Paired-Dual Learning for Fast Training of Latent Variable Hinge-Loss MRFs

Stephen H. Bach[*] University of Maryland, College Park, MD
Bert Huang[*] Virginia Tech, Blacksburg, VA
Jordan Boyd-Graber University of Colorado, Boulder, CO
Lise Getoor University of California, Santa Cruz, CA
[*] Equal contributors.

Abstract

Latent variables allow probabilistic graphical models to capture nuance and structure in important domains such as network science, natural language processing, and computer vision. Naive approaches to learning such complex models can be prohibitively expensive—because they require repeated inferences to update beliefs about latent variables—so lifting this restriction for useful classes of models is an important problem. Hinge-loss Markov random fields (HL-MRFs) are graphical models that allow highly scalable inference and learning in structured domains, in part by representing structured problems with continuous variables. However, this representation leads to challenges when learning with latent variables. We introduce *paired-dual learning*, a framework that greatly speeds up training by using tractable entropy surrogates and avoiding repeated inferences. Paired-dual learning optimizes an objective with a pair of dual inference problems. This allows fast, joint optimization of parameters and dual variables. We evaluate on social-group detection, trust prediction in social networks, and image reconstruction, finding that paired-dual learning trains models as accurate as those trained by traditional methods in much less time, often before traditional methods make even a single parameter update.

1. Introduction

Latent variables can capture structure in complicated domains and have been used extensively in social and biological network analysis, Web analytics, computer vision, and many other domains that study large-scale, structured data. However, including latent variables sacrifices scalability for expressiveness because the values of latent variables are—by definition—unknown. Algorithms for learning with latent variables often require repeated inference to iteratively update parameters, and each inference alone can be expensive for a large model. For example, inference methods like Gibbs sampling and belief propagation require many iterations to converge, and learning methods like EM alternate between fully inferring latent variable values and updating parameters.

Latent variables are particularly valuable in rich, structured models, but the computational costs become even more challenging. Our contribution is a new learning framework for rich, structured, continuous latent-variable models that addresses this computational bottleneck. Our focus is on hinge-loss Markov random fields (HL-MRFs) (Bach et al., 2013b), a class of probabilistic graphical models that makes large-scale maximum a posteriori (MAP) inference highly efficient by representing structured domains with continuous variables. These models have been successfully applied to user attribute (Li et al., 2014) and trust (Huang et al., 2013; West et al., 2014) prediction in social networks, natural language semantics (Beltagy et al., 2014), and drug discovery (Fakhraei et al., 2014). Researchers have also begun to train HL-MRFs with latent variables for tasks such as group detection in social media (Bach et al., 2013a), online-education analytics (Ramesh et al., 2014), and automobile-traffic modeling (Chen et al., 2014). Like other approaches to learning with latent variables, these applications repeatedly solve inference problems to convergence for *each* update of the parameters. Removing this bottleneck is critical for retaining the existing scalability benefits of HL-MRFs when training with latent variables.

Overcoming the need for repeated inference requires contending with challenges that arise from a continuous representation, including the need for efficient alternatives to representing distributions over uncountable state spaces and evaluating irreducible integrals. For fully-supervised learning, large-margin methods can use

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

the dual of loss-augmented inference to form a joint convex minimization (Taskar et al., 2005; Meshi et al., 2010). Schwing et al. (2012) extended this idea to latent-variable learning for discrete MRFs, using a method specifically formulated to pass messages corresponding to the discrete states of the variables. While these methods are incompatible with continuous models, dualization is also a key to faster training of continuous models with latent variables.

In Section 3, we propose *paired-dual learning*, a framework that quickly trains HL-MRFs with latent variables by avoiding repeated inferences. Traditional methods for learning with latent variables require repeated inferences for two distributions to compute gradients. The unobserved random variables are grouped into two sets, those with training labels and those without, i.e., the latent variables. One distribution is joint over the labeled variables and the latent variables, and the other is over the latent variables conditioned on the labels. Paired-dual learning uses an equivalent variational learning objective that substitutes dual problems for the two corresponding inference problems, augmented with entropy surrogates to make the learning problem well-formed. We describe how to design suitable entropy surrogates that retain the useful properties of entropy while still admitting fast HL-MRF inference. We can therefore compute the gradient of the paired-dual learning objective with respect to the parameters using the intermediate states of inference, enabling a fast, block-coordinate joint optimization.

We show in Section 4 that paired-dual learning drastically reduces the time required for learning without sacrificing accuracy on three real-world problems: social-group detection, trust prediction in social networks, and image reconstruction. Paired-dual learning cuts training time by as much as 90%, often converging before traditional methods make a single update to the parameters.

2. Background

In this section, we review hinge-loss MRFs, the class of models for which we derive paired-dual learning. We also give an overview of MAP inference and variational learning with latent variables, which will serve as foundations for our framework.

2.1. Hinge-Loss MRFs

HL-MRFs are Markov random fields with hinge-loss potential functions defined over continuous variables.

Definition 1. *Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a vector of n variables and $\boldsymbol{x} = (x_1, \ldots, x_{n'})$ a vector of n' variables with joint domain $[0,1]^{n+n'}$. Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$ be m continuous potentials of the form $\phi_i(\boldsymbol{y}, \boldsymbol{x}) = (\max\{\ell_i(\boldsymbol{y}, \boldsymbol{x}), 0\})^{p_i}$, where ℓ_i is a linear function of $\boldsymbol{y}$ and $\boldsymbol{x}$ and $p_i \in \{1, 2\}$. Given a vector of nonnegative free parameters, i.e., weights, $\boldsymbol{w} = (w_1, \ldots, w_m)$, a* **hinge-loss Markov random field** *P over $\boldsymbol{y}$ and conditioned on $\boldsymbol{x}$ is a probability density function*

$$P(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{x}; \boldsymbol{w})} \exp\left(-\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})\right) ;$$
$$Z(\boldsymbol{x}; \boldsymbol{w}) = \int_{\boldsymbol{y}} \exp\left(-\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})\right) \, d\boldsymbol{y} .$$

HL-MRFs are very expressive. Hinge functions can model logic-like implications, in which one variable should be greater than another, and correlations, in which two variables are preferred to be close in value, by adding two hinge-loss potentials to make a distance function. The exponent p_i specifies the loss family.

2.2. MAP Inference for Hinge-Loss MRFs

HL-MRFs admit exact, highly scalable MAP inference that optimizes a dual to the inference objective, which is constructed via techniques called *consensus optimization* and the *alternating direction method of multipliers*, or ADMM (Boyd et al., 2011, and references therein). This dual problem is substituted into the learning objective to derive paired-dual learning, so we review it in this subsection. The convexity of the potentials and the non-negativity of the weights make MAP inference for HL-MRFs the following convex optimization:

$$\arg\max_{\boldsymbol{y} \in [0,1]^n} P(\boldsymbol{y}|\boldsymbol{x}; w) \equiv \arg\min_{\boldsymbol{y}} \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) . \quad (1)$$

For HL-MRFs, this objective can be solved efficiently using the following formulation as a consensus optimization. Let $f(\boldsymbol{v}) = \sum_{i=1}^m w_i \cdot \phi_i(\boldsymbol{v}^i)$ be a separable function with components corresponding to the potentials in Equation 1, and let $\boldsymbol{v} = \{\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m\}$ consist of local copies for variables $\boldsymbol{y}$ so that each term in $f(\boldsymbol{v})$ is a function of disjoint components of $\boldsymbol{v}$. To make the optimization over $f(\boldsymbol{v})$ equivalent to MAP inference, let $\bar{\boldsymbol{v}}$ be a vector of n consensus variables, each corresponding to entries in the HL-MRF variable vector $\boldsymbol{y}$, and let a consensus function $\mathbf{c}(\boldsymbol{v}, \bar{\boldsymbol{v}})$ be a linear operator that outputs a vector of differences between each pair of corresponding components of $\boldsymbol{v}$ and $\bar{\boldsymbol{v}}$. For example, the element $\mathbf{c}(\boldsymbol{v}, \bar{\boldsymbol{v}})_{(i,j)}$ is the difference between consensus variable $\bar{v}_j$ and its i-th local copy v^i_j. The function $\mathbf{c}$ can be viewed as the violations for the constraint that the local variables equal their corresponding consensus variables. Finally, let each component of $\boldsymbol{v}$ and $\bar{\boldsymbol{v}}$ be real valued and introduce a constraint function on $\bar{\boldsymbol{v}}$, $g(\bar{\boldsymbol{v}})$, which is 0 if $\bar{\boldsymbol{v}} \in [0,1]^n$ and ∞ otherwise. Then Equation 1 is equivalent to

$$\arg\min_{\boldsymbol{v}, \bar{\boldsymbol{v}}} f(\boldsymbol{v}) + g(\bar{\boldsymbol{v}}) \quad \text{such that} \quad \mathbf{c}(\boldsymbol{v}, \bar{\boldsymbol{v}}) = \mathbf{0} .$$

This consensus optimization formulation can be solved efficiently with ADMM, which provides strong convergence guarantees. ADMM relaxes the equality constraints of consensus optimization by introducing dual variables $\boldsymbol{\alpha}$, with one entry for each dimension of $\mathbf{c}(\boldsymbol{v}, \bar{\boldsymbol{v}})$, and forming the *augmented Lagrangian*

$$L(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}}) = f(\boldsymbol{v}) + g(\bar{\boldsymbol{v}}) + \boldsymbol{\alpha}^\top \mathbf{c}(\boldsymbol{v}, \bar{\boldsymbol{v}}) + \frac{\eta}{2} \|\mathbf{c}(\boldsymbol{v}, \bar{\boldsymbol{v}})\|^2$$

where $\eta > 0$ is a user-specified parameter. By alternating maximization of L with respect to $\boldsymbol{v}$ and $\bar{\boldsymbol{v}}$, and then updating $\boldsymbol{\alpha}$, ADMM converges to a MAP assignment to the HL-MRF variables $\boldsymbol{y} = \bar{\boldsymbol{v}}^\star$. For HL-MRF potentials, these updates can be done efficiently (Bach et al., 2013b).

2.3. Variational Learning with Latent Variables

Paired-dual learning quickly optimizes a standard learning objective, which we review in this subsection. When learning models with latent variables, the usual goal is to maximize the marginal likelihood of the labels $\hat{\boldsymbol{y}}$ given observed variables $\boldsymbol{x}$, marginalizing out over all possible configurations of latent variables $\boldsymbol{z}$. For a parameter setting w and any state of the latent variables $\boldsymbol{z}$, the log marginal likelihood can be expressed as a log ratio of joint and conditional likelihoods, which simplifies to the difference of two normalizing partition functions:

$$\log P\left(\hat{\boldsymbol{y}} | \boldsymbol{x}; w\right) = \log Z(\boldsymbol{x}, \hat{\boldsymbol{y}}; w) - \log Z(\boldsymbol{x}; w) \;.$$

Each of these partition functions has a variational form (Wainwright & Jordan, 2008), yielding the identity

$$\begin{aligned} &\log Z(\boldsymbol{x}, \hat{\boldsymbol{y}}; \boldsymbol{w}) - \log Z(\boldsymbol{x}; \boldsymbol{w}) \\ &= \min_{\rho \in \Delta(\boldsymbol{y}, \boldsymbol{z})} \max_{q \in \Delta(\boldsymbol{z})} \mathbb{E}_\rho \left[\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\right] - H(\rho) \\ &\qquad - \mathbb{E}_q \left[\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{z})\right] + H(q) \;, \end{aligned} \tag{2}$$

where ρ is a joint distribution over the $\boldsymbol{y}$ and $\boldsymbol{z}$ variables from the space of all joint distributions $\Delta(\boldsymbol{y}, \boldsymbol{z})$, q is a conditional distribution over the the $\boldsymbol{z}$ variables from the space of all conditional distributions $\Delta(\boldsymbol{z})$, and H is the entropy.

Using the variational form, Equation 2, regularized maximum likelihood is the following saddle-point optimization:

$$\begin{aligned} &\arg\min_{\boldsymbol{w}} \max_{\rho \in \Delta(\boldsymbol{y}, \boldsymbol{z})} \min_{q \in \Delta(\boldsymbol{z})} \\ &\frac{\lambda}{2} \|\boldsymbol{w}\|^2 - \mathbb{E}_\rho \left[\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\right] + H(\rho) \\ &\qquad + \mathbb{E}_q \left[\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{z})\right] - H(q) \end{aligned} \tag{3}$$

where $\lambda \geq 0$ is a tunable regularization parameter.[1] We solve the learning problem in its variational form because it enables principled approximations of intractable problems by restricting the spaces of distributions $\Delta(\boldsymbol{y}, \boldsymbol{z})$ and $\Delta(\boldsymbol{z})$.

A traditional approach for optimizing Equation 3 computes subgradients of the outer minimization over $\boldsymbol{w}$ by exactly solving the inner min-max and differentiating. Another approach iteratively solves the conditional inference over $\boldsymbol{z}'$, fixes $\boldsymbol{z}'$, and solves the remaining min-max over $\boldsymbol{w}$ and $\boldsymbol{y}, \boldsymbol{z}$ as a fully-observed maximum-likelihood estimation.[2] Each of these approaches performs a block coordinate ascent-descent that requires fully solving two (or more) inferences per iteration of the outer optimization.

3. Paired-Dual Learning

In this section, we present paired-dual learning, a framework for training HL-MRFs with latent variables. Optimizing the variational learning objective, Equation 3, is intractable because the expectations and entropies are irreducible integrals. Traditional methods approximate the objective by restricting the variational distributions ρ and q to tractable families, and we adopt this approach as well. However, traditional methods fit and refit ρ and q exactly before each update of the parameters $\boldsymbol{w}$. Paired-dual learning speeds up training by interleaving updates of $\boldsymbol{w}$ into dual optimizations over ρ and q. Dualizing these inference problems allows training to use the intermediate solutions produced by ADMM. To enable this interleaved joint optimization, we first construct surrogates for the entropy functions $H(\rho)$ and $H(q)$ so that, when the variational families $\Delta(\boldsymbol{y}, \boldsymbol{z})$ and $\Delta(\boldsymbol{z})$ are restricted to point estimates, fitting the distributions ρ and q is subsumed by MAP inference, while still preserving the desired properties of entropy functions in learning. To optimize over the model parameters $\boldsymbol{w}$, we consider the ADMM duals of both variational inference problems, forming a new saddle-point objective that can be differentiated with respect to $\boldsymbol{w}$ during intermediate stages of ADMM.

3.1. Tractable Entropy Surrogates

As with many continuous models, optimizing Equation 3 exactly for HL-MRFs is intractable because the expectations and the entropies are irreducible integrals. To remove this intractability, we first adopt the common approximation of restricting $\Delta(\boldsymbol{y}, \boldsymbol{z})$ and $\Delta(\boldsymbol{z})$ to tractable families of variational distributions. We restrict the variational families to point distributions, enabling highly scalable MAP inference techniques to optimize over them. In other words, the minimizing distribution $\rho^\star$ places all probability on the point $(\boldsymbol{y}, \boldsymbol{z})$ that minimizes $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - H(\rho)$, and

[1] We use L2 regularization in our derivations and experiments, but paired-dual learning is easily adapted to include any regularization function whose subdifferentials are computable.

[2] This strategy is equivalent to variational expectation maximization (EM), or “hard” EM if using point distributions, and it generalizes the standard approach for latent structured SVM.

$q^\star$ places all probability on the point $\boldsymbol{z}$ that minimizes $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{z}) - H(q)$. Moreover, the entropies $H(\rho)$ and $H(q)$ are always zero for point distributions, so finding $\rho^\star$ and $q^\star$ for a particular $\boldsymbol{w}$ are instances of MAP inference.

Using this approximation alone, Equation 3 always has a degenerate global optimum at $\boldsymbol{w} = \boldsymbol{0}$. This degeneracy reveals the importance of having nontrivial entropy terms to reward high-entropy states. To remove this degenerate solution, we need to include tractable surrogates for the entropies in Equation 3 that behave as the true entropies should: biasing the objective away from the labeled state so that stronger weights are necessary to produce good predictions. Therefore, the surrogate entropy and the weight-norm regularization will have opposite effects, removing the degenerate zero solution.

We can preserve this non-degeneracy effect without complicating MAP inference by choosing hinge functions as entropy surrogates and treating them as potentials with fixed weights. For example, if a HL-MRF variable y represents the degree to which a person is in each of two latent groups—with $y = 0.0$ being completely in a group and $y = 1.0$ being completely in the other—then, the following pair of squared-hinge potentials can act as a suitable entropy surrogate for the point distribution at y:

$$-w \left(\max\{y, 0\}^2 + \max\{1 - y, 0\}^2 \right) .$$

This entropy surrogate penalizes solutions where y deviates from 0.5, making the learning objective prefer models strong enough to push y towards one extreme. During learning, the associated parameter w is fixed, but during MAP inference the surrogate can be treated as another pair of hinge potentials, preserving the scalability of inference.

The function that acts as a surrogate does not need a probabilistic interpretation, and the appropriate choice of these surrogates can generalize the objectives of latent structured SVM (LSSVM) (Yu & Joachims, 2009) and variants of expectation maximization (EM). The LSSVM objective uses a loss between the current prediction $\boldsymbol{y}$ and the labels $\hat{\boldsymbol{y}}$ as a surrogate for $H(\rho)$ and no surrogate, i.e., 0, for $H(q)$. The ℓ_1 loss function can be represented with simple hinge functions, enabling HL-MRF inference (Bach et al., 2013b). We discuss these connections further in Section 5.

Let h be any surrogate entropy of point distributions. The tractable latent variable HL-MRF learning objective is

$$\begin{aligned}\arg\min_{\boldsymbol{w}} \; \max_{\boldsymbol{y},\boldsymbol{z}} \; \min_{\boldsymbol{z}'} \\ \frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) + h(\boldsymbol{y}, \boldsymbol{z}) \\ + \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{z}') - h(\hat{\boldsymbol{y}}, \boldsymbol{z}') \; .\end{aligned} \tag{4}$$

3.2. Joint Optimization

The traditional approaches involving repeatedly performing complete inference, i.e., finding $\boldsymbol{y}$, $\boldsymbol{z}$, and $\boldsymbol{z}'$ in Equation 4, can be very expensive in large-scale settings. Instead, we derive a method that exploits that HL-MRF inference can be solved via ADMM. In particular, this method enables optimization using partial solutions to inference. That is, the optimization can proceed before the inference optimization completes its computation.

We form a new joint optimization by rewriting Equation 4 with the corresponding augmented Lagrangians used to solve the inner optimizations. Let $L_{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}})$ be the augmented Lagrangian for minimizing $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - h(\boldsymbol{y}, \boldsymbol{z})$. We subscript the augmented Lagrangian with the parameters $\boldsymbol{w}$ to emphasize that it is also a function of the current parameters. Let $L'_{\boldsymbol{w}}(\boldsymbol{v}', \boldsymbol{\alpha}', \bar{\boldsymbol{v}}')$ be the analogous augmented Lagrangian for minimizing $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{z}') - h(\hat{\boldsymbol{y}}, \boldsymbol{z}')$. Substituting them into Equation 4, we write the equivalent paired-dual learning objective:

$$\begin{aligned}\arg\min_{\boldsymbol{w}} \; \max_{\boldsymbol{v},\bar{\boldsymbol{v}}} \; \min_{\boldsymbol{\alpha}} \; \min_{\boldsymbol{v}',\bar{\boldsymbol{v}}'} \; \max_{\boldsymbol{\alpha}'} \\ \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + L'_{\boldsymbol{w}}(\boldsymbol{v}', \boldsymbol{\alpha}', \bar{\boldsymbol{v}}') - L_{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}}) \; .\end{aligned} \tag{5}$$

Since the inner optimizations are guaranteed to converge to the global optima for fixed $\boldsymbol{w}$ (Boyd et al., 2011), Equations 4 and 5 are identical. With this view, we no longer need to solve the optimizations to completion as they appear in the primal Equations 4. Instead, a finer-grained block-coordinate optimization over the variables that appear in the paired-dual Equation 5, interleaving subgradient steps over $\boldsymbol{w}$ and ADMM iterations over the other variables, reaches an optimum more quickly.

This objective is non-convex, and determining whether any block-coordinate optimization scheme for it will converge is an open question. If the inner optimizations were solved to convergence between updates of $\boldsymbol{w}$, then the optimization provably converges as an instance of the concave-convex procedure (Yuille & Rangarajan, 2003), in the same manner as LSSVM (Yu & Joachims, 2009). Schwing et al. (2012) derived a convergent algorithm for training discrete Markov random fields with latent variables that dualizes the optimization over (discrete) $\boldsymbol{y}$ and $\boldsymbol{z}$ and interleaves updating the corresponding dual variables and the parameters $\boldsymbol{w}$—while still solving the optimization over $\boldsymbol{z}'$ to convergence at each iteration. This algorithm updates beliefs over discrete variables but is not applicable to the continuous, non-linear potentials of HL-MRFs. While no guarantees for paired-dual learning are known, it always converges in our diverse experiments (see Section 4).

Algorithm 1 Paired-Dual Learning

```
Input: model $P(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{x}; \boldsymbol{w})$, labeled data $\hat{\boldsymbol{y}}$,
    initial parameters $\boldsymbol{w}$
Form augmented Lagrangian $L_{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}})$
    for $\arg\min_{\boldsymbol{z},\boldsymbol{y}} \; \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - h(\boldsymbol{y}, \boldsymbol{z})$
Form augmented Lagrangian $L'_w(\boldsymbol{v}', \boldsymbol{\alpha}', \bar{\boldsymbol{v}}')$
    for $\arg\min_{\boldsymbol{z}'} \; \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{z}') - h(\hat{\boldsymbol{y}}, \boldsymbol{z}')$
for $t$ from 1 to $T$ do
    for $n$ from 1 to $N$ or until converged do
        $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \eta \mathbf{c}(\boldsymbol{v}, \bar{\boldsymbol{v}})$
        $\boldsymbol{v} \leftarrow \arg\min_{\boldsymbol{v}} L_{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}})$
        $\bar{\boldsymbol{v}} \leftarrow \arg\min_{\bar{\boldsymbol{v}}} L_{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}})$
    end for
    for $n$ from 1 to $N$ or until converged do
        $\boldsymbol{\alpha}' \leftarrow \boldsymbol{\alpha}' + \eta \mathbf{c}'(\boldsymbol{v}', \bar{\boldsymbol{v}}')$
        $\boldsymbol{v}' \leftarrow \arg\min_{\boldsymbol{v}'} L'_{\boldsymbol{w}}(\boldsymbol{v}', \boldsymbol{\alpha}', \bar{\boldsymbol{v}}')$
        $\bar{\boldsymbol{v}}' \leftarrow \arg\min_{\bar{\boldsymbol{v}}'} L'_{\boldsymbol{w}}(\boldsymbol{v}', \boldsymbol{\alpha}', \bar{\boldsymbol{v}}')$
    end for
    if $t > K$ then
        $\nabla_{\boldsymbol{w}} \leftarrow \nabla_{\boldsymbol{w}} \left[ \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + L'_{\boldsymbol{w}}(\boldsymbol{v}', \boldsymbol{\alpha}', \bar{\boldsymbol{v}}') - L_{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}}) \right]$
        Update $\boldsymbol{w}$ via $\nabla_{\boldsymbol{w}}$
    end if
end for
```

3.3. Learning Algorithm

The complete learning algorithm is summarized in Algorithm 1. We first construct the augmented Lagrangian $L_w(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}})$ for MAP inference in $P(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{x}; w)$ and the analogous augmented Lagrangian $L'_w(\boldsymbol{v}', \boldsymbol{\alpha}', \bar{\boldsymbol{v}}')$ for inference in $P(\boldsymbol{z} | \boldsymbol{x}, \hat{\boldsymbol{y}}; w)$, as described in Section 2.2. Then, at each iteration t, we first execute ADMM iterations, which update the Lagrangian $L_w(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}})$ by taking a step in the dual space over the variables $\boldsymbol{\alpha}$, then optimizing $\boldsymbol{v}$, and finally optimizing $\bar{\boldsymbol{v}}$. We limit ADMM to N iterations before moving on, where N is a user-specified parameter.[3] In our experiments, we found that higher values result in slower training, and in Section 4, we discuss results that suggest setting $N = 1$, i.e., single updates of all variables, provides the best speed and accuracy.

We then update the other Lagrangian $L'_{\boldsymbol{w}}(\boldsymbol{v}', \boldsymbol{\alpha}', \bar{\boldsymbol{v}}')$. At the end of each iteration t, we update $\boldsymbol{w}$ via the derivative of the joint objective, Equation 5. The gradients $\nabla_{\boldsymbol{w}}$ for $L_{\boldsymbol{w}}$ and $L'_{\boldsymbol{w}}$ are straightforward. The gradient for a potential ϕ is the potential function value at the current setting of the local copies $\boldsymbol{v}$ and $\boldsymbol{v}'$. This computation only differs from how one computes the gradient in the primal setting in that it is evaluated for variable copies that might not agree during this intermediate stage. Since the weights $\boldsymbol{w}$ do not interact with any of the dual terms in the augmented Lagrangian, these terms do not affect the gradient.

Naive interleaving of learning with inference could be implemented with early stopping and warm starting of ADMM inference. Without the paired-dual view, one could use the gradient of the primal objective using the consensus variables $\bar{\boldsymbol{v}}$ and $\bar{\boldsymbol{v}}'$ (or some other estimate of the inference variables), but these gradients would not correspond to Equation 5, or to any principled objective function. Instead, the paired-dual learning objective enables joint optimization of a principled objective, with gradient computations no more complicated than in the primal setting.

Finally, one can "warm up" the ADMM variables by updating $\boldsymbol{v}$, $\boldsymbol{\alpha}$, $\bar{\boldsymbol{v}}$, $\boldsymbol{v}'$, $\boldsymbol{\alpha}'$, and $\bar{\boldsymbol{v}}'$ for a few iterations before beginning to update the parameters $\boldsymbol{w}$. Setting warm-up parameter K greater than zero can improve the initial search direction for $\boldsymbol{w}$ by reducing the gap between the paired-dual gradient and the ADMM approximation for the initial setting of $\boldsymbol{w}$. In our experiments (Section 4), $K = 0$ often suffices, but for one task, using $K = 10$ produces a better start to optimization. The cost of this warmup is negligible, since learning often requires hundreds of ADMM iterations, but the benefits of taking a better initial gradient step can be significant in practice.

Variants of paired-dual learning easily fit into this framework. We can stop after a fixed number of iterations or when $\boldsymbol{w}$ has converged. We can transparently apply existing strategies for smoother gradient-based optimization, e.g., adaptive rescaling (Duchi et al., 2011) or averaging.

4. Experiments

In this section, we evaluate paired-dual learning by comparing it with traditional learning methods on real-world problems. We test two variants of paired-dual learning: the finest grained interleaving with only two ADMM iterations per weight update ($N = 1$) and a coarser grained 20 ADMM iterations per update ($N = 10$). We compare with *primal subgradient*, which evaluates subgradients of Equation 4 by solving the inner optimizations to convergence ($N = \infty$), and *expectation maximization* (EM), which fits the parameters via multiple subgradient descent steps for each point estimate of the latent variables $\boldsymbol{z}'$.

We consider three problems that publications have addressed using HL-MRFs: group detection in social media, social-trust prediction, and image reconstruction. For each problem, we build HL-MRFs that include latent variables and surrogate entropies, run each learning algorithm, and

[3] If $L_w(\boldsymbol{v}, \boldsymbol{\alpha}, \bar{\boldsymbol{v}})$ converges for the current setting of $\boldsymbol{w}$, we terminate the inner loop early. Therefore, each inner loop performs between 1 and N ADMM iterations at each outer iteration t. See Appendix F for more on ADMM convergence criteria.

evaluate on held-out test data. The iterations of ADMM constitute most of the computational cost during learning, so we measure the quality of the learned models as a function of the number of ADMM iterations taken during learning. Since each ADMM step is exactly the same amount of computation, regardless of the learning algorithm or the current model, the number of ADMM steps represents the computational cost, avoiding confounding factors such as differences in hardware used in these experiments. During each outer iteration of each algorithm, we store the current weights and later use these weights offline to measure the primal objective, Equation 4, and predictive performance on held out data. We provide high-level details on each experiment and defer additional details to the appendix.

For all four methods, we update weights using a standard subgradient descent approach for large-scale MRFs (e.g., Lowd & Domingos, 2007), in which we take steps in the direction dictated by the subgradient, scaled by the number of potentials sharing each weight, and return the final average weight vector over all iterations of learning. EM and primal subgradient solve inference problems to convergence for each update of the parameters, but we warm-start them at each iteration from the optima for the previous iteration to avoid artificially inflating their running times.

Discovering Latent Groups in Social Media Groups of people can form online around common traits, interests, or opinions. Often these groups are not explicitly defined in social media, but can be discovered by modeling group membership as latent variables that depend on user behavior. To test paired-dual learning on this task, we use the data of Bach et al. (2013a), who collected roughly 4.275M tweets from about 1.350M Twitter users, from a 48-hour window around the Venezuelan presidential election on Oct. 7, 2012. We model the supporters of the two candidates by introducing two latent groups.

We use a learning setup based on that of Bach et al. (2013a), who build a model that relates language usage and social interactions to latent group membership. The 20 most retweeted users in the data are considered *top users*. Others that interacted with a top user and used at least one hashtag are *regular users*, whose group affiliation are latent.

We construct HL-MRFs by introducing squared hinge-loss dependencies between each user's latent group and each hashtag, and each user's latent group to each top user. We then introduce dependencies between pairs of regular users for each online interaction they shared. These dependencies among users' latent groups makes the task a single, joint structured prediction. We treat hashtag usage and interactions with non-top users as observations $\boldsymbol{x}$, interactions with top users as labeled targets $\boldsymbol{y}$, and latent group membership as latent variables $\boldsymbol{z}$. The dependencies share parameters such that there is a parameter for each hashtag-group pair and each group-top-user interaction pair. We evaluate each model's ability to predict interactions with top users, measuring the area under the precision recall curve (AuPR) using ten folds of cross-validation. In this experiment, we set $K = 0$, immediately starting learning.

Paired-dual learning optimizes the objective value significantly faster than all other methods, and this faster optimization translates to the faster learning of a more accurate model on test data. In fact, the curves for primal subgradient and EM begin at their first parameter updates, so paired-dual learning reaches a high quality model before the primal methods have learned anything. The top row of Figure 4 plots the objective and AuPR for one fold and a scatter plot of the AuPR on all ten folds for a subset of the points. Full results are in Appendix B.

Latent User Attributes in Trust Networks HL-MRFs have recently been shown to be state-of-the-art tools for social-trust prediction, the task of predicting directed trust relationships between pairs of users in social networks. Huang et al. (2013) showed that HL-MRFs representing social psychological theories produce more accurate joint trust predictions than existing methods specifically designed for trust prediction. We augment their model, which is based on the social theory of structural balance, by using latent variables to model the user attributes of trustworthiness and willingness to trust. We describe here the additional latent variables and dependencies.

We introduce two latent attributes for each user, "trusting" and "trustworthy." We then introduce dependencies between each trusting property and all possible outgoing trust relationships in which the corresponding user participates, and between each trustworthy property and all possible incoming trust relationships. Full details on the model are in Appendix C. These latent properties act as aggregators, modeling the trends in each user's trust relationships.

We evaluate on a subsample of roughly 2,000 users of Epinions.com (Huang et al., 2013; Richardson et al., 2003). The task is to predict user-user trust ratings given the observed social network and partial observation of ratings. We again set $K = 0$ and perform eight-fold cross-validation, and we plot the objective and AuPR curves for held-out distrust relationships from one fold and a scatter plot of the AuPR for a subset of the points for all folds. (We show results for distrust relationships because they account for roughly 10% of all relationships and are therefore harder to predict with high precision and recall.)

The results again show a faster objective descent for paired-dual learning, which learns a high-accuracy model well before the other methods begin learning. Though it is not the purpose of our experiments, it is interesting to note that the

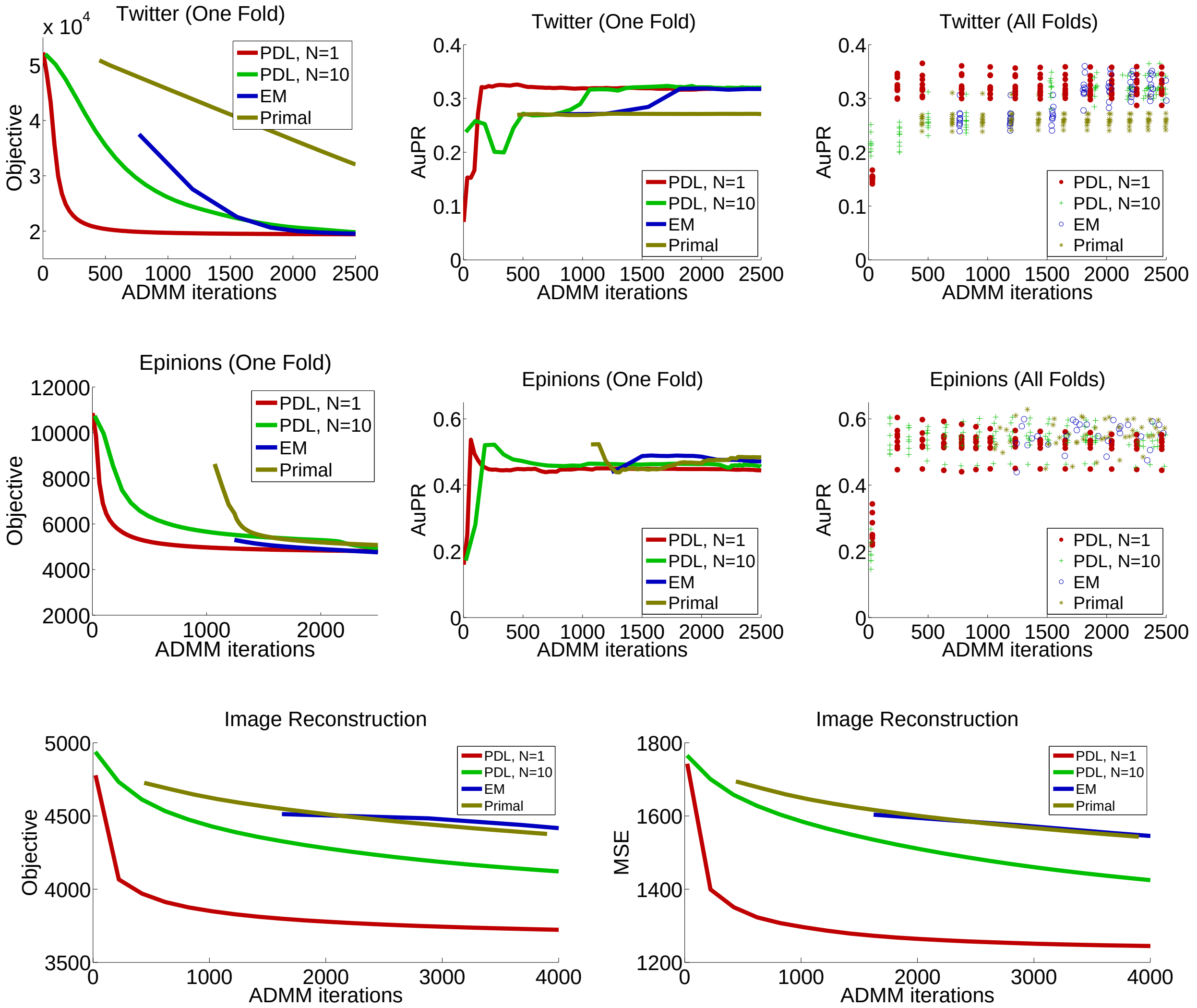

Figure 1. Objective score and performance with respect to ADMM iterations for one fold, as well as a subset of points for all folds. On all three problems—group detection, trust prediction, and image reconstruction—paired-dual learning (PDL) reduces the primal learning objective and improves predictive performance much faster than expectation maximization (EM) or primal subgradient (Primal), often reaching a good model before the existing algorithms complete their first parameter update. Full results are in Appendices B, C, and D.

AuPR scores on this data set are substantially better than those achieved in previous work using fully-supervised HL-MRFs. The full results are in Appendix C.

Image Reconstruction Reconstructing part of an obstructed image requires some amount of semantic understanding of physical objects that images depict. These latent semantics make it an ideal test setting for latent variable modeling. We follow the experimental setup of previous papers (Poon & Domingos, 2011; Bach et al., 2013b). Using the 400-image Olivetti face data set, we reveal the top half of each face image to the prediction algorithm, and task it with predicting the bottom half. Bach et al. (2013b) used fully-observed learning to fit non-latent, or "flat", HL-MRFs to this task, which were able to reconstruct images with mean-squared error comparable to the best latent-variable methods. These flat models had a large number of parameters for potentials between neighboring pixels and "mirror-image" pixels. Examining the outputs from these HL-MRFs reveals that the models relied heavily on trivial structural patterns, such as face symmetry. This reliance is especially obvious in the reconstructions by flat HL-MRFs for bottom-halves of faces, which seemed to mimic the shadows of mouths by reflecting blurry images of top-half eyes. Latent variables improve performance by learning actual facial structures, rather than exploiting trivial patterns. With all the parameters, variables, and dependencies in the model for each pixel, the efficiency of paired-dual learning becomes critical.

We use a simpler HL-MRF with a latent layer. We include squared hinge-loss potentials between six latent state variables and the input-half pixel intensities, rounded versions of the input pixels, and, finally, the output-half intensities. These potentials allow the values of the latent variables to mediate interactions between the inputs and outputs. We additionally include potentials between each latent state that prefer contiguous regions of latent states, a prior potential for each pixel to learn an average or background value, and a quadratic prior on all free variables, which serves as a surrogate entropy. The full model is listed in Appendix D. We omit any direct dependencies between output pixels to isolate the effectiveness of latent variable modeling.

We train on 50 randomly selected images from the first 350, and test on the last 50 images, as was done previously. Because of the higher dimensionality of these pixel-based models, we set $K = 10$, allowing the ADMM variables to warm up before updating the parameters $\boldsymbol{w}$. (These warmup ADMM iterations are included in the plots above.)

Again, paired-dual learning with one iteration of ADMM is significantly faster at optimizing the objective, which directly translates to a reduction in test error, while the primal methods and the more conservative 10-iteration paired-dual approach are much slower to improve the objective. The learned latent variable model fits latent states to archetypal face shapes, as visualized in Appendix D.

5. Related Approaches for Discrete Models

There exist many approaches to learning discrete, discriminative models with latent variables. Existing classes of probabilistic models include *hidden-unit conditional random fields* (van der Maaten et al., 2011), a class of undirected graphical models similar to linear conditional random fields, except that a latent variable mediates the interaction between each observation and target variable on the chain. This restricted structure allows the latent variables to be marginalized out during inference and learning but cannot express more complex dependencies. More expressive discriminative models have been trained via specialized inference algorithms designed for specific models (e.g., Kok & Domingos, 2007; Poon & Domingos, 2009). Another class of probabilistic models are *sum-product networks* (Poon & Domingos, 2011), or SPNs, which represent distributions as networks of sum and product operations. Interior nodes in an SPN have a natural interpretation as latent variables, and SPNs can be trained with EM.

The variational objective, Equation 4, relates to several important ideas in probabilistic inference and latent variable learning. For discrete MRFs, surrogates enable efficient and accurate inference (e.g., Heskes, 2006; Weiss et al., 2007; Wainwright & Jordan, 2008; Meshi et al., 2009). Especially for learning, no statistical interpretation of the surrogates is necessary. For example, using the family of point distributions and replacing the entropy with a distance metric between the point and the labels, we obtain the objective for LSSVM (Yu & Joachims, 2009). Similarly, using point expectations and using null surrogates, i.e., $h(\rho) = 0$, the objective becomes analogous to structured perceptron (Collins, 2002; Richardson & Domingos, 2006). Lastly, using tractable families of distributions for both the expectation and the entropies makes the learning objective that of variational EM (Neal & Hinton, 1999).

Replacing inference problems with duals to speed up learning has also been explored for discrete models. For fully-supervised settings, Taskar et al. (2005) dualize the loss-augmented inference problem as part of large-margin learning, making a joint quadratic program. Meshi et al. (2010) improve on this approach to use dual decomposition for LP relaxations of inference in discrete graphical models. Schwing et al. (2012) extend this idea to latent-variable models. By dualizing one of the two inference subroutines and passing messages corresponding to the discrete states, they speed up learning of discrete models with latent variables. Related to this line of work, Domke (2013) use dualization as part of a technique to reduce structured prediction to non-structured logistic regression.

The same principles behind paired-dual learning can be adapted for discrete models, and we are investigating the benefits of dualizing both inferences, as opposed to just one, as well as whether useful message-passing algorithms exist for the paired-dual objective in discrete models.

6. Conclusion

This paper presents a new framework for fast training of latent variable HL-MRFs. This contribution addresses a variety of challenges that arise in the training of these powerful continuous models. While traditional latent variable learning methods require full inferences to compute gradients of the learning objective, paired-dual learning evaluates gradients using incomplete dual inference optimizations. Therefore, it can learn without the expensive cost of repeated, full inference. We demonstrate our approach on a variety of real-world data sets, which show that paired-dual learning is able to train accurate models in a small fraction of the time required by traditional algorithms. This substantial speedup for training richly structured, continuous models with latent variables will further enable their application to large-scale, high-impact problems.

Acknowledgments

This work is supported by NSF grant IIS1218488 and IARPA via DoI/NBC contract number D12PC00337.

Boyd-Graber is supported by NSF grants IIS1320538, IIS1409287, and NCSE1422492. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Bach, S. H., Huang, B., and Getoor, L. Learning latent groups with hinge-loss Markov random fields. In *ICML Workshop on Inferning: Interactions between Inference and Learning*, 2013a.

Bach, S. H., Huang, B., London, B., and Getoor, L. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Uncertainty in Artificial Intelligence*, 2013b.

Beltagy, I., Erk, K., and Mooney, R. J. Probabilistic soft logic for semantic textual similarity. In *Annual Meeting of the Association for Computational Linguistics*, 2014.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 2011.

Chen, P., Chen, F., and Qian, Z. Road traffic congestion monitoring in social media with hinge-loss Markov random fields. In *IEEE International Conference on Data Mining (ICDM)*, 2014.

Collins, M. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *Empirical Methods in Natural Language Processing*, 2002.

Domke, J. Structured learning via logistic regression. In *Neural Information Processing Systems*, 2013.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

Fakhraei, S., Huang, B., Raschid, L., and Getoor, L. Network-based drug-target interaction prediction with probabilistic soft logic. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2014.

Heskes, T. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Journal of Artificial Intelligence Research*, 26:153–190, 2006.

Huang, B., Kimmig, A., Getoor, L., and Golbeck, J. A flexible framework for probabilistic models of social trust. In *Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction*, 2013.

Kok, S. and Domingos, P. Statistical predicate invention. In *International Conference on Machine Learning*, 2007.

Li, J., Ritter, A., and Jurafsky, D. Inferring user preferences by probabilistic logical reasoning over social networks. arXiv:1411.2679 [cs.SI], 2014.

Lowd, D. and Domingos, P. Efficient weight learning for Markov logic networks. In *Principles and Practice of Knowledge Discovery in Databases*, 2007.

Meshi, O., Jaimovich, A., Globerson, A., and Friedman, N. Convexifying the Bethe free energy. In *Uncertainty in Artificial Intelligence*, 2009.

Meshi, O., Sontag, D., Jaakkola, T., and Globerson, A. Learning efficiently with approximate inference via dual losses. In *International Conference on Machine Learning*, 2010.

Neal, R. and Hinton, G. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. (ed.), *Learning in graphical models*, pp. 355–368. MIT Press, 1999.

Poon, H. and Domingos, P. Unsupervised semantic parsing. In *Conference on Empirical Methods in Natural Language Processing*, 2009.

Poon, H. and Domingos, P. Sum-product networks: A new deep architecture. In *Uncertainty in Artificial Intelligence*, 2011.

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, Hal, and Getoor, L. Learning latent engagement patterns of students in online courses. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

Richardson, M. and Domingos, P. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

Richardson, M., Agrawal, R., and Domingos, P. Trust management for the semantic web. In Fensel, D., Sycara, K., and Mylopoulos, J. (eds.), *The Semantic Web - ISWC 2003*, volume 2870 of *Lecture Notes in Computer Science*, pp. 351–368. Springer Berlin / Heidelberg, 2003.

Schwing, A. G., Hazan, T., Pollefeys, M., and Urtasun, R. Efficient structured prediction with latent variables for general graphical models. In *International Conference on Machine Learning*, 2012.

Taskar, B., Chatalbashev, V., Koller, D., and Guestrin, C. Learning structured prediction models: A large margin approach. In *International Conference on Machine Learning*, 2005.

van der Maaten, L., Welling, M., and Saul, L. Hidden-unit conditional random fields. In *Artificial Intelligence and Statistics*, 2011.

Wainwright, M. and Jordan, M. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2), January 2008.

Weiss, Y., Yanover, C., and Meltzer, T. MAP estimation, linear programming and belief propagation with convex free energies. In *Uncertainty in Artificial Intelligence*, 2007.

West, R., Paskov, H. S., Leskovec, J., and Potts, C. Exploiting Social Network Structure for Person-to-Person Sentiment Analysis. *Transactions of the Association for Computational Linguistics (TACL)*, 2:297–310, 2014.

Yu, C. and Joachims, T. Learning structural SVMs with latent variables. In *International Conference on Machine Learning*, 2009.

Yuille, A. L. and Rangarajan, A. The concave-convex procedure. *Neural Comput.*, 15(4):915–936, 2003.