Supplementary Material - Efficient Structured Prediction with Latent Variables for General Graphical Models

1. Introduction

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

$$\frac{C}{p}\|w\|_p^p + \sum_{(x,y)\in\mathcal{D}}\left(\epsilon \ln \sum_{\hat{s}\in\mathcal{S}} \exp\left(\frac{w^\top \phi(x,\hat{s}) + \ell_{(x,y)}(\hat{s})}{\epsilon}\right) - \epsilon \ln \sum_{\hat{h}\in\mathcal{H}} \exp\left(\frac{w^\top \phi(x,(y,\hat{h})) + \ell^c_{(x,y)}((y,\hat{h}))}{\epsilon}\right)\right). \quad (1)$$

2. Proof of Claim 1

Claim 1 *The following function*

$$\frac{C}{p}\|w\|_p^p + \sum_{(x,y)}\left(\epsilon \ln \sum_{\hat{s}\in\mathcal{S}} \exp\left(\frac{w^\top \phi(x,\hat{s}) + \ell_{(x,y)}(\hat{s})}{\epsilon}\right) - \epsilon H(q_{(x,y)}) - \mathbb{E}_{q_{(x,y)}}[w^\top \phi(x,(y,\hat{h})) + \ell^c(x,(y,\hat{h}))]\right), \quad (2)$$

convex in w and $q_{(x,y)}$ separately, is an upper bound on Eq. (1), $\forall q_{(x,y)}(h) \in \Delta$, with Δ 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:

$$-\epsilon \ln \sum_{\hat{h}} \exp \frac{w^\top \phi(x,\hat{y},\hat{h}) + \ell^c_{(x,y)}((y,\hat{h}))}{\epsilon} = \min_{q_{(x,y)}\in\Delta} -\epsilon H(q_{(x,y)}) - \sum_{\hat{h}} q_{(x,y)}(\hat{h})(w^\top \phi(x,y,\hat{h}) + \ell^c_{(x,y)}((y,\hat{h}))), \quad (3)$$

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 $q_{(x,y)}$ $\forall (x,y)$ 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 $\phi_{(x,y),i}(s_i) = \ell_{(x,y),i}(x,s_i) + \sum_{r:i\in\mathbb{S}_r} w_r \phi_{r,i}(x,s_i)$ and $\phi_{(x,y),\alpha}(s_\alpha) = \ell_{(x,y),\alpha}(x,s_\alpha) + \sum_{r:\alpha\in E_r} w_r \phi_{r,\alpha}(x,s_\alpha)$.*

Proof: Assume that the each element of the feature vector ϕ decomposes into a graphical model structure, *i.e.*, the $r-$th element takes the following form:

$$\phi_r(x,s) = \sum_{\alpha\in E_r} \phi_{r,\alpha}(x,s_\alpha) + \sum_{i\in\mathbb{S}_r} \phi_{r,i}(x,s_i). \quad (4)$$

with E_r, $\mathbb{S}_r$ 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 $\mathbb{S}_r$ and factors from E_r. An edge connects a single node $i \in \mathbb{S}_r$ to a factor $\alpha \in E_r$ iff $i \in \alpha$. Consider the factor graph $G = \bigcup_r G_r$ where we define the set of neighbors $N(i) := \{\alpha : i \in \alpha \ \forall\ \alpha \in E\}$ and $N(\alpha) := \{i : i \in \alpha \ \forall\ i \in \mathbb{S}\}$.

Program 1 *Approximated structured prediction with latent variables*

$$
\begin{aligned}
\min_{d,\lambda,w} \quad & f_1 \left\{ \begin{array}{l} \frac{C}{2}\|w\|_2^2 + \sum_{(x,y)\in\mathcal{D}} \left(\sum_{i\in\mathbb{S}} \epsilon c_i \ln \sum_{s_i} \exp\left(\frac{\phi_{(x,y),i}(s_i) - \sum_{\alpha\in N(i)} \lambda_{(x,y),i\to\alpha}(s_i)}{\epsilon c_i} \right) + \right. \\ \left. \sum_{\alpha\in E} \epsilon c_\alpha \ln \sum_{s_\alpha} \exp\left(\frac{\phi_{(x,y),\alpha}(s_\alpha) + \sum_{i\in N(\alpha)} \lambda_{(x,y),i\to\alpha}(s_i)}{\epsilon c_\alpha} \right) \right) - \end{array} \right. \\
& f_2 \left\{ -\sum_r w_r \left(\sum_{(x,y)} \left(\sum_{i\in\mathbb{Y}} \phi_{r,i}(x,y_i) + \sum_{i\in\mathbb{H},h_i} \phi_{r,i}(x,h_i) d_{(x,y),i}(h_i) + \sum_{\alpha\in E,h_\alpha} \phi_{r,\alpha}(x,(y,h)_\alpha) d_{(x,y),\alpha}(h_\alpha) \right) \right) \right. \\
& f_3 \left\{ \begin{array}{l} -\sum_{(x,y)} \left(\sum_{i\in\mathbb{H},h_i} \ell^c_{(x,y),i}(x,h_i) d_{(x,y),i}(h_i) + \sum_{\alpha\in E_{\mathbb{H}},h_\alpha} \ell^c_{(x,y),\alpha}(x,(y,h)_\alpha) d_{(x,y),\alpha}(h_\alpha) \right) \\ -\sum_{(x,y)} \left(\sum_{i\in\mathbb{H}} \epsilon \hat{c}_i H(d_{(x,y),i}) + \sum_{\alpha\in E_{\mathbb{H}}} \epsilon \hat{c}_\alpha H(d_{(x,y),\alpha}) \right) \end{array} \right. \\
\text{s.t.} \quad & \left. \begin{array}{l} \sum_{h_\alpha\setminus h_i} d_{(x,y),\alpha}(h_\alpha) = d_{(x,y),i}(h_i) \quad \forall (x,y), i\in\mathbb{H}, \alpha\in N(i), h_i\in\mathcal{S}_i \\ d_{(x,y),i}, d_{(x,y),\alpha} \in \underline{\Delta} \end{array} \right\} := d_{(x,y)} \in \mathcal{C}_{(x,y)} \quad \forall (x,y)\in\mathcal{D}
\end{aligned}
$$

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

$$
\begin{aligned}
\ell_{(x,y)}(s) &= \sum_{i\in\mathbb{S}} \ell_{(x,y),i}(s_i) + \sum_{\alpha\in E} \ell_{(x,y),\alpha}(s_\alpha), && (5) \\
\ell^c_{(x,y)}(\hat{h}) &= \sum_{i\in\mathbb{H}} \ell^c_{(x,y),i}(\hat{s}_i) + \sum_{\alpha\in E_{\mathbb{H}}} \ell^c_{(x,y),\alpha}((y,\hat{h})_\alpha). && (6)
\end{aligned}
$$

Let $d_{(x,y),i}, d_{(x,y),\alpha}$ be the marginals of $q_{(x,y)}(h)$. Then,

$$
\begin{aligned}
\sum_{\hat{h}} \ell^c_{(x,y)}((y,\hat{h})) q_{(x,y)}(\hat{h}) = \ell^c_{(x,y)}(y) &= \sum_{i\in\mathbb{Y},y_i} \ell^c_{(x,y),i}(x,y_i) + \sum_{i\in\mathbb{H},h_i} \ell^c_{(x,y),i}(x,h_i) d_{(x,y),i}(h_i) + \\
&\quad + \sum_{\alpha\in E_{\mathbb{H}},h_\alpha} \ell^c_{(x,y),\alpha}(x,(y,h)_\alpha) d_{(x,y),\alpha}(h_\alpha) && (7) \\
\sum_{\hat{h}} \phi_r(x,y,\hat{h}) q_{(x,y)}(\hat{h}) = v_{(x,y),r} &= \sum_{i\in\mathbb{Y}} \phi_{r,i}(x,y_i) + \sum_{i\in\mathbb{H},h_i} \phi_{r,i}(x,h_i) d_{(x,y),i}(h_i) + \\
&\quad + \sum_{\alpha\in E,h_\alpha} \phi_{r,\alpha}(x,(y,h)_\alpha) d_{(x,y),\alpha}(h_\alpha) && (8) \\
v_r &= \sum_{(x,y)} v_{(x,y),r} && (9)
\end{aligned}
$$

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 $\hat{c}_i, \hat{c}_\alpha$. 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:

$$\begin{aligned}
\min_{w,d,\sigma} \quad & \frac{C}{p}\|w\|_p^p + \sum_{(x,y)\in\mathcal{D}} \Bigg(\epsilon \ln \sum_s \exp \frac{\sigma(x,s)+\ell_{(x,y)}(s)}{\epsilon} \underbrace{- \sum_r w_r v_{(x,y),r}}_{f_2} - \\
& \underbrace{-l^c_{(x,y)}(y) - \sum_{i\in\mathbb{H}} \epsilon \hat{c}_i H(d_{(x,y),i}) - \sum_{\alpha\in E_{\mathbb{H}}} \epsilon \hat{c}_\alpha H(d_{(x,y),\alpha})}_{f_3} \Bigg) \\
\text{s.t.} \quad & \forall (x,y), i\in\mathbb{H}, \alpha\in N(i), h_i \quad \sum_{h_\alpha\setminus h_i} d_{(x,y),\alpha}(h_\alpha) = d_{(x,y),i}(h_i) \\
& d_{(x,y),i}, d_{(x,y),\alpha} \in \underline{\Delta} \\
& \forall (x,y), s \quad \sigma(x,s) = w^\top \phi(x,s)
\end{aligned}$$

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

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

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

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

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

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

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

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

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,

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

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

$$u_r \quad = \quad \sum_{(x,y)} u_{(x,y),r} \tag{17}$$

The approximated dual program with counting numbers c_i, c_α is then

$$\begin{aligned}
\max_{b,u} \quad & \sum_{(x,y)} \sum_{i\in\mathbb{S}} \epsilon c_i H(b_{(x,y),i}) + \sum_{\alpha\in E} \epsilon c_\alpha H(b_{(x,y),\alpha}) + \sum_{i\in\mathbb{S},s_i} b_{(x,y),i}(s_i)\phi_{(x,y),i}(s_i) + \sum_{\alpha\in E,s_\alpha} b_{(x,y),\alpha}(s_\alpha)\phi_{(x,y),\alpha}(s_\alpha) - \\
& -\frac{C^{1-q}}{q}\sum_r |u_r - v_r|^q \qquad (18)\\
\text{s.t.} \quad & \forall \alpha\in E, i\in N(\alpha), s_i, (x,y) \quad \sum_{s_\alpha\setminus s_i} b_{(x,y),\alpha}(s_\alpha) = b_{(x,y),i}(s_i) \qquad (19)\\
& b_{(x,y),i}, b_{(x,y),\alpha} \in \Delta \qquad (20)
\end{aligned}$$

By maximizing the obtained Lagrangian with Lagrange multipliers $\lambda_{(x,y),i\to\alpha}(s_i)$ for the constraint given in Eq. (19) and after combining with f_3 and corresponding constraints on d we obtain the claim. □

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 λ, 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 $\hat{c}_\alpha > 0$. □

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 λ and w. When updating the beliefs d, we obtain the 'latent variable prediction problem' which requires solving

$$\begin{aligned}
\sum_{(x,y)\in\mathcal{D}} \min_{d_{(x,y)}} \quad & f_2(w,d) + f_3(d) \\
\text{s.t.} \quad & \forall (x,y)\in\mathcal{D} \quad d_{(x,y)} \in \mathcal{C}_{(x,y)}.
\end{aligned} \qquad (21)$$

Explicitly it reads as

$$\begin{aligned}
\sum_{(x,y)} \quad & \max_d \sum_{i\in\mathbb{H}} \epsilon\hat{c}_i H(d_{(x,y),i}) + \sum_{\alpha\in E_\mathbb{H}} \epsilon\hat{c}_\alpha H(d_{(x,y),\alpha}) + \sum_{i\in\mathbb{H},h_i} d_{(x,y),i}(h_i)\left(\sum_{r:i\in\mathbb{S}_r} w_r\phi_{r,i}(x,h_i) + l^c_{(x,y),i}(h_i)\right) + \\
& \sum_{\alpha\in E_\mathbb{H},h_\alpha} d_{(x,y),\alpha}(h_\alpha)\left(\sum_{r:\alpha\in E_r} w_r\phi_{r,\alpha}(x,(y,h)_\alpha) + l^c_{(x,y),\alpha}(x,(y,h)_\alpha)\right) \qquad (22)\\
\text{s.t.} \quad & \forall (x,y), i\in\mathbb{H}, \alpha\in N(i), h_i \quad \sum_{h_\alpha\setminus h_i} d_{(x,y),\alpha}(h_\alpha) = d_{(x,y),i}(h_i) \\
& \forall (x,y) \quad d_{(x,y),i}, d_{(x,y),\alpha} \in \Delta
\end{aligned}$$

This is a standard (convex) belief propagation task with local potentials $\phi^c_{(x,y),i}(h_i) = \sum_{r:i\in\mathbb{S}_r} w_r\phi_{r,i}(x,h_i) + l^c_{(x,y),i}(h_i)$ and clique potentials $\phi^c_{(x,y),\alpha}(h_\alpha) = \sum_{r:\alpha\in E_r} w_r\phi_{r,\alpha}(x,(y,h)_\alpha) + l^c_{(x,y),\alpha}(x,(y,h)_\alpha)$. We solve it by minimizing its unconstrained dual program using a message passing algorithm on the graph defined by the nodes $i\in\mathbb{H}$ and corresponding cliques. The algorithm is guaranteed to find the optimal solution for counting numbers $\hat{c}_i, \hat{c}_\alpha$ and $\epsilon > 0$.

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

$$\min_{w,\lambda} f_1(w,\lambda) + f_2(w,d). \qquad (23)$$

It explicitly reads as

$$\min_{\lambda,w} \quad \frac{C}{2}\|w\|_2^2 + \sum_{(x,y)\in\mathcal{D}} \left(\sum_{i\in\mathbb{S}} \epsilon c_i \ln \sum_{s_i} \exp\left(\frac{\phi_{(x,y),i}(s_i) - \sum_{\alpha\in N(i)} \lambda_{(x,y),i\to\alpha}(s_i)}{\epsilon c_i} \right) + \right. \tag{24}$$

$$\left. \sum_{\alpha\in E} \epsilon c_\alpha \ln \sum_{s_\alpha} \exp\left(\frac{\phi_{(x,y),\alpha}(s_\alpha) + \sum_{i\in N(\alpha)} \lambda_{(x,y),i\to\alpha}(s_i)}{\epsilon c_\alpha} \right) \right) -$$

$$-\sum_r w_r \underbrace{\left(\sum_{(x,y)} \left(\sum_{i\in\mathbb{Y}} \phi_{r,i}(x,y_i) + \sum_{i\in\mathbb{H},h_i} \phi_{r,i}(x,h_i) d_{(x,y),i}(h_i) + \sum_{\alpha\in E,h_\alpha} \phi_{r,\alpha}(x,(y,h)_\alpha) d_{(x,y),\alpha}(h_\alpha) \right) \right)}_{v_r}.$$

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

Lemma 1 *Given a node $i\in\mathbb{S}$ of the graphical model G, the optimal $\lambda_{(x,y),i\to\alpha}(s_i)$ $\forall\alpha\in N(i), s_i\in\mathcal{S}_i, (x,y)\in\mathcal{D}$ of Theorem 1 (resp. Eq. (23)) satisfies*

$$\lambda_{(x,y),i\to\alpha}(s_i) \propto \frac{c_\alpha}{c_i + \sum_{\alpha\in N(i)} c_\alpha} \left(\phi_{(x,y),i}(s_i) + \sum_{\alpha\in N(i)} \mu_{(x,y),\alpha\to i}(s_i) \right) - \mu_{(x,y),\alpha\to i}(s_i). \tag{25}$$

with

$$\mu_{(x,y),\alpha\to i}(s_i) = \epsilon c_\alpha \ln \sum_{s_\alpha\setminus s_i} \exp \frac{\phi_{(x,y),\alpha}(s_\alpha) + \sum_{u\in N(\alpha)\setminus i} \lambda_{(x,y),u\to\alpha}(s_u)}{\epsilon c_\alpha} \tag{26}$$

Proof:

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

$$\min_{\lambda_{(x,y),i\to\alpha}(s_i)} \quad \epsilon c_i \ln \sum_{s_i} \exp \frac{\phi_{(x,y),i}(s_i) - \sum_{\alpha\in N(i)} \lambda_{(x,y),i\to\alpha}(s_i)}{\epsilon c_i} + \tag{27}$$

$$\sum_{\alpha\in N(i)} \epsilon c_\alpha \ln \sum_{s_i} \exp \frac{\mu_{(x,y),\alpha\to i}(s_i) + \lambda_{(x,y),i\to\alpha}(s_i)}{\epsilon c_\alpha}$$

We find the optimal $\lambda_{(x,y),i\to\alpha}(s_i)$ $\forall\alpha\in N(i), s_i$ whenever the gradient vanishes which is achieved for

$$\lambda_{(x,y),i\to\alpha}(s_i) \propto \frac{c_\alpha}{c_i + \sum_{\alpha\in N(i)} c_\alpha} \left(\phi_{(x,y),i}(s_i) + \sum_{\alpha\in N(i)} \mu_{(x,y),\alpha\to i}(s_i) \right) - \mu_{(x,y),\alpha\to i}(s_i). \tag{28}$$

∎

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

$$\sum_{(x,y)} \left(\sum_{i\in\mathbb{S}_r,s_i} b_{(x,y),i}(s_i)\phi_{r,i}(x,s_i) + \sum_{\alpha\in E_r,s_\alpha} b_{(x,y),\alpha}(s_\alpha)\phi_{r,\alpha}(x,s_\alpha) \right) - v_r + C|w_r|^{p-1}\operatorname{sign}(w_r) \tag{29}$$

with

$$b_{(x,y),i}(s_i) \quad \propto \quad \exp\left(\frac{\phi_{(x,y),i}(s_i) - \sum_{\alpha\in N(i)} \lambda_{(x,y),i\to\alpha}(s_i)}{\epsilon c_i} \right)$$

$$b_{(x,y),\alpha}(s_\alpha) \quad \propto \quad \exp\left(\frac{\phi_{(x,y),\alpha}(s_\alpha) + \sum_{i\in N(\alpha)} \lambda_{(x,y),i\to\alpha}(s_i)}{\epsilon c_\alpha} \right)$$

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 $(x,y) \in \mathcal{D}, i \in \mathbb{S}$ (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 η that reduces $f_1 + f_2$ and update

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

Proof: This is a direct computation of the gradient w.r.t. w_r. □

Hence the complete algorithm reads as detailed in Program 2

References

Yuille, A. L. and Rangarajan, A. The Concave-Convex Procedure (CCCP). *Neural Computation*, 2003.