Bethe Projections for Non-Local Inference

Luke Vilnis[*] UMass Amherst luke@cs.umass.edu

David Belanger[*] UMass Amherst belanger@cs.umass.edu

Daniel Sheldon UMass Amherst sheldon@cs.umass.edu

Andrew McCallum UMass Amherst mccallum@cs.umass.edu

Abstract

Many inference problems in structured prediction are naturally solved by augmenting a tractable dependency structure with complex, non-local auxiliary objectives. This includes the mean field family of variational inference algorithms, soft- or hard-constrained inference using Lagrangian relaxation or linear programming, collective graphical models, and forms of semi-supervised learning such as posterior regularization. We present a method to discriminatively *learn* broad families of inference objectives, capturing powerful non-local statistics of the latent variables, while maintaining tractable and provably fast inference using non-Euclidean projected gradient descent with a distance-generating function given by the Bethe entropy. We demonstrate the performance and flexibility of our method by (1) extracting structured citations from research papers by learning soft global constraints, (2) achieving state-of-the-art results on a widely-used handwriting recognition task using a novel learned non-convex inference procedure, and (3) providing a fast and highly scalable algorithm for the challenging problem of inference in a collective graphical model applied to bird migration.

1 INTRODUCTION

Structured prediction has shown great success in modeling problems with complex dependencies between output variables. Practitioners often use undirected graphical models, which encode conditional dependency relationships via a graph. However, the tractability of exact inference in these models is limited by the graph's *treewidth*, often yielding a harsh tradeoff between model expressivity and tractability.

[*] Equal contribution.

Graphical models are good at modeling local dependencies between variables, such as the importance of surrounding context in determining the meaning of words or phrases. However, their sensitivity to cyclic dependencies often renders them unsuitable for modeling preferences for certain globally consistent states. For example, in the canonical NLP task of part-of-speech tagging, there is no clear way to enforce the constraint that every sentence have at least one verb without increasing the likelihood that *every* token is predicted to be a verb.

Concretely, exact marginal inference in a discrete graphical model can be posed as the following optimization problem

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu} \in \mathcal{M}}{\arg\min} \; -H(\boldsymbol{\mu}) - \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \,, \tag{1}$$

where $\boldsymbol{\mu}$ is a concatenated vector of node and clique marginals, $H(\boldsymbol{\mu})$ is the entropy, $\mathcal{M}$ is the marginal polytope, and $\boldsymbol{\theta}$ are parameters. Here we face a tradeoff: adding long-range dependencies directly to the model increases the clique size and thus the complexity of the problem and size of $\boldsymbol{\mu}$, rendering inference intractable. However, the linear scoring function $\boldsymbol{\theta}$ breaks down over cliques, preventing us from enforcing global regularities in any other way. In this work, we propose to augment the inference objective (1) and instead optimize

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu} \in \mathcal{M}}{\arg\min} -H(\boldsymbol{\mu}) - \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle + L_{\boldsymbol{\psi}}(\boldsymbol{\mu}). \tag{2}$$

Here, $L_{\boldsymbol{\psi}}$ is some arbitrary parametric function of the entire concatenated marginal vector, where $\boldsymbol{\psi}$ may depend on input features. Since $L_{\boldsymbol{\psi}}$ is non-linear, it can enforce many types of non-local properties. Interestingly, whenever $L_{\boldsymbol{\psi}}$ is convex, and whenever inference is easy in the underlying model, i.e., solving (1) is tractable, we can solve (2) using non-Euclidean projected gradient methods using the Bethe entropy as a distance-generating function. Unlike many message-passing algorithms, our procedure maintains primal feasibility across iterations, allowing its use as an *anytime* algorithm. Furthermore, for non-convex $L_{\boldsymbol{\psi}}$, we also show convergence to a local optimum of (2). Finally, we

present algorithms for discriminative learning of the parameters $\boldsymbol{\psi}$. In a slight abuse of terminology, we call $L_{\boldsymbol{\psi}}$ a *non-local energy function*.

Ours is not the first work to consider modeling global preferences by augmenting a tractable base inference objective with non-local terms. For example, generalized mean-field variational inference algorithms augment a tractable distribution (the Q distribution) with a non-linear, non-convex global energy function that scores terms in the full model (the P distribution) using products of marginals of Q (Wainwright & Jordan, 2008). This is one special case of our non-local inference framework, and we present algorithms for solving the problem for much more general $L_{\boldsymbol{\psi}}$, with compelling applications.

Additionally, the modeling utility provided by global preferences has motivated work in *dual decomposition*, where inference in loopy or globally-constrained models is decomposed into repeated calls to inference in tractable independent subproblems (Komodakis et al., 2007; Sontag et al., 2011). It has seen wide success due to its ease of implementation, since it reuses existing inference routines as black boxes. However, the technique is restricted to modeling linear constraints, imposed *a priori*. Similarly, these types of constraints have also been imposed on expectations of the posterior distribution for use in semi-supervised learning, as in *posterior regularization* and *generalized expectation* (Ganchev et al., 2010; Mann & McCallum, 2010). In contrast, our methods are designed to discriminatively learn expressive inference procedures, with minimal domain knowledge required, rather than regularizing inference and learning.

First, we provide efficient algorithms for solving the marginal inference problem (2) and performing MAP prediction in the associated distribution, for both convex and non-convex global energy functions. After that, we provide a learning algorithm for $\boldsymbol{\theta}$ and the parametrized $L_{\boldsymbol{\psi}}$ functions using an interpretation of (2) as approximate variational inference in a a probabilistic model. All of our algorithms are easy to implement and rely on simple wrappers around black-box inference subroutines.

Our experiments demonstrate the power and generality of our approach by achieving state-of-the-art results on several tasks. We extract accurate citations from research papers by learning discriminative global regularities of valid outputs, outperforming a strong dual decomposition-based baseline (Anzaroot et al., 2014). In a benchmark OCR task (Taskar et al., 2004), we achieve state-of-the-art results with a learned non-convex, non-local energy function, that guides output decodings to lie near dictionary words. Finally, our general algorithm for solving (2) provides large speed improvements for the challenging task of inference in chain-structured *collective graphical models* (CGMs), applied to bird migration (Sheldon & Dietterich, 2011).

2 BACKGROUND

Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote a set of discrete variables and $\mathbf{x}$ be a collection of input features. We define the conditional distribution $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) = \exp(\langle \boldsymbol{\theta}(\mathbf{x}), S(\mathbf{y})\rangle)/Z$, where $S(\mathbf{y})$ is a mapping from $\mathbf{y}$ to a set of sufficient statistics, $\boldsymbol{\theta}(\mathbf{x})$ is a differentiable vector-valued mapping, and $Z = \sum_{\mathbf{y}} \exp(\langle \boldsymbol{\theta}, S(\mathbf{y})\rangle)$. Conditional random fields (CRFs) assume that $(y_1, \ldots, y_n)$ are given a graph structure and $S(\mathbf{y})$ maps $\mathbf{y}$ to a 0-1 vector capturing joint settings of each clique (Lafferty et al., 2001). Going forward, we often suppress the explicit dependency of $\boldsymbol{\theta}$ on $\mathbf{x}$. For fixed $\boldsymbol{\theta}$, the model is called a Markov random field (MRF).

Given a distribution $P(\mathbf{y})$, define the expected sufficient statistics operator $\mu(P) = \mathbb{E}_P[S(\mathbf{y})]$. For the CRF statistics $S(\mathbf{y})$ above, $\boldsymbol{\mu}$ is a concatenated vector of node and clique marginals. Therefore, *marginal inference*, the task of finding the marginal distribution of $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$ over $\mathbf{y}$, is equivalent to computing the expectation $\mu(P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}))$.

For tree-structured graphical models, $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) \longleftrightarrow \boldsymbol{\mu}(P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}))$ is a bijection, though this is not true for general graphs. Furthermore, for trees the entropy $H(P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}))$ is equal to the Bethe entropy $H_{\mathcal{B}}(\boldsymbol{\mu}(P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})))$, defined, for example, in Wainwright & Jordan (2008). The *marginal polytope* $\mathcal{M}$ is the set of $\boldsymbol{\mu}$ that correspond to some $P_{\boldsymbol{\theta}}$.

As mentioned in the introduction, marginal inference can be posed as the optimization problem (1). MAP inference finds the joint setting $\mathbf{y}$ with maximum probability. For CRFs, this is equivalent to

$$\arg\min_{\mathbf{y}} \langle -\boldsymbol{\theta}(\mathbf{x}), S(\mathbf{y})\rangle \,. \tag{3}$$

For tree-structured CRFs, marginal and MAP inference can be performed efficiently using dynamic programming. Our experiments focus on such graphs. However, the inference algorithms we present can be extended to general graphs wherever marginal inference is tractable using a convex entropy approximation and a local polytope relaxation.

3 MARGINAL INFERENCE WITH NON-LOCAL ENERGIES

We move beyond the standard inference objective (1), augmenting it with a non-local energy term as in (2):

$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu}\in\mathcal{M}} -H_{\mathcal{B}}(\boldsymbol{\mu}) - \langle \boldsymbol{\theta}, \boldsymbol{\mu}\rangle + L_{\boldsymbol{\psi}}(\boldsymbol{\mu}).$$

Here, $L_{\boldsymbol{\psi}}$ is some arbitrary parametrized function of the marginals, and $\boldsymbol{\psi}$ may depend on input features $\mathbf{x}$.

Intuitively, we are augmenting the inference objective (1) by allowing it to optimize a broader set of tradeoffs – not

only between expected node scores, clique scores, and entropy, but also global functions of the marginals. To be concrete, in our citation extraction experiments (Section 8.1), for example, we employ the simple structure:

$$L_{\boldsymbol{\psi}}(\boldsymbol{\mu}) = \sum_j \boldsymbol{\psi}_j \ell_j(\boldsymbol{\mu}), \tag{4}$$

Where each ℓ_j is a univariate convex function and each $\boldsymbol{\psi}_j$ is constrained to be non-negative, in order to maintain the overall convexity of $L_{\boldsymbol{\psi}}$. We further employ

$$\ell_j(\boldsymbol{\mu}) = \tilde{\ell}_j \left(a_j^\top \boldsymbol{\mu}\right), \tag{5}$$

where a_j encodes a 'linear measurement' of the marginals and $\tilde{\ell}_j$ is some univariate convex function.

4 VARIATIONAL INTERPRETATION AND MAP PREDICTION

We next provide two complementary interpretations of (2) as variational inference in a class of tractable probability distributions over $\mathbf{y}$. They yield precisely the same variational expression. However, both are useful because the first helps motivate a MAP prediction algorithm, while the second helps characterize our learning algorithm in Section 7 as (approximate) variational EM.

Proposition 1. *For fixed $\boldsymbol{\theta}$ and $L_{\boldsymbol{\psi}}$, the output $\boldsymbol{\mu}^*$ of inference in the augmented objective* (2) *is equivalent to the output of standard inference* (1) *in an MRF with the same clique structure as our base model, but with a modified parameter $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \nabla L_{\boldsymbol{\psi}}(\boldsymbol{\mu}^*)$.*

Proof. Forming a Lagrangian for (2), the stationarity conditions with respect to the variable $\boldsymbol{\mu}$ are:

$$0 = -(\boldsymbol{\theta} - \nabla L_{\boldsymbol{\psi}}(\boldsymbol{\mu}^*)) - \nabla H_{\mathcal{B}}(\boldsymbol{\mu}^*) + \nabla_{\boldsymbol{\mu}} C(\boldsymbol{\mu}, \boldsymbol{\lambda}), \tag{6}$$

where $C(\boldsymbol{\mu}, \boldsymbol{\lambda})$ are collected terms relating to the marginal polytope constraints. The proposition follows because (6) is the same as the stationarity conditions for

$$\boldsymbol{\mu}^* = \operatorname*{arg\,min}_{\boldsymbol{\mu} \in \mathcal{M}} - \langle \boldsymbol{\theta} - \nabla L_{\boldsymbol{\psi}}(\boldsymbol{\mu}^*), \boldsymbol{\mu} \rangle - H_{\mathcal{B}}(\boldsymbol{\mu}). \ \square \tag{7}$$

Therefore, we can characterize a joint distribution over $\mathbf{y}$ by first finding $\boldsymbol{\mu}^*$ by solving (2) and then defining an MRF over $\mathbf{y}$ with parameters $\tilde{\boldsymbol{\theta}}$. Even more conveniently, our inference technique in Section 6 iteratively estimates $\tilde{\boldsymbol{\theta}}$ on the fly, namely via the dual iterate $\boldsymbol{\theta}_t$ in Algorithm 1.

Ultimately, in many prediction problems we seek a single output configuration $\mathbf{y}$ rather than an inferred distribution over outputs. Proposition 1 suggests a simple prediction procedure: first, find the variational distribution over $\mathbf{y}$ parametrized as an MRF with parameter $\tilde{\boldsymbol{\theta}}$. Then, perform MAP in this MRF. Assuming an available marginal inference routine for this MRF, we assume the tractability of MAP – for example using a dynamic program. We avoid predicting $\mathbf{y}$ by locally maximizing nodes' marginals, since this would not necessarily yield feasible outputs.

Instead of solving (2), we could have introduced global energy terms to the MAP objective (3) that act directly on values $S(\mathbf{y})$ rather than on expectations $\boldsymbol{\mu}$, as in (2). However, this yields a difficult combinatorial optimization problem for prediction and does not yield a natural way to learn the parametrization of the global energy. Section 8.1 demonstrates that using energy terms defined on marginals, and performing MAP inference in the associated MRF, performs as well or better than an LP technique designed to directly perform MAP subject to global penalty terms.

Our second variational interpretation characterizes $\boldsymbol{\mu}^*$ as a variational approximation to a complex joint distribution:

$$P_c(\mathbf{y}|\mathbf{x}) = (1/Z_{\boldsymbol{\theta},\boldsymbol{\psi}}) P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) P_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}). \tag{8}$$

We assume that isolated marginal inference in $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$ is tractable, while $P_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})$ is an alternative structured distribution over $\mathbf{y}$ for which we do not have an efficient inference algorithm. Specifically, we assume that (1) can be solved for $P_{\boldsymbol{\theta}}$. Furthemore, we assume that $P_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}) \propto \exp\left(L_{\boldsymbol{\psi}}(S(\mathbf{y}); \mathbf{x})\right)$, where $L_{\boldsymbol{\psi}}(\cdot; \mathbf{x})$ is a convex function, conditional on input features $\mathbf{x}$. Going forward, we will often surpress the dependence of $L_{\boldsymbol{\psi}}$ on $\mathbf{x}$. Above, $Z_{\boldsymbol{\theta},\boldsymbol{\psi}}$ is the normalizing constant of the combined distribution. Note that if L was linear, inference in both $P_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})$ and $P_c(\mathbf{y}|\mathbf{x})$ would be tractable, since the distribution would decompose over the same cliques as $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$.

Not surprisingly, (8) is intractable to reason about, due to the non-local terms in (2), so we approximate it with a variational distribution $Q(\mathbf{y})$. The connection between this variational approximation and Proposition 1 is derived in Appendix A. Here, we assume no clique structure on $Q(\mathbf{y})$, but show that minimizing a variational upper bound on $KL\left(Q(\mathbf{y})||P_c(\mathbf{y}|\mathbf{x})\right)$, for a given $\mathbf{x}$, yields a Q that is parametrized compactly as the MRF in Proposition 1. We discuss the relationship between this and general mean-field inference in Section 5.

Although the analysis of this section assumes convexity of $L_{\boldsymbol{\psi}}$, our inference techniques can be applied to non-convex $L_{\boldsymbol{\psi}}$, as discussed in Section 6.3, and our learning algorithm produces state-of-the-art results even in the non-convex regime for a benchmark OCR task.

5 RELATED MODELING TECHNIQUES

Mean field variational inference in undirected graphical models is a particular application of our inference framework, with a non-convex $L_{\boldsymbol{\psi}}$ (Wainwright & Jordan, 2008).

The technique estimates marginal properties of a complex joint distribution P using the clique marginals $\boldsymbol{\mu}$ of some tractable base distribution Q, not necessarily fully factorized. This induces a partitioning of the cliques of P into those represented directly by $\boldsymbol{\mu}$ and those where we define clique marginals as a product distribution of the relevant nodes' marginals in $\boldsymbol{\mu}$. To account for the energy terms of the full model involving cliques absent in the simple base model, the energy $\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle$ of the base model is augmented with an extra function of $\boldsymbol{\mu}$.

$$L(\mu) = -\sum_{c \in \mathcal{C}} \left\langle \boldsymbol{\theta}_c, \bigotimes_{n \in c} \mu_n \right\rangle \tag{9}$$

where $\mathcal{C}$ is the set of cliques not included in the tractable sub-model, $\boldsymbol{\theta}_c$ are the potentials of the original graphical model corresponding to the missing cliques, and $\bigotimes_n \mu_n$ represents a repeated outer (tensor) product of the node marginals for the nodes in those cliques.

Note $L(\boldsymbol{\mu})$ is non-linear and non-convex. Our work generalizes (9) by allowing arbitrary non-linear interaction terms between components of $\boldsymbol{\mu}$. This is very powerful – for example, in our citation extraction experiments in Section 8.1, expressing these global terms in a standard graphical model would require many factors touching all variables. Local coordinate ascent mean-field can be frustrated by these rigid global terms. Our gradient-based method avoids these issues by updating all marginals simultaneously.

Dual decomposition is a popular method for performing MAP inference in complex structured prediction models by leveraging repeated calls to MAP in tractable submodels (Komodakis et al., 2007; Sontag et al., 2011). The family of models solvable with dual decomposition is limited, however, because the terms that link the submodels must be expressible as linear constraints. Similar MAP techniques (Ravikumar et al., 2010; Aguiar et al., 2011; Fu & Banerjee, 2013) based on the alternating direction method of multipliers (**ADMM**) can be adapted for marginal inference, in problems where marginal inference in submodels is tractable. However, the non-local terms are defined as linear functions on settings of graphical model nodes, while our non-linear $L_{\boldsymbol{\psi}}(\boldsymbol{\mu})$ terms provide practitioners with an expressive means to learn and enforce regularities of the inference output.

Posterior regularization (PR) (Ganchev et al., 2010), **learning from measurements** (LFM) Liang et al. (2009) , and **generalized expectations** (GE) (Mann & McCallum, 2010), are a family of closely-related techniques for performing unsupervised or semi-supervised learning of a conditional distribution $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$ or a generative model $P_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{y})$ using expectation-maximization (EM), where the E-step for latent variables $\mathbf{y}$ does not come directly from inference in the model, but instead from projection onto a set of expectations obeying global regularity properties. In PR and GE, this yields a projection objective of the form (2), where the $L_{\boldsymbol{\psi}}$ terms come from a Lagrangian relaxation of regularity constraints, and $\boldsymbol{\psi}$ corresponds to dual variables. Originally, PR employed linear constraints on marginals, but He et al. (2013) extend the framework to arbitrary convex differentiable functions. Similarly, in LFM such an inference problem arises because we perform posterior inference assuming that the observations $\mathbf{y}$ have been corrupted under some noise model. Tarlow & Zemel (2012) also present a method for learning with certain forms of non-local losses in a max-margin framework.

Our goals are very different than the above learning methods. We do not impose non-local terms $L_{\boldsymbol{\psi}}$ in order to regularize our learning process or allow it to cope with minimal annotation. Instead, we use $L_{\boldsymbol{\psi}}$ to increase the expressivity of our model, performing inference for every test example, using a different $\boldsymbol{\psi}$, since it depends on input features. Since we are effectively 'learning the regularizer,' on fully-labeled data, our learning approach in Section 7 differs from these methods. Finally, unlike these frameworks, we employ non-convex $L_{\boldsymbol{\psi}}$ terms in some of our experiments. The algorithmic consequences of non-convexity are discussed in Section 6.3.

6 OPTIMIZING THE NON-LOCAL MARGINAL INFERENCE OBJECTIVE

We now present an approach to solving (2) using non-Euclidean projected gradient methods, which require access to a procedure for marginal inference in the base distribution (which we term the *marginal oracle*), as well as access to the gradient of the energy function $L_{\boldsymbol{\psi}}$. We pose these algorithms in the *composite minimization* framework, which gives us access to a wide variety of algorithms that are discussed in the supplementary material.

6.1 CONVEX OPTIMIZATION BACKGROUND

Before presenting our algorithms, we review several definitions from convex analysis (Rockafellar, 1997).

We call a function φ *σ-strongly convex* with respect to a norm $\|\cdot\|_P$, if for all $x, y \in \text{dom}(\varphi)$,

$$\varphi(y) \geq \varphi(x) + \nabla\varphi(x)^T(y - x) + \frac{\sigma}{2}\|y - x\|_P^2.$$

Proposition 2 (e.g. Beck & Teboulle (2003))**.** *The negative entropy function $-H(x) = \sum_i x_i \log x_i$ is 1-strongly convex with respect to the 1-norm $\|\cdot\|_1$ over the interior of the simplex Δ (restricting dom(H) to **int**(Δ)).*

Given a smooth and strongly convex function φ, we can also define an associated generalized (asymmetric) distance measure called the *Bregman divergence* (**?**) generated by φ,

$$B_\varphi(x, x_0) = \varphi(x) - \varphi(x_0) - \langle \nabla\varphi(x_0), x - x_0 \rangle .$$

Algorithm 1 Bethe-RDA

Input: parameters $\boldsymbol{\theta}$, energy function $L_{\boldsymbol{\psi}}(\mu)$
set $\boldsymbol{\theta}_0 = \boldsymbol{\theta}$
set μ_0 to prox-center MARGINAL-ORACLE($\boldsymbol{\theta}_0$)
$\bar{g}_0 = 0$
repeat
  $\beta_t =$ constant ≥ 0
  $\bar{g}_t = \frac{t-1}{t}\bar{g}_{t-1} + \frac{1}{t}\nabla L(\mu_t)$
  $\boldsymbol{\theta}_t = \boldsymbol{\theta} - \frac{t}{t+\beta_t}\bar{g}_t$
  $\mu_t =$ MARGINAL-ORACLE($\boldsymbol{\theta}_t$)
until CONVERGED(μ_t, μ_{t-1})

For example, the KL divergence is the Bregman divergence associated to the negative entropy function, and the squared Euclidean distance is its own associated divergence.

Composite minimization (**?**) is a family of techniques for minimizing functions of the form $h = f + R$, where we have an oracle that allows us to compute minimizations over R in closed form (usually R here takes the form of a regularizer). Problems of this form are often solved with an algorithm called *proximal gradient*, which minimizes $h(x)$ over some convex set X using:

$$x_{t+1} = \arg\min_{x \in X} \ \langle \nabla f(x_t), x \rangle + \frac{1}{2\eta_t}\|x - x_t\|_2^2 + R(x),$$

for some decreasing sequence of learning rates η_t. Note that because of the requirement $x \in X$, proximal gradient generalizes projected gradient descent – since unconstrained minimization might take us out of the feasible region X, computing the update requires projecting onto X.

But there is no reason to use the squared Euclidean distance when computing our updates and performing the projection. In fact, the squared term can be replaced by any Bregman divergence. This family of algorithms includes the *mirror descent* and *dual averaging* algorithms (Beck & Teboulle, 2003; Nesterov, 2009).

We base our projected inference algorithms on *regularized dual averaging* (RDA) (Xiao, 2010). The updates are:

$$x_{t+1} = \arg\min_{x \in X} \ \langle \bar{g}_t, x \rangle + \frac{\beta_t}{t}\varphi(x) + R(x), \qquad (10)$$

where $\bar{g}_t = \frac{1}{t}\sum_k^t \nabla f(x_k)$ is the average gradient of f encountered so far. One benefit of RDA is that it does not require the use of a learning rate parameter ($\beta_t = 0$) when using a strongly convex regularizer. RDA can be interpreted as doing a projection onto X using the Bregman divergence generated by the strongly convex function $\varphi + R$.

6.2 OUR ALGORITHM

These non-Euclidean proximal algorithms are especially helpful when we are unable to compute a projection in terms of Euclidean distance, but can do so using a different Bregman divergence. We will show that this is exactly the case for our problem of projected inference: the marginal oracle allows us to project in terms of KL divergence.

However, to maintain tractability we avoid using the entropy function H on the exponentially-large simplex Δ, and instead optimize over the structured, factorized marginal polytope $\mathcal{M}$ and its corresponding structured Bethe entropy $H_{\mathcal{B}}$. For tree-structured models, H and $H_{\mathcal{B}}$ have identical values, but different inputs. It remains to show the strong convexity of $-H_{\mathcal{B}}$ so we can use it in RDA.

Proposition 3. *For trees with n nodes, the negative Bethe entropy function $-H_{\mathcal{B}}$ is $\frac{1}{2}(2n-1)^{-2}$-strongly convex with respect to the 2-norm over the interior of the marginal polytope $\mathcal{M}$.*

Proof. Consequence of Lemma 1 in Fu & Banerjee (2013).

With these definitions in hand, we present Bethe-RDA projected inference Algorithm 1. This algorithm corresponds to instantiating (10) with $R = -H_{\mathcal{B}} - \langle \boldsymbol{\theta}, \mu \rangle$ and $\varphi = -H_{\mathcal{B}}$. Note the simplicity of the algorithm when choosing $\beta_t = 0$. It is intuitively appealing that the algorithm amounts to no more than calling our marginal inference oracle with iteratively modified parameters.

Proposition 4. *For convex energy functions and convex $-H_{\mathcal{B}}$, the sequence of primal averages of Algorithm 1 converges to the optimum of the variational objective* (2) *with suboptimality of $O(\frac{ln(t)}{t})$ at time t.*

Proof. This follows from Theorem 3 of Xiao (2010) along with the strong convexity of $-H_{\mathcal{B}}$. □

If we have more structure in the energy functions, specifically a Lipschitz-continuous gradient, we can modify the algorithm to use Nesterov's acceleration technique and achieve a convergence of $O(\frac{1}{t^2})$. Details can be found in Appendix D. Additionally, in practice these problems need not be solved to optimality and give stable results after a few iterations, as demonstrated in Figure 8.1.

6.3 INFERENCE WITH NON-CONVEX, NON-LOCAL ENERGIES

An analogy can be made here to loopy belief propagation – even in the case of non-convex loss functions (and even non-convex entropy functions with associated inexact marginal oracles), the updates of our inference (and learning) algorithms are well-defined. Importantly, since one of our motivations for developing non-local inference was to generalize mean field inference, and the additional penalty terms are non-convex in that case, we would like our algorithms to work for the non-convex case as well.

Algorithm 2 Learning with non-local energies

Input: examples $\mathbf{x}_i, \mathbf{y}_i$ and inference oracle MARG() for distributions with the clique structure of $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$.
Output: parameters $(\boldsymbol{\theta}, \boldsymbol{\psi})$ for $P_c(\mathbf{y}|\mathbf{x})$.
repeat
  //E-Step
  for all $(\mathbf{x}_i, \mathbf{y}_i)$ **do**
    $\boldsymbol{\mu}_i \leftarrow$ (Algorithm 1) // using $\boldsymbol{\theta}, \boldsymbol{\psi}$ and MARG()
    $\rho_i \leftarrow$ (Proposition 5) // using $\boldsymbol{\psi}, \boldsymbol{\mu}_i$
    // note $Q_i(\mathbf{y}_i)$ is a CRF with potentials $\boldsymbol{\theta} + \rho_i$.
  end for
  //M-Step (gradient-based learning of CRF parameters)
  repeat
    $m_i \leftarrow \text{MARG}(Q_i)\ \forall j$ //standard CRF inference
    $\nabla_{\boldsymbol{\theta}} \leftarrow \sum_i S(\mathbf{y}_i) - m_i$
    $\nabla_{\boldsymbol{\psi}} \leftarrow \sum_i \frac{d\rho_i}{d\boldsymbol{\psi}}^\top (S(\mathbf{y}_i) - m_i)$
    $\boldsymbol{\theta} \leftarrow$ Gradient-Step$(\boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}})$
    $\boldsymbol{\psi} \leftarrow$ Gradient-Step$(\boldsymbol{\psi}, \nabla_{\boldsymbol{\psi}})$
  until converged
until converged OR iter $>$ max_iters

Algorithm 3 Doubly-stochastic learning with $L_{\boldsymbol{\psi}}$ given by a sum of scalar functions of linear measurements (5).

Input: examples $\mathbf{x}_i, \mathbf{y}_i$ and MARGINAL-ORACLE() for distributions with the clique structure of $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$.
Output: parameters $(\boldsymbol{\theta}, \boldsymbol{\psi})$ for $P_c(\mathbf{y}|\mathbf{x})$.
repeat
  sample $(\mathbf{x}_i, \mathbf{y}_i)$ randomly
  $\boldsymbol{\mu}_i \leftarrow$ (Algorithm 1)
  $\nabla_{\boldsymbol{\theta}} \leftarrow S(\mathbf{y}_i) - \boldsymbol{\mu}_i$
  $\nabla_{\boldsymbol{\psi}_j} \leftarrow \nabla \ell_j(\boldsymbol{\mu}_i) a_j^\top (S(\mathbf{y}_i) - \boldsymbol{\mu}_i)$
  $\boldsymbol{\theta} \leftarrow$ Gradient-Step$(\boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}})$
  $\boldsymbol{\psi} \leftarrow$ Gradient-Step$(\boldsymbol{\psi}, \nabla_{\boldsymbol{\psi}})$
until converged OR iter $>$ max_iters

Unlike loopy belief propagation, however, since we derive our algorithms in the framework of composition minimization, we have access to a wealth of theoretical guarantees. Based on results from the theory of optimization with first-order surrogate loss functions (Mairal, 2013), in Appendix C we propose a small modification to Algorithm 1 with an asymptotic convergence condition even for non-convex energies. In practice we find that the unmodified Algorithm 1 also works well for these problems, and experimentally in Section 8.2, we see good performance in both inference and learning with non-convex energy functions.

7 LEARNING MODELS WITH NON-LOCAL ENERGIES

We seek to learn the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ of the underlying CRF base model and $L_{\boldsymbol{\psi}}$, respectively. Let $S = \{\mathbf{y}_i, \mathbf{x}_i\}$ be n training examples. Let $Q(\mathbf{y}_i; \boldsymbol{\mu}_i)$ be the variational distribution for $\mathbf{y}_i$ resulting from applying Proposition 1. Namely, $Q(\mathbf{y}_i; \boldsymbol{\mu}_i)$ is an MRF with parameters

$$\boldsymbol{\rho_i} \coloneqq \boldsymbol{\theta} - \nabla_{\boldsymbol{\mu}} L_{\boldsymbol{\psi}}(\boldsymbol{\mu}_i). \tag{11}$$

We employ the notation $Q(\mathbf{y}_i; \boldsymbol{\mu}_i)$ to highlight the role of $\boldsymbol{\mu}_i$: for a given $(\mathbf{y}_i, \mathbf{x}_i)$ pair, the family of variational distributions over $\mathbf{y}_i$ is indexed by possible values of $\boldsymbol{\mu}_i$ (recall we suppress the explicit dependence of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ on $\mathbf{x}$). Finally, define the shorthand $M = \{\mu_1, \ldots, \mu_n\}$.

$\boldsymbol{\psi}$ interacts with the data in a complex manner that prevents us from using standard learning techniques for the exponential family. Namely, we can not easily differentiate a likelihood with respect to $\boldsymbol{\psi}$, since this requires differentiating the output $\boldsymbol{\mu}$ of a convex optimization procedure, and the extra $L_{\boldsymbol{\psi}}$ term in (2) prevents the use of conjugate duality relationships available for the exponential family. We could have used automatic methods to differentiate the iterative inference procedure (Stoyanov et al., 2011; Domke, 2012), but found our learning algorithm works well.

We employ a variational learning algorithm, presented in Algorithm 2, alternately updating the parameters M of our tractable CRF-structured variational distributions, and updating the parameters $(\boldsymbol{\theta}, \boldsymbol{\psi})$ assuming the following surrogate likelihood given by these CRF approximations:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}; M) = \sum_i \log Q(y_i; \boldsymbol{\mu}_i). \tag{12}$$

Given $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, we update M using Algorithm 1. Given M, we update $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ by taking a single step in the direction of the gradient of the surrogate likelihood (12). We avoid taking more than one gradient step, since the gradients for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ depend on M and an update to $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ will break the property that $\boldsymbol{\mu}(Q(\mathbf{y}; \boldsymbol{\mu}_i)) = \boldsymbol{\mu}_i$. Therefore, we recompute $\boldsymbol{\mu}_i$ every time we update the parameters.

Overall, it remains to show how to compute gradients of (12). For $\boldsymbol{\theta}$, we have the standard CRF likelihood gradient (Sutton & McCallum, 2006):

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\psi}; M) = \sum_i S(\mathbf{y}_i) - \boldsymbol{\mu}_i\ . \tag{13}$$

For $\boldsymbol{\psi}$, we have:

$$\nabla_{\boldsymbol{\psi}} L(\boldsymbol{\theta}, \boldsymbol{\psi}; M) = \sum_i \frac{d\boldsymbol{\rho_i}}{d\boldsymbol{\psi}} \frac{d}{d\boldsymbol{\rho_i}} \log Q(\mathbf{y}_i; \boldsymbol{\mu}_i). \tag{14}$$

From (11), $\frac{d}{d\boldsymbol{\rho_i}} \log Q(\mathbf{y}_i; \boldsymbol{\mu}_i)$ is also $S(\mathbf{y}_i) - \boldsymbol{\mu}_i$and

$$\frac{d\boldsymbol{\rho_i}}{d\boldsymbol{\psi}} = \frac{d}{d\boldsymbol{\psi}} \frac{d}{d\boldsymbol{\mu}} L_{\boldsymbol{\psi}}(\boldsymbol{\mu}) \tag{15}$$

Clearly, this depends on the structure of $L_{\boldsymbol{\psi}}$. Consider the parametrization (4). With this, we have:

$$\frac{\partial}{\partial \boldsymbol{\psi}_j} \frac{d}{d\boldsymbol{\mu}} L_{\boldsymbol{\psi}}(\boldsymbol{\mu}) = \nabla \ell_j(\boldsymbol{\mu}) \frac{d}{d\boldsymbol{\mu}} \ell_j(\boldsymbol{\mu}) \tag{16}$$

Therefore, we have $\frac{\partial}{\partial \boldsymbol{\psi}_j} \log Q(\mathbf{y}_i; \boldsymbol{\mu}_i) = \nabla \ell_j(\boldsymbol{\mu}) \frac{d}{d\boldsymbol{\mu}} \ell_j(\boldsymbol{\mu})^\top (S(\mathbf{y}) - \boldsymbol{\mu}_i)$. For linear measurements (5), this amounts to

$$\nabla \ell(\boldsymbol{\mu}) \left(a_j^\top S(\mathbf{y}) - a_j^\top \boldsymbol{\mu}_i \right). \tag{17}$$

This has a simple interpretation: the gradient with respect to $\boldsymbol{\psi}_j$ equals the gradient of the scalar loss ℓ_j at the current marginals $\boldsymbol{\mu}_j$ times the difference in linear measurements between the ground truth labels and the inferred marginals.

Algorithm 2 has an expensive double-loop structure. In practice it is sufficient to employ a 'doubly-stochastic' version given in Algorithm 3, where we sample a training example $(\mathbf{x}_i, \mathbf{y}_i)$ and use this to only perform a single gradient step on $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. To demonstrate the simplicity of implementing our learning algorithm, we avoid any abstract derivative notation in Algorithm 3 by specializing it to the case of (17). In our experiments, however, we sometimes do not use linear measurements. Overall, all our experiments use the fast doubly-stochastic approach of Algorithm 3 solely, since it performs well. In general, our learning algorithms are not guaranteed to converge because we approximate the complex interaction between $\boldsymbol{\psi}$ and $\boldsymbol{\mu}$ with alternating updates. In practice, however, terminating after a fixed number of iterations yields models that generalize well.

Finally, recall that the notation $L_{\boldsymbol{\psi}}(\boldsymbol{\mu}_i)$ suppresses the potential dependence of $\boldsymbol{\psi}$ on $\mathbf{x}_i$. We assume each $\boldsymbol{\psi}_j$ is a differentiable function of features of $\mathbf{x}_i$. Therefore, in our experiments where $\boldsymbol{\psi}$ depends on $\mathbf{x}_i$, we perform gradient updates for the parametrization of $\boldsymbol{\psi}(\mathbf{x})$ via further application of the chain rule.

8 EXPERIMENTS

8.1 CITATION EXTRACTION

Model	F1
Our Baseline	94.47
Non-local Energies	95.47
Baseline (Anzaroot et al., 2014)	94.41
Soft-DD (Anzaroot et al., 2014)	95.39

Table 1: Comparison of F1 scores on Citation Extraction dataset. We compare MAP inference F1 scores of our non-local energy model and the specialized dual decomposition model of Anzaroot et al. (2014). Both variants learn global regularities that significantly improve performance.

We first apply our algorithm to the NLP task of performing text field segmentation on the UMass citation dataset (Anzaroot & McCallum, 2013), which contains strings of citations from research papers, segmented into fields (author, title, etc.). Our modeling approach, closely follows Anzaroot et al. (2014), who extract segmentations using a linear-chain segmentation model, to which they add a large set of 'soft' linear global regularity constraints.

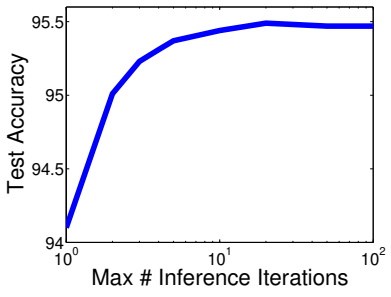

Figure 1: Citation extraction F1 when limiting maximum number of test-time inference iterations. Most of our accuracy gain is captured within the first 5-10 iterations.

Let $\mathbf{y}$ be a candidate labeling. Imagine, for example, that we constrain predicted segmentations to have no more predicted last names than first names. Then, the numbers of first and last names can be computed by linear measurements $a_{\text{first}}^\top S(\mathbf{y})$ and $a_{\text{last}}^\top S(\mathbf{y})$, respectively. A hard constraint on $\mathbf{y}$ would enforce $a_{\text{first}}^\top S(\mathbf{y}) - a_{\text{last}}^\top S(\mathbf{y}) = 0$. This is relaxed in Anzaroot et al. (2014) to a penalty term

$$c\ell_h \left(a_{\text{first}}^\top S(\mathbf{y}) - a_{\text{last}}^\top S(\mathbf{y}) \right) \tag{18}$$

that is added to the MAP inference objective, where $\ell_h(x) = \max(1 - x, 0)$ is a hinge function. For multiple soft constraints, the overall prediction problem is

$$\arg\min_{\mathbf{y}} \langle -\boldsymbol{\theta}, S(\mathbf{y}) \rangle + \sum_j c_j \ell_h \left(a_j^\top S(\mathbf{y}) \right), \tag{19}$$

where $\boldsymbol{\theta}$ are the parameters of the underlying linear-chain model. They use a dual decomposition style algorithm for solving (19), that crucially relies on the specific structure of the hinge terms ℓ_h. They learn the c_j for hundreds of 'soft constraints' using a perceptron-style algorithm.

We consider the same set of measurement vectors a_j, but impose non-local terms that act on *marginals* $\boldsymbol{\mu}$ rather than specific values $\mathbf{y}$. Further, we use *smoothed* hinge functions, which improve the convergence rate of inference (Rennie, 2005). We find the variational distribution by solving the marginal inference version of (19), an instance of our inference framework with linear measurements (5):

$$\arg\min_{\boldsymbol{\mu}} \langle -\boldsymbol{\theta}, \boldsymbol{\mu} \rangle - H_{\mathcal{B}}(\boldsymbol{\mu}) + \sum_j c_j \ell_h \left(a_j^\top \boldsymbol{\mu} \right), \tag{20}$$

As in Anzaroot et al. (2014), we first learn chain CRF parameters $\boldsymbol{\theta}$ on the training set. Then, we learn the c_j parameters on the development set, using Algorithm 3, and tune hyperparameters for development set performance. At both

train and test time, we ignore any terms in (20) for which $c_j < 0$.

We present our results in Table 1, measuring segment-level F1. We can see that our baseline chain has slightly higher accuracy than the baseline approach of Anzaroot et al. (2014), possibly due to optimization differences. Our augmented model (Non-Local Energies) matches and very slightly beats their soft dual decomposition (Soft-DD) procedure. This is especially impressive because they employ a specialized linear-programming solver and learning algorithm adapted to the task of MAP inference under hinge-loss soft constraints, whereas we simply plug in our general learning and inference algorithms for non-local structured prediction – applicable to any set of energy functions.

Our comparable performance provides experimental evidence for our intuition that preferences about MAP configurations can be expressed (and "relaxed") as functions of expectations. Anzaroot et al. (2014) solve a penalized MAP problem directly, while our prediction algorithm first finds a distribution satisfying these preferences, and then performs standard MAP inference in that distribution.

Finally, in Figure 1 we present results demonstrating that our algorithm's high performance can be obtained using only 5-10 calls per test example to inference in the underlying chain model. In Section B, we analyze the empirical convergence behavior of Algorithm 1.

8.2 HANDWRITING RECOGNITION

N-Grams	2	3	4	5	6
Accuracy	85.02	96.20	97.21	98.27	98.54

Table 2: Character-wise accuracy of Structured Prediction Cascades (Weiss et al., 2012) on OCR dataset.

Model	Accuracy
2-gram (base model)	84.93
$L^u_{\boldsymbol{\psi}}$	94.01
$L^u_{\boldsymbol{\psi}}$ (MM)	94.96
$L^w_{\boldsymbol{\psi}}$	98.26
$L^w_{\boldsymbol{\psi}}$ (MM)	**98.83**
55-Class Classifier (MM)	86.06

Table 3: Character-wise accuracy of our baselines, and models using learned non-local energies on Handwriting Recognition dataset. Note that word classifier baseline is also given in character-wise accuracy for comparison.

We next apply our algorithms to the widely-used handwriting recognition dataset of Taskar et al. (2004). We follow the setup of Weiss et al. (2012), splitting the data into 10 equally sized folds, using 9 for training and one to test. We report the cross-validation results across all 10 folds.

The *structured prediction cascades* of Weiss et al. (2012) achieve high performance on this dataset by using extremely high order cliques of characters (up to 6-grams), for which they consider only a small number of candidate outputs. Their state-of-the-art results are reproduced in Table 2. The excellent performance of these large-clique models is consequence of the fact that the data contains only 55 unique words, written by 150 different people. Once the model has access to enough higher-order context, the problem becomes much easier to solve.

With this in mind, we design two non-convex, non-local energy functions. These energies are intended to regularize our predictions to lie close to known elements of the vocabulary. Our base model is a standard linear-chain CRF with image features on the nodes, and no features on the bigram edge potentials. Let $U(\mu) = \sum_n \mu_n$ be a function that takes the concatenated vector of node and edge marginals and sums up all of the node marginals, giving the global unigram expected sufficient statistics. Let $\{u_i\} = \{U(\mu(y_i))\}$ indicate the set of all such unique vectors when applying U to the train set empirical sufficient statistics for each data case y_i. Simply, this gives 55 vectors u_i of length 26 containing the unigram counts for each unique word in the train set.

Our intuition is that we would like to be able to "nudge" the results of inference in our chain model by pulling the inferred $U(\mu)$ to be close to one of these global statistics vectors. We add the following non-convex non-local energy function to the model:

$$L^u_{\boldsymbol{\psi}}(\mu) = \boldsymbol{\psi} \min_i \|u_i - U(\mu)\|_1. \tag{21}$$

We learn two variants of this model, which differently parametrize the dependence of $\boldsymbol{\psi}$ on $\mathbf{x}$. The first has a single bias feature on the non-local energy. The second conditions on a global representation of the sequence: concretely, we approximate the RBF *kernel mean map* (MM) (Smola et al., 2007) using random Fourier features (RFF) (Rahimi & Recht, 2007). This simply involves multiplying each image feature vector in the sequence by a random matrix with ~ 1000 rows, applying a pointwise non-linearity, and taking $\boldsymbol{\psi}$ to be a linear function of the average vector.

Results of these experiments can be seen in Table 3. Adding the non-local energy brings our performance well above the baseline bigram chain model, and our training procedure is able to give substantially better performance when $\boldsymbol{\psi}$ depends on the above input features.

The energy $L^u_{\boldsymbol{\psi}}$, based on unigram sufficient statistics, is not able to capture the relative ordering of letters in the vocabulary words, which the structured prediction cascades models do capture. This motivates us to consider another energy function. Let $\{w_i\} = \{\mu_n(y_i)\}$ be the set of unique vectors of concatenated node marginal statistics for the train set. This gives 55 vectors of length $l_i * 26$, where l_i is

s	625	10k	50k
Our Method	0.19	2.7	14
IP	2.8	93	690

Table 4: Comparison of runtime (in seconds, averaged over 10 trials) between the interior point solver (IP) of Sheldon et al. (2013) v.s. Algorithm 1 on different CGM problem sizes s, the cardinality of the edge potentials in the underlying graphical model, where marginal inference is $O(s)$.

the length of the ith distinct train word. Next, we define a different energy function to add to our base chain model:

$$L^w_{\boldsymbol{\psi}}(\mu) = \boldsymbol{\psi} \min_i \|w_i - \mu\|_1. \quad (22)$$

Once again we implement featurized and non-featurized versions of this model. As noted in structured prediction cascades, giving the model access to this level of high-order structure in the data makes the inference problem extremely easy. Our model outperforms the best structured prediction cascades results, and we note again an improvement from using the featurized over the non-featurized $\boldsymbol{\psi}$.

Of course, since the dataset has only 55 actual labels, and some of those are not valid for different input sequences due to length mismatches, this is arguably a classification problem as much as a structured prediction problem. To address this, we create another baseline, which is a constrained 55-class logistic regression classifier (constrained to only allow choosing output classes with appropriate lengths given the input). We use our same global mean-map features from the $L^*_{\boldsymbol{\psi}}$ (MM) variants of the structured model and report these results in Table 3. We also tune the number of random Fourier features as a hyperparameter to give the classifier as much expressive power as possible. As we can see, the performance is still significantly below the best structured models, indicating that the interplay between local and global structure is important.

8.3 COLLECTIVE GRAPHICAL MODELS

Next, we demonstrate that that our proximal gradient-based inference framework dramatically speeds up approximate inference in *collective graphical models* (CGMs) (Sheldon & Dietterich, 2011). CGMs are a method for structured learning and inference with noisy aggregate observation data. The large-scale dependency structure is represented via a graphical model, but the nodes represent not just single variables, but aggregate sufficient statistics of large sets of underlying variables, corrupted by some noise model. In previous work, CGMs have been successfully applied to modeling bird migration. Here, the base model is a linear chain representing a time series of bird locations. Each observed variable corresponds to counts from bird watchers in different locations. These observations are assumed to be Poisson distributed with rate proportional to the true count of birds present. The CGM MAP task is to infer the underlying migration patterns.

Sheldon et al. (2013) demonstrate that MAP in CGMs is NP-hard, *even for trees*, but that approximate MAP can be performed by solving a problem of the form (2):

$$\mu^* = \arg\max_{\boldsymbol{\mu}} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle + H_{\mathcal{B}}(\boldsymbol{\mu}) + \sum_i^n P_i(\mu_i | \boldsymbol{\psi} \mathbf{y}_i) \quad (23)$$

where P_i are (concave) Poisson log-likelihoods and each $\mathbf{y}_i$ is an observed bird count.

For the case where the underlying CGM graph is a tree, the 'hard EM' learning algorithm of Sheldon et al. (2013) is the same as Algorithm 2 specialized to their model. Therefore, Sheldon et al. (2013) provide additional experimental evidence that our alternating surrogate-likelihood optimization works well in practice.

The learning procedure of Sheldon et al. (2013) is very computationally expensive because they solve instances of (23) using an interior-point solver in the inner loop. For the special case of trees, Algorithm 1 is directly applicable to (23). Using synthetic data and code obtained from the authors, we compare their generic solver to Algorithm 1 for solving instances of (23). In Table 4, we see that our method achieves a large speed-up with no loss in solution accuracy (since it solves the same convex problem).

9 DISCUSSION AND FUTURE WORK

Our results show that our inference and learning framework allows for tractable modeling of non-local dependency structures, resistant to traditional probabilistic formulations. By approaching structured modeling not via independence assumptions, but as arbitrary penalty functions on the marginal vectors $\boldsymbol{\mu}$, we open many new modeling possibilities. Additionally, our generic gradient-based inference method can achieve substantial speedups on pre-existing problems of interest. In future work, we will apply our framework to new problems and new domains.

ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by DARPA under agreement number FA8750-13-2-0020, and in part by NSF grant #CNS-0958392. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Aguiar, Pedro, Xing, Eric P, Figueiredo, Mário, Smith, Noah A, and Martins, André. An augmented lagrangian approach to constrained map inference. In *ICML*, 2011.

Anzaroot, Sam and McCallum, Andrew. A new dataset for fine-grained citation field extraction. In *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.

Anzaroot, Sam, Passos, Alexandre, Belanger, David, and McCallum, Andrew. Learning soft linear constraints with application to citation field extraction. In *ACL*, 2014.

Beck, Amir and Teboulle, Marc. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Domke, Justin. Generic methods for optimization-based modeling. In *AISTATS*, 2012.

Duchi, John, Shalev-Shwartz, Shai, Singer, Yoram, and Tewari, Ambuj. Composite objective mirror descent. In *COLT*, 2010.

Fu, Qiang and Banerjee, Huahua Wang Arindam. Bethe-admm for tree decomposition based parallel map inference. In *UAI*, 2013.

Ganchev, Kuzman, Graça, Joao, Gillenwater, Jennifer, and Taskar, Ben. Posterior regularization for structured latent variable models. *JMLR*, 99:2001–2049, 2010.

He, L., Gillenwater, J., and Taskar, B. Graph-Based Posterior Regularization for Semi-Supervised Structured Prediction. In *CoNLL*, 2013.

Komodakis, Nikos, Paragios, Nikos, and Tziritas, Georgios. Mrf optimization via dual decomposition: Message-passing revisited. In *IEEE ICCV*, 2007.

Lafferty, John, McCallum, Andrew, and Pereira, Fernando CN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

Liang, Percy, Jordan, Michael I, and Klein, Dan. Learning from measurements in exponential families. In *ICML*, 2009.

Mairal, Julien. Optimization with first-order surrogate functions. In *ICML*, 2013.

Mann, Gideon S and McCallum, Andrew. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *JMLR*, 11:955–984, 2010.

Nesterov, Yurii. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Parikh, Neal and Boyd, Stephen. Proximal algorithms. *Foundations and Trends in Optimization*, 2013.

Rahimi, Ali and Recht, Benjamin. Random features for large-scale kernel machines. In *NIPS*, 2007.

Ravikumar, Pradeep, Agarwal, Alekh, and Wainwright, Martin J. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *JMLR*, 11:1043–1080, 2010.

Rennie, Jason DM. Smooth hinge classification, 2005.

Rockafellar, R Tyrell. *Convex Analysis*, volume 28. Princeton University Press, 1997.

Sheldon, Daniel, Sun, Tao, Kumar, Akshat, and Dietterich, Thomas G. Approximate inference in collective graphical models. In *ICML*, 2013.

Sheldon, Daniel R and Dietterich, Thomas G. Collective graphical models. In *NIPS*, 2011.

Smola, Alex, Gretton, Arthur, Song, Le, and Schölkopf, Bernhard. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Sontag, David, Globerson, Amir, and Jaakkola, Tommi. Introduction to dual decomposition for inference. *Optimization for Machine Learning*, 1:219–254, 2011.

Stoyanov, Veselin, Ropson, Alexander, and Eisner, Jason. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *AISTATS*, 2011.

Sutton, Charles and McCallum, Andrew. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, pp. 93–128, 2006.

Tarlow, Daniel and Zemel, Richard S. Structured output learning with high order loss functions. In *AISTATS*, 2012.

Taskar, Ben, Carlos, Guestrin, and Koller, Daphne. Max-margin markov networks. In *NIPS*, 2004.

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

Weiss, D., Sapp, B., and Taskar, B. Structured Prediction Cascades. *ArXiv e-prints*, August 2012.

Xiao, Lin. Dual averaging methods for regularized stochastic learning and online optimization. *JMLR*, 11:2543–2596, 2010.

Supplementary Material

A Variational Approximation

During learning, reasoning about $P_c(\mathbf{y}|\mathbf{x})$ in (8) is difficult, due to the intractability of $Z_{\boldsymbol{\theta},\boldsymbol{\psi}}$. In response, we approximate it with a variational distribution:

$$Q(\mathbf{y}) = \arg\min_{Q'} F(Q'; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}), \tag{24}$$

where

$$\begin{aligned} F(Q') &= KL(Q'(\mathbf{y})||P_c(\mathbf{y}|\mathbf{x})) \\ &= -H(Q') - \mathbb{E}_{Q'}[\langle \boldsymbol{\theta}, S(\mathbf{y})\rangle] + \mathbb{E}_{Q'}[L_{\boldsymbol{\psi}}(S(\mathbf{y}))] \\ &\leq -H(Q') - \langle \boldsymbol{\theta}, \boldsymbol{\mu}(Q')\rangle + L_{\boldsymbol{\psi}}(\boldsymbol{\mu}(Q')). \end{aligned} \tag{25}$$

Given $\mathbf{x}$, $\boldsymbol{\theta}$, and $\boldsymbol{\psi}$, we select Q by minimizing the convex upper bound (25), which follows from Jensen's inequality.

So far, we have not assumed any structure on Q. Next, we show that the minimizer of (25) is a MRF with the same clique structure as $P_{\boldsymbol{\theta}}$. This provides an alternative derivation of the techniques in Section 4.

Let $q_{\mathbf{y}}$ denote the probability under Q of a given joint configuration $\mathbf{y}$. There are exponentially-many such $q_{\mathbf{y}}$, and $H(Q)$ is the entropy on the simplex $-\sum_{\mathbf{y}} q_{\mathbf{y}} \log(q_{\mathbf{y}})$. Since Q minimizes (25), we have the following stationarity condition for every $q_{\mathbf{y}}$:

$$\frac{d}{dq_{\mathbf{y}}} \left[-H(Q_\phi) - q_{\mathbf{y}} \log(P_{\boldsymbol{\theta}}(y|\mathbf{x})) + L_{\boldsymbol{\psi}}(\boldsymbol{\mu}(Q_\phi))\right] + \lambda = 0 \tag{26}$$

Here, λ is a dual variable for the constraint $\sum_{\mathbf{y}} q_{\mathbf{y}} = 1$. Rearranging, we have:

$$Q(\mathbf{y}) = \tag{27}$$

$$(1/Z) P_{\boldsymbol{\theta}}(y|\mathbf{x}) \exp\left(-\left(\frac{d}{d\boldsymbol{\mu}} L_{\boldsymbol{\psi}}(\boldsymbol{\mu}(Q))\right)^\top \left(\frac{d}{dq_{\mathbf{y}}} \boldsymbol{\mu}(Q)\right)\right), \tag{28}$$

where Z is a normalizing constant.

Proposition 5. *There exists a vector ρ such that the quantity $\left(\frac{d}{d\boldsymbol{\mu}} L_{\boldsymbol{\psi}}(\boldsymbol{\mu}(Q))\right)^\top \left(\frac{d}{dq_{\mathbf{y}}} \boldsymbol{\mu}(Q)\right) = \rho^\top S(\mathbf{y})$ for all $q_{\mathbf{y}}$. Furthermore, ρ is a simple, closed-form function of $\boldsymbol{\mu}(Q)$.*

Proof. We have $\frac{d}{dq_{\mathbf{y}}} \boldsymbol{\mu}(Q) = S(\mathbf{y})$, since $\boldsymbol{\mu}(Q) = \sum_{\mathbf{y}} q_{\mathbf{y}} S(\mathbf{y})$. Therefore, $\rho = \frac{d}{d\boldsymbol{\mu}} L_{\boldsymbol{\psi}}(\boldsymbol{\mu}(Q))$. □

Corollary 1. *Since $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) \propto \langle \boldsymbol{\theta}, S(\mathbf{y})\rangle$, Proposition 5 implies $Q(\mathbf{y})$ is an MRF with the same clique decomposition as $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$.*

So far, Q is implicitly defined in terms of its own marginals $\boldsymbol{\mu}(Q)$. Since we assume $P_{\boldsymbol{\theta}}$ and $P_{\boldsymbol{\psi}}$ have the same sufficient statistics $S(\mathbf{y})$, we can use the Bethe entropy representation $H(Q) = H_{\mathcal{B}}(\mu(Q))$. This transforms (25) to the augmented inference problem (2). Therefore, we can directly solve for $\boldsymbol{\mu}(Q)$, which can then be used to provide a closed-form expression for the CRF distribution Q.

B Additional Experiments

In Figure 2, we examine the convergence behavior of our algorithm on the citation dataset. This demonstrates that our inference procedure converges quite quickly except for a small number of difficult cases, where the global energy and the local evidence are in significant disagreement.

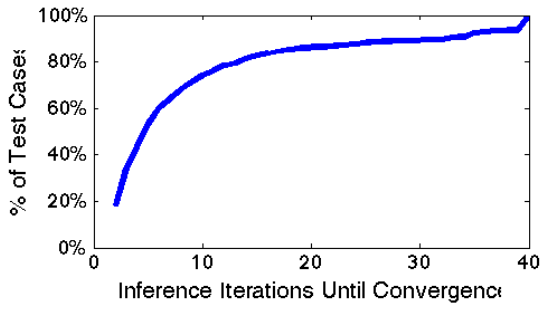

Figure 2: The number of iterations taken for inference to converge on test set citations, as a percentage of the total number of test cases. Number of iterations is capped at 40. We can see that the distribution is long tailed. Inference converges within 40 iterations for 93.7 of examples, and each example takes an average of 9.8 iterations to converge.

Algorithm 4 Bethe-MD

Input: parameters $\boldsymbol{\theta}$, energy function $L(\mu)$, learning rate sequence $\{\eta_t\}$
set μ_0 to prox-center MARGINAL-ORACLE($\boldsymbol{\theta}$)
repeat
  $g_t = \nabla H_{\mathcal{B}}(\mu_{t-1}) + \eta_t \nabla L(\mu_{t-1})$
  $\mu_t = \text{MARGINAL-ORACLE}(\frac{1}{1+\eta_t}(\eta_t \boldsymbol{\theta} - g_t))$
until CONVERGED(μ_t, μ_{t-1})

C Non-Convex Energies and Composite Mirror Descent

We introduce a small modification of Algorithm 1, along with a rough proof sketch of its convergence even in the case of non-convex energy functions. Because it leans heavily on significant prior work in optimization, it is hard to give a self-contained proof of the results in this section, and our argument takes the form of a proof sketch that appeals to these other works. However, the basic argument simply combines the strong convexity of $H_{\mathcal{B}}$ and its associated Bregman divergence, along with the results of Mairal (2013) for the case of composite minimization of non-convex functions using the Euclidean Bregman divergence, and the fact that the local updates performed using entropy $H_{\mathcal{B}}$ as a distance-generating function have a log-barrier function for the constraint set $\mathcal{M}$, effectively bounding the norm of the gradient of $H_{\mathcal{B}}$ when restricted to the set of iterates actually visited during optimization.

While Algorithm 1 was built on the framework of regularized dual averaging (RDA), we introduce a slightly different formulation based on *composite mirror descent* (COMID) (Duchi et al., 2010). Like RDA, COMID is a gradient method for minimizing functions of the form $h = f + R$. At each time step t, COMID makes the update

$$w_{t+1} = \arg\min_{w} \langle \nabla f(w_t), w \rangle + \frac{1}{\eta_t} B_{\varphi}(w, w_t) + R(w) \tag{29}$$

where φ is some strongly convex function and B_{φ} is its associated Bregman divergence. In Algorithm 4, we present an instantiation of composite mirror descent for our inference problem.

At first glance, this seems significantly different from our original Algorithm 1, but remembering that $\nabla H_{\mathcal{B}}(\mu_t) = \boldsymbol{\theta}_t$ because of conjugate duality of the exponential family, we can see that it actually only a corresponds to a slight re-weighting of the iterates of Algorithm 1.

First, we give Algorithm 4 similar guarantees in the convex setting as we did for Algorithm 1.

Proposition 6. *For convex energy functions and convex $-H_{\mathcal{B}}$, given the learning rate sequence $\eta_t = \frac{1}{\lambda t}$, where λ is the strong convexity of $-H_{\mathcal{B}}$, the sequence of primal averages of Algorithm 4 converges to the optimum of the variational objective* (2) *with suboptimality of $O(\frac{ln(t)}{t})$ at time t.*

Proof. This follows from a standard online-to-batch conversion, along with the strong convexity of $H_{\mathcal{B}}$ and Theorem 7 of Duchi et al. (2010). □

Now, having introduced composite mirror descent in (29), will lean heavily on the framework for optimization with first-order surrogate losses of Mairal (2013) to show that these types of algorithms should converge even in the non-convex case. We now recall a few definitions from that work.

First, we define the *asymptotic stationary point* condition, which gives us a notion of convergence in the non-convex optimization case.

Definition 1 (Asymptotic Stationary Point (Mairal, 2013))**.** *For a sequence $\{\boldsymbol{\theta}_n\}_{n\geq 0}$, and differentiable function f, we say it satisfies an asymptotic stationary point condition if*

$$\lim_{n\to+\infty} \|\nabla f(\boldsymbol{\theta}_n)\|_2 = 0$$

We call a function L-strongly smooth if L is a bound on the largest eigenvalue of the Hessian – this tells us how the norm of the gradient changes. This is also known as a L-Lipschitz continuous gradient. Now we recall the notion of a *majorant first-order surrogate function*.

Definition 2 (Majorant First-Order Surrogate (Mairal, 2013))**.** *A function g: $\mathbb{R}^p \to \mathbb{R}$ is a majorant first-order surrogate of f near κ when the following conditions are satisfied*

- *Majorant: we have $g \geq f$.*
- *Smoothness: the approximation error $h = g - f$ is differentiable, and its gradient is L-Lipschitz continuous, moreover, we have $h(\kappa) = 0$ and $\nabla h(\kappa) = 0$*

We denote by $\mathcal{S}_L(f, \kappa)$ the set of such surrogates.

Now we recall the majorant first-order surrogate property for the composite minimization step in the case of Euclidean Bregman divergence (Euclidean distance).

Proposition 7 (Proximal Gradient Surrogates (Mairal, 2013))**.** *Assume that $h = f + R$ where f is differentiable with an L-Lipschitz gradient. Then, h admits the following majorant surrogate in $\mathcal{S}_{2L}(f, \kappa)$:*

$$g(\boldsymbol{\theta}) = f(\kappa) + \nabla f(\kappa)^\top(\boldsymbol{\theta} - \kappa) + \frac{L}{2}\|\boldsymbol{\theta} - \kappa\|_2^2 + R(\boldsymbol{\theta}) \tag{30}$$

We can use this result to establish a majorant property for the composite mirror descent surrogate (29) given a strongly convex and strongly smooth Bregman divergence.

Proposition 8 (Composite Mirror Descent Surrogates)**.** *Assume that $h = f + R$ where f is differentiable with an L-Lipschitz gradient, φ is a σ-strongly convex and γ-strongly smooth function, and B_φ is its Bregman divergence. Then, h admits the following majorant surrogate in $\mathcal{S}_{L+L\frac{\gamma}{\sigma}}(f, \kappa)$:*

$$g(\boldsymbol{\theta}) = f(\kappa) + \nabla f(\kappa)^\top(\boldsymbol{\theta} - \kappa) + \frac{L}{2\sigma}B_\varphi(\boldsymbol{\theta}, \kappa) + R(\boldsymbol{\theta}) \tag{31}$$

Proof. By the definition of strong convexity and the Bregman divergence, (31) upper bounds (30), so it is a majorant of h. Additionally, by the additive property of strong smoothness, we get the strong smoothness constant for the surrogate. □

However, small technical conditions keep Proposition 8 from applying directly to our case. The Bethe entropy $H_\mathcal{B}$, and thus its associated Bregman divergence, is not strongly smooth – its gradient norm is unbounded as we approach the corners of the marginal polytope. However, it is *locally Lipschitz* – every point in the domain has a neighborhood for which the function is Lipschitz. In practice, since the $-H_\mathcal{B}$ mirror descent updates have a barrier function for the constraint set $\mathcal{M}$, our iterative algorithm will never get too close to the boundary of the polytope and it is effectively strongly smooth for purposes of our minimization algorithm. This is not a rigorous argument, but is both intuitively plausible and born out in experiments.

Proposition 9. *The sequence of iterates w_t from Algorithm 4, when bounded away from the corners of the marginal polytope constraint set $\mathcal{M}$, and for appropriate choice of learning rates $\{\eta_t\}$, convex $-H_\mathcal{B}$, and L-strongly smooth (but possibly non-convex) energy function L_ψ, satisfies an asymptotic stationary point condition.*

Algorithm 5 Accelerated Bethe-RDA

Input: parameters $\boldsymbol{\theta}$, energy function $L(\mu)$
set μ_0 to prox-center MARGINAL-ORACLE($\boldsymbol{\theta}$)
set $\nu_0 = \mu_0$
$\bar{g}_0 = 0$
repeat
  $c_t = \frac{2}{t+1}$
  $u_t = (1 - c_t)\mu_{t-1} + c_t\nu_{t-1}$
  $\bar{g}_t = (1 - c_t)\bar{g}_{t-1} + c_t\nabla L(u_t)$
  $\nu_t = \text{MARGINAL-ORACLE}(\frac{t(t+1)}{4L+t(t+1)}(\boldsymbol{\theta} - \bar{g}_t))$
  $\mu_t = (1 - c_t)\mu_{t-1} + c_t\nu_t$
until CONVERGED(μ_t, μ_{t-1})

Proof. This follows from application of Proposition 8, and noting that Algorithm 4 corresponds to the generalized surrogate-minimization scheme in Algorithm 1 of Mairal (2013). The asymptotic stationary point condition then follows from Proposition 2.1 of Mairal (2013). The appropriate learning rates $\{\eta_t\}$ must be chosen by the Lipschitz constant of the gradient of $L_{\boldsymbol{\psi}}$, as well as the effective Lipschitz constant of the gradient of $H_{\mathcal{B}}$, given how far we are bounded from the edge of the constraint set (this effective smoothness constant is determined by the norm of our parameter vector $\boldsymbol{\theta}$). □

In this section we have given a rough proof sketch for the asymptotic convergence of our inference algorithms even in the case of non-convex energies. Our heuristic argument for the effective smoothness of the entropy $H_{\mathcal{B}}$ is the most pressing avenue for future work, but we believe it could be made rigorous by examining the norm of the parameter vector and how it contributes to the "sharpness" of the barrier function for the mirror descent iterates.

D Accelerated Bethe-RDA

If we have L-strongly smooth losses (L is a bound on the largest eigenvalue of the Hessian), we can use an accelerated dual averaging procedure to obtain an even faster convergence rate of $O(\frac{1}{t^2})$. Let D be the diameter of the marginal polytope as measured by the strongly convex distance-generating function $H_{\mathcal{B}}$ (using its associated Bregman divergence.) Then Algorithm 5 gives us a convergence rate of $4LD^2/t^2$ by Corollary 7 of Xiao (2010).