Institut
Mines-Télécom

Primal-dual coordinate descent

A Coordinate Descent Primal-Dual Algorithm with Large Step Size and Possibly Non-Separable Functions

Olivier Fercoq and Pascal Bianchi

Problem

Minimize the convex function

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

- f, g, h convex
- f is differentiable
- $\text{prox}_{\tau,g}$ and prox_{σ,h^*} are easy to compute

$$\text{prox}_{\tau,g}(y) = \arg\min_{x \in \mathbb{R}^n} g(x) + \underbrace{\frac{1}{2}\sum_{i=1}^n \frac{1}{\tau_i}(x^{(i)} - y^{(i)})^2}_{\frac{1}{2}\|x-y\|^2_{\tau^{-1}}}$$

- $M : \mathbb{R}^n \to \mathbb{R}^m$ is linear

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

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

Examples: Lasso

A is a dictionary
b is a vector we want to explain
Solution x: a sparse representation of b

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1$$

- $f(x) = \frac{1}{2}\|Ax - b\|_2^2$
- $g(x) = \lambda \|x\|_1$
- $h(x) = 0$

Examples: L_1 + TV regularized regression

Each line of A: a 3D image of brains
b_j: the result of experiment j
x: which regions explain best the result of the experiments?

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

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

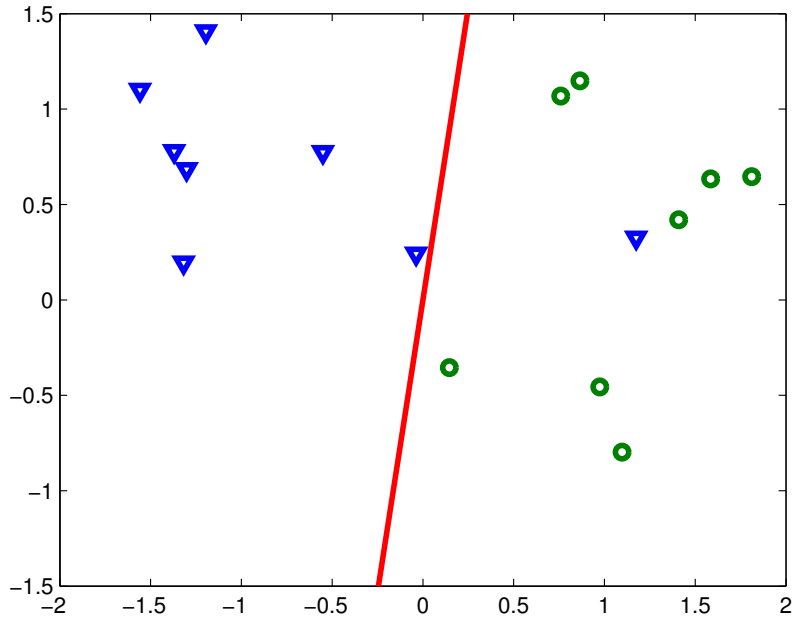
Examples: dual of Support Vector Machines

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

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

$$b^T x = 0$$

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



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:
  $\forall x \in \mathbb{R}^n, \forall i \in \{1, \dots, n\}, \forall t \in \mathbb{R},$

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

Coordinate descent

At iteration k:

1. Choose randomly a coordinate i_{k+1}
2. $\bar{x}_{k+1} = \text{prox}_{\tau,g}\left(x_k - \tau \nabla f(x_k)\right)$
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 $\tau_i < \dfrac{2}{\beta_i}$ 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]

$$y_{k+1} = \text{prox}_{\sigma h^*}\left(y_k + \sigma M x_k\right)$$
$$x_{k+1} = \text{prox}_{\tau g}\left(x_k - \tau M^*(2y_{k+1} - y_k) - \tau \nabla f(x_k)\right)$$

- Generalizes Chambolle-Pock, ADMM and proximal gradient
- Converges to a saddle point of the Lagrangian
- Convergence as soon as $\tau < \dfrac{1}{\frac{L(\nabla f)}{2} + \sigma \|M\|^2}$

  (fixed point of a firmly nonexpansive operator)

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

[Bianchi, Hachem & Iutzeler, 2014]

- We take Vũ-Condat's fixed point operator

$$T\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \overbrace{\mathrm{prox}_{\sigma h^*}\left(y + \sigma M x\right)}^{\bar{y}} \\ \mathrm{prox}_{\tau g}\left(x - \tau M^*(2\bar{y} - y) - \tau \nabla f(x)\right) \end{pmatrix}$$

- 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 $\tau < \dfrac{1}{\frac{L(\nabla f)}{2} + \sigma \|M\|^2}$

Duplication trick

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

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

We have $D(m)SK = M$. We define $\overline{h} = h \circ (D(m)S)$:

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

$$\text{prox}_{m\sigma,\overline{h}^*}(\boldsymbol{y}) = (\mathbf{1}_{m_1} \otimes \text{prox}^{(1)}_{\sigma,h^*}(S(\boldsymbol{y})), \ldots, \mathbf{1}_{m_p} \otimes \text{prox}^{(p)}_{\sigma,h^*}(S(\boldsymbol{y})))$$

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:
  $\forall x \in \mathbb{R}^n, \forall i \in \{1, \dots, n\}, \forall t \in \mathbb{R},$

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

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

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:

   $$x^i_{k+1} = \overline{x}^i_{k+1}$$
   $$y^j_{k+1} = \overline{y}^j_{k+1}$$

3. Otherwise, set $x^i_{k+1} = x^i_k$, $y^j_{k+1} = y^j_k$

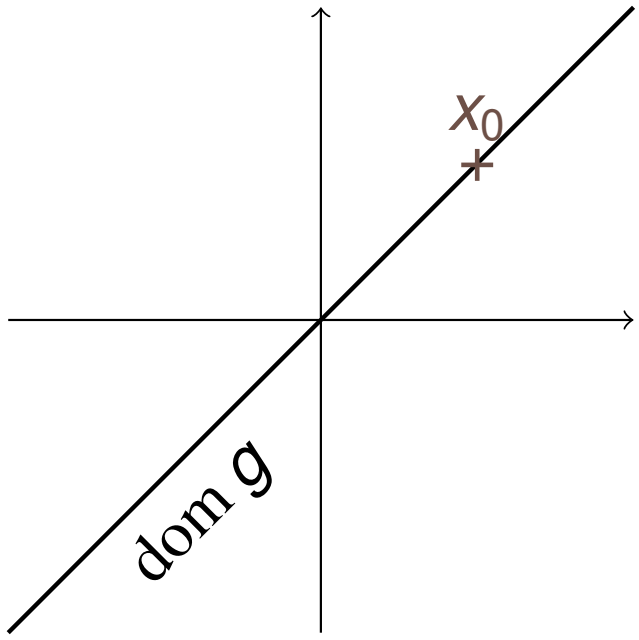
Iterates need not be feasible

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):



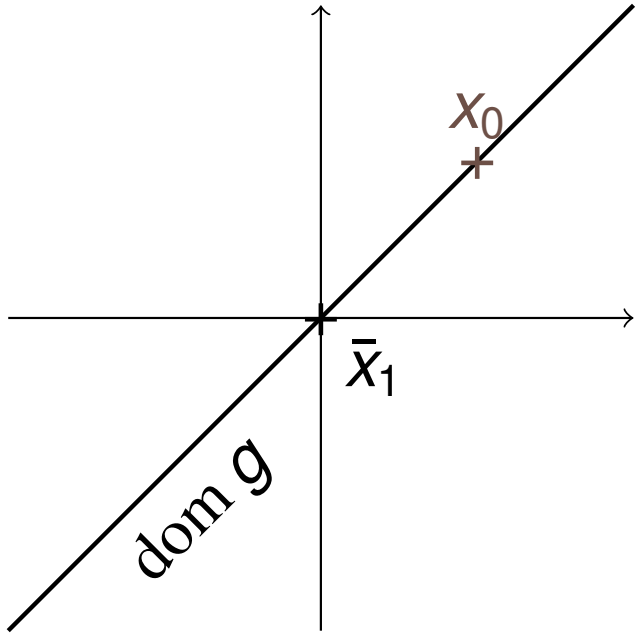
Iterates need not be feasible

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 $i_1 = 1$



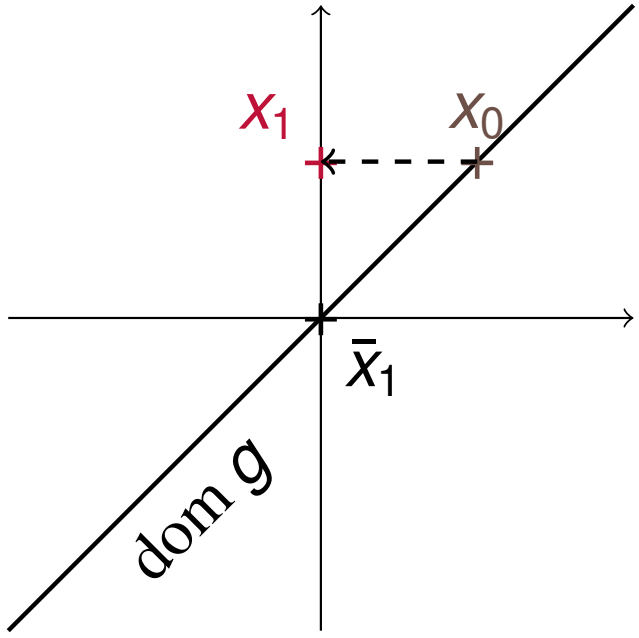
Iterates need not be feasible

Example:

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

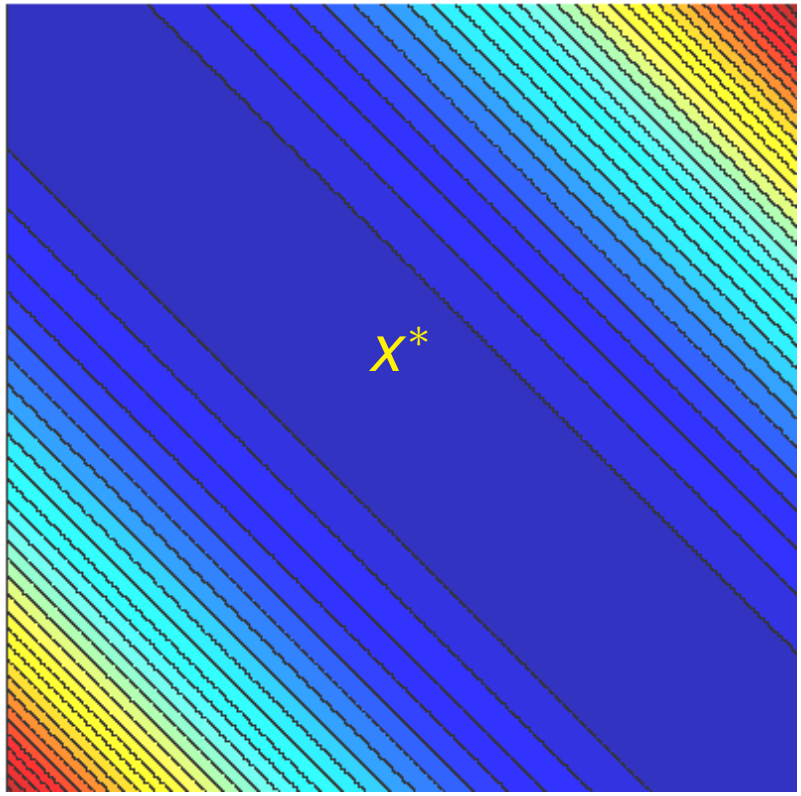
  $g(x) = \begin{cases} 0 & \text{if } x^1 = x^2 \\ +\infty & \text{if } x^1 \neq x^2 \end{cases}$

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



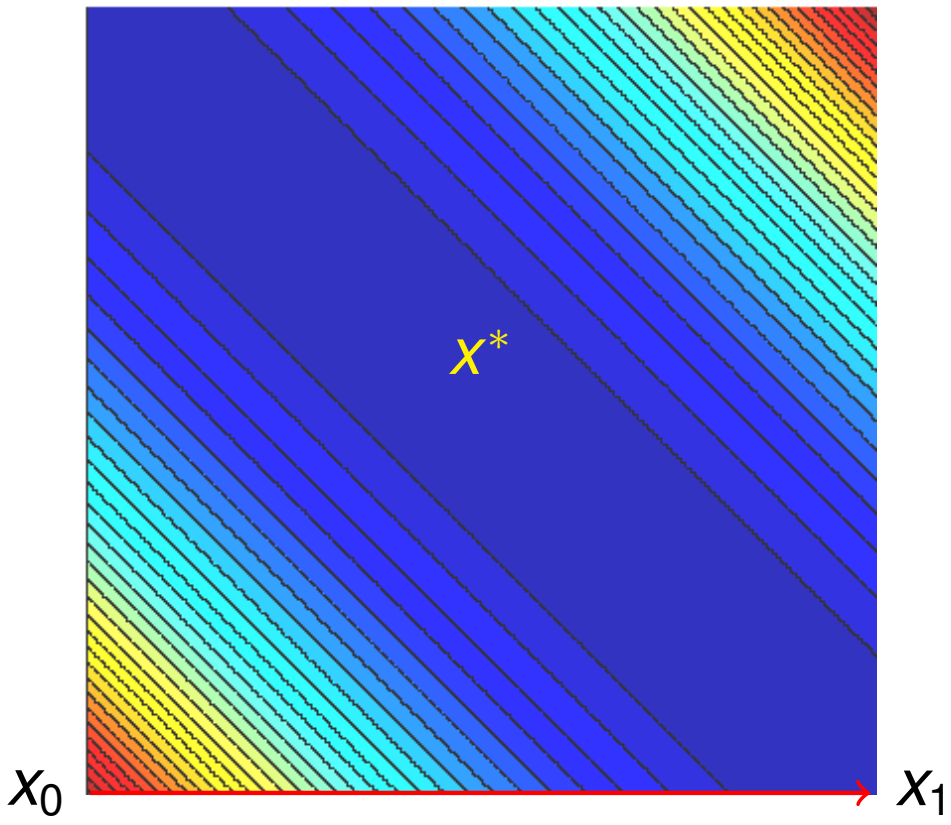
Distance to optimum may not change

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



Distance to optimum may not change

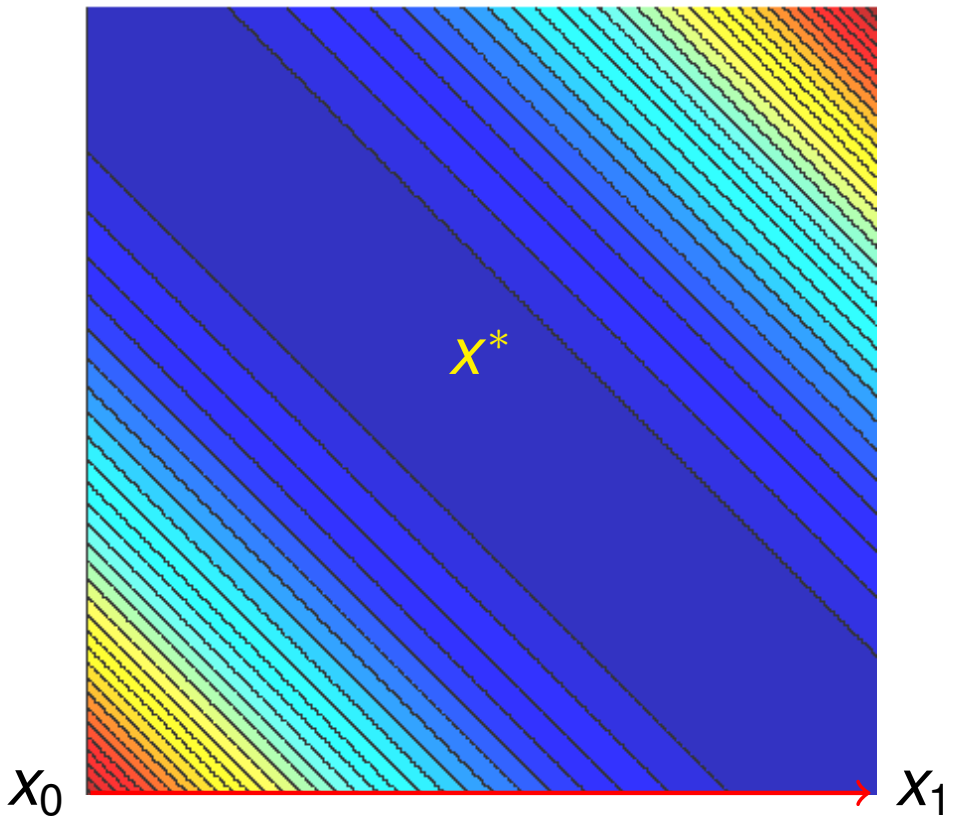
$$f(x) = \frac{1}{2}(x_1 + x_2 - 1)^2, \qquad L(\nabla f) = 2, \ \beta_i = 1$$



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

Distance to optimum may increase

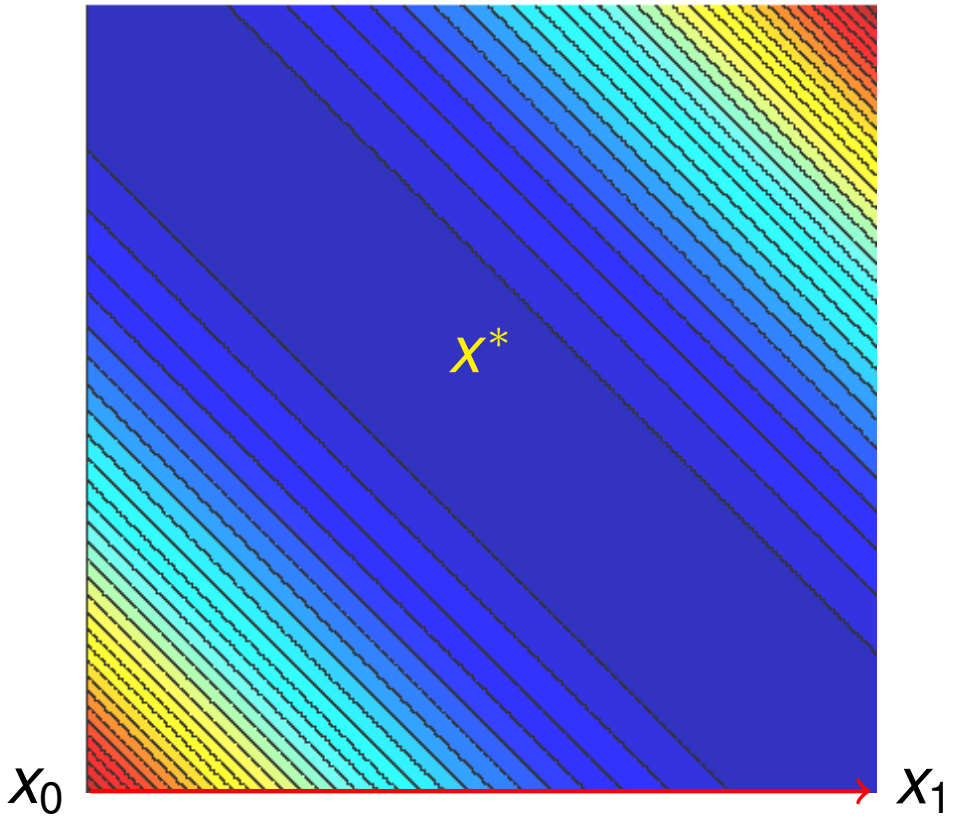
$$f(x) = \frac{1}{2}(x_1 + x_2 + x_3 - 1)^2, \qquad L(\nabla f) = 3, \ \beta_i = 1$$



$$\mathbf{E}[\|x_1 - x_*\|^2] = \frac{2}{3} > \|x_0 - x_*\|^2 = \frac{1}{3}$$

Distance to optimum may increase a lot

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



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

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 the Lagrangian* $\mathcal{L}(x, y) = f(x) + g(x) + \langle Ax, y \rangle - h^*(y)$ *such that*

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

Proof: A stochastic Lyapunov function is

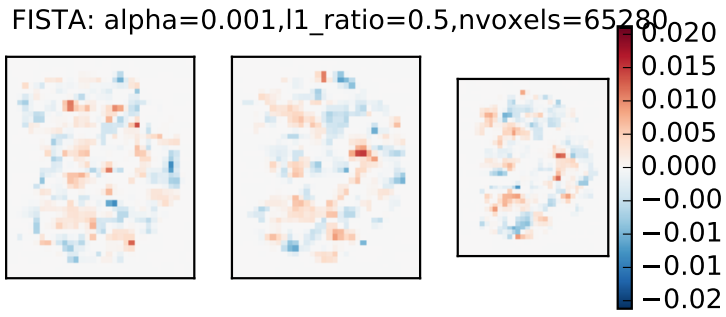
$S_k = f(x_k) - f(x_*) - \langle \nabla f(x_*), x_k - x_* \rangle + \frac{1}{2} \| z_k - z_* \|_P^2$

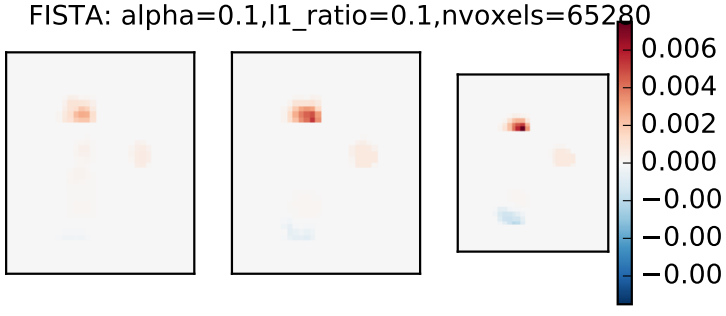
L_1 + TV regularized least squares

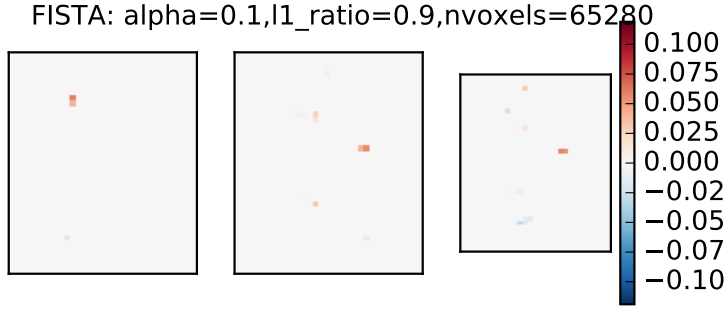
$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \alpha(r\|x\|_1 + (1-r)\|\nabla x\|_{2,1})$

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

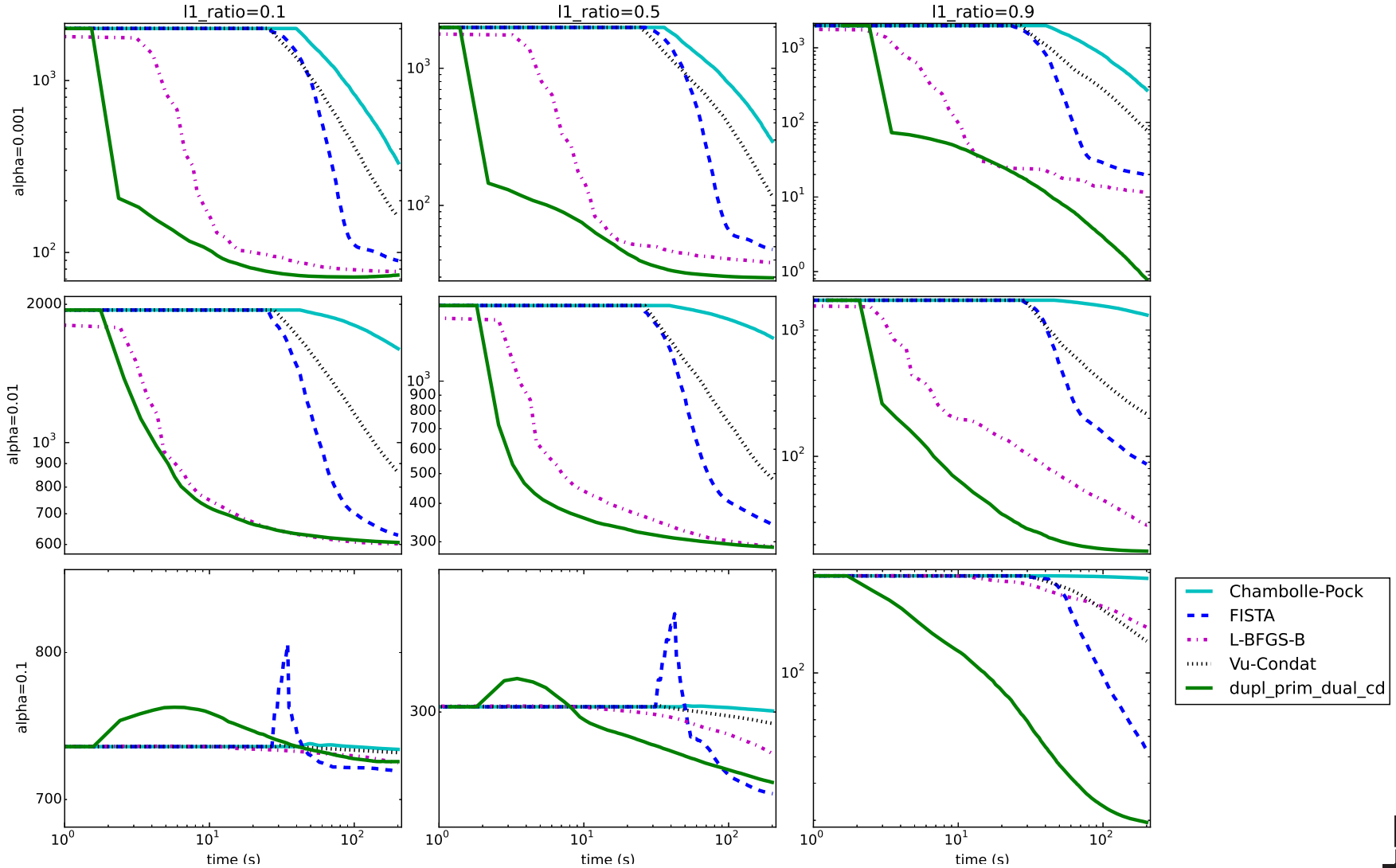
∇: 195,840 $\times$ 65,280 sparse matrix (3D gradient)







Comparison of algorithms on L_1 + TV regularized least squares

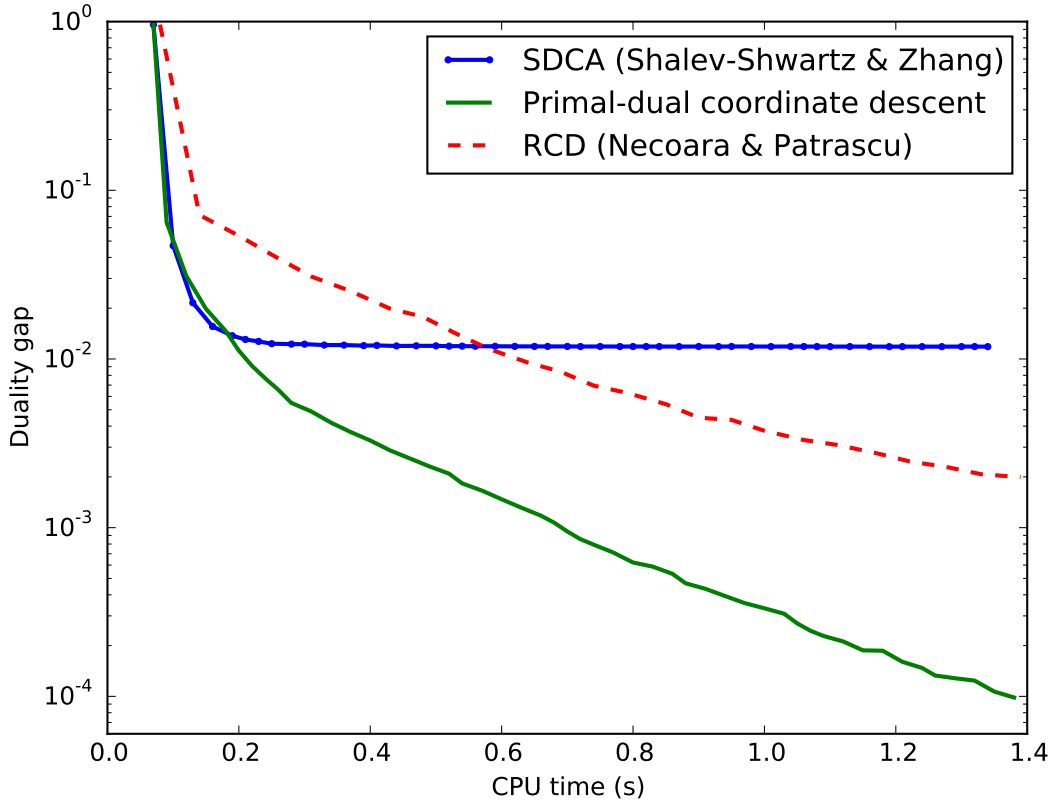


Dual SVM

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

RCV1 dataset: $A = 47{,}236 \times 20{,}242$ matrix, nnz = 0.157%

$C = \frac{1}{n}$, $\lambda = \frac{1}{n}$, 100 passes through the data

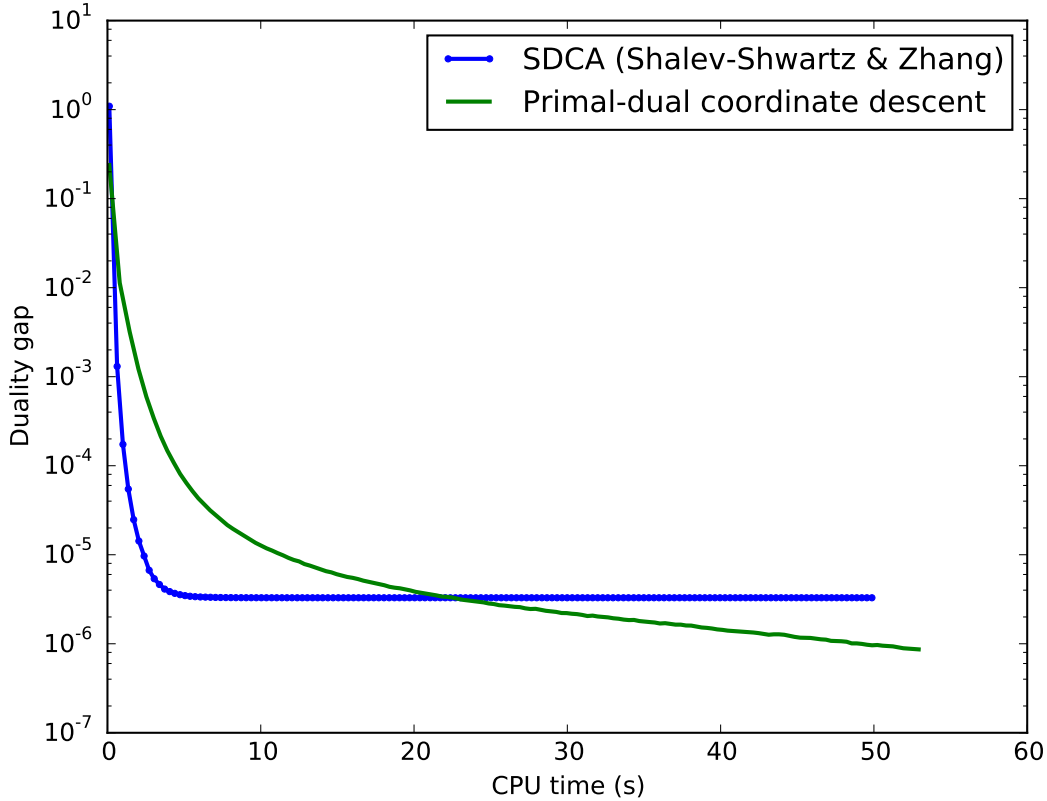


Larger SVM problem

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

KDD 2009: $A = 86{,}825 \times 50{,}000$ matrix, nnz = 1.79%

$\lambda = \frac{1}{n}$, $C_i \in \{C^+, C^-\}$, 300 passes



Conclusion

Summary

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

Open questions

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