Primal-dual coordinate descent

Olivier Fercoq

Joint work with P. Bianchi & W. Hachem

15 July 2015

Problem

Minimize the convex function

$$\min_{x \in \mathbb{R}^n} f(x) + g(x) + h(Mx)$$

- f, g, h convex
- f is differentiable
- $(I + \tau \partial g)^{-1}$ and $(I + \sigma \partial h^*)^{-1}$ are easy to compute
- $M : \mathbb{R}^n \to \mathbb{R}^m$ is linear

Equivalent to saddle point problem if $0 \in \mathrm{ri}(M \mathrm{dom}\, g - \mathrm{dom}\, h)$

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} f(x) + g(x) - h^*(y) + \langle Mx, y \rangle$$

Examples: Lasso

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda \|x\|_1$$

- $f(x) = \|Ax - b\|_2^2$
- $g(x) = \lambda \|x\|_1$
- $h(x) = 0$

Examples: dual of Support Vector Machines

$$\min_{x\in\mathbb{R}^n} \frac{1}{2\lambda n^2}\sum_{j=1}^{m}\left(\sum_{i=1}^{n} b_i A_{ji} x^{(i)}\right)^2 - \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

$$\text{st}: x \in [0,1]$$

$$b^T x = 0$$

- $f(x) = \frac{1}{2\lambda n^2}\left(\sum_{i=1}^{N} b_i A_{ji} x^{(i)}\right)^2 - \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$
- $g(x) = I_{[0,1]^N}(x)$
- $h(y) = I_{b^\perp}(y)$
- $Mx = x$

Examples: L_1 + TV regularized regression

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \alpha_1 \|x\|_1 + \alpha_2 \|\nabla x\|_{2,1}$$

- $f(x) = \|Ax - b\|_2^2$
- $g(x) = \alpha_1 \|x\|_1$
- $\nabla =$ discrete gradient
- $h(y) = \alpha_2 \|y\|_{2,1} = \alpha_2 \sum_{i=1}^{n} \sqrt{y_{i,1}^2 + y_{i,2}^2 + y_{i,3}^2}$

Part 1: Classical coordinate descent

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

- $h = 0$
- g is separable
- f has a coordinate-wise Lipschitz gradient:
  $\forall x \in \mathbb{R}^n, \forall i \in \{1, \ldots, n\}, \forall t \in \mathbb{R},$

$$|\nabla_i f(x + t e_i) - \nabla_i f(x)| \leq \beta_i |t|$$

Coordinate descent

At iteration k:

1. Choose randomly a coordinate i_{k+1}
2. $\bar{x}_{k+1} = (I + \tau \partial g)^{-1}\big(x_k - \tau \nabla f(x_k)\big)$
3. Update: $x_{k+1} = \begin{cases} \bar{x}^i_{k+1} & \text{if } i = i_{k+1} \\ x^i_k & \text{if } i \neq i_{k+1} \end{cases}$

Remarks

- As g is separable, one only needs to compute $\bar{x}^{i_{k+1}}_{k+1}$
- Convergence if $\dfrac{1}{\tau_i} < \dfrac{\beta_i}{2}$ for all i [Richtárik, Takáč, 2011]

Part 2: Stochastic primal-dual coordinate descent

$$\min_{x \in \mathbb{R}^n} f(x) + g(x) + h(Mx)$$

- ∇f is $L(\nabla f)$-Lipschitz
- g and h need not be separable
- $M \in \mathbb{R}^{m \times n}$

Vũ-Condat's algorithm

[Vũ + Condat, 2013]

$$y_{k+1} = (I + \sigma \partial h^*)^{-1}\big(y_k + \sigma M x_k\big)$$
$$x_{k+1} = (I + \tau \partial g)^{-1}\big(x_k - \tau M^*(2y_{k+1} - y_k) - \tau \nabla f(x_k)\big)$$

- Generalizes Chambolle-Pock, ADMM and proximal gradient
- Converges to a saddle point of the Lagrangian
- Proof: fixed point algorithm for a firmly nonexpansive operator

Vũ-Condat's algorithm

$$y_{k+1} = (I + \sigma \partial h^*)^{-1}\big(y_k + \sigma M x_k\big)$$
$$x_{k+1} = (I + \tau \partial g)^{-1}\big(x_k - \tau M^*(2y_{k+1} - y_k) - \tau \nabla f(x_k)\big)$$

- The fixed point operator is:

$$T\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \overbrace{(I + \sigma \partial h^*)^{-1}\big(y + \sigma M x\big)}^{\bar{y}} \\ (I + \tau \partial g)^{-1}\big(x - \tau M^*(2\bar{y} - y) - \tau \nabla f(x)\big) \end{pmatrix}$$

- Convergence as soon as $\dfrac{1}{\tau} - \sigma \|M\|^2 < \dfrac{L(\nabla f)}{2}$

Coordinate descent for firmly nonexpansive operators

[Bianchi, Hachem & Iutzeler + Combettes & Pesquet, 2014]

At iteration k:

1. Choose randomly a (block of) coordinate i_{k+1}
2. $\bar{z}_{k+1} = T(z_k)$
3. Update: $z_{k+1} = \begin{cases} \bar{z}^i_{k+1} & \text{if } i = i_{k+1} \\ z^i_k & \text{if } i \neq i_{k+1} \end{cases}$

Convergence: if $T = \alpha R + (1-\alpha) I$ where $\alpha \in (0,1)$ and R is nonexpansive in a separable norm

Primal-dual coordinate descent

- We take Vũ-Condat's fixed point operator

$$T\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \overbrace{(I + \sigma\partial h^*)^{-1}\big(y + \sigma M x\big)}^{\bar{y}} \\ (I + \tau\partial g)^{-1}\big(x - \tau M^*(2\bar{y} - y) - \tau\nabla f(x)\big) \end{pmatrix}$$

- Assume that M is block diagonal and define blocks of primal-dual variables $z = (y, x)$ accordingly
- The algorithm is

$$z_{k+1} = \begin{cases} T_i(z_k) & \text{if } i = i_{k+1} \\ z_k^i & \text{if } i \neq i_{k+1} \end{cases}$$

- Convergence as soon as $\dfrac{1}{\tau} - \sigma\|M\|^2 < \dfrac{L(\nabla f)}{2}$

Speed of convergence

Theorem

Assume that for all $i \in \{1, \ldots, n\}$, $p_i = \mathbf{P}(i_{k+1} = i) > 0$ *and* $\tau^{-1} > \sigma \|M\|^2 + \frac{L(\nabla f)}{2}$.
Denote $\hat{\bar{x}}_{k+1} = \frac{1}{k}(\sum_{l=1}^{k+1} \bar{x}_l)$ *and* $\hat{\bar{y}}_{k+1} = \frac{1}{k}(\sum_{l=1}^{k+1} \bar{y}_l)$.
Then for all $(y, x) \in \mathbb{R}^m \times \mathbb{R}^n$,

$$\begin{aligned}
&\mathbf{E}[\mathcal{L}(\hat{\bar{x}}_{k+1}, y) - \mathcal{L}(x, \hat{\bar{y}}_{k+1})] \\
&\qquad \leq \frac{1}{2k} \max\left(1, 1 + (\frac{1}{\kappa} - 1)\frac{1}{1 - 1/(2\kappa)}\right) \|z - z_0\|^2_{p^{-1}P}
\end{aligned}$$

where $\kappa = (\tau^{-1} - \sigma\|M\|^2)/L(\nabla f) > \frac{1}{2}$ *and* $P = \begin{bmatrix} \tau^{-1} I & M \\ M^T & \sigma^{-1} I \end{bmatrix}$

Duplication trick

$$M = \begin{pmatrix} M_{1,1} & M_{1,2} & 0 \\ 0 & M_{2,2} & 0 \\ M_{3,1} & M_{3,2} & M_{3,3} \end{pmatrix} \longrightarrow K = \begin{pmatrix} M_{1,1} & 0 & 0 \\ 0 & M_{1,2} & 0 \\ 0 & M_{2,2} & 0 \\ M_{3,1} & 0 & 0 \\ 0 & M_{3,2} & 0 \\ 0 & 0 & M_{3,3} \end{pmatrix}$$

Define $S = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$ and $D(m) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$

We have $D(m)SK = M$. We define $\overline{h} = h \circ (D(m)S)$:
$\text{prox}_{m\sigma,\overline{h}^*}(\mathbf{y}) = (\mathbf{1}_{m_1} \otimes \text{prox}^{(1)}_{\sigma,h^*}(S(\mathbf{y})), \ldots, \mathbf{1}_{m_p} \otimes \text{prox}^{(p)}_{\sigma,h^*}(S(\mathbf{y})))$

Part 3: Stochastic primal-dual coordinate descent with long steps

Comparison of coordinate descent algorithms

Classical	Primal-dual
$f(x) + \sum_{i=1}^{n} g_i(x^i)$ g separable	$f(x) + g(x) + h(Kx)$ g and h non-separable ✔ M: additional coupling ✔
$\lvert\nabla_i f(x + te_i) - \nabla_i f(x)\rvert \leq \beta_i \lvert t\rvert$ Longer steps ✔	Bases on operator T: $\beta = L(\nabla f)$
Speed in $O(1/k)$ ✔	Speed in $O(1/k)$ ✔

Combine advantages of both approaches

$$\min_{x \in \mathbb{R}^n} f(x) + g(x) + h(Mx)$$

- f has a coordinate-wise Lipschitz gradient:
  $\forall x \in \mathbb{R}^n, \forall i \in \{1, \ldots, n\}, \forall t \in \mathbb{R},$

$$|\nabla_i f(x + te_i) - \nabla_i f(x)| \leq \beta_i |t|$$

- g and h need not be separable
- $M \in \mathbb{R}^{m \times n}$

Iterates need not be feasible

Example:

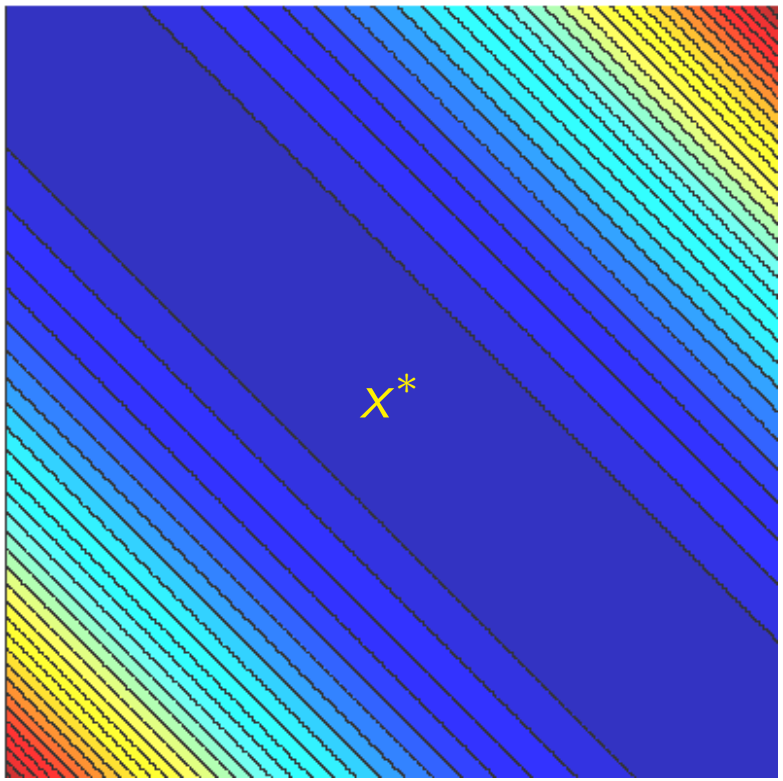
- $f(x) = \frac{1}{2}\|x\|^2$

  $g(x) = \begin{cases} 0 & \text{if } x^1 = x^2 \\ +\infty & \text{if } x^1 \neq x^2 \end{cases}$

  $h = 0$
- Start with $x_0 = [1, 1]$ (feasible):
- $\bar{x}_1 = [0, 0]$
- Let $i_1 = 1$, we get $x_1 = [0, 1]$ (unfeasible)

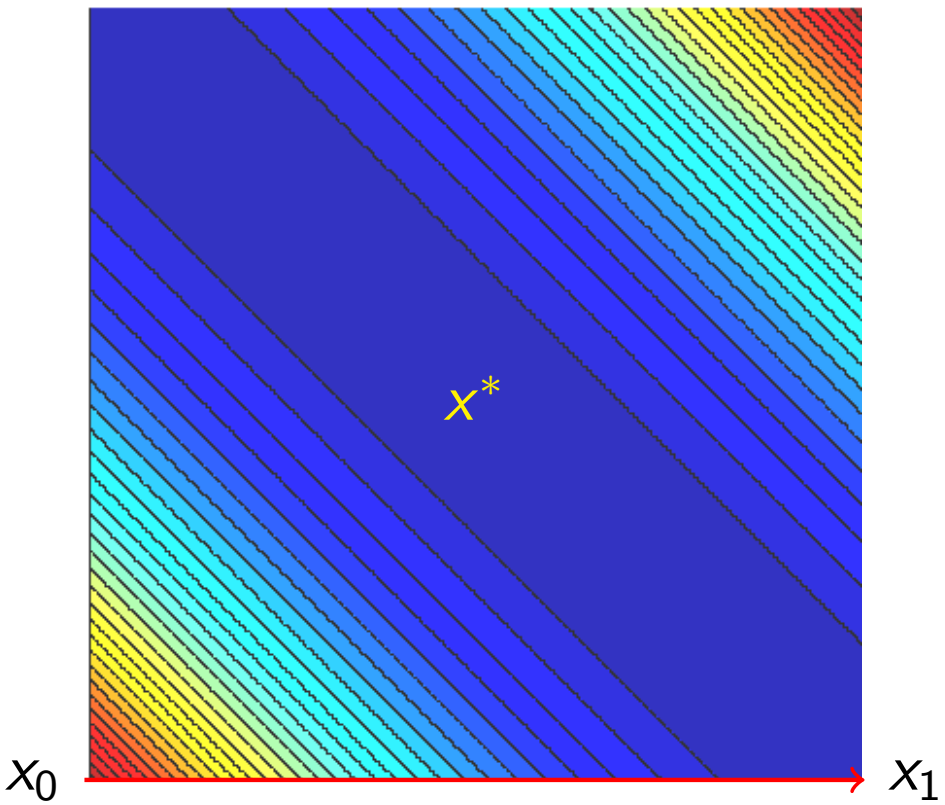
Distance to optimum may not change

$$f(x) = \frac{1}{2}(x_1 + x_2 - 1)^2, \qquad L(\nabla f) = 2, \;\; \beta_i = 1$$



Distance to optimum may not change

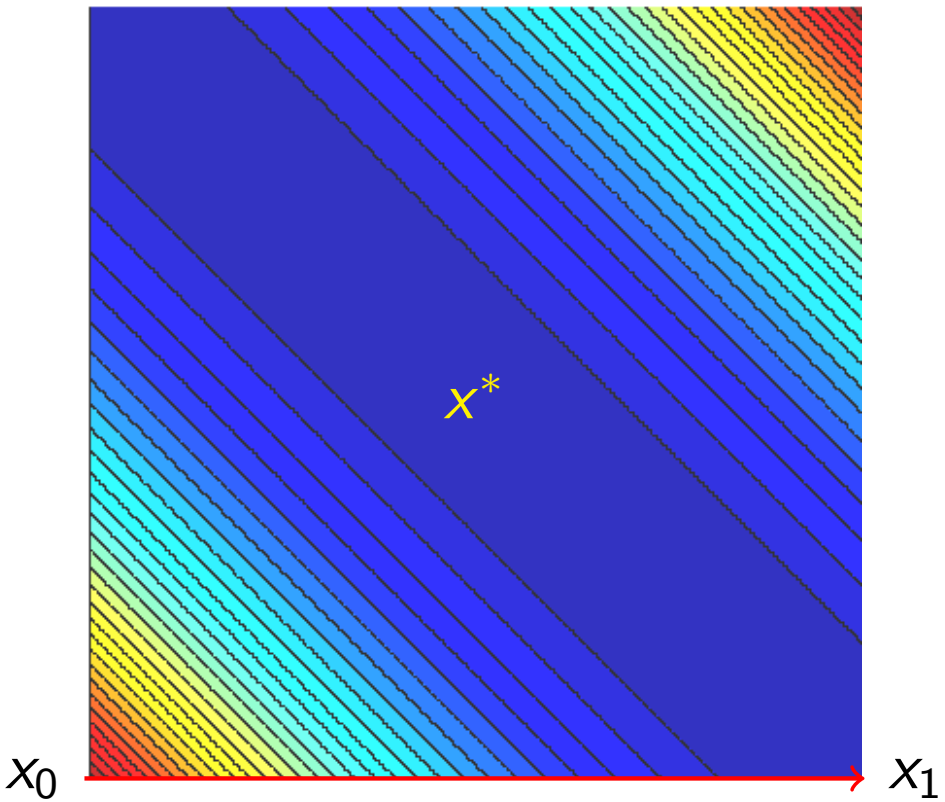
$$f(x) = \frac{1}{2}(x_1 + x_2 - 1)^2, \qquad L(\nabla f) = 2, \;\; \beta_i = 1$$



$$\mathbf{E}[\|x_1 - x_*\|^2] = \tfrac{1}{2} = \|x_0 - x_*\|^2$$

Distance to optimum may increase

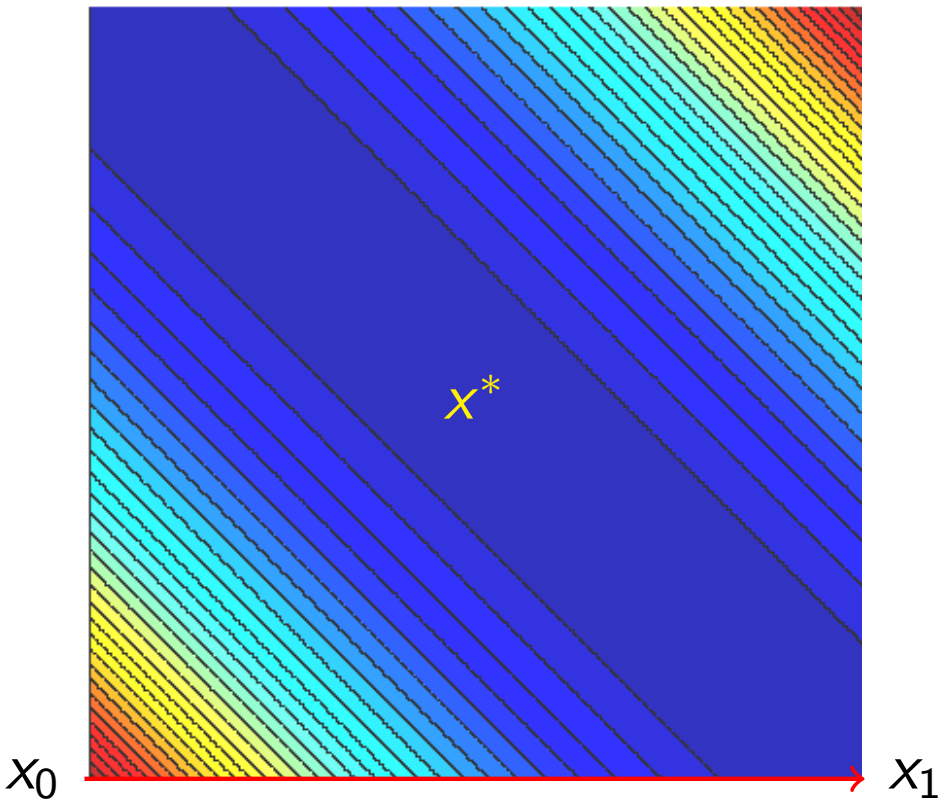
$$f(x) = \frac{1}{2}(x_1 + x_2 + x_3 - 1)^2, \qquad L(\nabla f) = 3, \;\; \beta_i = 1$$



$$\mathbf{E}[\|x_1 - x_*\|^2] = \tfrac{2}{3} > \|x_0 - x_*\|^2$$

Distance to optimum may increase a lot

$$f(x) = \frac{1}{2}\left(\sum_{i=1}^{n} x_i - 1\right)^2, \qquad L(\nabla f) = n, \;\; \beta_i = 1$$



$$\mathbf{E}[\|x_1 - x_*\|^2] = \tfrac{n}{n-1} \gg \|x_0 - x_*\|^2 = \tfrac{1}{n}$$

New algorithm

[Suppose that M is block diagonal (duplication trick available)]

At iteration k:

1. $\overline{y}_{k+1} = \text{prox}_{\sigma,h^*}(y_k + D(\sigma)Mx_k)$

   $\overline{x}_{k+1} = \text{prox}_{\tau,g}(x_k - D(\tau)(\nabla f(x_k) + M^*(2\overline{y}_{k+1} - y_k)))$

2. For $i = i_{k+1}$ and $\forall j \ : \ M_{j,i_{k+1}} \neq 0$, update:

$$x_{k+1}^{(i)} = \overline{x}_{k+1}^{(i)}$$
$$y_{k+1}^{(j)} = \overline{y}_{k+1}^{(j)}$$

3. Otherwise, set $x_{k+1}^{(i)} = x_k^{(i)}$, $y_{k+1}^{(j)} = y_k^{(j)}$

Convergence

Theorem

If for all $i \in \{1, \dots, n\}$, $\mathbf{P}(i_{k+1} = i) = 1/n$ *and*

$$\tau_i < \frac{1}{\beta_i + \rho\left(\sum_{j \in J(i)} \sigma_j M_{j,i}^* M_{j,i}\right)}$$

then there exists a saddle point (x^*, y^*) *of* $\mathcal{L}$ *such that*

$$\lim_{k \to \infty} (x_k, y_k) = (x_*, y_*)$$

Proof: We consider the Lyapunov function

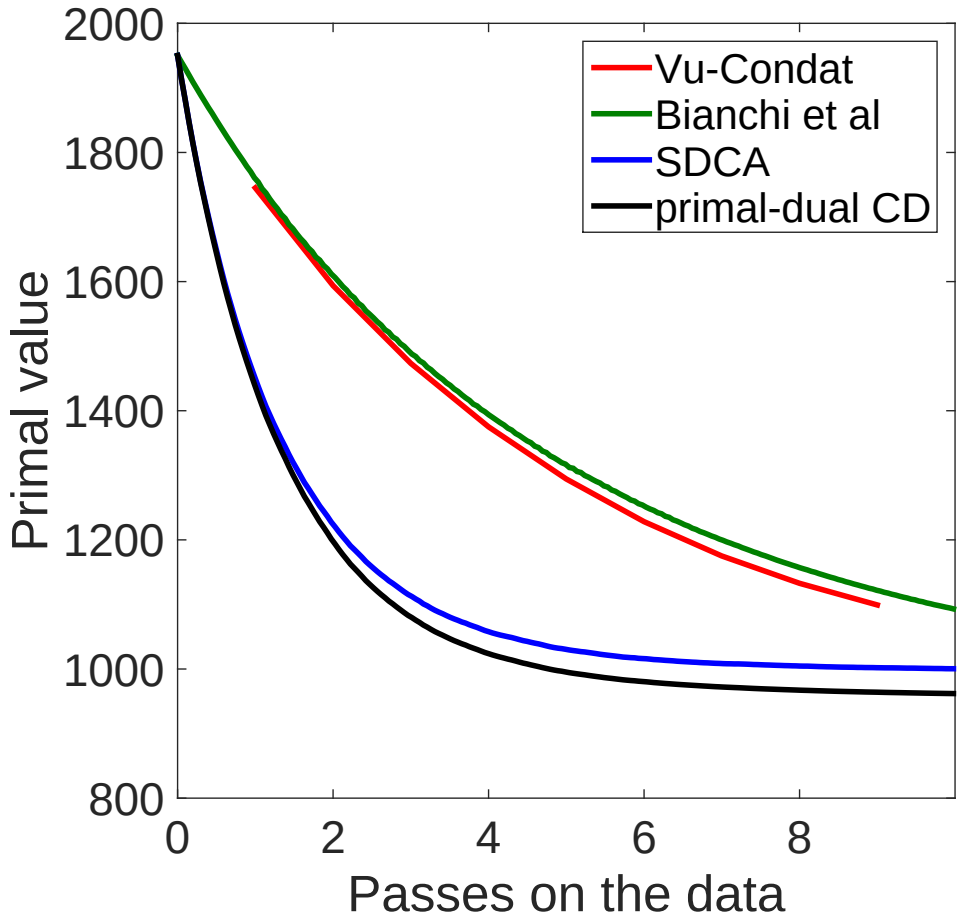
$$S_k = f(x_k) - f(x_*) - \langle \nabla f(x_*), x_k - x_* \rangle + \frac{1}{2} \|z_k - z_*\|_P$$

Dual SVM

$$\max_{x \in \mathbb{R}^n} -\frac{1}{2}\|AD(b)x\|_2^2 + e^T x - I_{[0,C]^n}(x) - I_{b^\perp}(x)$$

RCV1 dataset: $A = 47{,}236 \times 20{,}242$ matrix, $n = 47{,}236$

$C = 0.1$

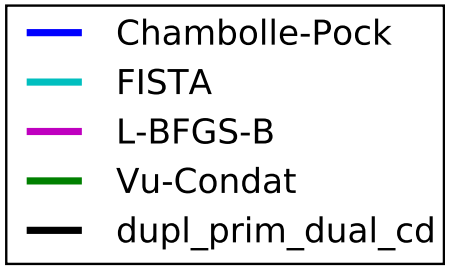


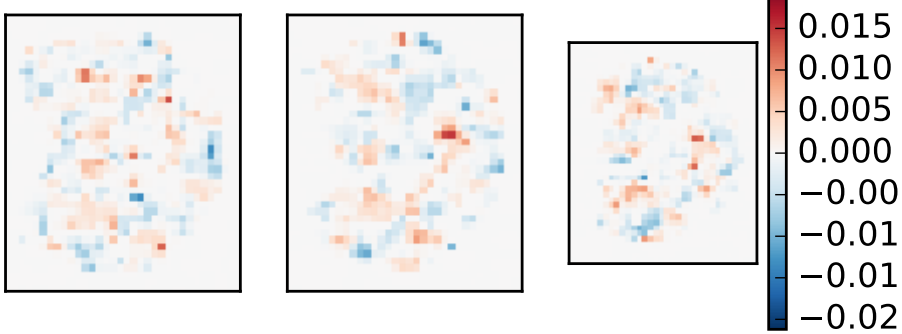
L_1 + TV regularized least squares

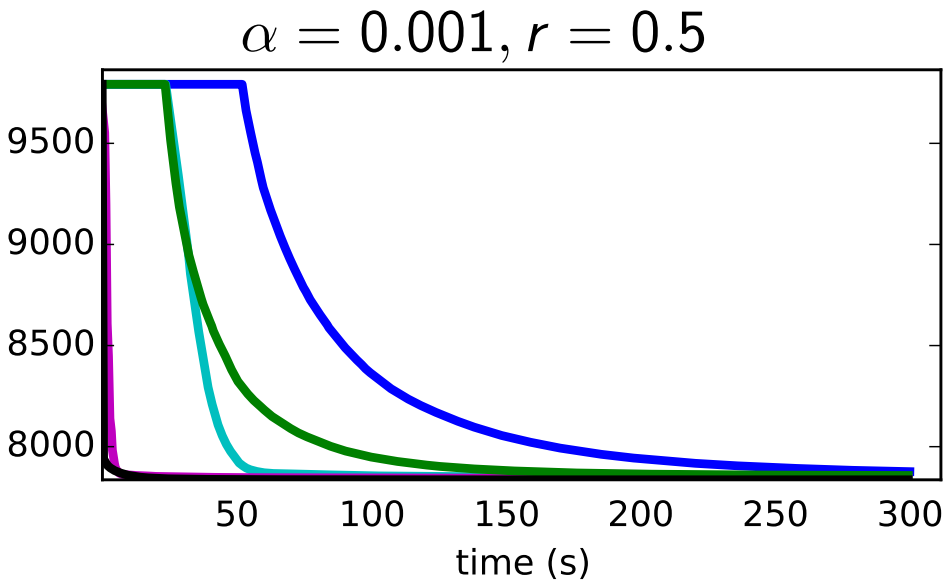
$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \alpha(r\|x\|_1 + (1-r)\|\nabla x\|_{2,1})$

A: 768 $\times$ 65,280 dense matrix

∇: 195,840 $\times$ 65,280 sparse matrix (3D gradient)







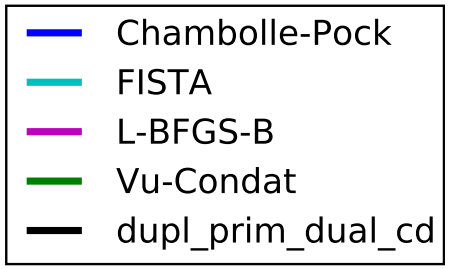
Lightly regularized problem

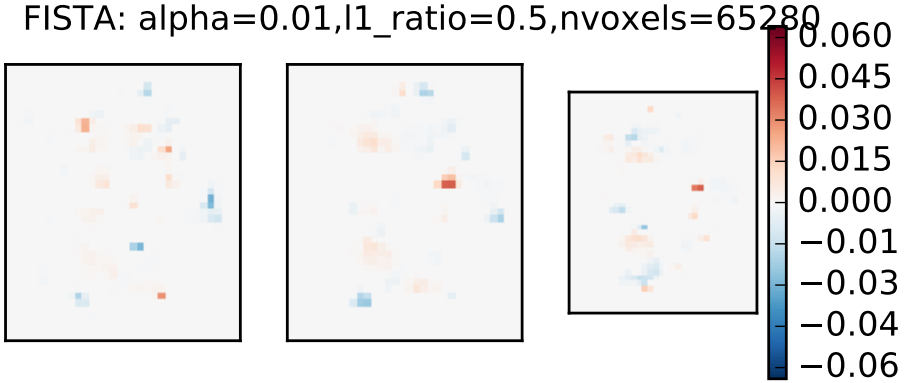
L_1 + TV regularized least squares

$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \alpha(r\|x\|_1 + (1-r)\|\nabla x\|_{2,1})$

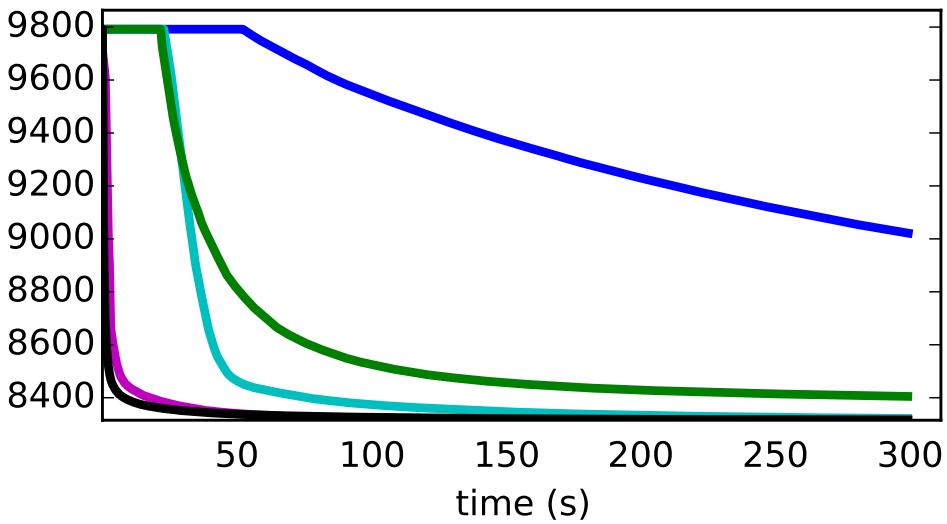
A: 768 $\times$ 65,280 dense matrix

∇: 195,840 $\times$ 65,280 sparse matrix (3D gradient)





$\alpha = 0.01, r = 0.5$



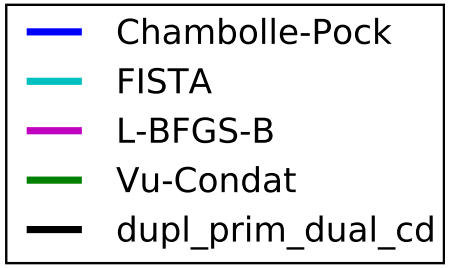
Mediumly regularized problem

L_1 + TV regularized least squares

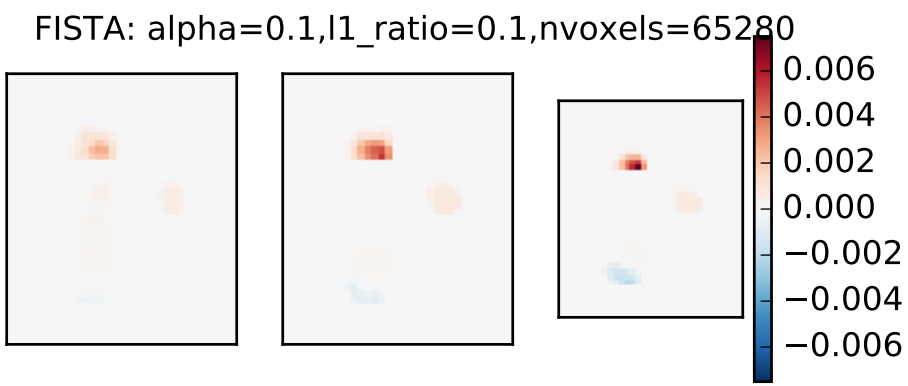
$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \alpha(r\|x\|_1 + (1-r)\|\nabla x\|_{2,1})$

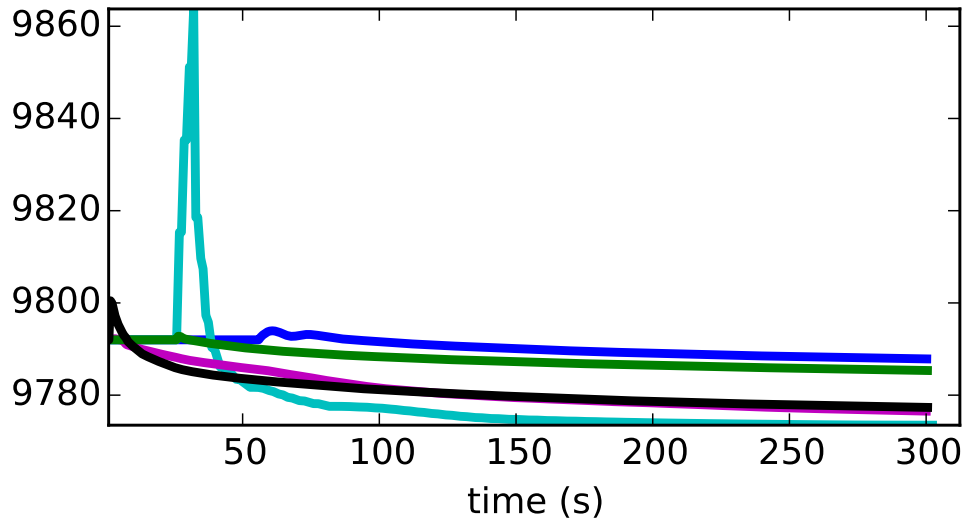
A: 768 $\times$ 65,280 dense matrix

∇: 195,840 $\times$ 65,280 sparse matrix (3D gradient)



$\alpha = 0.1, r = 0.1$





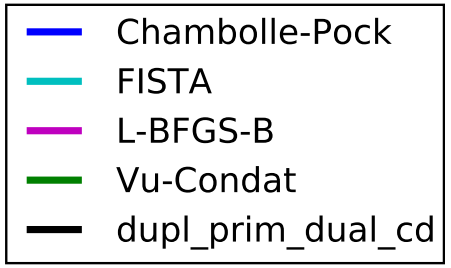
Strongly TV-regularized problem

L_1 + TV regularized least squares

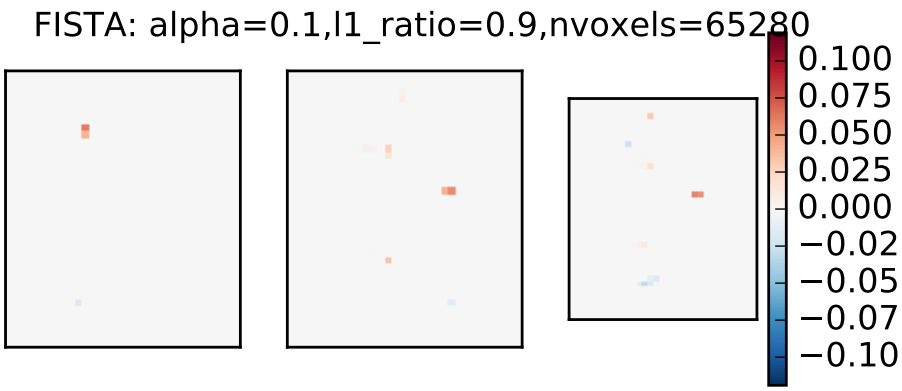
$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \alpha(r\|x\|_1 + (1-r)\|\nabla x\|_{2,1})$

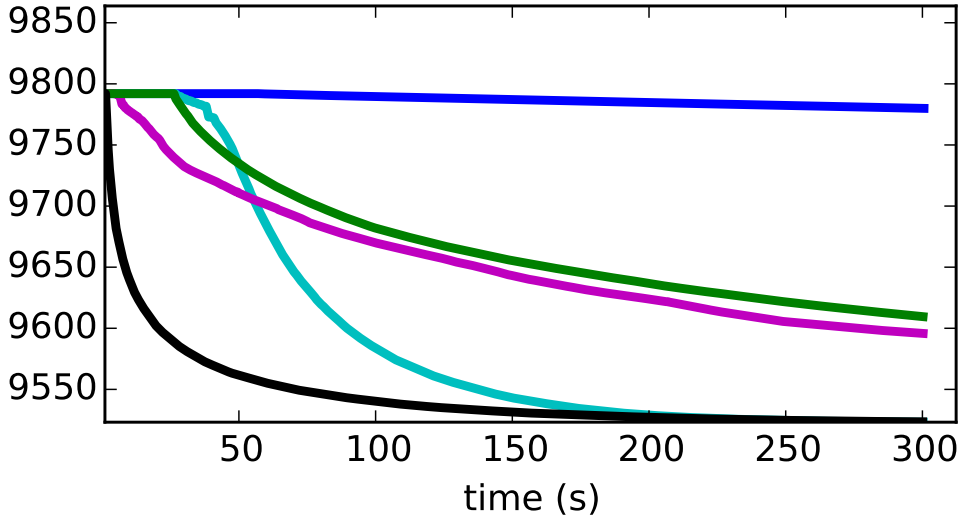
A: 768 $\times$ 65,280 dense matrix

∇: 195,840 $\times$ 65,280 sparse matrix (3D gradient)



$\alpha = 0.1, r = 0.9$





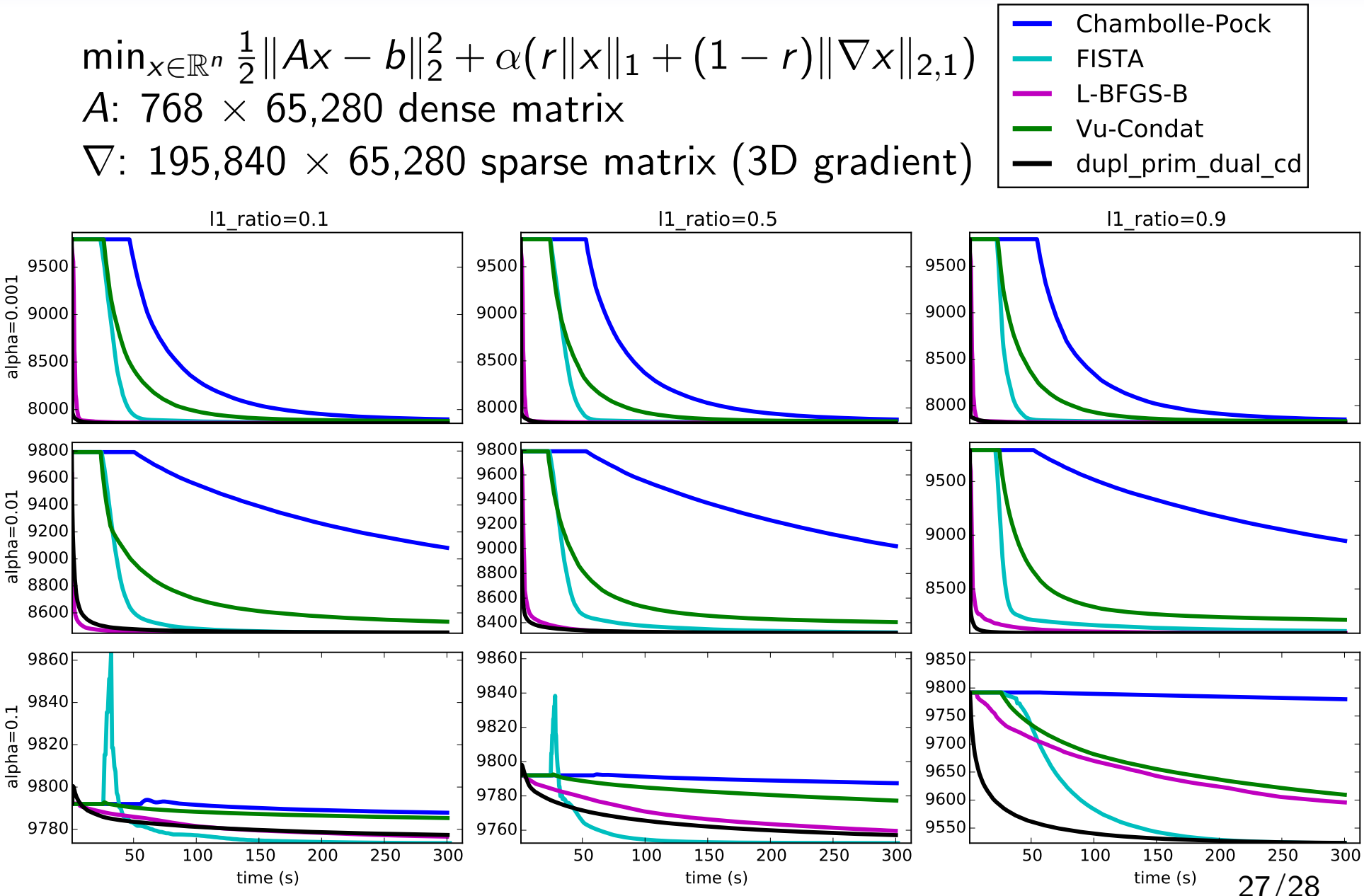
Strongly L_1-regularized problem

L_1 + TV regularized least squares

$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \alpha(r\|x\|_1 + (1-r)\|\nabla x\|_{2,1})$

A: 768 $\times$ 65,280 dense matrix

∇: 195,840 $\times$ 65,280 sparse matrix (3D gradient)



Conclusion

Summary

- Genuine coordinate descent method with non-separable and non-smooth convex function
- Promising numerical results

Open questions

- Long steps
  - Non uniform probabilities
  - Speed of convergence
  - Replace β_i by $\beta_i/2$
- Non-ergodic speed of convergence
- Longer steps for non-separable proximal operators: eg. MISO, projection on $\{x \ : \ x^1 = \ldots = x^n\}$