# Robust Block Coordinate Descent (RCD)

Kimon Fountoulakis and Rachael Tappenden

School of Mathematics

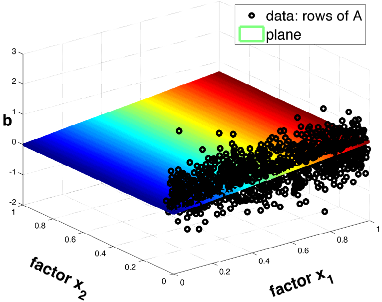
University of Edinburgh

Edinburgh SIAM Student Chapter Conference

# Standard examples in optimization

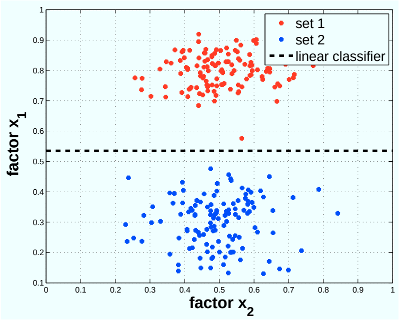
## Data fitting

$$\text{minimize } \gamma\|x\|_1 + \frac{1}{2}\|Ax - b\|_2^2$$




## Binary classification

$$\text{minimize } \gamma\|x\|_1 + \sum_{i=1}^{m} \log(1 + e^{-b_i x^\intercal a_i})$$



# Problem formulation

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

- $x \in \mathbb{R}^N$, $f(x) : \mathbb{R}^N \to \mathbb{R}$, $\Psi(x) : \mathbb{R}^N \to \mathbb{R}$

**Assumptions**

- $f$ is smooth (possibly) convex function
- $\Psi$ is a (possibly) nonsmooth convex function

**Plenty of data**

- $N$ is very large. i.e. of order millions or billions

# Numerical methods in convex optimization

Build a convex function $Q$ that locally approximates $F$ at a point $x$:

- $Q(y; x) \approx F(y)$ for $y$ close to $x$
- $Q(x; x) = F(x)$

**General framework**

1: Given $x_0$ (an initial guess)
2: For $k = 0, 1, 2, \ldots$
3: Approximately solve the subproblem

$$y^* \approx \arg\min_y \; Q(y; x_k)$$

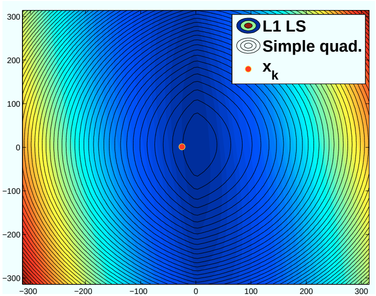
4: Set $x_{k+1} := y^*$

# Examples of local convex approximations

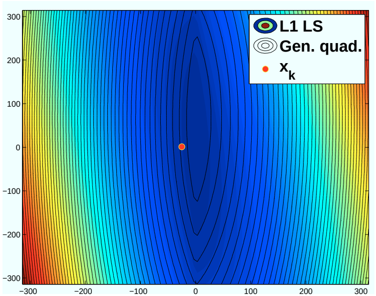
*L1 Least Squares (L1 LS):* minimize $\underbrace{\gamma\|x\|_1}_{\Psi} + \underbrace{1/2\|Ax - b\|_2^2}_{f}$

- **Simple quadratic** (majority of modern algorithms)
  $Q(y; x_k) := \gamma\|y\|_1 + f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{L}{2}\langle y - x_k, y - x_k\rangle$

- **General quad.**
  $Q(y; x_k) := \gamma\|y\|_1 + f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2}\langle y - x_k, H_k(y - x_k)\rangle$




$L = \lambda_{max}(A^\intercal A)$




$H_k = diag(A^\intercal A)$

# Trade-off between simple and general quadratic approximations

| **Type** | **Inexpensive step** | **Good approximation** |
|---|---|---|
| **Simple quad.:** | ✔ | |
| **Gen. quad.:** | | ✔ |

**Aim: control the trade off**


**Three ways in Robust Block Coordinate Descent (RCD)**

- Inexpensively choose $H_k$ such that it approximates the structure of $f$
- Dimensionality reduction: update only a block of coordinates
- Solve approximately the subproblem over the chosen block of coordinates

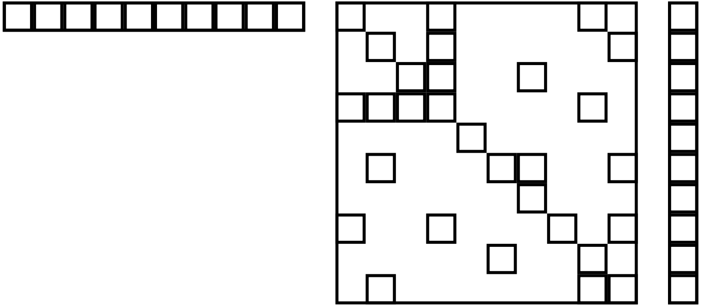
# Trade offs in RCD: construction of quadratic

1) Construct any $H_k \succ 0$ such that $H_k \approx \nabla^2 f(x_k)$. No need to store $H_k$, we only need a process to perform matrix-vector products with it.

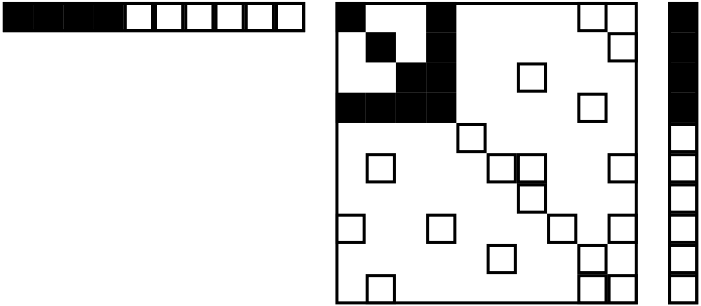
2) Construct a local convex model

$$Q(y; x_k) := \gamma\Psi(y) + f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2}\langle y - x_k, H_k(y - x_k)\rangle$$

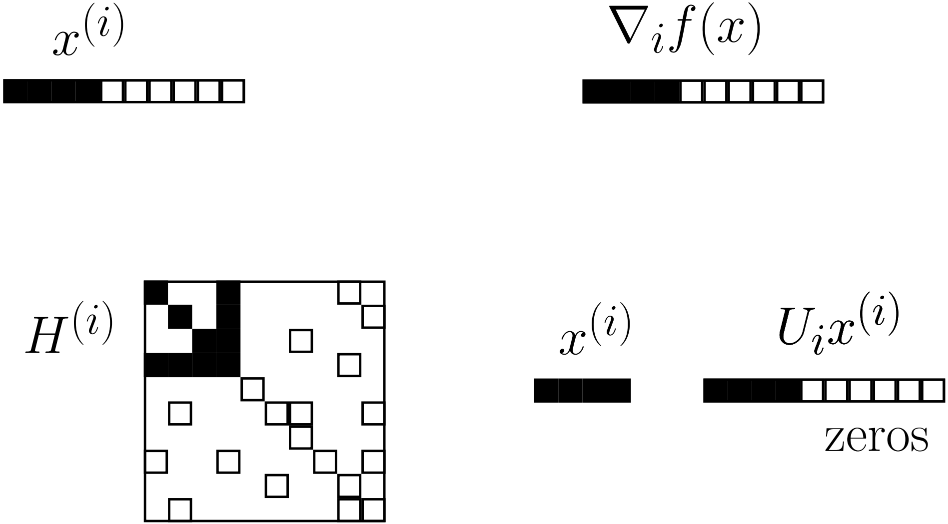
# Dimensionality reduction



# Dimensionality reduction



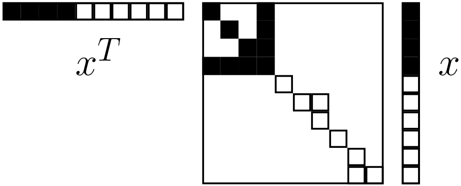
# Dimensionality reduction: block notation



# Dimensionality reduction: a smaller subproblem

**Assumption: $\Psi$ is block separable**

$\Psi(x) =$




$$\color{red}{\Psi_1(x^{(1)})} + \Psi_2(x^{(2)}) + \Psi_3(x^{(3)}) + \Psi_4(x^{(4)}) + \Psi_5(x^{(5)})$$

**Reformulation of local approximation and subproblem**

$$Q_i(x_k^{(i)} + t^{(i)}; x_k) := \gamma \Psi_i(x_k^{(i)} + t^{(i)}) + f(x_k) + \langle \nabla_i f(x_k), t^{(i)} \rangle + \frac{1}{2} \langle t^{(i)}, H_k^{(i)} t^{(i)} \rangle$$

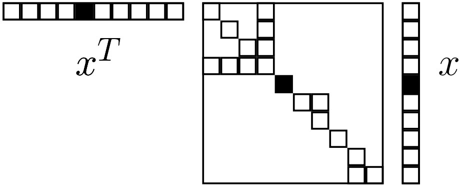
$$t_k^{(i)} \approx \arg\min_{t^{(i)}} Q_i(x_k^{(i)} + t^{(i)}; x_k)$$

The subproblem has smaller dimensions than $N$!

# Dimensionality reduction: a smaller subproblem

**Assumption: $\Psi$ is block separable**

$\Psi(x) =$




$$\Psi_1(x^{(1)}) + \Psi_2(x^{(2)}) + \Psi_3(x^{(3)}) + \Psi_4(x^{(4)}) + \Psi_5(x^{(5)})$$

**Reformulation of local approximation and subproblem**

$$Q_i(x_k^{(i)} + t^{(i)}; x_k) := \gamma \Psi_i(x_k^{(i)} + t^{(i)}) + f(x_k) + \langle \nabla_i f(x_k), t^{(i)} \rangle + \frac{1}{2} \langle t^{(i)}, H_k^{(i)} t^{(i)} \rangle$$

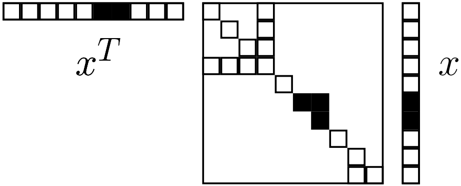
$$t_k^{(i)} \approx \arg\min_{t^{(i)}} Q_i(x_k^{(i)} + t^{(i)}; x_k)$$

The subproblem has smaller dimensions than $N$!

# Dimensionality reduction: a smaller subproblem

**Assumption: $\Psi$ is block separable**

$$\Psi(x) =$$




$$\Psi_1(x^{(1)}) + \Psi_2(x^{(2)}) + \Psi_3(x^{(3)}) + \Psi_4(x^{(4)}) + \Psi_5(x^{(5)})$$

**Reformulation of local approximation and subproblem**

$$Q_i(x_k^{(i)} + t^{(i)}; x_k) := \gamma \Psi_i(x_k^{(i)} + t^{(i)}) + f(x_k) + \langle \nabla_i f(x_k), t^{(i)} \rangle + \frac{1}{2} \langle t^{(i)}, H_k^{(i)} t^{(i)} \rangle$$

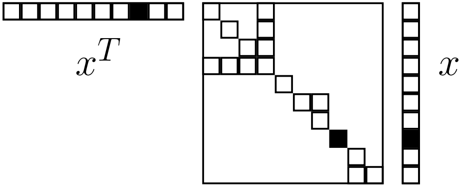
$$t_k^{(i)} \approx \arg\min_{t^{(i)}} Q_i(x_k^{(i)} + t^{(i)}; x_k)$$

The subproblem has smaller dimensions than $N$!

# Dimensionality reduction: a smaller subproblem

**Assumption: $\Psi$ is block separable**

$$\Psi(x) =$$




$$\Psi_1(x^{(1)}) + \Psi_2(x^{(2)}) + \Psi_3(x^{(3)}) + \Psi_4(x^{(4)}) + \Psi_5(x^{(5)})$$

**Reformulation of local approximation and subproblem**

$$Q_i(x_k^{(i)} + t^{(i)}; x_k) := \gamma \Psi_i(x_k^{(i)} + t^{(i)}) + f(x_k) + \langle \nabla_i f(x_k), t^{(i)} \rangle + \frac{1}{2} \langle t^{(i)}, H_k^{(i)} t^{(i)} \rangle$$

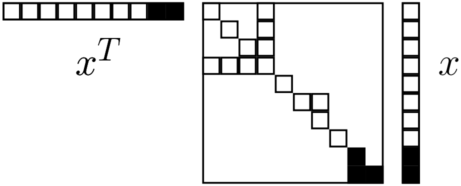
$$t_k^{(i)} \approx \arg\min_{t^{(i)}} Q_i(x_k^{(i)} + t^{(i)}; x_k)$$

The subproblem has smaller dimensions than $N$!

# Dimensionality reduction: a smaller subproblem

**Assumption: $\Psi$ is block separable**

$\Psi(x) =$




$$\Psi_1(x^{(1)}) + \Psi_2(x^{(2)}) + \Psi_3(x^{(3)}) + \Psi_4(x^{(4)}) + \Psi_5(x^{(5)})$$

**Reformulation of local approximation and subproblem**

$$Q_i(x_k^{(i)} + t^{(i)}; x_k) := \gamma \Psi_i(x_k^{(i)} + t^{(i)}) + f(x_k) + \langle \nabla_i f(x_k), t^{(i)} \rangle + \frac{1}{2} \langle t^{(i)}, H_k^{(i)} t^{(i)} \rangle$$

$$t_k^{(i)} \approx \arg\min_{t^{(i)}} Q_i(x_k^{(i)} + t^{(i)}; x_k)$$

The subproblem has smaller dimensions than $N$!

# Trade offs in RCD: inexact solution of subproblem

**Interpretation:** solve subproblem until

- direction $t_k^{(i)}$ reduces $Q_i$ compared to zero direction, and
- direction $t_k^{(i)}$ is closer to optimality than zero direction.

**First condition:** decrease of local model

$$Q_i(x_k^{(i)} + t_k^{(i)}; x_k) < Q_i(x_k^{(i)}; x_k)$$

**Second condition:** decrease distance from optimality of the local model

$$\|g_i(x_k^{(i)} + t_k^{(i)}; x_k)\|_2 \leq \eta_k^i \|g_i(x_k^{(i)}; x_k)\|_2, \quad \eta_k^i \in [0, 1)$$

Think $g_i(x_k^{(i)} + t^{(i)}; x_k)$ as the gradient of $Q_i$ at $t^{(i)}$ (if $\Psi$ is nonsmooth $g_i$ is not a gradient, but it is a measure for the distance from optimality).

# RCD

1: **Input:** Choose $x^0$ and $\theta \in (0, 1/2)$

2: **Loop:** For $k = 1, 2, ...,$ until termination criteria are met

3: Sample block of coordinates $i$ with probability $p_i > 0$

4: Calculate direction $t_k^{(i)}$ by approximately solving

$$t_k^{(i)} \approx \arg\min_{t^{(i)}} Q_i(x_k^{(i)} + t^{(i)}; x_k)$$

5: Backtracking line search along direction $t_k^{(i)}$ starting from $\alpha = 1$. That is, find $\alpha \in (0, 1]$ such that a sufficient decrease condition is satisfied (explained in next slide).

6: Set $x_{k+1}^{(i)} := x_k^{(i)} + \alpha t_k^{(i)}$

# Global convergence of RCD

**Theory:** RCD converges to first order stationary points! Convexity of $f$ is not necessary.

**Line search:** Find a step size $\alpha \in (0, 1]$ such that:

- the decrease of $F$ (objective function) is proportional to the decrease of its *block* first order approximation.

Let $\ell_i(x_k^{(i)} + t_k^{(i)}; x_k)$ be the block first order approximation of $F$:

$$\ell_i(x_k^{(i)} + t^{(i)}; x_k) := \Psi_i(x_k^{(i)} + t^{(i)}) + f(x_k) + \langle \nabla_i f(x_k), t^{(i)} \rangle$$

For $\theta \in (0, 1/2)$, find a step size $\alpha \in (0, 1]$ such that

$$F(x_k) - F(x_k + \alpha U_i t_k^{(i)}) \geq \theta \left( \ell_i(x_k^{(i)}; x_k) - \ell_i(x_k^{(i)} + \alpha t_k^{(i)}; x_k) \right).$$

# Local convergence of RCD: unit step sizes

**Theory:** There exists a neighbourhood of the optimal solution in which line search will accept unit step sizes for any chosen $i$.

**Assumptions**

- $f$ is strongly convex
- Block Lipschitz continuity of $\nabla^2 f(x)$

$$\|\nabla_i^2 f(x + U_i t^{(i)}) - \nabla_i^2 f(x)\|_2 \leq M_i \|t^{(i)}\|_2 \quad \forall i, x$$

**Stronger inexactness condition for the subproblem**

- $Q_i(x_k^{(i)}; x_k) - Q_i(x_k^{(i)} + t_k^{(i)}; x_k) > \xi \left( \ell_i(x_k^{(i)}; x_k) - \ell_i(x_k^{(i)} + \alpha t_k^{(i)}; x_k) \right)$

# Local rate of convergence

**Theory:** block quadratic convergence rate

- If $\eta_k^i = \min\{1/2, \|g_i(x_k^{(i)}; x_k)\|_2\}$ in $\|g_i(x_k^{(i)} + t_k^{(i)}; x_k)\|_2 \leq \eta_k^i \|g_i(x_k^{(i)}; x_k)\|_2$

Then, $\|g_i(x_k^{(i)}; x_k)\|_2$ has a quadratic rate of convergence in expectation.

**Theory:** block superlinear convergence rate

- if $\eta_k^i \to 0$ for $k \to \infty$

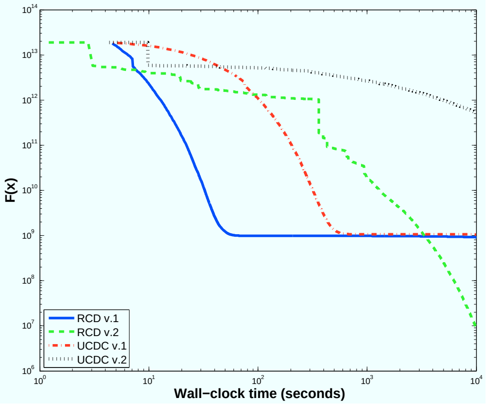
Then $\|g_i(x_k^{(i)}; x_k)\|_2$ has a superlinear rate of convergence in expectation.

# Numerical experiments: synthetic L1 least squares

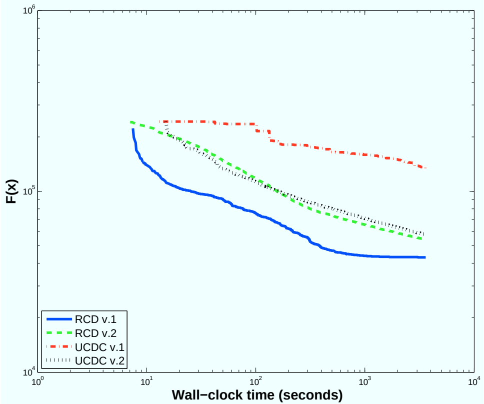
**Solvers**

- UCDC v.1: Single coordinate descent with simple quadratic
- UCDC v.2: Block coordinate descent with simple quadratic
- RCD v.1: $H_k^{(i)} := diag(\nabla_i^2 f(x_k))$. RCD v.2: $H_k := \nabla_i^2 f(x_k)$

**Instance info:** $N = 2^{21}$, $m = N/4$, $\text{nnz}(A) = 10^{-4} mN$ and Blocks $\approx 10^{-2} N$.



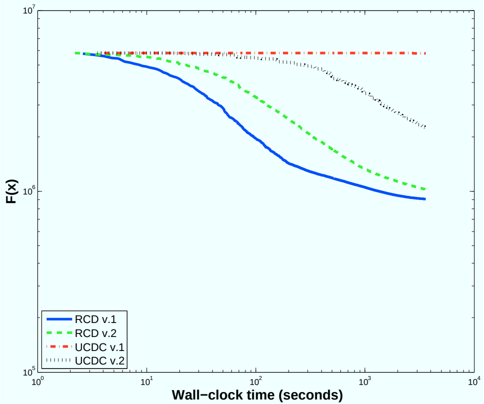
# Numerical experiments: real world binary classification




**Info**

Name: webspam

$N \approx 16$ million, $m \approx 0.02N$




**Info**

Name: kdd2010 (algebra)

$N \approx 20$ million, $m \approx 0.4N$

Thank you!

**Paper:** K. Fountoulakis and R. Tappenden. Robust block coordinate descent. *Technical Report ERGO-14-010*, 2014

**Software:** `http://www.maths.ed.ac.uk/~kfount/` (only for reproduction of the presented experiments)