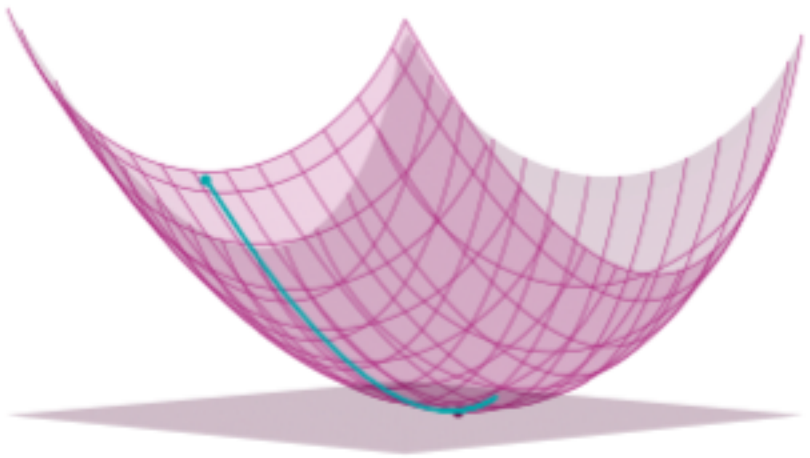


Block Coordinate Descent and Exact Minimization

Jelena Diakonikolas
Boston University

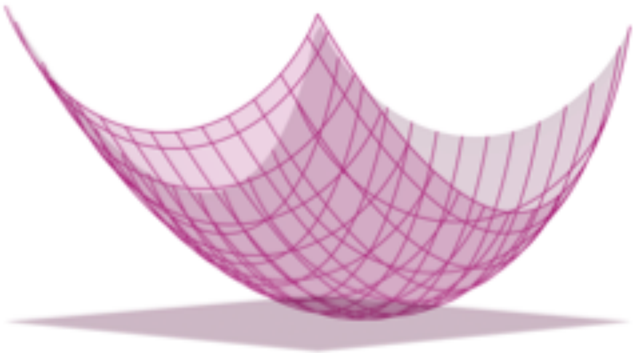
joint work with Lorenzo Orecchia (BU)

Workshop on Local Algorithms
June 2018

Full-Gradient First-Order Convex Optimization

Unconstrained convex minimization:

$$\boxed{\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in \mathbb{R}^N \end{array}}$$



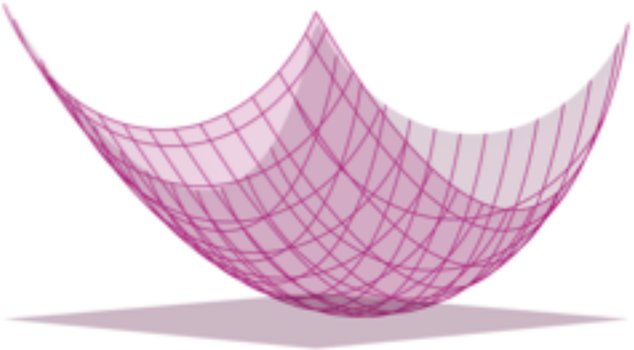
First-order blackbox (oracle) model:

$\mathbf{x} \in \mathbb{R}^N$ $\longrightarrow$

$\longrightarrow$ $\nabla f(\mathbf{x}),\ f(\mathbf{x})$

History

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in \mathbb{R}^N \end{array}$$



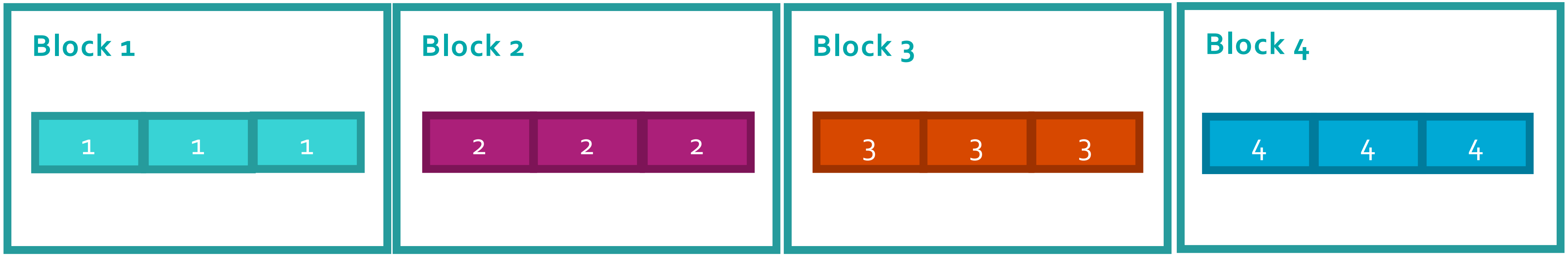
- Methods with optimal iteration complexity in various settings are well-known:
  - Gradient descent – folklore
  - Nemirovski's mirror descent [Nemirovski, Yudin'83]
  - Nesterov's accelerated method (AGD) [Nesterov'83]
  - Frank-Wolfe methods [Frank, Wolfe'56]
  - ... and many more – books: [Bubeck'14], [Sra, Nowozin, Wright'11]
- Typical complexity of an iteration is near-linear in the input size, few iterations
- Particularly attractive for large-scale problems; broad applications in machine learning and TCS

Block Coordinate Descent: Setting

- Fix a partition of the vector of variables into n blocks:

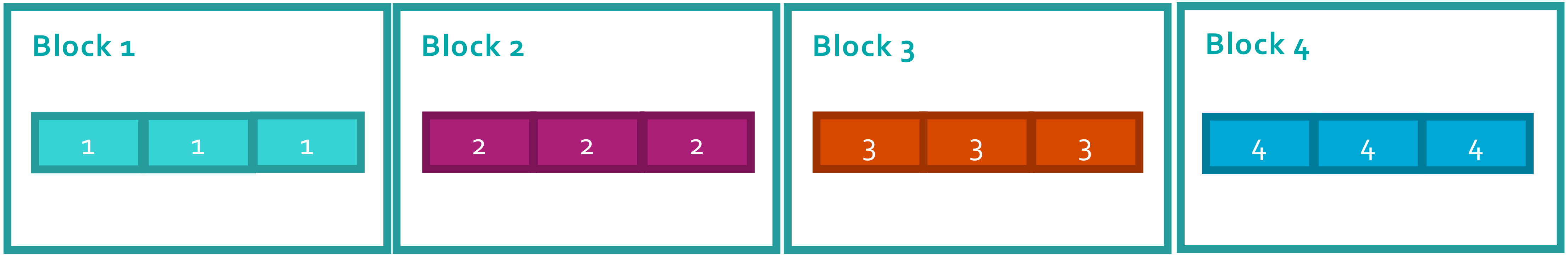
Block Coordinate Descent: Setting

- Fix a partition of the vector of variables into n blocks:



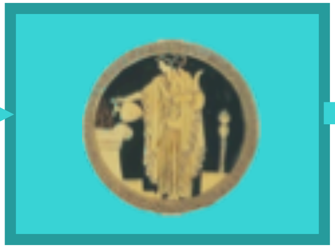
Block Coordinate Descent: Setting

- Fix a partition of the vector of variables into n blocks:



- Assume access to two types of oracles:

$\mathbf{x} \in \mathbb{R}^n, i$



$\nabla_i f(\mathbf{x})$

first-order oracle

$\mathbf{x} \in \mathbb{R}^n, i$

$\arg\min_{\mathbf{u}} f(\mathbf{x}_1, .., \underbrace{\mathbf{u}}_{i}, ..., \mathbf{x}_n)$

minimization oracle

Assumptions about the Problem

- **Assumptions**:
  - Function is differentiable and L-smooth:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y}$$

  - Each block i is L_i-smooth:

$$\|\nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{y})\|_* \leq L_i\|\mathbf{x}_i - \mathbf{y}_i\|, \ \forall \mathbf{x}, \mathbf{y}, \text{ where } \ \mathbf{y}_k = \mathbf{x}_k \ \text{ for } \ k \neq i$$

  - Block n is "least" smooth, possibly with $L_n = \infty$:

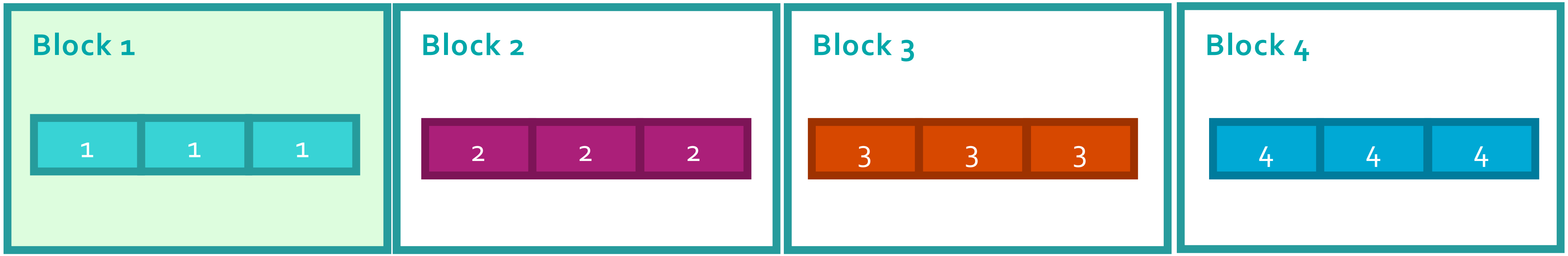
$$L_n = L_{\max} = \max_{1 \leq i \leq n} L_i$$

Note that if $L_n = \infty$, then it must be $L = \infty$!

Basic (Nonaccelerated) Methods

Cyclic Block Coordinate Descent

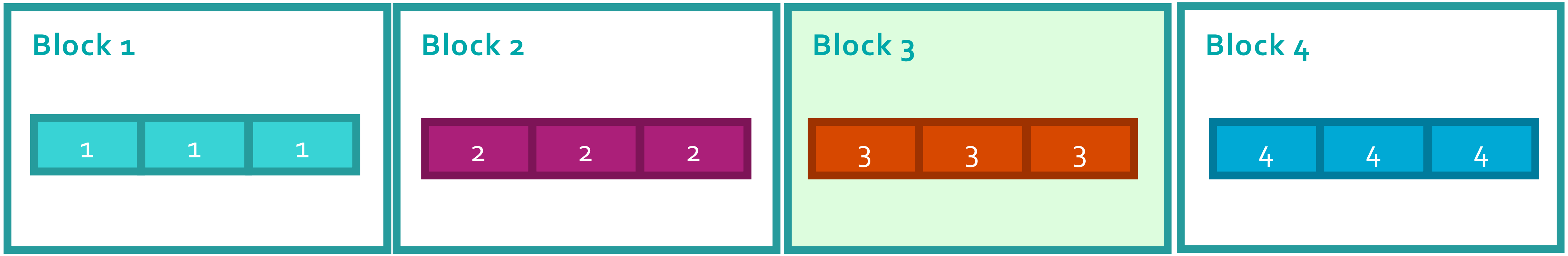
- Almost a folklore method [Ortega & Rheinboldt, 1970]
- Fix a (possibly random) permutation of the blocks
- Take either exact minimization or a gradient step over a block selected in the cyclic order



Example order: 1, 3, 2, 4

Cyclic Block Coordinate Descent

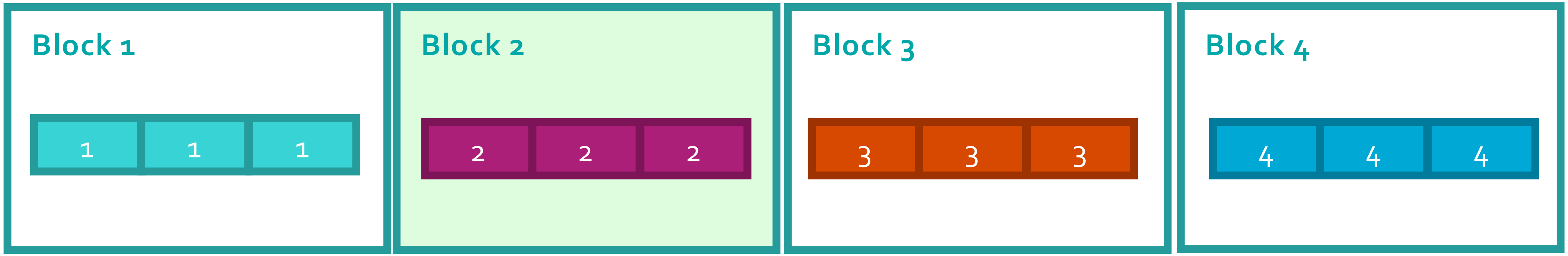
- Almost a folklore method [Ortega & Rheinboldt, 1970]
- Fix a (possibly random) permutation of the blocks
- Take either exact minimization or a gradient step over a block selected in the cyclic order



Example order: 1, 3, 2, 4

Cyclic Block Coordinate Descent

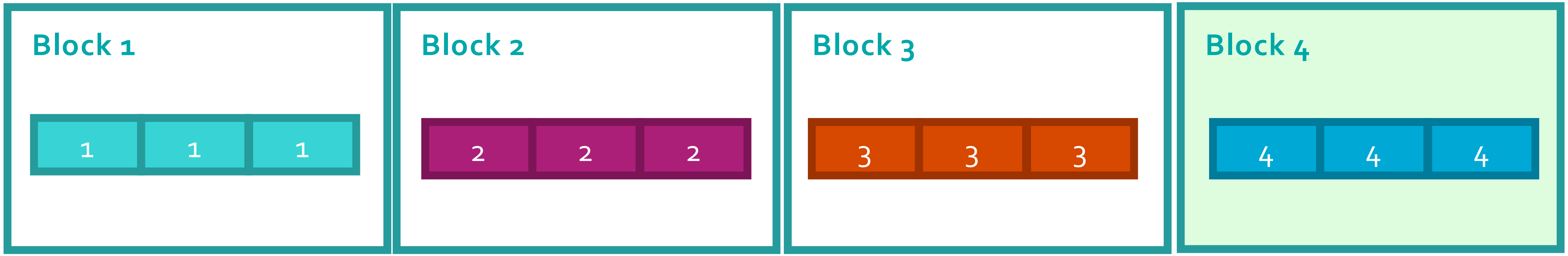
- Almost a folklore method [Ortega & Rheinboldt, 1970]
- Fix a (possibly random) permutation of the blocks
- Take either exact minimization or a gradient step over a block selected in the cyclic order



Example order: 1, 3, 2, 4

Cyclic Block Coordinate Descent

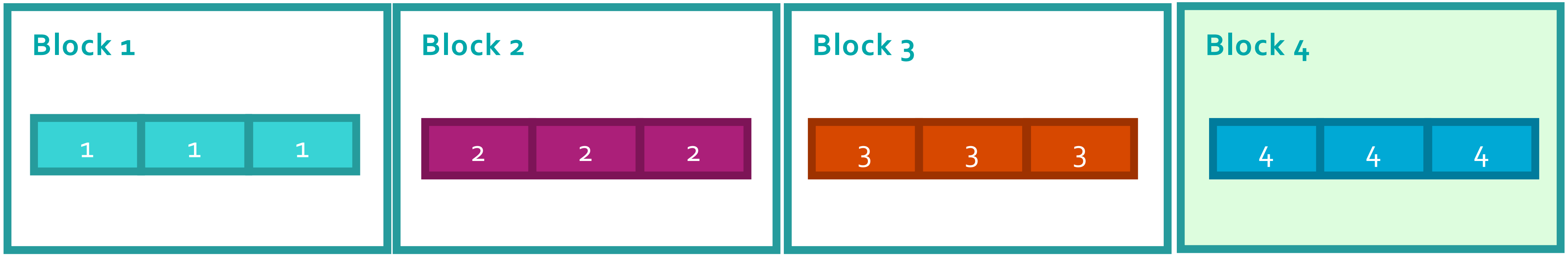
- Almost a folklore method [Ortega & Rheinboldt, 1970]
- Fix a (possibly random) permutation of the blocks
- Take either exact minimization or a gradient step over a block selected in the cyclic order



Example order: 1, 3, 2, 4

Cyclic Block Coordinate Descent

- Almost a folklore method [Ortega & Rheinboldt, 1970]
- Fix a (possibly random) permutation of the blocks
- Take either exact minimization or a gradient step over a block selected in the cyclic order



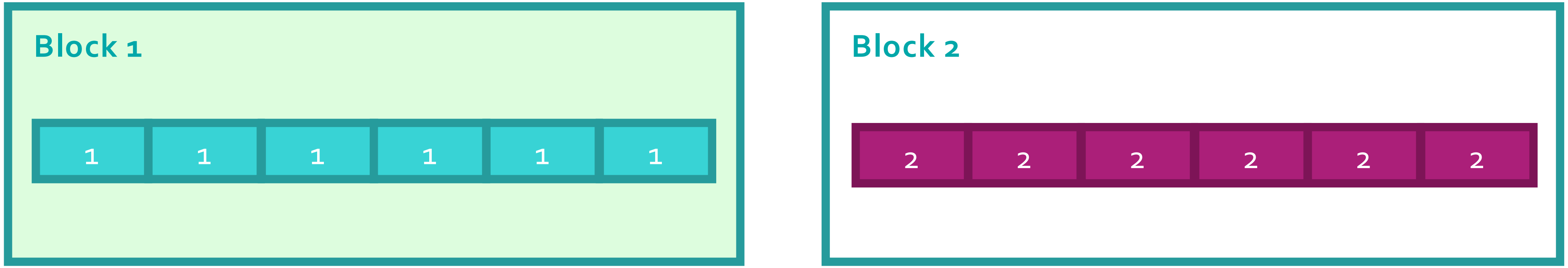
Example order: 1, 3, 2, 4

Dependence of the optimality gap on smoothness parameters:

$L_n + \frac{\min(nL^2, (\sum_{i=1}^{n} L_i)^2)}{L_{\min}}$ [Sun, Hong'15], Ln^3 [Hong, Wang, Razaviyayn, Luo'17]

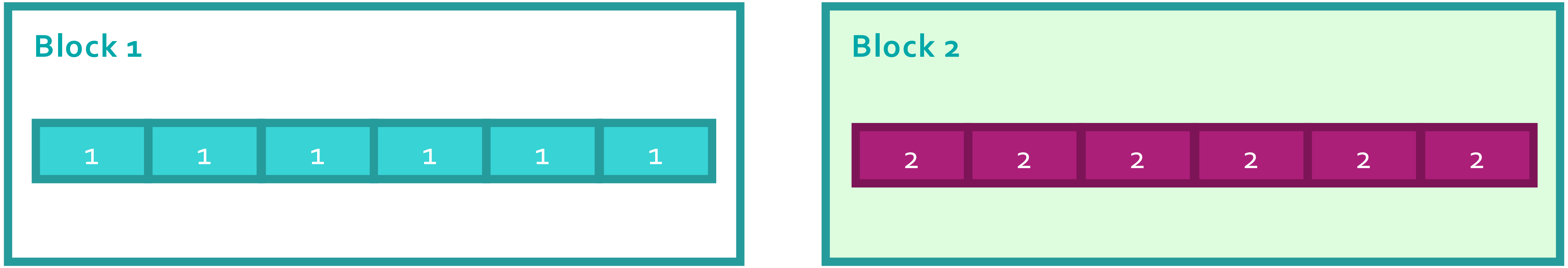
Alternating Minimization

- A special case of cyclic BCD when there are only two blocks
- Exact minimization on the less smooth block; exact minimization or gradient descent step on the other block



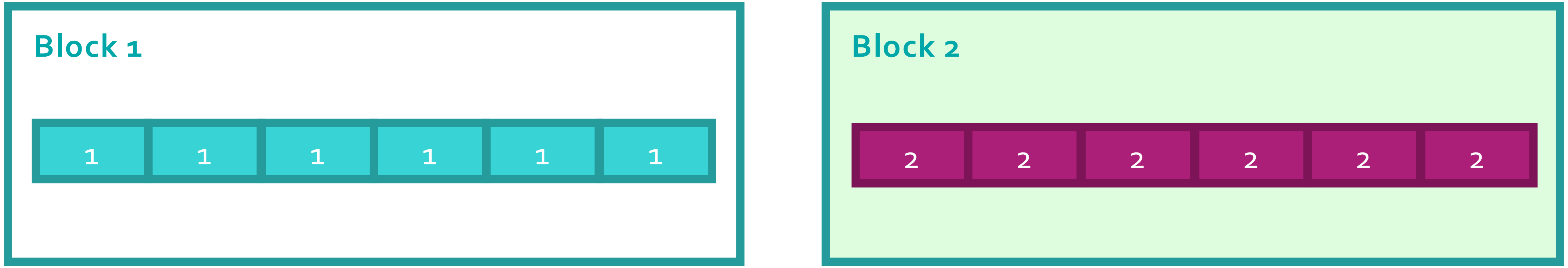
Alternating Minimization

- A special case of cyclic BCD when there are only two blocks
- Exact minimization on the less smooth block; exact minimization or gradient descent step on the other block



Alternating Minimization

- A special case of cyclic BCD when there are only two blocks
- Exact minimization on the less smooth block; exact minimization or gradient descent step on the other block



Dependence of the optimality gap on smoothness parameters:

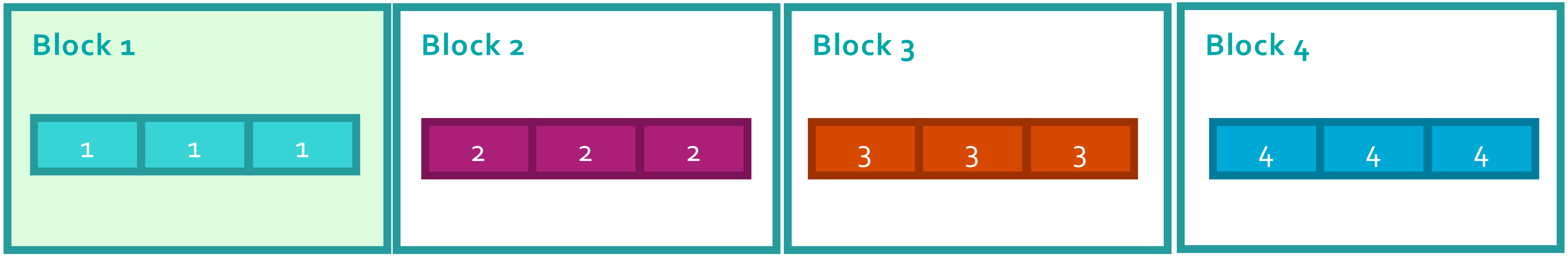
$$\min(L_1, L_2) \quad \text{[Beck'15]}$$

Randomized Block Coordinate Descent

- Introduced by [Nesterov, 2012]
- Fix a probability distribution $\{p_i\}_{i=1}^{n}$ over the blocks
- Select block i with probability p_i and do a gradient descent step or exact minimization over it

i

1

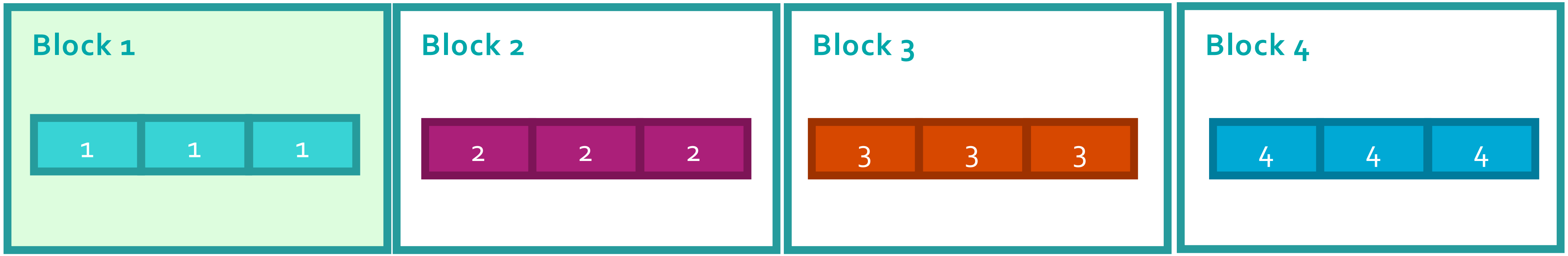


Randomized Block Coordinate Descent

- Introduced by [Nesterov, 2012]
- Fix a probability distribution $\{p_i\}_{i=1}^{n}$ over the blocks
- Select block i with probability p_i and do a gradient descent step or exact minimization over it

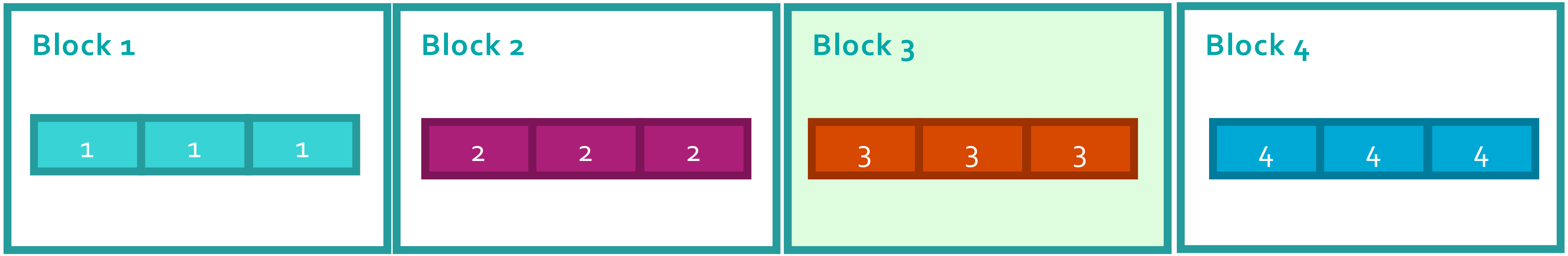
i

1



Randomized Block Coordinate Descent

- Introduced by [Nesterov, 2012]
- Fix a probability distribution $\{p_i\}_{i=1}^{n}$ over the blocks
- Select block i with probability p_i and do a gradient descent step or exact minimization over it

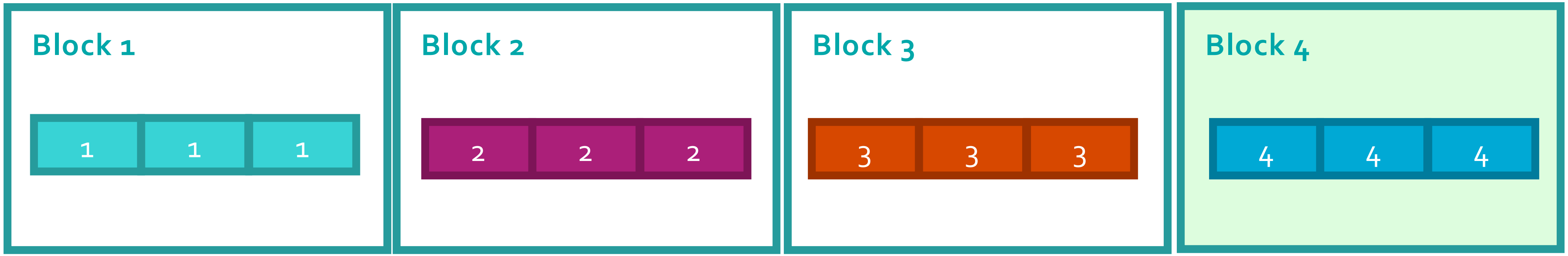


Randomized Block Coordinate Descent

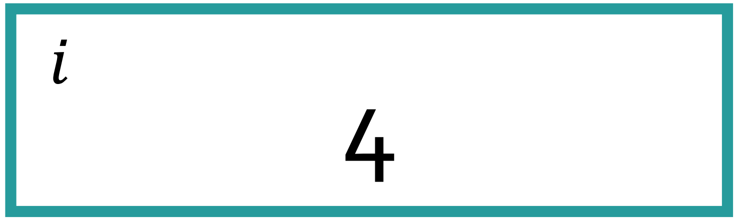
- Introduced by [Nesterov, 2012]
- Fix a probability distribution $\{p_i\}_{i=1}^{n}$ over the blocks
- Select block i with probability p_i and do a gradient descent step or exact minimization over it

i

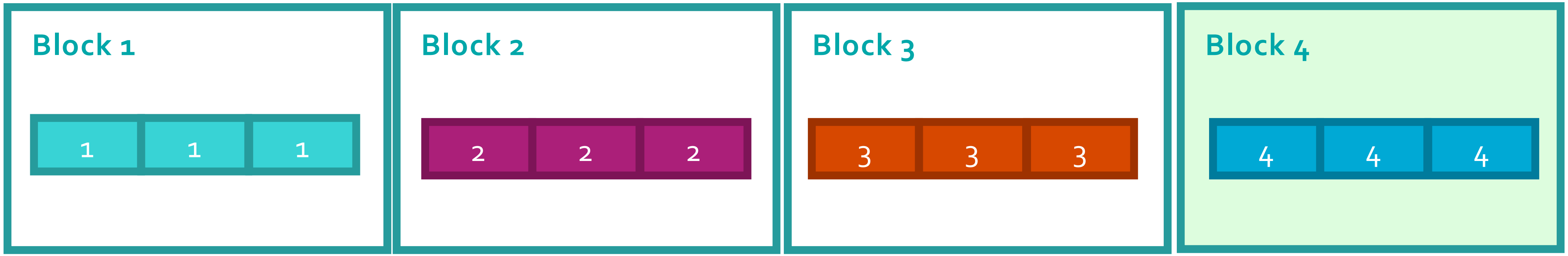
4



Randomized Block Coordinate Descent



- Introduced by [Nesterov, 2012]
- Fix a probability distribution $\{p_i\}_{i=1}^{n}$ over the blocks
- Select block i with probability p_i and do a gradient descent step or exact minimization over it



Dependence of the optimality gap on smoothness parameters:

$\sum_{i=1}^{n} L_i$ for $p_i \sim L_i$ [Nesterov'12]

Dependence of the optimality gap on smoothness parameters:

- Cyclic block coordinate descent (n blocks):

$$L_n + \frac{\min(nL^2, (\sum_{i=1}^n L_i)^2)}{L_{\min}} \text{ [Sun, Hong'15]}, \quad Ln^3 \text{ [Hong, Wang, Razaviyayn, Luo'17]}$$

- Randomized block coordinate descent (n blocks):

$$\sum_{i=1}^n L_i \text{ for } p_i \sim L_i \text{ [Nesterov'12]}$$

- Alternating minimization (2 blocks):

$$\min(L_1, L_2) \text{ [Beck'15]}$$

So far, only alternating minimization (two blocks) can avoid paying for the less-smooth block

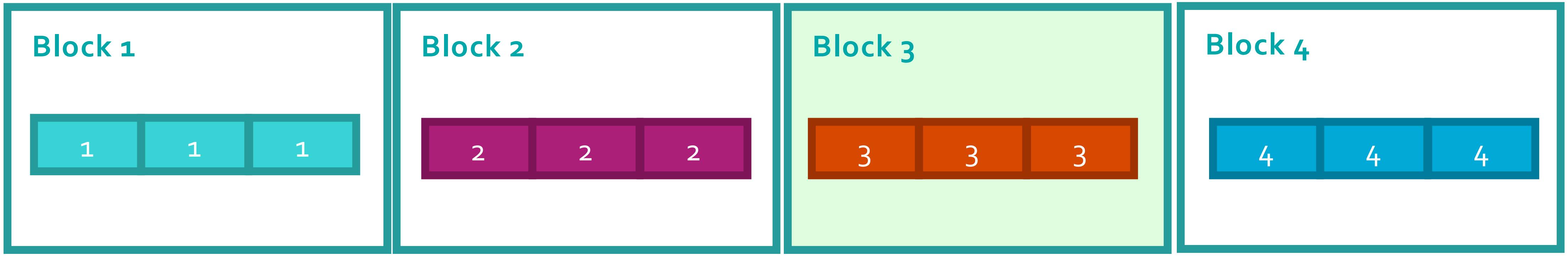
Q. Can we avoid paying for the least-smooth block when there are more than two blocks?

Alternating Randomized Block Coordinate Descent

- … or how to avoid paying for the least smooth (non-smooth) block
- Fix a probability distribution $\{p_i\}_{i=1}^{n-1}$ over blocks $1, 2, \ldots, n-1$
- Do a gradient descent (or exact min) step over block i, then exact minimization over block n

i

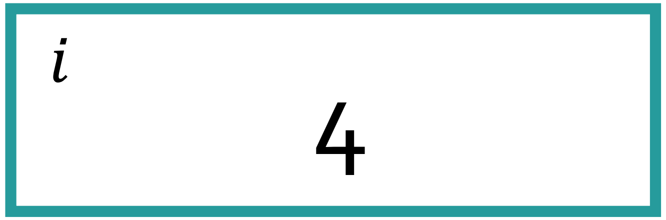
3

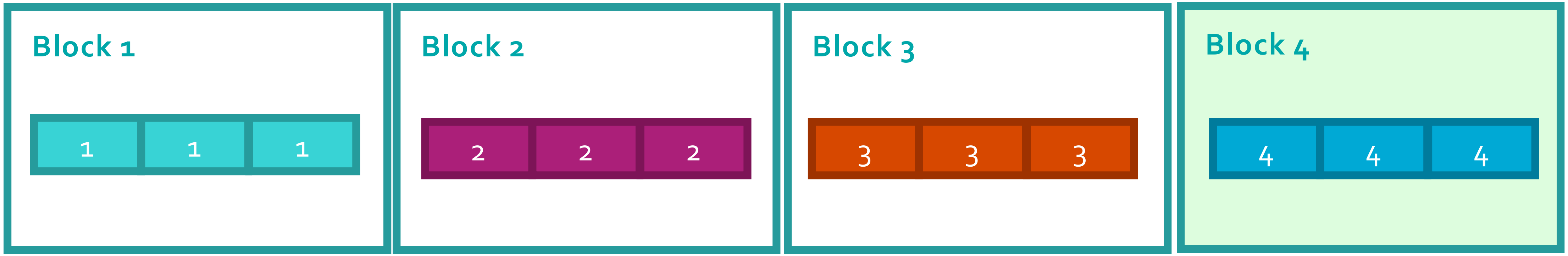


*D, Orecchia, "Alternating Randomized Block Coordinate Descent," in Proc. ICML'18, 2018, to appear.

Alternating Randomized Block Coordinate Descent

- … or how to avoid paying for the least smooth (non-smooth) block
- Fix a probability distribution $\{p_i\}_{i=1}^{n-1}$ over blocks $1, 2, \ldots, n-1$
- Do a gradient descent (or exact min) step over block i, then exact minimization over block n

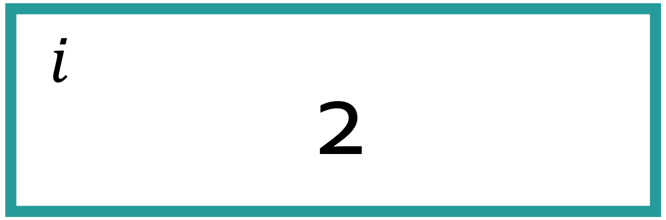


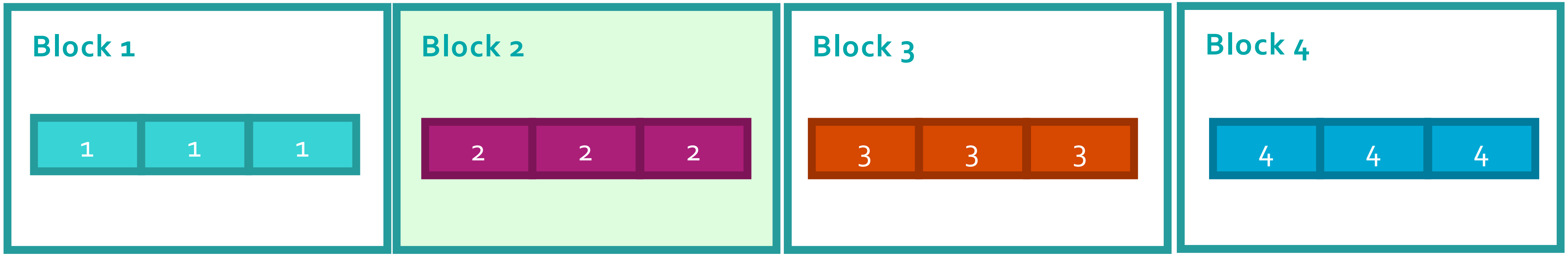


*D, Orecchia, "Alternating Randomized Block Coordinate Descent," in Proc. ICML'18, 2018, to appear.

Alternating Randomized Block Coordinate Descent

- … or how to avoid paying for the least smooth (non-smooth) block
- Fix a probability distribution $\{p_i\}_{i=1}^{n-1}$ over blocks $1, 2, \ldots, n-1$
- Do a gradient descent (or exact min) step over block i, then exact minimization over block n





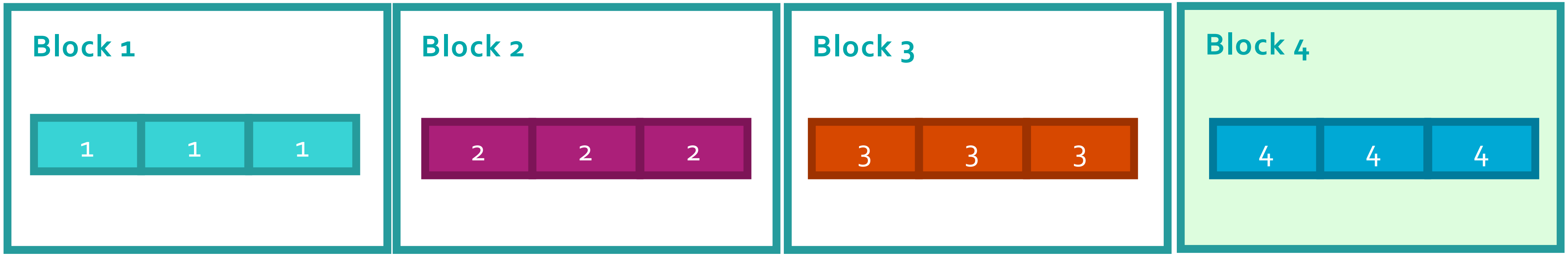
*D, Orecchia, "Alternating Randomized Block Coordinate Descent," in Proc. ICML'18, 2018, to appear.

Alternating Randomized Block Coordinate Descent

- … or how to avoid paying for the least smooth (non-smooth) block
- Fix a probability distribution $\{p_i\}_{i=1}^{n-1}$ over blocks $1, 2, \dots, n-1$
- Do a gradient descent (or exact min) step over block i, then exact minimization over block n

i

4



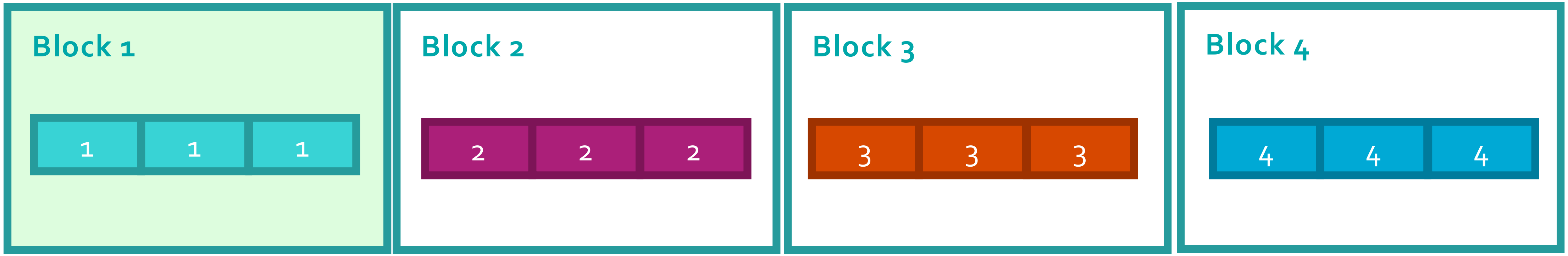
*D, Orecchia, "Alternating Randomized Block Coordinate Descent," in Proc. ICML'18, 2018, to appear.

Alternating Randomized Block Coordinate Descent

- … or how to avoid paying for the least smooth (non-smooth) block
- Fix a probability distribution $\{p_i\}_{i=1}^{n-1}$ over blocks $1, 2, \ldots, n-1$
- Do a gradient descent (or exact min) step over block i, then exact minimization over block n

i

1



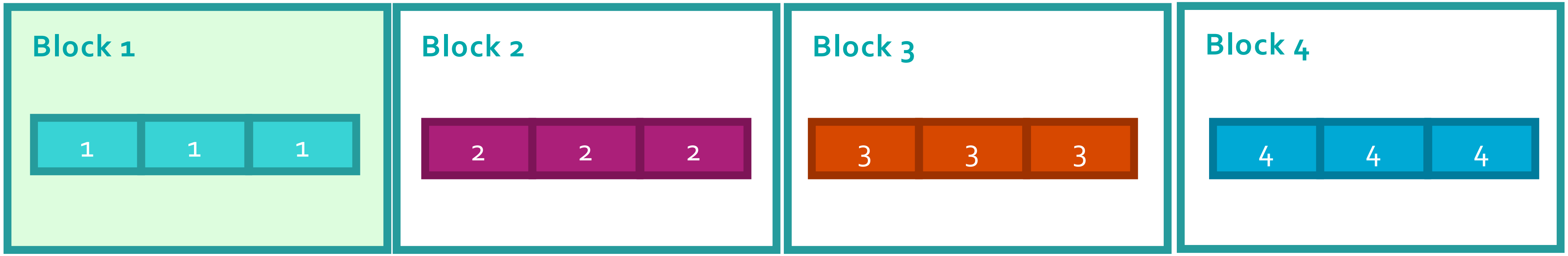
*D, Orecchia, "Alternating Randomized Block Coordinate Descent," in Proc. ICML'18, 2018, to appear.

Alternating Randomized Block Coordinate Descent

- ... or how to avoid paying for the least smooth (non-smooth) block
- Fix a probability distribution $\{p_i\}_{i=1}^{n-1}$ over blocks $1, 2, \ldots, n-1$
- Do a gradient descent (or exact min) step over block i, then exact minimization over block n

i

1



Generalizes randomized BCD and alternating minimization

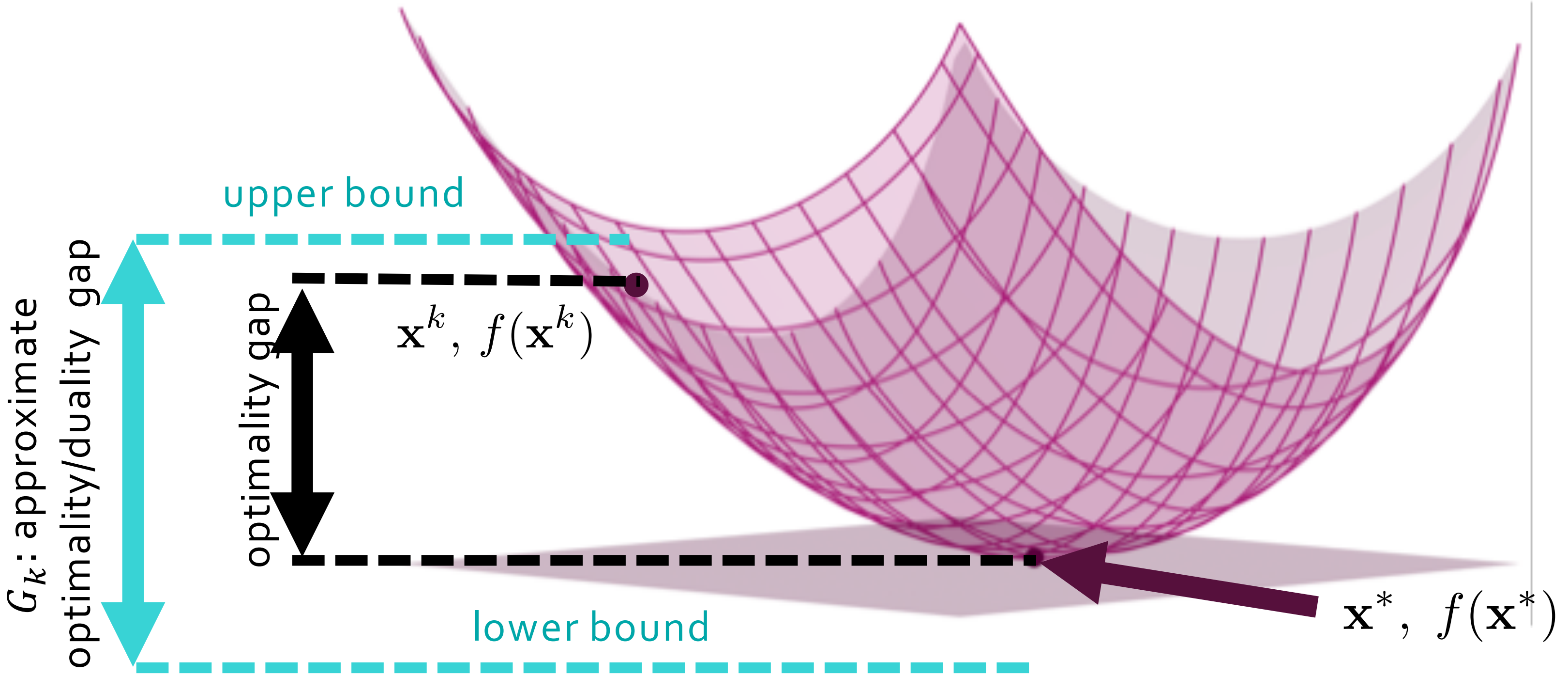
Dependence of the optimality gap on smoothness parameters: $\sum_{i=1}^{n-1} L_i$ for $p_i \sim L_i$ **(no dependence on L_n!)**

Possible to accelerate: gives the same convergence time as the fastest known accelerated BCD – NUACDM [Allen-Zhu, Qu, Richtárik, Yuan'16], except without any dependence on L_n

***D**, Orecchia, "Alternating Randomized Block Coordinate Descent," in Proc. ICML'18, 2018, to appear.

Convergence Analysis: Main Ideas

- Extension of Approximate Duality Gap Technique [**D**, Orecchia, 2017]



$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \leq G_k$. Let A_k be an increasing (rate) function of iteration count k. Then, if $\boxed{A_k G_k \leq A_{k-1} G_{k-1}, \forall k}$

$$\boxed{f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \leq G_k \leq \frac{A_0 G_0}{A_k}}$$

Convergence Analysis: Main Ideas

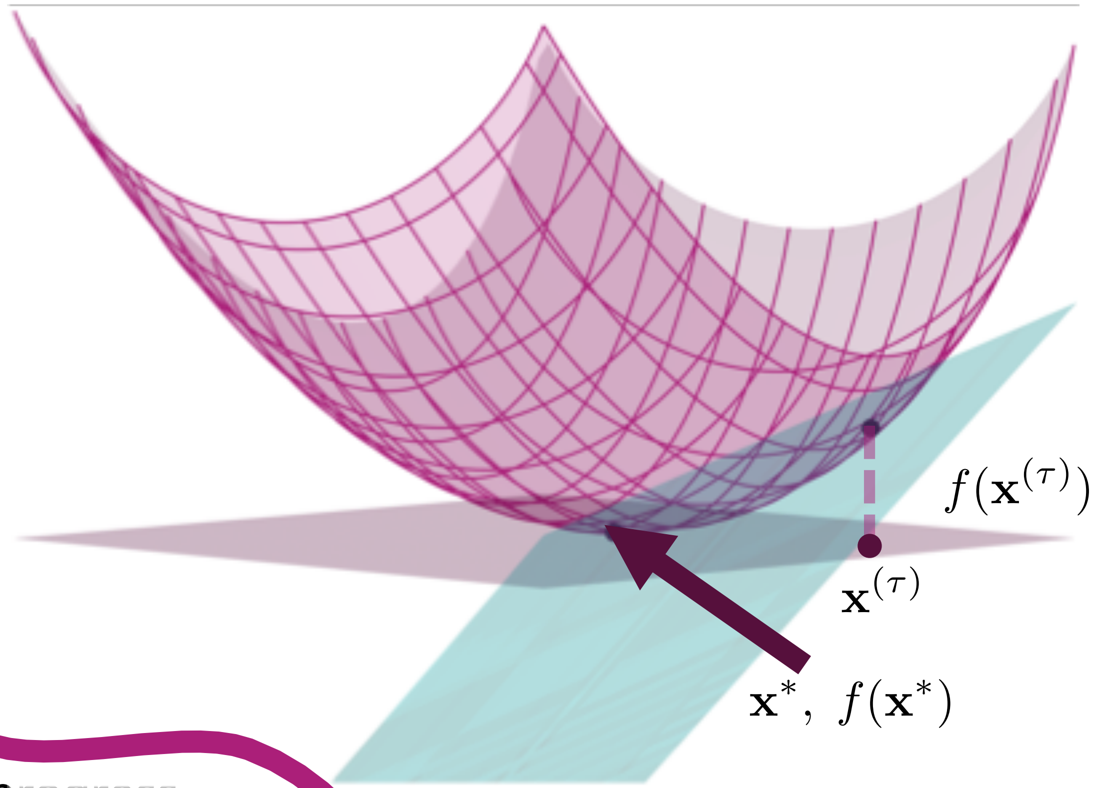
- Lower bound uses the following:
  - by convexity (and differentiability), $\forall \boldsymbol{x}_k$:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k \rangle$$

$$= f(\mathbf{x}^k) + \sum_{i=1}^{n} \langle \nabla_i f(\mathbf{x}^k), \mathbf{x}_i^* - \mathbf{x}_i^k \rangle$$

$$= f(\mathbf{x}^k) + \sum_{i=1}^{\boxed{n-1}} \langle \nabla_i f(\mathbf{x}^k), \mathbf{x}_i^* - \mathbf{x}_i^k \rangle$$

because $\nabla_i f(\mathbf{x}^k) = \mathbf{0}$



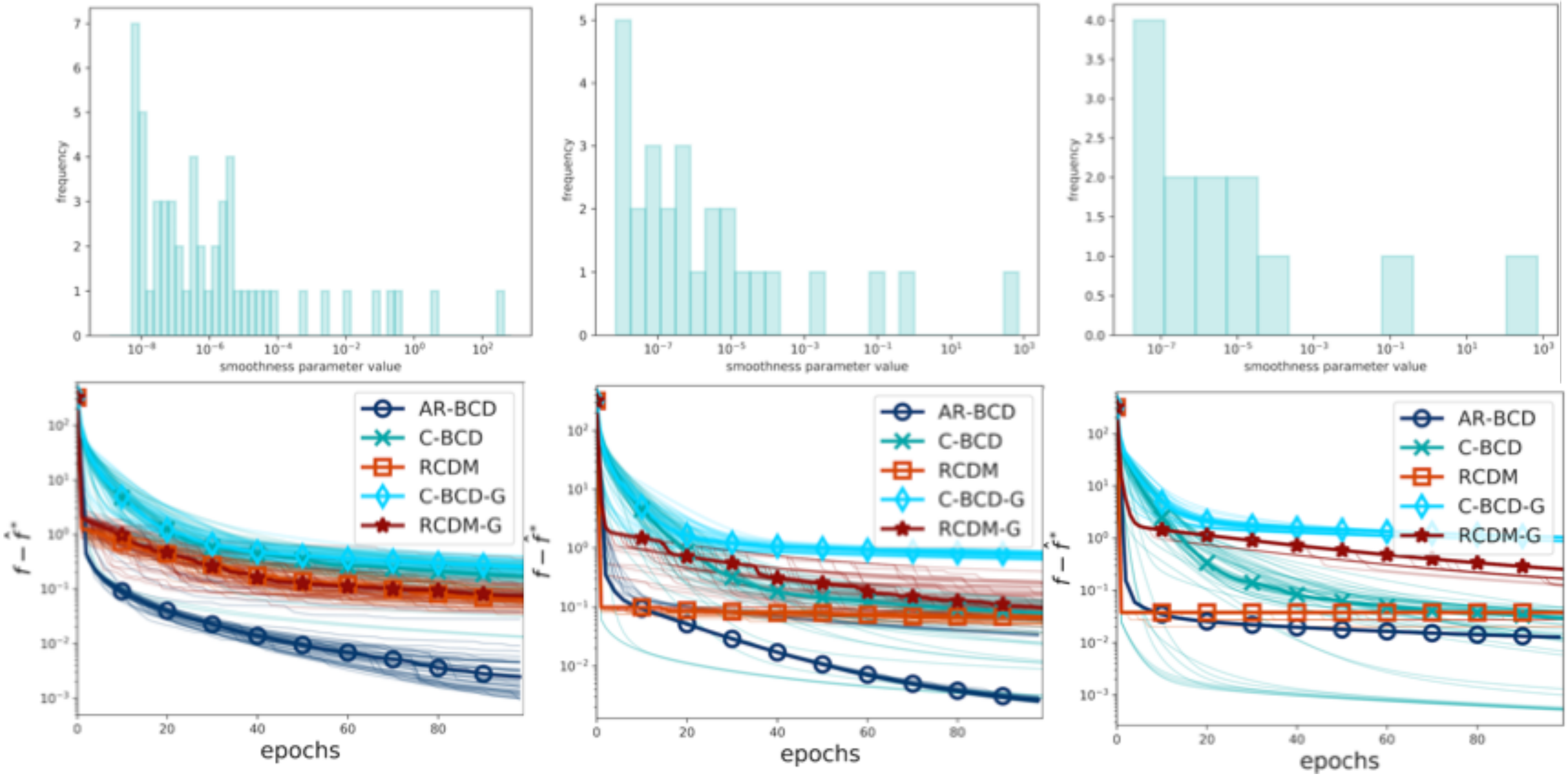
- Upper bound uses the (block) gradient descent progress:

$$f(\mathrm{Grad}_i(\mathbf{x}^k)) \leq f(\mathbf{x}^k) \boxed{- \frac{1}{2L_i} \|\nabla f(\mathbf{x}_k)\|_*^2}$$

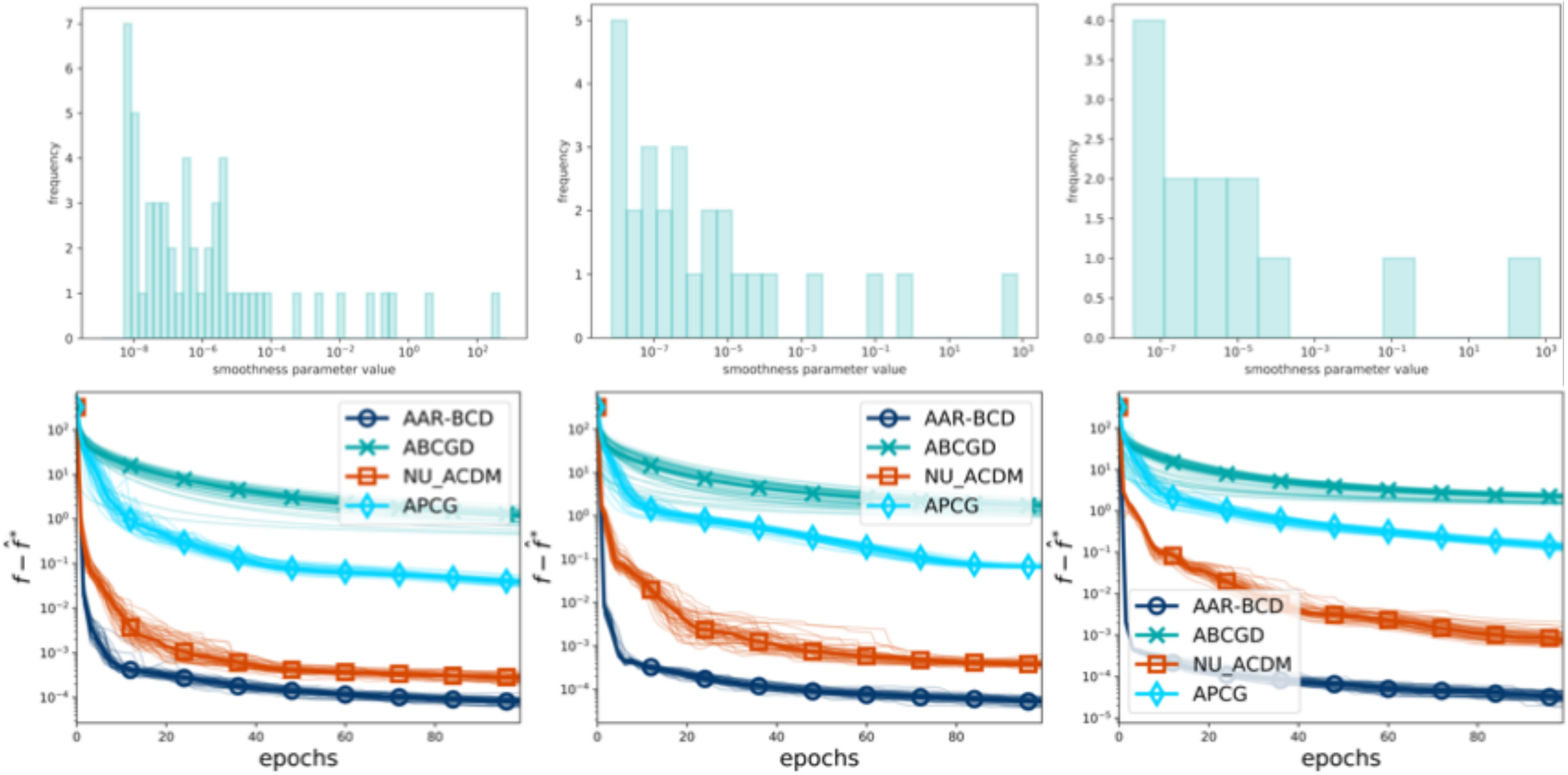
can sample only over the first $n-1$ (i.e., “smoother”) blocks

Numerical Experiments

Experiments: Linear Regression on BlogFeedback Dataset



Experiments: Linear Regression on BlogFeedback Dataset



Summary

- A novel block coordinate descent method (and its accelerated version) that can handle a completely non-smooth block (structured non-smoothness)

- The method outperforms existing methods if one block has much worse (but finite) smoothness parameter than the remaining ones

- **Ongoing work:**
  - **Extension to the smooth and strongly convex setting**
  - **Extension to the composite non-smooth setting**
  - **Improved convergence bounds for randomized BCD with exact minimization**
- **Open question:**
  - **We need to know which block is the least smooth to not pay for it. Is it possible to relax this?**

jelena@jelena-diakonikolas.com
www.jelena-diakonikolas.com

Thank you!