# Adaptive Stochastic Dual Coordinate Ascent for Conditional Random Fields


Rémi Le Priol  Alexandre Piché  Simon Lacoste-Julien

Montreal Institute for Learning Algorithms, Université de Montréal


## Overview

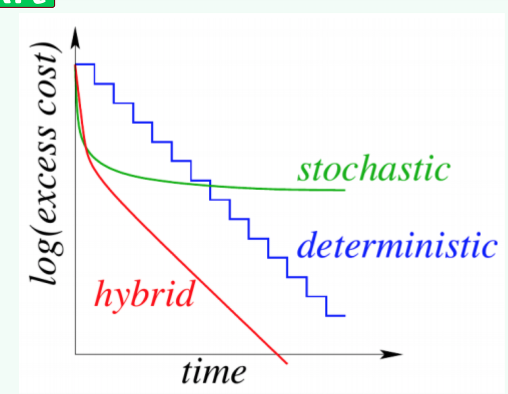
### Goal

Fast and exact optimization of Conditional Random Fields

### State of the Art

Variance reduced methods (hybrid):

- Stochastic Average Gradient (SAG) [3]
- Online Exponentiated Gradient (OEG) [1]
- Stochastic Dual Coordinate Ascent (SDCA) [this work]

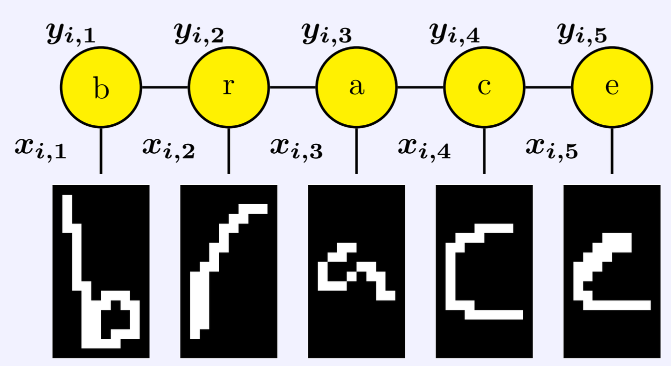



### Contributions

- Adapt the algorithm SDCA to the CRF setting.
- Accelerate SDCA with an adaptive sampling strategy (with proof).
- Get state of the art optimization speed on sparse datasets.

## Background

### Structured Prediction




a) OCR

| sample x | The | Chicago | Bulls | won. |
|---|---|---|---|---|
| label y | 0 | B-Organization | I-Organization | 0 |
| independent guess | 0 | B-Location | 0 | 0 |

b) NER

Data: inputs $x_i \in \mathcal{X}$ and structured labels $y_i \in \mathcal{Y}$ (e.g. sequence) for $i \in \{1, \dots, n\}$.

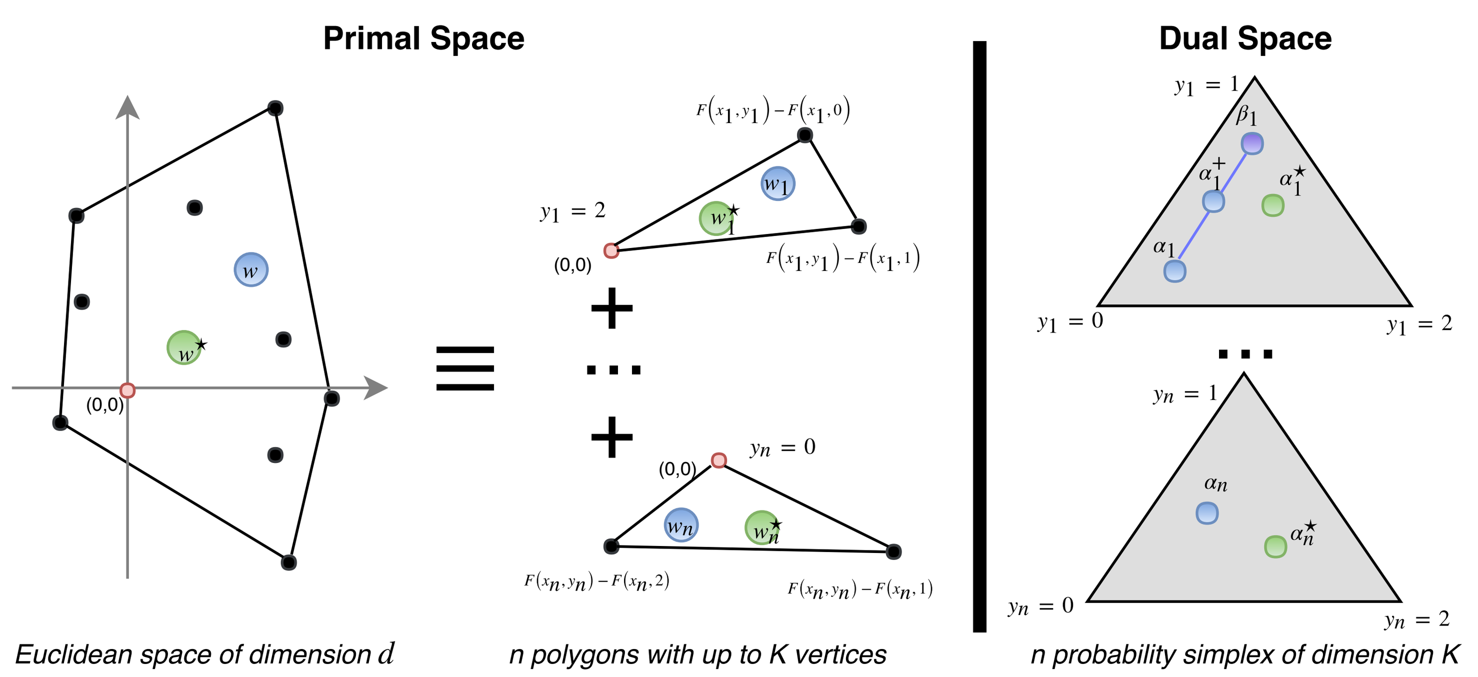
### Conditional Random Fields

- **Model:** exponential family with sufficient statistic $F$: $p(y|x; \boldsymbol{w}) \propto \exp(\boldsymbol{w}^T F(x, y))$.
- **Goal:** learn the optimal parameter $\boldsymbol{w}^*$.
- **Problem:** partition function $= \sum_y e^{\boldsymbol{w}^T F(x,y)} =$ intractable sum over $\mathcal{Y}$ (huge set)
- **Solution:** graphical model of $y|x$ so that $F$ decomposes as a sum over the cliques of the graph. For a sequential graph with $T$ nodes as in OCR:

$$F(x, y) = \sum_{t=1}^{T} F_t(x_t, y_t) + \sum_{t=1}^{T-1} F_t(y_t, y_{t+1}) \tag{1}$$

Use message passing to evaluate the partition function and marginals.

## Parameters




## Optimization Problem

### Primal and Dual

- **Primal:** minimize the $l_2$-regularized negative log-likelihood.

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \mathcal{P}(w) := \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} -\log p(y_i|x_i; \boldsymbol{w}) \tag{2}$$

- **Dual:** maximize the entropy-regularized negative squared loss.

$$\boxed{\max_{\boldsymbol{\alpha}} \mathcal{D}(\boldsymbol{\alpha}) := -\frac{\lambda}{2}\|\hat{w}(\boldsymbol{\alpha})\|^2 + \frac{1}{n}\sum_{i=1}^{n} H(\boldsymbol{\alpha}_i)} \tag{3}$$

- $\forall i, \boldsymbol{\alpha}_i \in \Delta_{|\mathcal{Y}|}$ is a distribution over the labels $\mathcal{Y}$.
- Dual weights $\hat{w}(\boldsymbol{\alpha})$ and entropy $H$ are defined by:

$$\hat{w}(\boldsymbol{\alpha}) := \frac{1}{\lambda n}\sum_i \left(F(x_i, y_i) - \mathbb{E}_{y \sim \alpha_i}[F(x_i, y)]\right), \quad H(\alpha_i) := -\sum_{y \in \mathcal{Y}} \alpha_i(y) \log(\alpha_i(y)) \, .$$

### Properties

- Primal probabilities $\hat{\alpha}_i(\boldsymbol{w}) = p(.|x_i; \boldsymbol{w})$. Dual weights $\hat{w}(\boldsymbol{\alpha})$.
- **Fixed point** property: $\hat{w}(\boldsymbol{\alpha}^\star) = \boldsymbol{w}^\star$ and $\hat{\alpha}(\boldsymbol{w}^\star) = \boldsymbol{\alpha}^\star$.
- **Duality gap:** $g(\boldsymbol{w}, \boldsymbol{\alpha}) := \mathcal{P}(\boldsymbol{w}) - \mathcal{D}(\boldsymbol{\alpha}) \geq 0$ and $g(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star) = 0$.

$$\boxed{g(\hat{w}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = \frac{1}{n}\sum_i D_{KL}(\boldsymbol{\alpha}_i || \hat{\alpha}_i(\hat{w}(\boldsymbol{\alpha}))) = \frac{1}{n}\sum_i g_i} \tag{4}$$

- Idea: the *individual duality gap* $g_i$ is the sub-optimality of data point $i$.

## SDCA for CRF

### Prox-SDCA algorithm

- Primal-dual method for Empirical Risk Minimization. Store $(\boldsymbol{\alpha}, \hat{w}(\boldsymbol{\alpha}))$.
- At each step, update a random block $\alpha_i$ to maximize $\mathcal{D}(\boldsymbol{\alpha})$.
- Guaranteed ascent direction: $\delta_i = \hat{\alpha}_i(\hat{w}(\boldsymbol{\alpha})) - \alpha_i$

$$\boxed{\alpha_i^+ \leftarrow \alpha_i + \gamma\delta_i = (1-\gamma)\alpha_i + \gamma\hat{\alpha}_i(\hat{w}(\boldsymbol{\alpha}))}, \quad \gamma \in [0, 1] \tag{5}$$

- Exact line search on step-size $\gamma \in [0, 1]$ with Newton-Raphson.

### Adaptation to CRFs

- **Problem:** $\alpha_i$ has dimension $|\mathcal{Y}| \gg 1$.
- **Hypothesis:** the CRF graph has a **junction tree** $(\mathcal{C}, \mathcal{S})$. (Similar to [1].)
- **Solution:** Replace joint probability $\alpha_i$ by marginal probabilities $\mu_{i,C}$ on maximal cliques of the graph $C \in \mathcal{C}$. Dual objective can be expressed only with these marginals.

$$\alpha_i(y) = \frac{\Pi_{C \in \mathcal{C}}\, \mu_{i,C}(y_C)}{\Pi_{S \in \mathcal{S}}\, \mu_{i,S}(y_S)} \quad \text{implies} \quad \boxed{H(\alpha_i) = \sum_{C \in \mathcal{C}} H(\mu_{i,C}) - \sum_{S \in \mathcal{S}} H(\mu_{i,S}) =: \tilde{H}(\mu_i) \, .}$$

$$F(x, y) = \sum_{C \in \mathcal{C}} F_C(x, y_C) \quad \text{implies} \quad \boxed{\hat{w}(\boldsymbol{\alpha}) = \sum_{C \in \mathcal{C}} \hat{w}_C(\boldsymbol{\mu}_C) =: \hat{w}(\boldsymbol{\mu}) \, .}$$

- Run message passing to infer $\hat{\mu}_{i,C}(\boldsymbol{w})(y_C) = p(y_C = .|x_i; \boldsymbol{w})$.

### Adaptive Sampling strategy

- Previous work [2] : sample proportionally to $\|\delta_i\|_1 = \|\boldsymbol{\alpha}_i - \hat{\alpha}_i(\hat{w}(\boldsymbol{\alpha}))\|_1$.
- Problem: $\|\delta_i\|_1$ cannot be expressed with the marginals.
- Our strategy: sample proportionally to individual duality gaps $g_i$.
- Strong convexity: proof of acceleration by a factor $\chi(\boldsymbol{g})^2 = \frac{1}{n}\sum_i g_i^2 / \left(\frac{1}{n}\sum_i g_i\right)^2 \geq 1$ .

## Pseudocode

Initialize $\boldsymbol{w} \leftarrow 0$ and $\mu_{i,C}(y_C) \leftarrow \mathbb{1}_{\{y_C = y_{i,C}\}}, \forall i, \forall C \in \mathcal{C}, \forall y_C \in \mathcal{Y}_C$.
(Optional) Let $g_i \leftarrow 1000, \forall i$ and $\bar{g} \leftarrow \frac{1}{n}\sum_i g_i$. {duality gap estimate}
**while** $\bar{g}$ >required precision **do**
  Sample $i$ in $\{1, \dots, n\}$ uniformly at random OR proportionally to $g_i$.
  Let $\nu_{i,C}(y_C) \leftarrow p(y_C|x_i; \boldsymbol{w}), \forall C \in \mathcal{C}$ {**message passing oracle**}
  (Optional) Let $g_i \leftarrow \tilde{D}(\mu_i||\nu_i)$ {individual duality gap}
  Let $\delta_i \leftarrow \nu_i - \mu_i$ {dual ascent direction}
  Let $\boldsymbol{v}_i \leftarrow \hat{w}(\delta_i) = \frac{1}{\lambda n}\sum_{C \in \mathcal{C}} \mathbb{E}_{\mu_{i,C}}[F(x_i, y_C)] - \mathbb{E}_{\nu_{i,C}}[F(x_i, y_C)]$ {primal direction}
  Solve $\gamma^* = \arg\max_{\gamma \in [0,1]} \tilde{H}(\mu_i + \gamma\delta_i) - \frac{\lambda n}{2}\|\boldsymbol{w} + \gamma\boldsymbol{v}_i\|^2$ {line search}
  Update $\mu_i \leftarrow \mu_i + \gamma^*\delta_i$
  Update $\boldsymbol{w} \leftarrow \hat{w}(\boldsymbol{\mu}) = \boldsymbol{w} + \gamma^*\boldsymbol{v}_i$
**return** $\boldsymbol{w}$

## Experiments

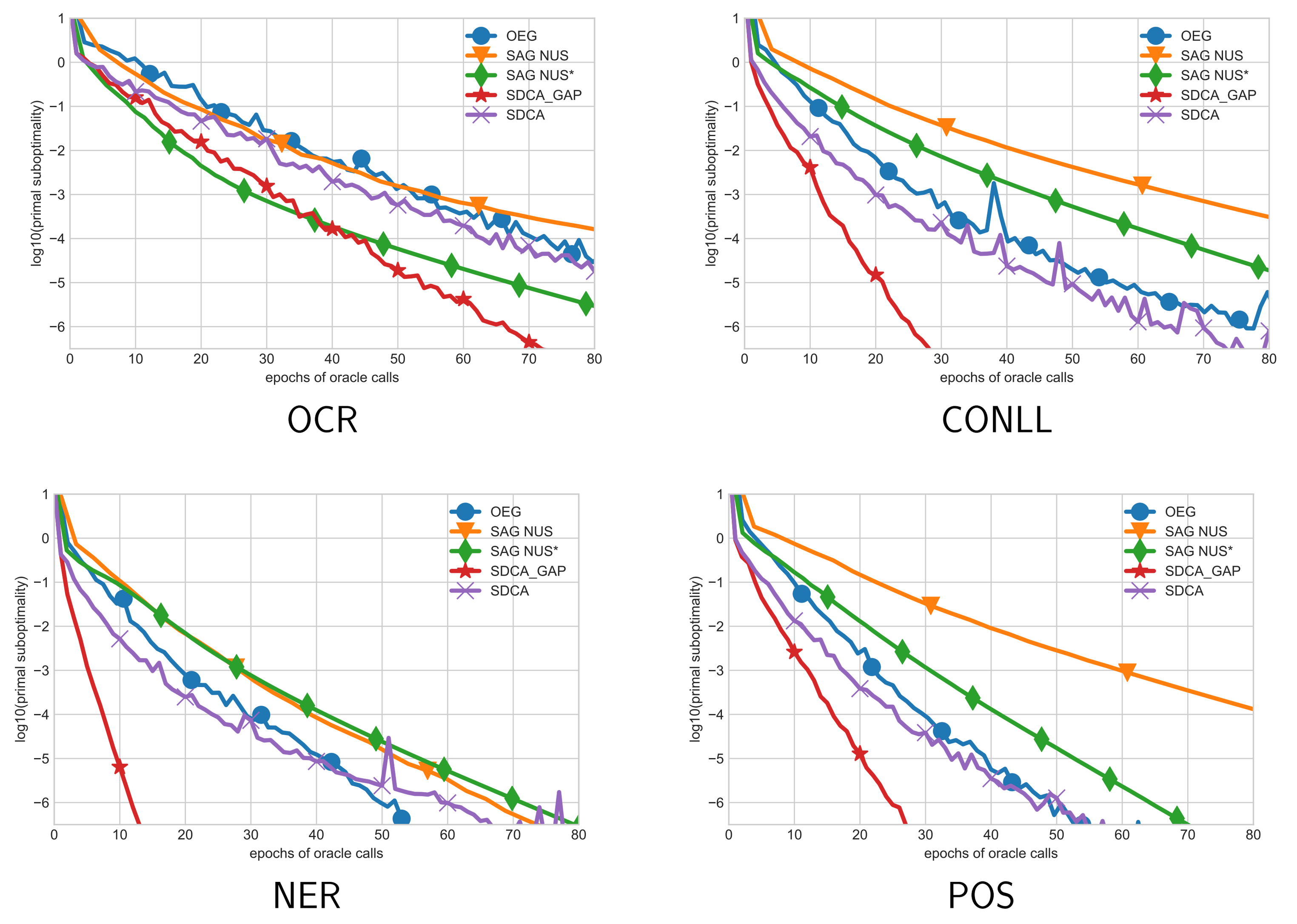



Figure 1: Primal sub-optimality as a function of the number of oracle calls (left). SDCA refers to uniform sampling. SDCA-GAP refers to sampling proportionally to the gaps 80% of the time. SAG-NUS performs a line search at every iteration. SAG-NUS* implements a line-search skipping strategy.

- OCR has dense features (images). All methods perform comparatively.
- CONLL, NER and POS are language understanding tasks with sparse features. Dual methods OEG and SDCA tend to perform better.
- The gap sampling strategy gives a significant advantage to SDCA.

## Take-away

- Dual variance reduced methods can be applied to structured problems.
- Duality gap sampling and exact line search makes them fast on sparse datasets.

## References


[1] M. Collins, A. Globerson, T. Koo, X. Carreras, and P. L. Bartlett. Exponentiated gradient algorithms for conditional random fields and max-margin Markov networks. *Journal of Machine Learning Research*, 2008.

[2] D. Csiba, Z. Qu, and P. Richtárik. Stochastic dual coordinate ascent with adaptive probabilities. In *ICML*, 2015.

[3] M. Schmidt, R. Babanezhad, M. Ahmed, A. Defazio, A. Clifton, and A. Sarkar. Non-uniform stochastic average gradient method for training conditional random fields. In *AISTATS*, 2015.