1. Introduction

After having introduced the temperature parameter $\epsilon$ we want to minimize the following cost function w.r.t. model parameters $w$:

$$
\frac{C}{p} ||w||_p^p + \sum_{(x,y) \in D} \left( \epsilon \ln \sum_{s \in S} \exp \left( \frac{w^T \phi(x, s) + \ell(x,y)(s)}{\epsilon} \right) - \epsilon \ln \sum_{h \in H} \exp \left( \frac{w^T \phi(x, (y, h)) + \ell^c(x, (y, h))}{\epsilon} \right) \right),
$$

(1)

2. Proof of Claim 1

Claim 1 The following function

$$
\frac{C}{p} ||w||_p^p + \sum_{(x,y) \in D} \left( \epsilon \ln \sum_{s \in S} \exp \left( \frac{w^T \phi(x, s) + \ell(x,y)(s)}{\epsilon} \right) - \epsilon H(q(x,y)) - \mathbb{E}_{q(x,y)}[w^T \phi(x, (y, h)) + \ell^c(x, (y, h))] \right),
$$

(2)

convex in $w$ and $q(x,y)$ separately, is an upper bound on Eq. (1), $\forall q(x,y) \in \Delta$, with $\Delta$ denoting the probability simplex, $H$ indicating the entropy and $\mathbb{E}$ referring to the expectation w.r.t. the stated distribution. The bound holds with equality for the $q^*_{(x,y)}(h)$ minimizing this cost function.

Proof: The partition function is the conjugate dual of the entropy, hence:

$$
-\epsilon \ln \sum_{h} \exp \left( \frac{w^T \phi(x, \hat{h}) + \ell^c(x, (y, \hat{h}))}{\epsilon} \right) = \min_{q(x,y) \in \Delta} -\epsilon H(q(x,y)) - \sum_{h} q(x,y)(\hat{h}) (w^T \phi(x, (y, \hat{h})) + \ell^c(x, (y, \hat{h}))),
$$

(3)

to obtain the problem stated in Eq. (2). It is easy to see that the cost function given in Eq. (2) is convex in $w$ and $q(x,y)$ separately. However not jointly convex in $w$ and $q(x,y)$. Neglecting minimization w.r.t. $q(x,y)$ results in an upper bound to the original problem. The original problem is attained for optimal $q^*_{(x,y)}$. $\square$

3. Proof of Theorem 1

Theorem 1 The approximation of the program in Eq. (2) takes the form given in Program 1 where $\phi_{(x,y), i}(s_i) = \ell_{(x,y), i}(x, s_i) + \sum_{r : i \in E_r} w_{r, i} \phi_{r, i}(x, s_i)$ and $\phi_{(x,y), \alpha}(s_\alpha) = \ell_{(x,y), \alpha}(x, s_\alpha) + \sum_{r : \alpha \in E_r} w_{r, \alpha} \phi_{r, \alpha}(x, s_\alpha)$.

Proof: Assume that the each element of the feature vector $\phi$ decomposes into a graphical model structure, i.e., the $r^\text{th}$ element takes the following form:

$$
\phi_{r}(x, s) = \sum_{\alpha \in E_r} \phi_{r, \alpha}(x, s_\alpha) + \sum_{i \in S_r} \phi_{r, i}(x, s_i).
$$

(4)

with $E_r, S_r$ the sets of factors and variables. Note that each feature is described by a bipartite factor graph $G_r$ with nodes originating from the variable set $S_r$ and factors from $E_r$. An edge connects a single node $i \in S_r$ to a factor $\alpha \in E_r$ if $i \in \alpha$. Consider the factor graph $G = \bigcup_r G_r$ where we define the set of neighbors $N(i) := \{ \alpha : i \in \alpha \land \alpha \in E \}$ and $N(\alpha) := \{ i : i \in \alpha \land i \in S \}$.
Supplementary Material

Program 1 Approximated structured prediction with latent variables

\[
\begin{align*}
\min_{d,h,w} & \left\{ \frac{C}{2} \|w\|^2 + \sum_{(x,y) \in \mathcal{D}} \left( \sum_{i \in \mathcal{S}} \epsilon c_i \ln \sum_{s_i} \exp \left( \frac{\phi(x,y),i(s_i) - \sum_{a \in N(i)} \lambda(x,y),i \rightarrow a(s_i)}{\epsilon c_i} \right) + \sum_{a \in E} \epsilon c_a \ln \sum_{s_a} \exp \left( \frac{\phi(x,y),a(s_a) + \sum_{i \in N(a)} \lambda(x,y),i \rightarrow a(s_i)}{\epsilon c_a} \right) \right) \right\} \\
& \text{s.t.} \quad \sum_{h \in \mathcal{H}} d(x,y),i(h_i) = d(x,y),i(h_i) \quad \forall (x,y), i \in \mathcal{H}, \alpha \in N(i), h_i \in \mathcal{S}_i \quad \Rightarrow d(x,y) \in \mathcal{C}(x,y) \quad \forall (x,y) \in \mathcal{D}
\end{align*}
\]

In many applications the loss functions \( \ell \) and \( \ell^c \) factorize in a similar way and are easily included into the graphical model \( G \), i.e.,

\[
\begin{align*}
\ell(x,y)(s) &= \sum_{i \in \mathcal{S}} \ell(x,y),i(s_i) + \sum_{a \in E} \ell(x,y),a(s_a), \quad \text{(5)} \\
\ell^c(x,y)(h) &= \sum_{i \in \mathcal{H}} \ell^c(x,y),i(s_i) + \sum_{a \in E^c} \ell^c(x,y),a((y, h)_a). \quad \text{(6)}
\end{align*}
\]

Let \( d(x,y),i, d(x,y),a \) be the marginals of \( q(x,y)(h) \). Then,

\[
\begin{align*}
\sum_{h} \ell^c(x,y)((y, h))q(x,y)(h) &= \ell^c(x,y)(y) = \sum_{i \in \mathcal{Y}, y_i} \ell^c(x,y),i(x, y_i) + \sum_{i \in \mathcal{H}, h_i} \ell^c(x,y),i(x, h_i)d(x,y),i(h_i) + \sum_{a \in E^c} \ell^c(x,y),a((y, h)_a) \quad \text{(7)} \\
\sum_{h} \phi_r(x, y, h)q(x,y)(h) &= \nu(x,y),r = \sum_{i \in \mathcal{Y}} \phi_r,i(x, y_i) + \sum_{i \in \mathcal{H}} \phi_r,i(x, h_i)d(x,y),i(h_i) + \sum_{a \in E} \phi_r,a(x, (y, h)_a)d(x,y),a(h_a) \quad \text{(8)}
\end{align*}
\]

Note that \( \ell^c(x,y),i(x, y_i) \) has no effect and can be neglected. We approximate the entropy via local entropy terms and introduce counting numbers \( \hat{c}_i, \hat{c}_a \). Moreover, we approximate the marginal polytope via the local polytope
given by the marginalization and simplex constraints. All in all we obtain the following approximated program:

$$
\min_{w,d,\sigma} \frac{C}{p} \|w\|_p^p + \sum_{(x,y) \in D} \left( \epsilon \ln \sum_s \exp \frac{\sigma(x,s) + \ell_{x,y}(s)}{\epsilon} - \sum_r w_r v_{x,y,r} \right) \left( f_2 \right)
$$

$$
- \ell_{x,y}(y) - \frac{\epsilon c_i H(d_{x,y},i)}{\epsilon} - \frac{\epsilon c_{i\alpha} H(d_{x,y},\alpha)}{\epsilon} \left( f_3 \right)
$$

s.t. \forall (x,y), i \in \mathbb{H}, \alpha \in N(i), h_i \sum_{h_{\alpha}\setminus h_i} d_{x,y},\alpha(h_{\alpha}) = d_{x,y},i(h_i)

$$
d_{x,y},i, d_{x,y},\alpha \in \Delta \quad \forall (x,y), s \quad \sigma(x,s) = w^T \phi(x,s)
$$

For clarity of the presentation let us now address minimization w.r.t. $w$ and $\sigma$ without regards to $f_3$ and the marginalization constraints on $d$ required for global consistency, i.e.,

$$
\min_{w,\sigma} \sum_{(x,y)} \epsilon \ln \sum_s \exp \frac{\sigma(x,s) + \ell_{x,y}(s)}{\epsilon} - \sum_r w_r v_s + \frac{C}{p} \|w\|_p^p \tag{10}
$$

s.t. \forall (x,y), s \quad \sigma(x,s) = w^T \phi(x,s) \tag{11}

In a subsequent step we minimize the Lagrangian w.r.t. the primal variables $w$ and $\sigma$, i.e.,

$$
\sum_{(x,y)} \min_{\sigma} \left( \epsilon \ln \sum_s \exp \frac{\sigma(x,s) + \ell_{x,y}(s)}{\epsilon} - \sum_s p(x,y)(s)\sigma(x,s) \right) \tag{12}
$$

$$
\min_{w} \left( \frac{C}{p} \|w\|_p^p + w^T \left( \sum_{(x,y),s} p(x,y)(s)\phi(x,s) - v \right) \right) \tag{13}
$$

Note that we have introduced Lagrangian multipliers $p_{x,y}(s) \forall (x,y), s$. Analytically carrying out the minimization, we obtain the following dual problem

$$
\max_{p(x,y) \in \Delta} \sum_{(x,y)} \epsilon H(p(x,y)) + \sum_s p(x,y)(s)(w^T \phi(x,s) + \ell_{x,y}(s)) - \frac{C^{1-q}}{q} \sum_r \left| \sum_{(x,y),s} p(x,y)(s) \phi_r(x,s) - v_{x,y,r} \right|^q \tag{14}
$$

Similar to our previous argument we assume that the probability distribution $p(x,y)(s)$ is defined via marginals $b_{x,y},i, b_{x,y},\alpha$. Again we approximate the marginal polytope via a local one given by the marginalization and simplex constraints. Hence we obtain,

$$
\sum_s \ell_{x,y}(s)p_{x,y}(s) = l_{x,y} = \sum_{i \in S, s_i} \ell_{x,y},i(s_i)b_{x,y},i(s_i) + \sum_{\alpha \in E, s_{\alpha}} \ell_{x,y},\alpha(s_{\alpha})b_{x,y},\alpha(s_{\alpha}) \tag{15}
$$

$$
\sum_s \phi_r(x,s)p_{x,y}(s) = u_{x,y,r} = \sum_{i \in S, s_i} \phi_r,i(x,s_i)b_{x,y},i(s_i) + \sum_{\alpha \in E, s_{\alpha}} \phi_r,\alpha(x,s_{\alpha})b_{x,y},\alpha(s_{\alpha}) \tag{16}
$$

$$
u_r = \sum_{(x,y)} u_{x,y,r} \tag{17}$$
The approximated dual program with counting numbers $c_i$, $c_\alpha$ is then
\[
\max_{b,y} \sum_{(x,y) \in \mathcal{S}} c_i H(b(x,y),i) + \sum_{\alpha \in E} c_\alpha H(b(y,x),\alpha) + \sum_{i \in \mathcal{S}, s_i} b(x,y),i(s_i) \phi(x,y),i(s_i) + \sum_{\alpha \in E, s_\alpha} b(x,y),\alpha(s_\alpha) \phi(x,y),\alpha(s_\alpha) - \\
- \frac{C^1 - q}{q} \sum_r |u_r - v_r|^q \tag{18}
\]
subject to
\[
\forall \alpha \in E, i \in N(\alpha), s_i, (x,y) \sum_{s_\alpha \setminus s_i} b(x,y),\alpha(s_\alpha) = b(x,y),i(s_i) \tag{19}
\]
and
\[
b(x,y),i, b(x,y),\alpha \in \Delta \tag{20}
\]
By maximizing the obtained Lagrangian with Lagrange multipliers $\lambda(x,y),i \rightarrow \alpha(s_i)$ for the constraint given in Eq. (19) and after combining with $f_3$ and corresponding constraints on $d$ we obtain the claim. $\square$

4. Proof of Claim 2

Claim 2 Algorithm 1 is guaranteed to decrease the cost function of Program 1 at every iteration and guaranteed to converge to a minimum or a saddle point for $\epsilon, c_i, c_\alpha, \hat{c}_i, \hat{c}_\alpha > 0$.

Proof: Recalling Theorem 5 in (Yuille & Rangarajan, 2003) we notice that alternating optimization of the approximated program in Theorem 1 w.r.t. $d$ and $\lambda, w$ is equivalent to CCCP, which is guaranteed to converge to a stationary point if the respective functions are convex. Convexity is ensured for $\epsilon, c_i, c_\alpha, \hat{c}_i$ and $\hat{c}_\alpha > 0$. $\square$

5. Algorithmic Details

Similar to a CCCP approach (Yuille & Rangarajan, 2003) we address optimization of Program 1 by alternating solving two tasks. One optimization considers only the beliefs $d$, while the other one operates on $\lambda$ and $w$. When updating the beliefs $d$, we obtain the ‘latent variable prediction problem’ which requires solving

\[
\sum_{(x,y) \in \mathcal{D}} \min_{d(x,y)} f_2(w, d) + f_3(d) \tag{21}
\]
subject to
\[
\forall (x,y) \in \mathcal{D} \quad d(x,y) \in C(x,y). 
\]
Explicitly it reads

\[
\sum_{(x,y)} \max_d \sum_{\hat{c}_i \in \mathcal{E}} c_i H(d(x,y),i) + \sum_{\alpha \in E_\mathcal{E}} \epsilon_\alpha H(d(x,y),\alpha) + \sum_{i \in \mathcal{E}, h_i} d(x,y),i(h_i) \left( \sum_{r,i \in \mathcal{E}, h_r} w_r \phi_r,i(x, h_i) + l_r^i(x,y),\alpha(x, y, h_\alpha) \right) \tag{22}
\]
subject to
\[
\forall (x,y) \quad d(x,y),i, d(x,y),\alpha \in \Delta \tag{23}
\]
This is a standard (convex) belief propagation task with local potentials $\phi_r^i(x,y),i(h_i) = \sum_{i \in \mathcal{S}_r} w_r \phi_r,i(x, h_i) + l_r^i(x,y),i(h_i)$ and clique potentials $\phi_r^i(x,y),\alpha(h_\alpha) = \sum_{r \in \mathcal{S}_r} w_r \phi_r,\alpha(x, y, h_\alpha) + l_r^i(x,y),\alpha(x, y, h_\alpha)$. We solve it by minimizing its unconstrained dual program using a message passing algorithm on the graph defined by the nodes $i \in \mathcal{E}$ and corresponding cliques. The algorithm is guaranteed to find the optimal solution for counting numbers $c_i$, $c_\alpha$ and $\epsilon > 0$.

To optimize for model parameters $w$ and messages $\lambda$ we gradually solve the following unconstrained problem:

\[
\min_{w, \lambda} f_1(w, \lambda) + f_2(w, d). \tag{23}
\]
It explicitly reads as

\[ \min_{\lambda,w} \frac{C}{2} \|w\|^2 + \sum_{(x,y) \in D} \left( \sum_{i \in S} \varepsilon_c \ln \sum_{s_i} \exp \left( \frac{\phi(x,y),i(s_i) - \sum_{\alpha \in N(i)} \lambda(x,y),i-\alpha(s_i)}{\varepsilon_c} \right) \right) + \sum_{\alpha \in E} \varepsilon_c \ln \sum_{s_\alpha} \exp \left( \frac{\phi(x,y),\alpha(s_\alpha) + \sum_{i \in N(\alpha)} \lambda(x,y),i-\alpha(s_i)}{\varepsilon_c} \right) \]  

(24)

\[ - \sum r \cdot \left( \sum_{i \in E} \phi_{r,i}(x,y_i) + \sum_{i \in h_i} \phi_{r,i}(x,h_i) + \sum_{\alpha \in E} \phi_{r,\alpha}(x,(y,h),\alpha) \right) \]  

This is an approximated structured prediction task with empirical means \( v_r \).

**Lemma 1** Given a node \( i \in S \) of the graphical model \( G \), the optimal \( \lambda(x,y),i-\alpha(s_i) \) \( \forall \alpha \in N(i), s_i \in S_i, (x,y) \in D \) of Theorem 1 (resp. Eq. (23)) satisfies

\[ \lambda(x,y),i-\alpha(s_i) \propto \frac{c_\alpha}{c_1 + \sum_{\alpha \in N(i)} c_\alpha} \left( \phi(x,y),i(s_i) + \sum_{\alpha \in N(i)} \mu(x,y),\alpha-\alpha(s_i) \right) - \mu(x,y),\alpha-\alpha(s_i). \]  

(25)

with

\[ \mu(x,y),\alpha-\alpha(s_i) = \varepsilon_c \alpha \ln \sum_{s_\alpha \setminus s_i} \exp \frac{\phi(x,y),\alpha(s_\alpha) + \sum_{u \in N(\alpha) \setminus i} \lambda(x,y),u-\alpha(s_u)}{\varepsilon_c} \]  

(26)

**Proof:**

To update the messages we take every \( (x,y),i \in S \) and obtain an analytic solution of the first order optimality condition w.r.t. \( \lambda(x,y),i-\alpha(s_i) \) \( \forall \alpha \in N(i), s_i \). We obtain the following simplified optimization problem:

\[ \min_{\lambda(x,y),i-\alpha(s_i)} \]  

\[ \varepsilon_c \sum_{s_i} \exp \frac{\phi(x,y),i(s_i) - \sum_{\alpha \in N(i)} \lambda(x,y),i-\alpha(s_i)}{\varepsilon_c} + \sum_{\alpha \in N(i)} \sum_{s_i} \exp \frac{\mu(x,y),\alpha-\alpha(s_i) + \lambda(x,y),i-\alpha(s_i)}{\varepsilon_c} \]  

(27)

We find the optimal \( \lambda(x,y),i-\alpha(s_i) \) \( \forall \alpha \in N(i), s_i \) whenever the gradient vanishes which is achieved for

\[ \lambda(x,y),i-\alpha(s_i) \propto \frac{c_\alpha}{c_1 + \sum_{\alpha \in N(i)} c_\alpha} \left( \phi(x,y),i(s_i) + \sum_{\alpha \in N(i)} \mu(x,y),\alpha-\alpha(s_i) \right) - \mu(x,y),\alpha-\alpha(s_i). \]  

(28)

\[ \square \]

**Lemma 2** The gradient of the approximated program given in Theorem 1 (resp. Eq. (23)) w.r.t. \( w_r \) equals

\[ \sum_{(x,y)} \left( \sum_{i \in E \setminus s_i} b_{(x,y),i}(s_i) \phi_{r,i}(x,s_i) + \sum_{\alpha \in E \setminus s_i} b_{(x,y),\alpha}(s_\alpha) \phi_{r,\alpha}(x,s_\alpha) \right) - v_r + C |w_r|^{p-1} \text{sign}(w_r) \]  

(29)

with

\[ b_{(x,y),i}(s_i) \propto \exp \left( \frac{\phi(x,y),i(s_i) - \sum_{\alpha \in N(i)} \lambda(x,y),i-\alpha(s_i)}{\varepsilon_c} \right) \]  

(30)

\[ b_{(x,y),\alpha}(s_\alpha) \propto \exp \left( \frac{\phi(x,y),\alpha(s_\alpha) + \sum_{i \in N(\alpha)} \lambda(x,y),i-\alpha(s_i)}{\varepsilon_c} \right) \]  

(31)
Program 2 Message passing algorithm for approximated structured prediction with latent variables

Repeat

1. Solve ‘latent variable prediction problem’ (Program (22)) until convergence

2. For each \((x,y) \in D, i \in \mathbb{S}\) (by Lemma 1):

   \[
   \forall \alpha \in N(i), s_i \quad \lambda_{(x,y),i \to \alpha}(s_i) \propto \frac{c_{\alpha}}{c_i + \sum_{\alpha' \in N(i)} c_{\alpha'}} \left( \phi_{(x,y),i}(s_i) + \sum_{\alpha' \in N(i)} \mu_{(x,y),\alpha' \to i}(s_i) \right) - \mu_{(x,y),\alpha \to i}(s_i)
   \]

3. For each \(r\) (by Lemma 2): find a stepsize \(\eta\) that reduces \(f_1 + f_2\) and update

   \[
   w_r \leftarrow w_r - \eta \left( \sum_{(x,y) \in \mathbb{S}_r, s_i} b_{(x,y),i}(s_i) \phi_{r,i}(x,s_i) + \sum_{\alpha \in E_r, s_\alpha} b_{(x,y),\alpha}(s_\alpha) \phi_{r,\alpha}(x,s_\alpha) \right) - v_r + C|w_r|^{p-1} \text{sign}(w_r)
   \]

Proof: This is a direct computation of the gradient w.r.t. \(w_r\). \(\Box\)

Hence the complete algorithm reads as detailed in Program 2

References