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.

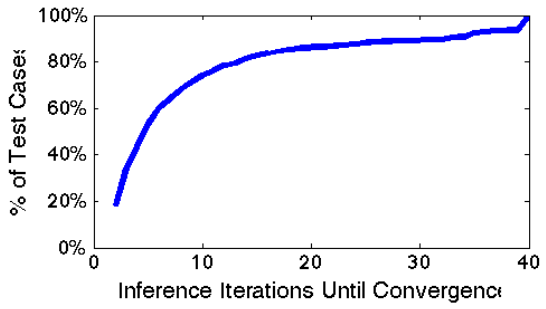

Figure 2: The number of iterations taken for inference to converge on test set citations, as a percentage of the total number of test cases. Number of iterations is capped at 40. We can see that the distribution is long tailed. Inference converges within 40 iterations for 93.7 of examples, and each example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