Block Conditional Gradient Algorithms

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

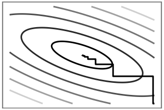
Séminaire MIAT INRA
September 23 2016

Context: large scale convex optimization

Two old ideas have received renewed attention in the past years:

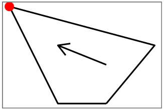
Block decomposition:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$$



Linear oracles:

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

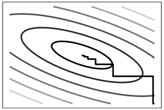


Context: large scale convex optimization

Two old ideas have received renewed attention in the past years:

Block decomposition:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$$

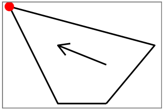


Coordinate descent:

- Large dimension
- Distributed data

Linear oracles:

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



Conditional gradient:

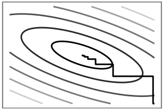
- "Complex constraints"
- Primal-dual interpretation

Context: large scale convex optimization

Two old ideas have received renewed attention in the past years:

Block decomposition:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$$

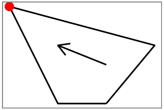


Coordinate descent:

- Large dimension
- Distributed data

Linear oracles:

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



Conditional gradient:

- "Complex constraints"
- Primal-dual interpretation

Theoretical properties and empirical performances?

Scope of the presentation

- Most results in the litterature 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).

Scope of the presentation

- Most results in the litterature 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, $\mathcal{C}_1$ with L-Lipschitz gradient over $X \subset \mathbb{R}^n$ which is convex and compact.

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

Main idea

Optimization setting: $f : \mathbb{R}^n \to \mathbb{R}$ is convex, $\mathcal{C}_1$ with L-Lipschitz gradient over $X \subset \mathbb{R}^n$ which is convex and compact.

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

Start with $\mathbf{x}^0 \in X$

$$\mathbf{p}^k \in \operatorname{argmax}_{\mathbf{y} \in X} \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{y} \rangle$$

$$\mathbf{x}^{k+1} = (1 - \alpha^k)\mathbf{x}^k + \alpha^k \mathbf{p}^k \qquad 0 \le \alpha^k \le 1$$

Main idea

Optimization setting: $f : \mathbb{R}^n \to \mathbb{R}$ is convex, $\mathcal{C}_1$ with L-Lipschitz gradient over $X \subset \mathbb{R}^n$ which is convex and compact.

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

Start with $\mathbf{x}^0 \in X$

$$\mathbf{p}^k \in \operatorname{argmax}_{\mathbf{y} \in X} \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{y} \rangle$$
$$\mathbf{x}^{k+1} = (1 - \alpha^k)\mathbf{x}^k + \alpha^k \mathbf{p}^k \qquad 0 \leq \alpha^k \leq 1$$

Step size:

- $\alpha^k = \frac{2}{k+2}$ *Open loop*
- $\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{y} \in [\mathbf{x}^k, \mathbf{p}^k]} f(\mathbf{y})$ *Exact line search*
- $\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{y} \in [\mathbf{x}^k, \mathbf{p}^k]} Q(\mathbf{x}^k, \mathbf{y})$ *Approximate line search*

$$f(\mathbf{y}) \leq Q(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$$

(tangent quadratic upper bound, descent Lemma).

Historical remarks

Fifty years ago:

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

Historical remarks

Fifty years ago:

- First appearance for quadratic programs (Frank, Wolfe, 1956).
- $f(\mathbf{x}^k) - \bar{f} = O(1/k)$ (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.

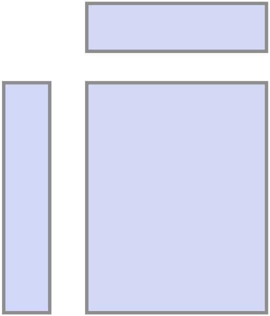
Why is it interesting?

- $O(1/k^2)$ 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

Convex duality

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

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

Convex duality

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

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

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

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

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

Convex duality

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

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

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

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

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

- 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

$$g : \mathbf{w} \to \max_{\mathbf{x} \in X} \langle \mathbf{x}, \mathbf{w} \rangle \qquad (\mathbf{x} \in \text{argmax} \Leftrightarrow \mathbf{x} \in \partial g(\mathbf{w}))$$

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

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

Primal subgradient and dual conditional gradient

$$g : \mathbf{w} \to \max_{\mathbf{x} \in X} \langle \mathbf{x}, \mathbf{w} \rangle \qquad (\mathbf{x} \in \operatorname{argmax} \Leftrightarrow \mathbf{x} \in \partial g(\mathbf{w}))$$

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

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

A conditional gradient step in the dual:

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

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

Primal subgradient and dual conditional gradient

$$g : \mathbf{w} \to \max_{\mathbf{x} \in X} \langle \mathbf{x}, \mathbf{w} \rangle \qquad (\mathbf{x} \in \operatorname{argmax} \Leftrightarrow \mathbf{x} \in \partial g(\mathbf{w}))$$

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

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

A conditional gradient step in the dual:

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

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

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

Primal subgradient and dual conditional gradient

$$g: \mathbf{w} \to \max_{\mathbf{x} \in X} \langle \mathbf{x}, \mathbf{w} \rangle \qquad (\mathbf{x} \in \operatorname{argmax} \Leftrightarrow \mathbf{x} \in \partial g(\mathbf{w}))$$

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

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

A conditional gradient step in the dual:

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

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!

Primal subgradient and dual conditional gradient

- The primal-dual interpretation holds in much more general settings (Bach 2015).
- Primal-dual convergence analysis, $\min_{i=1,\dots,k} P(\mathbf{w}^i) + D(\mathbf{x}^i) = O(1/k)$ (Jaggi 2013).
- Automatic step size tuning for subgradient descent in the primal.

Outline

$L2$ regularized ERM

Consider a problem of the form:

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

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

$L2$ regularized ERM

Consider a problem of the form:

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

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

Binary SVM: dataset $(\mathbf{a}_i, l_i) \in \mathbb{R}^m \times \{-1, 1\}, i = 1, \ldots, N$

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

- Prediction: $l(\mathbf{a}, \mathbf{w}) = \text{argmax}_{l\in\{-1,1\}} l\mathbf{a}^T\mathbf{w} = \text{sign}(\mathbf{a}^T\mathbf{w})$.
- Convex surrogate of the empirical risk: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(l(\mathbf{a}_i, \mathbf{w}) = l_i)$

$L2$ regularized ERM: dual block conditional gradient

The dual has a separable block structure: $\mathbf{x}_i \in X, i = 1, \ldots, N$. Start with $\mathbf{x}_i^0 \in X, i = 1, \ldots, N$, and iterate for $k \in \mathbb{N}$ and $i = 1, \ldots, N$

$$\begin{aligned} \mathbf{p}_i^k &\in \operatorname{argmax}_{\mathbf{y} \in X} \left\langle \nabla_{\mathbf{x}_i} D(\mathbf{x}^k), \mathbf{x}_i^k - \mathbf{y} \right\rangle \\ \mathbf{x}_i^{k+1} &= (1 - \alpha_i^k)\mathbf{x}_i^k + \alpha_i^k \mathbf{p}_i^k \qquad\qquad 0 \leq \alpha_i^k \leq 1 \end{aligned}$$

$L2$ regularized ERM: dual block conditional gradient

The dual has a separable block structure: $\mathbf{x}_i \in X, i = 1, \ldots, N$. Start with $\mathbf{x}_i^0 \in X, i = 1, \ldots, N$, and iterate for $k \in \mathbb{N}$ and $i = 1, \ldots, N$

$$
\begin{aligned}
\mathbf{p}_i^k &\in \operatorname{argmax}_{\mathbf{y} \in X} \left\langle \nabla_{\mathbf{x}_i} D(\mathbf{x}^k), \mathbf{x}_i^k - \mathbf{y} \right\rangle \\
\mathbf{x}_i^{k+1} &= (1 - \alpha_i^k)\mathbf{x}_i^k + \alpha_i^k \mathbf{p}_i^k \qquad\qquad 0 \leq \alpha_i^k \leq 1
\end{aligned}
$$

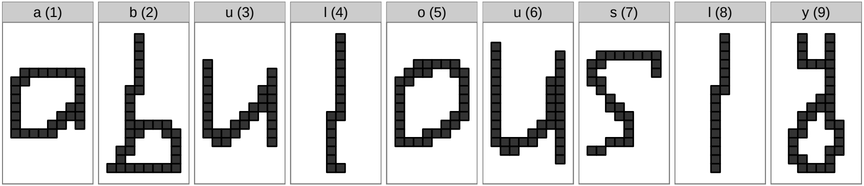
Mainly three way to choose blocks:

- 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 ...).

$$
\mathbf{p}_i^k \in \partial g \left(-A_i \mathbf{w}^k + \mathbf{b}_i \right)
$$

Structured output learning and structured SVM



Structured output learning and structured SVM

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

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

Structured output learning and structured SVM

Dataset: $(\mathbf{a}_i, l_i) \in \mathcal{A} \times \mathcal{L}, i = 1, \dots, N$. $\mathcal{L}$ is discrete and structured:

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

Binary SVM:

- $\mathcal{L} = \{-1, 1\}$.
- $\phi(\mathbf{a}, l) = l\mathbf{a}$.
- Δ 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: $(\mathbf{a}_i, l_i) \in \mathcal{A} \times \mathcal{L}, i = 1, \ldots, N$. $\mathcal{L}$ is discrete and structured:

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

Label sequence learning:

- $\mathcal{L}$ is the set of possible words over an alphabet.
- ϕ is inspired by HMM (unary and binary terms over a chain)
- Δ is the Hamming distance.
- Prediction (or decoding) is done by dynamic programming (Viterbi algorithm).

Structured output learning and structured SVM

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

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

Empirical risk: $\mathbf{w} \to \sum_{i=1}^{N} \Delta(l_i, l(\mathbf{a}_i, \mathbf{w}))$.

Label sequence learning:

- $\mathcal{L}$ is the set of possible words over an alphabet.
- ϕ is inspired by HMM (unary and binary terms over a chain)
- Δ is the Hamming distance.
- Prediction (or decoding) is done by dynamic programming (Viterbi algorithm).

Structured output learning and structured SVM

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

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

Convex relaxation: $$\mathbf{w} \to \sum_{i=1}^{N} \max_{l \in \mathcal{L}} \left\{ \Delta(l_i, l) - \langle \mathbf{w}, \phi(\mathbf{a}_i, l) - \phi(\mathbf{a}_i, l_i) \rangle \right\}.$$

Label sequence learning:

- $\mathcal{L}$ is the set of possible words over an alphabet.
- ϕ is inspired by HMM (unary and binary terms over a chain)
- Δ is the Hamming distance.
- Prediction (or decoding) is done by dynamic programming (Viterbi algorithm).

Structured output learning and structured SVM

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

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

Structured SVM: $$\min_{\mathbf{w} \in \mathbb{R}^m} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{N} \max_{l \in \mathcal{L}} \{\Delta(l_i, l) - \langle \mathbf{w}, \phi(\mathbf{a}_i, l) - \phi(\mathbf{a}_i, l_i) \rangle\}$$

Label sequence learning:

- $\mathcal{L}$ is the set of possible words over an alphabet.
- ϕ is inspired by HMM (unary and binary terms over a chain)
- Δ 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 $|\mathcal{L}|$).

Outline

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 ∇D.

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 ∇D.

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

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

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 ∇D.

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

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

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

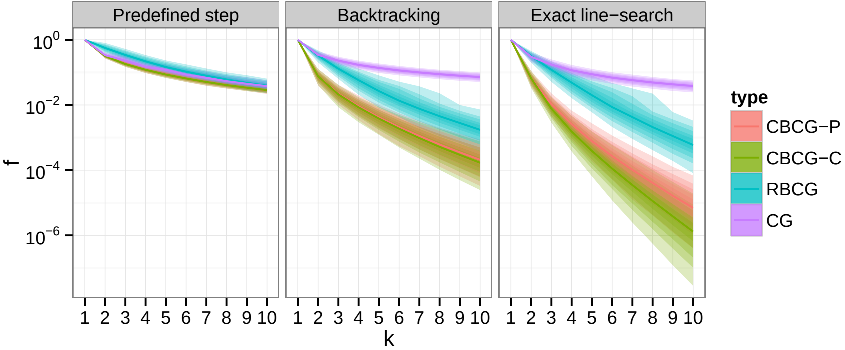
$$\text{Approximate line search} : O\left(\frac{1}{\tilde{k}} LB^2 N \frac{L}{\beta}\right)$$

$$\text{Open loop } \left(\alpha_i^{\tilde{k}} = \frac{2}{\tilde{k}+2}\right) : O\left(\frac{1}{\tilde{k}} LB^2 \sqrt{N}\right)$$

where β is the smallest block Lipschitz modulus of ∇D (variations constrained to a single blocks).

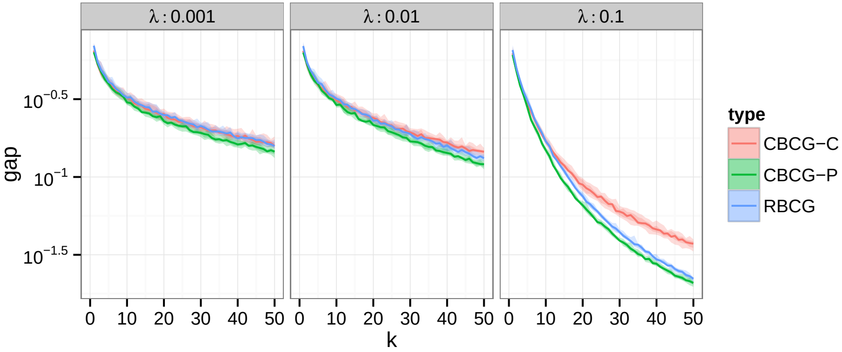
Results on synthetic problems

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



Results on structural SVM

Handwritten words recognition.



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.