SDCA-Powered Inexact Dual Augmented Lagrangian Method for Fast CRF Learning

Shell Xu Hu
LIGM (UMR 8049), UPE
École des Ponts ParisTech
hus@imagine.enpc.fr

Guillaume Obozinski
LIGM (UMR 8049), UPE
École des Ponts ParisTech
guillaume.obozinski@enpc.fr

Abstract

We propose an efficient *dual* augmented Lagrangian formulation to learn conditional random fields (CRF). Our algorithm, which can be interpreted as an inexact gradient descent algorithm on the multiplier, does not require to perform global inference iteratively, and requires only a fixed number of stochastic clique-wise updates at each epoch to obtain a sufficiently good estimate of the gradient w.r.t. the Lagrange multipliers. We prove that the proposed algorithm enjoys global linear convergence for both the primal and the dual objectives. Our experiments show that the proposed algorithm outperforms state-of-the-art baselines in terms of speed of convergence.

1 Introduction

Learning in graphical models has historically relied on the computation of the (sub)gradient of the log-likelihood w.r.t. to the canonical parameters, which requires to solve a MAP or probabilistic inference problem at each iteration. This approach is slow given that the inference problem is itself computationally expensive. The difficulty of inference and learning in graphical models is related to the fact that the log-partition function is in general intractable.

Recent progress on the optimization problems whose objective is a large finite sum of convex terms has shown that they could be optimized very efficiently by stochastic algorithms that sample one term at a time (Defazio et al., 2014; Roux et al., 2012; Shalev-Shwartz and Zhang, 2016). It turns out that the dual objective of the maximum likelihood estimation of CRF (a.k.a. the maximum entropy principle) decomposes additively over all cliques if a decomposable entropy surrogate is used. Even though this dual formulation has a potential to take advantage of stochastic algorithms, and can be optimized without resorting to solve a global inference on the entire graph per iteration, all dual parameters (i.e. *mean parameters*) are coupled by the *marginal polytope* constraints, which are in general intractable. Even its most commonly used relaxation, namely the *local consistency polytope*, is itself in practice difficult to optimize over. Recently, Meshi et al. (2015a,b) proposed to replace the marginalization constraints, which are part of the local consistency polytope, by quadratic penalty terms. The relaxed problem has then only separable constraints over the cliques that makes it possible to use efficient block coordinate optimization schemes.

Following these ideas, we consider a dual formulation for CRF learning in which the marginalization constraints are replaced by an augmented Lagrangian term, and the intractable Shannon entropy is replaced by a quadratic surrogate so that stochastic dual coordinate ascent (SDCA) can be used to optimize over the mean parameters, with similar guarantees as in Shalev-Shwartz and Zhang (2016). We finally show that by periodically updating the Lagrangian multipliers as we are optimizing the relaxed dual, we can gradually enforce the marginalization constraints, while retaining global linear convergence. In terms of the primal problem associated with the Lagrange multipliers, our algorithm is an inexact gradient descent algorithm using stochastic approximation of the multiplier gradients.

Our paper is organized as follows. We review CRF learning in Section 3. A dual augmented Lagrangian formulation is presented in Section 4. The proposed algorithm is presented in Section 5, followed by its convergence analysis in Section 6. Finally, we present experiments on three applications in Section 7 (Most notations used in the paper can be found in Appendix F).

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

2 Related Work

Due to the independent interest of inference problem in discrete graphical models, in particular in computer vision, a significant amount of work has been devoted to develop efficient approximate inference algorithms (Komodakis et al., 2007; Martins et al., 2015; Savchynskyy et al., 2011; Sontag et al., 2008). However, the learning problem is not necessarily easier (can even fail to converge) with an approximate inference approach as the subroutine (Kulesza and Pereira, 2007).

There is a large body of research on efficient algorithms for structured learning. For the max-margin formulation, the fastest algorithms to date rely on block coordinate Frank-Wolfe updates (Lacoste-Julien et al., 2013; Meshi et al., 2015b; Tang et al., 2016). Using dual decomposition in the inner inference problem, Hazan and Urtasun (2010); Komodakis (2011); Meshi et al. (2010) proposed to solve the classical saddle-point formulation for structured learning problem with algorithms that alternate between message passing and model parameter updates. Going further Meshi et al. (2015b); Yen et al. (2016) work on a purely dual formulation to enable clique-wise updates. For maximum likelihood learning, exponentiated gradient and its block variants can be applied (Collins et al., 2008). Other recent work have relied on incremental algorithms (Schmidt et al., 2015) and the fact that the Gauss-Southwell rule can be applied efficiently for coordinate descent in some forms of graphical models (Nutini et al., 2015).

The BCMM algorithm of Hong et al. (2014) which uses stochastic block coordinate updates inside ADMM inspired our approach. But our algorithm performs multiple passes over all blocks before updating the multiplier; and we prove stronger convergence rates.

We list related structured learning methods with their main characteristics in Table 1 in Appendix B.5.

Yen et al. (2016) is the most similar work to ours: the proposed algorithm constructs greedily an (initially sparse) working set of cliques, which is incremented at each epoch, while we perform stochastic updates on all cliques and possibly several passes over the data between each update of all Lagrange multipliers. Also, our work is leveraging the connection with SDCA, and we prove both linear convergence in the primal and the dual whereas Yen et al. (2016) prove only linear convergence in the dual. Finally, our algorithm is outperforming other methods in experiments.

3 CRF Learning

A discrete *conditional random field* (CRF) is a family of conditional distributions over a vector of discrete random variables $Y := (Y_1, \dots, Y_m)$ given the observation X. The form of the CRF is assumed to be a product of *local functions* (a.k.a. *factors* or *clique functions*) that each depends on only a small number of random variables (i.e. a *clique*). If there exists multiple cliques that share the same local function, then we group cliques by *clique types*. Specifically, let $w_\tau \in \mathbb{R}^{d_\tau}$ be the parameter vector associated with the clique type $\tau \in \mathcal{T}$, where $\mathcal{T}$ is the set of clique types. Let $\mathcal{C}$ denote the set of all cliques, and $\mathcal{C}_\tau$ the set of cliques of type $\tau \in \mathcal{T}$. Note that each clique c has a unique clique type, which we denote by τ_c. With these notations the density function of the CRF can be written as

$$p(y|x;w) := \frac{1}{Z(x,w)} \prod_{\tau \in \mathcal{T}} \prod_{c \in \mathcal{C}_\tau} \exp\Big(\langle w_\tau, \phi_c(x, y_c)\rangle\Big),$$

where $w = (w_\tau)_{\tau \in \mathcal{T}}$; we denoted $Z(x,w)$ the *partition function* and $\phi_c(x, y_c) \in \mathbb{R}^{d_{\tau_c}}$ the *feature map* for clique c. Since all random variables are discrete, we use a one-hot vector $y_i \in \mathcal{Y}_i := \{u \in \{0,1\}^{k_i} : \|u\|_1 = 1\}$ to represent the value of Y_s. Here k_i is the cardinality of $\mathcal{Y}_i$. For a clique c, the value for the corresponding random variables is $y_c = \otimes_{i \in c}\, y_i \in \mathcal{Y}_c := \bigotimes_{i \in c} \mathcal{Y}_i$, where $\otimes$ (resp. $\bigotimes$) denotes the tensor product of vectors (resp. of spaces). Similarly, $y \in \mathcal{Y}$ is of the form $y = \otimes_{i \in \mathcal{V}}\, y_i$. W.l.o.g., we consider in the paper only cliques of size at most 2, that is $\mathcal{C} = \mathcal{V} \cup \mathcal{E}$, with $\mathcal{V}$ and $\mathcal{E}$ respectively the set of nodes and of edges of the graph; the framework generalizes easily to higher-order cliques. Notations used in the paper are listed in Appendix F.

3.1 CRF as exponential family

Given a sample $(x^{(n)}, y^{(n)})$, for each clique c, let $\eta_c^{(n)}(w) := [\langle w_{\tau_c}, \phi_c(x^{(n)}, y_c)\rangle : y_c \in \mathcal{Y}_c]$; then a *natural parameter* for the exponential family form of the conditional distribution $p(y \mid x^{(n)})$ is $\eta^{(n)}(w) := [\eta_c^{(n)}(w) : c \in \mathcal{C}]$. The associated *sufficient statistics* is $T(y) := [y_c : c \in \mathcal{C}]$, and $\langle \eta^{(n)}(w), T(y)\rangle = \sum_c \langle \eta_c^{(n)}(w), y_c\rangle$. With these notations, $p(y \mid x^{(n)})$ has the exponential family form:

$$p\big(y \mid \eta^{(n)}(w)\big) = \exp\Big[\langle \eta^{(n)}(w), T(y)\rangle - F\big(\eta^{(n)}(w)\big)\Big],$$

where $F(\eta) := \log \sum_y \exp\langle \eta, T(y)\rangle = \log Z(x^{(n)}, w)$ is the log-partition function.

Given i.i.d. samples $\{(x^{(n)}, y^{(n)})\}_{1 \le n \le N}$, the maximum likelihood estimator for w is computed by the maximizing $\sum_n \log p(y^{(n)} \mid x^{(n)}; w)$. Using the exponential family representation, we can rewrite this problem in two equivalent forms:

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

and $\min_w \sum_{n=1}^N F\big(\theta^{(n)}(w)\big)$ with $\theta^{(n)}(w)$ another natural parameter obtained via the affine transformation $\theta^{(n)}(w) = \eta^{(n)}(w) - \langle \eta^{(n)}(w), T(y^{(n)})\rangle \mathbf{1}$. Alternatively, by defining $\Psi^{(n)}$ as a sparse block matrix with $|\mathcal{T}| \times |\mathcal{C}|$ blocks, whose (τ_c, c)-th block is the matrix $\Psi_c^{(n)} \in \mathbb{R}^{d_{\tau_c} \times k_c}$ with

$$\Psi_c^{(n)} = [\phi_c(x^{(n)}, y_c) - \phi_c(x^{(n)}, y_c^{(n)}) \colon y_c \in \mathcal{Y}_c],$$

we have $\theta_c^{(n)}(w) = \Psi_c^{(n)\intercal} w_{\tau_c}$ and $\theta^{(n)}(w) = \Psi^{(n)\intercal} w$.

W.l.o.g., we assume $N = 1$ and drop the superscript (n) from now on, since one may view N graphs as a single large graph with several connected components.

Regularized maximum likelihood estimation with a regularization constant $\lambda > 0$ is thus formulated as

$$\min_w F\big(\theta(w)\big) + \frac{\lambda}{2}\|w\|_2^2. \tag{1}$$

In order to extend this formulation to cover as well max-margin learning (i.e., structured SVMs), we consider the loss-augmented CRF learning introduced by Pletscher et al. (2010) and Hazan and Urtasun (2010), which leads to a slightly generalized formulation:

$$\min_w \gamma F\big(\tfrac{1}{\gamma}\theta_\ell(w)\big) + \frac{\lambda}{2}\|w\|_2^2, \tag{2}$$

where $\theta_\ell(w) := \theta(w) + \ell$ is then the natural parameter, with $\ell = \big[[\ell_c(y_c^\star, y_c) \colon y_c \in \mathcal{Y}_c] \colon c \in \mathcal{C}\big]$ the user-defined loss and $\gamma \in (0, +\infty)$ the temperature hyperparameter. For a derivation for the loss-augmented CRF see Appendix A.

It is well known that the cost of gradient descent to optimize either (1) or (2) (for $\gamma > 0$) is prohibitive since $\nabla_{w_\tau} F(\theta(w)) = \sum_{c \in \mathcal{C}_\tau} \Psi_c \mathbb{E}_\theta[Y_c]$ involves an expectation over the exponentially large space $\mathcal{Y}$. To exploit the underlying structure of the function F it is useful to work on the dual problem. Indeed, since F is convex, it has a variational representation based on conjugate duality:

$$F(\theta) = \max_\mu \langle \mu, \theta \rangle - F^*(\mu),$$

where F^* is the Fenchel conjugate of F, and the dual variable μ called the *mean parameter* is defined by $\mu = (\mu_c)_{c \in \mathcal{C}}$ with $\mu_c = \mathbb{E}_\theta[Y_c]$. The set of valid mean parameters form the so called *marginal polytope* $\mathcal{M}$, which is defined as the convex hull of $\{T(y) : y \in \mathcal{Y}\}$. Moreover, if let $H_{\text{Shannon}}(\mu)$ denote the Shannon entropy of a CRF with mean parameter μ, it is a classical result (Wainwright, 2008, Thm 3.4) that

$$F^*(\mu) = -H_{\text{Shannon}}(\mu) + \iota_{\mathcal{M}}(\mu),$$

where $\iota_{\mathcal{M}}(\mu)$ equal to 0 if $\mu \in \mathcal{M}$ and $+\infty$ otherwise.

4 Relaxed Formulations

In this section, we derive general relaxed dual, primal and corresponding saddle-point formulations for the CRF learning problem: first, we use the classical local polytope relaxation (Sec. 4.1). Second, we further relax the marginalization constraints via an augmented Lagrangian (Sec. 4.2). Third, we propose a surrogate for the entropy, which is decomposable, and retains good properties even when the aforementioned constraints are relaxed (Sec. 4.3). The resulting formulation is convex and is amenable to fast optimization algorithm that are presented in Section 5.

4.1 Classical local polytope relaxation

Both $\mathcal{M}$ and $H_{\text{Shannon}}(\mu)$ are in general intractable due to the exponentially large structured-output space $\mathcal{Y}$ and they are typically replaced by decomposable surrogates.

It is common to relax $\mathcal{M}$ to the *local consistency polytope* (Wainwright, 2008)

$$\mathcal{L} := \Big\{\mu \in \mathcal{I} \colon \sum_{y_j \in \mathcal{Y}_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i), \forall \{i,j\} \in \mathcal{E}, \forall y_i\Big\},$$

where $\mathcal{I}$ denotes the Cartesian product of *simplex constraints* on each clique. Note that $\mathcal{L} \supseteq \mathcal{M}$, since any set of true marginals must satisfy the simplex constraints and the *marginalization constraints*, but not vice versa. Equivalently, if we define $A_i = I_{k_i} \otimes \mathbf{1}_{k_i}^\intercal$, the equality constraints can be written in a matrix form as $\mu_i - A_i \mu_{ij} = 0$ for all $\{i,j\} \in \mathcal{E}$. Combining all equations, we have $A\mu = 0$, where A is a $|\mathcal{E}| \times |\mathcal{C}|$ block matrix (see Appendix F). So, we have equivalently $\mathcal{L} = \mathcal{I} \cap \{\mu \colon A\mu = 0\}$.

Since H_{Shannon} is also intractable for graphs with large tree-width, we will use an approximation H_{Approx} which will be constructed so as to be defined and concave on the whole set $\mathcal{I}$. We propose several entropy approximations suited to our needs in Section 4.3.

Definition 1. *Let $F_{\mathcal{I}}$ and $F_{\mathcal{L}}$ be the counterparts of F obtained by relaxing $\mathcal{M}$ to $\mathcal{I}$ and $\mathcal{L}$ respectively, which, in other words, are the Fenchel conjugates of $F_{\mathcal{I}}^*$ and $F_{\mathcal{L}}^*$ when these are defined by H_{Approx}:*

$$F_{\mathcal{I}}(\theta_\ell) := \max_\mu \langle \mu, \theta_\ell \rangle - F_{\mathcal{I}}^*(\mu),$$

$$F_{\mathcal{L}}(\theta_\ell) := \max_\mu \langle \mu, \theta_\ell \rangle - F_{\mathcal{L}}^*(\mu),$$

where $F_{\mathcal{I}}^*(\mu) := -H_{\text{Approx}}(\mu) + \iota_{\mathcal{I}}(\mu)$ and $F_{\mathcal{L}}^*(\mu) := F_{\mathcal{I}}^*(\mu) + \iota_{\{A\mu=0\}}$.

Replacing F with $F_{\mathcal{L}}$ in (2) yields the relaxed primal

$$P(w) := \gamma F_{\mathcal{L}}\big(\tfrac{1}{\gamma}\theta_\ell(w)\big) + \frac{\lambda}{2}\|w\|_2^2. \tag{3}$$

The corresponding dual objective function is given by

$$D(\mu) := \langle \mu, \ell \rangle - \gamma F^*_{\mathcal{L}}(\mu) - \frac{1}{2\lambda}\|\Psi\mu\|_2^2. \quad (4)$$

See Appendix B.1 for a derivation.

4.2 A dual augmented Lagrangian

It is difficult to optimize $D(\mu)$, since the optimization requires some form of projection onto $\mathcal{L}$, which can be shown to be equivalent to perform graph-wise marginal inference (Collins et al., 2008). The difficulty is due to the coupling equality constraint $A\mu = 0$. Meshi et al. (2015b) proposed to relax $\iota_{\{A\mu=0\}}$ by a quadratic term $\frac{1}{2\rho}\|A\mu\|_2^2$, which corresponds to employ the *penalty method* (Bertsekas, 1982). They argue that it is not crucial to enforce exact $A\mu = 0$ in learning, since the relaxed problem works well in practice and enables an efficient optimization with only clique-wise updates. However, the penalty method is known to have issues associated with the choice of ρ: unless we use a carefully designed scheduling to update ρ, for a reasonably small ρ, the algorithm will be slow; on the other hand, using a large fixed value of ρ degrades the problem to independent logistic regression problems, and, thereby, leads to suboptimal solutions.

Instead, we propose to solve problem (4) as a saddle problem of the form $\max_\mu \min_\xi D_\rho(\mu, \xi)$ where D_ρ is the augmented Lagrangian

$$D_\rho(\mu, \xi) := \Big[\langle \ell, \mu \rangle - \gamma F^*_{\mathcal{I}}(\mu) + \langle \xi, A\mu \rangle\Big] - \Big[\frac{1}{2\rho}\|A\mu\|_2^2 + \frac{1}{2\lambda}\|\Psi\mu\|_2^2\Big], \quad (5)$$

with ξ is the Lagrangian multiplier and $\rho > 0$.

Using duality again, we can derive an associated relaxed primal objective

$$\tilde{P}_\rho(w, \delta, \xi) := \gamma F_{\mathcal{I}}\Big(\frac{\theta_\ell(w) + A^\intercal\delta}{\gamma}\Big) + \frac{\lambda}{2}\|w\|_2^2 + \frac{\rho}{2}\|\delta - \xi\|_2^2,$$

so that $\min_{(w,\delta)} \tilde{P}_\rho(w, \delta, \xi)$ is a primal problem associated with the dual problem $\max_\mu D_\rho(\mu, \xi)$.

Strong duality between these two problems yields a representer theorem

$$w^\star = -\frac{1}{\lambda}\Psi\mu^\star, \quad \delta^\star = \xi^\star - \frac{1}{\rho}A\mu^\star \quad (6)$$

which provides a duality gap

$$\mathrm{gap}(w, \delta, \mu, \xi) := \tilde{P}_\rho(w, \delta, \xi) - D_\rho(\mu, \xi)$$

for the convergence of the maximization of $D_\rho(\mu, \xi)$ with respect to μ. Moreover, it is easy to check that $\min_{\xi,\delta} \tilde{P}_\rho(w, \delta, \xi) = P(w)$ because $\min_\delta F_{\mathcal{I}}(\theta(w) + A^\intercal\delta) = F_{\mathcal{L}}(\theta(w))$ for any w (see Appendix B.2). This shows that $w^\star$ defined in (6) is also an optimum of the original primal problem $\min_w P(w)$. As a consequence, if a sequence μ^t converges to $\mu^\star$ then the corresponding $w^t = -\frac{1}{\lambda}\Psi\mu^t$ converges to a solution of (2). For more details, see Appendix B.

4.3 Gini entropy surrogate

We seek a concave entropy surrogate H_{Approx} that decomposes additively on the cliques. Since the constraint $A\mu = 0$ is relaxed, we need a surrogate well defined on the whole set $\mathcal{I}$. The Bethe entropy (Yedidia et al., 2005) is generally non-concave. Its concave counterparts, such as the tree-reweighted entropy (Wainwright et al., 2005) or the region-based entropy (London et al., 2015; Yedidia et al., 2005), are only concave on the local consistency polytope, but non-concave on $\mathcal{I}$.

Moreover, a generic difficulty with these entropies is that they do not have Lipschitz gradients, which prevents the direct application of proximal methods with usual quadratic proximity terms. We thus propose a coarse but convenient entropy surrogate of the form:

$$H_{\text{Approx}}(\mu) = \sum_{c \in \mathcal{C}} h_c(\mu_c) \quad \text{with} \quad h_c(\mu_c) := (1 - \|\mu_c\|_2^2).$$

Another surrogate with the same separable form is the second-order Taylor expansion of the *oriented tree-reweighted entropy* (OTRW, Globerson and Jaakkola, 2007) around the uniform distribution. This surrogate is also concave on $\mathcal{I}$ (although not strongly concave) and smooth. Preliminary experiments however did not show that using this more sophisticated entropy improved the results. See Appendix C for more details.

5 Algorithm

Given the form of the entropy surrogate proposed, D_ρ decomposes as a sum of convex separable terms over the block associated to cliques plus a smooth term:

$$\begin{aligned} D_\rho(\mu, \xi) &= -\sum_{c \in C} f^*_c(\mu_c) - r(\mu) \quad \text{with} \quad (7)\\ f^*_c(\mu_c) &:= -\gamma h_c(\mu_c) + \iota_{\triangle_c}(\mu_c)\\ r(\mu) &:= -\langle A^\intercal\xi + \ell, \mu \rangle + \frac{1}{2\lambda}\|\Psi\mu\|^2 + \frac{1}{2\rho}\|A\mu\|^2, \end{aligned}$$

where $\triangle_c := \{\mu_c \in \mathbb{R}^{d_c}_+ \mid \mu_c^\intercal \mathbf{1} = 1\}$ is the canonical simplex. It can thus be maximized efficiently by a block-coordinate proximal scheme, such as the proximal stochastic dual coordinate descent (SDCA, Shalev-Shwartz and Zhang, 2016), which has linear convergence guarantees both in the primal and the dual.

Algorithm 1 IDAL scheme

```
Input: $T_{\text{in}}, T_{\text{ex}}, \epsilon$
Initialize: $\hat{\mu}_c^0 = \frac{1}{k_c}\mathbf{1}$ for all $c \in \mathcal{C}$ and $\xi^1 = 0$
for $t = 1, \ldots, T_{\text{ex}}$ do
    $\hat{\mu}^t = \mathcal{A}(\hat{\mu}^{t-1}, T_{\text{in}}, t)$
    Stop if $G_t \leq \epsilon$ and $\|A\hat{\mu}^t\|^2 \leq \epsilon$
    $\xi^{t+1} = \xi^t - \frac{1}{L_d} A\hat{\mu}^t$
end for
Output: $\hat{\mu}^{T_{\text{ex}}}, \xi^{T_{\text{ex}}}$
```

Algorithm 2 SDCA version of $\mathcal{A}(\mu, T_{\text{in}}, t)$

```
$\mu^{t,0} = \mu$
for $s = 1, \ldots, T_{\text{in}}$ do
    Draw a clique $c$ uniformly at random
    $\mu_c^{t,s} = \text{Prox}_{\frac{1}{L_c} f_c^*}\big(\mu_c^{t,s-1} - \frac{1}{L_c}\nabla_{\mu_c} r(\mu^{t,s-1})\big)$
    $\mu_{-c}^{t,s} = \mu_{-c}^{t,s-1}$
end for
Output: $\mu^{t,T_{\text{in}}}$
```

To solve $\min_\xi \max_\mu D_\rho(\mu, \xi)$ we thus propose an algorithm similar to the block coordinate method of multipliers (BCMM) of Hong et al. (2014): perform dual stochastic block coordinate ascent (SDCA) on the variables μ_c to partially maximize $D_\rho(\mu, \xi)$ in μ and regularly take a gradient descent step in ξ. Our algorithm, is an inexact dual augmented Lagrangian (IDAL) method, in the sense that it is an inexact gradient descent algorithm on the function $\xi \mapsto d(\xi) := \max_\mu D_\rho(\mu, \xi)$. To be precise, if at epoch t, ξ takes the value ξ^t and $\hat{\mu}^{t-1}$ is the value of μ from the previous epoch, Algorithm 2 takes T_{in} stochastic block-coordinate proximal gradient steps on μ to obtain $\hat{\mu}^t$. Denoting L_c the Lispchitz constant of r w.r.t. μ_c, μ_c is then updated by a partial gradient step, and an application of the proximal operator of $\frac{1}{L_c} f_c^*$. Then, by Danskin's theorem[1], applied to equation (5), we have that $A\hat{\mu}^t$ is an approximate gradient of $d(\xi^t)$, and so, Algorithm 1 updates ξ with $\xi^{t+1} = \xi^t - \frac{1}{L_d} A\hat{\mu}^t$, where L_d is the Lispchitz constant of $d(\xi)$. As for the stopping criteria, we use $G_t := \text{gap}(w(\hat{\mu}^t), \delta(\hat{\mu}^t, \xi^t), \hat{\mu}^t, \xi^t) \leq \epsilon$ and $\|A\hat{\mu}^t\|^2 \leq \epsilon$, where $w(\hat{\mu}^t), \delta(\hat{\mu}^t, \xi^t)$ are defined via the representer theorem (6) (see Appendix B.4).

6 Convergence Analysis

In this section, we study the convergence rate of our algorithm. First, we show that if we use an iterative and linearly convergent algorithm $\mathcal{A}$ to approximately solve $\min_\mu D_\rho(\mu, \xi)$, and if we use warm starts, that is, following the notations of the previous section, we use $\hat{\mu}^{t-1}$ as the initial value to solve $\min_\mu D_\rho(\mu, \xi^t)$, then running $\mathcal{A}$ for a fixed number of iterations is sufficient to guarantee global linear convergence in the primal and in the dual. We show that SDCA or simple block-coordinate proximal gradient descent are applicable as the algorithm $\mathcal{A}$.

6.1 Conditions for global linear convergence

To study the convergence, we consider:

- $\bar{\mu}^t := \mu^\star(\xi^t) = \text{argmax}_\mu D_\rho(\mu, \xi^t)$.
- $\mu^{t,s}$, the value of μ after s inner steps at epoch t.
- $\hat{\mu}^t := \mu^{t,T_{\text{in}}}$ the value of μ at the end of epoch t.
- D_ρ-suboptimality: $\Delta_t^s := D_\rho(\bar{\mu}^t, \xi^t) - D_\rho(\mu^{t,s}, \xi^t)$, with at the end of each epoch $\hat{\Delta}_t := \Delta_t^{T_{\text{in}}} = \Delta_{t+1}^0$.
- d-suboptimality: $\Gamma_t := d(\xi^t) - d(\xi^\star)$.

Lemma 1 (Linear convergence of the outer iteration)**.** *Let $\mathcal{A}$ be an algorithm that approximately solves $\max_\mu D_\rho(\mu, \xi^t)$ in the sense that*

$$\exists \beta \in (0, 1), \qquad \mathbb{E}[\hat{\Delta}_t] \leq \beta\, \mathbb{E}[\Delta_t^0].$$

Then, $\exists \kappa \in (0, 1)$ characterizing $d(\xi)$ and $C > 0$, such that, if $\lambda_{\max}(\beta)$ is the largest eigenvalue of the matrix

$$M(\beta) = \begin{bmatrix} 6\beta & 3\beta \\ 1 & 1 - \kappa \end{bmatrix},$$

then after T_{ex} iterations of Algorithm 1 we have

$$\left\| \begin{matrix} \mathbb{E}[\hat{\Delta}_{T_{\text{ex}}}] \\ \mathbb{E}[\Gamma_{T_{\text{ex}}}] \end{matrix} \right\| \leq C\, \lambda_{\max}(\beta)^{T_{\text{ex}}} \left\| \begin{matrix} \mathbb{E}[\hat{\Delta}_0] \\ \mathbb{E}[\Gamma_0] \end{matrix} \right\|.$$

The constant κ in the theorem is of the form $\kappa = \frac{\tau}{L_d}$ with L_d the Lipschitz constant of $d(\xi)$ and τ a restricted strong convexity constant for $d(\xi)$ obtained by Hong and Luo (2017) (see Lemma D.3 in Appendix D.2).

Corollary 1. *If $\mathcal{A}$ is a linearly convergent algorithm with rate π and if it is run for T_{in} iterations, such that, for some β: $\lambda_{\max}(\beta) < 1$, we have $(1 - \pi)^{T_{\text{in}}} \leq \beta$, then $\mathbb{E}[\hat{\Delta}_t]$ and $\mathbb{E}[\Gamma_t]$ converge linearly to 0.*

Note that linear convergence of the expectations implies that Δ_t and Γ_t converge linearly to 0 almost surely, as a classical consequence of Markov's inequality and the Borel-Cantelli lemma. We will show in the next section that when $\mathcal{A}$ is SDCA it is linearly convergent.

Note that the convergence of the gaps Δ_t and Γ_t imply the linear convergence for the augmented Lagrangian formulation, in the following sense:

Corollary 2. *Let $D_\infty(\mu) := \langle \ell, \mu \rangle - \gamma F_{\mathcal{I}}^*(\mu) - \frac{1}{2\lambda}\|\Psi\mu\|_2^2$, so that we have $D(\mu) = D_\infty(\mu) - \iota_{\{A\mu=0\}}$. If Δ_t and Γ_t converge linearly to 0, then $|D_\infty(\hat{\mu}^t) - D_\infty(\mu^\star)|$ and $\|A\hat{\mu}^t\|_2^2$ both converge to 0 linearly.*

[1] (see e.g. Bertsekas, 1999, Prop. B.25)

Furthermore, if $\mathcal{A}$ is linearly convergent as in Corollary 2, the algorithm is linearly convergent in terms of the total number of inner steps (for SDCA this is the total number of clique updates) performed by algorithms $\mathcal{A}$ throughout:

Corollary 3. *With the notations of the previous corollary, for any $\beta \in (0, 1)$ such that $\lambda_{\max}(\beta) < 1$, it is possible to obtain $\mathbb{E}[\hat{\Delta}_t] \le \epsilon$ and $\mathbb{E}[\Gamma_t] \le \epsilon$ with a total number of inner iterations $T_{\text{tot}} := T_{\text{in}} T_{\text{ex}}$ such that*

$$T_{\text{tot}} \ge \frac{\log(\beta)}{\log \lambda_{\max}(\beta) \log(1 - \pi)} \log(\epsilon).$$

We show in Appendix D.4 that to have $\lambda_{\max}(\beta) < 1$ we should have $\beta = \alpha\kappa$ with $\alpha < \frac{1}{3(1+2\kappa)}$.

To reason in terms of rate, if the rate of convergence is r then we should have $T_{\text{tot}} \ge \frac{\log(\epsilon)}{\log(1-r)}$. So identifying the rate of convergence of the algorithm yields $r = 1 - \exp\left(\frac{\log(1-\pi)\log(\lambda_{\max}(\beta))}{\log(\beta)}\right)$. If α and κ are not too large, we can get a simplified expression for the rate, characterized as follows.

Corollary 4. *Let ${\Delta^\star_t}_{T_{\text{in}}+s} := \Delta^s_t + \Gamma_t$. If $\kappa < \frac{1}{2}$ and $\alpha = \frac{1}{12}$, if $T_{\text{in}} \ge \frac{\log(\alpha\kappa)}{\log(1-\pi)}$, then, there exist a constant $C' > 0$ such that after a total of s inner updates, we have*

$$\mathbb{E}[\Delta^\star_s] \le C' \left(1 - \frac{\kappa\pi}{2\log(12/\kappa)}\right)^s.$$

6.2 Convergence results with SDCA

Given the structure of D_ρ, if the functions f^*_c in (7) are strongly convex, a good candidate for $\mathcal{A}$ is stochastic dual coordinate ascent (SDCA). Indeed, the results of Shalev-Shwartz and Zhang (2016) show that

Proposition 1. *If $\mathcal{A}$ is SDCA, let $|\mathcal{C}|$ be the total number of cliques, σ_c the strong convexity constant of f^*_c, and L_c the Lipschitz constant of $\mu_c \mapsto r(\mu)$, then $\mathcal{A}$ is linearly convergent with rate $\pi = \min_{c \in \mathcal{C}} \frac{\sigma_c}{|\mathcal{C}|(\sigma_c + L_c)}$.*

Moreover SDCA allows us to bound the duality gap by the increase of D_ρ, which yields linear convergence in the primal.

Proposition 2. *Let $\hat{w}^t = w(\hat{\mu}^t)$. If $\mathcal{A}$ is SDCA, then*

$$\mathbb{E}[P(\hat{w}^t) - P(w^\star)] \le \frac{1}{\pi}\mathbb{E}[\hat{\Delta}_t] + \mathbb{E}[\Gamma_t].$$

For the sake of the natural surrogates for the entropy (like the Gini-OTRW entropy proposed in Appendix C), individual functions f^*_c are not strongly convex, although $-D_\rho$ is strongly convex, because the entropy surrogate is strongly concave on $\mathcal{L}$ and the term $\|A\mu\|^2$ is strongly convex on $\text{Ker}(A)^\perp$. In that case another decomposition is relevant: if σ is the strong convexity constant of $-D_\rho$, then let $\tilde{f}^*_c(\mu_c) = \iota_{\triangle_c}(\mu_c) + \sigma\|\mu_c\|^2_2$ and $\tilde{r}(\mu) = -H_{\text{Approx}}(\mu) + r(\mu) - \sigma\|\mu_c\|^2_2$. We again have $D_\rho(\mu) = -\sum_{c \in \mathcal{C}} \tilde{f}^*_c(\mu_c) - \tilde{r}(\mu)$, with $\tilde{f}^*_c$ strongly convex and $\tilde{r}$ convex and smooth. SDCA and its theory are here applicable again and guarantees that Proposition 1 and following hold. However, for the convergence in the primal a slightly different argument is needed.

Proposition 3. *Let $w^{t,s} = w(\mu^{t,s})$. If $\mathcal{A}$ is a linearly convergent algorithm and the function $\mu \mapsto -H_{\text{approx}} + \frac{1}{2\rho}\|A\mu\|^2_2$ is strongly convex, then $P(w^{t,s}) - P(w^\star)$ converges to 0 linearly.*

6.3 Discussion

Optimization with inexact gradients (Devolder et al., 2014) and inexact proximal operators (Schmidt et al., 2011) have been shown to yield the same convergence rate as their exact counterparts, provided that errors decrease at a certain rate. Linear convergence of an inexact augmented Lagrangian method in which both inner and outer optimizations use Nesterov's accelerated gradient descent is shown in Lan and Monteiro (2016). We use the same ideas, except that we leverage the large finite sum structure of the dual problem to use randomized algorithms. The use of warm-start is also similar to its use in the meta-algorithm proposed by Lin et al. (2017), who use inexact gradient descent on the Moreau-Yosida regularization of a non-smooth objective. In our context, this approach would actually be applicable by working on $P_\rho(w, \xi)$ instead of working in the dual. An investigation in this direction is of interest but beyond the scope of this paper.

7 Experiments

We evaluate our algorithm IDAL on three different CRF models including 1) a simulated Gaussian mixture Potts model with grid graph and two clique types (nodes and edges); 2) a semantic segmentation model with planar graph and two clique types (nodes and edges); 3) a multi-label classification model with fully-connected graph and unique clique type for all cliques.

We compare with algorithms using only clique-wise oracles for solving $\min_\xi \max_\mu D_\rho(\mu, \xi)$, namely, the soft-constrained block-coordinate Frank-Wolfe algorithm (SoftBCFW) by Meshi et al. (2015b) and the greedy direction method of multipliers (GDMM) algorithm by Yen et al. (2016). Note that SoftBCFW in fact solves only the special case $\max_\mu D_\rho(\mu, \xi \equiv 0)$, thus it will converge to a different point than IDAL. In addition, we include a third baseline for the special case using SDCA (referred as SoftSDCA). Since SoftBCFW and GDMM have been shown outperforming other baselines such as Lacoste-Julien et al. (2013), Meshi et al. (2010)

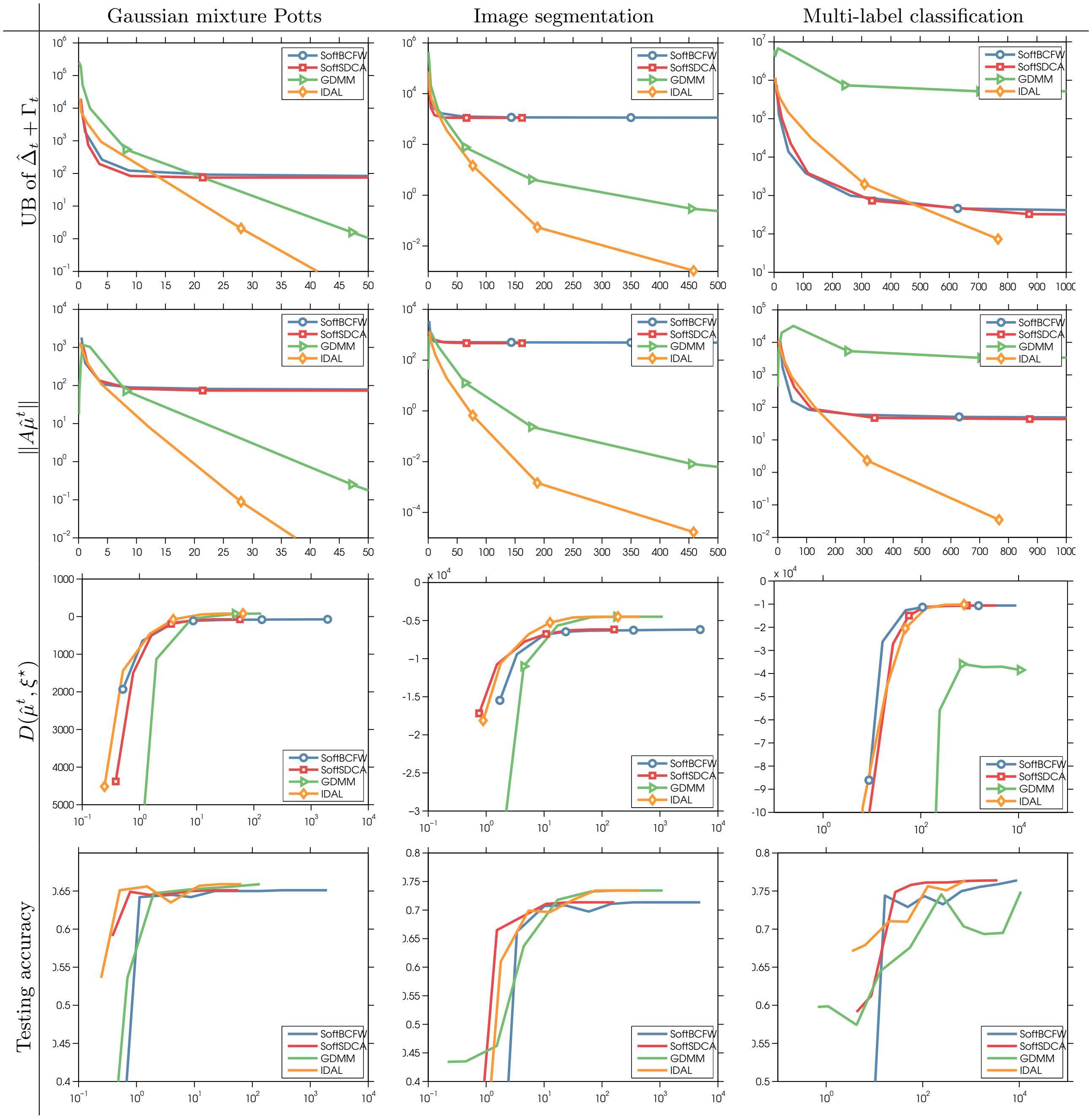

Figure 1: The comparison between IDAL and other baselines. For the choices of hyperparameters in terms of accuracy and speed, we set $(\lambda = 10, \rho = 1, \gamma = 1)$ for Gaussian mixture Potts, $(\lambda = 10, \rho = 1, \gamma = 10)$ for semantic segmentation and $(\lambda = 1, \rho = 0.1, \gamma = 1)$ for multi-label classification. The x-axis is running time in seconds.

and Hazan and Urtasun (2010), we will not make an extensive comparison for all these algorithms.

7.1 Setup

Gaussian mixture Potts models This is an extension of the Potts model given observations, whose conditional density function is defined via Bayes' rule $p(y|x) \propto p(x|y)p(y)$, with $p(y)$ a Potts distribution associated with a grid graph and parameterized by $w_{\text{binary}} \in \mathbb{R}^{k^2}$, and with $p(x|y) = \prod_i p(x_i|y_i)$ assumed to factorize into independent conditional Gaussian distributions with canonical parameters $w_{\text{unary}} \in \mathbb{R}^{2k}$, i.e., $p(x_i|y_i) \propto \exp(\langle w_{\text{unary}}(y_i), [x_i, x_i^2] \rangle)$. We consider a 10×10 grid graph with node cardinality $k = 5$. To generate the data, we first draw the label y from $p(y)$, and then the observation x_i is generated from the conditional Gaussian $p(x_i|y_i)$ for each node. The simulated dataset contains 100 samples and is equally divided for training and testing.

Semantic image segmentation We consider a typical CRF model used in computer vision for labeling image pixels with semantic classes. The graph is built upon clustering pixels into superpixels. Each superpixel defines a node. Two superpixels with a shared boundary define an edge. The CRF model takes the form $p(y|x) \propto \exp\big(\sum_i w_{\text{unary}}^\intercal \psi_i(x, y_i) + \sum_{i,j} w_{\text{binary}}^\intercal \psi_{ij}(x, y_i, y_j)\big)$, where $\psi_i(x, y_i)$ measure the intra-cluster compatibility within the superpixel i, and $\psi_{ij}(x, y_i, y_j)$ measure the inter-cluster compatibility between superpixels i and j. We conduct the experiment on the *MSRC-21* dataset introduced by Shotton et al. (2006), which has 21 classes, 335 training images and 256 testing images.

Multi-label classification The task for this problem is assigning each input vector a set of binary target labels. It is natural to model the inter-label dependencies by CRFs that treat each label as a node in a fully connected label graph. Following Finley and Joachims (2008), we define the CRF density function as $p(y|x) \propto \exp(\sum_i w_i^\intercal \phi_i(x, y_i) + \sum_{i,j} w_{ij}^\intercal \phi_{ij}(y_i, y_j))$, where the feature maps are specified as $\phi_i(x, y_i) = y_i \otimes x$ for each node and $\phi_{ij}(y_i, y_j) = y_i \otimes y_j$ for each edge. We conduct the experiments on the *Yeast* dataset[2], which contains 1500 training samples and 917 testing samples. Each sample has 14 labels and 103 attributes.

Hyperparameters In theory, T_{in} could be very large depending on the choice of α and the condition number π. We find that in practice only a relatively small T_{in} is needed. We empirically choose $T_{\text{in}} = \frac{1}{2}|\mathcal{C}|$. We set the number of outer iterations $T_{\text{ex}} = 3000$ and the stopping threshold $\epsilon = 10^{-3}$. The ranges of λ is pre-defined as $\{10, 1.0, 0.01, 0.001\}$ and the range of γ is $\{100.0, 10.0, 1.0, 0.001\}$. For each experiment, we choose the best λ and γ in terms of the validation accuracy and a reasonable running time (not all experiments finished in 3000 outer iterations). We set $\rho = 1.0$ or $\rho = 0.1$ as in Meshi et al. (2015b).

7.2 Results

To compare IDAL with GDMM, we use the criterion $P_\rho(\hat{w}^t, \hat{\delta}^t, \xi^t) - D_\rho(\hat{\mu}^t, \xi^t) + P_\rho(\hat{w}^t, \hat{\delta}^t, \xi^t) - D_\rho(\bar{\mu}^{T_{\text{ex}}}, \xi^{T_{\text{ex}}})$, which is an upper bound of the theoretical quantity $\hat{\Delta}_t + \Gamma_t$ that we analyzed. To compare IDAL with SoftBCFW, since $\xi = 0$ for SoftBCFW, we use the criterion $D_\rho(\hat{\mu}^t, \xi^\star)$, in which $\xi^\star$ is obtained from running IDAL to convergence. Besides, we also use the criteria $\|A\hat{\mu}^t\|^2$ (it measures the convergence of $d(\xi)$, since $\nabla d(\xi^t) \simeq A\hat{\mu}^t$) and the testing accuracy, which are applicable for all three algorithms. The results are shown in Figure 1.

There are several interesting points that we can say based on the results: 1) by tightening the marginalization constraints $A\mu = 0$, it does help to gain a better testing accuracy (e.g., IDAL gains small improvements over SoftBCFW); 2) based on the curves of $D_\rho(\mu, \xi^\star)$, we can see that it is key to approach $\mu^\star$ by first obtaining $\xi^\star$, which again shows the importance of enforcing exactness of the local consistency polytope; 3) IDAL is shown to be a faster algorithm than GDMM. One possible reason is that GDMM is in fact an active-set algorithm, which means the number of updated cliques at very beginning is insufficient comparing to IDAL. Based on our analysis, we have shown that the quality of the approximate gradient $A\hat{\mu}_t$ depends on T_{in}. Therefore, it is very likely that GDMM suffers from a slow convergence because of the poor gradients.

8 Conclusion

We proposed a relaxed dual augmented Lagrangian formulation for CRF learning, in which, thanks to dual decomposition, SDCA can be used to partially optimize over mean parameters in order to yield a sufficiently good approximation of the multiplier gradient. Our theoretical analysis shows that if warm-starts are leveraged and multiplier gradients are approximated with a linearly convergent algorithm, global linear convergence can be obtained. If SDCA is used, linear convergence is obtained both in the primal and for the convergence of the dual Lagrangian method.

Comparing to other baselines such as GDMM and SoftBCFW, our algorithm is faster in terms of the distance to the optimal objective function value (i.e. $\hat{\Delta}_t + \Gamma_t$) and the feasibility of the constraints $\|A\mu\|_2^2$.

It would be of interest to investigate the use of the same dual augmented Lagrangian formulation for both inference and learning, since according to Wainwright (2006), this should improve the performance.

In future work, we intend to investigate applications to other problems in machine learning, the use of Nesterov acceleration or quasi-Newton methods for multiplier updates, or the connection to other approaches based on Moreau-Yosida regularization.

Acknowledgments

We would like to thank Nikos Komodakis for useful discussions at early stages of this work. This research is partially supported by the CSTB and by ANR CHORUS research grant 13-MONU-0005-10.

[2] http://sourceforge.net/projects/mulan/files/datasets/yeast.rar

References

Bertsekas, D. P. (1982). The method of multipliers for equality constraints. In *Constrained optimization and Lagrange Multiplier methods*. Athena scientific.

Bertsekas, D. P. (1999). Nonlinear programming.

Collins, M., Globerson, A., Koo, T., Carreras, X., and Bartlett, P. L. (2008). Exponentiated gradient algorithms for conditional random fields and max-margin Markov networks. *JMLR*, 9:1775–1822.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654.

Devolder, O., Glineur, F., and Nesterov, Y. (2014). First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75.

Finley, T. and Joachims, T. (2008). Training structural SVMs when exact inference is intractable. In *International Conference on Machine Learning (ICML)*, pages 304–311.

Globerson, A. and Jaakkola, T. (2007). Convergent propagation algorithms via oriented trees. In *UAI*, pages 133–140.

Hazan, T. and Urtasun, R. (2010). A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, pages 838–846.

Hong, M., Chang, T.-H., Wang, X., Razaviyayn, M., Ma, S., and Luo, Z.-Q. (2014). A block successive upper bound minimization method of multipliers for linearly constrained convex optimization. *arXiv preprint arXiv:1401.7079*.

Hong, M. and Luo, Z.-Q. (2017). On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199.

Komodakis, N. (2011). Efficient training for pairwise or higher order CRFs via dual decomposition. In *CVPR*, pages 1841–1848.

Komodakis, N., Paragios, N., and Tziritas, G. (2007). MRF optimization via dual decomposition: Message-passing revisited. In *ICCV*, pages 1–8.

Kulesza, A. and Pereira, F. (2007). Structured learning with approximate inference. In *Advances in Neural Information Processing Systems*, pages 785–792.

Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P. (2013). Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, pages 53–61.

Lan, G. and Monteiro, R. D. (2016). Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547.

Lin, H., Mairal, J., and Harchaoui, Z. (2017). QuickeNing: A generic quasi-Newton algorithm for faster gradient-based optimization. *arXiv preprint arXiv:1610.00960*.

London, B., Huang, B., and Getoor, L. (2015). The benefits of learning with strongly convex approximate inference. In *ICML*, pages 410–418.

Martins, A. F. T., Figueiredo, M. A. T., Aguiar, P. M. Q., Smith, N. A., and Xing, E. P. (2015). AD3: Alternating directions dual decomposition for MAP inference in graphical models. *JMLR*, 16:495–545.

Meshi, O., Mahdavi, M., and Schwing, A. G. (2015a). Smooth and strong: MAP inference with linear convergence. In *NIPS*, pages 298–306.

Meshi, O., Sontag, D., Globerson, A., and Jaakkola, T. S. (2010). Learning efficiently with approximate inference via dual losses. In *ICML*, pages 783–790.

Meshi, O., Srebro, N., and Hazan, T. (2015b). Efficient training of structured SVMs via soft constraints. In *AISTATS*, pages 699–707.

Nutini, J., Schmidt, M., Laradji, I., Friedlander, M., and Koepke, H. (2015). Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *ICML*, pages 1632–1641.

Pletscher, P., Ong, C. S., and Buhmann, J. M. (2010). Entropy and margin maximization for structured output learning. In *ECML*, pages 83–98.

Roux, N. L., Schmidt, M., and Bach, F. R. (2012). A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671.

Savchynskyy, B., Kappes, J., Schmidt, S., and Schnörr, C. (2011). A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling. In *CVPR*, pages 1817–1823.

Schmidt, M., Babanezhad, R., Ahmed, M., Defazio, A., Clifton, A., and Sarkar, A. (2015). Non-uniform stochastic average gradient method for training conditional random fields. In *AIStats*, pages 819–828.

Schmidt, M., Le Roux, N., and Bach, F. R. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*, pages 1458–1466.

Shalev-Shwartz, S. and Zhang, T. (2016). Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1-2):105–145.

Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2006). Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*. Springer.

Sontag, D., Meltzer, T., Globerson, A., Jaakkola, T., and Weiss, Y. (2008). Tightening LP relaxations for MAP using message passing. In *UAI*, pages 503–510.

Tang, K., Ruozzi, N., Belanger, D., and Jebara, T. (2016). Bethe learning of graphical models via MAP decoding. In *AIStats*, pages 1096–1104.

Wainwright, M. J. (2006). Estimating the wrong graphical model: Benefits in the computation-limited setting. *JMLR*, 7(Sep):1829–1859.

Wainwright, M. J. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. (2005). A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312.

Yen, I. E.-H., Huang, X., Zhong, K., Zhang, R., Ravikumar, P. K., and Dhillon, I. S. (2016). Dual decomposed learning with factorwise oracle for structural SVM of large output domain. In *NIPS*, pages 5024–5032.