A Smoother Way To Train Structured Prediction Models

Krishna Pillutla, Vincent Roulet, Sham Kakade, Zaid Harchaoui
University of Washington

Overview

- Training structured prediction models is a **non-smooth optimization** problem involving **inference oracles**.
- We break the non-smoothness barrier for fast optimization with **smooth inference oracles** and **accelerated incremental algorithms**

Structured prediction

Structured outputs: such as chain of tags in named entity recognition

Spain's Nadal tops ATP ranking after French Open victory.

LOC PER x ORG x x MISC MISC x

Score function: $\phi(\cdot, \cdot; w)$ measures compatibility of output y for input x through

$$\phi(x, y; w) = \begin{cases} \text{high if } y \text{ is a good labeling for } x \\ \text{low if } y \text{ is a poor labeling for } x \end{cases}$$

Inference: finding best output

$$y^*(x; w) \in \operatorname*{argmax}_{y \in \mathcal{Y}} \phi(x, y; w)$$

Given by combinatorial algorithms, e.g., dynamic programming, graph cut/matching

Training: Find optimal w for $\phi(\cdot, \cdot; w)$, s.t. inference $y^*(x; w)$ is correct

Given a loss ℓ, use surrogate max-margin loss defined for input-output (x_i, y_i) as

$$f_i(w) = \max_{y' \in \mathcal{Y}} \psi_i(y'; w)$$

where $\quad \psi_i(y'; w) = \phi(x_i, y'; w) + \ell(y_i, y') - \phi(x_i, y_i; w)$

Optimization problem is

$$\min_{w \in \mathbb{R}^d} \left[F(w) = \frac{1}{n} \sum_{i=1}^n f_i(w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

where one obtains $v \in \partial f_i(w)$ by calling inference oracle

$$\operatorname*{argmax}_{y' \in \mathcal{Y}} \psi_i(y', w)$$

Smoothing

Composite objective: Rewrite $f_i = h \circ g_i$ where

$$h(z) = \max_{j \in \{1, \ldots, |\mathcal{Y}|\}} z_j, \quad \text{and} \quad g_i(w) = \big(\psi_i(y', w)\big)_{y' \in \mathcal{Y}}$$

Smoothing: Smooth max function $h(z) = \max_{\Delta^{|\mathcal{Y}|}} \langle u, z \rangle$ as

$$h_{\mu\omega}(z) = \max_{u \in \Delta^{|\mathcal{Y}|}} \{\langle z, u \rangle - \mu\omega(u)\}$$

where $\Delta^{|\mathcal{Y}|}$ is the simplex, $\mu > 0$, and ω is strongly convex

Smoothing type	$\omega(u)$	Smoothing computation
entropy	$H(u) = \langle u, \log u \rangle$	log-sum-exp
ℓ_2^2	$\ell_2^2(u) = \frac{1}{2}\|u\|_2^2$	projection on simplex

Approximates f_i to $O(\mu)$ by smooth max-margin loss

$$f_{i,\mu\omega} = h_{\mu\omega} \circ g_i$$

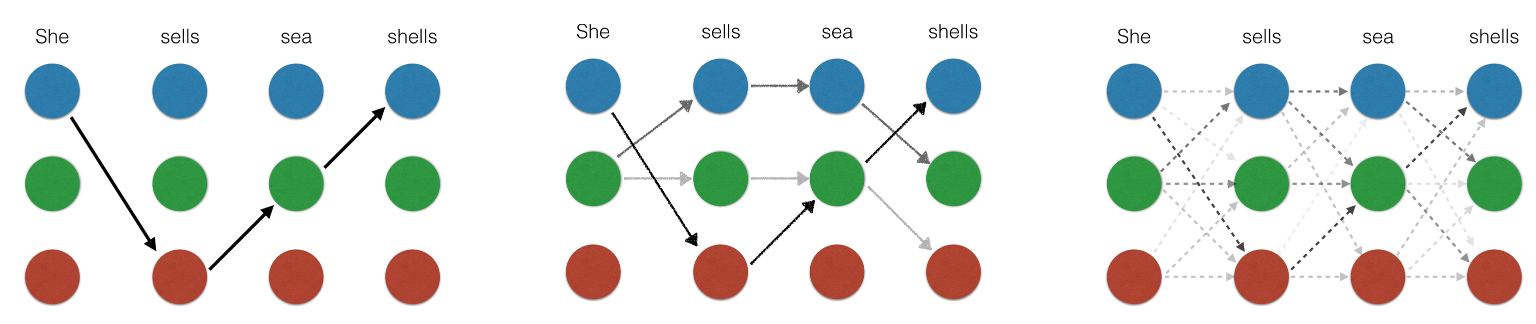
Smooth inference oracles

Max oracle: First order information on f

Top-K oracle: First order information on approximation of ℓ_2^2 smoothing of f

Exp oracle: First order information on entropy smoothing of f

Illustration on a chain graph



Non-smooth $\qquad$ ℓ_2^2 smoothing $\qquad$ Entropy smoothing

Computational complexity: in terms of $\mathcal{T}$: cost of max oracle & p: size of y

Max oracle	Top-K oracle		Exp oracle	
Algo	Algo	Time	Algo	Time
Max-product	Top-K max-product	$K\mathcal{T}\log K$	Sum-product	$\mathcal{T}$
Graph cut	BMMF	$pK\mathcal{T}$	Intractable	
Graph matching	BMMF	$K\mathcal{T}$	Intractable	
Branch and Bound	Top-K search	N/A	Intractable	

Here, BMMF is the Best Max-Marginal First algorithm (Yanover & Weiss 2003)

Convex structured prediction

Suppose score given by a predefined feature mapping $\Phi(x, y)$, such that

$$\phi(x, y; w) = \Phi(x, y)^\top w$$

is linear in w and the training problem is convex

Idea: Consider smoothed, regularized objectives

$$F_{\mu,\kappa}(w; z) = \frac{1}{n} \sum_{i=1}^n f_{i,\mu\omega}(w) + \frac{\lambda}{2}\|w\|_2^2 + \frac{\kappa}{2}\|w - z\|_2^2$$

centered on given z, solved by linearly convergent incremental method $\mathcal{M}$

Algorithm: Starting from $w_0 = z_0$, at each step k,

- Solve approximately using $\mathcal{M}$

$$w_{k+1} \approx \operatorname*{argmin}_w F_{\mu_k, \kappa_k}(w; z_k)$$

- Acceleration by extrapolation

$$z_{k+1} = w_k + \beta_k(w_{k+1} - w_k)$$

Convergence: Guaranteed to get approximate solution $F(w_k) - F^* \leq \epsilon$ after

$$\mathbb{E}(N) = \begin{cases} O\left(n + \sqrt{\frac{n}{\lambda\epsilon}}\right), & \text{if fixed smoothing} \\ O\left(n + \frac{1}{\lambda\epsilon}\right), & \text{if adaptive smoothing} \end{cases} \quad \text{iterations}$$

Deep structured prediction

Suppose score given by a learned feature mapping $\Phi(x, y; w_0)$, such that

$$\phi(x, y; w) = \Phi(x, y; w_0)^\top w_1$$

is non-linear in $w = (w_0, w_1)$ and the training problem is non-convex

Idea: Consider linear approximation of ψ_i around z as

$$\psi_i(y; w; z) = \psi_i(y; z) + \nabla_z \psi_i(y; z)(w - z) \quad \text{and} \quad f_i(w; z) = \max_{y \in \mathcal{Y}} \psi_i(y; w; z)$$

to get a regularized convex model

$$F_\gamma(w; z) = \frac{1}{n} \sum_{i=1}^n f_i(w; z) + \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{2\gamma}\|w - z\|_2^2$$

Algorithm: At each step k use convex solver to approximately solve at ϵ_k accuracy

$$w_{k+1} \approx \operatorname*{argmin}_w F_\gamma(w; w_k)$$

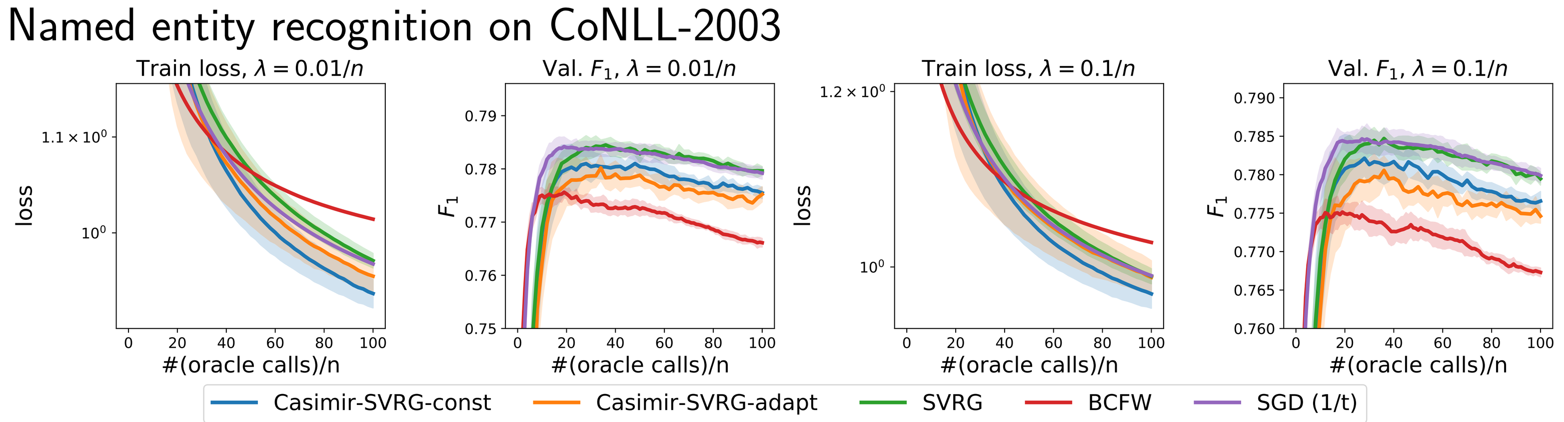
Convergence: Guaranteed to get a ϵ-near stationary point after

$$\mathbb{E}(N) = O\left(\frac{n}{\epsilon^2} + \frac{\sqrt{n}}{\epsilon^3}\right) \quad \text{iterations}$$

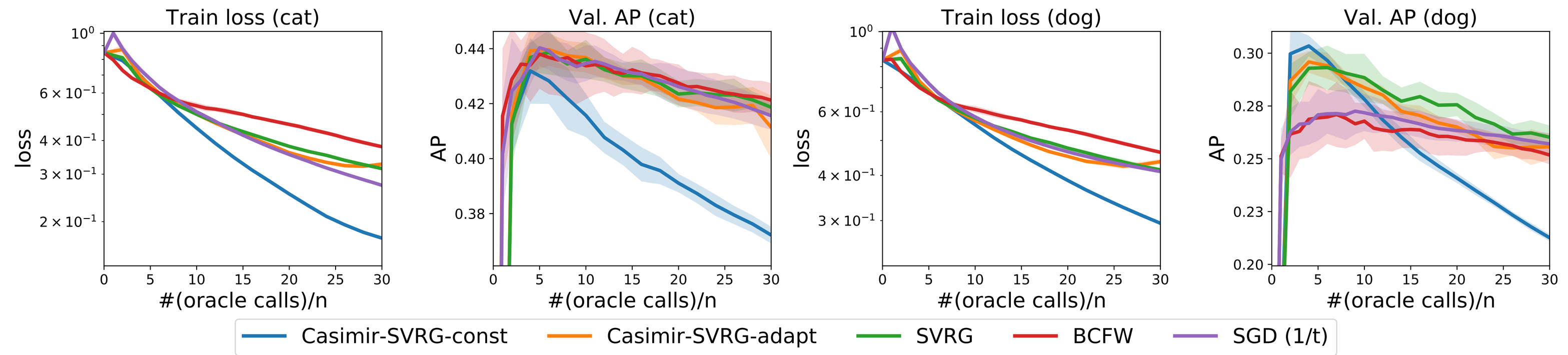
Numerical Experiments

Pre-defined feature map: convex problem, using structural SVMs

Named entity recognition on CoNLL-2003

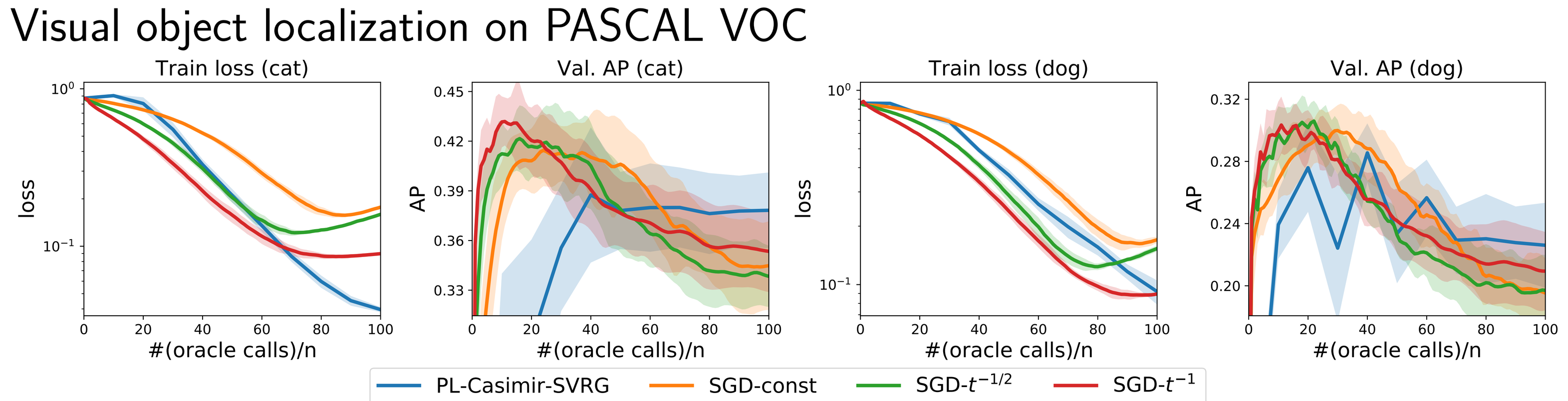


Visual object localization on PASCAL VOC



Learned feature map: non-convex problem, using convolutional neural networks

Visual object localization on PASCAL VOC



Casimir: `github.com/krishnap25/casimir`

References

Beck, A. and Teboulle, M. [2012], 'Smoothing and first order methods: A unified framework', *SIAM Journal on Optimization* **22**(2), 557–580

Lin, H., Mairal, J. and Harchaoui, Z. [2018], 'Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice', *Journal of Machine Learning Research* **18**(212), 1–54.