Improving Optimization-Based Approximate Inference by Clamping Variables

Junyao Zhao, Josip Djolonga, Sebastian Tschiatschek, Andreas Krause
Department of Computer Science, ETH Zürich
zhaoju@student.ethz.ch,{josipd,tschiats}@inf.ethz.ch,krausea@ethz.ch

Abstract

While central to the application of probabilistic models to discrete data, the problem of marginal inference is in general intractable and efficient approximation schemes need to exploit the problem structure. Recently, there have been efforts to develop inference techniques that do not necessarily make factorization assumptions about the distribution, but rather exploit the fact that sometimes there exist efficient algorithms for finding the MAP configuration. In this paper, we theoretically prove that for discrete multi-label models the bounds on the partition function obtained by two of these approaches, Perturb-and-MAP and the bound from the infinite Rényi divergence, can be only improved by clamping any subset of the variables. For the case of log-supermodular models we provide a more detailed analysis and develop a set of efficient strategies for choosing the order in which the variables should be clamped. Finally, we present a number of numerical experiments showcasing the improvements obtained by the proposed methods on several models.

1 INTRODUCTION

A key challenge in probabilistic inference is that of computing the normalizing partition function of unnormalized probability distributions, which would enable the evaluation of the probability of evidence and marginal probabilities. While for low tree-width graphs one can use the junction tree algorithm to normalize the distribution, in general the worst case time complexity is exponential in the number of random variables. Consequently, we very often have to resort to techniques for approximating the partition function. In this paper we will focus on a family of such techniques that make assumptions only about the optimization characteristics of the used energy function.

In this paper we will consider discrete probabilistic models $P(X_1, X_2, \ldots, X_N)$ defined over N random variables $X_1, \ldots, X_N$, such that each variable X_i takes on values in $\mathcal{L} = \{0, 1, \ldots, L-1\}$. We will assume that the distribution is given in the form

$$P(\mathbf{x}) = \frac{1}{\mathcal{Z}(f)} \exp(-f(\mathbf{x})), \tag{1}$$

where $f\colon \{0, 1, \ldots, L-1\}^n \to \mathbb{R}$ is an arbitrary *energy function*, and $\mathcal{Z}(f)$ is the *partition function*, which ensures normalization of the distribution. As already suggested, computing $\mathcal{Z}$ is provably hard, and is known to be #P-hard even for binary pairwise models [1, 2].

We make *no factorization assumptions* about f, and thus we cannot directly apply most existing algorithms and techniques relying on a log-linear representation [3]. Instead, we will assume that we can efficiently compute the MAP configuration under any linear perturbation. Specifically, we assume that we can efficiently solve

$$\underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x}) + \sum_{i=1}^{n} g_{i,x_i},$$

for any set of numbers $g_{i,x_i} \in \mathbb{R}$.

The above assumption has two important consequences. First, the energies satisfying this assumption are *closed under clamping*, i.e. if we fix any x_k to have value l, we can solve the problem $\min_{\mathbf{x}\colon x_k = l} f(\mathbf{x}) + \sum_{i=1}^{n} g_{i,x_i}$ by letting $g_{k,l} \to -\infty$. Second, there are at least two techniques for approximate marginal inference which can be directly applied under that sole assumption: *(i)* the Perturb-and-MAP framework of [4], and *(ii)* the minimization of the infinite Rényi divergence [5], known as L-FIELD for log-supermodular models [6, 7]. In this paper, we show both theoretically and experimentally that

these two properties can be safely combined to yield better approximate inference results. Specifically, we prove that performing inference after clamping can *only improve* the estimate of the partition function.

An important class of energy functions that satisfies the above assumption, and which we will also analyze as a special case in more detail, is that of *binary* ($L = 2$) log-supermodular distributions. Their energy has to satisfy $f(\mathbf{x}\vee\mathbf{y})+f(\mathbf{x}\wedge\mathbf{y}) \leq f(\mathbf{x})+f(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$, where $\wedge$ and $\vee$ denote element-wise minimum and maximum respectively. These energies can be minimized in polynomial time [8], very efficiently in many cases [9, 10]. The distributions corresponding to these energy functions have non-negative correlations [11], which also explains why they are sometimes referred to as *attractive*. They have been extensively used in computer vision for semantic image segmentation—both pairwise models (also known as graph-cuts) [12], but also models with complicated higher-order potentials [13].

Previous work. In this work we analyze the bound obtained from the infinite Rényi divergence [5], which has been used in [7] for log-supermodular models (in the same paper the authors show that this problem is equivalent to the L-FIELD bound of Djolonga and Krause [6]). We will also study the Perturb-and-MAP method, which was first proposed by Papandreou and Yuille [4] and then analyzed in more detail by Hazan and Jaakkola [14]. Recently, Shpakova and Bach [15] have drawn an interesting connection between these two techniques for log-supermodular models. The idea of improving approximate inference techniques by clamping has been studied by Weller and Jebara [16] and Weller and Domke [17]. In these papers, the authors have answered in the affirmative the question if clamping always helps for mean-field, tree-reweighted belief-propagation and traditional belief propagation (for log-supermodular models) [3]. Unfortunately, these inference techniques can not be easily applied to models with higher order factors without additional factorization assumptions, or a restriction to specific families [18, 19].

Contributions. We prove that clamping can only improve the partition function estimate obtained from Perturb-and-MAP and the bound arising from the infinite Rényi divergence for any discrete model. Furthermore, for log-supermodular models we propose heuristics for choosing which variables to clamp. Finally, we provide an empirical analysis to demonstrate the benefits of clamping and the usefulness of the proposed heuristics.

2 PRELIMINARIES

2.1 MATHEMATICAL SETTING

We will represent each configuration $\mathbf{X} = (X_1 = z_1, X_2 = z_2, \ldots, X_N = z_N)$ by a vector $\mathbf{x} \in \{0,1\}^{NL}$ using the one-hot 1-of-L encoding, i.e. each variable X_i has a corresponding block $\mathbf{x}_i \in \{0,1\}^L$ of $\mathbf{x}$ with a single 1 at the position corresponding to z_i, i.e. $x_{i,j} = [\![z_i = j]\!]$. We also define the set of all admissible vectors $\mathbf{x}$ as

$$\mathcal{X} = \{\mathbf{x} \in \{0,1\}^{NL} \mid \forall i\colon \mathbf{1}^T\mathbf{x}_i = 1\},$$

where $\mathbf{1}$ is the vector of all ones of length L. The partition function is $\mathcal{Z}(f) = \sum_{\mathbf{x}\in\mathcal{X}} \exp(-f(\mathbf{x}))$.

2.2 CLAMPING

The basic idea of clamping is as follows: after selecting any variable X_k, we can re-write the partition function as

$$\mathcal{Z}(f) = \sum_{\mathbf{x}} e^{-f(\mathbf{x})} = \sum_{l=1}^{L} \underbrace{\sum_{\mathbf{x}\in\mathcal{X}\colon \mathbf{x}_{k,l}=1} e^{-f(\mathbf{x})}}_{\mathcal{Z}_{k,l}},$$

i.e. $\mathcal{Z}(f)$ can be computed as the sum of the terms $\mathcal{Z}_{k,l}$, which correspond to the partition functions of distributions induced by the energy function f by fixing X_k to value l. To perform approximate inference, we can now approximate each term $\mathcal{Z}_{k,l}$ independently (this corresponds to performing approximate inference in L separate models) and add up these approximations.

In the remainder of the paper we will work with methods which *guarantee an upper bound* on the partition function. Specifically, if we apply these methods directly to $\mathcal{Z}$ we will obtain some upper bound $\widehat{\mathcal{Z}} \geq \mathcal{Z}$, and similarly applying them to the clamped problem will yield an estimate $\widehat{\mathcal{Z}}_{k,l} \geq \mathcal{Z}_{k,l}$. The important question that arises is if the above strategy will always improve the approximation. Specifically, while we do know that $\mathcal{Z} \leq \sum_{l=1}^{L} \widehat{\mathcal{Z}}_{k,l}$, it is in general not clear if

$$\mathcal{Z} \leq \sum_{l=1}^{L} \widehat{\mathcal{Z}}_{k,l} \leq \widehat{\mathcal{Z}},$$

which is exactly the question studied in §3 and §4.

Also, we would like to point out that while we can use $\sum_{l=1}^{L} \widehat{\mathcal{Z}}_{k,l}$ as an approximation to $\mathcal{Z}$, it is not clear how to obtain approximate marginals after clamping. To motivate the strategy we undertake in the experimental section, let us start by noting that

$$P(X_i = j) = \sum_{l=1}^{L} P(X_i = j \mid X_k = l) P(X_k = l).$$

Then, if in the above formula we approximate $P(X_k = l) = \mathcal{Z}_{k,l}/\mathcal{Z}$ by $\widehat{\mathcal{Z}}_{k,l}/\sum_{l=1}^{L}\widehat{\mathcal{Z}}_{k,l}$, as a natural approximation to $P(X_k = l)$ we will use the following quantity

$$\tau_{i,j} = \frac{\sum_{l=1}^{L}\widehat{\mathcal{Z}}_{k,l}\tau_{i,j}^{l}}{\sum_{l=1}^{L}\widehat{\mathcal{Z}}_{k,l}},$$

where $\tau_{i,j}^{l}$ is the approximation of $P(X_i = j)$ obtained from the sub-problem corresponding to $\widehat{\mathcal{Z}}_{k,l}$. These quantities will be exact if all approximations are also exact, and, as we show in §5, they do improve with the number of clamps when approximate inference is used.

2.3 THE INFINITE RÉNYI DIVERGENCE

The infinite Rényi divergence D_∞ between distributions Q and P is defined as [5]

$$D_\infty(P \,\|\, Q) = \sup_{\mathbf{x}\in\mathcal{X}} \log \frac{P(\mathbf{x})}{Q(\mathbf{x})} \geq 0.$$

Note that this is an *inclusive divergence* [20] in the sense that it will try to cover as much as possible from the distribution. In what follows we will focus our attention to completely factorized distributions Q. Specifically, we will assume that Q has the form

$$\log Q(\mathbf{x}) = -\mathbf{s}^T\mathbf{x} - \sum_{i=1}^{N}\log\sum_{j=1}^{L}\exp(-s_{i,j}),$$

for some parameters $\mathbf{s} \in \mathbb{R}^{NL}$. If we plug this Q and $P(\mathbf{x}) = \exp(-f(\mathbf{x}))/\mathcal{Z}$ in the definition above, we arrive at the following upper bound on the partition function

$$\sum_{i=1}^{N}\log\sum_{j=1}^{L}\exp(-s_{i,j}) + \sup_{\mathbf{x}\in\mathcal{X}}(\mathbf{s}^T\mathbf{x} - f(\mathbf{x})) \geq \log\mathcal{Z}.$$

Note that to minimize this (convex) bound we just need to be able to solve the perturbed MAP problem, where the perturbation is equal to $-\mathbf{s}^T\mathbf{x}$. Djolonga and Krause [7] analyze this bound and discuss algorithms for its optimization for the case when f is submodular.

2.4 PERTURB-AND-MAP

Hazan and Jaakkola [14] proposed using Perturb-and-MAP to estimate the partition function. The idea behind this method is to execute the following procedure several times: (i) perturb the energy by adding a random modular term, and (ii) find the MAP configuration under the perturbation. Then, if we repeatedly perform the above steps, we can obtain both an upper bound $\widehat{\mathcal{Z}}_P$ (in expectation) on $\mathcal{Z}$, and an estimate of the marginals (by treating the configurations found in (ii) as if they had come from the true distributions and computing sample averages). While one can have more complicated perturbation models, in this paper we will focus on the simplest perturbation, i.e. those that only modify the unary potentials, which can be easily applied given the optimization assumption we have made about the energy f. Formally, if $\cup_{i=1}^{N}\cup_{j=1}^{L}\{g_{i,j}(1), g_{i,j}(0)\}$ is a collection of i.i.d. Gumbel variables, we have that

$$\log\mathcal{Z} \leq \mathbb{E}_g[\max_{\mathbf{x}\in\mathcal{X}}\{\sum_{i=1}^{n}\sum_{j=1}^{L} g_{i,j}(x_{i,j}) - f(\mathbf{x})\}].$$

3 CLAMPING WITH D_∞

In this section, we will first prove that clamping can only improve the estimate of the partition function. Then, we will discuss the specific cases of binary and multi-label submodular functions. Finally, we will introduce some strategies for choosing the variables to be clamped.

To prove the main claim, let us first rewrite the D_∞ divergence into a form that is easier to manipulate. Remember from §2.3 that the upper bound has the form

$$\log\mathcal{Z} \leq \sum_{i=1}^{n}\log\sum_{j=1}^{L}\exp(-s_{i,j}) + \sup_{\mathbf{x}\in\mathcal{X}}(\mathbf{x}^T\mathbf{s} - f(\mathbf{x})). \tag{2}$$

As done in [7], we can introduce a new variable $-t$ to capture the supremum and re-write the problem as

$$\begin{aligned} \min_{\mathbf{s},t}\quad & \sum_{i=1}^{n}\log\sum_{j=1}^{L}\exp(-s_{i,j}) - t \\ \text{s.t.}\quad & \mathbf{x}^T\mathbf{s} + t \leq f(\mathbf{x}),\ \forall\mathbf{x}\in\mathcal{X}. \end{aligned} \tag{$\texttt{OPT}_\infty$}$$

The constraint set of the above problem, which is also known as the upper polyhedron [21], will be denoted as

$$U(f) = \{(\mathbf{s}, t) \mid \mathbf{x}^T\mathbf{s} + t \leq f(\mathbf{x}),\ \forall\mathbf{x}\in\mathcal{X}\}. \tag{3}$$

Now, if we clamp some variable $X_k = l$, the resulting sub-problem will have the form[1]

$$\begin{aligned} \min_{\mathbf{s},t}\quad & \sum_{i=1,i\neq k}^{n}\log\sum_{j=1}^{L}\exp(-s_{i,j}) - t \\ \text{s.t.}\quad & \forall\mathbf{x}\in\mathcal{X}\colon x_{k,l} = 0 \text{ we have that} \\ & \mathbf{x}^T\mathbf{s} + t \leq f(\mathbf{x} + \mathbf{e}_{k,l}). \end{aligned} \tag{$\texttt{OPT}_\infty^l$}$$

Note that the above optimization problem does not depend the variables in the block $\mathbf{s}_k$, which we have kept to simplify the notation. We will now prove that clamping can only improve the estimate of the partition function.

[1]The vector $\mathbf{e}_{k,l} \in \mathbb{R}^{NL}$ has all coordinates zero, except for the coordinate (k, l), which is equal to one.

Theorem 3.1. *For the D_∞ objective, clamping can only improve the estimate. Specifically, if $\widehat{\mathcal{Z}}$ is the optimal value for ($\texttt{OPT}_\infty$) and $\widehat{\mathcal{Z}}_l$ is the optimal value for ($\texttt{OPT}^l_\infty$) we have $\mathcal{Z} \leq \sum_{l=1}^{L} \widehat{\mathcal{Z}}_l \leq \widehat{\mathcal{Z}}$.*

Proof. The exponential of the objective of problem ($\texttt{OPT}_\infty$) can be rewritten as

$$\prod_{i=1}^{N}(\sum_{l=1}^{L} e^{-s_{i,l}})e^{-t} = \sum_{l=1}^{L} e^{-s_{k,l}-t} \cdot \prod_{i\neq k}(\sum_{j=1}^{L} e^{-s_{i,j}})$$

Then, it trivially follows that

$$\begin{aligned}\widehat{\mathcal{Z}} &= \min_{(\mathbf{s},t)\in U(f)} \sum_{l=1}^{L}(e^{-s_{k,l}-t} \cdot \prod_{i\neq k}(\sum_{j=1}^{L} e^{-s_{i,j}})) \\ &\geq \sum_{l=1}^{L} \underbrace{\min_{(\mathbf{s},t)\in U(f)} (e^{-s_{k,l}-t} \cdot \prod_{i\neq k}(\sum_{j=1}^{L} e^{-s_{i,j}}))}_{A_l}\end{aligned}$$

Let us now bound $\log A_l$, i.e. the optimal value of

$$\min_{(\mathbf{s},t)\in U(f)} -s_{k,l} - t + \sum_{i\neq k} \log(\sum_{j=1}^{L} e^{-s_{i,j}}).$$

We will show that any feasible pair $(\mathbf{s}, t)$ of this problem can be converted into a feasible pair $(\mathbf{s}', t')$ of problem ($\texttt{OPT}^l_\infty$) with the same objective value. This would in turn imply that $\log A_l \geq \log \widehat{\mathcal{Z}}_l$, which completes the proof. Define $t' = t + s_{k,l}$, and $\mathbf{s}'$ to be equal to $\mathbf{s}$ with the exception of $s'_{k,l}$, which is set to zero. It is obvious that plugging in $(\mathbf{s}', t')$ into ($\texttt{OPT}^l_\infty$) will yield the same objective, so we just have to prove that it is feasible. Let $\mathbf{x} \in \mathcal{X}$ have $x_{k,l} = 0$. Define $\mathcal{I} = \{0, 1, \ldots, N-1\} \times \{0, 1, \ldots, L-1\}$. The inequality in $U(f)$ corresponding to $\mathbf{x} + \mathbf{e}_{k,l}$ reads

$$\sum_{(i,j)\in\mathcal{I}-(k,l)} x_{i,j}s_{i,j} + s_{k,l} + t \leq f(\mathbf{x} + \mathbf{e}_{k,l}),$$

which with a little manipulation can be re-written as the inequality corresponding to $\mathbf{x}$ in ($\texttt{OPT}^l_\infty$), namely

$$\underbrace{\sum_{(i,j)\in\mathcal{I}-(k,l)} x_{i,j}s_{i,j} + \underbrace{0}_{s'_{k,l}}}_{\mathbf{x}^T\mathbf{s}'} + \underbrace{s_{k,l} + t}_{t'} \leq f(\mathbf{x} + \mathbf{e}_{k,l}),$$

which implies that $(\mathbf{s}', t')$ is feasible for ($\texttt{OPT}^l_\infty$).

In Appendix A.1 we provide a stronger result. i.e. that $A_l = \widehat{\mathcal{Z}}_l$ for submodular functions f. □

3.1 RELAXATIONS

Unfortunately, it can be in general very hard to work with $U(f)$ for arbitrary sets $\mathcal{X}$. For example, even if f is a submodular function, checking if $(\mathbf{s}, t)$ is a member of $U(f)$ requires minimizing $f(\mathbf{x}) - \mathbf{x}^T\mathbf{s}$ subject to $\mathbf{x} \in \mathcal{X}$, which is NP-hard if $L > 2$, as shown by Dahlhaus et al. [22]. However, it is of course tractable if we use $\overline{\mathcal{X}} = \{0, 1\}^{NL}$, which is exactly the approach taken by Zhang et al. [23]. Note that in this case the upper bound is still valid, because replacing $\mathcal{X}$ by $\overline{\mathcal{X}} \supseteq \mathcal{X}$ in Equation (2) can only increase the bound. We would like to point out that in the previous argument would still hold if we replace $\mathcal{X}$ by $\overline{\mathcal{X}}$. In other words, even if a relaxation is used in lieu of the true 1-of-L constraints, the obtained upper bound can only improve after clamping.

3.2 VARIABLE SELECTION FOR PROBABILISTIC SUBMODULAR MODELS

Let us consider the case when f is a submodular set function. This case has been analyzed in more detail by Djolonga and Krause [7]. In that paper, the authors show that for binary models we can simplify the optimization problem as follows. First, because the model is binary for any position i, we have $x_{i,1} = 1 - x_{i,0}$. Then, we can treat only the variables $\mathbf{x}_{i,1}$, and it can be further shown that ($\texttt{OPT}_\infty$) is equivalent to minimizing the following objective

$$\sum_{i=1}^{N} \log(1 + e^{-x_{i,1}}),$$

over the constraint that requires the concatenation $\mathbf{s}$ of the variables $x_{i,1}$ to be a member of the base-polytope [24]

$$B(f) = \{\mathbf{s} \mid \mathbf{1}^T\mathbf{s} = f(\mathbf{1}), \mathbf{x}^T\mathbf{s} \leq f(\mathbf{x}),\ \forall \mathbf{x} \in \{0, 1\}^n\}. \tag{4}$$

The fact that this object is very well understood (see e.g. [24, 25]) has enabled the development of efficient algorithms for minimizing the D_∞ upper bound. Here, we will also make use of one of the properties of the base polytope to design heuristics for choosing a good clamping order.

We propose two computationally efficient and effective strategies that build on the idea of clamping some random variable X_i whose corresponding optimization variable $s_{i,1}$ can "vary" the most. We quantify this, using the observation that all elements in the base polytope satisfy [25]

$$s_{i,1} \in [f(\mathbf{1}) - f(\mathbf{1} - \mathbf{e}_{i,1}), f(\mathbf{e}_{i,1})]. \tag{5}$$

Actually, the bound is tight in the sense that for both end-points of this interval there exists a vertex whose $(i, 1)$-th coordinate is exactly equal to it. Our experiments

show that this range has a strong correlation with the improvement we can make by clamping variable X_i.

The first heuristic that we propose is NAIVEMAXRANGE, that clamps the top k variables with the largest such intervals. In the experiments, we observe that this simple method outperforms random choice. Moreover, we can adaptively apply this strategy — instead of fixing all k variables to clamp in the beginning, we can first clamp the variable with the largest interval and then recursively apply the same strategy to the resulting sub-problems. We call this strategy BRANCHMAXRANGE. This strategy gave the best experimental results.

The BRANCHMAXRANGE algorithm is shown in Algorithm 1. Its input parameters are the energy function f, the number of variables to clamp k and a set of clampings $\mathcal{C}$, i.e. tuples of the form (i, j) denoting that the i^{th} variable is clamped to value j. For the first call, the set of clampings is empty. The output of the algorithm is the approximate partition function $\widehat{\mathcal{Z}}$ and a vector of approximate marginals $\mathbf{p}$. The algorithm recursively identifies the (non-clamped) variable with the largest interval in line 5. It then recurses by clamping the identified variable to both values it can take in lines 6 and 7. Finally, it aggregates the outputs of the recursive calls to compute the approximate partition function and the approximate marginals in lines 8–12. If all k variables are clamped in a call of BRANCHMAXRANGE, it performs approximate inference in line 2, e.g. by using L-FIELD or Perturb-and-MAP.

Algorithm 1: BRANCHMAXRANGE clamping for D_∞

Input: f, clampings $\mathcal{C}$, number of variables to clamp k
Output: $\widehat{\mathcal{Z}}(f)$, approximate marginals $\mathbf{p}$

1 **if** $|\mathcal{C}| = k$ **then**
2 | **return** *approx_method*$(f, \mathcal{C})$
3 **end**
4 $\mathcal{V} = [N] \setminus \cup_{(i,l)\in\mathcal{C}}\{i\}$ `// free variables`
5 $i = \arg\max_{i'\in\mathcal{V}} f(\mathbf{e}_{i',1}) - [f(\mathbf{1}) - f(\mathbf{1} - \mathbf{e}_{i',1})]$
6 $(\widehat{\mathcal{Z}}_0, \mathbf{p}^0) = \text{BRANCHMAXRANGE}(f, \mathcal{C} \cup \{(i, 0)\}, k)$
7 $(\widehat{\mathcal{Z}}_1, \mathbf{p}^1) = \text{BRANCHMAXRANGE}(f, \mathcal{C} \cup \{(i, 1)\}, k)$
8 $\widehat{\mathcal{Z}} = \widehat{\mathcal{Z}}_0 + \widehat{\mathcal{Z}}_1$
9 $p_i = \frac{\widehat{\mathcal{Z}}_1}{\widehat{\mathcal{Z}}}$
10 **for** $j \in \mathcal{V} \setminus \{i\}$ **do**
11 | $p_j = \frac{1}{\widehat{\mathcal{Z}}}[\widehat{\mathcal{Z}}_0 \cdot p_j^0 + \widehat{\mathcal{Z}}_1 \cdot p_j^1]$
12 **end**
13 **return** $(\widehat{\mathcal{Z}}, \mathbf{p})$

4 CLAMPING WITH PERTURB-AND-MAP

In this section, we will first prove that clamping can only improve the upper bound estimate from Perturb-and-MAP, and then propose an efficient strategy for selecting the variables to be clamped.

Recall that the bound on the partition function arising from this technique has the form

$$\log \mathcal{Z} \leq \log \widehat{\mathcal{Z}} = \mathbb{E}_g[\max_{\mathbf{x}\in\mathcal{X}}\{\sum_{i=1}^{n}\sum_{j=1}^{L} g_{i,j}(x_{i,j}) - f(\mathbf{x})\}],$$

where $\cup_{i=1}^{N}\cup_{j=1}^{L}\{g_{i,j}(1), g_{i,j}(0)\}$ is a collection of i.i.d. Gumbel random variables. Then we have that

$$\begin{aligned}
&\mathbb{E}_g[\max_{\mathbf{x}\in\mathcal{X}}\{\sum_{i,l} g_{i,l}(x_{i,l}) - f(\mathbf{x})\}] \\
=&\mathbb{E}_g[\max_{\mathbf{x}\in\mathcal{X}}\{\sum_{i,l}(g_{i,l}(1) - g_{i,l}(0))\mathbf{x}_{i,l} + g_{i,l}(0)) - f(\mathbf{x})\}] \\
=&\mathbb{E}_{\mathbf{z}}[\max_{\mathbf{x}\in\mathcal{X}}\{\sum_{i,l} z_{i,l}x_{i,l} - f(\mathbf{x})\}],
\end{aligned}$$

where $z_{i,l} = g_{i,l}(1) - g_{i,l}(0)$ is sampled from a logistic distribution. The last equality is by the linearity of expectation, the fact that the Gumbel random variables have a zero mean, and the fact that the difference of two Gumbel variables is a logistic random variable [4]. To simplify notation, we introduce the following set

$$\mathcal{X}_{-i} = \{\Pi_i(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\},$$

where the operation Π_i zeroes out the block $\mathbf{x}_i$. We can then rewrite this bound as

$$\mathbb{E}_{\mathbf{z}}[\max_{l=1,2,\ldots,L}\{\max_{\mathbf{x}\in\mathcal{X}_{-i}}\{\mathbf{z}^T\mathbf{x} - f(\mathbf{x} + \mathbf{e}_{i,l}) + z_{i,l}\}\}].$$

If we clamp $x_{i,l}$ to zero and one, we will obtain the following upper bound on $\log \mathcal{Z}$

$$\begin{aligned}
\log\big[\exp(\underbrace{\mathbb{E}_{\mathbf{z}}[\max_{\mathbf{x}\in\mathcal{X},\mathbf{x}_{i,l}=0}\{\mathbf{z}^T\mathbf{x} - f(\mathbf{x})\}]}_{\widehat{\mathcal{Z}}_-}) \\
+ \exp(\underbrace{\mathbb{E}_{\mathbf{z}}[\max_{\mathbf{x}\in\mathcal{X}_{-i}}\{\mathbf{z}^T\mathbf{x} - f(\mathbf{x} + \mathbf{e}_{i,l})\}]}_{\widehat{\mathcal{Z}}_+})\big].
\end{aligned} \tag{6}$$

Note that both $\widehat{\mathcal{Z}}$ and $\widehat{\mathcal{Z}}_+ + \widehat{\mathcal{Z}}_-$ upper bound the true partition function. The main question is if the second bound dominates the first one, which we show in the following theorem to be indeed the case.

Theorem 4.1. *For multi-label models, if we clamp any arbitrary variable $x_{i,l}$ we have that $\mathcal{Z} \leq \widehat{\mathcal{Z}}_+ + \widehat{\mathcal{Z}}_- \leq \widehat{\mathcal{Z}}$.*

Proof. Before we start with the proof, we would like to explain the strategy that we will use. Instead of showing that clamping X_i helps, we will instead show that

clamping $X_{i,l}$ helps, i.e. if we treat $X_{i,l}$ as a variable itself. Then, clamping X_i is equivalent to L consecutive clamps of the variables $X_{i,1}, X_{i,2}, \ldots, X_{i,L}$. As we will show that each of these clamps can only improve the estimate on the partition function, then it must be true that clamping X_i can only improve the estimate.

Let us denote by $\mathbf{z}_{-i,l}$ the vector which has all coordinates of $\mathbf{z}$ except its i,l-th coordinate. Then, for ease of readability let us define

$$G(\mathbf{z}_{-i,l}) = \max_{\mathbf{x}\in\mathcal{X}, \mathbf{x}_{i,l}=0} \{\mathbf{z}^T\mathbf{x} - f(\mathbf{x})\} - \max_{\mathbf{x}\in\mathcal{X}_{-i}} \{\mathbf{z}^T\mathbf{x} - f(\mathbf{x}+\mathbf{e}_{i,l})\}.$$

If we define $(x)_+ = \max(x, 0)$, we have that

$$\begin{aligned}
\log \frac{\widehat{\mathcal{Z}}_-}{\widehat{\mathcal{Z}}} &= -\mathbb{E}_{\mathbf{z}}[(z_{i,l} - G(\mathbf{z}_{-i,l}))_+] \\
&= -\mathbb{E}_{\mathbf{z}_{-i,l}}[\int_{G(\mathbf{z}_{-i,l})}^{+\infty} p(z_{i,l}) \cdot (z_{i,l} - G(\mathbf{z}_{-i,l}))dz_{i,l}] \\
&= \mathbb{E}_{\mathbf{z}_{-i,l}}[-\log(1 + e^{-G(\mathbf{z}_{-i,l})})] \\
&\leq \log \mathbb{E}_{\mathbf{z}_{-i,l}}[\frac{1}{1 + e^{-G(\mathbf{z}_{-i,l})}}],
\end{aligned}$$

where the last equality is known for logistic distributions (see e.g. [15]), while the inequality is due to Jensen's inequality. Note the RHS does not depend on $z_{i,l}$, hence this is indeed a function of $\mathbf{z}_{-i,l}$. We can analogously prove that

$$\log \frac{\widehat{\mathcal{Z}}_+}{\widehat{\mathcal{Z}}} \leq \log \mathbb{E}_{\mathbf{z}_{-i,l}}[\frac{1}{1 + e^{G(\mathbf{z}_{-i,l})}}].$$

Combining these two inequalities we obtain

$$\begin{aligned}
\frac{\widehat{\mathcal{Z}}_- + \widehat{\mathcal{Z}}_+}{\widehat{\mathcal{Z}}} &\leq \mathbb{E}_{\mathbf{z}_{-i,l}}[\frac{1}{1 + e^{-G(\mathbf{z}_{-i,l})}} + \frac{1}{1 + e^{G(\mathbf{z}_{-i,l})}}] \\
&= 1,
\end{aligned}$$

which we had to show. □

Independently of this paper and very recently, Balog et al. [26] derived this result for the Weibull or Fréchet upper bounds, and this actually implies the same result for Perturb-and-MAP because the limit of the Weibull or Fréchet upper bounds, as their distribution parameter approaches 0, is equal to the Perturb-and-MAP upper bound.

4.1 SPEEDING UP GREEDY CLAMPING

Unfortunately, coming up with an easily computable heuristic for clamping variables for Perturb-and-MAP is non-trivial. One possible idea would be to greedily choose a variable that results in the biggest improvement. Unfortunately, if we use M samples for each clamped subproblem we will need to solve in total $O(NLM)$ MAP estimations, which, even though trivially parallelizable, can be still prohibitively expensive.

We will now propose a simple approximation to the greedy strategy, that can sometimes take significantly less time to run and gave promising results in the experiments we performed in §5. Let us start by introducing a related problem, which is that of computing the following quantities

$$\texttt{MIN-MARG}^*_{i,l} = \max_{\mathbf{x}\in\mathcal{X}:\, x_{i,l}=1} \mathbf{s}^T\mathbf{x} - f(\mathbf{x}),$$

for all i, l and some fixed $\mathbf{s} \in \mathbb{R}^{NL}$, which are also known as *min-marginals*. A large family of models where this optimization can be done more efficiently than doing N separate optimization problems are graph-representable submodular functions, as shown by Kohli and Torr [27]. For these models fixing the value of a variable is equivalent to adding a single edge of infinite capacity to the corresponding node, and the authors have developed an algorithm that can more efficiently compute all min-marginals by re-using some intermediate results.

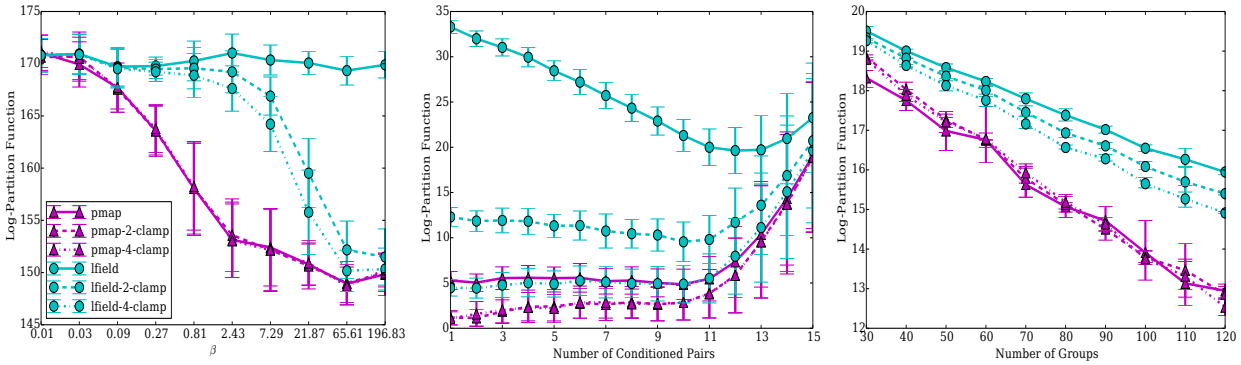
Note that after clamping variable X_k to value l we have to evaluate

$$\mathbb{E}_{\mathbf{s}}[\max_{\mathbf{x}\in\mathcal{X}:\, x_{j,l}=1} \mathbf{s}^T\mathbf{x} - f(\mathbf{x})],$$

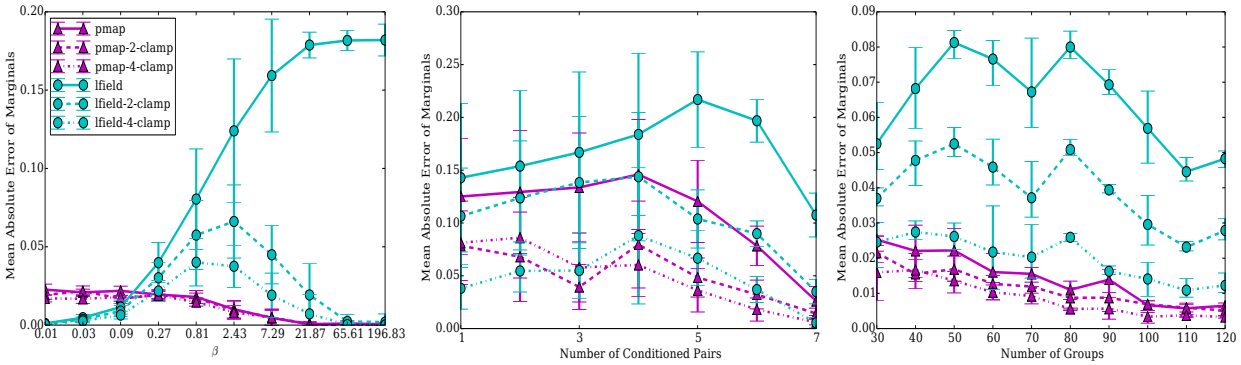
by drawing a sample $\mathcal{S}_l = \{\mathbf{s}_{l,1}, \mathbf{s}_{l,2}, \ldots, \mathbf{s}_{l,M}\}$ and solving the M resulting optimization problems. What we suggest then is to estimate these quantities using the min-marginals by tying the samples, i.e. by drawing a single sample $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M\}$ and setting $\mathcal{S}_l = \mathcal{S}$. Then, we can make use of the faster min-marginal computation from Kohli and Torr [27], but the samples are not independent anymore. However, as substantiated by our experimental results in §5, this approach, which we call Perturb-and-Min-Marginals, works well in practice and performs better than randomly clamping variables.

5 EXPERIMENTS

In this section we want to showcase the following: (1) demonstrate that clamping indeed improves the bounds on the log-partition function, (2) analyze the effect on the estimated marginals, (3) compare the performance of various variable selection strategies. Because in the experiments we focus our attention on submodular models, the minimization of the D_∞ divergence turns into the L-FIELD method of [6], whose code and experimental setup we reuse here. For (1) and (2), we run Perturb-and-MAP (with 200 random samples, labelled `pmap`) and L-FIELD

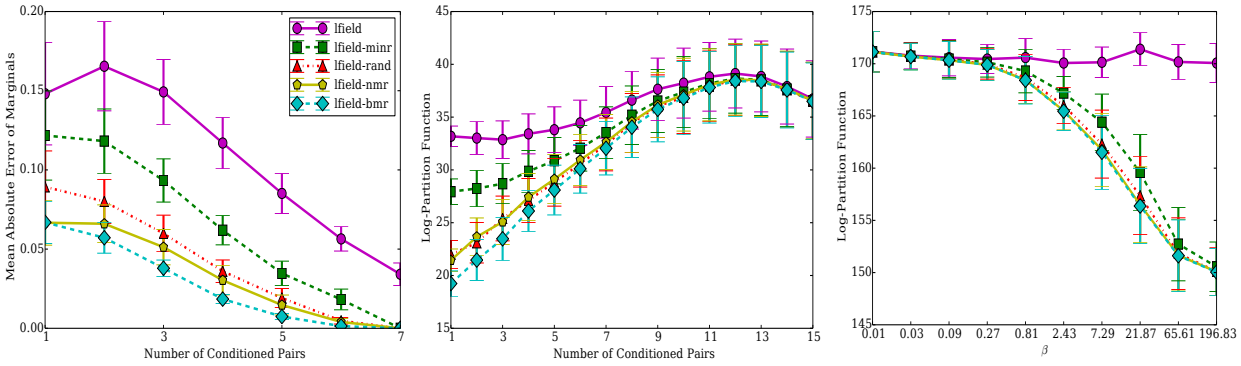


(a) 10×10 Grid cuts (b) Cond. pairs ($n = 50, c = 0.1$) (c) R.c. ($m = 30, \alpha = 0.2, \beta = \frac{1}{30}$)



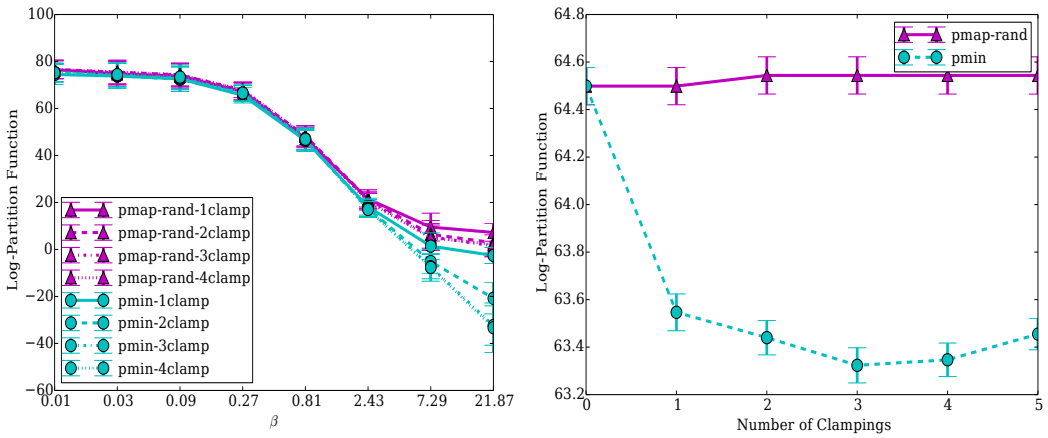
(d) 4×4 Grid cuts (e) Cond. pairs ($n = 20, c = 0.1$) (f) R.c. ($m = 10, \alpha = 0.2, \beta = \frac{1}{10}$)

Figure 1: In the above plots we show the effects on the estimated partition function (first row) and marginals (second row). We can see that clamping improves the estimates on both $\mathcal{Z}$ and the marginals. Further experiments with different parameter settings can be found in Fig. 4 in appendix.



(a) Cond. pairs ($n = 20, c = 1$) (b) Cond. pairs ($n = 50, c = 1$) (c) 10×10 Grid cuts

Figure 2: Comparison of the proposed clamping strategies for L-FIELD. As evident from the plots, `bmr` consistently outperforms the other proposed alternatives.



(a) 10×10 Grid cuts

(b) 10×10 G.c. ($\alpha = 0.05, \beta = 0.01$)

Figure 3: Comparison of the proposed clamping strategies for Perturb-and-MAP. The plots show that `pmin` substantially outperforms `pmap-rand`.

after 2 and 4 clamps. For (3), we test different heuristics for variable selection: `bmr` (BRANCHMAXRANGE), `nmr` (NAIVEMAXRANGE) and `rand` (random selection). We also show that Perturb-and-Min-Marginals (`pmin`) does indeed improve over choosing the variables randomly (`pmap-rand`). Finally, to show that using the interval sizes for L-FIELD does make sense, we also include the strategy that chooses variables with the smallest interval size, denoted by `minr`, which we expect to perform poorly. We used the following models, similar to the setup in [6]. As a mincut solver we used the algorithm by Boykov and Kolmogorov [28].

- *Grid cuts.* The first class of models we experiment on are grid-structured pairwise models, i.e. $P(\mathbf{x}) \propto \exp(-\sum_{\{i,j\}\in E} \beta' [\![\mathbf{x}_i \neq \mathbf{x}_j]\!] - \sum_i z_i)$, where E are the grid-structured edges. We sampled $\beta' \sim \text{Unif}([0, \beta])$ and $z_i \sim \text{Unif}([-1, +1])$, i.e. $P(\mathbf{x})$ is an attractive (log-supermodular) Ising model.

- *Conditioned pairs.* The model has the same functional form as before, but the graph is complete and the edge weights are generated as follows. We first sample two centers from $\mathcal{N}([3, 3], I)$ and $\mathcal{N}([-3, -3], I)$ respectively. Then, around each center we sample n points. These $2n$ points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{2n}\}$ are assigned to the elements, and the weight between elements i and j is set to $e^{-c\|\mathbf{x}_i - \mathbf{x}_j\|}$. Then, for $k = 1, 2, \ldots, K$, we perform inference on the posterior distribution after conditioning that k elements from the first cluster are in A and k elements from the second cluster are not contained in A,

- *Random covers.* Motivated by the P^n potentials from vision [13], we generate models with higher-order potentials as follows. We first sample k vectors $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_k$ of size m from $\{0, 1\}^N$ uniformly at random. Then, we use $f(\mathbf{x}) = \beta \cdot \sum_{i=1}^{k} (\frac{\|\mathbf{x}_i \wedge \mathbf{x}\|_1^\alpha}{\|\mathbf{x}_i\|_1^\alpha}) + \mathbf{z}^T\mathbf{x}$, where $\mathbf{z} \sim \text{Unif}([-1, 1]^N)$, which is submodular for $\alpha \in [0, 1]$ and $\beta \geq 0$. We would like to point out that this is a higher-order model because in the i-th factor a total of $\|\mathbf{x}_i\|_1$ variables participate.

The results for different numberd of clampings are shown in Fig. 1, while the performance of the different heuristics for choosing the clamping order can be seen in Fig. 2 and Fig. 3. We can see that clamping *does* improve the estimate on the partition function, and significantly so for L-FIELD. The marginals are likewise generally improved. We can also see that the proposed `bmr` heuristic outperforms the proposed baselines. Moreover, note that if we use the reverse order (`minr`) we obtain results worse than random, thus providing more evidence towards the hypothesis that the possible improvement is related to the "variability" of the corresponding optimization variable. Furthermore, the Perturb-and-Min-Marginals heuristic outperforms random selection consistently. Finally, as one can see, Perturb-and-MAP often gives better estimate of the partition function compared to L-FIELD, but in practice L-FIELD is typically much faster than Perturb-and-MAP, hence it is interesting to compare clamping of these two methods in terms of runtime. However, we can not expect the performance of L-FIELD after clamping to exceed the performance of Perturb-and-MAP, simply because of the large performance gap between these two methods (and because every additional clamping for L-

FIELD roughly doubles the runtime). Nevertheless, we believe that there are cases where MAP queries are rather expensive such that Perturb-and-MAP is infeasible while L-FIELD still enjoys a relatively low computational complexity.

6 CONCLUSION

Perturb-and-MAP and the minimization of the Rényi infinite divergence are approximate inference techniques whose application depends only on the ability to optimize the energy function under a linear perturbation. Since this class of functions is also closed under clamping, it is a natural question to ask if these techniques can be combined without harming the obtained bounds. In this paper we have answered this question in the affirmative, and moreover provided heuristics for choosing the clamping order. Finally, in a set of experiments we have shown the benefits of clamping for these techniques.

Acknowledgements. The research was partially supported by ERC StG 307036 and a Google European PhD Fellowship. This work was done in part while Andreas Krause was visiting the Simons Institute for the Theory of Computing.

References

[1] M. Jerrum and A. Sinclair. "Polynomial-time approximation algorithms for the Ising model". *SIAM Journal on computing* 22.5 (1993), pp. 1087–1116.

[2] L. A. Goldberg and M. Jerrum. "The complexity of ferromagnetic Ising with local fields". *Combinatorics, Probability and Computing* 16.01 (2007), pp. 43–61.

[3] M. J. Wainwright and M. I. Jordan. "Graphical models, exponential families, and variational inference". *Foundations and Trends in Machine Learning* 1.1–2 (2008), pp. 1–305.

[4] G. Papandreou and A. L. Yuille. "Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models". *ICCV*. 2011.

[5] T. Van Erven and P. Harremos. "Rényi divergence and Kullback-Leibler divergence". *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820.

[6] J. Djolonga and A. Krause. "From MAP to Marginals: Variational Inference in Bayesian Submodular Models". *Neural Information Processing Systems (NIPS)*. 2014.

[7] J. Djolonga and A. Krause. "Scalable Variational Inference in Log-supermodular Models". *International Conference on Machine Learning (ICML)*. 2015.

[8] M. Grötschel, L. Lovász, and A. Schrijver. "The ellipsoid method and its consequences in combinatorial optimization". *Combinatorica* 1.2 (1981), pp. 169–197.

[9] P. Stobbe and A. Krause. "Efficient Minimization of Decomposable Submodular Functions". *NIPS*. 2010.

[10] S. Jegelka, F. Bach, and S. Sra. "Reflection methods for user-friendly submodular optimization". *NIPS*. 2013.

[11] S. Karlin and Y. Rinott. "Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions". *Journal of Multivariate Analysis* 10.4 (1980), pp. 467–498.

[12] Y. Boykov, O. Veksler, and R. Zabih. "Fast approximate energy minimization via graph cuts". *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23.11 (2001), pp. 1222–1239.

[13] P. Kohli, P. H. Torr, et al. "Robust higher order potentials for enforcing label consistency". *International Journal of Computer Vision* 82.3 (2009), pp. 302–324.

[14] T. Hazan and T. S. Jaakkola. "On the Partition Function and Random Maximum A-Posteriori Perturbations". *ICML*. 2012.

[15] T. Shpakova and F. Bach. "Parameter Learning for Log-supermodular Distributions". *arXiv preprint arXiv:1608.05258* (2016).

[16] A. Weller and T. Jebara. "Clamping variables and approximate inference". *Advances in Neural Information Processing Systems*. 2014, pp. 909–917.

[17] A. Weller and J. Domke. "Clamping Improves TRW and Mean Field Approximations". *AISTATS*. 2016.

[18] D. Tarlow, K. Swersky, R. S. Zemel, and R. P. Adams. "Fast Exact Inference for Recursive Cardinality Models". *Proceedings of the Twenty-Eighth Conference Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2012.

[19] V. Vineet, J. Warrell, and P. H. Torr. "Filter-based mean-field inference for random fields with higher-order terms and product label-spaces". *International Journal of Computer Vision* 110.3 (2014), pp. 290–307.

[20] T. Minka. *Divergence measures and message passing*. Tech. rep. Microsoft Research, 2005.

[21] R. Iyer and J. Bilmes. “Polyhedral aspects of Submodularity, Convexity and Concavity”. *arXiv preprint arXiv:1506.07329* (2015).

[22] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour, and M. Yannakakis. “The complexity of multiterminal cuts”. *SIAM Journal on Computing* 23.4 (1994), pp. 864–894.

[23] J. Zhang, J. Djolonga, and A. Krause. “Higher-Order Inference for Multi-class Log-supermodular Models”. *International Conference on Computer Vision (ICCV)*. 2015.

[24] S. Fujishige. *Submodular functions and optimization*. Annals of Discrete Mathematics vol. 58. 2005.

[25] F. Bach. “Learning with submodular functions: a convex optimization perspective”. *Foundations and Trends® in Machine Learning* 6.2-3 (2013).

[26] M. Balog, N. Tripuraneni, Z. Ghahramani, and A. Weller. “Lost Relatives of the Gumbel Trick”. *arXiv preprint arXiv:1706.04161* (2017).

[27] P. Kohli and P. H. Torr. “Measuring uncertainty in graph cut solutions”. *Computer Vision and Image Understanding* 112.1 (2008), pp. 30–38.

[28] Y. Boykov and V. Kolmogorov. “An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision”. *IEEE transactions on pattern analysis and machine intelligence* 26.9 (2004), pp. 1124–1137.