

---

# Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials

Supplementary Material

---

**Philipp Krähenbühl**  
 Computer Science Department  
 Stanford University  
 philkr@cs.stanford.edu

**Vladlen Koltun**  
 Computer Science Department  
 Stanford University  
 vladlen@cs.stanford.edu

This document provides detailed derivations for the mean field approximation and our learning algorithm.

## 1 Mean Field Approximation

Let's recall the general fully connected CRF model,

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i < j} \psi_p(x_i, x_j). \quad (1)$$

Throughout this supplement, the indices  $i$  and  $j$  range from 1 to  $N$ . The pairwise edge potential  $\psi_p(x_i, x_j)$  is defined as a linear combination of Gaussian kernels  $k^{(m)}(\mathbf{f}_i, \mathbf{f}_j)$ :

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^K w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j). \quad (2)$$

The Gibbs distribution is given by

$$P(\mathbf{X}) = \frac{1}{Z} \tilde{P}(\mathbf{X}) = \frac{1}{Z} \exp \left( \sum_i \psi_u(x_i) + \sum_{i < j} \psi_p(x_i, x_j) \right) \quad (3)$$

where the partition function is defined as  $Z = \sum_{\mathbf{x}} \tilde{P}(\mathbf{x})$ .

Let's define an approximate distribution  $Q(\mathbf{X}) = \prod_i Q_i(X_i)$  as a product of independent marginals  $Q_i(\mathbf{X}_i)$  over each variable in the CRF. For notational clarity we use  $Q_i(X_i)$  to denote the marginal over variable  $X_i$ , rather than the more commonly used  $Q(X_i)$ .

The mean field approximation models a distribution  $Q(\mathbf{X})$  that minimizes the KL-divergence  $\mathbf{D}(Q\|P)$  [1]:

$$\begin{aligned} \mathbf{D}(Q\|P) &= \sum_{\mathbf{x}} Q(\mathbf{x}) \log \left( \frac{P(\mathbf{x})}{Q(\mathbf{x})} \right) \\ &= \sum_{\mathbf{x}} Q(\mathbf{x}) \log P(\mathbf{x}) - \sum_{\mathbf{x}} Q(\mathbf{x}) \log Q(\mathbf{x}) \\ &= \mathbf{E}_{\mathbf{U} \sim Q} [\log P(\mathbf{U})] - \mathbf{E}_{\mathbf{U} \sim Q} [\log Q(\mathbf{U})] \\ &= \mathbf{E}_{\mathbf{U} \sim Q} [\log \tilde{P}(\mathbf{U})] - \mathbf{E}_{\mathbf{U} \sim Q} [\log Z] - \sum_i \mathbf{E}_{U_i \sim Q} [\log Q(U_i)] \\ &= -\mathbf{E}_{\mathbf{U} \sim Q} [E(\mathbf{U})] - \sum_i \mathbf{E}_{U_i \sim Q} [\log Q_i(U_i)] - \log Z \end{aligned} \quad (4)$$

$\mathbf{E}_{\mathbf{U} \sim Q}$  refers to the expected value under the distribution  $Q$ . We use the fact that the Shannon entropy  $\mathbf{E}_{\mathbf{U} \sim Q}[\log Q(\mathbf{U})] = \sum_i \mathbf{E}_{U_i \sim Q_i}[\log Q_i(U_i)]$  decomposes when  $Q(X) = \prod_i Q_i(X_i)$ , due to linearity of expectation.

The marginal  $Q_i(x_i)$  that minimizes the KL-divergence is found by analytically minimizing a Lagrangian that consists of all terms in  $\mathbf{D}(Q\|P)$  plus Lagrange multipliers assuring the marginals  $Q_i(X_i)$  are probability distributions. Detailed derivations and a proof of convergence can be found in Chapter 11.5 of Koller and Friedman [1]. For brevity of exposition we will only present the final update equation:

$$Q_i(x_i) = \frac{1}{Z_i} \exp \left\{ -\psi_u(x_i) - \sum_{j \neq i} \mathbf{E}_{U_j \sim Q_j} [\psi_p(x_i, U_j)] \right\} \quad (5)$$

Substituting the definition of the pairwise potential (Eq. 2) into the mean field update in Equation 5 yields the following formulation of the update equation, which is used in the paper.

$$\begin{aligned} Q_i(x_i = l) &= \frac{1}{Z_i} \exp \left\{ -\psi_u(x_i) - \sum_{j \neq i} \mathbf{E}_{U_j \sim Q_j} \left[ \mu(l, U_j) \sum_{m=1}^K w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) \right] \right\} \\ &= \frac{1}{Z_i} \exp \left\{ -\psi_u(x_i) - \sum_{m=1}^K w^{(m)} \sum_{j \neq i} \mathbf{E}_{U_j \sim Q_j} \left[ \mu(l, U_j) k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) \right] \right\} \\ &= \frac{1}{Z_i} \exp \left\{ -\psi_u(x_i) - \sum_{m=1}^K w^{(m)} \sum_{j \neq i} \sum_{l' \in \mathcal{L}} Q_j(l') \mu(l, l') k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) \right\} \\ &= \frac{1}{Z_i} \exp \left\{ -\psi_u(x_i) - \sum_{l' \in \mathcal{L}} \mu(l, l') \sum_{m=1}^K w^{(m)} \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l') \right\} \end{aligned} \quad (6)$$

We make use of linearity of expectation and rearrange terms such that the message passing is the innermost step, and the compatibility transform is the outermost.

## 2 Mean-field Learning

To efficiently learn the symmetric label compatibility function for our model we use maximum likelihood estimation (MLE). The objective of MLE is to find a set of parameters that maximizes the log-likelihood of the model given training images  $\mathcal{I}$  and their ground truth segmentations  $\mathcal{T}^{(n)}$ :

$$\begin{aligned} \ell(\mu : \mathcal{T}^{(n)}, \mathcal{I}^{(n)}) &= \log P(\mathbf{X} = \mathcal{T}^{(n)} | \mathcal{I}^{(n)}, \mu) \\ &= -E(\mathcal{T}^{(n)} | \mathcal{I}^{(n)}, \mu) - \log Z(\mathcal{I}^{(n)}, \mu) \end{aligned} \quad (7)$$

The global partition function  $Z$  couples all parameters and variables, making it intractable to analytically maximize  $\ell$ . However, it can be shown that the partition function is convex and hence the log-likelihood function is concave [1]. MLE can thus be performed with gradient-based optimization techniques. We now consider the gradient of  $\ell$ :

$$\begin{aligned} \frac{\partial}{\partial \mu_{a,b}} \ell(\mu : \mathcal{T}^{(n)}, \mathcal{I}^{(n)}) &= \frac{\partial}{\partial \mu_{b,a}} \ell(\mu : \mathcal{T}^{(n)}, \mathcal{I}^{(n)}) \\ &= -\frac{\partial}{\partial \mu_{a,b}} E(\mathcal{T}^{(n)} | \mathcal{I}^{(n)}, \mu) - \frac{\partial}{\partial \mu_{a,b}} \log Z(\mathcal{I}^{(n)}, \mu) \\ &= -\sum_m \frac{1}{2} \sum_{i \neq j} k^{(m)}(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) \frac{\partial}{\partial \mu_{a,b}} \mu(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) - \frac{1}{Z} \frac{\partial}{\partial \mu_{a,b}} Z(\mathcal{I}^{(n)}, \mu) \end{aligned} \quad (8)$$

For values  $(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) \neq (a, b)$  or  $(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) \neq (b, a)$  the first term evaluates to 0. We can replace  $\frac{\partial}{\partial \mu_{a,b}} \mu(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) = 1_{a=\mathcal{T}_i^{(n)}} 1_{b=\mathcal{T}_j^{(n)}} + 1_{b=\mathcal{T}_i^{(n)}} 1_{a=\mathcal{T}_j^{(n)}}$ , where  $1_{[\cdot]}$  is the indicator function.

The second expression yields a very similar result

$$\begin{aligned}
\frac{1}{Z} \frac{\partial}{\partial \mu_{a,b}} Z(\mathcal{I}^{(n)}, \mu) &= \frac{1}{Z} \sum_{\mathbf{X}} \frac{\partial}{\partial \mu_{a,b}} \tilde{P}(\mathbf{X}|\mathcal{I}^{(n)}, \mu) \\
&= \frac{1}{Z} \sum_{\mathbf{X}} \frac{\partial}{\partial \mu_{a,b}} \exp(-E(\mathbf{X}|\mathcal{I}^{(n)}, \mu)) \\
&= - \sum_{\mathbf{X}} \frac{1}{Z} \exp(-E(\mathbf{X}|\mathcal{I}^{(n)}, \mu)) \frac{\partial}{\partial \mu_{a,b}} E(\mathbf{X}|\mathcal{I}^{(n)}, \mu) \\
&= - \sum_{\mathbf{X}} P(\mathbf{X}) \sum_m w^{(m)} \frac{1}{2} \sum_{i \neq j} k^{(m)}(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) \left( \mathbf{1}_{a=\mathcal{T}_i^{(n)}} \mathbf{1}_{b=\mathcal{T}_j^{(n)}} + \right. \\
&\quad \left. \mathbf{1}_{b=\mathcal{T}_i^{(n)}} \mathbf{1}_{a=\mathcal{T}_j^{(n)}} \right) \\
&= - \sum_{\mathbf{X}} P(\mathbf{X}) \sum_m w^{(m)} \frac{1}{2} \left( \sum_{i \neq j} k^{(m)}(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) \mathbf{1}_{a=\mathcal{T}_i^{(n)}} \mathbf{1}_{b=\mathcal{T}_j^{(n)}} + \right. \\
&\quad \left. \sum_{j \neq i} k^{(m)}(\mathcal{T}_j^{(n)}, \mathcal{T}_i^{(n)}) \mathbf{1}_{a=\mathcal{T}_j^{(n)}} \mathbf{1}_{b=\mathcal{T}_i^{(n)}} \right) \\
&= - \sum_{\mathbf{X}} P(\mathbf{X}) \sum_m w^{(m)} \sum_{i \neq j} k^{(m)}(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) \mathbf{1}_{a=\mathcal{T}_i^{(n)}} \mathbf{1}_{b=\mathcal{T}_j^{(n)}} \tag{9}
\end{aligned}$$

We make use of the CRF symmetry  $\psi_p(x_i, x_j) = \psi_p(x_j, x_i)$ , such that  $\sum_{i < j} \psi_p(x_i, x_j) = \frac{1}{2} \sum_{i \neq j} \psi_p(x_i, x_j)$ .

This expression is intractable for exact computation. We therefore approximate  $P$  using the mean-field approximation  $Q$  presented in the previous section:

$$\begin{aligned}
&\frac{1}{Z} \frac{\partial}{\partial \mu_{a,b}} Z(\mathcal{I}^{(n)}, \mu) \\
&\approx \sum_{\mathbf{X}} Q(\mathbf{X}) \sum_m w^{(m)} \sum_{i \neq j} k^{(m)}(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) \mathbf{1}_{a=X_i} \mathbf{1}_{b=X_j} \\
&= \sum_m w^{(m)} \sum_{i \neq j} k^{(m)}(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) \sum_{\mathbf{X}} Q(\mathbf{X}/\{X_i, X_j\}) \mathbf{1}_{a=X_i} Q_i(X_i) \mathbf{1}_{b=X_j} Q_j(X_j) \\
&= \sum_m w^{(m)} \sum_{i \neq j} k^{(m)}(\mathcal{T}_i^{(n)}, \mathcal{T}_j^{(n)}) Q_i(a) Q_j(b) \tag{10}
\end{aligned}$$

Due to the definition of  $Q$ , the marginalization  $\sum_{\mathbf{X}/\{X_i, X_j\}} Q(\mathbf{X}/\{X_i, X_j\}) = 1$  and  $\mathbf{1}_{a=X_i} Q_i(X_i)$  equals 0 for  $a \neq X_i$ , thus  $\sum_{X_i} \mathbf{1}_{a=X_i} Q_i(X_i) = Q_i(a)$ .

Rearranging the terms of Equation 10 and substituting them into Equation 8 produces the final gradient

$$\begin{aligned}
\frac{\partial}{\partial \mu(a,b)} \ell_n(\mu : \mathcal{I}^{(n)}, \mathcal{T}^{(n)}) &\approx \sum_m w^{(m)} \left( - \sum_i \mathcal{T}_i^{(n)}(a) \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) \mathcal{T}_j^{(n)}(b) \right. \\
&\quad \left. + \sum_i Q_i(a) \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_i(b) \right) \tag{11}
\end{aligned}$$

where  $\mathcal{T}^{(n)}(a)$  is a binary image in which the  $i$ th variable  $\mathcal{T}_i^{(n)}(a)$  is defined to be  $\mathbf{1}_{\mathcal{T}_i^{(n)}=a}$ .

### 3 Computing the KL-divergence

This section briefly outlines how the KL-divergence  $\mathbf{D}(Q||P)$  can be estimated up to a constant  $\log Z$  using high-dimensional filtering. The KL-divergence can be used to analyze the convergence of the mean field approximation.

In Equation 4,  $\log Z$  is a constant depending only on the image  $\mathcal{I}$  and the CRF parameters. The partition function is independent of the actual assignment  $\mathbf{x}$  and can thus be ignored when evaluate the convergence rate. The Shannon entropy  $\sum_i \mathbf{E}_{U_i \sim Q_i} [\log Q_i(U_i)]$  consist of only local terms, which can be computed efficiently given  $Q$ .

The computationally expensive part is evaluating the expected value of the Gibbs Energy  $E(\mathbf{X})$

$$\begin{aligned} \mathbf{E}_{\mathbf{U} \sim Q}[E(\mathbf{U})] &= \mathbf{E}_{\mathbf{U} \sim Q} \left[ \sum_i \psi_u(U_i) + \sum_{i < j} \psi_p(U_i, U_j) \right] \\ &= \sum_i \mathbf{E}_{U_i \sim Q_i} [\psi_u(U_i)] + \sum_{i < j} \mathbf{E}_{U_i \sim Q_i, U_j \sim Q_j} [\psi_p(U_i, U_j)] \end{aligned} \quad (12)$$

The first expression can be evaluated in linear time by summing up all expected values of the unary potentials  $\psi_u$ . The second expression in its current form requires a summation over all pairs of variables, which is again computationally intractable. We can however formulate the second term as a filtering operation:

$$\begin{aligned} \sum_{i < j} \mathbf{E}_{U_i \sim Q_i, U_j \sim Q_j} [\psi_p(U_i, U_j)] &= \frac{1}{2} \sum_i \mathbf{E}_{U_i \sim Q_i} \left[ \sum_{j \neq i} \mathbf{E}_{U_j \sim Q_j} [\psi_p(U_i, U_j)] \right] \\ &= \frac{1}{2} \sum_{m=1}^K w^{(m)} \sum_i \mathbf{E}_{U_i \sim Q_i} \left[ \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) \mathbf{E}_{U_j \sim Q_j} [\mu(U_i, U_j)] \right] \end{aligned}$$

where  $\mathbf{E}_{U_j \sim Q_j} [\mu(U_i, U_j)]$  is the compatibility transformation and  $\sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) \mathbf{E}_{U_j \sim Q_j} [\mu(U_i, U_j)]$  can be evaluated using high-dimensional filtering.

### References

- [1] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. [1, 2](#)