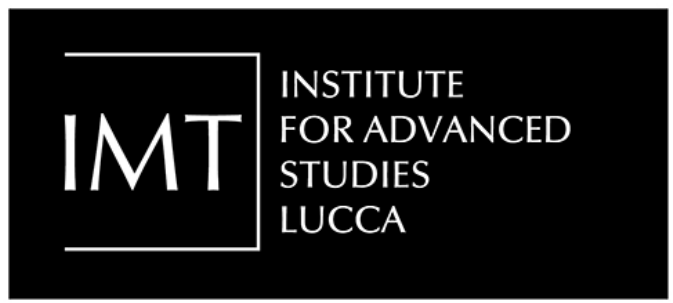
Splitting Envelopes
Accelerated Second-order Proximal Methods

Panos Patrinos
(joint work with Lorenzo Stella, Alberto Bemporad)

September 8, 2014



Outline

- ✓ forward-backward envelope (FBE)
- ✓ forward-backward Newton method (FBN)
- ✓ dual FBE and Augmented Lagrangian
- ✓ alternating minimization Newton method (AMNM)
- ✓ Douglas Rachford envelope (DRE)
- ✓ accelerated Douglas Rachford splitting (ADRS)

based on

1. P. Patrinos and A. Bemporad. *Proximal Newton methods for convex composite optimization*. In Proc. 52nd IEEE Conference on Decision and Control (CDC), pages 2358-2363, Florence, Italy, 2013.
2. P. Patrinos, L. Stella, and A. Bemporad, *Forward-backward truncated Newton methods for convex composite optimization.* submitted, arXiv:1402.6655, 2014.
3. P. Patrinos, L. Stella, and A. Bemporad. *Douglas-Rachford splitting: complexity estimates and accelerated variants*. In Proc. 53rd IEEE Conference on Decision and Control (CDC), Los Angeles, CA, arXiv:1407.6723, 2014.
4. L. Stella, P. Patrinos, and A. Bemporad, *Alternating minimization Newton method for separable convex optimization*, 2014 (submitted).

fixed point implementation for MPC

- ✓ A. Guiggiani, P. Patrinos, and A. Bemporad. *Fixed-point implementation of a proximal Newton method for embedded model predictive control*. In 19th IFAC, South Africa, 2014.

Convex composite optimization

$$\text{minimize} \;\; F(x) = f(x) + g(x)$$

- ✓ $f : \mathbb{R}^n \to \mathbb{R}$ convex, twice continuously differentiable with

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^n$$

- ✓ $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ convex, nonsmooth with inexpensive proximal mapping

$$\text{prox}_{\gamma g}(x) = \underset{z \in \mathbb{R}^n}{\arg\min} \left\{ g(z) + \tfrac{1}{2\gamma} \|z - x\|^2 \right\}$$

- ✓ **many problem classes**: QPs, cone programs, sparse least-squares, rank minimization, total variation minimization,. . .
- ✓ **applications**: control, system identification, signal processing, image analysis, machine learning,. . .

Proximal mappings

$$\mathrm{prox}_{\gamma g}(x) = \underset{z \in \mathbb{R}^n}{\arg\min} \left\{ g(z) + \tfrac{1}{2\gamma}\|z - x\|^2 \right\} \qquad \gamma > 0$$

- ✓ **resolvent** of maximal monotone operator ∂g

$$\mathrm{prox}_{\gamma g}(x) = (I + \gamma \partial g)^{-1}(x)$$

- ✓ single-valued and (firmly) **nonexpansive**
- ✓ explicitly computable for many functions (see Parikh, Boyd '14, Combettes, Pesquet '10)
- ✓ reduces to **projection** when g is indicator of convex set

$$\mathrm{prox}_{\gamma \delta_C}(x) = \Pi_C(x)$$

- ✓ $z = \mathrm{prox}_{\gamma g}(x)$ is **implicit** a subgradient step $(0 \in \partial g(z) + \gamma^{-1}(z - x))$

$$z = x - \gamma v \qquad v \in \partial g(z)$$

Proximal Minimization Algorithm

$$\text{minimize} \;\; g(x), \qquad g : \mathbb{R}^n \to \overline{\mathbb{R}} \text{ closed proper convex}$$

given $x^0 \in \mathbb{R}^n$, repeat

$$x^{k+1} = \text{prox}_{\gamma g}(x^k) \qquad \gamma > 0$$

- ✓ fixed point iteration for optimality conditions

$$0 \in \partial g(x^\star) \iff x^\star \in (I + \gamma \partial g)(x^\star) \iff x^\star = \text{prox}_{\gamma g}(x^\star)$$

- ✓ special case of proximal point algorithm (Martinet '70, Rockafellar '76)
- ✓ converges under very general conditions
- ✓ mostly conceptual algorithm

Moreau envelope

Moreau envelope of closed proper convex $g : \mathbb{R}^n \to \overline{\mathbb{R}}$

$$g^\gamma(x) = \inf_{z \in \mathbb{R}^n} \left\{ g(z) + \tfrac{1}{2\gamma}\|z - x\|^2 \right\}, \qquad \gamma > 0$$

✓ g^γ is **real-valued, convex, differentiable** with $1/\gamma$-Lipschitz gradient

$$\nabla g^\gamma(x) = (1/\gamma)(x - \mathrm{prox}_{\gamma g}(x))$$

✓ minimizing nonsmooth g is equivalent to minimizing smooth g^γ

✓ **proximal minimization algorithm = gradient method for** g^γ

$$x^{k+1} = x^k - \gamma \nabla g^\gamma(x^k)$$

✓ can use any method of unconstrained smooth minimization for g^γ

Forward-Backward Splitting (FBS)

$$\text{minimize} \;\; F(x) = f(x) + g(x)$$

- ✓ **optimality condition:** $x_\star \in \mathbb{R}^n$ is optimal if and only if

$$x_\star = \text{prox}_{\gamma g}(x_\star - \gamma \nabla f(x_\star)), \qquad \gamma > 0$$

- ✓ **forward-backward splitting** (aka **proximal gradient**)

$$x^{k+1} = \text{prox}_{\gamma g}(x^k - \gamma \nabla f(x^k)), \qquad \gamma \in (0, 2/L_f)$$

- ✓ FBS is a **fixed point iteration**
- ✓ $g = 0$: gradient method, $g = \delta_C$: gradient projection, $f = 0$: prox min
- ✓ accelerated versions (Nesterov)

Forward-Backward Envelope

$$x - \mathrm{prox}_{\gamma g}(x - \gamma \nabla f(x)) = 0$$

✓ use $\mathrm{prox}_{\gamma g}(y) = y - \gamma \nabla g^{\gamma}(y)$ for $y = x - \gamma \nabla f(x)$

$$\gamma \nabla f(x) + \gamma \nabla g^{\gamma}(x - \gamma \nabla f(x)) = 0$$

✓ multiply with $\gamma^{-1}(I - \gamma \nabla^2 f(x))$ (positive definite for $\gamma \in (0, 1/L_f)$)

✓ gradient of the **F**orward **B**ackward **E**nvelope (FBE)

$$F_{\gamma}^{\mathrm{FB}}(x) = f(x) - \tfrac{\gamma}{2}\|\nabla f(x)\|_2^2 + g^{\gamma}(x - \gamma \nabla f(x))$$

✓ alternative expression for FBE

$$F_{\gamma}^{\mathrm{FB}}(x) = \inf_{z \in \mathbb{R}^n} \{ \underbrace{f(x) + \langle \nabla f(x), z - x \rangle}_{\text{linearize } f \text{ around } x} + g(z) + \tfrac{1}{2\gamma}\|z - x\|^2 \}$$

Properties of FBE

- ✓ stationary points of F_γ^{FB} = minimizers of F
- ✓ reformulates original **nonsmooth** problem into a **smooth** one

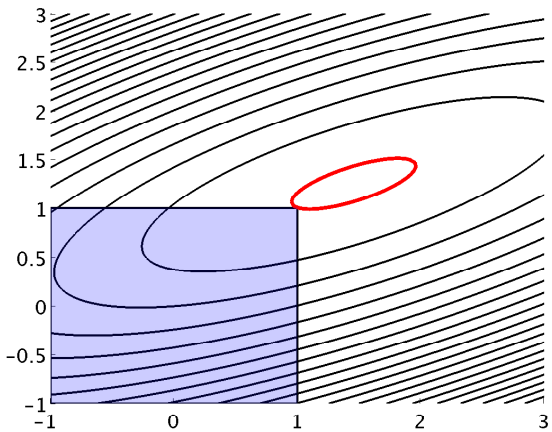
$$\operatorname*{minimize}_{x\in\mathbb{R}^n} \; F_\gamma^{\mathrm{FB}}(x) \qquad \textbf{equivalent to} \qquad \operatorname*{minimize}_{x\in\mathbb{R}^n} \; F(x)$$

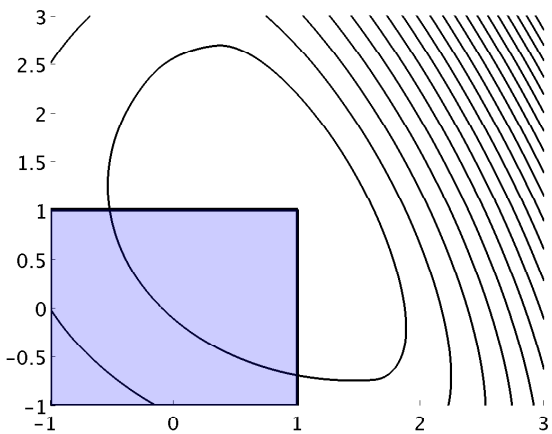
- ✓ F_γ^{FB} is **real-valued, continuously differentiable**

$$\nabla F_\gamma^{\mathrm{FB}}(x) = \gamma^{-1}(I - \gamma\nabla^2 f(x))(x - \operatorname{prox}_{\gamma g}(x - \gamma\nabla f(x))$$

- ✓ FBS is a **variable metric gradient method** for FBE

$$x^{k+1} = x^k - \gamma D_k^{-1}\nabla F_\gamma^{\mathrm{FB}}(x^k)$$





Forward-Backward Newton Method (FBN)

Input: $x^0 \in \mathbb{R}^n$, $\gamma \in (0, 1/L_f)$, $\sigma \in (0, 1/2)$

for $k = 0, 1, 2, \ldots$ **do**

Newton direction

Choose $H^k \in \hat{\partial}^2 F_\gamma^{\mathrm{FB}}(x^k)$. Compute d^k by solving (approximately)

$$H^k d = -\nabla F_\gamma^{\mathrm{FB}}(x^k)$$

Line search

Compute stepsize by backtracking

$$F_\gamma^{\mathrm{FB}}(x^k + \tau_k d^k) \leq F_\gamma^{\mathrm{FB}}(x^k) + \sigma\tau_k \langle \nabla F_\gamma^{\mathrm{FB}}(x^k), d^k \rangle$$

Update: $x^{k+1} = x^k + \tau_k d^k$

end

Linear Newton approximation

FBE is C^1 but not C^2

$$Hd = -\nabla F_\gamma^{\mathrm{FB}}(x) \qquad \text{where}$$

$$\nabla F_\gamma^{\mathrm{FB}}(x) = \gamma^{-1}(I - \gamma\nabla^2 f(x))(x - \mathrm{prox}_{\gamma g}(x - \gamma\nabla f(x))$$

and $\hat{\partial}^2 F_\gamma(x)$ is an **approximate generalized Hessian**

$$H = \gamma^{-1}(I - \gamma\nabla^2 f(x))(I - P(I - \gamma\nabla^2 f(x))) \in \hat{\partial}^2 F_\gamma(x)$$

where $P \in \underbrace{\partial_C(\mathrm{prox}_{\gamma g})}_{\text{Clarke's generalized Jacobian}}(x - \gamma\nabla f(x))$

- ✓ preserves all favorable properties of the Hessian for C^2 functions
- ✓ "Gauss-Newton" generalized Hessian: we omit 3rd order terms

Generalized Jacobians of proximal mappings

✓ $\partial_C \operatorname{prox}_{\gamma g}(x)$ is the following set of matrices (Clarke, 1983)

$$\operatorname{conv}\left\{\begin{array}{l}\text{limits of (ordinary) Jacobians for every sequence that converges}\\ \text{to } x\text{, consisting of points where } \operatorname{prox}_{\gamma g} \text{ is differentiable}\end{array}\right\}$$

- $\operatorname{prox}_{\gamma g}(x)$ simple to compute $\Longrightarrow P \in \partial_C(\operatorname{prox}_{\gamma g})(x)$ for free
- g (block) separable $\Longrightarrow P \in \partial_C(\operatorname{prox}_{\gamma g})(x)$ (block) diagonal

example–ℓ_1 norm more examples in Patrinos, Stella, Bemporad (2014)

✓ $g(x) = \|x\|_1$ $\qquad \operatorname{prox}_{\gamma f}(x)_i = \begin{cases} x_i + \gamma, & \text{if } x_i \leq -\gamma, \\ 0, & \text{if } -\gamma \leq x_i \leq \gamma \\ x_i - \gamma, & \text{if } x_i \geq \gamma \end{cases}$

✓ $P \in \partial_C(\operatorname{prox}_{\gamma g})(x)$ are diagonal matrices with

$$P_{ii} = \begin{cases} 1, & \text{if } i \in \{i \mid |x_i| > \gamma\}, \\ \in [0,1], & \text{if } i \in \{i \mid |x_i| = \gamma\}, \\ 0, & \text{if } i \in \{i \mid |x_i| < \gamma\}. \end{cases}$$

Convergence of FBN

✓ every limit point of $\{x^k\}$ converges to $\arg\min_{x \in \mathbb{R}^n} F(x)$

✓ all $H \in \hat{\partial}^2 F_\gamma(x_\star)$ nonsingular $\Longrightarrow$ **Q-quadratic asymptotic rate**

extension: **FBN II**

✓ apply **FB** step after a **Newton step**

✓ same asymptotic rate + **global complexity estimates**

- non-strongly convex f: **sublinear rate** for $F(x^k) - F(x_\star)$
- strongly convex f: **linear rate** for $\left\{ \begin{array}{c} F(x^k) - F(x_\star) \\ \|x^k - x_\star\|^2 \end{array} \right\}$

FBN–CG

large problems

conjugate gradient (CG) on regularized Newton system until

$$\| \underbrace{(H^k + \delta_k I)d^k + \nabla F_\gamma^{\mathrm{FB}}(x^k)}_{\text{residual}} \| \leq \eta_k \|\nabla F_\gamma^{\mathrm{FB}}(x^k)\|$$

with $\eta_k = O(\|\nabla F_\gamma^{\mathrm{FB}}(x^k)\|)$, $\delta_k = O(\|\nabla F_\gamma^{\mathrm{FB}}(x^k)\|)$

properties

- ✓ **no need to form $\nabla^2 f(x)$ and H^k – only matvec products**
- ✓ same convergence properties

Box-constrained convex programs

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & \ell \leq x \leq u \end{aligned}$$

Newton direction solves

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\langle d, \nabla^2 f(x^k) d\rangle + \langle \nabla f(x^k), d\rangle \\ \text{subject to} \quad & d_i = \ell_i - x_i^k, i \in \beta_1, \quad d_i = u_i - x_i^k, i \in \beta_2 \end{aligned}$$

where

$$\begin{aligned} \beta_1 &= \{i \mid x_i^k - \gamma \nabla_i f(x^k) \leq \ell_i\} \quad \text{estimate of } x_i^\star = \ell_i \\ \beta_2 &= \{i \mid x_i^k - \gamma \nabla_i f(x^k) \geq u_i\} \quad \text{estimate of } x_i^\star = u_i \end{aligned}$$

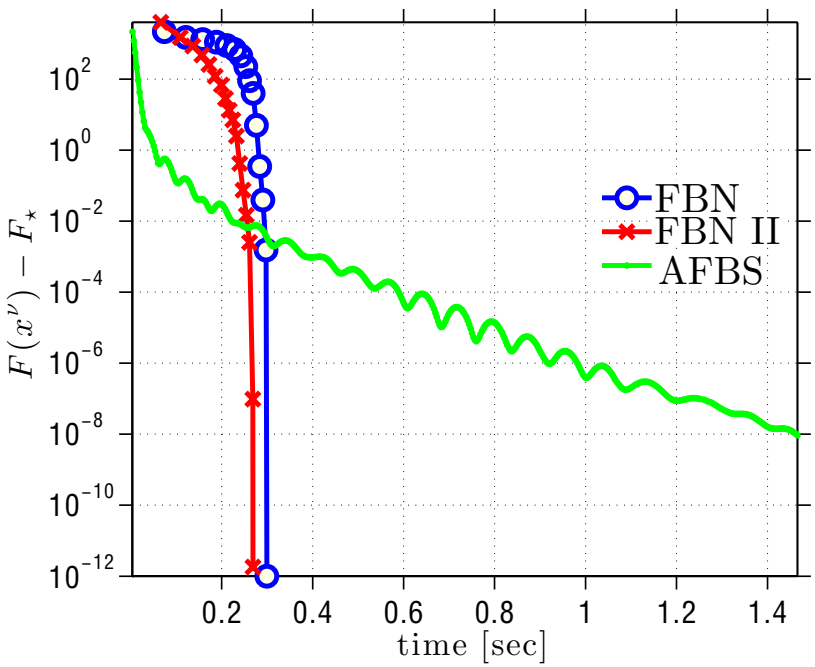
Newton system becomes

$$Q_{\delta\delta} d_\delta = -(\nabla_\delta f(x^k) + \nabla_{\delta\beta} f(x^k) d_\beta), \quad (\beta = \beta_1 \cup \beta_2, \ \delta \setminus \beta = [n])$$

Example

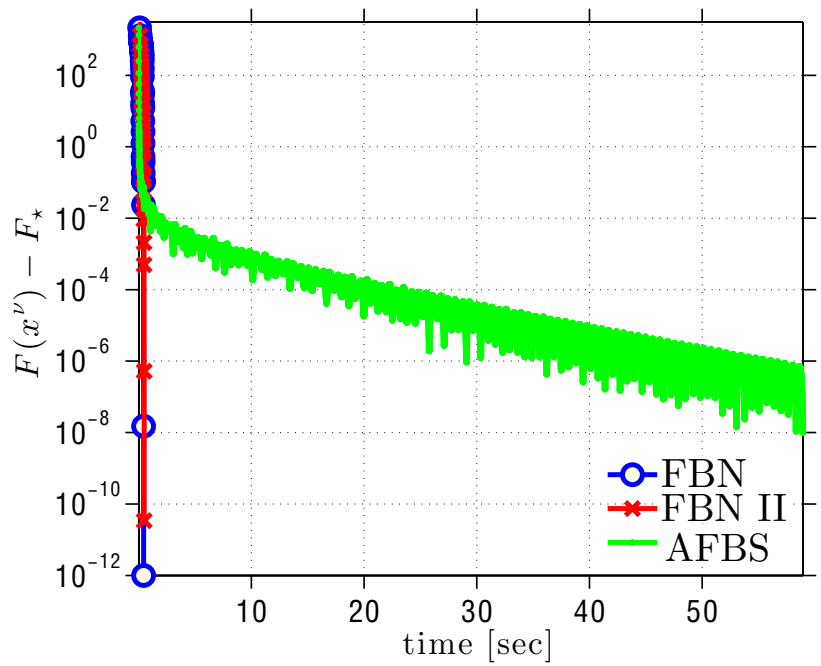
$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\langle x, Qx\rangle + \langle q, x\rangle, \qquad n = 1000 \\ \text{subject to} \quad & \ell \le x \le u \end{aligned}$$

$\mathrm{cond}(Q) = 10^4$



GUROBI: 4.87 sec, CPLEX: 3.73 sec

$\mathrm{cond}(Q) = 10^8$



GUROBI: 5.96 sec, CPLEX: 4.83 sec

FBN: much less sensitive to bad conditioning

Sparse least-squares

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

✓ Newton system becomes

$$\begin{aligned} d_\beta &= -x_\beta \\ A_{\cdot\delta}^\top A_{\cdot\delta} d_\delta &= -[A_{\cdot\delta}^\top (A_{\cdot\delta} x_\delta - b) + \lambda \operatorname{sign}(x_\delta - \gamma \nabla_\delta f(x))] \end{aligned}$$

✓ δ is an estimate of the nonzero components of $x_\star$

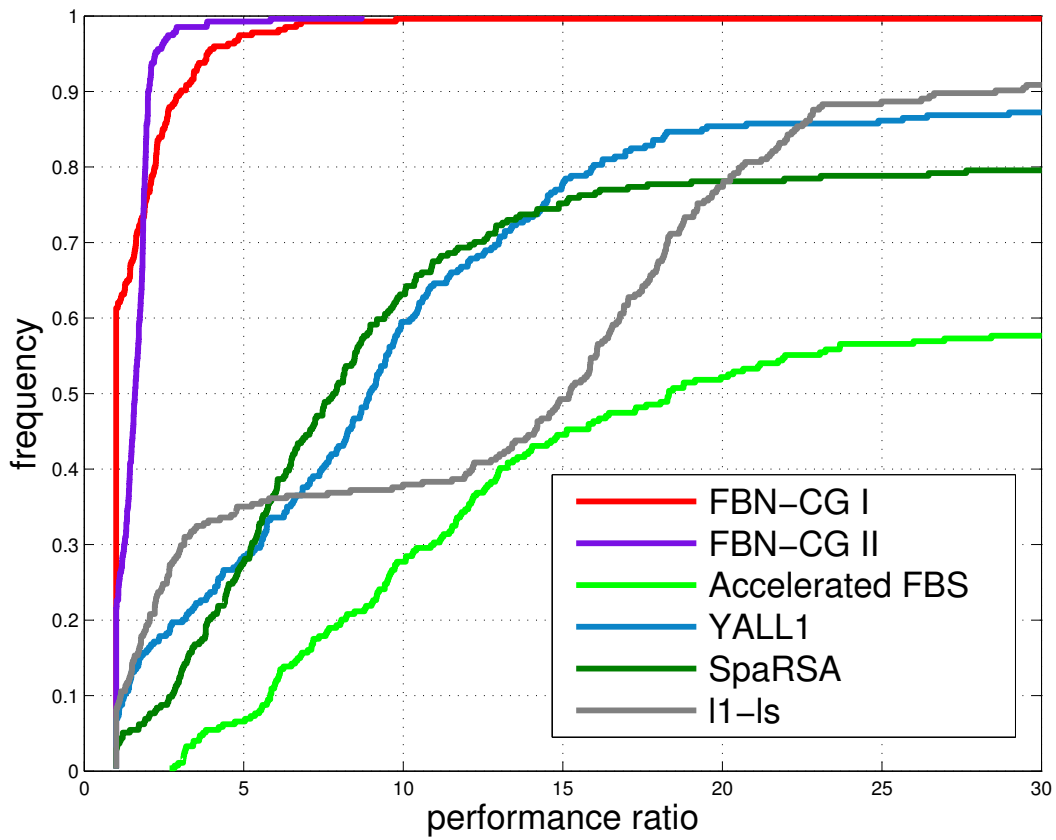
$$\delta = \{i \mid |x_i - \gamma \nabla_i f(x)| > \lambda\gamma\}$$

✓ close to solution δ small

FBN Methods are robust

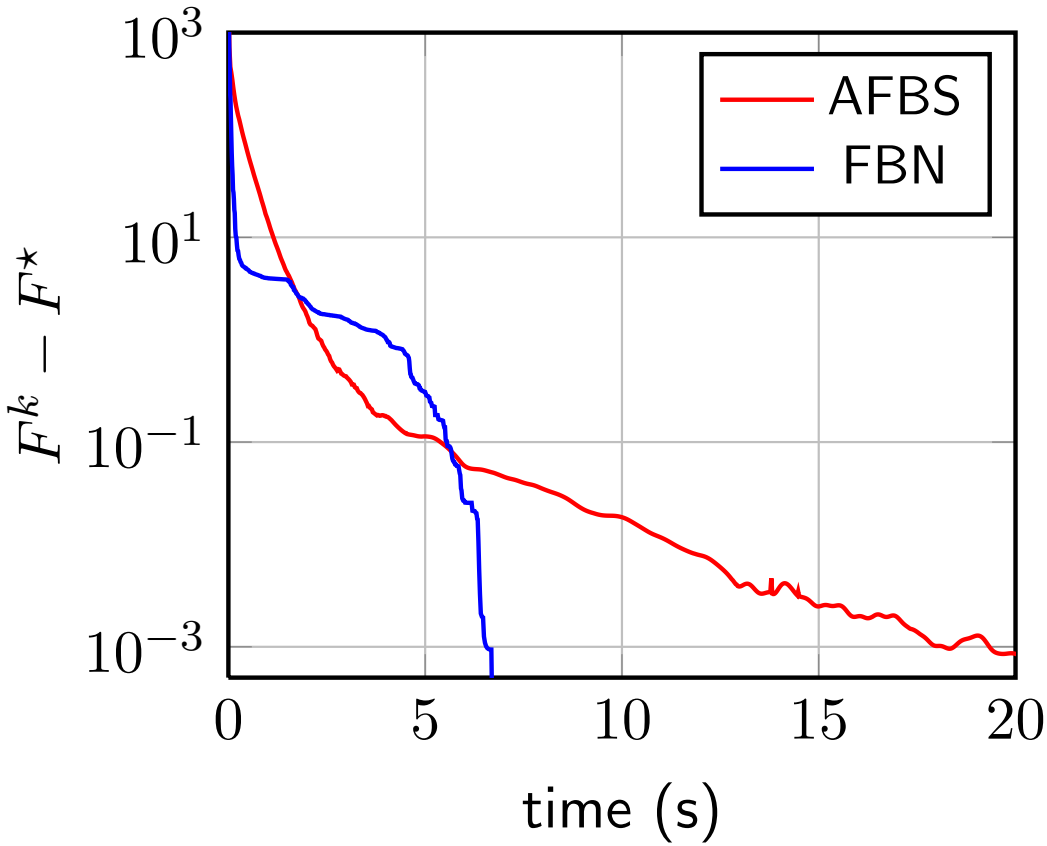
- ✓ 548 datasets taken from

  `http://wwwopt.mathematik.tu-darmstadt.de/spear/`
- ✓ compared against AFBS, YALL1 (ADDM-based), SpaRSA (Barzilai-Borwein), l1-ls (interior point)
- ✓ performance plot: point (x, y) indicates algorithm is at most x times slower in a fraction of y problems



Sparse Logistic Regression

$$\underset{x,y}{\text{minimize}} \sum_{i=1}^{m} \log\left(1 + e^{-b_i(a_i^\intercal x + y)}\right) + \lambda\|x\|_1, \quad \lambda > 0$$



Augmented Lagrangians and Moreau envelopes

$$\begin{array}{lll} \text{minimize} & f(x) & f : \mathbb{R}^n \to \overline{\mathbb{R}} \text{ convex (can be nonsmooth)}, \\ \text{subject to} & Ax = b & A \in \mathbb{R}^{p \times n} \end{array}$$

Augmented Lagrangian

$$L_\gamma(x, y) = f(x) + \langle y, Ax - b \rangle + \tfrac{\gamma}{2} \|Ax - b\|^2$$

Augmented Lagrangian method (ALM or Method of Multipliers)

$$x^k = \inf_{x \in \mathbb{R}^n} L_\gamma(x, y^k) \qquad \text{Hestenes (1969), Powell (1969)}$$
$$y^{k+1} = y^k + \gamma(Ax^k - b)$$

ALM = proximal minimization for dual Rockafellar (1973,1976)

= gradient method for Moreau envelope of dual

Separable convex problems

$$\begin{array}{rll} \text{minimize} & f(x)+g(z), & f,g \text{ convex (can be nonsmooth)} \\ \text{subject to} & Ax+Bz=b & A\in\mathbb{R}^{p\times n}, B\in\mathbb{R}^{p\times m} \end{array}$$

- ✓ f is **strongly convex** with convexity parameter μ_f
- ✓ f, g are nice, e.g. separable. coupling introduced through constraints
- ✓ **ALM** not amenable to decomposition: x and z updates are coupled

Alternating Minimization Method (AMM): FBS applied to the dual

$$\begin{aligned} x^{k+1} &= \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \; L_0(x,z^k,y^k) \\ z^{k+1} &= \operatorname*{arg\,min}_{z\in\mathbb{R}^m} \; L_\gamma(x^{k+1},z^k,y^k) \qquad\qquad \gamma\in(0,2\mu_f/\|A\|^2) \\ y^{k+1} &= y^k+\gamma(Ax^{k+1}+Bz^{k+1}-b) \end{aligned}$$

Augmented Lagrangian

$$L_\gamma(x,z,\lambda)=f(x)+g(z)+\langle y,Ax+Bz-b\rangle+\tfrac{\gamma}{2}\|Ax+Bz-b\|^2$$

Dual FBE

FBE for dual problem is augmented Lagrangian!

$$F_\gamma^{\mathrm{FB}}(y) = L_\gamma(x(y), z(y), y)$$

where $x(y)$, $z(y)$ are the **AMM updates**

$$x(y) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \; f(x) + \langle y, Ax \rangle$$

$$z(y) = \operatorname*{arg\,min}_{z \in \mathbb{R}^m} \; g(z) + \langle y, Bz \rangle + \tfrac{\gamma}{2} \|Ax(y) + Bz - b\|^2$$

Connection between AMM and FBE

dual problem is equivalent to

$$\underset{y\in\mathbb{R}^p}{\text{maximize}}\ F_\gamma^{\mathrm{FB}}(y) = L_\gamma(x(y), z(y), y) \qquad \gamma \in \left(0, \mu_f/\|A\|^2\right)$$

✓ $f \in C^2(\mathbb{R}^n) \implies F_\gamma^{\mathrm{FB}} \in C^1(\mathbb{R}^p)$

$$\nabla F_\gamma^{\mathrm{FB}}(y) = \overbrace{\left(I - \gamma A(\nabla^2 f(x(y)))^{-1}A^\top\right)}^{D(y)}(Ax(y) + Bz(y) - b)$$

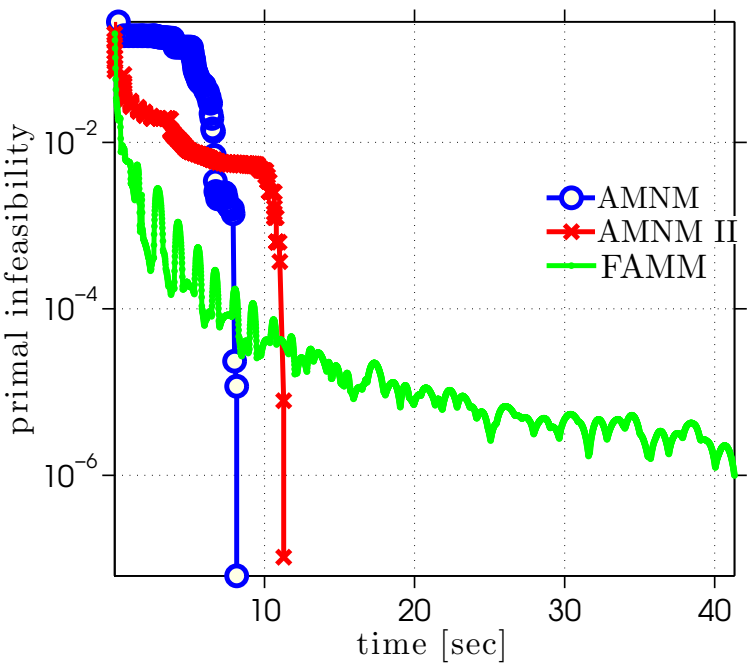
✓ **AMM** as a **variable metric gradient method** on **dual FBE**

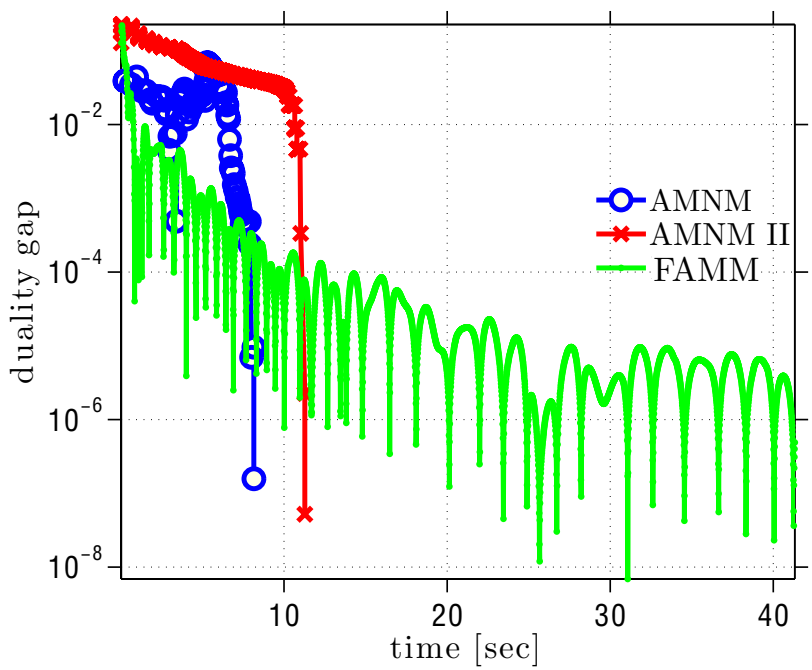
$$y^{k+1} = y^k + \gamma D(y^k)^{-1}\nabla F_\gamma^{\mathrm{FB}}(y^k)$$

✓ **AMNM: FBN to dual**

Strictly convex QPs

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\langle x, Qx\rangle + \langle q, x\rangle & \text{cond}(Q) = 10^4 \\ \text{subject to} & \ell \leq Ax \leq u & A \in \mathbb{R}^{2000\times 1000} \end{array}$$





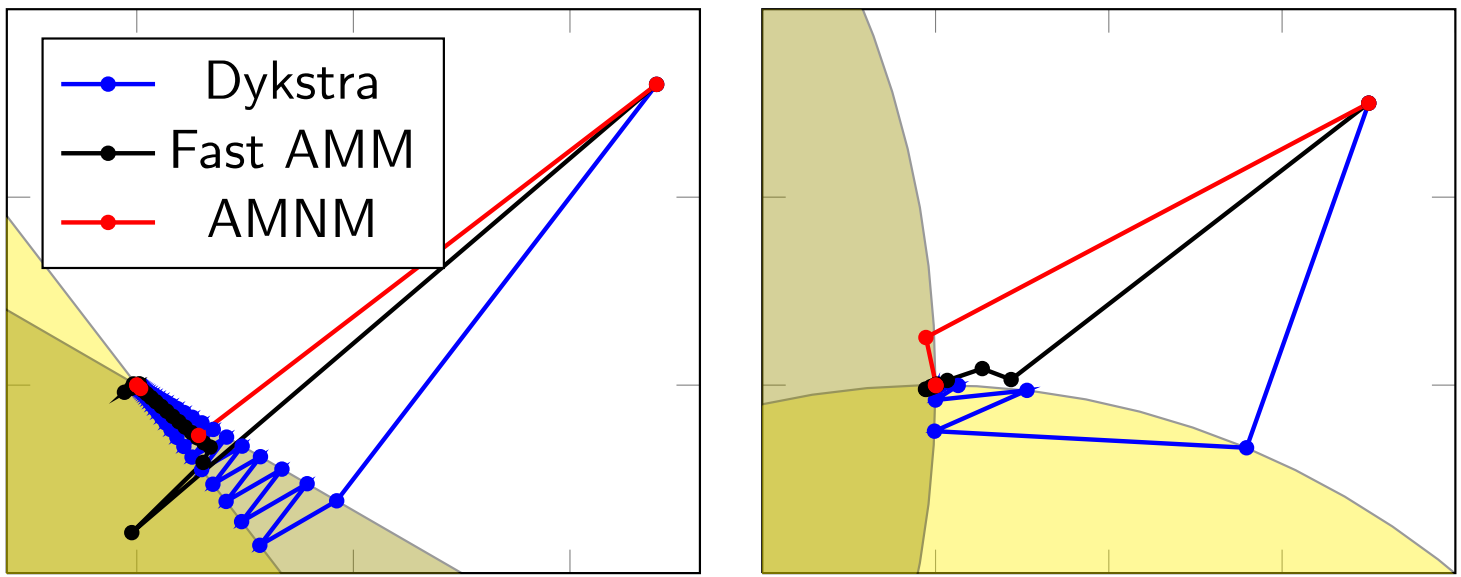
GUROBI: 6.7 s **CPLEX**: 60.8 s **Fast AMM**: 41 s **AMNM**: 8.1 s **AMNM II**:11.3 s

Projection onto Convex Sets

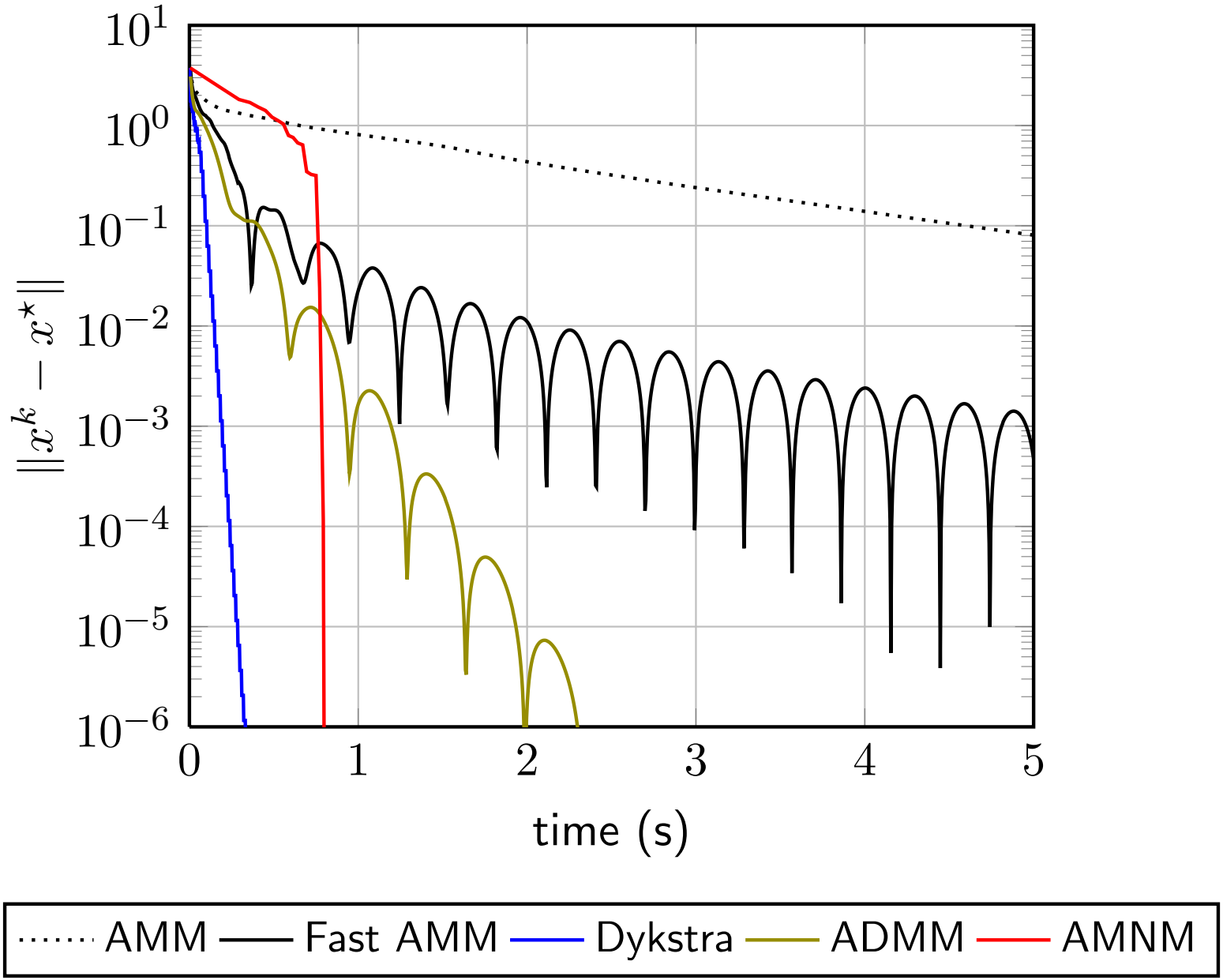
$$\underset{x}{\text{minimize}} \ \frac{1}{2}\|x-p\|^2 + \sum_{i=1}^{m} \delta_{C_i}(z_i)$$

$$\text{subject to } x = z_i, \quad i = 1, \ldots, M$$

- ✓ δ_C is the **indicator** of C
- ✓ $x^\star$ is the **projection** of p onto $C_1 \cap \ldots \cap C_m$



Random problem with $m = 100$ random hyperplanes in $\mathbb{R}^{120}$



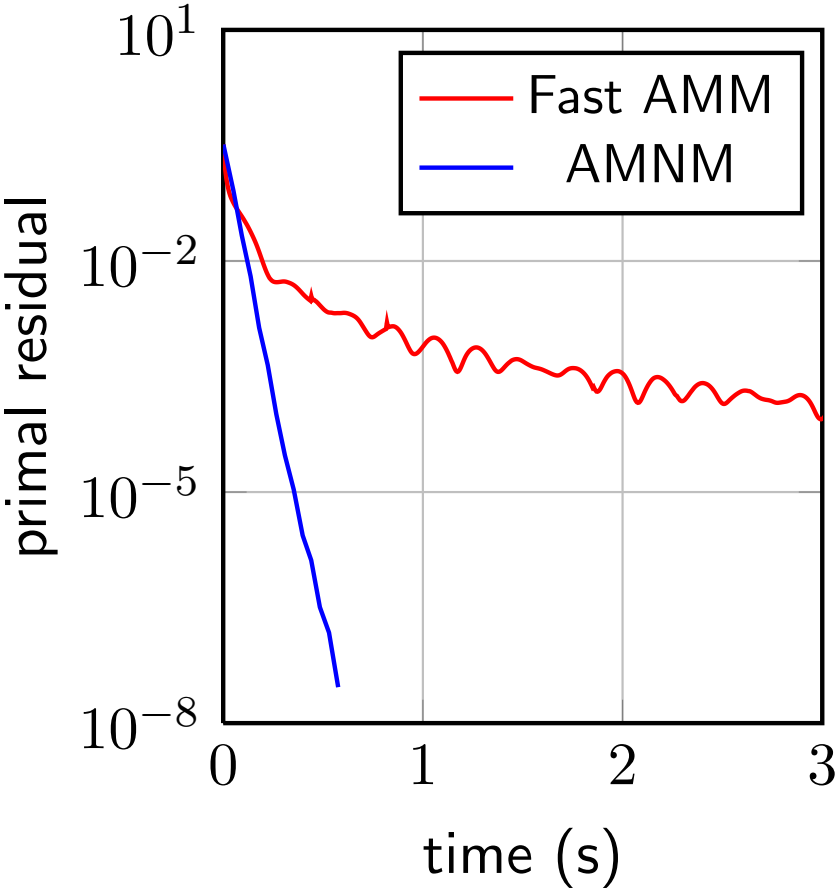
Distributed MPC

$$\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{M} \sum_{t=1}^{N-1} \left(\|\xi_i(t)\|_{Q_i}^2 + \|u_i(t)\|_{R_i}^2 \right) + \|\xi_i(N)\|_{P_i}^2 \\
\text{subject to} \quad & \xi_i(0) = \xi_i^0, && i \in \mathbb{N}_{[1,M]} \\
& \xi_i(t+1) = \sum_{j \in \mathcal{N}_i} \Phi_{ij} \xi_j(t) + \Gamma_{ij} u_j(t), && t \in \mathbb{N}_{[0,N-1]}, i \in \mathbb{N}_{[1,M]} \\
& (\xi_i(t), u_i(t)) \in Y_i, && t \in \mathbb{N}_{[0,N-1]}, i \in \mathbb{N}_{[1,M]} \\
& \xi_i(N) \in Z_i, && i \in \mathbb{N}_{[1,M]}
\end{aligned}$$

- ✓ solve an optimal control problem over a network of agents
- ✓ local constraint sets Y_i, Z_i are simple, coupling in the dynamics
- ✓ can handle complicated local and coupled constraints as well

DMPC simulations

$M = 100$ subsystems
23600 vars, 20600 cons



average over 20 instances
8 states, 3 inputs, $N = 10$

	FAMM	AMNM	
M	local	local	global
5	29.5	2.1	2.1
10	47.0	2.2	2.1
20	65.7	2.4	2.4
50	104.8	3.3	3.3
100	139.2	3.9	3.8
200	159.3	4.4	4.3

communication rounds
(in thousands)

Douglas-Rachford Splitting

$$\text{minimize } F(x) = f(x) + g(x)$$

- ✓ **optimality condition**: $x_\star$ is optimal if and only if $x_\star = \text{prox}_{\gamma f}(\tilde{x})$, where $\tilde{x}$ solves

$$\text{prox}_{\gamma g}(2\,\text{prox}_{\gamma f}(x) - x) - \text{prox}_{\gamma f}(x) = 0$$

- ✓ DRS

$$\begin{aligned} y^k &= \text{prox}_{\gamma f}(x^k) \\ z^k &= \text{prox}_{\gamma g}(2y^k - x^k) \\ x^{k+1} &= x^k + \lambda_k(z^k - y^k) \end{aligned}$$

- ✓ $\gamma > 0$ and $\lambda_k \in [0, 2]$ with $\sum_{k\in\mathbb{N}} \lambda_k(2 - \lambda_k) = +\infty$
- ✓ DRS is a **relaxed fixed point iteration**

ADMM

$$\begin{array}{rll} \text{minimize} & f(x) + g(z), & f, g \text{ convex (can be nonsmooth)} \\ \text{subject to} & Ax + Bz = b & A \in \mathbb{R}^{p \times n}, B \in \mathbb{R}^{p \times m} \end{array}$$

Alternating **D**irection **M**ethod of **M**ultipliers

$$\begin{aligned} x^{k+1} &= \underset{x \in \mathbb{R}^n}{\arg\min}\ L_\gamma(x, z^k, y^k) \\ z^{k+1} &= \underset{z \in \mathbb{R}^m}{\arg\min}\ L_\gamma(x^{k+1}, z, y^k) \\ y^{k+1} &= y^k + \gamma(Ax^{k+1} + Bz^{k+1} - b) \end{aligned}$$

DRS applied to the dual (Eckstein, Bertsekas, 1992)

Douglas Rachford Envelope

assume f convex, $C^2 \implies$ Moreau envelope f^γ is C^2

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n$$

✓ optimality condition

$$\operatorname{prox}_{\gamma f}(x) - \operatorname{prox}_{\gamma g}(2\operatorname{prox}_{\gamma f}(x) - x) = 0$$

✓ use $\nabla h^\gamma(x) = \gamma^{-1}(x - \operatorname{prox}_{\gamma h}(x))$

$$\nabla f^\gamma(x) + \nabla g^\gamma(x - 2\gamma \nabla f^\gamma(x)) = 0$$

✓ multiply by $(I - 2\gamma\nabla^2 f^\gamma(x))$, $\gamma \in (0, 1/L_f)$ and "integrate"

Douglas **R**achford **E**nvelope (DRE)

$$F_\gamma^{\mathrm{DR}}(x) = f^\gamma(x) - \gamma\|\nabla f^\gamma(x)\|^2 + g^\gamma(x - 2\gamma\nabla f^\gamma(x))$$

Properties of DRE

✓ $\gamma \in (0, 1/L_f)$: miniziming DRE $=$ minimizing nonsmooth F

$$\begin{aligned} \inf \;\; F(x) &= \inf \;\; F_\gamma^{\mathrm{DR}}(x) \\ \arg\min \;\; F(x) &= \mathrm{prox}_{\gamma f}(\arg\min \;\; F_\gamma^{\mathrm{DR}}(x)) \end{aligned}$$

✓ $f \in C^2(\mathbb{R}^n) \implies$ DRE is C^1 on $\mathbb{R}^n$

✓ f quadratic $\implies$ DRE is convex with $(1/\gamma)$-Lipschitz gradient

Connection between DRE and FBE

✓ partial linearization of F around x

$$\ell_F(z;x) = f(x) + \langle \nabla f(x), z - x \rangle + g(z)$$

✓ FBE is

$$F_\gamma^{\mathrm{FB}}(x) = \min_{z \in \mathbb{R}^n} \left\{ \ell_F(z;x) + \tfrac{1}{2\gamma}\|z - x\|^2 \right\}$$

✓ DRE is [use $\nabla f^\gamma(x) = \gamma^{-1}(x - \mathrm{prox}_{\gamma f}(x)) = \nabla f(\mathrm{prox}_{\gamma f}(x))$]

$$F_\gamma^{\mathrm{DR}}(x) = \min_{z \in \mathbb{R}^n} \left\{ \ell_F(z; \mathrm{prox}_{\gamma f}(x)) + \tfrac{1}{2\gamma}\|z - \mathrm{prox}_{\gamma f}(x)\|^2 \right\}$$

✓ DRE is equal to FBE evaluated at $\mathrm{prox}_{\gamma f}(x)$

$$F_\gamma^{\mathrm{DR}}(x) = F_\gamma^{\mathrm{FB}}(\mathrm{prox}_{\gamma f}(x))$$

Connection between DRS and FBS

✓ FBS (with relaxation)

$$x^{k+1} = x^k + \lambda_k \left(\text{prox}_{\gamma g} \left(x^k - \gamma \nabla f(x^k) \right) - x^k \right)$$

✓ DRS

$$y^k = \text{prox}_{\gamma f}(x^k)$$
$$x^{k+1} = x^k + \lambda_k \left(\text{prox}_{\gamma g}(2y^k - x^k) - y^k \right)$$

✓ $f \in C^1(\mathbb{R}^n)$

$$y^k = \text{prox}_{\gamma f}(x^k) = x^k - \gamma \nabla f(\text{prox}_{\gamma f}(x^k))$$

✓ DRS becomes

$$y^k = \text{prox}_{\gamma f}(x^k)$$
$$x^{k+1} = x^k + \lambda_k \left(\text{prox}_{\gamma g} \left(y^k - \gamma \nabla f(y^k) \right) - y^k \right)$$

✓ **DRS** iteration = **FBS** iteration at "shifted" point $y^k = \text{prox}_{\gamma f}(x^k)$

DRS as a variable metric method

DRS

$$x^{k+1} = x^k + \lambda_k \left(\mathrm{prox}_{\gamma g}(2\,\mathrm{prox}_{\gamma f}(x^k) - x^k) - \mathrm{prox}_{\gamma f}(x^k) \right)$$

gradient of DRE

$$\nabla F_\gamma^{\mathrm{DR}}(x) = \left(I - 2\gamma \nabla^2 f^\gamma(x^k) \right) \left(\mathrm{prox}_{\gamma f}(x^k) - \mathrm{prox}_{\gamma g}(2\,\mathrm{prox}_{\gamma f}(x^k) - x^k) \right)$$

DRS: variable metric method applied to DRE

$$x^{k+1} = x^k - \lambda_k D^k \nabla F_\gamma^{\mathrm{DR}}(x^k)$$

where $D^k = (I - 2\gamma \nabla^2 f^\gamma(x^k))^{-1}$

- ✓ relaxation parameter λ_k of DRS $\implies$ stepsize for gradient method
- ✓ can use backtracking for selecting λ_k

DRS – complexity estimates

- ✓ assume f quadratic $\implies$ DRE is convex for $\gamma \in (0, 1/L_f)$
- ✓ DRS is preconditioned gradient method under change of variables

$$x = Sw \qquad\qquad S = D^{1/2}$$

- ✓ convergence rate of DRS $\qquad \lambda_k = \lambda = (1 - \gamma L_f)/(1 + \gamma L_f)$

$$F(z^{k+1}) - F_\star \leq \frac{1}{(2\gamma\lambda)k}\|x^0 - \tilde{x}\|^2$$

- ✓ optimal prox-parameter γ for DRE

$$\gamma_\star = \frac{\sqrt{2} - 1}{L_f}$$

- ✓ linear convergence rate if F is strongly convex

Accelerated DRS

- ✓ DRE is **convex** with $(1/\gamma)$-Lipschitz gradient for f **quadratic** and $\gamma \in (0, 1/L_f)$
- ✓ **Nesterov's FGM** applied to (preconditioned) DRE: $x^0 = x^{-1} \in \mathbb{R}^n$

$$
\begin{aligned}
u^k &= x^k + \beta_k(x^k - x^{k-1}) \\
y^k &= \mathrm{prox}_{\gamma f}(u^k) \\
z^k &= \mathrm{prox}_{\gamma g}(2y^k - u^k) \\
x^{k+1} &= u^k + \lambda(z^k - y^k)
\end{aligned}
$$

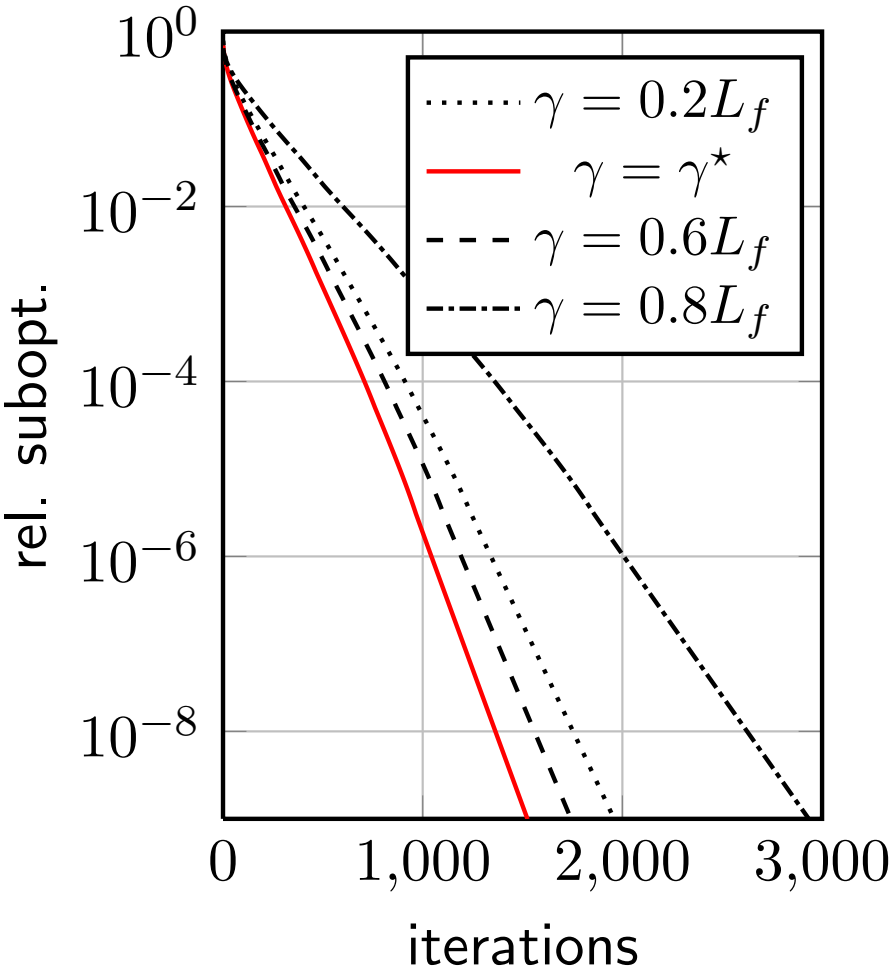
- ✓ $\lambda = (1 - \gamma L_f)/(1 + \gamma L_f)$. can choose $\beta_k = \frac{k-1}{k+2}$
- ✓ convergence rate is $O(1/k^2)$

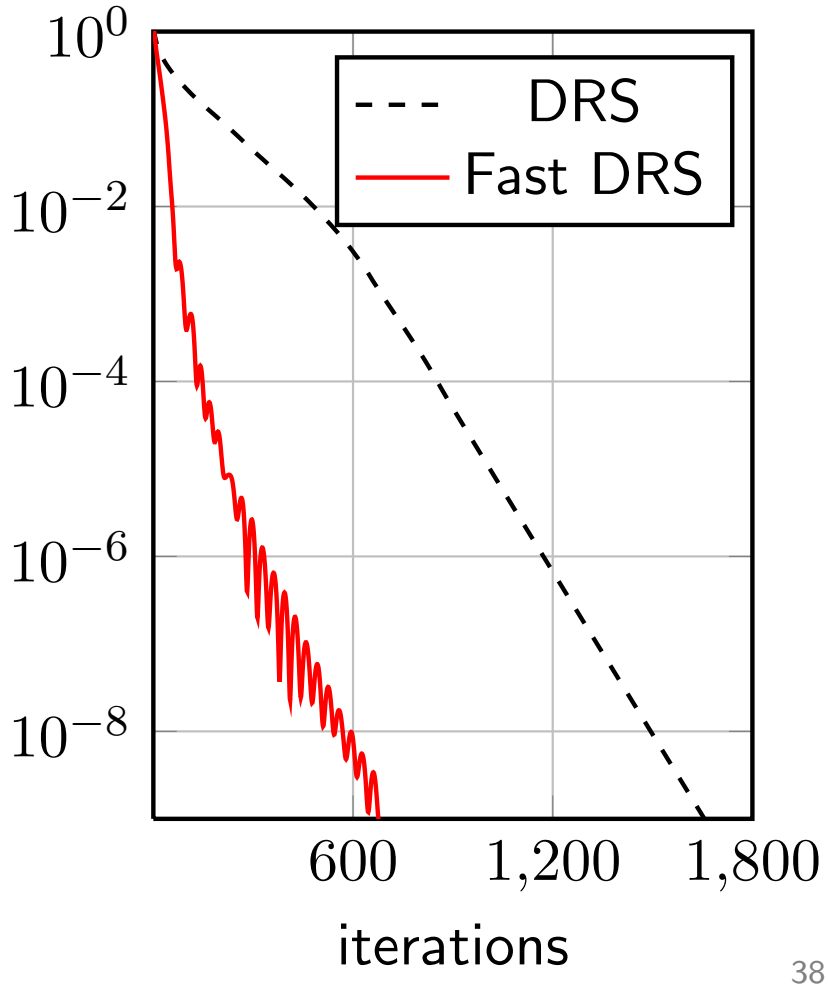
$$
F(z^k) - F_\star \leq \frac{4}{\gamma\lambda(k+2)^2}\|x^0 - \tilde{x}\|^2.
$$

- ✓ linear convergence for f or g strongly convex

Sparse least-squares

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





Take home message I

- ✓ proximal minimization = gradient method on Moreau envelope (70's)
- ✓ **FBS** = (variable metric) gradient method on **FBE** (this talk)
- ✓ **DRS** = (variable metric) gradient method on **DRE** (this talk)

Take home message II

✓ ALM = proximal minimization for dual

$$\text{Moreau envelope of dual} = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} L_\gamma(x, y)$$

(Rockafellar, 1973)

✓ **AMM** = **FBS** for dual

$$\textbf{FBE of dual} = L_\gamma(x(y), z(y), y)$$

where

$$x(y) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} L_0(x, z, y) \qquad z(y) = \operatorname*{arg\,min}_{z \in \mathbb{R}^n} L_\gamma(x(y), z, y)$$

to conclude

✓ interpretation of operator splitting algorithms as gradient methods

✓ splitting envelopes can lead to new exciting algorithms

✓ examples: FBN, AMNM and ADRS