Block-Coordinate Frank-Wolfe for Structural SVMs

Martin Jaggi [a] Simon Lacoste-Julien [b] Mark Schmidt [b] Patrick Pletscher [c]

[a] CMAP - ERC SIPA project, École Polytechnique, CNRS UMR 7641, Paris, France
[b] INRIA - SIERRA project-team, École Normale Supérieure, CNRS UMR 8548, Paris, France
[c] Machine Learning Laboratory, ETH Zurich, Switzerland

Short Summary

Motivation

Despite their wider applicability, optimization of *structural* SVMs remains challenging.

Contributions

New ***block-coordinate*** variant of the classic ***Frank-Wolfe algorithm***

(for convex optim. with block-separable constraints)

Giving a new simple ***online*** algorithm for structural SVMs, with primal-dual convergence rate, outperforming existing solvers in practice

Advantages

- The ***optimal step-size*** can be computed in closed-form (no parameter tuning)
- ***Duality gap*** guarantee, (e.g. as a stopping criterion)
- Allows use of ***approximate maximization oracles*** (weakest / most general oracle)

Frank-Wolfe *(or conditional gradient)*

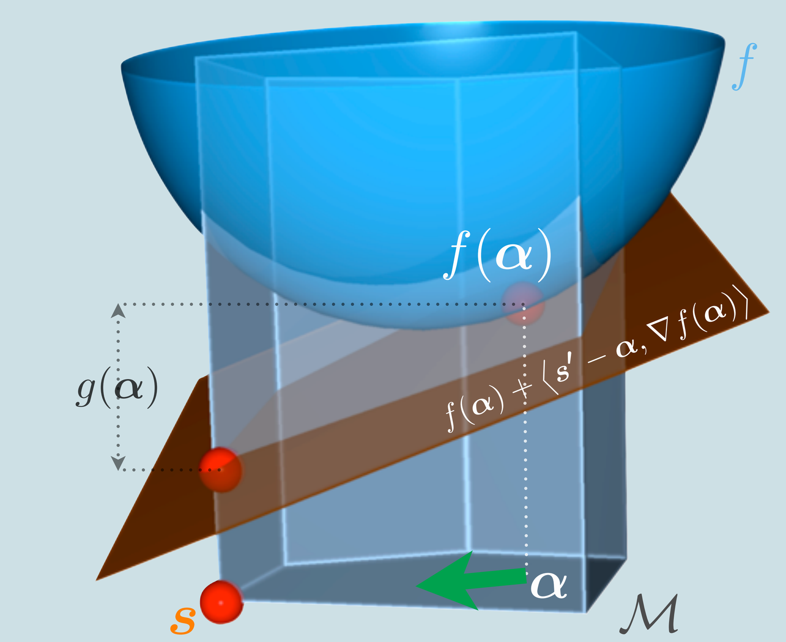
Constrained Convex Optimization over a compact domain

$$\min_{\alpha \in \mathcal{M}} f(\alpha)$$

Algorithm 1 Frank-Wolfe

Let $\alpha^{(0)} \in \mathcal{M}$
for $k = 0 \ldots K$ **do**
  Compute $s := \underset{s' \in \mathcal{M}}{\operatorname{argmin}} \left\langle s', \nabla f(\alpha^{(k)}) \right\rangle$
  Let $\gamma := \frac{2}{k+2}$, or find the optimal γ
  Update $\alpha^{(k+1)} := (1-\gamma)\alpha^{(k)} + \gamma s$
end for

Idea: Minimize a *linear approximation*



Convergence:
Error $\leq \frac{2C_f}{k+2}$
after k steps.

(also in **duality gap**, and with **inexact subproblems**)

Duality Gap
$g(\alpha)$ = efficient certificate for approximation quality

Sparse Iterates!

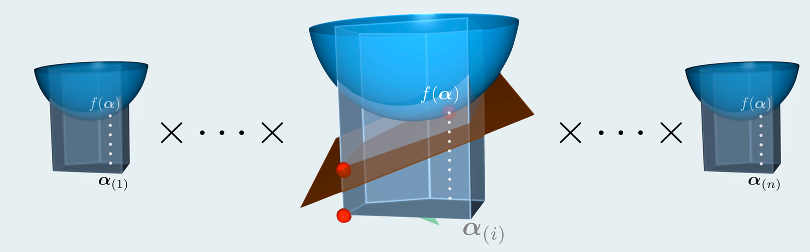
Constant bounded by the Lipschitz constant L_f of the gradient, $C_f \leq L_f \, diam(\mathcal{M})^2$

Block-Coordinate Frank-Wolfe

Problem: Minimize a convex function over block-separable compact constraints

$$\min_{\substack{\alpha \in \mathcal{M}^{(1)} \times \ldots \times \mathcal{M}^{(n)} \\ \mathcal{M}^{(1)} \subset \mathbb{R}^{m_1}, \ \ldots, \ \mathcal{M}^{(n)} \subset \mathbb{R}^{m_n}}} f(\alpha) \qquad \alpha = (\alpha_{(1)}, \ldots, \alpha_{(n)})$$

Idea: Combine Coordinate Descent with cheaper Frank-Wolfe steps



(pick one single block at random, and perform a Frank-Wolfe step affecting only this block)

Algorithm 3 Block-Coordinate Frank-Wolfe **new**

Let $\alpha^{(0)} \in \mathcal{M} = \mathcal{M}^{(1)} \times \ldots \times \mathcal{M}^{(n)}$
for $k = 0 \ldots K$ **do**
  Pick $i \in_{u.a.r.} [n]$
  Find $s_{(i)} := \underset{s'_{(i)} \in \mathcal{M}^{(i)}}{\operatorname{argmin}} \left\langle s'_{(i)}, \nabla_{(i)} f(\alpha^{(k)}) \right\rangle$
  Let $\gamma := \frac{2n}{k+2n}$, or find the optimal γ
  Update $\alpha^{(k+1)}_{(i)} := \alpha^{(k)}_{(i)} + \gamma (s_{(i)} - \alpha^{(k)}_{(i)})$
end for

Convergence:
Error $\leq \frac{2nC_f^{prod}}{k+2n}$
after k steps.

(also in **duality gap**, and with **inexact subproblems**)

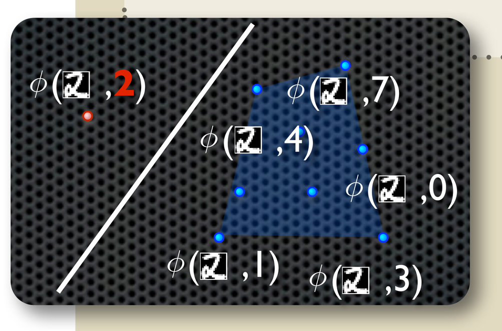
The constant C_f^{prod} can be *much* smaller than C_f. (For structural SVM, $nC_f^{prod} \approx C_f$)

Structural SVM

Structured Prediction

Goal: Given a joint "structured" feature map $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, construct a good linear classifier of the form

$$h_w(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \langle w, \phi(x, y) \rangle$$



Large margin separation

Maximization oracle *(loss augmented decoding)*

Primal

$$\min_w \quad \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \underbrace{\max_{y \in \mathcal{Y}} \Big\{ L(y_i, y) - \big\langle w, \underbrace{\phi(x_i, y_i) - \phi(x_i, y)}_{=: \psi_i(y)} \big\rangle \Big\}}_{= \text{ structured hinge loss}}$$

primal-dual correspondence $w = A\alpha$

Dual

$$\min_{\alpha \in \mathbb{R}^{n \cdot |\mathcal{Y}|}} \quad f(\alpha) := \frac{\lambda}{2}\|A\alpha\|^2 - b^T\alpha$$
$$\text{s.t.} \quad \sum_{y \in \mathcal{Y}} \alpha_i(y) = 1 \quad \forall i \in [n]$$
$$\text{and} \quad \alpha_i(y) \geq 0 \quad \forall i \in [n], \forall y \in \mathcal{Y}$$

block-structure!

Challenge: exponential # of variables

$A := \left\{ \frac{1}{\lambda n} \psi_i(y) \in \mathbb{R}^d \;\middle|\; i \in [n], y \in \mathcal{Y} \right\}$ $\qquad b := \left(\frac{1}{n} L_i(y) \right)_{i \in [n], y \in \mathcal{Y}}$

Optimization of the Structural SVM Dual

Key Insight: Frank-Wolfe step = Maximization oracle

Batch Frank-Wolfe:

Duality gap $\leq \varepsilon$ after $O\left(\frac{R^2}{\lambda\varepsilon}\right)$ iterations
(iteration cost: ***n*** *oracle calls)*

Block-Coordinate Frank-Wolfe:

Duality gap $\leq \varepsilon$ after $O\left(\frac{R^2}{\lambda\varepsilon}\right)$ iter.
(iteration cost: ***one*** *oracle call)*

Relation with Batch Subgradient

Can interpret batch subgradient (in the primal) as classic Frank-Wolfe (in the dual)

Relation with Cutting Plane

Can interpret cutting plane (SVM[struct], bundle methods) as a Frank-Wolfe variant, giving a simpler convergence proof

Relation with Stochastic Subgradient (SGD)

Same cheap iteration cost, but we have stronger primal-dual guarantees, more robustness, no step-size tuning, and faster in experiments

Algorithm 4 BCFW for Structural SVM

Let $w^{(0)} := w_i^{(0)} := 0$
for $k = 0 \ldots K$ **do**
  Pick $i \in_{u.a.r.} [n]$
  Solve $y_i^* := \underset{y \in \mathcal{Y}_i}{\operatorname{argmax}} \; H_i(y; w^{(k)})$
  Let $w_s := \frac{1}{\lambda n} \psi_i(y_i^*)$
  Let $\gamma := \frac{2n}{k+2n}$, or find the optimal γ
  Update $w_i^{(k+1)} := (1-\gamma) w_i^{(k)} + \gamma w_s$
  Update $w^{(k+1)} := w^{(k)} + w_i^{(k+1)} - w_i^{(k)}$
end for

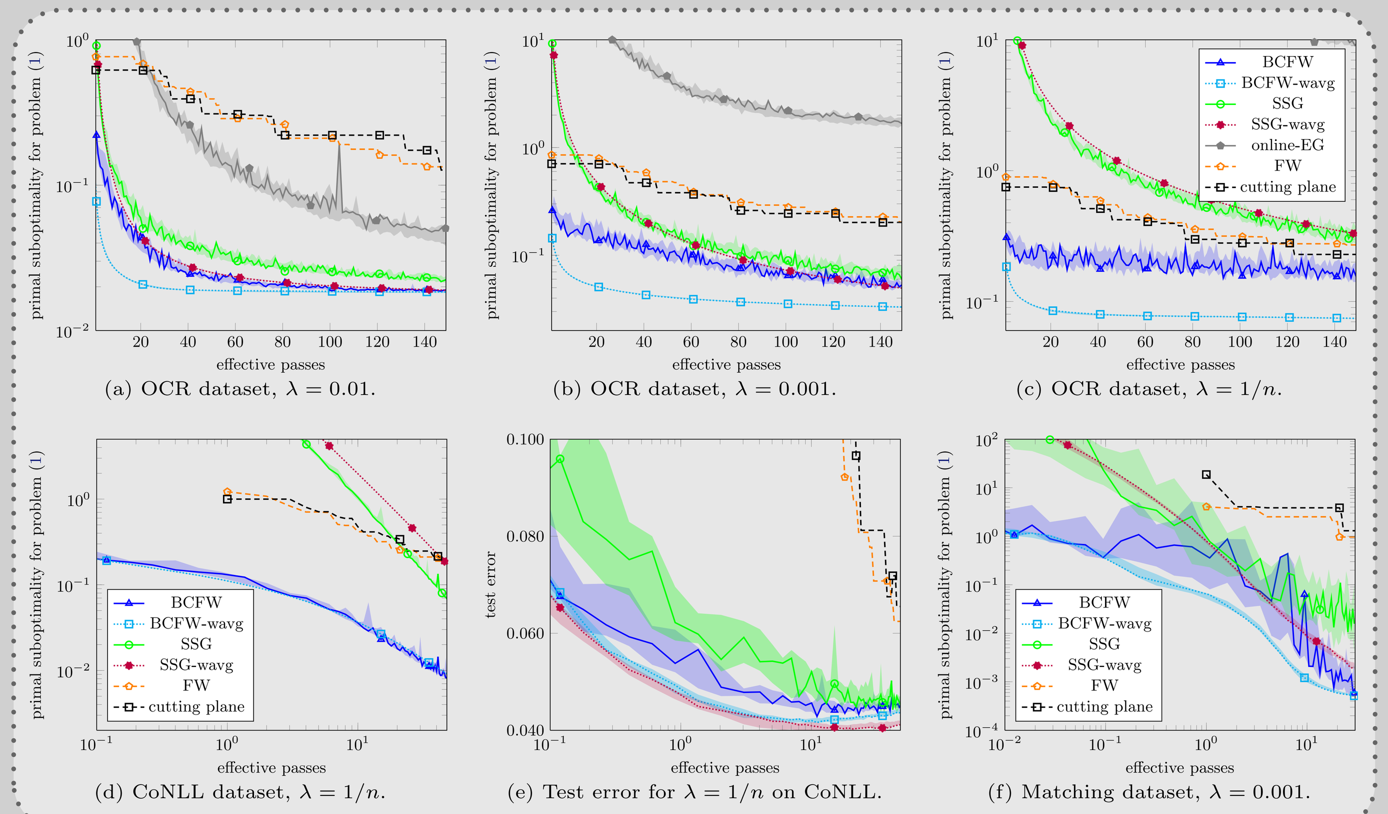
Related Work

Table 1. Convergence rates given in the *number of calls to the oracles* for different optimization algorithms for the structural SVM objective (1) in the case of a Markov random field structure, to reach a specific accuracy ε measured for different types of gaps, in term of the number of training examples n, regularization parameter λ, size of the label space $|\mathcal{Y}|$, maximum feature norm $R := \max_{i,y} \|\psi_i(y)\|_2$ (some minor terms were ignored for succinctness). Table inspired from (Zhang et al., 2011). Notice that only stochastic subgradient and our proposed algorithm have rates independent of n.

Optimization algorithm	Online	Primal/Dual	Type of guarantee	Oracle type	# Oracle calls
dual extragradient (Taskar et al., 2006)	no	primal-"dual"	saddle point gap	Bregman projection	$O\left(\frac{nR\log\lvert\mathcal{Y}\rvert}{\lambda\varepsilon}\right)$
online exponentiated gradient (Collins et al., 2008)	yes	dual	expected dual error	expectation	$O\left(\frac{(n+\log\lvert\mathcal{Y}\rvert)R^2}{\lambda\varepsilon}\right)$
excessive gap reduction (Zhang et al., 2011)	no	primal-dual	duality gap	expectation	$O\left(nR\sqrt{\frac{\log\lvert\mathcal{Y}\rvert}{\lambda\varepsilon}}\right)$
BMRM (Teo et al., 2010)	no	primal	≥primal error	maximization	$O\left(\frac{nR^2}{\lambda\varepsilon}\right)$
1-slack SVM-Struct (Joachims et al., 2009)	no	primal-dual	duality gap	maximization	$O\left(\frac{nR^2}{\lambda\varepsilon}\right)$
stochastic subgradient (Shalev-Shwartz et al., 2010)	yes	primal	primal error w.h.p.	maximization	$\tilde{O}\left(\frac{R^2}{\lambda\varepsilon}\right)$
this paper: stochastic block-coordinate Frank-Wolfe	yes	primal-dual	expected duality gap	maximization	$O\left(\frac{R^2}{\lambda\varepsilon}\right)$ Thm. 3

Experimental Results

dataset		n	d
OCR	sequence labeling	6251	4028
CoNLL	POS sequence labeling	8936	1643026
Matching	word alignment	5000	82



(a) OCR dataset, $\lambda = 0.01$. (b) OCR dataset, $\lambda = 0.001$. (c) OCR dataset, $\lambda = 1/n$.
(d) CoNLL dataset, $\lambda = 1/n$. (e) Test error for $\lambda = 1/n$ on CoNLL. (f) Matching dataset, $\lambda = 0.001$.

comparing block-coordinate Frank-Wolfe (BCFW) to stochastic subgradient (SSG), online exponentiated gradient (EG), batch Frank-Wolfe (FW) and cutting plane (wavg = weighted averaging of the iterates)