Proofs for Multi-Modal Mean-Fields via Cardinality-Based Clamping

This document provides technical details and proofs related to Section 5. We first prove the approximation of the KL-Divergence used in Eq. 9. Then, we show that the problem that we are trying to solve in Eq. 9, the minimization of the KL-Divergence, is actually equivalent to the one solved by [5], namely, finding an approximation to the log-partition function. It eventually justifies the benchmark experiments ran in 7.1. Finally, we justify the Gaussian approximation used in the case of large clamping groups in 5.2.1-(2).

1 Minimising the KL-Divergence

Let us see how the KL-Divergence between Q_{MM} and P of Eq. 3 can be minimised with respect to the parameters m_k and to the distributions Q_k, leading to Eq. 9. We reformulate the minimisation problem up to a constant approximation factor of order $\epsilon \log(\epsilon)$.

First, remember that our minimisation problem enforces the near-disjointness condition,

$$\forall k \neq k' \sum_{\mathbf{x} \in \mathcal{X}'_k} Q_k(\mathbf{x}) \leq \epsilon \ , \tag{1}$$

between the elements of the mixture.

Let us then prove the following useful Lemma.

Lemma 1.1 *For all mixture element $k \leq K$,*

$$\sum_{\mathbf{x} \in \mathcal{X}} Q_k(\mathbf{x}) \log \left(\sum_{k' \leq K} m_{k'} Q_{k'}(\mathbf{x}) \right) = \sum_{\mathbf{x} \in \mathcal{X}} Q_k(\mathbf{x}) \log \left(m_k Q_k(\mathbf{x}) \right) + \mathcal{O}(\epsilon \log \epsilon) \ . \tag{2}$$

Proof Let k be the index of a mixture component $k \leq K$, and let us denote the approximation error

$$\delta_k = \sum_{\mathbf{x} \in \mathcal{X}} Q_k(\mathbf{x}) \log \left(\sum_{k' \leq K} m_{k'} Q_{k'}(\mathbf{x}) \right) - \sum_{\mathbf{x} \in \mathcal{X}} Q_k(\mathbf{x}) \log \left(m_k Q_k(\mathbf{x}) \right) \ . \tag{3}$$

Then, we use the near-disjointness condition to bound δ_k,

$$\delta_k \leq \underbrace{\sum_{\mathbf{x} \in \mathcal{X}_k} Q_k(\mathbf{x}) \log \left(1 + \frac{\sum_{k' \neq k} m_{k'} Q_{k'}(\mathbf{x})}{Q_k(\mathbf{x})} \right)}_{I} + \underbrace{\sum_{\mathbf{x} \in \mathcal{X} \setminus \mathcal{X}_k} Q_k(\mathbf{x}) \log \left(1 + \frac{\sum_{k' \neq k} m_{k'} Q_{k'}(\mathbf{x})}{Q_k(\mathbf{x})} \right)}_{J} \tag{4}$$

We first use the well known inequality $\log(1+x) \leq x$ in order to upper bound I,

$$I \leq \sum_{\mathbf{x} \in \mathcal{X}_k} Q_k(\mathbf{x}) \frac{\sum\limits_{k' \neq k} m_{k'} Q_{k'}(\mathbf{x})}{Q_k(\mathbf{x})} \tag{5}$$

$$\leq \sum_{k' \neq k} \sum_{\mathbf{x} \in \mathcal{X}_k} m_{k'} Q_{k'}(\mathbf{x}) \tag{6}$$

$$\leq \sum_{k' \neq k} \epsilon \tag{7}$$

$$\leq \mathcal{O}(\epsilon) \ . \tag{8}$$

The second term, J, can then be upper-bounded using the fact that the $m_{k'}$ and $Q_{k'}$ are mixture weights and probabilities and hence $\sum\limits_{k' \neq k} m_{k'} Q_{k'}(\mathbf{x}) \leq 1$ for all $\mathbf{x}$. Therefore,

$$J \leq \sum_{\mathbf{x} \in \mathcal{X} \setminus \mathcal{X}_k} Q_k(\mathbf{x}) \log \left(1 + \frac{1}{Q_k(\mathbf{x})} \right) \tag{9}$$

$$\leq \sum_{\mathbf{x} \in \mathcal{X} \setminus \mathcal{X}_k} -Q_k(\mathbf{x}) \log \left(Q_k(\mathbf{x}) \right) \tag{10}$$

$$\leq \sum_{k' \neq k} \sum_{\mathbf{x} \in \mathcal{X}_{k'}} -Q_k(\mathbf{x}) \log \left(Q_k(\mathbf{x}) \right) \ . \tag{11}$$

Furthermore, for all $k' \neq k$, the near-disjointness condition enforces that $\sum\limits_{\mathbf{x} \in \mathcal{X}_{k'}} Q_k(\mathbf{x}) \leq \epsilon$. Under this constraint, on each of the subsets $\mathcal{X}_{k'}$, the maximal entropy is reached if $Q_k(\mathbf{x}) = \frac{\epsilon}{\mid \mathcal{X}_{k'} \mid}$ for all $\mathbf{x}$ in $\mathcal{X}_{k'}$. And, therefore

$$\sum_{\mathbf{x} \in \mathcal{X}_{k'}} -Q_k(\mathbf{x}) \log \left(Q_k(\mathbf{x}) \right) \leq \epsilon \log \left(\frac{\mid \mathcal{X}'_k \mid}{\epsilon} \right) \tag{12}$$

$$\leq \mathcal{O}(\epsilon \log \epsilon) + \mathcal{O}(\epsilon) \ , \tag{13}$$

where the factor $\log(\mid \mathcal{X}_k \mid)$, which is of the order of the number of variables, has been integrated in the constant.

Hence,

$$J \leq \sum_{k' \neq k} \sum_{\mathbf{x} \in \mathcal{X}_{k'}} -Q_k(\mathbf{x}) \log \left(Q_k(\mathbf{x}) \right) \tag{14}$$

$$\leq \mathcal{O}(\epsilon \log \epsilon) + \mathcal{O}(\epsilon) \ , \tag{15}$$

$$\tag{16}$$

which terminates the proof.

We can then move on to the minimisation of the KL-Divergence

$$\min_{\hat{m},\hat{q}} \mathrm{KL}(Q_{MM}\|P) = \min_{\hat{m},\hat{q}} \sum_{\mathbf{x}\in\mathcal{X}} \sum_{k\leq K} Q_{MM}(\mathbf{x}) \log\left(\frac{Q_{MM}(\mathbf{x})}{P(\mathbf{x})}\right) \tag{17}$$

$$= \min_{\hat{m},\hat{q}} \sum_{\mathbf{x}\in\mathcal{X}} \sum_{k\leq K} Q_{MM}(\mathbf{x}) \log\left(\frac{Q_{MM}(\mathbf{x})}{e^{-E(\mathbf{x})}}\right) + \log(Z) \tag{18}$$

$$= \min_{\hat{m},\hat{q}} \sum_{k\leq K} \sum_{\mathbf{x}\in\mathcal{X}} m_k Q_k(\mathbf{x}) \log\left(\frac{\sum\limits_{k'\leq K} m_{k'} Q_{k'}(\mathbf{x})}{e^{-E(\mathbf{x})}}\right) + \log(Z) \tag{19}$$

$$= \min_{\hat{m},\hat{q}} \sum_{k\leq K} \sum_{\mathbf{x}\in\mathcal{X}} m_k Q_k(\mathbf{x}) \log\left(\frac{m_k Q_k(\mathbf{x})}{e^{-E(\mathbf{x})}}\right) + \log(Z) + \mathcal{O}(\epsilon\log\epsilon) \tag{20}$$

$$= \min_{\hat{m}} \left[\sum_{k\leq K} m_k \log m_k + \sum_{k\leq K} \min_{q_k} \sum_{\mathbf{x}\in\mathcal{X}} m_k Q_k(\mathbf{x}) \log\left(\frac{Q_k(\mathbf{x})}{e^{-E(\mathbf{x})}}\right)\right] + \log(Z) + \mathcal{O}(\epsilon\log\epsilon) \tag{21}$$

$$= \min_{\hat{m}} \sum_{k\leq K} m_k \log(m_k) - \sum_{k\leq K} m_k A_k + \log(Z) + \mathcal{O}(\epsilon\log\epsilon)\ , \tag{22}$$

where,

$$A_k = \max_{q_i^k, i=1\ldots N} \sum_{\mathbf{x}\in\mathcal{X}} Q_k(\mathbf{x}) \log\left(\frac{e^{-E(\mathbf{x})}}{Q_k(\mathbf{x})}\right)\ .$$

Equation 20 is obtained using Lemma 1.1.

Assuming that we are able to compute A_k, for all k, the minimisation of this KL-Divergence with respect to parameters m_k, under the nomalisation constraint

$$\sum_{k\leq K} m_k = 1\ , \tag{23}$$

is then straightforward and leads to

$$m_k = \frac{e^{A_k}}{\sum\limits_{k'\leq K} e^{A_{k'}}}\ . \tag{24}$$

2 Equivalence with the approximation of the partition function.

The recent work of [5], that we use as a baseline, looks for the best heuristic to choose the clamping variables. They measure the quality of the approximation through the closeness of the estimated partition function, which they compute as the sum of MF approximated partition functions for each component of the mixture, to the true one. We will now see that this problem is strictly equivalent to the minimisation of the KL-Divergence of Eq. 9.

Indeed, replacing 23 in 22, we directly obtain that

$$\mathrm{KL}(Q_{MM}\|P) = \log(\sum_{k'\leq K} e^{A_{k'}}) + \log(Z) + \mathcal{O}(\epsilon\log\epsilon) \tag{25}$$

$$= \log(Z) - \log(\tilde{Z}) + \mathcal{O}(\epsilon\log\epsilon)\ , \tag{26}$$

where,

$$\tilde{Z} = \sum_{k' \leq K} e^{A_{k'}} \ , \tag{27}$$

is precisely the approximation of the partition function Z proposed by [5]. In other terms, it is the sum of local variational lower-bounds on clamped subsets of the state space.

3 Gaussian approximation to the cardinality constraint.

In the following, we explain the Gaussian approximation of the cardinality constraint used in 5.2.1-(2) and in our application to Semantic Segmentation. Let us consider the case where we generate only two modes modelled by $Q_1(\mathbf{x}) = \prod q_i^1(x_i)$ and $Q_2(\mathbf{x}) = \prod q_i^2(x_i)$ and we seek to estimate the q_i^1 probabilities. The q_i^2 probabilities are evaluated similarly.

Recall that, each A_k is obtained through the constrained MF optimisation problem

$$\begin{aligned} \max_{q_i^k, i=1\ldots N} \quad & \sum_{\mathbf{x}\in\mathcal{X}} Q_k(\mathbf{x}) \log\left(\frac{e^{-E(\mathbf{x})}}{Q_k(\mathbf{x})}\right) \\ \text{s.t.} \quad & Q_1\left(\sum_{u=1\ldots L} \mathbb{1}(\mathbf{X}_{i_u} = v_u) < C\right) \leq \epsilon \ . \end{aligned} \tag{28}$$

Under the probability Q_1, $\sum_{u=1\ldots L} \mathbb{1}(\mathbf{X}_{i_u} = v_u)$ is a sum of independent binary random variables that are non identically distributed, in other words, a Poisson Binomial Distribution. In the general case, there is no closed-form formula for computing the Cumulative Distribution Function of such a distribution from the individual marginals parametrising Q_1. However, when L is large (≥ 10), the Gaussian approximation is good enough.

Therefore, we use a Gaussian approximation to replace the cardinality constraint by

$$\sum_{u\in\{1\ldots L\}} q_{i_u}^1(v_u) < C + \sigma F^{-1}(1-\epsilon) \ , \tag{29}$$

where F is the Gaussian cumulative distribution fonction and σ^2 the variance, which, in theory should be

$$\sigma^2 = \sum_{u\in\{1\ldots L\}} q_{i_u}^1(v_u)(1 - q_{i_u}^1(v_u)) \ , \tag{30}$$

but which can be either upper-bounded by $\frac{L}{4}$ or re-estimated at the beginning of each Lagrangian iteration.

In short, we replace the untractable higher order constraint 28, by a simple one involving only the sum of the MF parameters $q_{i_u}^1(v_u)$.

References

[1] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua. Multiple Object Tracking Using K-Shortest Paths Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):1806–1819, 2011.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *International Conference for Learning Representations*, 2015.

[3] P. Krähenbühl and V. Koltun. Parameter Learning and Convergent Inference for Dense Random Fields. In *International Conference on Machine Learning*, pages 513–521, 2013.

[4] J. W. Suurballe. Disjoint Paths in a Network. *Networks*, 4:125–145, 1974.

[5] A. Weller and J. Domke. Clamping improves trw and mean field approximations. In *Advances in Neural Information Processing Systems*, 2015. 1, 3, 4

[6] S. Zheng, S. Jayasumana, B. Romera-paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional Random Fields as Recurrent Neural Networks. In *International Conference on Computer Vision*, 2015.