Supplementary Material for "A Principled Deep Random Field Model for Image Segmentation"

Pushmeet Kohli
Microsoft Research
Cambridge, UK

Anton Osokin
Moscow State University
Moscow, Russia

Stefanie Jegelka
UC Berkeley
Berkeley, CA, USA

In this supplement, we provide details on the multi-label model and also prove some of the theoretical results in the main paper.

1. Multi-label models

Let $\mathcal{L}$ be the set of all labels that a node can take. We will denote labels $\mathfrak{a} \in \mathcal{L}$ by fractional characters.

The multi-label extension of the directed cooperative cut energy that is defined in the main paper is

$$\Psi_g(\mathbf{x}) = \sum_{\mathfrak{a} \in \mathcal{L}} F_g\Big(\sum_{(i,j) \in g} \psi_{\mathfrak{a}}(x_i, x_j) \Big), \tag{1}$$

where the pairwise function $\psi_{\mathfrak{a}}$ is defined as (see also [3, Sec. 5.4.1])

$$\psi_{\mathfrak{a}}(x_i, x_j) = \begin{cases} \theta_{ij} & \text{if } x_i = \mathfrak{a} \text{ and } x_i \neq x_j \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

As before, we introduce group variables $h_{g,\mathfrak{a}}$, indexed by edge groups and labels, because we have one function F_g for each label. If $F_g(a) = \min\{a, T\}$ is the truncation with one breakpoint, then we can write

$$\Psi_{g,\mathfrak{a}}(\mathbf{x}) = \min_{h_{g,\mathfrak{a}} \in \{0,1\}} h_{g,\mathfrak{a}} \left(\sum_{(i,j) \in g} \psi_{\mathfrak{a}}(x_i, x_j) \right) + (1 - h_{g,\mathfrak{a}}) T. \tag{3}$$

If the vector $\mathbf{h}$ of variables $h_{g,\mathfrak{a}}$ is fixed, then the entire energy becomes a sum of pairwise potentials of the form

$$\psi_{i,j}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ \theta_{ij}(\mathfrak{a}) + \theta_{ji}(\mathfrak{b}) & \text{if } x_i = \mathfrak{a}, x_j = \mathfrak{b}, \end{cases} \tag{4}$$

where $\theta_{ij}(\mathfrak{a})$ is the weight of $\psi_{\mathfrak{a}}(x_i, x_j)$ under the current assignment of $h_{g,\mathfrak{a}}$. In the above case, $\theta_{ij}(\mathfrak{a}) = \theta_{ij} h_{g,\mathfrak{a}}$.

We will perform expansion moves [1] with such a potential. For a given assignment $\mathbf{x} \in \mathcal{L}^n$, an expansion move with respect to label $\mathfrak{a}$ is allowed to change any label of $\mathbf{x}$ to $\mathfrak{a}$, but can make no other changes. Formally, define the set

$$\mathcal{X}(\mathbf{x}, \mathfrak{a}) = \{\mathbf{y} \in \mathcal{L}^n \mid y_i = x_i \text{ or } y_i = \mathfrak{a} \text{ for all } 1 \leq i \leq n\}. \tag{5}$$

An *expansion move* with respect to a label $\mathfrak{a}$ and the current labeling $\mathbf{x}$ finds

$$\mathbf{y}' \in \operatorname*{arg\,min}_{\mathbf{y} \in \mathcal{X}(\mathbf{x}, \mathfrak{a})} E(\mathbf{y}). \tag{6}$$

A *stationary point* $\mathbf{x}'$ with respect to all labels is one that for all labels $\mathfrak{a} \in \mathcal{L}$ satisfies $\mathbf{x}' \in \arg\min_{\mathbf{y} \in \mathcal{X}(\mathbf{x}, \mathfrak{a})} E(\mathbf{y})$.

For pairwise potentials such as 4, expansion moves can be computed as the minimum cut in a graph. This proves Lemma 2 in the main paper:

Lemma 2. *The multi-label model can be reduced to a non-submodular pairwise model analogous to the binary model. If $|\mathcal{L}|$ and $|\mathcal{G}|$ are constants, then, with the help of $|\mathcal{L}||\mathcal{G}|$ auxiliary variables, we can compute an exact expansion move in polynomial time.*

In the sequel, we will denote the vector of all $h_{g,\mathfrak{a}}$ variables by $\mathbf{h}$. We re-state Theorem 1 in the main paper in a bit more detail:

Theorem 1. *Let $\mathbf{x}^* \in \arg\min E(\mathbf{x})$ be an optimal MAP labeling for a cooperative cut energy composed of terms of the form (1). For a given assignment $\mathbf{h} \in \{0,1\}^{|\mathcal{G}||\mathcal{L}|}$ of the group variables, let $\mathbf{x}(\mathbf{h})$ be a stationary point of the expansion moves with respect to all labels. Then*

$$\min_{\mathbf{h} \in \{0,1\}^{|\mathcal{G}|}} E(\mathbf{x}(h)) \leq 2cE(\mathbf{x}^*), \tag{7}$$

where $c = \max_{\mathfrak{a},\mathfrak{b} \in \mathcal{L}, g \in \mathcal{G}} F'_{\mathfrak{a}}(0) / F'_{\mathfrak{b}}(\sum_{(i,j) \in g} \theta_{ij})$ is the ratio of the largest and smallest slopes of F.

Proof. For each $\mathbf{h}$, we find a labeling $\mathbf{x}(\mathbf{h})$ that is a stationary point with respect to all labels. In the end, we will take the best of the solutions $\mathbf{x}(\mathbf{h})$ that we found. This can be done via a variant of the graph cut algorithm in [1]. To ease

notation in the proof, we introduce the notation $E_h(\mathbf{x}, \mathbf{h})$ for the energy function that is a function of $\mathbf{h}$ (instead of minimizing over it as in Equation 3. With this notation, $E(\mathbf{x}) = \min_{\mathbf{h}} E_h(\mathbf{x}, \mathbf{h})$.

An adaptation of Theorem 6.6 in [1] implies that

$$E_h(\mathbf{x}(\mathbf{h}), \mathbf{h}) \leq 2c \min_{\mathbf{y} \in \mathcal{L}^n} E_h(\mathbf{y}, \mathbf{h}). \tag{8}$$

The constant c arises since the slopes of F (determined by the assignment of $\mathbf{h}$) scale the pairwise weights θ_{ij}, and this leads to label-sensitive pairwise potentials in the framework of [1]. Let $\mathbf{h}^*$ be the optimal assignment of $\mathbf{h}$ for the optimal solution $\mathbf{x}^*$, i.e., $E_h(\mathbf{x}^*, \mathbf{h}^*) = E(\mathbf{x}^*)$. Since the bound (8) holds for all assignments $\mathbf{h}$, we get that

$$\begin{aligned}
&\min_{\mathbf{h} \in \{0,1\}^{|\mathcal{G}|}} E(\mathbf{x}(\mathbf{h})) && (9)\\
&= \min_{\mathbf{h} \in \{0,1\}^{|\mathcal{G}|}} \min_{\mathbf{h}' \in \{0,1\}^{|\mathcal{G}|}} E_h(\mathbf{x}(\mathbf{h}), \mathbf{h}') && (10)\\
&\leq \min_{\mathbf{h} \in \{0,1\}^{|\mathcal{G}|}} E_h(\mathbf{x}(\mathbf{h}), \mathbf{h}) && (11)\\
&\leq E_h(\mathbf{x}(\mathbf{h}^*), \mathbf{h}^*) && (12)\\
&\leq 2c \min_{\mathbf{y} \in \mathcal{L}^n} E_h(\mathbf{y}, \mathbf{h}^*) = 2cE_h(\mathbf{x}^*). && (13)
\end{aligned}$$

□

2. Arbitrary monotone concave functions

We consider the energy

$$E(\mathbf{x}) = \sum_i \psi_i(x_i) + \sum_{g \in \mathcal{G}} \Psi_g(\mathbf{x}), \text{ where} \tag{14}$$

$$\Psi_g(\mathbf{x}) = F_g\Big(\sum_{(i,j) \in \mathcal{E}_g} \psi_{ij}(x_i, x_j)\Big). \tag{15}$$

We make the following assumptions:

1. the pairwise potentials are of the form $\psi_{ij}(x_i, x_j) = \theta_{ij}|x_i - x_j|_+ \geq 0$ or $\psi_{ij}(x_i, x_j) = \theta_{ij}|x_i - x_j| \geq 0$;
2. the functions $F_g : \mathbb{R}_+ \to \mathbb{R}_+$ are nonnegative, monotone increasing scalar concave functions that satisfy $F_g(\lambda y) \leq \lambda F_g(y)$ for all $y \geq 0$;
3. the energy E is nonnegative.
4. $|\mathcal{G}|$ is constant.

We re-state Lemma 1 from the main paper:

Lemma 1. *If the energy (14) satisfies (1)-(4), then there is an FPTAS for minimizing this energy,* i.e.*, there is an algorithm that runs in time polynomial in $1/\epsilon$ and n and returns a solution $\mathbf{x}$ with $E(\mathbf{x}) \leq (1+\epsilon)E(\mathbf{x}^*)$, where $\mathbf{x}^*$ is the optimizing MAP assignment.*

Proof. We will treat the unary potentials as an additional edge group (this is the group of terminal edges), and set $k = |\mathcal{G}| + 1$. Let M be such that the energy functions takes values between $1/M$ and M. We create a set of slopes $\mathcal{A} = \{\alpha = rw \mid r \in \mathcal{R}, w \in \mathcal{W}\}$, where $\mathcal{R} = \{2^0, 2^1, \ldots, 2^{\lceil \log_2 M \rceil}\}$ and $\mathcal{W} = \{1, 2, \ldots \lceil \frac{2(k-1)}{\epsilon} \rceil\}$. We will essentially represent Ψ_g by a piecewise linear function with pieces

$$\widehat{\Psi}_g(\mathbf{x}; \alpha) = \sum_{(i,j) \in g} \alpha \psi_{ij}(x_i, x_j) \tag{16}$$

with slopes $\alpha \in \mathcal{A}$.

To see how such a function connects to an approximation by functions $\Psi'(\mathbf{x}) = \min\{\beta \sum_{(i,j) \in g} \psi_{ij}(x_i, x_j), T\}$, observe that fixing $\mathbf{h}$ in the algorithm in the main paper corresponds to assigning a "slope" $\sum_\ell \beta_\ell h_\ell$ to each edge group. We find a minimizer for each slope, and, among those minimizers, select the one minimizing the actual energy. We will see that doing the same for the slopes α will suffice. The pieces (16) can be written in terms of the functions $\Psi'(\mathbf{x}) = \min\{\beta \sum_{(i,j) \in g} \psi_{ij}(x_i, x_j), T\}$. To do so, we sort the slopes in $\mathcal{A}$ in increasing order, and number them $\alpha_1, \alpha_2, \ldots$. The corresponding β_i are then $\beta_1 = \alpha_1$ and $\beta_i = \alpha_i - \alpha_{i-1}$ $(i > 1)$, so that $\alpha_j = \sum_{i \leq j} \beta_i$. (Those β_i are only needed for the conceptual connection.)

Recall that a particular assignment $\mathbf{h}$ in the algorithm in the main paper corresponds to assigning a slope $\alpha_g \in \mathcal{A}$ to each group. We hence imitate the algorithm by computing the minimizers $\mathbf{x}(\mathbf{a})$ for all $\mathbf{a} \in \mathcal{A}^{\mathcal{G}'}$ and for

$$\widehat{E}(\mathbf{x}; \mathbf{a}) = \sum_{g \in \mathcal{G}'} \widehat{\Psi}_g(\mathbf{x}; \alpha_g), \tag{17}$$

where $\mathcal{G}'$ is the extended set of edge groups that includes the extra group for the unary potentials. We then evaluate the energy E at each assignment $\mathbf{x}(\mathbf{a})$, and choose the best among those assignments.

To analyze this strategy, we will take the cut viewpoint and draw connections to multi-objective optimization to be able to use ideas from [5]. For the cut viewpoint, we introduce binary variables $y_{ij} = |x_i - x_j|$. Each assignment $\mathbf{x}$ corresponds to a cut $\mathbf{y}$ and vice versa.

We can write the energy as

$$E(\mathbf{x}) = \sum_{g \in \mathcal{G}'} F_g\Big(\sum_{(i,j) \in \mathcal{E}} \theta_{ij}^g y_{ij}\Big), \tag{18}$$

where $\theta_{ij}^{(g)} = 0$ if $(i,j) \notin \mathcal{G}'$. Written in this form, the energy can be viewed as a function combining k linear objectives $\theta^{(g)}\mathbf{y} = \sum_{ij} \theta_{ij}^{(g)} y_{ij}$ into one objective by using a concave function $F(\theta^{(1)}\mathbf{y}, \ldots, \theta^{(k)}\mathbf{y}) = \sum_g F_g(\theta^{(g)}\mathbf{y})$.

A theorem in [2] says that the set $\{\mathbf{x}(\mathbf{a}) \mid \mathbf{a} \in \mathcal{A}^{\mathcal{G}'}\}$ is an *ϵ-approximate convex Pareto-optimal front* corresponding to the k linear functions $\theta^{(g)}\mathbf{y}$ for the constraint that

$\mathbf{y}$ is the indicator vector of a cut. An approximate convex Pareto-optimal set C_ϵ is a set of solutions such that for each feasible[1] $\mathbf{y} \in [0,1]^{\mathcal{E}}$ there exists an $y' \in C_\epsilon$ such that $\theta^{(g)}\mathbf{y}' \leq \theta^{(g)}\mathbf{y}$ for all $g \in \mathcal{G}'$.

Lemma 3.2 in [5] states that the convex hull of C_ϵ contains a $(1+\epsilon)$-optimal point. Since our energy is concave, the minimum over the convex hull is attained at a corner point, and therefore the search over all corner points (*i.e.* over the set C_ϵ) will yield a $(1+\epsilon)$-approximate solution.

Finally, the cardinality $|\mathcal{A}|$ (which determines the number of optimization problems to solve) is polynomial in $1/\epsilon$ and n. □

3. Details on experiments

Here, we provide some more details on how the potentials in the experiments were computed.

The unary potentials are computed by fitting a Gaussian mixture model with 5 components to pixels of seed regions. We added user scribbles to the MSRC data for the multi-label experiments to be compatible with the binary label experiments.

We use an 8-neighbor graph structure and contrast-dependent Potts pairwise potentials $\theta_{ij} = 2.5 + 47.5\exp(-0.5\|I_i - I_j\|^2/\sigma)$, where σ is the mean of the color gradients in the image.

As in [4], the edge groups were defined by defining a 3-dimensional feature vector $\phi(i,j) = I_i - I_j$ for each edge (i,j) using the pixel RGB values I_i. We then cluster the edges using these features (using k-means), and each cluster becomes a group $g \in \mathcal{G}$. The discount functions F were the same in the binary and multi-label case:

$$\Psi_g(\mathbf{x}) = \lambda \min\left\{\sum_{ij\in\mathcal{E}_g} \theta_{ij}(x_i - x_j)_+, \sum_{ij\in\mathcal{E}_g} \alpha\theta_{ij}(x_i - x_j)_+ + \theta_g\right\} \tag{19}$$

For the multi-label functions, we adapt this function according to Section 1 in this supplement and make it label-dependent.

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001. 1, 2

[2] I. Diakonikolas and M. Yannakakis. Succinct approximate convex pareto curves. In *SODA*, 2008. 2

[3] S. Jegelka. *Combinatorial Problems with Submodular Coupling in Machine Learning and Computer Vision*. PhD thesis, ETH Zurich, 2012. 1

[4] S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: Coupling edges in graph cuts. In *CVPR*, pages 1897–1904, 2011. 3

[5] S. Mittal and A. Schulz. An FPTAS for optimizing a class of low-rank functions over a polytope. *Optimization online*, 2011. 2, 3

[1] The "feasible set" here is the convex hull of all cut indicator vectors