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 opt. with block-separable constraints)

Frank-Wolfe (or conditional gradient)

Constrained Convex Optimization over a compact domain

Algorithm 1 Frank-Wolfe:

\[ \min_{\alpha \in \mathcal{M}} f(\alpha) \]

\[ \text{for } i = 1, \ldots, K \text{ do} \]

\[ \gamma = \frac{1}{\|w_i\|^2} \text{ or find the optimal } \gamma \]

\[ \text{Update } \alpha^{(k+1)} = (1 - \gamma)\alpha^{(k)} + \gamma y_i \text{ for } y_i \text{ s.t. } \alpha_i^{(k)} = 0 \]

Duality Gap
\[ g(\alpha) = \text{efficient certificate for approximation quality} \]

Sparse Iterates!

Optimization of the Structural SVM Dual

Key Insight:
Frank-Wolfe step = Maximization oracle

Batch Frank-Wolfe:
Duality gap \( \leq \) after \( O\left(\frac{\lambda^2}{\varepsilon^2}\right) \) iterations (iteration cost: \( n \) oracle calls)

Relaxation with Batch Subgradient
Can interpret batch subgradient (in the primal) as Frank-Wolfe (in the dual)

Relation with Cutting Plane
Can interpret cutting plane (SVM++, bundle methods) as a Frank-Wolfe variant, giving a simpler convergence proof

Block-Coordinate Frank-Wolfe:
Duality gap \( \leq \) after \( O\left(\frac{\lambda^2}{\varepsilon^2}\right) \) iter. (iteration cost: one oracle call)

Relaxation 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

Related Work

- Online-Dual Frank-Wolfe for Structured SVMs
- Stochastic Frank-Wolfe
- Block-Coordinate Frank-Wolfe

Algorithm 4 OXFW for Structured SVMs

Let \( \alpha^{(0)} \in \mathcal{M} \) and \( \lambda^2 \leq \lambda_0^2 \)

\[ \text{for } k = 0, 1, 2, \ldots \text{ do} \]

\[ \text{Pick } \epsilon, \alpha \text{ s.t. } f(\alpha(\epsilon, \alpha)) \leq f(\alpha) \]

\[ \text{Solve } y \text{ s.t. } R(y) = \min \{ R(y) | f(\alpha(y)) \leq f(\alpha) \} \]

\[ \text{Update } \gamma = \frac{1}{\|w_i\|^2} \text{ or find the optimal } \gamma \]

\[ \text{Update } \alpha^{(k+1)} = (1 - \gamma)\alpha^{(k)} + \gamma y_i \text{ for } y_i \text{ s.t. } \alpha_i^{(k)} = 0 \]

Convergence:
Error \( \leq \frac{\lambda^2}{\varepsilon^2} \) after \( k \) steps

Duality gap \( g(\alpha) = \text{efficient certificate for approximation quality} \)

Sparse Iterates!