient and Hessian now change to

$$f(x+h) = f(x) + \left\langle V^{-1}\nabla f(x), h\right\rangle_V + \tfrac{1}{4}\left\langle (V^{-1}\nabla^2 f(x) + \nabla^2 f(x)V^{-1})h, h\right\rangle + o(\|h\|_V^2).$$

- In the classical Newton method : the descent direction is the gradient computed w.r.t. the scalar product defined by $V = \nabla^2 f(x)$ (the hessian is the unit matrix).
- Yet another look at gradient descent : it amounts to solving (up to renormalization)

$$\min_{d\in\mathbb{R}^N} \langle \nabla f(x), d\rangle + \frac{1}{2}\|d\|^2 \ .$$

- Why not something else than unit ball, e.g., an ellipsoid.
- Scaling by V induces a change of coordinate system (and metric), and if such a change adjusts to the geometry of the problem $\Rightarrow$ better convergence (e.g. inverse of Hessian).
- With this standpoint :

$$d = \operatorname*{argmin}_{d\in\mathbb{R}^N} \langle \nabla f(x), d\rangle + \frac{1}{2}\|d\|_V^2 = -V^{-1}\nabla f(x) \ .$$

Generic Variable Metric Scheme

Use a varying V_n to adjust it to the geometry along the trajectory.

Denote $H_n = V_n^{-1}$.

Initialization : Choose an initial $x^{(0)} \in \mathrm{dom}(f)$ and sdp matrix H_0.

Main iteration : Construct a sequence of iterates $(x^{(n)})_{n\in\mathbb{N}}$ as follows :

repeat

- Compute the H_n-anti-gradient descent direction :

$$d^{(n)} = -H_n \nabla f(x^{(n)}).$$

- Use a fixed, variable or line search to get the descent step size γ_n.
- Update the iterate :

$$x^{(n+1)} = x^{(n)} + \gamma_n d^{(n)} \ .$$

- Update the sdp matrix $H_n \to H_{n+1}$.

until *convergence* ;

Output : $x^{(n)}$.

Generic Variable Metric Scheme

- Enjoys global and local convergence guarantees.
- Quadratic or superlinear convergence rates in some situations.
- Examples of metrics (remember $H_n = V_n^{-1}$) :
  - Newton : $H_n = \left(\nabla^2 f(x^{(n)})\right)^{-1}$.
  - Quasi-Newton : recursive refinement of V_n to approximate the hessian satisfying the secant condition

$$H_n(\nabla f(x^{(n)}) - \nabla f(x^{(n-1)})) = x^{(n)} - x^{(n-1)} \ .$$

    - Barzilai-Borwein : $H_n = \tau_n I$.
    - Broyden family (BFGS, DFP) : rank-2 update $H_{n+1} = H_n + \sum_{i=1}^{2} u_i^{(n)} {u_i^{(n)}}^T$.
    - SR1 : symmetric rank-1 update $H_{n+1} = H_n + u^{(n)} {u^{(n)}}^T$.
    - Limited memory variants : e.g. LMSRr, L-BFGS, CG.

Variable metric for the non-smooth case

$$\min_{x\in\mathbb{R}^N} f(x), \quad f \in \Gamma_0(\mathbb{R}^N)$$

- Semi-smooth (quasi)-Newton methods :
  - Does not exploit the structure of the problem.
  - Construct a simple slanting function is challenging in high dimension.
- Active sets :
  - Identify activity via a subproblem.
  - Very simple f.
- Variable metric proximal point algorithm :

$$x^{(n+1)} = (1-\lambda_n)x^{(n)} + \lambda_n(I + \gamma_n V_n^{-1}\partial f)^{-1}(x^{(n)})\ , \lambda_n \in]0,1], \gamma_n > 0\ .$$

  - But what if f is not simple, e.g. constrained problem, $\min f + \sum_i g_i$.

Variable metric splitting

$$\min_{x \in \mathbb{R}^N} f(x) + \sum_i g_i(x)$$

- Assumptions :
  - $f, g_i : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\},\ f, g_i \in \Gamma_0(\mathbb{R}^N)$;
  - $f \in C^1(\mathbb{R}^N)$ with β-Lipschitz gradient, all g_i's are simple ;
  - Domain qualification condition(s) ;
  - Set of minimizers $\mathcal{M}^\star \neq \emptyset$ (e.g. coercivity).
- Requirements :
  - Exploit the (composite) additive structure of the objective.
  - Exploit the properties of the individual functions : g_i simple (closed-form proximity operator) and f smooth.
  - Deal with large scale data.
  - Avoid nested algorithms.

Variable metric forward-backward splitting

$$\min_{x \in \mathbb{R}^N} f(x) + g(x)$$

- Forward-Backward splitting (non-relaxed) :

$$x^{(n+1)} = (I + \gamma_n \partial g)^{-1} \left(x^{(n)} - \gamma_n \nabla f(x^{(n)}) \right) \ , \gamma_n \in]0, 2/\beta[\ .$$

- Variable metric Forward-Backward splitting (non-relaxed) :

$$x^{(n+1)} = (I + V_n^{-1} \partial g)^{-1} \left(x^{(n)} - V_n^{-1} \nabla f(x^{(n)}) \right) \ .$$

- More generally, find the zeros of :

$$0 \in Ax + Bx$$

  - $A, B : \mathcal{H} \to 2^{\mathcal{H}}$ are maximal monotone operators ;
  - A single-valued with $\beta A \in \mathcal{A}(\frac{1}{2})$, B simple ;
  - $\mathrm{zer}(A + B) \neq \emptyset$.

Variable metric splitting

Main questions and challenges

- Convergence guarantees.
- Convergence Rates.
- Construct attractive metrics :
  - Computational load and storage of V_n^{-1}.
  - Easy implementation of the implicit step : $(I + V_n^{-1}\partial g)^{-1}$.
    - In general, $(I + V_n^{-1}\partial g)^{-1}$ as difficult to compute as solving the original problem.
  - Fast algorithms for large scale problems.
  - Maintain convergence guarantees.

Variable metric splitting

- Forward-Backward [Chen and Rockafellar 97, Tseng and Yun 09, Lolito et al. 09, Combettes and Vu 12] :
  - arbitrary exogenous metrics ;
  - optimization and monotone inclusions ;
  - finite-dimensional or infinite dimensional settings ;
  - convergence ;
  - no fast algorithm (implicit step).
- Forward-Backward with specific (simple) metrics or problems [e.g. Bonnetini et al. 09, Salzo and Villa 11, Schmidt et al. 11, Lee et al. 12] :
  - most of them finite-dimensional optimization setting ;
  - some of them only special problems (e.g. box or linear constraints) ;
  - some lack convergence guarantees ;
  - solve a subproblem (implicit step).
- Forward-Backward with pre-conditioning [e.g. Elad 06, Wright et al. 09, Vonesch et al. 09, Goldstein and Setzer 11] :
  - finite-dimensional optimization setting ;
  - some of them only special problems (e.g. specific operators) ;
  - some lack convergence guarantees.

Outline

- Convex analysis with variable metric.
- Convergence.
- LMSRr proximal splitting scheme.
- Applications.
- Extensions and conclusion.

Outline

- Convex analysis with variable metric.
- Convergence.
- LMSRr proximal splitting scheme.
- Applications.
- Extensions and conclusion.

Proximity operator

Definition *Let* $\mathcal{H} = (\mathbb{R}^N, \langle \cdot, \cdot \rangle)$ *equipped with the usual Euclidean scalar product* $\langle x, y\rangle$ *and associated norm* $\|x\| = \sqrt{\langle x, x\rangle}$. *For* $V \in \mathbb{R}^{N\times N}$ *sdp, let* $\mathcal{H}_V = (\mathbb{R}^N, \langle \cdot, \cdot \rangle_V)$ *with the scalar product* $\langle x, y\rangle_V = \langle x, Vy\rangle$ *and norm* $\|x\|_V$ *corresponding to the metric induced by* V.

Definition (Proximity operator in $\mathcal{H}$ **[J.-J. Moreau 1962])** *Let* $g \in \Gamma_0(\mathcal{H})$. *Then, for every* $x \in \mathcal{H}$, *the function* $z \mapsto \frac{1}{2}\|x - z\|^2 + g(z)$ *achieves its infimum at a unique point denoted by* $\mathrm{prox}_g\, x$. *The single-valued operator* $(I + \partial g)^{-1} = \mathrm{prox}_g : \mathcal{H} \to \mathcal{H}$ *thus defined is the proximity operator of* g.

Definition (Proximity operator $\mathcal{H}_V$**)** *We define* $\mathrm{prox}_g^V(x) = (\mathrm{I}_{\mathcal{H}_V} + V^{-1}\partial g)^{-1}(x) = \underset{z\in\mathbb{R}^N}{\mathrm{argmin}}\, \frac{1}{2}\|x - z\|_V^2 + g(z)$ *the proximity operator of* g *w.r.t. the norm endowing* $\mathcal{H}_V$.

Computing prox_g^V *is difficult in general even if* prox_g *is available.*

NSC of a minimum

$$\min_{x \in \mathbb{R}^N} f(x) + g(x)$$

Theorem *Let* $f \in \Gamma_0(\mathcal{H}) \cap C^1(\mathbb{R}^N)$ *and* $g \in \Gamma_0(\mathcal{H})$ *as defined before, and* $V \succeq a\mathrm{I}_{\mathcal{H}}$, $a > 0$. *Then, for* $\gamma > 0$, *the following are equivalent :*

(i) $x^\star \in \mathcal{M}^\star$.

(ii) $x^\star = \operatorname{prox}_{\gamma g}^V \circ \left(\mathrm{I}_{\mathcal{H}} - \gamma V^{-1} \nabla f\right)(x^\star)$.

(iii) $x^\star = \left(\frac{V + \gamma \partial g}{a}\right)^{-1} \circ \left(\frac{V - \gamma \nabla f}{a}\right)(x^\star)$.

NSC of a minimum

$$\min_{x \in \mathbb{R}^N} f(x) + g(x)$$

Theorem *Let* $f \in \Gamma_0(\mathcal{H}) \cap C^1(\mathbb{R}^N)$ *and* $g \in \Gamma_0(\mathcal{H})$ *as defined before, and* $V \succeq a\mathrm{I}_{\mathcal{H}}$*,* $a > 0$*. Then, for* $\gamma > 0$*, the following are equivalent :*

(i) $x^\star \in \mathcal{M}^\star$*.*

(ii) $x^\star = \mathrm{prox}^V_{\gamma g} \circ \left(\mathrm{I}_{\mathcal{H}} - \gamma V^{-1} \nabla f\right)(x^\star)$ *.*

(iii) $x^\star = \left(\frac{V + \gamma \partial g}{a}\right)^{-1} \circ \left(\frac{V - \gamma \nabla f}{a}\right)(x^\star)$*.*

- FB stems from a fixed point equation.

Proximal calculus in $\mathcal{H}_V$

Lemma (Moreau identity in $\mathcal{H}_V$) *Let $g \in \Gamma_0(\mathcal{H})$, then for any $x \in \mathcal{H}$*

$$\operatorname{prox}^V_{\rho g^*}(x) + \rho V^{-1} \circ \operatorname{prox}^{V^{-1}}_{g/\rho} \circ V(x/\rho) = x, \forall\, 0 < \rho < +\infty \; .$$

Corollary

$$\operatorname{prox}^V_{g}(x) = x - V^{-1} \circ \operatorname{prox}^{V^{-1}}_{g^*} \circ V(x) \; .$$

- If prox^V_g is easy to compute, then so is $\operatorname{prox}^{V^{-1}}_{g^*}$.

(hint : Moreau identity and Sherman-Morrison inversion lemma.)

Proximal calculus in $\mathcal{H}_V$

Lemma (Moreau identity in $\mathcal{H}_V$) *Let $g \in \Gamma_0(\mathcal{H})$, then for any $x \in \mathcal{H}$*

$$\operatorname{prox}^V_{\rho g^*}(x) + \rho V^{-1} \circ \operatorname{prox}^{V^{-1}}_{g/\rho} \circ V(x/\rho) = x, \forall\, 0 < \rho < +\infty\ .$$

Corollary

$$\operatorname{prox}^V_g(x) = x - V^{-1} \circ \operatorname{prox}^{V^{-1}}_{g^*} \circ V(x)\ .$$

Conclusion

- If prox^V_g is easy to compute, then so is $\operatorname{prox}^{V^{-1}}_{g^*}$.

(hint : Moreau identity and Sherman-Morrison inversion lemma.)

Continuity properties of operators

Lemma (Proximity operator) *Let V be sdp with $bI \succeq V \succeq aI$, $b \geq a > 0$, and $g \in \Gamma_0(\mathcal{H})$. Then*

(i) $\mathrm{prox}_{\gamma g}^{V}$ *is firmly non-expansive on* $\mathcal{H}_V$*, hence non-expansive,* $\gamma > 0$.

(ii) $\left(\frac{V+\gamma\partial g}{a}\right)^{-1}$ *is firmly non-expansive on* $\mathcal{H}$*, hence non-expansive,* $\gamma > 0$.

(iii) *For* W *sdp,*

$$\left\|\mathrm{prox}_{\gamma g}^{W}(x) - \mathrm{prox}_{\gamma g}^{V}(y)\right\| \leq b/a\left(\|W - V\|\,\|x\| + \|x - y\|\right) + b/a\left\|\mathrm{prox}_{\gamma g}^{V}(0)\right\| .$$

Continuity properties of operators

Lemma (Proximity operator) *Let V be sdp with $bI \succeq V \succeq aI$, $b \geq a > 0$, and $g \in \Gamma_0(\mathcal{H})$. Then*

(i) $\operatorname{prox}_{\gamma g}^{V}$ *is firmly non-expansive on $\mathcal{H}_V$, hence non-expansive, $\gamma > 0$.*

(ii) $\left(\frac{V+\gamma\partial g}{a}\right)^{-1}$ *is firmly non-expansive on $\mathcal{H}$, hence non-expansive, $\gamma > 0$.*

(iii) For W sdp,

$$\left\|\operatorname{prox}_{\gamma g}^{W}(x) - \operatorname{prox}_{\gamma g}^{V}(y)\right\| \leq b/a \left(\|W - V\| \|x\| + \|x - y\|\right) + b/a \left\|\operatorname{prox}_{\gamma g}^{V}(0)\right\| .$$

Continuity properties of operators

Lemma (Proximity operator) *Let V be sdp with $bI \succeq V \succeq aI$, $b \geq a > 0$, and $g \in \Gamma_0(\mathcal{H})$. Then*

(i) $\mathrm{prox}_{\gamma g}^{V}$ *is firmly non-expansive on $\mathcal{H}_V$, hence non-expansive, $\gamma > 0$.*

(ii) $\left(\frac{V+\gamma\partial g}{a}\right)^{-1}$ *is firmly non-expansive on $\mathcal{H}$, hence non-expansive, $\gamma > 0$.*

(iii) For W sdp,

$$\left\|\mathrm{prox}_{\gamma g}^{W}(x) - \mathrm{prox}_{\gamma g}^{V}(y)\right\| \leq b/a\left(\|W - V\|\,\|x\| + \|x - y\|\right) + b/a\left\|\mathrm{prox}_{\gamma g}^{V}(0)\right\| .$$

Lemma (Gradient operator) *Let V be sdp with $bI \succeq V \succeq aI$, $b \geq a > 0$, and $f \in \Gamma_0(\mathcal{H}) \cap C^1(\mathcal{H})$ with $\nabla f \in \beta$-Lip$(\mathcal{H})$. Then $(\mathrm{I}_{\mathcal{H}} - \gamma V^{-1}\nabla f)$ is a $\gamma\beta/(2a)$-averaged operator on $\mathcal{H}_V$, hence nonexpansive, $\forall \gamma \in]0, 2a/\beta[$.*

Outline

- Convex analysis with variable metric.
- Convergence.
- LMSRr proximal splitting scheme.
- Applications.
- Extensions and conclusion.

Variable Metric Forward-Backward

- Global convergence :
  - Assumptions on the metric and step size.
  - Rates (linear) with extra assumptions on the functions, e.g. strong convexity.
- Local convergence :
  - Assumptions on initialization.
  - Local assumptions on the smooth part.
  - Fast rates (quadratic or linear) for well-behaved metrics with extra smoothness assumptions.
- Even in the smooth case, it is very difficult to prove something good about local convergence of a quasi-Newton method.

Variable Metric Forward-Backward

Theorem (Global linear convergence) *Assume that either f or g is strongly, and that the variable metric forward-backward is run through a sequence of sdp matrices $V_n \to V$, and a step size $\gamma \in]0, \overline{\gamma}_V[$ with an appropriate $\overline{\gamma}_V$. Then the variable metric forward-backward converges to the (unique) minimizer $x^\star$ linearly.*

Theorem (Local linear convergence) *Assume that $f \in \Gamma_0(\mathbb{R}^N) \cap C^2(\mathbb{R}^N)$ with its Hessian being positive definite at $x^\star$. Assume that the variable metric forward-backward is run with sdp matrices $V_n \succeq a\mathrm{I}_{\mathcal{H}}$, $a > 0$, such that $\sup_n \left\| V_n - \nabla^2 f(x^{(n)}) \right\| \leq a\rho < 1$ or a fortiori $\sup_n \left\| V_n - \nabla^2 f(x^\star) \right\| \leq \rho a$, $\rho < 1$. If the method is started sufficiently close to $x^\star$, then it converges to $x^\star$ linearly.*

Newton Forward-Backward

Theorem *Assume that* $f \in \Gamma_0(\mathbb{R}^N) \cap C^2(\mathbb{R}^N)$ *with its Hessian being Lipschitz continuous.*

(i) Local convergence : If the Hessian is positive definite at $x^\star$*, then the Newton forward-backward, started sufficiently close to* $x^\star$*, converges to* $x^\star$ *quadratically.*

(ii) Global convergence : If f *is strongly convex, the Newton forward-backward with an appropriate step size or line search converges to (the unique)* $x^\star$ *linearly.*

Quasi-Newton Forward-Backward

Theorem *Assume that $f \in \Gamma_0(\mathbb{R}^N) \cap C^2(\mathbb{R}^N)$ with its Hessian being positive definite at $x^\star$ and Lipschitz continuous. If the matrices V_n converge superlinearly to $\nabla^2 f(x^\star)$, then the quasi-Newton forward-backward, started sufficiently close to $x^\star$, converges to $x^\star$ superlinearly.*

- Under appropriate assumptions, SR1 satisfies the above requirements [Conn et al. 91, Byrd et al. 96].

Outline

- Convex analysis with variable metric.
- Convergence.
- LMSRr proximal splitting scheme.
- Applications.
- Extensions and conclusion.

LMSR-r FB

Zero-memory SRr update

$$H_n = D_n + \sum_{i=1}^{r} u_i^{(n)} {u_i^{(n)}}^T, \ D_n \text{ diagonal}$$

Require: $x_0 \in \mathrm{dom}(f+g)$, Lipschitz constant estimate β of ∇f, stopping criterion ϵ

for $n = 1, 2, 3, \ldots$ **do**

$s^{(n)} \leftarrow x^{(n)} - x^{(n-1)}$

$y^{(n)} \leftarrow \nabla f(x^{(n)}) - \nabla f(x^{(n-1)})$

Compute $H_n = D_n + \sum_{i=1}^{r} u_i^{(n)} {u_i^{(n)}}^T$ (see shortly), and set $V_n = H_n^{-1}$.

Compute the rank-r proximity operator (see shortly)

$$\hat{x}^{(n+1)} \leftarrow \mathrm{prox}_g^{V_n}(x^{(n)} - H_n \nabla f(x^{(n)}))$$

$p^{(n)} \leftarrow \hat{x}^{(n+1)} - x^{(n)}$ and terminate if $\|p^{(n)}\| < \epsilon$

Line search along the ray $x^{(n)} + \theta p^{(n)}$ to determine $x^{(n+1)}$, or choose $\theta = 1$.

end for

LMSR-r proximity operator

Theorem *Let* $g \in \Gamma_0(\mathcal{H})$ *and* $V = D + \sum_{i=1}^{r} u_i u_i^T$, *where* $D \succ 0$ *is diagonal. Then,*

(i)

$$\mathrm{prox}_g^V(x) = D^{-1/2} \circ \mathrm{prox}_{g \circ D^{-1/2}} \circ D^{1/2}(x - D^{-1} U\alpha) \ ,$$

where $U = (u_1, \cdots, u_r)$ *and* $\alpha \in \mathbb{R}^r$ *is the unique root of*

$$p(\alpha) = U^T \left(x - D^{-1/2} \circ \mathrm{prox}_{g \circ D^{-1/2}} \circ D^{1/2}(x - D^{-1} U\alpha) \right) + B\alpha \ ,$$

where $B = U^T Q^+ U$ *is sdp.*

(ii) $p : \mathbb{R}^r \longrightarrow \mathbb{R}^r$ *is a Lipschitz continuous mapping.*

(iii) If $\mathrm{prox}_{g \circ D^{-1/2}}$ *is Newton differentiable with generalized derivative* G, *then so is the mapping* h *with a generalized derivative* $g : \mathbb{R}^r \longrightarrow \mathbb{R}^{r \times r}$

$$g(\alpha) = U^T (Q^+ + D^{-1/2} \circ G(D^{1/2} x - \alpha D^{-1/2} u) \circ D^{-1/2}) U \ .$$

LMSR-r proximity operator

Theorem *Let* $g \in \Gamma_0(\mathcal{H})$ *and* $V = D + \sum_{i=1}^r u_i u_i^T$, *where* $D \succ 0$ *is diagonal. Then,*

Much simpler

(i)

$$\text{prox}_g^V(x) = D^{-1/2} \circ \boxed{\text{prox}_{g \circ D^{-1/2}}} \circ D^{1/2}(x - D^{-1}U\alpha) \ ,$$

where $U = (u_1, \cdots, u_r)$ *and* $\alpha \in \mathbb{R}^r$ *is the unique root of*

$$p(\alpha) = U^T \left(x - D^{-1/2} \circ \text{prox}_{g \circ D^{-1/2}} \circ D^{1/2}(x - D^{-1}U\alpha) \right) + B\alpha \ ,$$

where $B = U^T Q^+ U$ *is sdp.*

(ii) $p : \mathbb{R}^r \to \mathbb{R}^r$ *is a Lipschitz continuous mapping.*

(iii) If $\text{prox}_{g \circ D^{-1/2}}$ *is Newton differentiable with generalized derivative* G, *then so is the mapping* h *with a generalized derivative* $g : \mathbb{R}^r \to \mathbb{R}^{r \times r}$

$$g(\alpha) = U^T (Q^+ + D^{-1/2} \circ G(D^{1/2}x - \alpha D^{-1/2}u) \circ D^{-1/2})U \ .$$

LMSR-r proximity operator

Theorem *Let* $g \in \Gamma_0(\mathcal{H})$ *and* $V = D + \sum_{i=1}^r u_i u_i^T$, *where* $D \succ 0$ *is diagonal. Then,*

Much simpler

(i)

$$\operatorname{prox}_g^V(x) = D^{-1/2} \circ \boxed{\operatorname{prox}_{g \circ D^{-1/2}}} \circ D^{1/2}(x - D^{-1}U\alpha) ,$$

where $U = (u_1, \cdots, u_r)$ *and* $\alpha \in \mathbb{R}^r$ *is the unique root of*

$$\boxed{p(\alpha) = U^T \left(x - D^{-1/2} \circ \operatorname{prox}_{g \circ D^{-1/2}} \circ D^{1/2}(x - D^{-1}U\alpha) \right) + B\alpha ,}$$

where $B = U^T Q^+ U$ *is sdp.*

Much lower-dimensional problem

(ii) $p : \mathbb{R}^r \to \mathbb{R}^r$ *is a Lipschitz continuous mapping.*

(iii) If $\operatorname{prox}_{g \circ D^{-1/2}}$ *is Newton differentiable with generalized derivative* G, *then so is the mapping* h *with a generalized derivative* $g : \mathbb{R}^r \to \mathbb{R}^{r \times r}$

$$g(\alpha) = U^T (Q^+ + D^{-1/2} \circ G(D^{1/2}x - \alpha D^{-1/2}u) \circ D^{-1/2}) U .$$

LMSR-r proximity operator

Theorem *Let* $g \in \Gamma_0(\mathcal{H})$ *and* $V = D + \sum_{i=1}^r u_i u_i^T$, *where* $D \succ 0$ *is diagonal. Then,*

Much simpler

(i)

$$\mathrm{prox}_g^V(x) = D^{-1/2} \circ \boxed{\mathrm{prox}_{g\circ D^{-1/2}}} \circ D^{1/2}(x - D^{-1}U\alpha) ,$$

where $U = (u_1, \cdots, u_r)$ *and* $\alpha \in \mathbb{R}^r$ *is the* *unique* *root of*

$$\boxed{p(\alpha) = U^T \left(x - D^{-1/2} \circ \mathrm{prox}_{g\circ D^{-1/2}} \circ D^{1/2}(x - D^{-1}U\alpha) \right) + B\alpha ,}$$

where $B = U^T Q^+ U$ *is sdp.*

Much lower-dimensional problem

(ii) $p : \mathbb{R}^r \to \mathbb{R}^r$ *is a Lipschitz continuous mapping.*

(iii) If $\mathrm{prox}_{g\circ D^{-1/2}}$ *is Newton differentiable with generalized derivative* G, *then so is the mapping* h *with a generalized derivative* $g : \mathbb{R}^r \to \mathbb{R}^{r\times r}$

$$\boxed{g(\alpha) = U^T(Q^+ + D^{-1/2} \circ G(D^{1/2}x - \alpha D^{-1/2}u) \circ D^{-1/2})U .}$$

Semi-smooth Newton (Superlinear)

LMSR-1 proximity operator

Corollary *Let* $g \in \Gamma_0(\mathcal{H})$ *and* $V = D + uu^T$*, where* $D \succ 0$ *is diagonal. Then,*

(i)

$$\operatorname{prox}_g^V(x) = D^{-1/2} \circ \operatorname{prox}_{g \circ D^{-1/2}}(D^{1/2}x - v) \ ,$$

where $v = \alpha D^{-1/2} u$ *and* α *is the unique root of*

$$p(\alpha) = \left\langle u, x - D^{-1/2} \circ \operatorname{prox}_{g \circ D^{-1/2}} \circ D^{1/2}(x - \alpha D^{-1} u) \right\rangle + \alpha \ .$$

(ii) $p : \mathbb{R} \to \mathbb{R}$ *is a Lipschitz continuous and strictly increasing function on* $\mathbb{R}$*.*

(iii) If $\operatorname{prox}_{g \circ D^{-1/2}}$ *is Newton differentiable with generalized derivative* G*, then so is the mapping* h *with a generalized derivative* g

$$g(\alpha) = 1 + \left\langle u, D^{-1/2} \circ G(D^{1/2}x - \alpha D^{-1/2} u) \circ D^{-1/2} u \right\rangle \ .$$

LMSR-1 proximity operator

Corollary *Let* $g \in \Gamma_0(\mathcal{H})$ *and* $V = D + uu^T$, *where* $D \succ 0$ *is diagonal. Then,*

Much simpler

(i)

$$\operatorname{prox}_g^V(x) = D^{-1/2} \circ \boxed{\operatorname{prox}_{g \circ D^{-1/2}}}(D^{1/2}x - v) \ ,$$

where $v = \alpha D^{-1/2} u$ *and* α *is the unique root of*

$$p(\alpha) = \left\langle u, x - D^{-1/2} \circ \operatorname{prox}_{g \circ D^{-1/2}} \circ D^{1/2}(x - \alpha D^{-1} u) \right\rangle + \alpha \ .$$

(ii) $p : \mathbb{R} \to \mathbb{R}$ *is a Lipschitz continuous and strictly increasing function on* $\mathbb{R}$.

(iii) If $\operatorname{prox}_{g \circ D^{-1/2}}$ *is Newton differentiable with generalized derivative* G, *then so is the mapping* h *with a generalized derivative* g

$$g(\alpha) = 1 + \left\langle u, D^{-1/2} \circ G(D^{1/2}x - \alpha D^{-1/2}u) \circ D^{-1/2}u \right\rangle \ .$$

LMSR-1 proximity operator

Corollary *Let $g \in \Gamma_0(\mathcal{H})$ and $V = D + uu^T$, where $D \succ 0$ is diagonal. Then,*

Much simpler

(i)

$$\mathrm{prox}_g^V(x) = D^{-1/2} \circ \boxed{\mathrm{prox}_{g \circ D^{-1/2}}}(D^{1/2}x - v) \; ,$$

where $v = \alpha D^{-1/2}u$ and α is the unique root of

1D problem

$$\boxed{p(\alpha) = \left\langle u, x - D^{-1/2} \circ \mathrm{prox}_{g \circ D^{-1/2}} \circ D^{1/2}(x - \alpha D^{-1}u) \right\rangle + \alpha \; .}$$

(ii) $p : \mathbb{R} \to \mathbb{R}$ *is a Lipschitz continuous and strictly increasing function on* $\mathbb{R}$*.*

(iii) If $\mathrm{prox}_{g \circ D^{-1/2}}$ is Newton differentiable with generalized derivative G, then so is the mapping h with a generalized derivative g

$$g(\alpha) = 1 + \left\langle u, D^{-1/2} \circ G(D^{1/2}x - \alpha D^{-1/2}u) \circ D^{-1/2}u \right\rangle \; .$$

LMSR-1 proximity operator

Corollary *Let $g \in \Gamma_0(\mathcal{H})$ and $V = D + uu^T$, where $D \succ 0$ is diagonal. Then,*

Much simpler

(i)

$$\operatorname{prox}_g^V(x) = D^{-1/2} \circ \boxed{\operatorname{prox}_{g \circ D^{-1/2}}}(D^{1/2}x - v) \,,$$

where $v = \alpha D^{-1/2}u$ and α is the unique root of

1D problem

$$\boxed{p(\alpha) = \left\langle u, x - D^{-1/2} \circ \operatorname{prox}_{g \circ D^{-1/2}} \circ D^{1/2}(x - \alpha D^{-1}u) \right\rangle + \alpha \,.}$$

(ii) $p : \mathbb{R} \to \mathbb{R}$ *is a Lipschitz continuous and strictly increasing function on* $\mathbb{R}$*.*

(iii) If $\operatorname{prox}_{g \circ D^{-1/2}}$ *is Newton differentiable with generalized derivative* G*, then so is the mapping* h *with a generalized derivative* g

$$\boxed{g(\alpha) = 1 + \left\langle u, D^{-1/2} \circ G(D^{1/2}x - \alpha D^{-1/2}u) \circ D^{-1/2}u \right\rangle \,.}$$

Semi-smooth Newton
(Superlinear)
or exact

LMSR-1 proximity operator: separable case

Corollary *Assume that $g \in \Gamma_0(\mathcal{H})$ is separable, i.e. $g(x) = \sum_{i=1}^N g_i(x_i)$, and $V = D + uu^T$, where $D = \mathrm{diag}(d_i) \succ 0$. Then,*

$$\mathrm{prox}_f^V(x) = \left(\mathrm{prox}_{g_i/d_i}(x_i - v_i/d_i)\right)_i ,$$

where $v = \alpha u$ and α is the unique root of

$$h(\alpha) = \left\langle u, x - \left(\mathrm{prox}_{g_i/d_i}(x_i - \alpha u_i/d_i)\right)_i \right\rangle + \alpha ,$$

which is a Lipschitz continuous and strictly increasing function on $\mathbb{R}$.

LMSR-1 proximity operator: separable case

Corollary *Assume that $g \in \Gamma_0(\mathcal{H})$ is separable, i.e. $g(x) = \sum_{i=1}^N g_i(x_i)$, and $V = D + uu^T$, where $D = \mathrm{diag}(d_i) \succ 0$. Then,*

Separable

$$\mathrm{prox}_f^V(x) = \Big(\mathrm{prox}_{g_i/d_i}(x_i - v_i/d_i)\Big)_i \,,$$

where $v = \alpha u$ and α is the unique root of

$$h(\alpha) = \Big\langle u, x - \Big(\mathrm{prox}_{g_i/d_i}(x_i - \alpha u_i/d_i)\Big)_i \Big\rangle + \alpha \,,$$

which is a Lipschitz continuous and strictly increasing function on $\mathbb{R}$.

LMSR-1 proximity operator: separable case

Corollary *Assume that* $g \in \Gamma_0(\mathcal{H})$ *is separable, i.e.* $g(x) = \sum_{i=1}^{N} g_i(x_i)$, *and* $V = D + uu^T$, *where* $D = \mathrm{diag}(d_i) \succ 0$. *Then,*

Separable

$$\mathrm{prox}_f^V(x) = \Big(\mathrm{prox}_{g_i/d_i}(x_i - v_i/d_i)\Big)_i ,$$

where $v = \alpha u$ *and* α *is the unique root of*

1D problem

$$h(\alpha) = \Big\langle u, x - \Big(\mathrm{prox}_{g_i/d_i}(x_i - \alpha u_i/d_i)\Big)_i \Big\rangle + \alpha ,$$

which is a Lipschitz continuous and strictly increasing function on $\mathbb{R}$.

LMSR-1 proximity operator: separable case

Corollary *Assume that $g \in \Gamma_0(\mathcal{H})$ is separable, i.e. $g(x) = \sum_{i=1}^N g_i(x_i)$, and $V = D + uu^T$, where $D = \mathrm{diag}(d_i) \succ 0$. Then,*

Separable

$$\mathrm{prox}_f^V(x) = \Big(\mathrm{prox}_{g_i/d_i}(x_i - v_i/d_i)\Big)_i ,$$

where $v = \alpha u$ and α is the unique root of

1D problem

$$h(\alpha) = \Big\langle u, x - \Big(\mathrm{prox}_{g_i/d_i}(x_i - \alpha u_i/d_i)\Big)_i\Big\rangle + \alpha ,$$

which is a Lipschitz continuous and strictly increasing function on $\mathbb{R}$.

Proposition *Assume that for $1 \le i \le N$, prox_{g_i} is piecewise affine on $\mathbb{R}$ with $k_i \ge 1$ segments , i.e.*

$$\mathrm{prox}_{g_i}(x_i) = a_j x_i + b_j, \quad t_j \le x_i \le t_{j+1}, j \in \{1, \ldots, k_i\} .$$

Let $k = \sum_{i=1}^N k_i$. Then $\mathrm{prox}_g^V(x)$ can be obtained exactly by sorting at most the k real values $\left(\frac{d_i}{u_i}(x_i - t_j)\right)_{i\in\{1,\ldots,N\}, j\in\{1,\ldots,k_i\}}$ which costs $O(k \log k)$.

LMSR-1 proximity operator: separable case

Corollary *Assume that $g \in \Gamma_0(\mathcal{H})$ is separable, i.e. $g(x) = \sum_{i=1}^{N} g_i(x_i)$, and $V = D + uu^T$, where $D = \operatorname{diag}(d_i) \succ 0$. Then,*

Separable

$$\operatorname{prox}_f^V(x) = \Big(\operatorname{prox}_{g_i/d_i}(x_i - v_i/d_i)\Big)_i \,,$$

where $v = \alpha u$ and α is the unique root of

1D problem

$$h(\alpha) = \Big\langle u, x - \Big(\operatorname{prox}_{g_i/d_i}(x_i - \alpha u_i/d_i)\Big)_i \Big\rangle + \alpha \,,$$

which is a Lipschitz continuous and strictly increasing function on $\mathbb{R}$.

Proposition *Assume that for $1 \le i \le N$, $\operatorname{prox}_{g_i}$ is piecewise affine on $\mathbb{R}$ with $k_i \ge 1$ segments , i.e.*

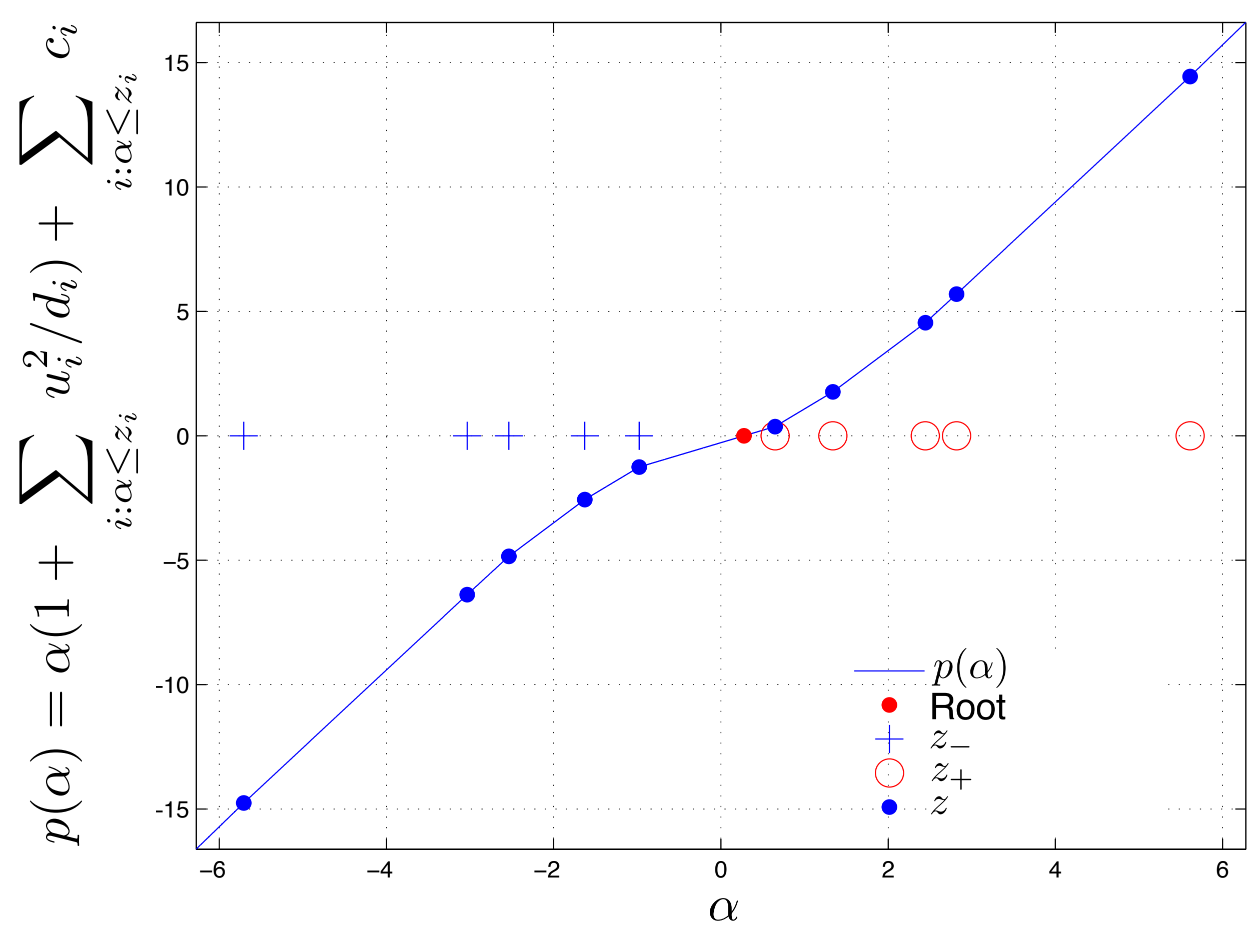
$$\operatorname{prox}_{g_i}(x_i) = a_j x_i + b_j, \quad t_j \le x_i \le t_{j+1}, j \in \{1, \ldots, k_i\} \,.$$

Let $k = \sum_{i=1}^{N} k_i$. Then $\operatorname{prox}_g^V(x)$ can be obtained exactly by sorting at most the k real values $\left(\frac{d_i}{u_i}(x_i - t_j)\right)_{i \in \{1,\ldots,N\}, j \in \{1,\ldots,k_i\}}$ which costs $O(k \log k)$.

LMSR-1 separable case: Examples

Function g	Algorithm
ℓ_1-norm	Separable : exact in $O(N \log N)$
Hinge	Separable : exact in $O(N \log N)$
ℓ_∞-ball	Separable : exact in $O(N \log N)$
Box constraint	Separable : exact in $O(N \log N)$
Positivity constraint	Separable : exact in $O(N \log N)$
Linear constraint	Nonseparable : exact in $O(N \log N)$
ℓ_1-ball	Nonseparable : SSN and $\mathrm{prox}_{g \circ D^{-1/2}}$ is $O(N \log N)$
ℓ_∞-norm	Nonseparable : Moreau-identity
Canonical simplex	Nonseparable : SSN and $\mathrm{prox}_{g \circ D^{-1/2}}$ is $O(N \log N)$
$\max$ function	Nonseparable : Moreau-identity

$$g(x) = \lambda \|x\|_1$$



$$z = \mathtt{sort}(d_i x_i / u_i \pm \lambda / |u_i|)$$

LMSR-1 metric update

$y^{(n)} = \nabla f(x^{(n)}) - \nabla f(x^{(n-1)}), \quad s^{(n)} = x^{(n)} - x^{(n-1)}$.

$\tau_n \leftarrow \frac{\left\langle s^{(n)}, y^{(n)} \right\rangle}{\left\| y^{(n)} \right\|^2}$ {Barzilai-Borwein step length}

$D_n \leftarrow \gamma \tau_n \mathrm{I}_{\mathcal{H}}, 0 < \gamma < 1$ {Diagonal part}

$u^{(n)} \leftarrow (s^{(n)} - D_n y^{(n)}) / \sqrt{\left\langle s^{(n)} - D_n y^{(n)}, y^{(n)} \right\rangle}$. {Rank-1 part vector}

return $H_n = D_n + u^{(n)} {u^{(n)}}^T$ {$V_n = H_n^{-1}$ by Sherman-Morrison lemma}

H_n satisfies the quasi-Newton *secant condition* :

$$H_n y^{(n)} = s^{(n)} .$$

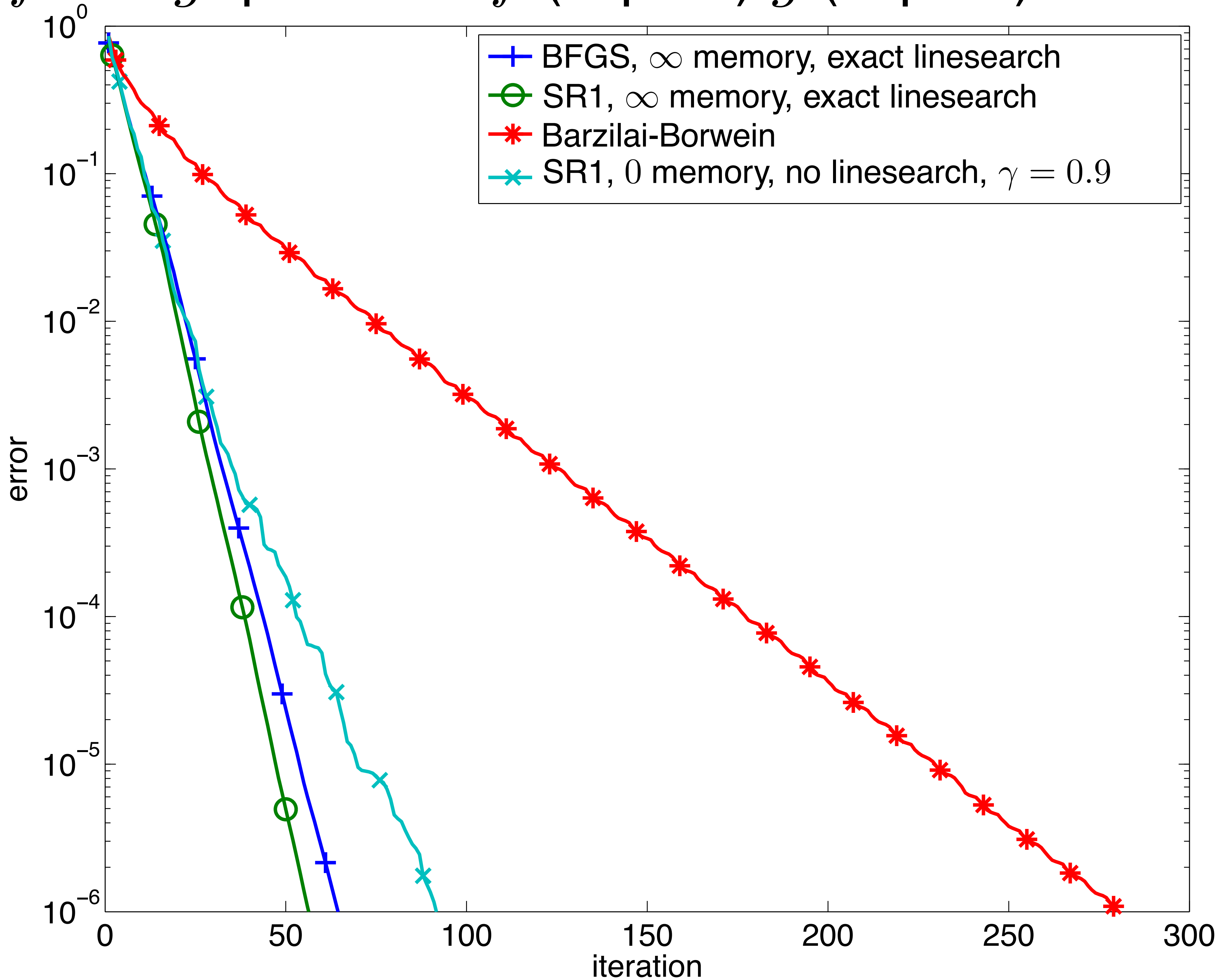
Outline

- Convex analysis with variable metric.
- Convergence.
- LMSRr proximal splitting scheme.
- Applications.
- Extensions and conclusion.

Does the rank-1 term matter ?

$$\min_{x \in \mathbb{R}^N} f(x) + g(x)$$

Both f and g quadratic. f (explicit) g (implicit). $N = 1000$.



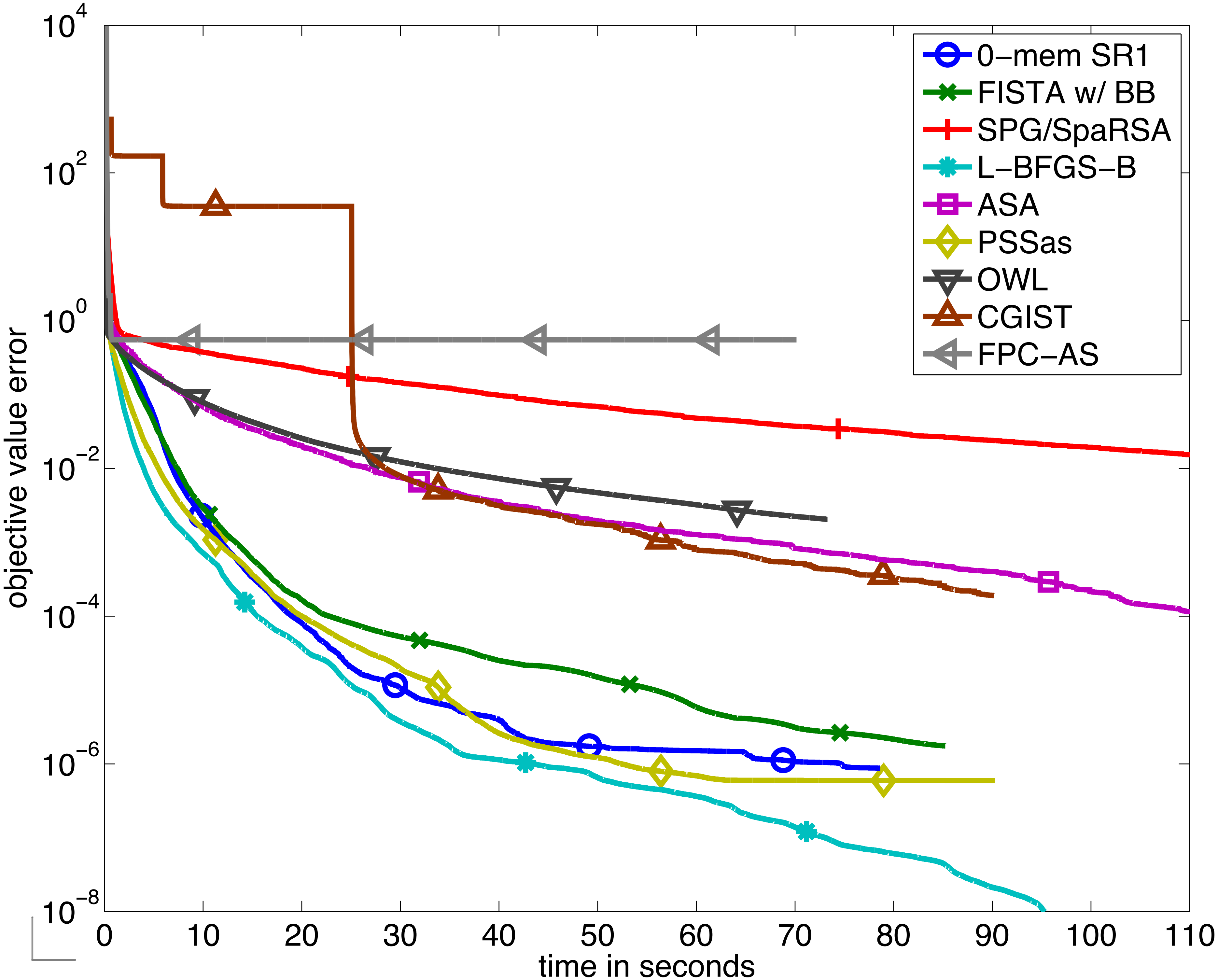
Comparisons

- First-order methods :
  SpaRSA with BB [Wright et al. 2009].
  FISTA [Nesterov 1983, Beck and Teboulle 2009].
- "1.5"-order methods. Most use active-set strategy :
  L-BFGS-B [Byrd et al. 1995].
  ASA "Active Set Algorithm" [Hager and Zhang 2006].
  CGIST "CG + IST" [Goldstein and Setzer 2011].
  FPC-AS "FPC + Active Set" [Wen et al. 2010].
  PSSas "Projected Scaled Sub-gradient + Active Set" [Schmidt et al. 2007].
  OWL "Orthant-wise Learning" [Andrew and Gao 2007].

Test 1: Lasso

$$\min_{x \in \mathbb{R}^N} \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1$$

$$A \in \mathbb{R}^{1500 \times 3000},\ A_{ij} \sim_{\text{iid}} \mathcal{N}(0,1).$$

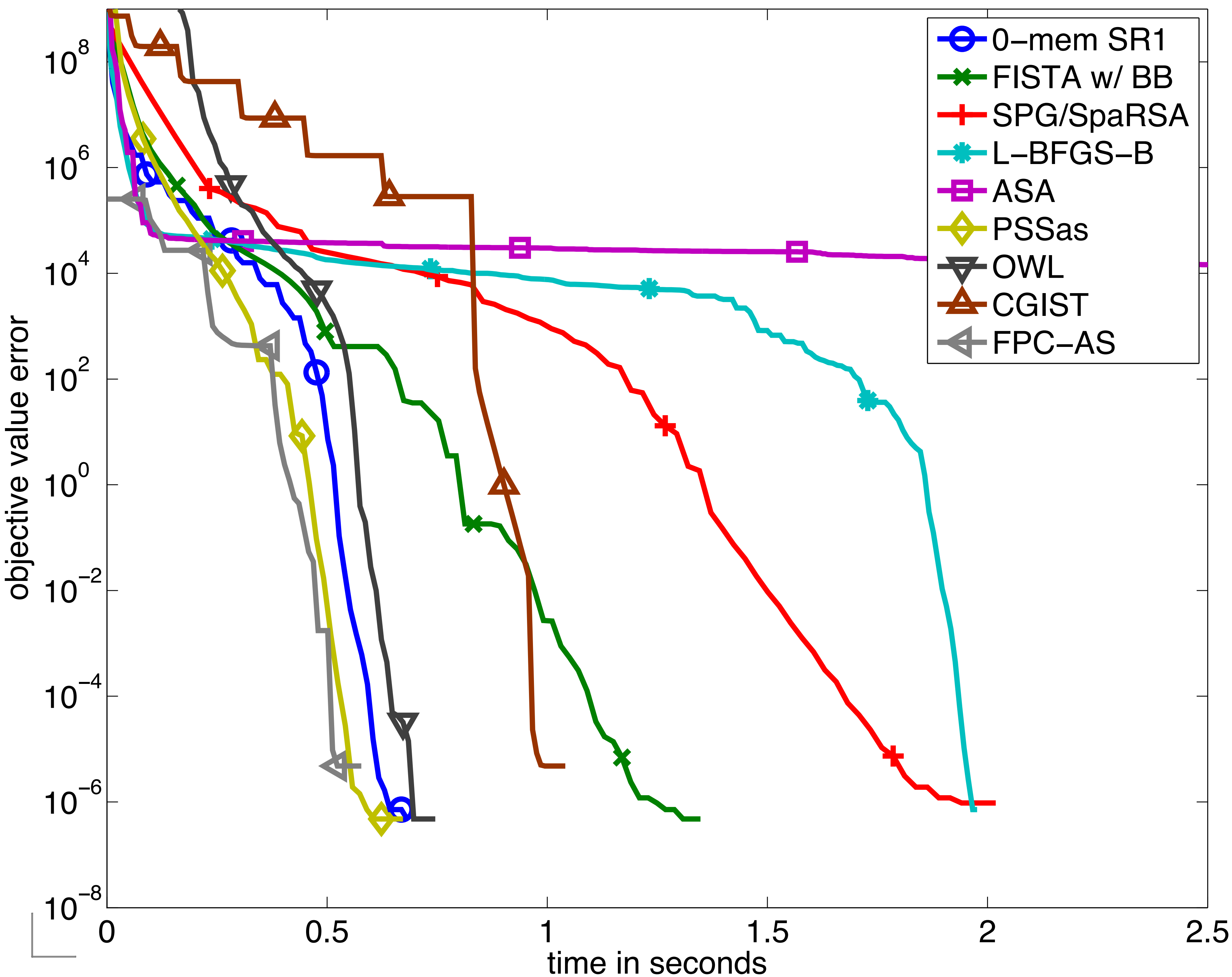


fastest	L-BFGS-B
	0-mem SR1
	PSSas
	FISTA
	ASA
	CGIST
	OWL
	SpaRSA
slowest	FPC-AS

Test 2: Lasso

$$\min_{x \in \mathbb{R}^N} \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1$$

A 3D discrete differential operator, $N = 2197$.



fastest	FPC-AS
	PSSas
	0-mem SR1
	OWL
	FISTA
	CGIST
	SpaRSA
	L-BFGS-B
slowest	ASA

Outline

- Convex analysis with variable metric.
- Convergence.
- LMSRr proximal splitting scheme.
- Applications.
- Extensions and conclusion.

Extensions

- Variable metric GFB splitting [Raguet-F.-Peyré 2011] to solve

$$\min_{x \in \mathcal{H}} f(x) + g_i \circ L_i(x) \ .$$

$\nabla f \in \beta\text{-Lip}(\mathcal{H})$, and $\forall i$, g_i simple and L_i bounded linear operator on $\mathcal{H}$.

- Monotone inclusions :

$$0 \in A(x) + B(x)$$

A and B maximal monotone, A merely Lipschitz (or skewed monotone). This would cover primal-dual splitting.

- Completely non-smooth case : pre-conditioning.
- Inexact versions : robustness to errors.
- Other quasi-Newton metrics with favorable structure.

Take away messages

- Variable metric proximal splitting.
- A new accelerated quasi-Newton forward-backward algorithm.
- Convergence guarantees and rates (special instances).
- An efficient LMSR1 metric construction.
- A new result on the calculation of proximity operators in this metric.
- A Fast solver for large-scale problems.
- Convergence guarantees for LMSRr.
- Many extensions.

Papers and code available
http://www.greyc.ensicaen.fr/~jfadili

Thanks
Any questions ?