Supplementary Material: Revisiting Uncertainty in Graph Cut Solutions

Daniel Tarlow
Dept. of Computer Science
University of Toronto
dtarlow@cs.toronto.edu

Ryan P. Adams
School of Engineering and Applied Sciences
Harvard University
rpa@seas.harvard.edu

Analysis of Kohli-Torr as Approximate Gibbs Distribution Inference

Here we treat the min-marginal uncertainty of Kohli & Torr (KT) [1] as a procedure for computing approximate marginal probabilities under the Gibbs distribution that is standard in conditional random field (CRF) models. In the very simple case where there are no interactions between variables in the model (i.e., a "unary only" model), the approximate KT marginals match the true marginals. The main result here is that even after adding just submodular pairwise interactions to the model, the marginals estimated by KT can be off by a factor that is exponentially large in the number of variables in the network.

The crux of the problem (which will be made more precise below) is fairly simple: if we approximate the sum of probabilities over a subset of configurations by instead maximizing over the probabilities of states in the subset, how bad can the approximation be? The analysis simply considers how to parameterize models such that the summation required to compute marginals gives as different value as possible from the maximizations used by KT.

We are concerned with this question, because when learning with the KT procedure, approximate gradients are calculated using these approximate marginals. This analysis adds some theoretical explanation for the empirical behavior of the KT learning procedure in our experiments.

We will consider models over a set of binary variables $\boldsymbol{y} \in \mathcal{Y} = \{0,1\}^D$. We assume we have an energy function $E(\boldsymbol{y})$ (for simplicity dropping the dependence on input $\boldsymbol{x}$ and the explicit dependence on parameters $\boldsymbol{w}$), then under the Gibbs distribution, the probability of a configuration $\boldsymbol{y}$ is

$$p'(\boldsymbol{y}) = \exp\{-E(\boldsymbol{y})\} \tag{1}$$

$$\mathcal{Z} = \sum_{\boldsymbol{y}' \in \mathcal{Y}} p'(\boldsymbol{y}') \tag{2}$$

$$p(\boldsymbol{y}) = \frac{1}{\mathcal{Z}} p'(\boldsymbol{y}). \tag{3}$$

The marginal probability for y_d under the Gibbs distribution is

$$p(y_d) = \sum_{\hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = y_d} p(\hat{\boldsymbol{y}}) \tag{4}$$

$$= \frac{\sum_{\hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = y_d} \exp\{-E(\hat{\boldsymbol{y}})\}}{\sum_{y'_d \in \{0,1\}} \sum_{\hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = y'_d} \exp\{-E(\boldsymbol{y}')\}}. \tag{5}$$

That is, we are summing over the probability of all configurations $\boldsymbol{y}'$ that are consistent with the given setting of y_d.

The KT approximation is to define

$$\tilde{p}(y_d) = \frac{\exp\{-\min_{\hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = y_d} E(\hat{\boldsymbol{y}})\}}{\sum_{y'_d \in \{0,1\}} \exp\{-\min_{\hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = y'_d} E(\hat{\boldsymbol{y}})\}} \tag{6}$$

$$= \frac{\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = y_d} \exp\{-E(\hat{\boldsymbol{y}})\}}{\sum_{y'_d \in \{0,1\}} \max_{\hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = y'_d} \exp\{-E(\hat{\boldsymbol{y}})\}}. \tag{7}$$

Comparing Eq. 5 to Eq. 7, we see that they are identical, except the maximizations in Eq. 7 are replaced by sums to get the expression Eq. 5 for the true marginals. For convenience, let $\tilde{p}'$ be the unnormalized value of the maximization:

$$\tilde{p}'(y_d) = \max_{\hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = y_d} \exp\{-E(\hat{\boldsymbol{y}})\}. \tag{8}$$

and similarly

$$p'(y_d) = \sum_{\hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = y_d} \exp\{-E(\hat{\boldsymbol{y}})\}. \tag{9}$$

The question we ask in the following is, "how large can we make the approximation factor $f = \tilde{p}(y_d)/p(y_d)$, for some d and y_d?" In other words, in the worst case, how bad can the approximation of KT be?

0.1. General Models

We begin with the simplest case, where there are no restrictions on the parameterization of the energy function. In this case, we have the freedom to assign an independent probability value to each of the 2^D joint configurations $\boldsymbol{y}$, so long as we restrict the probabilities to be non-negative, and the sum to be equal to 1.

Without loss of generality, suppose that the KT approximate marginal is $\tilde{p}(y_d = 1) = \lambda$, and thus $\tilde{p}(y_d = 0) = 1 - \lambda$, for $\lambda \in [0, 1]$. This implies that

$$\frac{\tilde{p}'(y_d = 1)}{\tilde{p}'(y_d = 0) + \tilde{p}'(y_d = 1)} = \lambda. \tag{10}$$

Again without loss of generality, we can further assume that $\tilde{p}'(y_d = 0) + \tilde{p}'(y_d = 1) = 1$, so

$$\tilde{p}'(y_d = 1) = \lambda, \text{ and} \tag{11}$$
$$\tilde{p}'(y_d = 0) = 1 - \lambda. \tag{12}$$

Let S_0 be the set of states where $y_d = 0$:

$$S_0 = \{\hat{\boldsymbol{y}} \mid \hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = 0\} \tag{13}$$

and define S_1 similarly:

$$S_1 = \{\hat{\boldsymbol{y}} \mid \hat{\boldsymbol{y}} \in \mathcal{Y}, \hat{y}_d = 1\} \tag{14}$$

So far, we know that

$$\max_{\hat{\boldsymbol{y}} \in S_1} p'(\hat{\boldsymbol{y}}) = \lambda \tag{15}$$
$$\max_{\hat{\boldsymbol{y}} \in S_0} p'(\hat{\boldsymbol{y}}) = 1 - \lambda. \tag{16}$$

The size of these sets are $|S_0| = |S_1| = 2^{D-1}$, so the largest that $p'(y_d = 1)$ could be is $\lambda \cdot 2^{D-1}$. The proof of this is straightforward: suppose that $p'(y_d = 1) > \lambda \cdot 2^{D-1}$. Then there must be some configuration $\boldsymbol{y} \in S_1$ such that $p'(\boldsymbol{y}) > \lambda$. But this contradicts Eq. 15. Finally, $p'(y_d = 1) = \lambda \cdot 2^{D-1}$ is achievable by assigning $p'(\boldsymbol{y}) = \lambda$ for all $\boldsymbol{y} \in S_1$.

The smallest that $p'(y_d = 0)$ could be is $1 - \lambda$, because probabilities are non-negative, and at least one configuration $\boldsymbol{y} \in S_0$ must have $p'(\boldsymbol{y}) = 1 - \lambda$ to be consistent with Eq. 16.

Since S_0 and S_1 are disjoint, the minimum achievable value of $p(y_d = 0)$ consistent with Eq. 15 and Eq. 16 comes from simultaneously maximizing $p'(y_d = 1)$ and minimizing $p'(y_d = 0)$. This yields

$$p(y_d = 0) = \frac{p'(y_d = 0)}{p'(y_d = 0) + p'(y_d = 1)} \tag{17}$$
$$= \frac{1 - \lambda}{1 - \lambda + \lambda \cdot 2^{D-1}}. \tag{18}$$

The approximation factor is then:

$$\frac{\tilde{p}(y_d = 0)}{p(y_d = 0)} = 1 - \lambda + \lambda \cdot 2^{D-1}, \tag{19}$$

which is exponentially large in D for constant values of λ (say $\lambda = .5$).

0.2. Pairwise Submodular Models

We now consider the case where we restrict the parameterization, so that we cannot choose arbitrary values for the probability assigned to a joint assignment $\boldsymbol{y}$. Instead, we require that the model be a pairwise graphical model with only submodular interactions. In this section, we show that there are such models where the same basic strategy as before is applicable, and thus the approximation factor is still exponentially bad.

Consider a pairwise model where there is a single root, y_r, which is shares an edge with all other variables. There are no other edges, so this is also a tree-structured model. Let $\boldsymbol{\theta}$ be the set of potentials, and define the unnormalized probability as a typical graphical model, where the only unary potential is on the root:

$$p'(\boldsymbol{y}) = \theta_r(y_r) \prod_d \theta_{rd}(y_r, y_d). \tag{20}$$

Let $\theta_r(1) = \lambda$, $\theta_r(0) = 1 - \lambda$ for $\lambda \in [0, 1]$, and parameterize pairwise interactions for all d as follows:

$$\theta_{rd}(y_r = 0, y_d = 0) = 1 \tag{21}$$
$$\theta_{rd}(y_r = 0, y_d = 1) = 0 \tag{22}$$
$$\theta_{rd}(y_r = 1, y_d = 0) = 1 \tag{23}$$
$$\theta_{rd}(y_r = 1, y_d = 1) = 1. \tag{24}$$

This is clearly a submodular interaction. Now, if y_r is constrained to be 1, there are 2^{D-1} joint configurations with support, each getting an unnormalized probability of λ. If y_r is constrained to 0, then there is only a single joint configuration with support (the all 0's assignment), and it gets unnormalized probability of $1-\lambda$. This model is then equivalent to the model constructed in the previous section, and thus the approximation factor is also exponentially bad.

Bootstrap Analysis Results

See Fig. 1 and Fig. 2.

References

[1] P. Kohli and P. H. S. Torr. Measuring uncertainty in graph cut solutions. *Computer Vision and Image Understanding*, 112(1):30–38, 2008. 1

		Log Lik	Accuracy	AUC
Aero	KT-final	-.35±.005 (-.29±.004)	87.3±.2 (89.7±.1)	.85±.018 (.87±.020)
	KT-f_{best}	-.32±.007 (-.28±.005)	88.1±.3 (89.9±.2)	.85±.018 (.87±.019)
	Ours	**-.26±.007** (**-.24±.005**)	**88.9±.3** (**90.4±.2**)	**.90±.014** (**.90±.014**)
Car	KT-final	-.48±.008 (-.65±.008)	84.1±.3 (78.7±.3)	.69±.036 (**.65±.022**)
	KT-f_{best}	-.39±.007 (**-.51±.006**)	**86.2±.3** (80.0±.3)	.66±.029 (.62±.023)
	Ours	**-.35±.009** (**-.51±.006**)	86.1±.5 (**81.4±.3**)	**.76±.035** (**.65±.032**)
Cow	KT-final	-.54±.009 (-.64±019)	79.7±.3 (75.9±.8)	.76±.040 (.77±.038)
	KT-f_{best}	-.47±.004 (-.52±.008)	80.9±.3 (76.7±.5)	.66±.040 (.65±.038)
	Ours	**-.38±.005** (**-.41±.006**)	**82.5±.4** (**79.9±.4**)	**.84±.023** (**.82±.034**)
Dog	KT-final	-.52±.009 (-.45±.004)	81.8±.4 (84.7±.1)	.64±.029 (.66±.027)
	KT-f_{best}	-.43±.004 (-.38±.003)	**84.1±.2** (86.7±.1)	.62±.023 (.66±.026)
	Ours	**-.38±.004** (**-.34±.004**)	84.0±.2 (**86.8±.2**)	**.76±.025** (**.79±.028**)

Figure 1. Results for binary models. Format is "Train (Test)". ± indicates the standard deviation of the result under bootstrap resamplings.

		$\cap/\cup$ **Before**	$\cap/\cup$ **After**	**Change**
Aero	KT-final	63.4±2.2	63.9±2.1	.5±.3
	KT-f_{best}	60.7±2.1	61.8±1.9	1.1±.4
	Ours	**64.1±2.2**	**66.6±2.0**	**2.5±.7**
Car	KT-final	**43.5±1.4**	44.2±1.2	.7±.3
	KT-f_{best}	40.7±1.0	43.3±.9	2.6±.5
	Ours	41.8±1.0	**45.9±1.0**	**4.1±.8**
Cow	KT-final	**51.2±3.7**	51.7±3.6	.5±.5
	KT-f_{best}	40.7±1.8	47.1±2.7	**6.4±2.0**
	Ours	46.5±1.9	**52.8±2.8**	6.3±1.8
Dog	KT-final	**52.1±2.6**	52.0±2.5	-.1±.5
	KT-f_{best}	44.5±1.1	47.2±1.2	2.7±.6
	Ours	48.9±1.4	**55.3±1.9**	**6.4±1.3**

Figure 2. Test results for maximizing surrogate expected $\frac{\cap}{\cup}$ score. "Before" corresponds to predicting the mode of Q; "After" is the prediction from our expected score maximization routine. ± indicates the standard deviation of the result under bootstrap resamplings.