Variable Clamping for Optimization-Based Inference

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 exploit the optimization properties of the distribution to obtain principled inference methods that can work in models of very large orders. In this paper, we first theoretically prove that for binary log-supermodular models the bounds on the partition function obtained by two of these approaches, Perturb-and-MAP and L-FIELD, can be always improved by clamping a subset of the variables. Furthermore, we find that for Perturb-and-MAP the bound also improves for general binary models, which are not necessarily log-supermodular. Moreover, we provide a set of heuristic strategies for choosing the clamping order and present experimental results showcasing the improvements obtained by the proposed methods on several models.

1 Introduction

In this paper we consider models of the form $P(\mathbf{x}) = \frac{1}{\mathcal{Z}(f)} \exp(-f(\mathbf{x}))$, where $\mathbf{x} \in \{0,1\}^n$ is a binary random vector, $f\colon \{0,1\}^n \to \mathbb{R}$ is an arbitrary *energy function*, and $\mathcal{Z}(f)$ is the *partition function*. Unfortunately, computing $\mathcal{Z}$ is known to be #P-hard, even for pairwise models $f(\mathbf{x}) = \mathbf{x}^T W \mathbf{x} + \mathbf{a}^T \mathbf{x}$ [1, 2], where $W \in \mathbb{R}^{n \times n}$ and $\mathbf{a} \in \mathbb{R}^n$. To be in line with the existing literature, in the remaining of this paper we will treat these distributions as equivalently being defined over subsets of $V = \{1, 2, \ldots, n\}$ using the natural bijection $\mathbf{x} \leftrightarrow \{i \mid x_i = 1\} \equiv A \subseteq V$, i.e. we will write them as $P(A) = \frac{1}{\mathcal{Z}(F)} \exp(-F(A))$, where $\mathcal{Z}(F) = \sum_{A \subseteq V} \exp(-F(A))$.

Of special interest are those distributions that allow for the efficient *minimization* of $F(A) + z(A)$ for *any* function $z(A) = \sum_{i \in A} z_i$ for some $z_i \in \mathbb{R}$. The functions of the form $z(A) = \sum_{i \in A} z_i$ are known as modular, and they can be seen as the equivalent of linear functions in the discrete domain—they are uniquely represented by a vector $\mathbf{z} \in \mathbb{R}^n$ and will be treated as both modular functions and vectors. Perhaps the most commonly used family that has the above property is the class of log-supermodular distributions. Their energy has to satisfy $F(A \cup B) + F(A \cap B) \leq F(A) + F(B)$ for all $A, B \subseteq V$. They can be minimized in polynomial time [3], and efficiently so in many cases [4, 5]. The above inequality implies a non-negative correlation between the variables [6], which also explains why they are also sometimes called *attractive*. They have been extensively used in computer vision for semantic image segmentation — both pairwise models (also known as graph-cuts) [7], but also models with complicated higher-order potentials [8].

The problem of variational inference in log-supermodular models has been first addressed by Djolonga and Krause [9], who have developed the L-FIELD variational inference methods. Recently, Shpakova and Bach [10] have drawn an interesting connection between L-FIELD and the Perturb-and-MAP inference method [11] (explained in more detail in §3). The main insight underlying Perturb-and-MAP is that we can both approximate the marginals and obtain an *upper* bound on the partition function by repeatedly optimizing a perturbed version of the energy function, i.e. $f(\mathbf{x}) + \mathbf{z}^T \mathbf{x}$ for randomly chosen functions $\mathbf{z}$. In this paper we focus on the two approximate inference techniques L-FIELD and Perturb-and-MAP and investigate the effect of clamping a subset of the variables on the approximation properties.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

The idea of clamping (i.e, fixing the value) of random variables to improve approximate inference techniques has been studied by Weller and Jebara [12] and Weller and Domke [13]. The basic approach is as follows: given that $\mathcal{Z}(F) = \sum_{A \subseteq V} e^{-F(A)} = \sum_{A \subseteq V \setminus \{i\}} e^{-F(A)} + \sum_{A \subseteq V \setminus \{i\}} e^{-F(A \cup \{i\})}$, we can approximate each term separately and add up these approximations. Specifically, if we define $F_{+i} \colon 2^{V \setminus \{i\}} \to \mathbb{R}$ and $F_{-i} \colon 2^{V \setminus \{i\}} \to \mathbb{R}$ as $F_{+i}(A) = F(A \cup \{i\})$ and $F_{-i}(A) = F(A)$, we have that $\mathcal{Z}(F) = \mathcal{Z}(F_{+i}) + \mathcal{Z}(F_{-i})$. Hence, we have to perform inference in two separate models: F_{+i} where the clamped variable is always set to one (the element is always included), and F_{-i} where the variable is fixed to zero (the element is excluded). The important question that arises is if the above strategy will always improve the approximation. In [12, 13], the authors have answered this question in the affirmative for mean-field, tree-reweighted belief-propagation and traditional belief propagation (for log-supermodular models). Unfortunately, these inference techniques can not be easily used in models with higher order factors without additional assumptions. In this paper, we close the gap, by showing that clamping always improves the estimates of the partition function for Perturb-and-MAP and L-FIELD, methods which weakly depend on the model order.

2 Clamping with L-FIELD

Djolonga and Krause [9] proposed variational methods for optimizing both lower and upper bounds on $\mathcal{Z}(F)$. Their methods depend only on the submodularity of F, and they obtain their bounds by exploiting the differential structure of submodular functions. Because submodular functions are closed under clamping, i.e. F is submodular so are F_{+i} and F_{-i} (see e.g. [14])[1]. we can apply the methods in the aforementioned paper to the subproblems of computing $\mathcal{Z}(F_{+i})$ and $\mathcal{Z}(F_{-i})$. In the following sections, we show that for both bounds clamping can only improve the estimate of $\mathcal{Z}$.

Minimizing the upper bound can be seen as a divergence minimization method [15], as it corresponds to minimizing the Rényi divergence $D_\infty(P \parallel Q) = \log \max_{A \subseteq V} P(A)/Q(A)$ [16] over all factorized distributions $Q(A) = \prod_{i \in A} q_i \prod_{i \notin A} (1 - q_i)$, where $q_i \in [0, 1]$. In the case of log-supermodular distributions, this optimization problem can be equivalently rewritten as [15]

$$\min_{\mathbf{s} \in B(F)} \sum_{i=1}^{n} \log(1 + e^{-s_i}), \text{ where } B(F) = \{\mathbf{s} \in \mathbb{R}^n \mid s(V) = F(V), \forall A \subseteq V : s(A) \leq F(A)\}$$

is known as the *base polytope* of F and has been analyzed extensively in the literature on submodular minimization [17, 14]. The optimal value of the above optimization problem will be denoted by $\widehat{\mathcal{Z}}_U(F)$. Because the optimization problem is smooth and we can efficiently solve linear programs over $B(F)$ in time $O(n \log n)$ [18], we can optimize it using the Frank-Wolfe algorithm [19, 20].

Remember that $\mathcal{Z}(F) = \mathcal{Z}(F_{+i}) + \mathcal{Z}(F_{-i})$. We propose to approximate this quantity with $\widehat{\mathcal{Z}}_U(F_{+i}) + \widehat{\mathcal{Z}}_U(F_{-i})$, which is immediately seen to be an upper bound on $\mathcal{Z}(F)$.

Theorem 2.1. *For log-supermodular models we have* $\mathcal{Z}(F) \leq \widehat{\mathcal{Z}}_U(F_{+i}) + \widehat{\mathcal{Z}}_U(F_{-i}) \leq \widehat{\mathcal{Z}}_U(F)$.

The proof of this theorem is provided in the appendix. In the appendix we further show that the extension [21] of the above bound to multi-label problems has the same property, i.e. clamping can also only improve the estimate of the partition function. Moreover, in [9], the authors also use the properties of submodular functions to obtain lower bounds on $\mathcal{Z}$. Likewise, as proven in the appendix, clamping can only improve the lower bounds.

Variable order selection. Now that we have shown that clamping can only improve the estimate, we can proceed to the question on *which variables* we should clamp. Our strategies are based on the following observation: given that the bound is an optimization problem over $B(F)$, it might make sense to clamp those variables that can "vary" the most. We quantify this, using the observation that all elements in $B(F)$ satisfy

$$s_i \in [F(V) - F(V \setminus \{i\}), F(\{i\})] \tag{1}$$

for the i-th coordinate (see e.g. [17]). Actually, the bound is tight in the sense there are points that take on both ends of the interval. Our experiments show that this range has a strong correlation with the

[1] Typically F_{+i} is defined as $F_{+i}(A) = F(A) - F(\{i\})$ to make sure that F_{+i} is normalized as $F_{+i}(\emptyset) = 0$, but this is of course w.l.o.g.

improvement we can make by clamping variable i. The first heuristic that we propose is `MaxRange`, that proposes for clamping 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`, and it is what gave the best experimental results (explained in more detail as Algorithm 1 in the appendix).

3 Clamping with Perturb-and-MAP

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). Formally, we have that $\log \mathcal{Z} \leq \log \widehat{\mathcal{Z}}_P(F) = \mathbb{E}_{\mathbf{z}}[\max_{A \subseteq V} z(A) - F(A)]$, where each coordinate of $\mathbf{z}$ is sampled independently from a logistic distribution [22]. Then,

$$\begin{aligned}\widehat{\mathcal{Z}}_P(F) &= \exp(\mathbb{E}_{\mathbf{z}}[\max_{A \in V}\{z(A) - f(A)\}]) \\ &= \exp(\mathbb{E}_{\mathbf{z}}[\max\{\max_{A \in V\setminus\{i\}}\{z_{-i}(A) - f(A)\}, \max_{A \in V\setminus\{i\}}\{z_{-i}(A) + z_i - f(A \cup \{i\})\}\}]) \\ &= \exp(\mathbb{E}_{\mathbf{z}}[\max\{\max_{A \in V\setminus\{i\}}\{z_{-i}(A) - f(A)\}, z_i + \max_{A \in V\setminus\{i\}}\{z_{-i}(A) - f(A \cup \{i\})\}\}]).\end{aligned} \tag{2}$$

If we clamp each variable separately, we will obtain the following estimate

$$\begin{aligned}\mathcal{Z}(F) &= \mathcal{Z}(F_{+i}) + \mathcal{Z}(F_{+i}) \leq \widehat{\mathcal{Z}}_P(F_{+i}) + \widehat{\mathcal{Z}}_P(F_{+i}) \\ &= \exp(\mathbb{E}_{\mathbf{z}_{-i}}[\max_{A \in V\setminus\{i\}}\{z_{-i}(A) - f(A)\}]) + \exp(\mathbb{E}_{\mathbf{z}_{-i}}[\max_{A \in V\setminus\{i\}}\{z_{-i}(A) - f(A \cup \{i\})\}])\end{aligned} \tag{3}$$

where $\mathbf{z}_{-i}$ denotes the restricted modular function over the ground set $V \setminus \{i\}$. The following theorem, which holds **without the assumption that F is submodular**, shows that in expectation we will always obtain a stronger bound.

Theorem 3.1. *For binary models we have that $\mathcal{Z}(F) \leq \widehat{\mathcal{Z}}_P(F_{+i}) + \widehat{\mathcal{Z}}_P(F_{-i}) \leq \widehat{\mathcal{Z}}_P(F)$.*

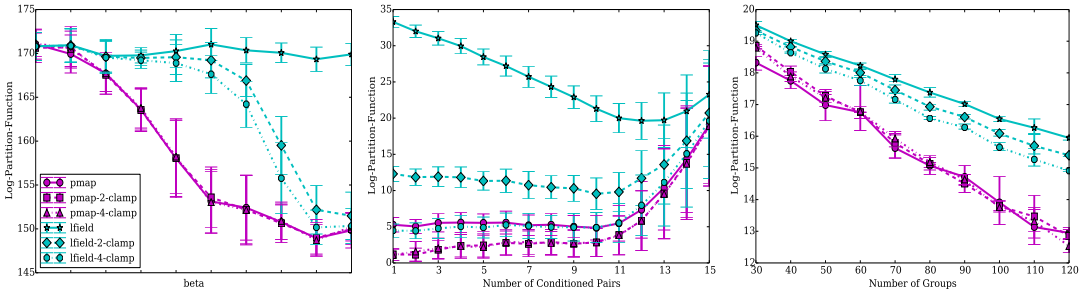
4 Experiments

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 for L-FIELD. For (1) and (2), we run Perturb-and-MAP (with 200 random samples, labelled `pmap`) and L-FIELD after 2 and 4 clamps. For (3), we test different heuristics for variable selection: `bmr` (`BranchMaxRange`), `nmr` (`NaiveMaxRange`), `rand` (random selection). Finally, to show that the strategy based on the intervals 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.

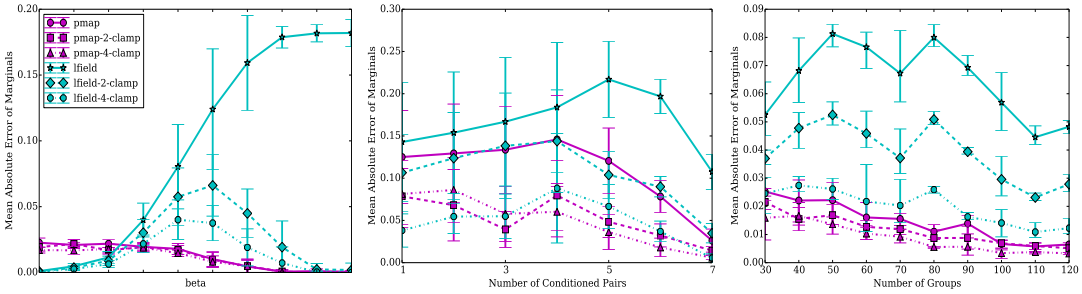
– *Grid cuts.* The first class of models we experiment on are grid-structured pairwise models, i.e. $P(A) \propto \exp(-\sum_{\{i,j\} \in E} \beta[\![A \cap \{i,j\} = 1]\!] - \sum_i z_i)$, where E are the grid-structured edges. We sampled $\beta \sim \text{Unif}([0, \beta])$ and $z_i \sim \text{Unif}([-1, +1])$, i.e. $P(A)$ is an attractive 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, 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 [8], we generate models with higher-order potentials as follows. We first choose k subsets $G_1, G_2 \ldots, G_k$ of size m in V at random. Then,

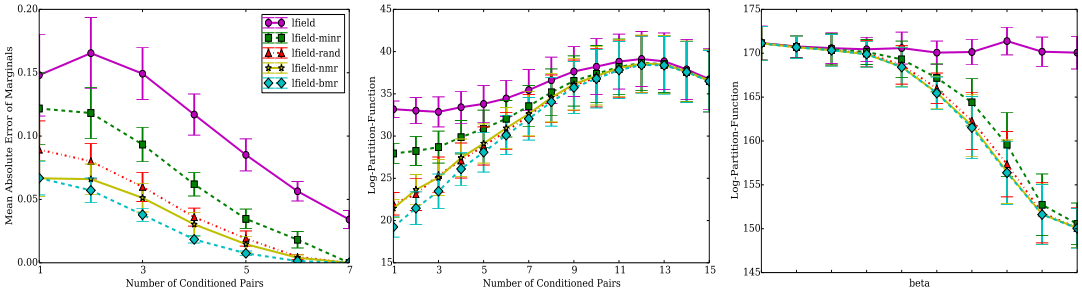


(a) 10×10 Grid cuts (b) Cond. pairs ($n = 50, c = 0.1$) (c) R.c. ($m = 30, \alpha = 0.2, \beta = \frac{1}{30}$)



(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. 3 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.

we use $F(A) = \beta \cdot \sum_{i=1}^{k} (\frac{|A \cap G_i|^{\alpha}}{|G_i|^{\alpha}}) + \sum_{i \in A} z_i$, where $z_i \sim \text{Unif}([-1, 1])$, which is submodular for $\alpha \in [0, 1]$ and $\beta \geq 0$.

The results from different number of clampings are shown in Fig. 1, while the performance of the different heuristics for choosing the order can be seen in Fig. 2. We can see that clamping *does* improve the estimate on the partition faction, 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.

5 Conclusion

We have shown that by clamping variables we can improve the Perturb-and-MAP and L-FIELD approximate inference techniques — both in theory and in a set of experiments.

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

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

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

[6] 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.

[7] 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.

[8] 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.

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

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

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

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

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

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

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

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

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

[18] J. Edmonds. "Submodular functions, matroids, and certain polyhedra". *Combinatorial structures and their applications* (1970), pp. 69–87.

[19] M. Frank and P. Wolfe. "An algorithm for quadratic programming". *Naval Res. Logist. Quart.* (1956).

[20] M. Jaggi. "Revisiting Frank-Wolfe: Projection-free sparse convex optimization". *ICML*. 2013.

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

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

6 Appendix

Algorithm 1: Branch-Max-Range

Input : f, V, number of clamping t
Output : $\widehat{\mathcal{Z}}(f)$, approximate marginals p
if $t == 0$ **then**
  return *approx_method*(f, V);
end
$i = \arg\max_{j \in V} f(\{j\}) - (f(V) - f(V \setminus \{j\}))$;
$(\widehat{\mathcal{Z}}(f_{+i}), p_+) =$ Branch-Max-Range$(f_{+i}, V \setminus \{i\}, t-1)$;
$(\widehat{\mathcal{Z}}(f_{-i}), p_-) =$ Branch-Max-Range$(f_{-i}, V \setminus \{i\}, t-1)$;
$\widehat{\mathcal{Z}}(f) = \widehat{\mathcal{Z}}(f_{+i}) + \widehat{\mathcal{Z}}(f_{-i})$;
$p_i = \frac{\widehat{\mathcal{Z}}(f_{+i})}{\widehat{\mathcal{Z}}(f)}$;
for $j \in V \setminus \{i\}$ **do**
  $p_j = \frac{\widehat{\mathcal{Z}}(f_{+i}) \cdot p_+ + \widehat{\mathcal{Z}}(f_{-i}) \cdot p_-}{\widehat{\mathcal{Z}}(f)}$;
end
return $(\widehat{\mathcal{Z}}(f), p)$;

In this appendix, we present the proof for all the main results in this paper. From now on, we notate two operations that preserve submodularity, (a) **contraction**: $F_X(A) = F(X \cup A) - F(X), A \subseteq V \setminus X$, (b) **restriction**: $F^X(A) = F(A), A \subseteq X$.

6.1 L-FIELD with Multi-label models

Instead of showing the proof for the binary case, we directly prove more general results for multi-label models, and the theorem for the binary models immediately follows. We show how to handle the general case where each variable can take one of L different values $\{1, 2, \cdots, L\}$. We will represent each random variable X_i by L distinct elements $V_i = \{v_{i,1}, v_{i,2}, \cdots, v_{i,L}\}$ corresponding to the values it can take. The idea is that if for example $v_{i,5}$ is chosen, then this corresponds to X_i taking on the value 5. This is also known as the 1-of-L encoding. To make sure that the variable can take only a single value, we will add a constraint M_i that forces the distribution to assign non-zero mass only to those configurations that select exactly one element from V_i. Formally stated, $M_i = \{A : |A \cap V_i| = 1\}$ and our final constraint is $M = \cap_{i=1}^N M_i$. The partition function is $\mathcal{Z}(F) = \sum_{A \in M} \exp(-F(A))$, and as before, we denote by $\widehat{\mathcal{Z}}_U(F)$ its upper bound. According to Formula (2) in [21], $\log \widehat{\mathcal{Z}}_U(F) = \min_{s \in B(F)} \sum_{i=1}^N \log \sum_{j=1}^L \exp(-s_{i,j})$. Define $M_{-k} = \cap_{i=1, i \neq k}^N M_i$ and $V_{-k} = \cup_{i=1, i \neq k}^N V_i$.

$$\mathcal{Z}(F) = \sum_{A \in M} \exp(-F(A)) = \sum_{l=1}^{L} \sum_{A \in M, v_{k,l} \in A} \exp(-F(A)) = \sum_{l=1}^{L} \underbrace{\sum_{A \in V_{-k} \cap M_{-k}} \exp(-F(A \cup \{v_{k,l}\}))}_{\mathcal{Z}(F_{+v_{k,l}})} \tag{4}$$

We can calculate the upper bound of $\mathcal{Z}(F_{+v_{k,l}})$, denoted by $\widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$, using the L-FIELD method, namely $\log \widehat{\mathcal{Z}}_U(F_{+v_{k,l}}) = \min_{s \in B(F_{\{v_{k,l}\}})} \sum_{i=1}^N \log \sum_{j=1}^L \exp(-s_{i,j}) - F(\{v_{k,l}\})$. Obviously $\mathcal{Z}(F) = \sum_l \mathcal{Z}(F_{+v_{k,l}}) \leq \sum_l \widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$. Hence $\sum_l \widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$ is a valid upper bound for the partition function, and this is exactly the new upper bound we want to use for $\mathcal{Z}(F)$ after clamping variable k. The following theorem shows that this is a better upper bound than $\widehat{\mathcal{Z}}_U(F)$.

Theorem 6.1. *For a multi-label log-supermodular model, after clamping arbitrary variable k, $\sum_l \widehat{\mathcal{Z}}_U(F_{+v_{k,l}}) \leq \widehat{\mathcal{Z}}_U(F)$.*

To see this, we decompose the objective used for computing $\widehat{\mathcal{Z}}_U(F)$, which is just the exponential of the objective for $\log \widehat{\mathcal{Z}}_U(F)$ as follows:

$$\exp(\sum_{i=1}^{N} \log \sum_{l=1}^{L} e^{-s_{i,l}}) = \prod_{i=1}^{N} (\sum_{l=1}^{L} e^{-s_{i,l}}) = \sum_{l=1}^{L} (e^{-s_{k,l}} \cdot \prod_{i \neq k} (\sum_{j=1}^{L} e^{-s_{i,j}})) \tag{5}$$

Define $\widehat{\mathcal{Z}}_U^l(F) = \min_{s \in B(F)} e^{-s_{k,l}} \cdot \prod_{i \neq k} (\sum_{j=1}^{L} e^{-s_{i,j}})$, then we know

$$\begin{aligned} \widehat{\mathcal{Z}}_U(F) &= \min_{s \in B(F)} \sum_{l=1}^{L} (e^{-s_{k,l}} \cdot \prod_{i \neq k} (\sum_{j=1}^{L} e^{-s_{i,j}})) \\ &\geq \sum_{l=1}^{L} \min_{s \in B(F)} (e^{-s_{k,l}} \cdot \prod_{i \neq k} (\sum_{j=1}^{L} e^{-s_{i,j}})) \\ &= \sum_{l} \widehat{\mathcal{Z}}_U^l(F) \end{aligned} \tag{6}$$

We will prove a stronger result, namely $\forall l, \widehat{\mathcal{Z}}_U^l(F) = \widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$, and hence $\widehat{\mathcal{Z}}_U(F) \geq \sum_l \widehat{\mathcal{Z}}_U^l(F) = \sum_l \widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$.

Lemma 6.2. $\widehat{\mathcal{Z}}_U^l(F) = \widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$.

Proof. This is equivalent to proving that $\log \widehat{\mathcal{Z}}_U^l(F) = \log \widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$. Later we will show that $\log \widehat{\mathcal{Z}}_U^l(F)$, the minimum of $-s_{k,l} + \sum_{i \neq k} \log \sum_{j=1}^{L} (1 + e^{-s_{i,j}})$ in $B(F)$, can still be achieved if we fix $s_{k,l} = F(\{v_{k,l}\})$. We assume this is true, hence we can replace $s_{k,l}$ with $F(\{v_{k,l}\})$ in $B(F)$ and get the following explicit form.

$$\begin{cases} \sum\limits_{v_{i,j} \in A} s_{i,j} + F(\{v_{k,l}\}) \leq F(A \cup \{v_{k,l}\}), & \forall A \subset V \setminus \{v_{k,l}\} \\ \sum\limits_{v_{i,j} \in A} s_{i,j} \leq F(A), & \forall A \subseteq V \setminus \{v_{k,l}\} \\ \sum\limits_{v_{i,j} \in V \setminus \{v_{k,l}\}} s_{i,j} + F(\{v_{k,l}\}) = F(V) \end{cases}$$

Notice that the second constraint is redundant, because the first inequality requires $\forall A \subset V \setminus \{v_{k,l}\}, \sum_{v_{i,j} \in A} s_{i,j} \leq F(A \cup \{v_{k,l}\}) - F(\{v_{k,l}\})$ and $F(A \cup \{v_{k,l}\}) - F(\{v_{k,l}\}) \leq F(A)$ by submodularity, and for the same reason the last equality fulfills the second inequality when $A = V \setminus \{v_{k,l}\}$. Thus we can remove the second constraint in above inequality system.
Now we write the explicit form of $B(F_{\{v_{k,l}\}})$ as follows.

$$\begin{cases} \sum\limits_{v_{i,j} \in A} s_{i,j} \leq F(A \cup \{v_{k,l}\}) - F(\{v_{k,l}\}), & \forall A \subset V \setminus \{v_{k,l}\} \\ \sum\limits_{v_{i,j} \in V \setminus \{v_{k,l}\}} s_{i,j} = F(V) - F(\{v_{k,l}\}) \end{cases}$$

Observe that this is the same as $B(F)$ when $s_i = F(\{i\})$. Hence the feasible regions of two minimization problem are exactly the same. Furthermore, since we fix $s_{k,l} = F(\{v_{k,l}\})$, the objective of $\log \widehat{\mathcal{Z}}_U^l(F)$ changes into $-F(\{v_{k,l}\}) + \sum_{i \neq k} \log \sum_{j=1}^{L} (1 + e^{-s_{i,j}})$, which is again the same as the objective of $\widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$. Therefore $\log \widehat{\mathcal{Z}}_U^l(F) = \log \widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$, which implies $\widehat{\mathcal{Z}}_U^l(F) = \widehat{\mathcal{Z}}_U(F_{+v_{k,l}})$. □

Lemma 6.3. *By adding $s_{k,l} = F(\{v_{k,l}\})$ to the constraint set, the result of the optimization problem $\min_{s \in B(F)} -s_{k,l} + \sum_{i \neq k} \log \sum_{j=1}^{L} (1 + e^{-s_{i,j}})$ will not change.*

Proof. First we define $g(s) = \sum_{i \neq k} \log \sum_{j=1}^{L} (1 + e^{-s_{i,j}}) - s_{k,l}$. Then we have

$$\begin{cases} \frac{\partial g}{\partial s_{k,l}} = -1 \\ \frac{\partial g}{\partial s_{i,j}} = \frac{-\exp(-s_{i,j})}{\sum_{j'=1}^{L} (1 + \exp(-s_{i,j'}))} > -1, \forall v_{i,j} \in V \setminus \{v_{k,l}\} \end{cases} \tag{7}$$

Hence it is easy to see exchange $\Delta \geq 0$ between $s_{i,j}$ and $s_{k,l}$, i.e. $s'_{k,l} = s_{k,l} + \Delta, s'_{i,j} = s_{i,j} - \Delta$, can only decrease the objective. Therefore, given optimal solution s^*, we can get a solution at least as good as s^* by setting $s'_{k,l} = s^*_{k,l} + \Delta$ and $s'_{i,j} = s^*_{i,j} - \Delta$ for arbitrary $(i,j) \neq (k,l)$. We can exploit this property to change $s^*_{k,l}$ into $F(\{v_{k,l}\})$, but we need to guarantee that every exchange results in a feasible solution. Hence we need to deal with exchange capacity $\hat{c}(s; v_{k,l}, e') = \min\{F(A) - s(A), \forall A \supseteq \{v_{k,l}\}, e' \notin A\}$($e'$ denotes the element to exchange with $v_{k,l}$). Let $S_{e'} \cup \{v_{k,l}\}$ be the set that achieves $\hat{c}(s; v_{k,l}, e')$, we know $e' \notin S_{e'} \cup \{v_{k,l}\}$. We propose the following procedure *exchange*, and we will prove later this algorithm will make $s'_{k,l} = F(\{v_{k,l}\})$. Since we already proved that exchange always results in better solution, this will finish the proof.

procedure *exchange*():
Initiate $U = V \setminus \{v_{k,l}\}, s = s^*$;
While $U \neq \emptyset$:
    Arbitrarily pick $e' \in U$;
    $s_{k,l} \leftarrow s_{k,l} + \hat{c}(s; v_{k,l}, e')$;
    $s_{e'} \leftarrow s_{e'} - \hat{c}(s; v_{k,l}, e')$;
    $U = U \cap S_{e'}$
end;

We first show that after one exchange with e' the new modular function s' is tight at $S_{e'} \cup \{v_{k,l}\}$.

$$\begin{aligned} s'_{k,l} &= s_{k,l} + \hat{c}(s; v_{k,l}, e') \\ &= s_{k,l} + F(S_{e'} \cup \{v_{k,l}\}) - s(S_{e'} \cup \{v_{k,l}\}) \\ &= s_{k,l} + F(S_{e'} \cup \{v_{k,l}\}) - s(S_{e'}) - s_{k,l} \\ &= F(S_{e'} \cup \{v_{k,l}\}) - s'(S_{e'}) \\ \Rightarrow & s'(S_{e'} \cup \{v_{k,l}\}) = F(S_{e'} \cup \{v_{k,l}\}) \end{aligned} \tag{8}$$

Because s' is tight at $S_{e'} \cup \{v_{k,l}\}$, the element picked next round must be the element in $S_{e'}$ such that the next exchange also results in a feasible solution, otherwise the next exchange will break the tight upper bound for s' at $S_{e'} \cup \{v_{k,l}\}$ since we only increase $s'_{k,l}$. This is why we let $U = U \cap S_{e'}$ in the algorithm. It is also obvious that once s' is tight at $S_{e'} \cup \{v_{k,l}\}$, it will always be tight at $S_{e'} \cup \{v_{k,l}\}$. Moreover, notice that $e' \notin S_{e'}$ but $e' \in U$, hence the intersection operation always strictly decreases the size of U in each round. Therefore, algorithm will terminate and U will definitely turn into $\emptyset$. The final U is $\cap_{e'} S_{e'}$, hence $\cap_{e'}(S_{e'} \cup \{v_{k,l}\}) = (\cap_{e'} S_{e'}) \cup \{v_{k,l}\} = \{v_{k,l}\}$. Since the final s' is tight at each $S_{e'} \cup \{v_{k,l}\}$ and it is well-known result that the intersection of tight sets is also tight. Therefore the final s' is tight at $\{v_{k,l}\}$, i.e. $s'_{k,l} = F(\{v_{k,l}\})$, which completes the proof. □

6.2 Clamping Improves the Lower Bound in Binary Models

From the proof of Lemma 4 in [9] we know that the lower bound of log partition function we get from optimizing over bar supergradient is $\log \widehat{\mathcal{Z}}_L(F) = \max_{X \in V} -F(X) + \sum_{j \in X} \log(1 + e^{F(V) - F(V \setminus \{j\})}) + \sum_{j \notin X} \log(1 + e^{-F(\{j\})})$. After clamping i, we also apply this method to get the lower bound for $\mathcal{Z}(F_{\{i\}})$ and $\mathcal{Z}(F^{V \setminus \{i\}})$, denote them by $\widehat{\mathcal{Z}}_L(F_{\{i\}})$ and $\widehat{\mathcal{Z}}_L(F^{V \setminus \{i\}})$ respectively. It is obvious that $\mathcal{Z}(f) = e^{-F(\{i\})} \cdot \mathcal{Z}(F_{\{i\}}) + \mathcal{Z}(F^{V \setminus \{i\}}) \geq e^{-F(\{i\})} \cdot \widehat{\mathcal{Z}}_L(F_{\{i\}}) + \widehat{\mathcal{Z}}_L(F^{V \setminus \{i\}})$, so if $\widehat{\mathcal{Z}}_L(F) \leq e^{-F(\{i\})} \cdot \widehat{\mathcal{Z}}_L(F_{\{i\}}) + \widehat{\mathcal{Z}}_L(F^{V \setminus \{i\}})$, then $e^{-F(\{i\})} \cdot \widehat{\mathcal{Z}}_L(F_{\{i\}}) + \widehat{\mathcal{Z}}_L(F^{V \setminus \{i\}})$ is a better lower bound.

Theorem 6.4. $\widehat{\mathcal{Z}}_L(F) \leq e^{-F(\{i\})} \cdot \widehat{\mathcal{Z}}_L(F_{\{i\}}) + \widehat{\mathcal{Z}}_L(F^{V \setminus \{i\}})$.

Proof. We take the exponent of $\log \widehat{\mathcal{Z}}_L(F)$, then $\widehat{\mathcal{Z}}_L(F) = \max_{X \in V} e^{-F(X)} \prod_{j \in X}(1 + e^{F(V) - F(V \setminus \{j\})}) \prod_{j \notin X}(1 + e^{-F(\{j\})})$. We split it into two cases. First, if $i \in X^*$, where X^*

is the optimal element set for bar supergradient, we know

$$\begin{aligned}\widehat{\mathcal{Z}}_L(F) &= \max_{X\in V} e^{-F(X)} \prod_{j\in X\setminus\{i\}} (1+e^{F(V)-F(V\setminus\{j\})})(1+e^{F(V)-F(V\setminus\{i\})}) \prod_{j\notin X}(1+e^{-F(\{j\})}) \\ &= \max_{X\in V} \underbrace{e^{-F(X)} \prod_{j\in X\setminus\{i\}} (1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X}(1+e^{-F(\{j\})})}_{A_1} \\ &\quad + \underbrace{e^{F(V)-F(V\setminus\{i\})-F(X)} \prod_{j\in X\setminus\{i\}} (1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X}(1+e^{-F(\{j\})})}_{B_1}\end{aligned} \tag{9}$$

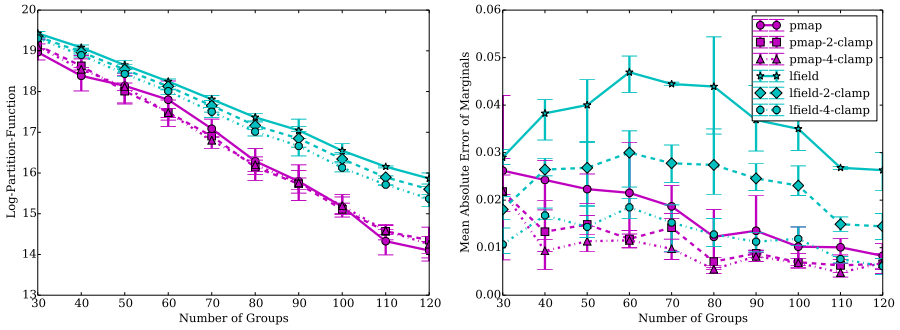
Otherwise, if $i \notin X^*$

$$\begin{aligned}\widehat{\mathcal{Z}}_L(F) &= \max_{X\in V\setminus\{i\}} e^{-F(X)} \prod_{j\in X} (1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X\cup\{i\}}(1+e^{-F(\{j\})})(1+e^{-F(\{i\})}) \\ &= \max_{X\in V\setminus\{i\}} \underbrace{e^{-F(X)} \prod_{j\in X} (1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X\cup\{i\}}(1+e^{-F(\{j\})})}_{A_2} \\ &\quad + \underbrace{e^{-F(X)-F(\{i\})} \prod_{j\in X} (1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X\cup\{i\}}(1+e^{-F(\{j\})})}_{B_2}\end{aligned} \tag{10}$$

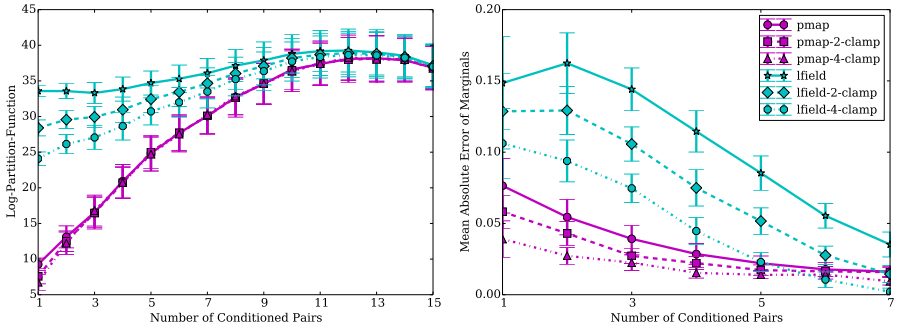
Since $\widehat{\mathcal{Z}}_L(F_{\{i\}}) = \max_{X\in V\setminus\{i\}} e^{F(\{i\})-F(X\cup\{i\})} \prod_{j\in X}(1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X}(1+e^{-F(\{i,j\})+F(\{i\})})$ and $\widehat{\mathcal{Z}}_L(F^{V\setminus\{i\}}) = \max_{X\in V\setminus\{i\}} e^{-F(X)} \prod_{j\in X}(1+e^{F(V\setminus\{i\})-F(V\setminus\{i,j\})}) \prod_{j\notin X}(1+e^{-F(\{j\})})$, we explicitly write the lower bound after clamping as follows.

$$\begin{aligned}&e^{-F(\{i\})}\widehat{\mathcal{Z}}_L(F_{\{i\}}) + \widehat{\mathcal{Z}}_L(F^{V\setminus\{i\}}) \\ &= \max_{X\in V\setminus\{i\}} \underbrace{e^{-F(X\cup\{i\})} \prod_{j\in X}(1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X}(1+e^{F(\{i\})-F(\{i,j\})})}_{A} \\ &\quad + \max_{X\in V\setminus\{i\}} \underbrace{e^{-F(X)} \prod_{j\in X}(1+e^{F(V\setminus\{i\})-F(V\setminus\{i,j\})}) \prod_{j\notin X}(1+e^{-F(\{j\})})}_{B}\end{aligned} \tag{11}$$

We claim that if $i \in X^*$, $A \geq A_1, B \geq B_1$, hence $A+B \geq A_1+B_1$, and if $i \notin X^*$, $B \geq A_2, A \geq B_2$, hence $A+B \geq A_2+B_2$. If this is true, then the expected result follows.
Let $X = X^*\setminus\{i\}$ when $i \in X^*$, then $A_1 = e^{-F(X\cup\{i\})} \prod_{j\in X}(1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X\cup\{i\}}(1+e^{-F(\{j\})})$. We compare A_1 with $A = e^{-F(X\cup\{i\})} \prod_{j\in X}(1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X}(1+e^{F(\{i\})-F(\{i,j\})})$. Since $F(\{i\})-F(\{i,j\}) \geq -F(\{j\})$ by diminishing return, it is easy to see $A \geq A_1$. On the other hand, $B_1 = e^{F(V)-F(V\setminus\{i\})-F(X\cup\{i\})} \prod_{j\in X}(1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X\cup\{i\}}(1+e^{-F(\{j\})})$. We compare this with $B = e^{-F(X)} \prod_{j\in X}(1+e^{F(V\setminus\{i\})-F(V\setminus\{i,j\})}) \prod_{j\notin X}(1+e^{-F(\{j\})})$. Since $F(V)-F(V\setminus\{i\})-F(X\cup\{i\}) \leq -F(X)$ and $F(V)-F(V\setminus\{j\}) \leq F(V\setminus\{i\})-F(V\setminus\{i,j\})$ by diminishing return, it follows that $B \geq B_1$.
Let $X = X^*$ when $i \notin X^*$, then $B_2 = e^{-F(X)-F(\{i\})} \prod_{j\in X}(1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X\cup\{i\}}(1+e^{-F(\{j\})})$. We compare this with A. Since $-F(X)-F(\{i\}) \leq -F(X\cup\{i\})$ and $-F(\{j\}) \leq F(\{i\})-F(\{i,j\})$, $A \geq B_2$ follows. Moreover, $A_2 = e^{-F(X)} \prod_{j\in X}(1+e^{F(V)-F(V\setminus\{j\})}) \prod_{j\notin X\cup\{i\}}(1+e^{-F(\{j\})})$, hence $B \geq A_2$ follows because $F(V)-F(V\setminus\{j\}) \leq F(V\setminus\{i\})-F(V\setminus\{i,j\})$. This completes the proof. □



(a) random covers ($m = 30, \alpha = 0.5, \beta = \frac{1}{30}$) (b) random covers ($m = 10, \alpha = 0.5, \beta = \frac{1}{10}$)



(c) conditioned pairs ($n = 50, c = 1$) (d) conditioned pairs ($n = 20, c = 1$)

Figure 3: Additional experiments on random covers and conditioned pairs with different parameters. Still we can see that clamping improves the estimates on both $\mathcal{Z}$ and the marginals.

6.3 Proof of Theorem 3.1

Proof.

$$
\begin{aligned}
\frac{\widehat{\mathcal{Z}}_P(F_{+i}) + \widehat{\mathcal{Z}}_P(F_{-i})}{\widehat{\mathcal{Z}}_P(F)} &= \exp(\mathbb{E}_z[(\max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A)\} - (\max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A\cup\{i\})\} + z_i)_-]) + \\
&\exp(\mathbb{E}_z[(\max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A\cup\{i\})\} + z_i - \max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A)\})_-]) \\
&= \exp(-\mathbb{E}_z[(\max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A\cup\{i\})\} - \max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A)\} + z_i)_+]) + \\
&\exp(\mathbb{E}_z[(\max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A\cup\{i\})\} - \max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A)\} + z_i)_-]) \\
&= \exp(-\mathbb{E}_z[(z_i - G(z_{-i}))_+]) + \exp(\mathbb{E}_z[(z_i - G(z_{-i}))_-]) \\
&= \exp(-\mathbb{E}_{z_{-i}}[\int_{G(z_{-i})}^{+\infty} p(z_i)\cdot(z_i - G(z_{-i}))dz_i]) + \exp(\mathbb{E}_{z_{-i}}[\int_{-\infty}^{G(z_{-i})} p(z_i)\cdot(z_i - G(z_{-i}))dz_i]) \\
&= \exp(\mathbb{E}_{z_{-i}}[-\log(1 + e^{-G(z_{-i})})]) + \exp(\mathbb{E}_{z_{-i}}[-\log(1 + e^{G(z_{-i})})]) \\
&\leq \mathbb{E}_{z_{-i}}[\frac{1}{1 + e^{-G(z_{-i})}}] + \mathbb{E}_{z_{-i}}[\frac{1}{1 + e^{G(z_{-i})}}] \text{ by Jensen's inequality,} \\
&= \mathbb{E}_{z_{-i}}[\frac{1}{1 + e^{-G(z_{-i})}} + \frac{1}{1 + e^{G(z_{-i})}}] \text{ by linearity of expectation,} \\
&= 1,
\end{aligned}
\tag{12}
$$

where we define $G(z_{-i}) = \max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A)\} - \max_{A\in V\setminus\{i\}}\{z_{-i}(A) - f(A\cup\{i\})\}$ for ease of readability. □