Learning to Cluster Using High Order Graphical Models with Latent Variables
(Supplemental Material)

Nikos Komodakis
University of Crete, Computer Science Department
http://www.csd.uoc.gr/~komod

Abstract

This document provides technical proofs for all theorems in the main paper.

1. Proofs

Lemma 1 *Let $\hat{\mathbf{x}}^{k,p}$, $\hat{\mathbf{x}}^{k,C}$ be binary minimizers of the energy functions $\bar{E}^k_p$, $\bar{E}^k_C$. Define $f^k_{pq} \equiv f_{pq}(\mathbf{y}^k)$, $\hat{X}^k_q \equiv \hat{x}^{k,C}_{qq} + \sum_p \hat{x}^{k,p}_{qq}, \forall q \in C$. Update (30) then reduces to*

$$\begin{bmatrix} \mathbf{w} \\ \lambda^k_{pq} \\ \lambda^k_{Cq} \end{bmatrix} -= s_t \begin{bmatrix} \tau \nabla J(w) + \sum_k \delta^k_{\mathbf{w}} \\ \frac{\hat{X}^k_q}{|S^k|+1} - \hat{x}^{k,p}_{qq} \\ \frac{\hat{X}^k_q}{|S^k|+1} - \hat{x}^{k,C}_{qq} \end{bmatrix}, \tag{36}$$

where $\delta^k_{\mathbf{w}} = \sum_{p,q} x^k_{pq} f^k_{pq} - \sum_{p \neq q} \hat{x}^{k,p}_{pq} f^k_{pq} - \frac{\sum_q \hat{X}^k_q f^k_{qq}}{|S^k|+1}$.

Note: If $J(\mathbf{w})$ is non-differentiable (*e.g.*, if $J(\mathbf{w}) = ||\mathbf{w}||_1$) then $\nabla J(\mathbf{w})$ should refer to a subgradient of $J(\cdot)$ at $\mathbf{w}$.

Proof. Update (30) requires computing a subgradient of the objective function (28) with respect to $\mathbf{w}$, $\boldsymbol{\lambda}^k$ (for a fixed $\mathbf{x}^k$). To this end, we need to compute the corresponding subgradient for each of the terms $\bar{\mathcal{L}}_{\bar{E}^k_p}(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k)$ and $\bar{\mathcal{L}}_{\bar{E}^k_C}(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k)$ that are included in function (28). By definition (21) it holds that[1]

$$\bar{\mathcal{L}}_{\bar{E}^k_p}(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) = \bar{E}^k_p(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) - \min_{\mathbf{x}} \bar{E}^k_p(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) \tag{37}$$

$$= \bar{E}^k_p(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) + \max_{\mathbf{x}} \left(-\bar{E}^k_p(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) \right) \tag{38}$$

A subgradient for a pointwise maximum function $g(\mathbf{w}, \boldsymbol{\lambda}^k) = \max_{\mathbf{x}} g_{\mathbf{x}}(\mathbf{w}, \boldsymbol{\lambda}^k)$, where each $g_{\mathbf{x}}(\cdot, \cdot)$ is convex and differentiable, is given by $\nabla g_{\hat{\mathbf{x}}}(\mathbf{w}, \boldsymbol{\lambda}^k)$ for any $\hat{\mathbf{x}}$ that satisfies $g(\mathbf{w}, \boldsymbol{\lambda}^k) = g_{\hat{\mathbf{x}}}(\mathbf{w}, \boldsymbol{\lambda}^k)$, *i.e.*, $\max_{\mathbf{x}} g_{\mathbf{x}}(\mathbf{w}, \boldsymbol{\lambda}^k) = g_{\hat{\mathbf{x}}}(\mathbf{w}, \boldsymbol{\lambda}^k)$. Since function $-\bar{E}^k_p(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ is linear (and hence both convex and differentiable) with respect to $\mathbf{w}$, $\boldsymbol{\lambda}^k$, a subgradient of function $\bar{\mathcal{L}}_{\bar{E}^k_p}(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k)$ (with respect to $\mathbf{w}$, $\boldsymbol{\lambda}^k$) will thus equal

$$\nabla \bar{E}^k_p(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) - \nabla \bar{E}^k_p(\hat{\mathbf{x}}^{k,p}; \mathbf{w}, \boldsymbol{\lambda}^k) \ , \tag{39}$$

where $\hat{\mathbf{x}}^{k,p}$ denotes a binary minimizer of function $\bar{E}^k_p(\cdot; \mathbf{w}, \boldsymbol{\lambda}^k)$. Therefore, based on (39) and the fact that $d^k_{pq} = \mathbf{w}^T f^k_{pq}$, a

[1] Note that both here and in the main paper all vectors of CRF variables $\mathbf{x}$ are *always* assumed to be *integral*. Therefore, in order to reduce notational clutter we often omit stating this integrality constraint when using such vectors (*e.g.*, we simply write $\min_{\mathbf{x}}$ instead of $\min_{\{\mathbf{x}:\mathbf{x} \text{ has integral components}\}}$).

subgradient of $\bar{\mathcal{L}}_{\bar{E}^k_p}$ will have components $\delta\mathbf{w}^{k,p}$, $\{\delta\lambda^{k,p}_q\}_q$ (corresponding to variables $\mathbf{w}$, $\{\lambda^k_{pq}\}_q$ respectively) given by

$$\delta\mathbf{w}^{k,p} = \sum_{q:q\neq p} x^k_{pq} f^k_{pq} + \sum_q \frac{x^k_{qq} f^k_{qq}}{|S^k|+1} - \left(\sum_{q:q\neq p} \hat{x}^{k,p}_{pq} f^k_{pq} + \sum_q \frac{\hat{x}^{k,p}_{qq} f^k_{qq}}{|S^k|+1} \right) \tag{40}$$

$$\delta\lambda^{k,p}_q = x^k_{qq} - \hat{x}^{k,p}_{qq} \quad . \tag{41}$$

Similarly, we can prove that a subgradient of function $\bar{\mathcal{L}}_{\bar{E}^k_C}(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k)$ will have components $\delta\mathbf{w}^{k,C}$, $\{\delta\lambda^{k,C}_q\}_{q\in C}$ (corresponding to variables $\mathbf{w}$, $\{\lambda^k_{Cq}\}_{q\in C}$ respectively) given by

$$\delta\mathbf{w}^{k,C} = \sum_{q\in C} \frac{x^k_{qq} f^k_{qq}}{|S^k|+1} - \sum_{q\in C} \frac{\hat{x}^{k,C}_{qq} f^k_{qq}}{|S^k|+1} \tag{42}$$

$$\delta\lambda^{k,C}_q = x^k_{qq} - \hat{x}^{k,C}_{qq} \quad , \ \forall q \in C \tag{43}$$

where $\hat{\mathbf{x}}^{k,C}$ denotes a binary minimizer of function $\bar{E}^k_C(\cdot; \mathbf{w}, \boldsymbol{\lambda}^k)$.

Therefore, a total subgradient of the objective function (28) will have components $\delta\mathbf{w}$, $\delta\lambda^{k,p}_q$, $\delta\lambda^{k,C}_q$ (corresponding to variables $\mathbf{w}$, λ^k_{pq}, λ^k_{Cq} respectively), where

$$\delta\mathbf{w} = \tau\nabla J(\mathbf{w}) + \sum_k \left(\sum_{p\in S^k} \delta\mathbf{w}^{k,p} + \sum_{C\in\mathcal{C}^k} \delta\mathbf{w}^{k,C} \right) \overset{(40),(42)}{=} \tau\nabla J(\mathbf{w}) + \sum_k \delta^k_{\mathbf{w}} \quad . \tag{44}$$

Furthermore, projection onto the set $\boldsymbol{\Lambda}^k = \big\{\boldsymbol{\lambda}^k : \sum_{p\in S^k} \lambda^k_{pq} + \lambda^k_{Cq} = 0\,,\ \forall C \in \mathcal{C}^k, q \in C\big\}$ simply requires to first subtract the average $\frac{\sum_{p\in S^k} \delta\lambda^{k,p}_q + \delta\lambda^{k,C}_q}{|S^k|+1} \overset{(41),(43)}{=} x^k_{qq} - \frac{X^k_q}{|S^k|+1}$ from each of the elements $\{\delta\lambda^{k,p}_q\}_p$, $\delta\lambda^{k,C}_q$ before applying the updates $\mathbf{w} -= s_t\delta\mathbf{w}$, $\lambda^k_{pq} -= s_t\delta\lambda^{k,p}_q$, $\lambda^k_{Cq} -= s_t\delta\lambda^{k,C}_q$ (where s_t is the multiplier used during the t-th iteration). This is easily seen to lead to updates (36), which concludes the proof of the lemma. □

Lemma 2 *Let* $[a]_+ \equiv \max(a, 0)$, $[a]_- \equiv \min(a, 0)$.

1. *For fixed* p, *let* $\theta^k_q \equiv \frac{\bar{u}^k_{qq}(1)}{|S^k|+1} + \lambda^k_{pq}, \forall q$ *and let us define* $\bar{\theta}^k_q \equiv \bar{u}^k_{pq}(1) + [\theta^k_q]_+, \forall q \neq p$ *and* $\bar{\theta}^k_p = \theta^k_p$. *A minimizer* $\hat{\mathbf{x}}$ of $\bar{E}^k_p(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ *can be computed as follows:*

$$\forall q \neq p,\ \hat{x}_{qq} \leftarrow [\theta^k_q < 0] \tag{45}$$

$$\forall q,\ \hat{x}_{pq} \leftarrow [q = \bar{q}],\ \textit{where}\ \bar{q} = \arg\min_q \bar{\theta}^k_q \tag{46}$$

2. *For fixed* $C \in \mathcal{C}^k$, *let* $\theta^k_q \equiv \frac{\bar{u}^k_{qq}(1)}{|S^k|+1} + \lambda^k_{Cq}, \forall q \in C$. *A minimizer* $\hat{\mathbf{x}}$ of $\bar{E}^k_C(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ *is given by*

$$\forall q \in C,\ \hat{x}_{qq} = \begin{cases} [\theta^k_q < \alpha], & \textit{if}\ 2\alpha + \sum_{q'\in C} [\theta^k_{q'} - \alpha]_- < 0 \\ 0, & \textit{otherwise} \end{cases} \tag{47}$$

Proof. 1. It holds that

$$\bar{E}^k_p(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) = \sum_{q:q\neq p} \bar{u}^k_{pq}(x_{pq}) + \sum_q \Big(\frac{\bar{u}^k_{qq}(x_{qq})}{|S^k|+1} + \lambda^k_{pq} x_{qq} \Big) + \sum_q \bar{\phi}_{pq}(x_{pq}, x_{qq}) + \bar{\phi}_p(\mathbf{x}_p) - \beta \tag{48}$$

$$= \sum_{q:q\neq p} \bar{u}^k_{pq}(1) x_{pq} + \sum_q \theta^k_q x_{qq} + \sum_q \bar{\phi}_{pq}(x_{pq}, x_{qq}) + \bar{\phi}_p(\mathbf{x}_p) - \beta \tag{49}$$

$$= \sum_{q:q\neq p} \bar{u}^k_{pq}(1) x_{pq} + \sum_q \big(\theta^k_q x_{qq} + \bar{\phi}_{pq}(x_{pq}, x_{qq}) \big) + \bar{\phi}_p(\mathbf{x}_p) - \beta \quad , \tag{50}$$

where $\bar{\phi}_{pq}(x_{pq}, x_{qq}) = \delta(x_{pq} \le x_{qq})$, $\bar{\phi}_p(\mathbf{x}_p) = \delta\left(\sum_q x_{pq} = 1\right)$ and $\delta(\cdot)$ equals 0 if the expression in parenthesis is satisfied and ∞ otherwise.

Due to the term $\theta_q^k x_{qq}$, it is easy to see that if we set $x_{qq} = 1$ for any $q \neq p$ then the value of the function $\bar{E}_p^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ will decrease if and only if it holds $\theta_q^k < 0$. Therefore, to minimize $\bar{E}_p^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ we must set

$$\hat{x}_{qq} = [\theta_q^k < 0], \; \forall q \neq p\,. \tag{51}$$

Furthermore, the fact that the components of an optimal solution $\hat{\mathbf{x}}$ must belong to $\{0, 1\}$ in conjunction with the form of the potential $\bar{\phi}_p(\mathbf{x}_p) = \delta\left(\sum_q x_{pq} = 1\right)$ impose the constraint that we must set equal to 1 exactly one of the variables in the set $\{\hat{x}_{pq}\}_q$. If we set variable $\hat{x}_{pq}$ (with $q \neq p$) equal to 1 then the cost we must pay is $\bar{u}_{pq}^k(1)$, due to the term $\bar{u}_{pq}^k(1)\hat{x}_{pq}$, plus $[\theta_q^k]_+$, due to the term $\theta_q^k \hat{x}_{qq} + \bar{\phi}_{pq}(\hat{x}_{pq}, \hat{x}_{qq})$ that requires also setting $\hat{x}_{qq} = 1$ (note that we are paying $[\theta_q^k]_+$ and not θ_q^k because for $q \neq p$ if $\theta_q^k < 0$ then $\hat{x}_{qq}$ is set to 1 anyway due to (51) and thus no extra cost is paid in this case). On the other hand, if we set $\hat{x}_{pp} = 1$ then the cost we must pay is θ_p^k due to the term $\theta_p^k \hat{x}_{pp}$. Therefore, for any q, the cost we pay if we choose to set $\hat{x}_{pq} = 1$ is given by $\bar{\theta}_q^k$. As a result, we should set $\hat{x}_{pq} = [q = \bar{q}]$, where $\bar{q} = \arg\min_q \bar{\theta}_q^k$.

2. Energy $\bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ can be expressed as

$$\bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) = \sum_{q \in C} \left(\frac{\bar{u}_{qq}^k(x_{qq})}{|S^k| + 1} + \lambda_{Cq}^k x_{qq} \right) + \bar{\phi}_C(\mathbf{x}_C) \tag{52}$$

$$= \sum_{q \in C} \theta_q^k x_{qq} + \bar{\phi}_C(\mathbf{x}_C) \tag{53}$$

$$= \sum_{q \in C} \theta_q^k x_{qq} - \alpha \Big| 1 - \sum_{q \in C} x_{qq} \Big|\,. \tag{54}$$

We will consider two cases:

(a) The minimizer of function $\bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ is given by $\hat{\mathbf{x}} = \mathbf{0}$ (*i.e.*, none of the binary variables $\{\hat{x}_{qq}\}_{q \in C}$ is equal to 1). In this case the minimum of function $\bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ must equal

$$\text{OPT}_1 = -\alpha\,. \tag{55}$$

(b) The minimizer of function $\bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ is given by $\hat{\mathbf{x}} \neq \mathbf{0}$. In this case at least one of the binary variables $\{\hat{x}_{qq}\}_{q \in C}$ will equal 1 and so $\bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ can be written as

$$\bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) = \sum_{q \in C} \theta_q^k x_{qq} - \alpha \Big| 1 - \sum_{q \in C} x_{qq} \Big| \tag{56}$$

$$= \sum_{q \in C} \theta_q^k x_{qq} - \alpha \left(\sum_{q \in C} x_{qq} - 1 \right) \tag{57}$$

$$= \sum_{q \in C} \left(\theta_q^k - \alpha \right) x_{qq} + \alpha\,. \tag{58}$$

Therefore, the minimizer $\hat{\mathbf{x}}$ will be given by

$$\hat{x}_{qq} = [\theta_q^k < \alpha] \tag{59}$$

and so the optimum value of $\bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k)$ will equal

$$\text{OPT}_2 = \sum_{q \in C} [\theta_q^k - \alpha]_- + \alpha\,. \tag{60}$$

To conclude the proof, it suffices to notice that the second case will hold true if and only if

$$\mathrm{OPT}_2 < \mathrm{OPT}_1 \Leftrightarrow \sum_{q \in C} [\theta_q^k - \alpha]_- + \alpha < -\alpha \Leftrightarrow \sum_{q \in C} [\theta_q^k - \alpha]_- + 2\alpha < 0 \,. \tag{61}$$

□

Lemma 3: Minimizations (27) and (28) in the main paper are equivalent.

Proof. It holds that

$$\min_{\{\mathbf{x}^k \in \mathcal{X}(\mathcal{C}^k)\}, \mathbf{w}} \tau J(\mathbf{w}) \; + \sum_k \left(\bar{E}^k(\mathbf{x}^k; \mathbf{w}) - \mathcal{R}^k(\mathbf{w}) \right) \tag{62}$$

$$\stackrel{(26)}{=} \min_{\{\mathbf{x}^k \in \mathcal{X}(\mathcal{C}^k)\}, \mathbf{w}} \tau J(\mathbf{w}) \; + \sum_k \left(\bar{E}^k(\mathbf{x}^k; \mathbf{w}) - \max_{\boldsymbol{\lambda}^k \in \boldsymbol{\Lambda}^k} \left(\sum_p \min_{\mathbf{x}} \bar{E}_p^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) + \sum_C \min_{\mathbf{x}} \bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) \right) \right) \tag{63}$$

$$= \min_{\{\mathbf{x}^k \in \mathcal{X}(\mathcal{C}^k)\}, \mathbf{w}, \{\boldsymbol{\lambda}^k \in \boldsymbol{\Lambda}^k\}} \tau J(\mathbf{w}) + \sum_k \left(\bar{E}^k(\mathbf{x}^k; \mathbf{w}) - \sum_p \min_{\mathbf{x}} \bar{E}_p^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) - \sum_C \min_{\mathbf{x}} \bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) \right) \tag{64}$$

$$\stackrel{(25)}{=} \min_{\{\mathbf{x}^k \in \mathcal{X}(\mathcal{C}^k)\}, \mathbf{w}, \{\boldsymbol{\lambda}^k \in \boldsymbol{\Lambda}^k\}} \tau J(\mathbf{w}) + \sum_k \Bigg(\sum_p \bar{E}_p^k(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) + \sum_C \bar{E}_C^k(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) - \sum_p \min_{\mathbf{x}} \bar{E}_p^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) - \sum_C \min_{\mathbf{x}} \bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) \Bigg) \tag{65}$$

$$= \min_{\{\mathbf{x}^k \in \mathcal{X}(\mathcal{C}^k)\}, \mathbf{w}, \{\boldsymbol{\lambda}^k \in \boldsymbol{\Lambda}^k\}} \tau J(\mathbf{w}) + \sum_k \sum_p \left(\bar{E}_p^k(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) - \min_{\mathbf{x}} \bar{E}_p^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) \right) + \sum_k \sum_C \left(\bar{E}_C^k(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) - \min_{\mathbf{x}} \bar{E}_C^k(\mathbf{x}; \mathbf{w}, \boldsymbol{\lambda}^k) \right) \tag{66}$$

$$= \min_{\{\mathbf{x}^k \in \mathcal{X}(\mathcal{C}^k)\}, \mathbf{w}, \{\boldsymbol{\lambda}^k \in \boldsymbol{\Lambda}^k\}} \tau J(\mathbf{w}) + \sum_k \sum_p \bar{\mathcal{L}}_{\bar{E}_p^k}(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) + \sum_k \sum_C \bar{\mathcal{L}}_{\bar{E}_C^k}(\mathbf{x}^k; \mathbf{w}, \boldsymbol{\lambda}^k) \;\; , \tag{67}$$

which concludes the proof. □