# An Inexact Dual Augmented Lagrangian Method for Fast CRF Learning

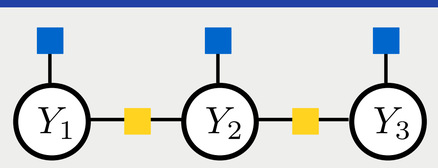
Shell Xu Hu (hus@imagine.enpc.fr) and Guillaume Obozinski (guillaume.obozinski@enpc.fr)

IMAGINE, LIGM, Université Paris-Est and École des Ponts ParisTech

## 1. Introduction

- ***Problem:*** *Maximum likelihood estimation of discrete conditional random fields (CRF).*
- ***Formulation:*** *Dual augmented Lagrangian formulation with variational relaxation of the intractable Shannon entropy and the marginal polytope.*
- ***Algorithm:*** *Inexact dual augmented Lagrangian (IDAL) method, which requires only clique-wise updates (i.e. stochastic coordinate descent) to stochastically approximate the gradient of the Lagrangian multipliers.*

## 2. Conditional Random Fields as Exponential Family




$\mathcal{T} = \{\text{node (blue)}, \text{edge (yellow)}\}$
$\mathcal{C} = \{1, 2, 3, (1,2), (1,3)\}$
$y_{12} = y_1 \otimes y_2$ (one-hot vectors)

The joint distribution over the random variables $Y := [Y_1, \dots, Y_S]$ given the observation $X$ is

$$p(y \mid x; w) := \frac{1}{Z(x,w)} \prod_{\tau \in \mathcal{T}} \prod_{c \in \mathcal{C}_\tau} \exp\Big(\langle w_\tau, \phi_c(x, y_c)\rangle\Big) = \exp\Big[\langle \eta(w,x), T(y)\rangle - F\big(\eta(w,x)\big)\Big].$$

- *Shared parameters*: $w_\tau$ for all cliques with type $\tau$. $w := [w_\tau \colon \forall \tau \in \mathcal{T}]$.
- *Feature map*: $\phi_c(x, y_c)$, which can be a column of the matrix $\Phi_c$, and then paste $\Phi_c$ as the $(\tau_c, c)$-block to form the big matrix $\Phi$.
- *Natural parameter*: $\eta(w,x) := \Phi^\intercal w$. When the context is clear, we will omit the dependency on $x$ and write $\eta(w)$.
- *Sufficient statistics*: $T(y) := [y_c \colon \forall c \in \mathcal{C}]$.
- *Log-partition function*: $F(\eta(w)) := \log \sum_y \exp(\langle \eta(w), T(y)\rangle) = \log Z(w,x)$.

## 3. Maximum Likelihood Estimation

Given i.i.d. samples $\{(x^{(n)}, y^{(n)})\}_{1 \le n \le N}$,

$$\max_w \sum_{n=1}^N \Big[\langle \eta^{(n)}(w), T(y^{(n)})\rangle - F\big(\eta^{(n)}(w)\big)\Big] \Leftrightarrow \min_w \sum_{n=1}^N F\Big(\theta^{(n)}(w) := \Psi^{(n)\intercal} w\Big),$$

where $\theta(w)$ is another natural parameter with $\Psi$ defined similarly as $\Phi$.

- **Computational issue:** $\nabla_{w_\tau} F(\eta(w)) = \sum_{c \in \mathcal{C}_\tau} \Phi_c \mathbb{E}_\theta[y_c] \longleftarrow \mathbb{E}_\theta[y_c], \forall c \in \mathcal{C}_\tau$ $\longleftarrow$ marginal inference at each iteration.

## 4. Variational Relaxation

It is a classical result[3] that

$$F(\theta) = \max_\mu \Big[\langle \mu, \theta\rangle - F^*(\mu)\Big] \qquad F^*(\mu) = -H_{\text{Shannon}}(\mu) + \iota_{\mathcal{M}}(\mu).$$

- *Marginal polytope* $\mathcal{M}$: $\mathcal{M}^\circ := \{\mu \colon \mu = \nabla_\theta F(\theta) = \mathbb{E}_\theta[T(Y)] \text{ for some } \theta\}$.
- **Theoretical issue**: Both $\mathcal{M}$ and $H_{\text{Shannon}}(\mu)$ are in general intractable due to the exponentially large structured-output space.
- *Local marginal polytope*:

$$\mathcal{L} := \Big\{\mu \in \mathcal{I} \mid \forall (s,t) \in E, \mu_s - A_{st}\mu_{st} = 0\Big\} = \mathcal{I} \cap \{\mu \colon A\mu = 0\},$$

  where $\mathcal{I}$ denote the (Cartesian) product of simplices of each clique.
- *Oriented tree-reweighted entropy*:

$$H_{\text{OTRW}}(\mu) = \sum_{s \in V} \rho_s H(\mu_s) + \sum_{s \to t} \rho_{s|t}\Big[H(\mu_{st}) - H(A_{st}\mu_{st})\Big],$$

  which is concave in $\mathcal{I}$ and strongly concave in $\mathcal{L}$.
- *Relaxed* $F$ by $\mathcal{M} \to \mathcal{L}$ and $H_{\text{Shannon}} \to H_{\text{OTRW}}$:

$$F_{\mathcal{L}}(\theta) := \max_\mu \Big[\langle \mu, \theta\rangle - F^*_{\mathcal{L}}(\mu)\Big] \qquad F^*_{\mathcal{L}}(\mu) := -H_{\text{OTRW}}(\mu) + \iota_{\mathcal{I}}(\mu) + \iota_{A\mu=0}.$$

## 5. Inference-Free Formulation

The primal and dual of the regularized MLE (strong duality holds):

$$\text{MLE (P)}: \quad \min_w F_{\mathcal{L}}\big(\theta(w)\big) + \frac{\lambda}{2}\|w\|_2^2 \quad \big(\text{involves marginal inference}\big)$$

$$\text{MaxEnt (D)}: \quad \max_\mu -F^*_{\mathcal{L}}(\mu) - \frac{1}{2\lambda}\|\Psi\mu\|_2^2 \quad \big(\text{involves projections on } \mathcal{L}\big)$$

- The key to avoid running marginal inference (or projections on $\mathcal{L}$) is to use the augmented Lagrangian formulation for $A\mu = 0$:

$$d(\xi) := \max_\mu \Big[D_\rho(\mu, \xi) := \underbrace{H_{\text{OTRW}}(\mu) - \iota_{\mathcal{I}}(\mu) + \langle \xi, A\mu\rangle}_{\text{block-separable \& concave}} \underbrace{-\frac{1}{2\rho}\|A\mu\|_2^2 - \frac{1}{2\lambda}\|\Psi\mu\|_2^2}_{\textit{smooth}}\Big].$$

- For fixed $\xi$, it is natural to optimize $D_\rho(\mu, \xi)$ by stochastic coordinate ascent, so only clique-wise updates are needed. The scheme is almost identical to that of SDCA[2] for regularized loss minimization.

## 6. IDAL Method

The idea is to solve $\min_\xi d(\xi)$ by inexact gradient descent with warm restarts.

- *Exact gradient*: $\nabla d(\xi) = A\overline{\mu}$, where $\overline{\mu} = \arg\max_{\mu \in \mathcal{I}} D_\rho(\mu, \xi)$.
- *Inexact gradient*: $\tilde{\nabla} d(\xi) = A\hat{\mu}$, where $\hat{\mu}$ is the result after running $T_{\text{in}}$ steps of
  1: Draw a clique $c$ uniformly at random.
  2: Update $\mu_c$ with a projected gradient step on $\mu_c \mapsto D_\rho([\mu_c, \mu_{-c}], \xi)$
- *Warm restart*: After each gradient update of $\xi$, we initialize $\mu$ by the previously approximate solution $\hat{\mu}$.

### Theoretical Analysis

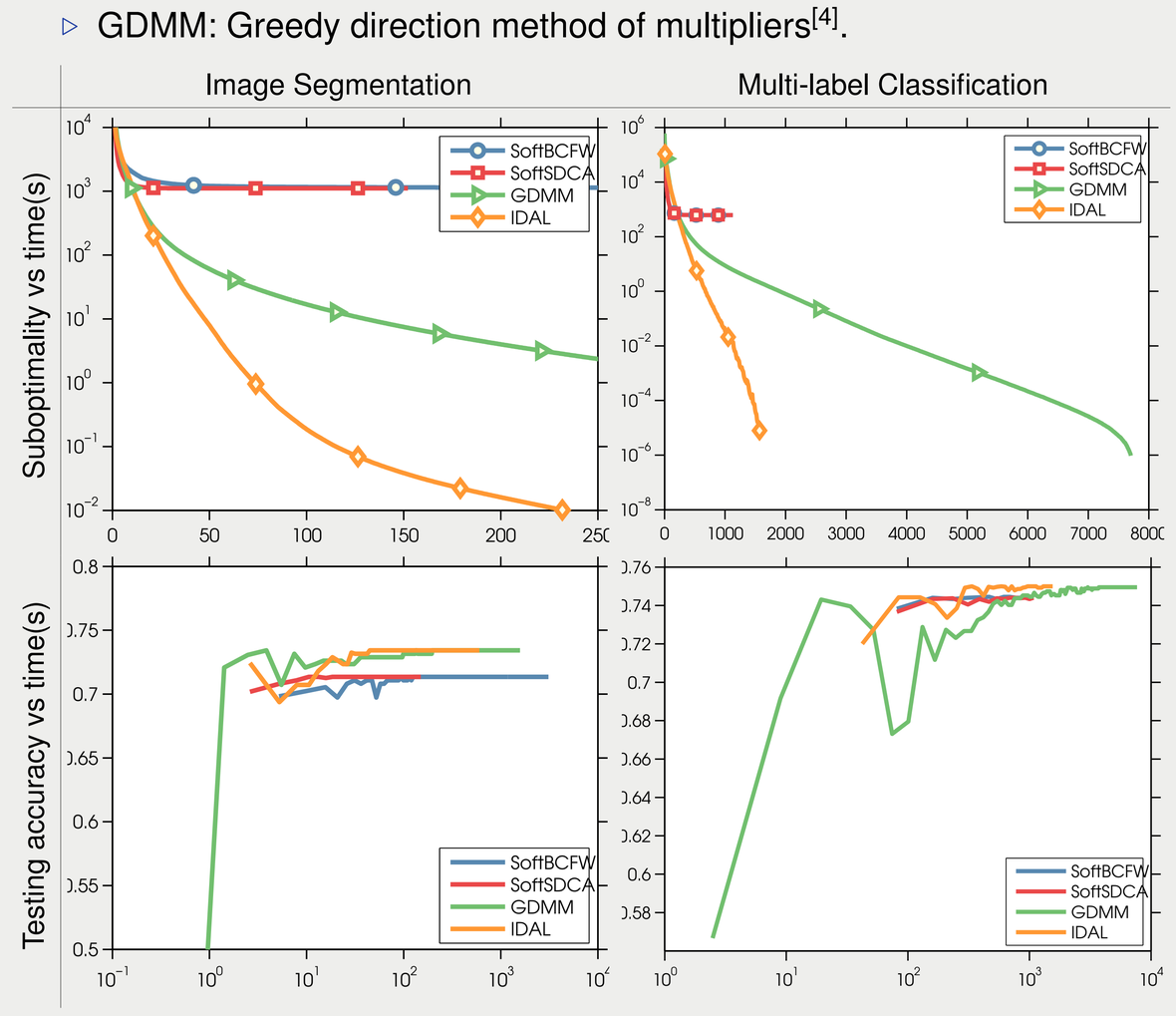
- *Suboptimalities*: $\Gamma_t = d(\xi^t) - d(\xi^*)$, $\hat{\Delta}_t := D_\rho(\overline{\mu}^t, \xi^t) - D_\rho(\hat{\mu}^t, \xi^t)$.
- $d(\xi)$ is $L$-smooth, which guarantees $d(\xi^t) - d(\xi^{t+1})$ is lower bounded.
- There exists a $\tau > 0$, such that $\Gamma_t \le \frac{1}{2\tau}\|g_t\|^2$ holds.
- SDCA on $\mu$ ensure $\hat{\Delta}_t \le (1-\pi)^{T_{\text{in}}}\Delta_t^0$ for some condition number $\pi$.
- It is sufficient to run $T_{\text{in}} > \frac{\log\frac{1}{\alpha} + \log\frac{L}{\tau}}{\pi}$ iterations on $\mu$ for some constant $\alpha \in (0,1)$ to guarantee that, after $T_{\text{ex}}$ iterations on $\xi$, we have

$$\left\|\begin{matrix}\hat{\Delta}_{T_{\text{ex}}} \\ \Gamma_{T_{\text{ex}}}\end{matrix}\right\| \le \sigma_{\max}(M)^{T_{\text{ex}}} C \left\|\begin{matrix}\hat{\Delta}_0 \\ \Gamma_0\end{matrix}\right\|, \text{ where } M = \begin{bmatrix}\frac{7\alpha\tau}{L} & \frac{3\alpha\tau}{L} \\ 1 & 1 - \frac{\tau}{L}\end{bmatrix}, C > 0.$$

  Therefore, if $\alpha$ is chosen so that $\sigma_{\max}(M) < 1$, the algorithm is **linearly convergent** with rate $\sigma_{\max}(M)$.

## 7. Experiments

- *Baselines* using clique-wise updates:
  - SoftBCFW/SDCA: Soft-constrained block-coordinate Frank-Wolfe[1] / stochastic dual coordinate ascent for a special case $\max_\mu D_\rho(\mu, \xi = 0)$.
  - GDMM: Greedy direction method of multipliers[4].




## References

[1] O. Meshi, N. Srebro, and T. Hazan. "Efficient Training of Structured SVMs via Soft Constraints". In: *AISTATS*. 2015.

[2] S. Shalev-Shwartz and T. Zhang. "Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization". In: *ICML*. 2014.

[3] M. J. Wainwright. "Graphical models, exponential families, and variational inference". In: *Foundations and Trends in Machine Learning* (2008).

[4] I. H. Yen et al. "Dual Decomposed Learning with Factorwise Oracle for Structural SVM of Large Output Domain". In: *Advances in Neural Information Processing Systems*. 2016, pp. 5024–5032.