E. Pauwels
joint work with A. Beck and S. Sabach.

Séminaire MIAT INRA
September 23 2016
Two old ideas have received renewed attention in the past years:

**Block decomposition:**

\[
x = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}
\]

**Linear oracles:**

\[
\min_{x \in X} \langle x, c \rangle
\]
Two old ideas have received renewed attention in the past years:

**Block decomposition:**

\[
\mathbf{x} = \begin{pmatrix}
    \mathbf{x}_1 \\
    \\
    \\
    \mathbf{x}_N
\end{pmatrix}
\]

**Linear oracles:**

\[
\min_{\mathbf{x} \in X} \langle \mathbf{x}, \mathbf{c} \rangle
\]

**Coordinate descent:**

- Large dimension
- Distributed data

**Conditional gradient:**

- “Complex constraints”
- Primal-dual interpretation
Two old ideas have received renewed attention in the past years:

**Block decomposition:**

\[
\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}
\]

**Linear oracles:**

\[
\min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \mathbf{c} \rangle
\]

**Coordinate descent:**
- Large dimension
- Distributed data

**Conditional gradient:**
- “Complex constraints”
- Primal-dual interpretation

Theoretical properties and empirical performances?
Scope of the presentation

- Most results in the literature hold for random block selection rules.
- Lacoste-Julien and co-authors analyzed the random block conditional gradient method (RBCG).
  - *Block-Coordinate Frank-Wolfe Optimization for Structural SVMs* (ICML 2013)
- We propose a convergence analysis for the cyclic block variant (CBCG).
Most results in the literature hold for random block selection rules.
Lacoste-Julien and co-authors analyzed the random block conditional gradient method (RBCG).

- *Block-Coordinate Frank-Wolfe Optimization for Structural SVMs* (ICML 2013)

We propose a convergence analysis for the cyclic block variant (CBCG).

This presentation: focus on machine learning related aspects
- General introduction to linear oracle based optimization methods.
- Specification to (regularized) empirical risk minimization (ERM).
- Details about the application to structured SVM.

(Taskar et. al., 2003 – Tsochantaridis et. al., 2005)
Outline

1. Context

2. Conditional Gradient algorithm

3. CG and convex duality

4. Block CG and $L_2$ regularized ERM

5. Results
Main idea

**Optimization setting:** \( f : \mathbb{R}^n \to \mathbb{R} \) is convex, \( C_1 \) with \( L \)-Lipschitz gradient over \( X \subseteq \mathbb{R}^n \) which is convex and compact.

\[
\bar{f} := \min_{x \in X} f(x)
\]
Main idea

**Optimization setting:** $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex, $C_1$ with $L$-Lipschitz gradient over $X \subset \mathbb{R}^n$ which is convex and compact.

$$\bar{f} := \min_{x \in X} f(x)$$

Start with $x^0 \in X$

$$p^k \in \arg\max_{y \in X} \langle \nabla f(x^k), x^k - y \rangle$$

$$x^{k+1} = (1 - \alpha^k)x^k + \alpha^k p^k \quad 0 \leq \alpha^k \leq 1$$
Main idea

Optimization setting: \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is convex, \( C_1 \) with \( L \)-Lipschitz gradient over \( X \subset \mathbb{R}^n \) which is convex and compact.

\[
\bar{f} := \min_{x \in X} f(x)
\]

Start with \( x^0 \in X \)

\[
p^k \in \arg\max_{y \in X} \langle \nabla f(x^k), x^k - y \rangle
\]

\[
x^{k+1} = (1 - \alpha^k)x^k + \alpha^k p^k \quad 0 \leq \alpha^k \leq 1
\]

Step size:

- \( \alpha^k = \frac{2}{k+2} \) \hspace{2cm} \text{Open loop}
- \( x^{k+1} = \arg\min_{y \in [x^k, p^k]} f(y) \) \hspace{2cm} \text{Exact line search}
- \( x^{k+1} = \arg\min_{y \in [x^k, p^k]} Q(x^k, y) \) \hspace{2cm} \text{Approximate line search}

\[
f(y) \leq Q(x, y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2
\]

(tangent quadratic upper bound, descent Lemma).
Historical remarks

Fifty years ago:

- First appearance for quadratic programs (Frank, Wolfe, 1956).
- $f(x^k) - \bar{f} = O(1/k)$ (Polyak, Dunn, Dem’Yanov . . . , 60’s).
- For any $\epsilon > 0$, it cannot be $O(1/k^{1+\epsilon})$ (Canon, Cullum, Polyak, 60’s).

Recent developments (illustrations follow):

- Revival for large scale problems.
- Primal dual interpretation (Bach 2015) and convergence analysis (Jaggi 2013)
- Block decomposition variants (Lacoste-Julien et al. 2013)
Historical remarks

Fifty years ago:

- First appearance for quadratic programs (Frank, Wolfe, 1956).
- $f(x^k) - \bar{f} = O(1/k)$ (Polyak, Dunn, Dem’Yanov . . . , 60’s).
- For any $\epsilon > 0$, it cannot be $O(1/k^{1+\epsilon})$ (Canon, Cullum, Polyak, 60’s).

Recent developments (illustrations follow):

- Revival for large scale problems.
- Primal dual interpretation (Bach 2015) and convergence analysis (Jaggi 2013)
- Block decomposition variants (Lacoste-Julien et al. 2013)
Why is it interesting?

- $O(1/k^2)$ can be achieved by using projections (Beck, Teboulle 2009).
- Conditional Gradient does not compete in practice.
Why is it interesting?

- $O(1/k^2)$ can be achieved by using projections (Beck, Teboulle 2009).
- Conditional Gradient does not compete in practice.

In some situations, projection does not constitute a practical alternative. Linear programs on convex sets attain their value at extreme points.
Why is it interesting?

- $O(1/k^2)$ can be achieved by using projections (Beck, Teboulle 2009).
- Conditional Gradient does not compete in practice.

In some situations, projection does not constitute a practical alternative. Linear programs on convex sets attain their value at extreme points.

**Trace norm:**
For $M \in \mathbb{R}^{m \times n}$, $\|M\|_* = \sum_i \sigma_i$, where $\{\sigma_i\}$ is the set of singular values of $M$.

- Projection on the trace norm ball is a thresholding of singular values $\rightarrow$ full SVD.
- Linear programming on the trace norm ball is finding the largest singular value $\rightarrow$ leading singular vector.
Outline

1. Context

2. Conditional Gradient algorithm

3. CG and convex duality

4. Block CG and $L_2$ regularized ERM

5. Results
Recall that $X$ is convex and compact. Define its support function $g: \mathbb{R}^n \to \mathbb{R}^n$

$$g: \mathbf{w} \to \max_{\mathbf{x} \in X} \langle \mathbf{x}, \mathbf{w} \rangle$$
Recall that $X$ is convex and compact. Define its support function $g : \mathbb{R}^n \to \mathbb{R}^n$

$$g : \mathbf{w} \to \max_{\mathbf{x} \in X} \langle \mathbf{x}, \mathbf{w} \rangle$$

Given $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$, consider the problems

$$\bar{p} = \min_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{2} \| \mathbf{w} \|^2 + g(-A\mathbf{w} + b) \quad (= P(\mathbf{w}))$$

$$\bar{d} = \min_{\mathbf{x} \in X} \frac{1}{2} \| A^T \mathbf{x} \|^2 - \langle \mathbf{x}, \mathbf{b} \rangle \quad (= D(\mathbf{x}))$$
Convex duality

Recall that $X$ is convex and compact. Define its support function $g : \mathbb{R}^n \to \mathbb{R}^n$

$$g : \mathbf{w} \mapsto \max_{\mathbf{x} \in X} \langle \mathbf{x}, \mathbf{w} \rangle$$

Given $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$, consider the problems

$$\bar{p} = \min_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{2} \| \mathbf{w} \|_2^2 + g(-A \mathbf{w} + b) \quad (= P(\mathbf{w}))$$

$$\bar{d} = \min_{\mathbf{x} \in X} \frac{1}{2} \| A^T \mathbf{x} \|_2^2 - \langle \mathbf{x}, b \rangle \quad (= D(\mathbf{x}))$$

- Weak duality: for any $\mathbf{w} \in \mathbb{R}^m$ and $\mathbf{x} \in X$,
  $$P(\mathbf{w}) + D(\mathbf{x}) \geq 0$$

- Strong duality holds
  $$\bar{p} + \bar{d} = 0$$
Primal subgradient and dual conditional gradient

\[ g: \mathbf{w} \rightarrow \max_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \mathbf{w} \rangle \]
\[ (\mathbf{x} \in \arg\max \iff \mathbf{x} \in \partial g(\mathbf{w})) \]

\[ \bar{p} = \min_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{2} \| \mathbf{w} \|_2^2 + g(-A\mathbf{w} + \mathbf{b}) \]
\[ (= P(\mathbf{w})) \]

\[ \bar{d} = \min_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \| A^T \mathbf{x} \|_2^2 - \langle \mathbf{x}, \mathbf{b} \rangle \]
\[ (= D(\mathbf{x})) \]
Primal subgradient and dual conditional gradient

\[ g : \mathbf{w} \rightarrow \max_{x \in X} \langle x, \mathbf{w} \rangle \quad (x \in \text{argmax} \iff x \in \partial g(\mathbf{w})) \]

\[ \bar{p} = \min_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{2} \| \mathbf{w} \|^2_2 + g(-A\mathbf{w} + \mathbf{b}) \quad (= P(\mathbf{w})) \]

\[ \bar{d} = \min_{x \in X} \frac{1}{2} \| A^T x \|^2_2 - \langle x, \mathbf{b} \rangle \quad (= D(x)) \]

A conditional gradient step in the dual:

\[ p^k : \max_{y \in X} \langle A A^T x^k - \mathbf{b}, x^k - y \rangle = \| A^T x^k \|^2_2 - \langle \mathbf{b}, x^k \rangle + g(-A A^T x^k + \mathbf{b}) \]

\[ = P(A^T x^k) + D(x^k) \]
Primal subgradient and dual conditional gradient

\[ g: \mathbf{w} \rightarrow \max_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \mathbf{w} \rangle \quad (\mathbf{x} \in \arg\max \Leftrightarrow \mathbf{x} \in \partial g(\mathbf{w})) \]

\[ \bar{p} = \min_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{2} \|\mathbf{w}\|_2^2 + g(-A\mathbf{w} + \mathbf{b}) \quad (= P(\mathbf{w})) \]

\[ \bar{d} = \min_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \|A^T \mathbf{x}\|_2^2 - \langle \mathbf{x}, \mathbf{b} \rangle \quad (= D(\mathbf{x})) \]

A conditional gradient step in the dual:

\[ p^k: \max_{\mathbf{y} \in \mathcal{X}} \langle A A^T \mathbf{x}^k - \mathbf{b}, \mathbf{x}^k - \mathbf{y} \rangle = \|A^T \mathbf{x}^k\|_2^2 - \langle \mathbf{b}, \mathbf{x}^k \rangle + g(-A A^T \mathbf{x}^k + \mathbf{b}) \]

\[ = P(A^T \mathbf{x}^k) + D(\mathbf{x}^k) \]

Consider the primal variable \( \mathbf{w}^k = A^T \mathbf{x}^k \): we have \( p^k \in \partial g(-A\mathbf{w}^k + \mathbf{b}) \).

\[ \mathbf{w}^{k+1} - \mathbf{w}^k = \alpha^k A^T (\mathbf{x}^k + p^k) = -\alpha^k \partial P(\mathbf{w}^k) \]
Primal subgradient and dual conditional gradient

\[ g: \mathbf{w} \rightarrow \max_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \mathbf{w} \rangle \quad (\mathbf{x} \in \text{argmax} \iff \mathbf{x} \in \partial g(\mathbf{w})) \]

\[ \bar{p} = \min_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{2}\|\mathbf{w}\|_2^2 + g(-A\mathbf{w} + \mathbf{b}) \quad (= P(\mathbf{w})) \]

\[ \bar{d} = \min_{\mathbf{x} \in \mathcal{X}} \frac{1}{2}\|A^T\mathbf{x}\|_2^2 - \langle \mathbf{x}, \mathbf{b} \rangle \quad (= D(\mathbf{x})) \]

A conditional gradient step in the dual:

\[ \mathbf{p}^k : \max_{\mathbf{y} \in \mathcal{X}} \langle AA^T\mathbf{x}^k - \mathbf{b}, \mathbf{x}^k - \mathbf{y} \rangle = \|A^T\mathbf{x}^k\|_2^2 - \langle \mathbf{b}, \mathbf{x}^k \rangle + g(-AA^T\mathbf{x}^k + \mathbf{b}) = P(A^T\mathbf{x}^k) + D(\mathbf{x}^k) \]

Consider the primal variable \( \mathbf{w}^k = A^T\mathbf{x}^k \): we have \( \mathbf{p}^k \in \partial g(-A\mathbf{w}^k + \mathbf{b}) \).

\[ \mathbf{w}^{k+1} - \mathbf{w}^k = \alpha^k A^T(-\mathbf{x}^k + \mathbf{p}^k) = -\alpha^k \partial P(\mathbf{w}^k) \]

Implicit subgradient steps in the primal!
The primal-dual interpretation holds in much more general settings (Bach 2015).

Primal-dual convergence analysis, \( \min_{i=1,\ldots,k} P(w^i) + D(x^i) = O(1/k) \) (Jaggi 2013).

Automatic step size tuning for subgradient descent in the primal.
Outline

1. Context
2. Conditional Gradient algorithm
3. CG and convex duality
4. Block CG and $L_2$ regularized ERM
5. Results
Consider a problem of the form:

$$\bar{p} = \min_{w \in \mathbb{R}^m} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{N} \sum_{i=1}^{N} g(-A_iw + b_i)$$

$$\bar{d} = \min_{x_i \in X, i=1, \ldots, N} \frac{\lambda}{2} \left\| \frac{1}{N\lambda} \sum_{i=1}^{N} A_i^T x_i \right\|_2^2 - \frac{1}{N} \sum_{i=1}^{N} \langle x_i, b_i \rangle$$
Consider a problem of the form:

$$\bar{p} = \min_{w \in \mathbb{R}^m} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{N} \sum_{i=1}^{N} g(-A_i w + b_i) \quad (= P(w))$$

$$\bar{d} = \min_{x_i \in X, i=1,\ldots,N} \frac{\lambda}{2} \left\| \frac{1}{N} \sum_{i=1}^{N} A_i^T x_i \right\|_2^2 - \frac{1}{N} \sum_{i=1}^{N} \langle x_i, b_i \rangle \quad (= D(x))$$

**Binary SVM:** dataset \((a_i, l_i) \in \mathbb{R}^m \times \{-1, 1\}, i = 1, \ldots, N\)

$$P(w) = \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - l_i a_i^T w)$$

- Prediction: \(l(a, w) = \arg\max_{l \in \{-1, 1\}} l a^T w = \text{sign}(a^T w)\).
- Convex surrogate of the empirical risk: \(\frac{1}{N} \sum_{i=1}^{N} 1(l(a_i, w) = l_i)\)
The dual has a separable block structure: \( \mathbf{x}_i \in \mathcal{X}, i = 1, \ldots, N \). Start with \( \mathbf{x}_i^0 \in \mathcal{X}, i = 1, \ldots, N \), and iterate for \( k \in \mathbb{N} \) and \( i = 1, \ldots, N \)

\[
\mathbf{p}_i^k \in \operatorname{argmax}_{y \in \mathcal{X}} \langle \nabla_{\mathbf{x}_i} D(\mathbf{x}_i^k), \mathbf{x}_i^k - y \rangle
\]

\[
\mathbf{x}_{i}^{k+1} = (1 - \alpha_i^k) \mathbf{x}_i^k + \alpha_i^k \mathbf{p}_i^k \\
0 \leq \alpha_i^k \leq 1
\]
The dual has a separable block structure: \( x_i \in X, i = 1, \ldots, N \). Start with \( x_0 \in X, i = 1, \ldots, N \), and iterate for \( k \in \mathbb{N} \) and \( i = 1, \ldots, N \)

\[
\begin{align*}
\mathbf{p}_i^k & \in \arg\max_{y \in X} \langle \nabla_{x_i} D(x^k), x_i^k - y \rangle \\
x_i^{k+1} & = (1 - \alpha_i^k) x_i^k + \alpha_i^k \mathbf{p}_i^k \quad 0 \leq \alpha_i^k \leq 1
\end{align*}
\]

Mainly three way to choose blocks:

- Uniformly at random (Lacoste-Julien et al. 2013).
- Cyclic (Beck et al. 2015).
- Essentially cyclic, “random permutation” (Beck et al. 2015).

Primal interpretation: a subgradient method (stochastic, cyclic, etc . . . ).

\[
\mathbf{p}_i^k \in \partial g (-A_i \mathbf{w}_i^k + \mathbf{b}_i)
\]
<table>
<thead>
<tr>
<th>a (1)</th>
<th>b (2)</th>
<th>u (3)</th>
<th>l (4)</th>
<th>o (5)</th>
<th>u (6)</th>
<th>s (7)</th>
<th>l (8)</th>
<th>y (9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>!</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>!</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>!</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>!</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>!</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>!</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>!</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>!</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>!</td>
</tr>
</tbody>
</table>
Structured output learning and structured SVM

Dataset: \((a_i, l_i) \in A \times L, i = 1, \ldots, N\). \(L\) is discrete and structured:

- Feature function: \( \phi: A \times L \rightarrow \mathbb{R}^m \)
- Prediction \( l(a, w) = \arg\max_{l \in L} \langle w, \phi(a, l) \rangle \)
- Risk function \( \Delta: L^2 \rightarrow \mathbb{R}_+ \).
Structured output learning and structured SVM

Dataset: \((a_i, l_i) \in A \times L, i = 1, \ldots, N\). \(L\) is discrete and structured:

- Feature function: \(\phi: A \times L \rightarrow \mathbb{R}^m\)
- Prediction \(l(a, w) = \arg\max_{l \in L} \langle w, \phi(a, l) \rangle\)
- Risk function \(\Delta: L^2 \rightarrow \mathbb{R}_+\).

Binary SVM:

- \(L = \{-1, 1\}\).
- \(\phi(a, l) = la\).
- \(\Delta\) is the 0−1 loss
- Prediction is a sign (optimize over a set of size 2)
- The dual constraint set is a box (product of segments).
Structured output learning and structured SVM

Dataset: \((a_i, l_i) \in \mathcal{A} \times \mathcal{L}, i = 1, \ldots, N\). \(\mathcal{L}\) is discrete and structured:

- Feature function: \(\phi: \mathcal{A} \times \mathcal{L} \to \mathbb{R}^m\)
- Prediction \(\hat{l}(a, w) = \arg\max_{l \in \mathcal{L}} \langle w, \phi(a, l) \rangle\)
- Risk function \(\Delta: \mathcal{L}^2 \to \mathbb{R}_+\).

Label sequence learning:

- \(\mathcal{L}\) is the set of possible words over an alphabet.
- \(\phi\) is inspired by HMM (unary and binary terms over a chain)
- \(\Delta\) is the Hamming distance.
- Prediction (or decoding) is done by dynamic programming (Viterbi algorithm).
Structured output learning and structured SVM

Dataset: \((a_i, l_i) \in A \times L, i = 1, \ldots, N\). \(L\) is discrete and structured:

- Feature function: \(\phi: A \times L \rightarrow \mathbb{R}^m\)
- Prediction \(l(a, w) = \arg\max_{l \in L} \langle w, \phi(a, l) \rangle\)
- Risk function \(\Delta: L^2 \rightarrow \mathbb{R}_+\).

Empirical risk: \(w \rightarrow \sum_{i=1}^{N} \Delta(l_i, l(a_i, w))\).

Label sequence learning:

- \(L\) is the set of possible words over an alphabet.
- \(\phi\) is inspired by HMM (unary and binary terms over a chain)
- \(\Delta\) is the Hamming distance.
- Prediction (or decoding) is done by dynamic programming (Viterbi algorithm).
Structured output learning and structured SVM

Dataset: \((a_i, l_i) \in A \times L, i = 1, \ldots, N\). \(L\) is discrete and structured:

- Feature function: \(\phi: A \times L \to \mathbb{R}^m\)
- Prediction \(l(a, w) = \arg \max_{l \in L} \langle w, \phi(a, l) \rangle\)
- Risk function \(\Delta: L^2 \to \mathbb{R}_+\).

Convex relaxation: \(w \to \sum_{i=1}^{N} \max_{l \in L} \{ \Delta(l_i, l) - \langle w, \phi(a_i, l) - \phi(a_i, l_i) \rangle \} \).

Label sequence learning:

- \(L\) is the set of possible words over an alphabet.
- \(\phi\) is inspired by HMM (unary and binary terms over a chain)
- \(\Delta\) is the Hamming distance.
- Prediction (or decoding) is done by dynamic programming (Viterbi algorithm).
Structured output learning and structured SVM

Dataset: \((a_i, l_i) \in A \times L, i = 1, \ldots, N\). \(L\) is discrete and structured:

- Feature function: \(\phi: A \times L \rightarrow \mathbb{R}^m\)
- Prediction \(l(a, w) = \arg\max_{l \in L} \langle w, \phi(a, l) \rangle\)
- Risk function \(\Delta: L^2 \rightarrow \mathbb{R}_+\).

Structured SVM:
\[
\min_{w \in \mathbb{R}^m} \frac{\lambda}{2} \|w\|^2 + \sum_{i=1}^{N} \max_{l \in L} \left\{ \Delta(l_i, l) - \langle w, \phi(a_i, l) - \phi(a_i, l_i) \rangle \right\}
\]

Label sequence learning:

- \(L\) is the set of possible words over an alphabet.
- \(\phi\) is inspired by HMM (unary and binary terms over a chain)
- \(\Delta\) is the Hamming distance.
- Prediction (or decoding) is done by dynamic programming (Viterbi algorithm).
- The dual constraint set is a product of simplices (of size \(|L|\)).
1. Context

2. Conditional Gradient algorithm

3. CG and convex duality

4. Block CG and $L_2$ regularized ERM

5. Results
Convergence rates

- \( \tilde{k} \): number of effective passes through the \( N \) blocks.
- The rates are given for the duality gap.
- \( B \): diameter of the dual constraint set \( X \times X \times \ldots \times X \).
- \( L \): Lipschitz modulus of \( \nabla D \).

Random block: the rate relates to an expectation (Lacoste-Julien et al. 2013).

Cyclic block: deterministic rate (Beck et al. 2015).

Approximate line search:

\[
O \left( \frac{1}{\tilde{k}} (L B^2 + D(x_0)) \right)
\]

Open loop (\( \alpha \tilde{k} = 2\tilde{k} + 2 \)):

\[
O \left( \frac{1}{\tilde{k}} LB^2 \sqrt{N} \right)
\]

where \( \beta \) is the smallest block Lipschitz modulus of \( \nabla D \) (variations constrained to a single block).
Convergence rates

- $\tilde{k}$: number of effective passes through the $N$ blocks.
- The rates are given for the duality gap.
- $B$: diameter of the dual constraint set $X \times X \times \ldots \times X$.
- $L$: Lipschitz modulus of $\nabla D$.

Random block: the rate relates to an expectation (Lacoste-Julien et al. 2013).

$$O \left( \frac{1}{\tilde{k}} (LB^2 + D(x^0)) \right)$$
Convergence rates

- \( \tilde{k} \): number of effective passes through the \( N \) blocks.
- The rates are given for the duality gap.
- \( B \): diameter of the dual constraint set \( X \times X \times \ldots \times X \).
- \( L \): Lipschitz modulus of \( \nabla D \).

**Random block:** the rate relates to an expectation (Lacoste-Julien et al. 2013).

\[
O \left( \frac{1}{\tilde{k}} (LB^2 + D(x^0)) \right)
\]

**Cyclic block:** deterministic rate (Beck et al. 2015).

- Approximate line search: \( O \left( \frac{1}{\tilde{k}} LB^2 \frac{L}{N} \right) \)
- Open loop \( \left( \alpha_i^{\tilde{k}} = \frac{2}{\tilde{k} + 2} \right) \): \( O \left( \frac{1}{\tilde{k}} LB^2 \sqrt{N} \right) \)

where \( \beta \) is the smallest block Lipschitz modulus of \( \nabla D \) (variations constrained to a single blocks).
Results on synthetic problems

1000 random QP over the unit cube in $\mathbb{R}^{100}$ (normalized).

- Predefined step
- Backtracking
- Exact line-search

$k$ $f$

type
- CBCG–P
- CBCG–C
- RBCG
- CG
Results on structural SVM

Handwritten words recognition.

![Graph showing different values of λ and their impact on gap](image)
Conclusion regarding cyclic block selection rule

- One of the few attempts to analyse essentially cyclic methods.
- Huge gap compared to random selection.
- Efficient in practice.

Future directions:

- Gap between theory and practice
- Linear convergence
- Exact line search, inexact oracles
General conclusion

- Nice duality between constraint block decomposition and sequential methods for sums.
- Conditional gradient is “bad”, but it is good in settings for which nothing else is affordable.