Parameter Learning and Convergent Inference for Dense Random Fields – Supplementary Material

Philipp Krähenbühl PHILKR@CS.STANFORD.EDU
Vladlen Koltun VLADLEN@CS.STANFORD.EDU
Computer Science Department, Stanford University, Stanford, CA 94305 USA

1. Differentiability of Mean Field Marginals

We show that the mean field marginals $\mathbf{q}^{(t)}(\boldsymbol{\theta})$ are continuously differentiable as a function of the parameters $\boldsymbol{\theta}$ for each iteration t. We will first show that the mean field marginals at $t = 0$ are differentiable. During the proof we rely on the strict positivity of the mean-field marginals $\mathbf{q}^{(t)}(\boldsymbol{\theta}) > 0$ for proper probability distributions. We then use induction and the implicit function theorem to proof differentiability for $t > 0$.

We will show differentiability for the convergent parallel mean field algorithm, the concave cross-entropy relaxation can be derived analogously.

$t = 0$. The mean field marginals $\mathbf{q}^{(0)}$ is always initialized to the unary term. This unary term is either a constant or a linear function (logistic regression), both of which are continuously differentiable.

Positivity. Recall that any mean field marginal is bounded by $0 \le \mathbf{q}_{(i,l)} \le 1$. What we will show here is that strict positivity holds $\mathbf{q}_{(i,l)} > 0$ for any proper probability distribution. A proper probability distribution is a distribution with a finite Gibbs energy for any assignment. For such a distribution both the unary and pairwise terms are finite, and equivalently the result of message passing $\mathbf{e}$ and the label compatibilities $\boldsymbol{\mu}^{(m)}$ in Equation 8. By rewriting the KKT-conditions (8) we get

$$\log \mathbf{q}_i + \lambda \mathbf{1} = \left(\sum_{m=1}^{C} \boldsymbol{\mu}^{(m)} \right) \mathbf{q}_i - \mathbf{e}_i,$$

which is guaranteed to be a finite value. At least one label l the marginal is lower bounded by $\mathbf{q}_{(i,l)} \ge \frac{1}{M}$, where M is the number of labels. Which implies that for at least one label l: $\log \mathbf{q}_{(i,l)}$ is finite, and hence λ needs to be finite. A finite λ implies $\log \mathbf{q}_i$ is finite and hence $\mathbf{q}_i > 0$.

Induction. We will now show that for any differentiable mean field marginals $\mathbf{q}^{(t-1)}(\boldsymbol{\theta})$, the marginals $\mathbf{q}^{(t)}$ are differentiable. Consider the Jacobian of the KKT conditions (8) with respect to $\mathbf{q}$ and λ:

$$J = \begin{bmatrix} \operatorname{diag}\left(\frac{1}{\mathbf{q}_i}\right) - \boldsymbol{\mu} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix},$$

where $\boldsymbol{\mu} = \sum_{m=1}^{C} \boldsymbol{\mu}^{(m)}$ is by definition negative definite. The inverse of this Jacobian is

$$J^{-1} = \begin{bmatrix} H^{-1} - \frac{H^{-1}\mathbf{1}\mathbf{1}^\top H^{-1}}{\mathbf{1}^\top H^{-1}\mathbf{1}} & \frac{H^{-1}\mathbf{1}}{\mathbf{1}^\top H^{-1}\mathbf{1}} \\ \frac{\mathbf{1}^\top H^{-1}}{\mathbf{1}^\top H^{-1}\mathbf{1}} & -\frac{1}{\mathbf{1}^\top H^{-1}\mathbf{1}} \end{bmatrix},$$

where $H = \operatorname{diag}\left(\frac{1}{\mathbf{q}_i}\right) - \boldsymbol{\mu}$ is positive definite and invertible. Now assume that the marginals $\mathbf{q}^{(t-1)}$ are differentiable. The sum of messages $\mathbf{e}^{(t)}$ is differentiable, since it is a sum of products of differentiable functions. By the implicit function theorem, the function $\mathbf{q}^{(t)}$ is unique and continuously differentiable, since the KKT conditions are continuously differentiable, $\mathbf{e}^{(t)}$ is continuously differentiable, and the Jacobian of the KKT conditions is invertible. This completes the proof by induction.

2. Invertability of the linear system (12)

We rewrite the linear system (12) in matrix form

$$M \begin{bmatrix} \frac{\partial \mathbf{q}_i^{(t)}}{\partial \boldsymbol{\theta}} \\ \frac{\partial \lambda}{\partial \boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathbf{e}_i^{(t)}}{\partial \boldsymbol{\theta}} \\ 0 \end{bmatrix},$$

where

$$M = \begin{bmatrix} \operatorname{diag}\left(\frac{1}{\mathbf{q}_i}\right) & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}.$$

In order to show that (12) is invertible, we need to show that M is invertible. We do this by providing the closed form solution of the inverse

$$M^{-1} = \begin{bmatrix} \operatorname{diag} \mathbf{q}_i - \mathbf{q}_i \mathbf{q}_i^\top & \mathbf{q}_i \\ \mathbf{q}_i^\top & -1 \end{bmatrix}.$$

It can easily be verified that $M^{-1}M = MM^{-1} = I$, since by definition $\mathbf{1}^\top \mathbf{q}_i = 1$.

The inverse M^{-1} also gives us the closed form solution in Equation 13, with $A_i^{(t)} = \mathbf{q}_i \mathbf{q}_i^\top - \operatorname{diag} \mathbf{q}_i$. Note that $-A_i^{(t)}$ is simply the upper left block if M^{-1}.

3. Convergence example

We present a small illustrative example, where our previous inference algorithm (Krähenbühl & Koltun, 2011) fails, but the two convergent alternatives converge.

Consider a two variable binary CRF with variables X_1, X_2, features $f_1 = 0$, $f_2 = 0.5$, and a simple Potts label compatibility with weight 5. The unary term is defined as $\psi(x_1) = 1_{[x_1=1]}$ and $\psi(x_2) = -2_{[x_1=1]}$. The pairwise term is given by

$$\psi(x_1, x_2) = \exp(-\frac{1}{8})5_{[x_1 \neq x_2]}.$$

For this example (Krähenbühl & Koltun, 2011) osculates between $q_{(1,1)} = 0.96, q_{(2,1)} = 0.09$ and $q_{(1,1)} = 0.01, q_{(2,1)} = 0.99$, while every other step increases the KL-divergence. Both our algorithm converge to the optimal solution $q_{(1,1)} = 0.99, q_{(2,1)} = 0.99$, optimizing the KL-divergence or concave approximation respectively.

References

Krähenbühl, Philipp and Koltun, Vladlen. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011.