Dissolve^{struct} - A Library for Distributed Structured Prediction

Tribhuvanesh Orekondy, Martin Jaggi, Aurelien Lucchi

ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Data Analytics Lab
Machine Learning Institute

Short Summary

Motivation: *Structured* SVMs have wide applicability, but distributed optimization in machine learning is challenging.

Contributions

New ***open source*** library for training large scale structured SVMs. *(Available on github soon)*

- Same **simple interface** as SVM^{struct}
- allows **approximate inference**
- scales to very large systems and **commodity clusters**, Amazon EC2 and others.
- automatic **fault tolerance** by running on Spark

Interface: *(same as SVM^{struct})*

User provides 3 functions:

- *Feature Map (x,y)*
- *Loss function (y,y')*
- *(approx.) Max Oracle*

Distributed and data-parallel execution is handled automatically

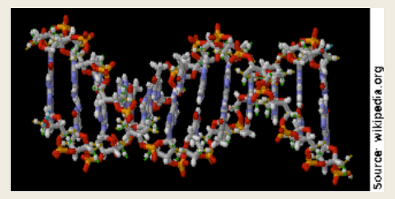
Many Applications

Text

- Parsing
- POS tagging, chunking
- sentence alignment
- named entity recognition

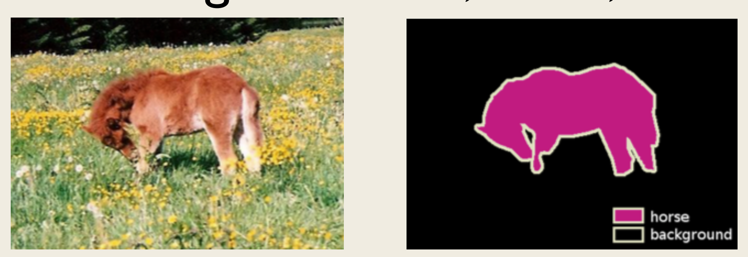
Biology

Protein structure & function prediction



Vision

Horse Segmentation, OCR, …



more?

- Scene understanding
- object localization & recognit.

We want your application!

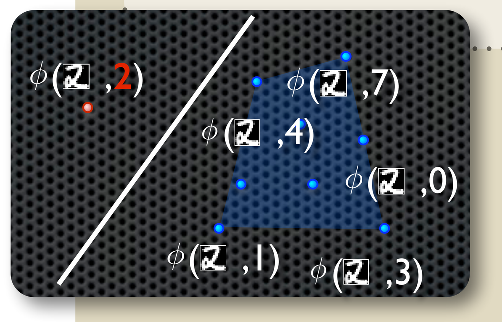
- 3D? max-cut? GPU decoding?

Structured 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

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



Large margin separation

Primal

maximization oracle *(loss augmented decoding)*

$$\min_{\boldsymbol{w}} \ \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \underbrace{\max_{\boldsymbol{y}\in\mathcal{Y}} \left\{ L(\boldsymbol{y}_i, \boldsymbol{y}) - \langle \boldsymbol{w}, \ \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i) - \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}) \rangle \right\}}_{=\text{ structured hinge loss } =: \psi_i(\boldsymbol{y})}$$

primal-dual correspondence $w = A\alpha$

Dual

$$\min_{\alpha\in\mathbb{R}^{n\cdot|\mathcal{Y}|}} \ f(\alpha) := \frac{\lambda}{2}\|A\alpha\|^2 - \boldsymbol{b}^T\alpha$$

$$\text{s.t.} \quad \textstyle\sum_{\boldsymbol{y}\in\mathcal{Y}} \alpha_i(\boldsymbol{y}) = 1 \quad \forall i \in [n]$$

$$\text{and} \quad \alpha_i(\boldsymbol{y}) \geq 0 \quad \forall i \in [n], \forall \boldsymbol{y} \in \mathcal{Y}$$

block-structure!

Challenge: exponential # of variables

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

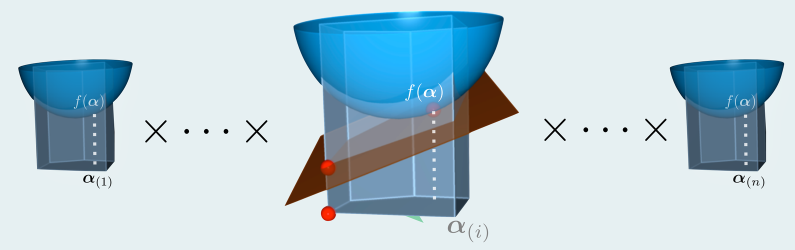
Block-Coordinate Frank-Wolfe

[ICML 2013]

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

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

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 4 BCFW for Structural SVM

new

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

iteration cost: ***one** oracle call*

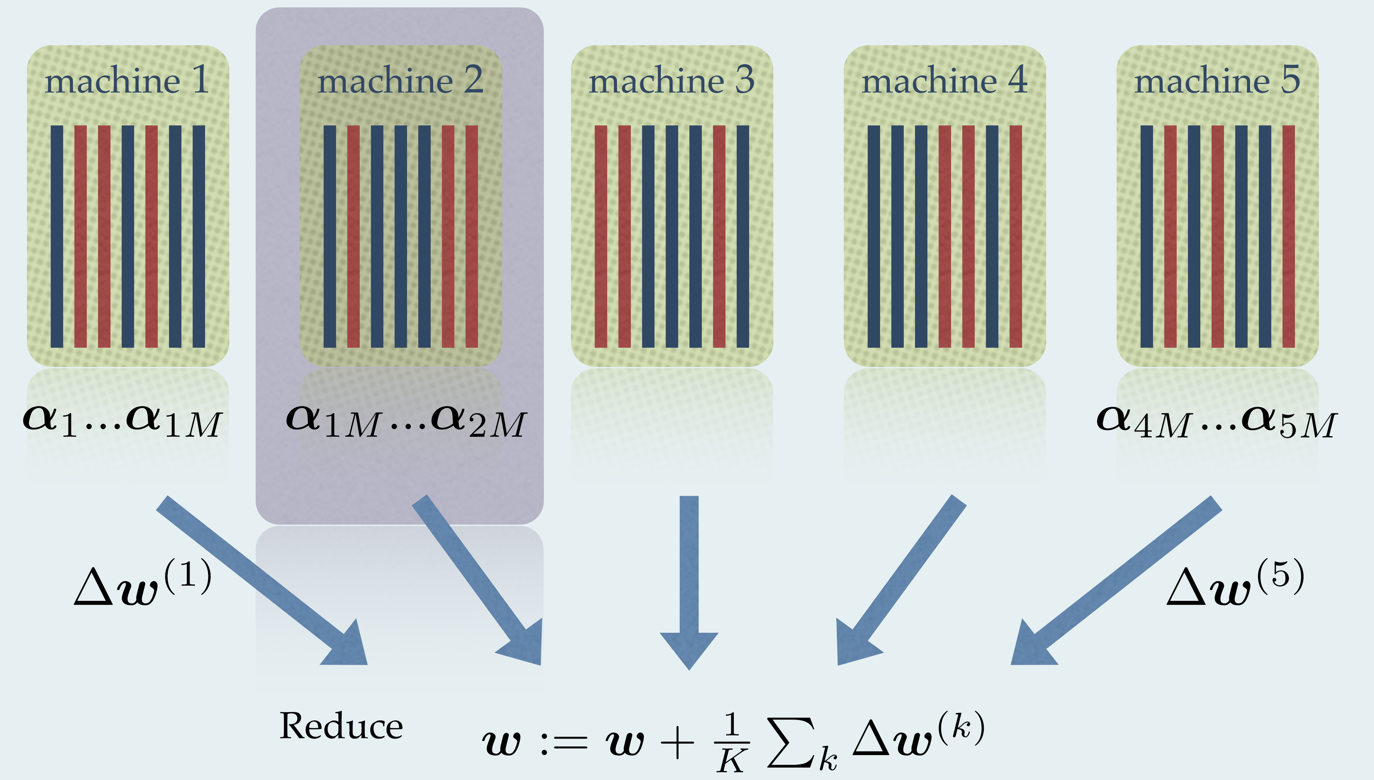
Error $\leq \varepsilon$ after $O\left(\frac{R^2}{\lambda\varepsilon}\right)$ iterations

Approximate Oracles *allowed!*

Communication Efficient Distributed Coordinate Ascent

CoCoA [NIPS 2014]

Fixing the **communication bottleneck** for distributed optimization in (generalized linear) machine learning.



Applications: State of the art in theory and practice, for SVM, Logistic Reg., Ridge Regression, Lasso / Sparse Reg.

Approximate Inference

inexact max-decoding oracles

Algorithm automatically detects if the oracle answer was accurate enough

Implemented Oracles so far:

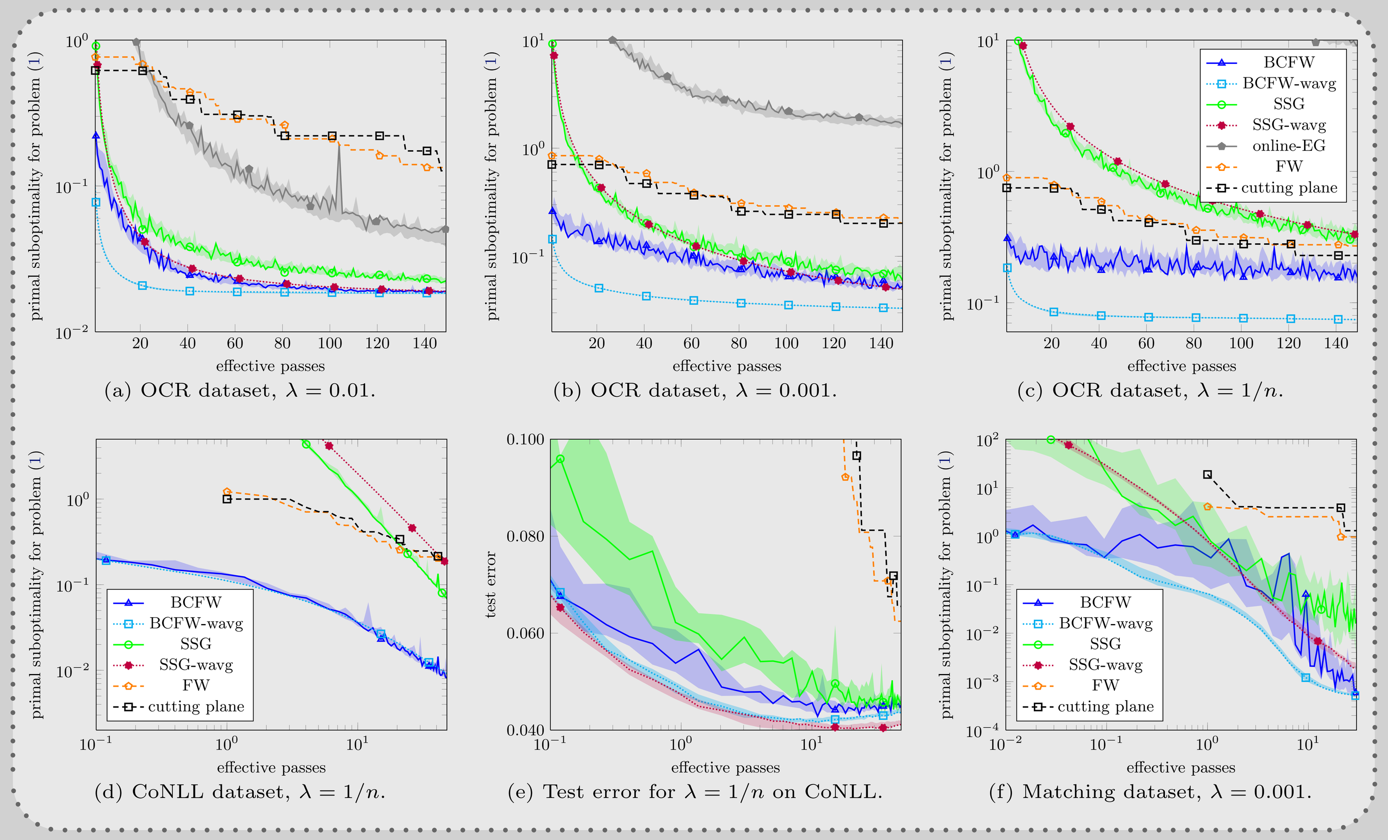
- Chain (Viterbi)
- Belief Propagation (using *Factorie* library)

Scala FACTORIE

add more oracles? need your help!

Single Machine Performance

dataset		n	d
OCR	sequence labeling	6251	4082
CoNLL	NLP chunking	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)