Distributed Message Passing for Large Scale Graphical Models
Supplementary Material

Alexander Schwing
ETH Zurich

Tamir Hazan
TTI Chicago

Marc Pollefeys
ETH Zurich

Raquel Urtasun
TTI Chicago

Abstract

In this paper we propose a distributed message-passing algorithm for inference in large scale graphical models. Our method can handle large problems efficiently by distributing and parallelizing the computation and the memory requirements. The convergence and optimality guarantees of recently developed message-passing algorithms are preserved by introducing new types of consistency messages, sent between the distributed computers. We demonstrate the effectiveness of our approach in the task of stereo reconstruction from high-resolution imagery, and show that inference is possible with more than 200 *labels in images larger than* 10 *MPixel. In this supplementary material we provide proofs for the claims in the paper.*

1. Proof of Claim 1

The problem of finding the maximum a-posteriori assignment (MAP) via an LP relaxation with entropy barrier functions and an encoded distributed architecture can be expressed as follows:

$$\max \sum_{s\in G_{\mathcal{P}}} \left(\sum_{\alpha\in G_s,\mathbf{x}_\alpha} b_\alpha^s(\mathbf{x}_\alpha)\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum_{i\in G_s,x_i} b_i^s(x_i)\theta_i(x_i)\right) + \epsilon \sum_{s\in G_{\mathcal{P}}} \left(\sum_{\alpha\in G_s} \hat{c}_\alpha H(\mathbf{b}_\alpha^s) + \sum_{i\in G_s} c_i H(\mathbf{b}_i^s)\right) \tag{1}$$

subject to:

$$\forall s, i, x_i, \alpha\in N(i), \quad \sum_{\mathbf{x}_\alpha\setminus x_i} b_\alpha^s(\mathbf{x}_\alpha) = b_i^s(x_i) \tag{2}$$

$$\forall s, \alpha\in N_{\mathcal{P}}(s), \mathbf{x}_\alpha, \quad b_\alpha^s(\mathbf{x}_\alpha) = b_\alpha(\mathbf{x}_\alpha) \tag{3}$$

Claim 1 *Set $\nu_{s\to\alpha} = 0$ for every $\alpha \notin G_{\mathcal{P}}$. Then the following program is the dual program for the distributed convex belief propagation in (1):*

$$\sum_{s,\alpha\in G_s} \epsilon\hat{c}_\alpha \ln \sum_{\mathbf{x}_\alpha} \exp\left(\frac{\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum\limits_{i\in N(\alpha)\cap s} \lambda_{i\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha)}{\epsilon\hat{c}_\alpha}\right) + \sum_i \epsilon c_i \ln \sum_{x_i} \exp\left(\frac{\theta_i(x_i) - \sum_{\alpha\in N(i)} \lambda_{i\to\alpha}(x_i)}{\epsilon c_i}\right)$$

subject to the constraints $\sum_{s\in N_{\mathcal{P}}(\alpha)} \nu_{s\to\alpha}(\mathbf{x}_\alpha) = 0$.

Proof: We first note that the constraints given in Eq. (2) have to hold $\forall s, \alpha \in G_s, i \in N(\alpha)\cap s, x_i$. Similarly, the constraints in (3) hold $\forall \alpha, s\in N_{\mathcal{P}}(\alpha), \mathbf{x}_\alpha$.

Using those equivalences and incorporating Lagrange multipliers $\lambda_{i\to\alpha}(x_i)$ and $\nu_{s\to\alpha}(\mathbf{x}_\alpha)$ for the constraints (2) and (3) respectively, the Lagrangian L for the program in (1) is given by

$$\begin{aligned} L \;=\; & \sum_{s,i\in G_s}\left[\epsilon c_i H(\mathbf{b}_i^s) + \sum_{x_i} b_i^s(x_i)\left(\theta_i(x_i) - \sum_{\alpha\in N(i)}\lambda_{i\to\alpha}(x_i)\right)\right] + \\ & + \sum_{s,\alpha\in G_s}\left[\epsilon\hat{c}_\alpha H(\mathbf{b}_\alpha^s) + \sum_{\mathbf{x}_\alpha} b_\alpha^s(\mathbf{x}_\alpha)\left(\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum_{i\in N(\alpha)\cap s}\lambda_{i\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha)\right)\right] \\ & - \sum_{\alpha,s\in N_{\mathcal{P}}(\alpha),\mathbf{x}_\alpha} b_\alpha(\mathbf{x}_\alpha)\nu_{s\to\alpha}(\mathbf{x}_\alpha). \end{aligned} \tag{4}$$

Note, that we added terms $\nu_{s\to\alpha}(\mathbf{x}_\alpha)\ \forall s,\alpha\in G_s$ for notational convenience. We therefore require $\nu_{s\to\alpha} = 0\ \forall\alpha\notin G_{\mathcal{P}}$.

The dual is given by

$$g = \sup_{b_i^s(x_i),b_\alpha^s(\mathbf{x}_\alpha),b_\alpha(\mathbf{x}_\alpha)} L\,.$$

As the Lagrangian (4) nicely decouples, we can compute the dual in a term by term manner. To this end we make use of the Fenchel theorem ($f^*(p) = \sup_x(px - f(x))$) and its scaling property ($f(x) = g(a\cdot x) \Rightarrow f^*(p) = g^*(p/a)$). We further note that the Fenchel-Conjugate-Dual of the negative entropy subject to simplex constraints ($f(\mathbf{x}) = \sum_i x_i\ln(x_i), s.t. x_i \geq 0, \sum_i x_i = 1$) is given by the log-sum-exp function ($f^*(\mathbf{x}) = \ln\sum_i\exp(x_i)$) (cf. [2]).

To compute the dual w.r.t. $b_\alpha(\mathbf{x}_\alpha)$ we solve

$$\forall\alpha,\mathbf{x}_\alpha,\quad \sup\nolimits_{b_\alpha(\mathbf{x}_\alpha)} -b_\alpha(\mathbf{x}_\alpha)\textstyle\sum_{s\in N_{\mathcal{P}}(\alpha)}\nu_{s\to\alpha}(\mathbf{x}_\alpha),$$

which is identical to zero if $\sum_{s\in N_{\mathcal{P}}(\alpha)}\nu_{s\to\alpha}(\mathbf{x}_\alpha) = 0$ and infinity otherwise.

We can therefore write the dual function g as

$$\begin{aligned} g \;=\; & \sum_i \epsilon c_i \ln\sum_{x_i}\exp\left(\frac{\theta_i(x_i) - \sum_{\alpha\in N(i)}\lambda_{i\to\alpha}(x_i)}{\epsilon c_i}\right) \\ & + \sum_{s,\alpha\in G_s}\epsilon\hat{c}_\alpha\ln\sum_{\mathbf{x}_\alpha}\exp\left(\frac{\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum_{i\in N(\alpha)\cap s}\lambda_{i\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha)}{\epsilon\hat{c}_\alpha}\right) \end{aligned} \tag{5}$$

subject to:

$$\forall\alpha,\mathbf{x}_\alpha,\quad \sum_{s\in N_{\mathcal{P}}(\alpha)}\nu_{s\to\alpha}(\mathbf{x}_\alpha) = 0, \tag{6}$$

$$\forall\alpha\notin G_{\mathcal{P}}, \nu_{s\to\alpha} = 0, \tag{7}$$

which proves the claim. □

The dual program is then obtained by $\min_{\lambda,\nu} g$ subject to above constraints. For minimization we perform block coordinate descent on g using the steps defined in the following claim.

2. Proof of Claim 2

Claim 2 *For every $s\in G_{\mathcal{P}}$, set $\mu_{\alpha\to i}(x_i)$ to be*

$$\epsilon\hat{c}_\alpha\ln\sum_{\mathbf{x}_\alpha\backslash x_i}\exp\left(\frac{\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum\limits_{j\in N(\alpha)\cap s\backslash i}\lambda_{j\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha)}{\epsilon\hat{c}_\alpha}\right)$$

then the block coordinate descent on $\lambda_{i\to\alpha}(x_i)$ takes the form

$$\lambda_{i\to\alpha}(x_i) = \frac{\hat{c}_\alpha}{\hat{c}_i}\left(\theta_i(x_i) + \sum_{\beta\in N(i)}\mu_{\beta\to i}(x_i)\right) - \mu_{\alpha\to i}(x_i),$$

where $\hat{c}_i = c_i + \sum_{\alpha \in N(i)} \hat{c}_\alpha$. *The block coordinate descent on* $\nu_{s\to\alpha}(\mathbf{x}_\alpha)$ *subject to the constraints takes the form:*

$$\nu_{s\to\alpha}(\mathbf{x}_\alpha) = \frac{1}{|N_{\mathcal{P}}(\alpha)|} \sum_{i\in N(\alpha)} \lambda_{i\to\alpha}(x_i) - \sum_{i\in N(\alpha)\cap s} \lambda_{i\to\alpha}(x_i)$$

The block coordinate descent steps above are guaranteed to converge for $\epsilon, c_\alpha, c_i \geq 0$*, and guaranteed to reach the optimum of (1) for* $\epsilon, c_\alpha, c_i > 0$.

Proof: First, we show the block coordinate descent on $\nu_{s\to\alpha}(\mathbf{x}_\alpha)$. To this end we assume $\lambda_{i\to\alpha}(x_i)\ \forall i, \alpha \in N(i), x_i$ to be fixed. We then choose a factor α and minimize the dual g w.r.t. $\nu_{s\to\alpha}(\mathbf{x}_\alpha)\ \forall s \in N_{\mathcal{P}}(\alpha), \mathbf{x}_\alpha$. Consider that part of the Lagrangian L^*_α of the dual program in (5) that contains variables $\nu_{s\to\alpha}(\mathbf{x}_\alpha)$ for a particularly chosen factor α, *i.e.*

$$L^*_\alpha = \epsilon\hat{c}_\alpha \ln \sum_{\mathbf{x}_\alpha} \exp\left(\frac{\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum_{i\in N(\alpha)\cap s} \lambda_{i\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha)}{\epsilon\hat{c}_\alpha}\right) + \sum_{\mathbf{x}_\alpha} \gamma_\alpha(\mathbf{x}_\alpha) \sum_{s\in N_{\mathcal{P}}(\alpha)} \nu_{s\to\alpha}(\mathbf{x}_\alpha),$$

Note that we have introduced Lagrange multipliers $\gamma_\alpha(\mathbf{x}_\alpha)$ for the constraints in (6). To obtain the stationary point of L^*_α, the derivative of L^*_α w.r.t. $\nu_{s\to\alpha}(\mathbf{x}_\alpha)\ \forall s \in N_{\mathcal{P}}(\alpha), \mathbf{x}_\alpha$ has to fulfill

$$\forall s \in N_{\mathcal{P}}(\alpha), \mathbf{x}_\alpha, \quad \frac{\partial L^*_\alpha}{\partial \nu_{s\to\alpha}(\mathbf{x}_\alpha)} = \frac{\exp\left(\frac{\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum_{i\in N(\alpha)\cap s} \lambda_{i\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha)}{\epsilon\hat{c}_\alpha}\right)}{\sum_{\mathbf{x}_\alpha} \exp\left(\frac{\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum_{i\in N(\alpha)\cap s} \lambda_{i\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha)}{\epsilon\hat{c}_\alpha}\right)} + \gamma_\alpha(\mathbf{x}_\alpha) = 0.$$

For simplicity we need to consider only the numerator, while taking one degree of freedom in the normalization. Taking the log of the numerator, and introducing some normalization constant $\beta(\mathbf{x}_\alpha)$, the above equation simplifies to

$$\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum_{i\in N(\alpha)\cap s} \lambda_{i\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha) = \beta(\mathbf{x}_\alpha). \tag{8}$$

Summing both sides over $s \in N_{\mathcal{P}}(\alpha)$ we compute this normalization constant via

$$\theta_\alpha(\mathbf{x}_\alpha) + \sum_{i\in N(\alpha)} \lambda_{i\to\alpha}(x_i) = \beta(\mathbf{x}_\alpha)|N_{\mathcal{P}}(\alpha)|.$$

We obtain the block coordinate descent updates for $\nu_{s\to\alpha}(\mathbf{x}_\alpha)$ given in the claim by plugging this result back into Eq. (8):

$$\nu_{s\to\alpha}(\mathbf{x}_\alpha) = \frac{1}{|N_{\mathcal{P}}(\alpha)|} \sum_{i\in N(\alpha)} \lambda_{i\to\alpha}(x_i) - \sum_{i\in N(\alpha)\cap s} \lambda_{i\to\alpha}(x_i).$$

The first part of above claim, *i.e*. the block coordinate descent steps on $\lambda_{i\to\alpha}(x_i)$ follow in an analoguous manner. We now assume $\nu_{s\to\alpha}(\mathbf{x}_\alpha)\ \forall \alpha, s \in N_{\mathcal{P}}(\alpha), \mathbf{x}_\alpha$ to be fixed. We then choose a node i and minimize the dual g w.r.t. $\lambda_{i\to\alpha}(x_i)$ $\forall \alpha \in N(i), x_i$. Consider again that part of the Lagrangian of the dual that contains variables $\lambda_{i\to\alpha}(x_i)\ \forall \alpha \in N(i), x_i$, *i.e.*

$$\begin{aligned} L^*_i \quad &= \quad \epsilon c_i \ln \sum_{x_i} \exp\left(\frac{\theta_i(x_i) - \sum_{\alpha\in N(i)} \lambda_{i\to\alpha}(x_i)}{\epsilon c_i}\right) \\ &\quad + \sum_{\alpha\in N(i)} \epsilon\hat{c}_\alpha \ln \sum_{\mathbf{x}_\alpha} \exp\left(\frac{\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum_{i\in N(\alpha)\cap s} \lambda_{i\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha)}{\epsilon\hat{c}_\alpha}\right). \end{aligned} \tag{9}$$

We subsumed all variables that are not considered when optimizing $\lambda_{i\to\alpha}(x_i)$ for a particularly chosen i in

$$\mu_{\alpha\to i}(x_i) = \epsilon\hat{c}_\alpha \ln \sum_{\mathbf{x}_\alpha\setminus x_i} \exp\left(\frac{\hat{\theta}_\alpha(\mathbf{x}_\alpha) + \sum_{j\in N(\alpha)\cap s\setminus i} \lambda_{j\to\alpha}(x_i) + \nu_{s\to\alpha}(\mathbf{x}_\alpha)}{\epsilon\hat{c}_\alpha}\right).$$

L_i^* given in Eq. (9) then simplifies to

$$L_i^* \quad = \quad \epsilon c_i \ln \sum_{x_i} \exp\left(\frac{\theta_i(x_i) - \sum_{\alpha \in N(i)} \lambda_{i\to\alpha}(x_i)}{\epsilon c_i}\right) + \sum_{\alpha \in N(i)} \epsilon \hat{c}_\alpha \ln \sum_{x_i} \exp\left(\frac{\mu_{\alpha\to i}(x_i) + \lambda_{i\to\alpha}(x_i)}{\epsilon \hat{c}_\alpha}\right).$$

Computation of the stationary point amounts to solving $\frac{\partial L_i^*}{\partial \lambda_{i\to\alpha}(x_i)} = 0 \; \forall \alpha \in N(i), x_i$, *i.e.* explicitly:

$$\frac{\exp\left(\frac{\mu_{\alpha\to i}(x_i)+\lambda_{i\to\alpha}(x_i)}{\epsilon\hat{c}_\alpha}\right)}{\sum_{x_i} \exp\left(\frac{\mu_{\alpha\to i}(x_i)+\lambda_{i\to\alpha}(x_i)}{\epsilon\hat{c}_\alpha}\right)} = \frac{\exp\left(\frac{\theta_i(x_i)-\sum_{\alpha\in N(i)}\lambda_{i\to\alpha}(x_i)}{\epsilon c_i}\right)}{\sum_{x_i} \exp\left(\frac{\theta_i(x_i)-\sum_{\alpha\in N(i)}\lambda_{i\to\alpha}(x_i)}{\epsilon c_i}\right)} \quad \forall \alpha \in N(i), x_i.$$

Again, the denominator can be neglected when introducing a normalization degree of freedom. We consequently obtain the proportionality

$$\frac{\mu_{\alpha\to i}(x_i) + \lambda_{i\to\alpha}(x_i)}{\hat{c}_\alpha} \propto \frac{\theta_i(x_i) - \sum_{\alpha\in N(i)} \lambda_{i\to\alpha}(x_i)}{c_i}. \tag{10}$$

Summing both sides $\forall \alpha \in N(i)$ results in

$$\sum_{\alpha\in N(i)} \lambda_{i\to\alpha}(x_i) \propto \frac{\sum_{\alpha\in N(i)} \hat{c}_\alpha}{c_i + \sum_{\alpha\in N(i)} \hat{c}_\alpha} \theta_i(x_i) - \frac{c_i}{c_i + \sum_{\alpha\in N(i)} \hat{c}_\alpha} \sum_{\alpha\in N(i)} \mu_{\alpha\to i}(x_i),$$

which, when plugged back into Eq. (10), gives

$$\lambda_{i\to\alpha}(x_i) \propto \frac{\hat{c}_\alpha}{\hat{c}_i}\left(\theta_i(x_i) + \sum_{\beta\in N(i)} \mu_{\beta\to i}(x_i)\right) - \mu_{\alpha\to i}(x_i)$$

as claimed initially. Note that proportionality is used to stabilize the algorithm, *e.g.*, by normalizing the messages.

Note that for $\epsilon, c_i, c_\alpha > 0$, g is convex and thus the block coordinate descent method is guaranteed to find the global optimum. When any of these variables is zero, Danskin's theorem states that its corresponding subgradient is the convex combination of the maximal assignments (cf. [1]). We are guaranteed to converge as the dual is lower bounded by the primal. We may however find a local optimum only, i.e., obtain a non-zero duality gap. □

By abbreviating $n_{i\to\alpha}(x_i) = \exp \lambda_{i\to\alpha}(x_i)$, $m_{\alpha\to i}(x_i) = \exp \mu_{\alpha\to i}(x_i)$ and $n_{s\to\alpha}(\mathbf{x}_\alpha) = \exp \nu_{s\to\alpha}(\mathbf{x}_\alpha)$ we obtain the *Distributed Convex Belief Propagation* algorithm given in the main body of the paper.

References

[1] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2^{nd} edition, 1999. 4

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 2