Approximation Guarantees for Sherali Adams Relaxations

Author 1
Institution 1

Author 2
Institution 2

Author 3
Institution 3

Abstract

We study the extension result.

1 Introduction

MRFs have been widely studied and are very important.

We show that the hierarchy of Sherali adams relaxations are intimately related to the use of fourier decomposition approaches for graphical models.

We also provide theoretical understanding of approximation guarantees for these LP relaxations.

2 Preliminaries

[**?**]. We study graphical models over n variables $\{x_1, \cdots, x_n\}$ where $x_i \in \mathcal{A}$. We denote the vector $(x_1, \cdots, x_n)$ as $x_{[n]}$ or simply as x. For any subset $S \subset [n]$ we use the notation x_S to denote the image of $x_{[n]}$ when projected to the subset of variables indexed by S. Let $w(S, x_S)$ be called a "weight" vector indexed by subsets of $[n]$ and assignment projections of $x_{[n]}$ onto S. For probabilisitc graphical models where the energy function can be written as $\exp(f(x_{[n]}))$ where $f\mathcal{A}^n \to \mathbb{R}$ is defined as:

$$f(x) = \sum_{S, x_s} w(S, x_s) \tag{1}$$

The probability distribution is specified by $p(x) = \exp(f(x))$. The MAP inference problem for p is equivalent to finding $x_* = \arg\max_{x \in \mathbb{X}^n} f(x)$. It can be shown that this problem is NP hard in general. Multiple approaches to get around this problem have been proposed. Most notably, to solve instead an LP relaxation of **??**.

Preliminary work. Under review by AISTATS 2018. Do not distribute.

The Shearli Adams hierarchy of LP relaxations and polytopes $\{\mathbb{L}_k\}_{k=2}^n$ provides a progressively stricter relaxation such that $\mathbb{L}_n \subset \cdots \mathbb{L}_2$ and solving the objective function over $\mathbb{L}_n$, also known as the marginal polytope $\mathbb{M}$ yields the right answer for **??**. We say that an LP relaxation is tight for a weight vector $w(S, x_S)$ if it has the same value as solving for $\max_x f(x)$. We will mostly restrict ourselves to binary graphical models where $\mathcal{A} = \{0, 1\}$. We work with a slightly more general version of the Sherali Adams hierarchy.

Definition 1 (Marginal Polytope)**.** We define the marginal polytope $M(\mathcal{X})$ over binary $\{0, 1\}$ variables $\mathcal{X}$ with $|\mathcal{X}| = n$ as $M(\mathcal{X}) \subset \mathbb{R}^{3^n}$ such that if $\mu \in M(\mathcal{X})$ then the dimensions of μ are indexed by (S, x) with $S \subset \{1, \cdots, n\}$ (S is a subset of the variables $\mathcal{X}$) and $x \in \{0, 1\}^{|S|}$ is an assignment for these variables and the following relationship must be satisfied:

$$\mu(S, x) = \mathbb{P}_\mu(S = x) \tag{2}$$

Each μ encodes (consistently) a probability distribution over $\mathcal{X}$ by specifiying the probabilities of all possible events in the sigma algebra generated by the random variables $\mathcal{X}$ (pairs (S, c)).

The MAP inference problem over a MRF with weight vector $w(S, x_S)$ is equivalent to solving the following LP:

$$\max_{\mu \in \mathbb{M}(\mathcal{X})} \langle \mu, w \rangle \tag{3}$$

For all $y \in \{0, 1\}^n$, define μ_y as:

$$\mu_y(S, x) = \begin{cases} 1 & \text{if } x = y_S \\ 0 & \text{o.w.} \end{cases} \tag{4}$$

These are the 2^n integral vertices of $M(\mathcal{X})$. Their convex combinations generate all of $M(\mathcal{X})$. These correspond to all deterministic distributions over these n binary random variables.

Definition 2 (Nested Set system)**.** Let $\mathcal{Y}$ be a base set, and $\mathcal{P}(\mathcal{Y})$ be its power set. A nested family of sets

$\mathcal{S} \subset \mathcal{P}(\mathcal{Y})$ is such that $S' \in \mathcal{S}$ for all $S' \subset S$ such that $S \in \mathcal{S}$. Notice that $\emptyset \in \mathcal{S}$ for any nested set system.

Let $\mathcal{X}$ be a set of variables with $|\mathcal{X}| = n$ and $\mathcal{S}$ a set system over $[n]$. Call S_r to the number of subsets of $\mathcal{S}$ of size r. The number of possible $\{0,1\}$ assignments for all subsets in $\mathcal{S}$ is $\mathcal{N}^{(2)}(\mathcal{S}) = \sum_{i=0}^{n} |S_i| * 2^i$.

If $\mathcal{S}$ is a nested set system, call $\mathcal{S}^{top} = \{S \in \mathcal{S} | \nexists S' \in \mathcal{S}, |S'| > |S| \text{ s.t. } S \subset S'\}$

Definition 3 (Generalized Sherali Adams Hierarchy)**.** Given a nested set system $\mathcal{S}$ over set $[n]$, let $\mathbb{L}_{\mathcal{S}}$ to be the *generalized sherali adams polytope* of vectors $\mu \in \mathbb{R}^{\mathcal{N}^{(2)}(\mathcal{S})}$ defined for $S \in \mathcal{S}$ and $x \in \{0,1\}^{|S|}$:

$$\mu(S, x) = \mathbb{P}_{S'}(S = x), \ \forall S' \in \mathcal{S}^{top} \tag{5}$$

Where $\mathbb{P}_{S'}(S = x)$ stands for the marginal probability of $S = x$ of the distribution indexed by S'. This definition implicitly specifies there must be consistency between the marginalizations from any two distinct $S, S'' \in \mathcal{S}^{top}$ down to a common set S such that $S \subset S'$ and $S \subset S''$.

Given a base set $\mathcal{X}$, define $\mathbb{B}_k$ to be the nested set system of all sets of size $\leq k$.

Definition 4 (Sherali Adams Hierarchy)**.** The $k-$th Sherali Adams polytope over set $[n]$ equals $\mathbb{L}_{\mathbb{B}_k}$. We will use the shorthand notation $\mathbb{L}_k$ instead.

Definition 5 (Pointed Sherali Adams Polytope)**.** Let $x \in [n]$. Define $\mathbb{B}_k^x = \mathbb{B}_k \cup \{\{x\} \cup s | s \in \mathbb{B}_k\}$. The pointed Sherali Adams Polytope over set $[n]$ with special variable x equals $\mathbb{L}_{\mathbb{B}_k^x}$. We will use the shorthand notation $\mathbb{L}_k^x$ instead.

Denote by $W_k = \{w | w(S, x) = 0 \forall S | |S| > k\}$. We refer to $w(S.x)$ for $|S| = k$ as the weights of degree k.

2.1 Related Work

Past ... have considered in depth the problem of when is it that a Sherali adams relaxation is tight with the intention of leveraging this knowledge in the ... of algorithms for inference.

3 Fourier Analysis and LP relaxations

In this section we work with domain $\{-1,1\}^n$ instead of $\{0,1\}^n$. In the next section we show how to translate

Definition 6 (Linearizations)**.** Let $\mathcal{F} \subset \{\{-1,1\} \to \mathbb{R}\}$ be a set of functions from $\{-1,1\}^n$ into the reals. Let $f : \{-1,1\}^n \to \mathbb{R}$ and $V \subset \{-1,1\}^n$. For any $f \in \mathcal{F}$, the optimization problem:

$$\max_{x \in \{-1,1\}^n} f(x) \tag{6}$$

Can be linearlized by introducing variables $y_v \in \mathbb{R}^D$ for some $D \in \mathbb{N}$ meant to represent the points in $\{-1,1\}^n$ and a linearlization $f^D \in \mathbb{R}^D$ of the objective function such that:

$$\langle f^D, y_v \rangle = f(v) \ \forall v \in \{-1,1\}^n \tag{7}$$

Given that we have linearized every $f \in \mathcal{F}$ and $\{-1,1\}^n$ we can consider the LP relaxation:

$$\mathcal{L}(f) = \max_{y \in P} \langle f^D, y \rangle \tag{8}$$

where P is any polytope such that $y_v \in P \ \forall v \in \{-1,1\}^n$.

We denote by the components of the set of linearizations as $(\{f^D | f \in \mathcal{F}\}, \{y_v | v \in \{-1,1\}^n\})$ as $\mathcal{L}(\mathcal{F})$. We use $\mathcal{L}(\mathcal{F})$ to denote the relaxation (the linearizations and polytope).

Definition 7 (Δ approximation)**.** [Consider heavily revising this definition] For a function class $\mathcal{F} \subset \{\{-1,1\}^n \to \mathbb{R}\}$ We say that an LP relaxation $\mathcal{L}(\mathcal{F})$ is a Δ approximation with respect to $\{s_f\}$ if for all $f(x) \in \mathcal{F}$, $\mathcal{L}(f) \leq s_f + \Delta$ where the family $\{s_f\}$ satisfies $\max_{x \in \{-1,1\}^n} f(x) \leq s_f$. When the family $\{s_f\}$ is omited we assume that $s_f = \max_{x \in \{-1,1\}^n} f(x)$ for all f.

Definition 8 (Size of an LP relaxation)**.** The size of an LP relaxation equals the number of constraints it has.

The space of functions $f : \{-1,1\}^n \to \mathbb{R}$ is a $\mathbb{R}^{2^n}$ hilbert space $L^2(\{-1,1\}^n)$ equipped with the inner product:

Definition 9. Let $f, g \in \{\{-1,1\}^n \to \mathbb{R}\}$, define the dot product:

$$\langle f, g \rangle_{\mathcal{L}^2(\{-1,1\}^n)} = \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} f(x) g(x) \tag{9}$$

Definition 10 (Fourier decomposition)**.** Let $x_S = \prod_{i \in S} x_i$. A function $f : \{-1,1\}^n \to \mathbb{R}$ has a unique decomposition of the form:

$$f(x) = \sum_{S \subset [n]} \hat{f}(S) x_S \tag{10}$$

Where $\{x_S\}_{S \subset [n]}$ is an orthonormal fourier basis of polynomials for $L^2(\{-1,1\}^n)$ under the dot product described above. By definition $\hat{f}(S) = \langle f, x_S \rangle_{L^2(\{-1,1\}^n)}$

Lemma 11 (Dot product and Fourier coefficients)**.** *Let $f(x) = \sum_{S\subset[n]} \hat{f}(S)x_S$ and $g(x) = \sum_{S\subset[n]} \hat{g}(S)x_S$. Then:*

$$\langle f(x), g(x)\rangle_{L^2(\{-1,1\}^n)} = \sum_{S\subset[n]} \hat{f}(S)\hat{g}(S) \qquad (11)$$

The following auxiliary theorem will be useful:

Theorem 12. *[From LP relaxations to polynomials and back] LP relaxation $\mathcal{L}(\mathcal{F})$ is a Δ approximation for $\mathcal{F}$ of size R with respect to $\{s_f\}$ if and only if there exist nonnegative functions $q_1, \cdots, q_R : \{-1,1\}^n \to \mathbb{R}_{\geq 0}$ such that for all $f \in \mathcal{F}$ with $\max_{x\in\{-1,1\}^n} f(x) \leq s_f$, ($s_f$ is an upper bound for the optimum of f, dependent on f) the function $s_f + \Delta - f(x)$ is a nonnegative combination of $q_1, \cdots, q_R$:*

$$s_f + \Delta - f(x) \in \{\lambda_0 + \sum_{i=1}^{R} \lambda_i q_i | \lambda_0, \cdots, \lambda_R \geq 0\} \quad (12)$$

The proof is in the appendix.

3.0.1 From $\{0,1\}$ to $\{-1,1\}$

Any MAP problem on binary variables $\{0,1\}$ can be turned into a maximization problem over $\{-1,1\}$ for an objective function in $L(\{-1,1\}^n)$.

Let w be a weight vector indexed by subsets $S \subset [n]$ and binary $\{0,1\}$ variables $z_1, \cdots, z_n$ with coordinates indexed by pairs (S, y_S) encoding the weight corresponding to the set of variables S if z_S agrees with configuration y_S. The weight vector $w \in \mathbb{R}^{3n}$. The objective function to maximize over possible assignments $(z_1, \cdots, z_n) \in \{0,1\}^n$ can be written as the following polynomial:

$$\begin{aligned} f(z) &= \sum_{S,y_S} w(S, y_S) * 1(y_S = z_S) \\ &= \sum_{S,y_S} w(S, y_S) * \prod_{i\in S}(1 - z_i - y_i + 2y_i z_i) \end{aligned}$$

Let $x_i = 2z_i - 1$ and define $\tilde{f}(x) = f\left(\frac{x_1+1}{2}, \cdots, \frac{x_n+1}{2}\right)$. The function $\tilde{f}$ is a polynomial over $\{-1,1\}^n$. Any maximizing configuration for $\tilde{f}$ can be turned into a maximizing configuration of f. The maximum degree of $\tilde{f}$ equals the maximum degree of f.

3.0.2 Sherali Adams as a functional Hierarchy

For any nested set System $\mathcal{S}$, the Sherali Adams polytope $\mathbb{L}_\mathcal{S}$ can be identified with a set of linear functionals acting over $\{0,1\}^n$ polynomials of degree at most $\max\{|S| \text{ s.t. } S \in \mathcal{S}^{top}\}$. Optimizing a weight vector w over all $\{0,1\}^n$ assignments of the underlying variables can be written as a polynomial objective f over $\{0,1\}^n$.

An element $\mu \in \mathbb{L}_\mathcal{S}$ acts over f linearly by mapping every monomial x_S to $\mu(S, 1\cdots 1)$. Similarly the Sherali Adams hierarchy has a similar description when considering $\{-1,1\}^n \to \mathbb{R}$ objectives. The $k-$th level of the Sherali Adams hierarchy can be treated as a set of linear functionals:

Definition 13 (Consider removing)**.** Given $f : \{-1,1\}^n \to \mathbb{R}$, let $\mathcal{V}(f) \subset [n]$ be $\mathcal{V}(f) = \cup_{S|\hat{f}(S)\neq 0} S$

Definition 14 ($\mathcal{S}-$sherali Adams functionals)**.** Let $\mathcal{S}$ be a nested set system. We define a $\mathcal{S}-$local expectation functional $\tilde{\mathbb{E}}_\mathcal{S}$ to be a linear functional on $L^2(\{-1,1\}^n)$ such that:

$$\begin{aligned} \tilde{\mathbb{E}}_\mathcal{S}[x_S] &= \mathbb{E}_{x_S\sim\mu_S}[x_S] && \forall S \in \mathcal{S} \\ \tilde{\mathbb{E}}[x_S] &= l_S && \text{for some } l_s \in \mathbb{R} \text{ o.w.} \end{aligned}$$

For every S there is a distribution over variables x_S whose expectation over x_S agrees with the pseudo expectation operator over the said variables as long as $S \in \mathcal{S}$.

Definition 15 ($k-$round Sherali Adams functionals)**.** If $\mathcal{S} = \mathbb{B}_k$ we call any $\tilde{\mathbb{E}}_\mathcal{S}$ a $k-$local expectation functional.

We denote by $\mathbb{L}_k$ the set of all $k-$local expectation functionals. This corresponds with the $k-$Sherali adams polytope as defined in Section [xxx].

Definition 16. An $r-$junta is a function $f : \{-1,1\}^n \to \mathbb{R}$ that depends solely on (at most) r variables $|\mathcal{V}(f)| \leq k$. Given a nested set system $\mathcal{S}$ over $[n]$, a $\mathcal{S}-$junta is a function f such that $\mathcal{V}(f) \in \mathcal{S}$.

The following Lemma follows directly from the definition.

Lemma 17. *An $\mathcal{S}-$Sherali Adams functional follows:*

$$\begin{aligned} \tilde{\mathbb{E}}_\mathcal{S}[1] &= 1 && \text{for the constant polynomial } 1 \\ \tilde{\mathbb{E}}_\mathcal{S}[P] &\geq 0 && \text{if } \mathcal{V}(P) \in \mathcal{S}, P \geq 0 \end{aligned}$$

Lemma 18 (Consider removing)**.** *A $k-$round Sherali Adams functional follows:*

$$\tilde{\mathbb{E}}_k[1] = 1 \text{ for the constant polynomial } 1$$
$$\tilde{\mathbb{E}}_k[P] \geq 0 \text{ for every nonnegative } k-\text{junta } P$$

Since the space $L^2(\{-1,1\}^n)$ is self dual, all linear functionals over $L^2(\{-1,1\}^n)$ can be identified with points in $L^2(\{-1,1\}^n)$. In this way we can say $\mathbb{L}_k \subset L^2(\{-1,1\}^n)$.

The later means that every $\mathcal{S}-$Sherali Adams functional has a Fourier decomposition:

$$\tilde{\mathbb{E}}_k = \sum_{S \subset [n]} \tilde{\mathbb{E}}_k[x_S] x_S$$

This also provides us with a natural embedding of $\tilde{\mathbb{E}}_k$ into $\mathbb{R}^{2^n}$ by associating every dimension of this space with a subset $S \subset [n]$ and representing every $\tilde{\mathbb{E}}_k$ by its vector of fourier coefficients.

Theorem 19. *Let $\mathcal{F}$ be a function class made of polynomials with maximum degree k. The $\mathcal{S}-$Sherali relaxation is a Δ approximation for the objective class $\mathcal{F}$ if and only if $\Delta + s_f - f$ is a sum of nonnegative $\mathcal{S}-$juntas for all $f \in \mathcal{F}$.*

The proof is in the appendix.

3.0.3 $\mathcal{S}-$Sherali Adams is the most powerful $\mathcal{S}$-local LP relaxation

Theorem 20. *Let $\mathcal{L}(\mathcal{F})$ be Δ approximation for the function class $\mathcal{F}$. By Theorem* **??**, *there exist nonnegative functions (which can be thought of as polynomials) $q_1, \cdots, q_R$ such that $s_f + \Delta - f(x)$ is a nonnegative combination of $\{q_i\}$. If all the $\{q_i\}$ follow $\mathcal{V}(g_i) \in \mathcal{S}$ then $\mathbb{L}_{\mathcal{S}}$ is also a Δ approximation for $\mathcal{F}$. In other words, among all relaxations all of whose constraints are $\mathcal{S}-$juntas, the $k-$Sherali Adams relaxation is the one with the smallest approximation error.*

Proof. Let $f \in \mathcal{F}$. By assumption there exist $\lambda_0(f), \cdots, \lambda_R(f) \geq 0$ such that:

$$s_f + \Delta - f(x) = \lambda_0 + \sum_{i=1}^{R} \lambda_i q_i \tag{13}$$

Let $\tilde{\mathbb{E}}$ be any $\mathcal{S}$-local pseudo expectation functional.

$$\tilde{\mathbb{E}}[s_f + \Delta - f(x)] = s_f + \Delta - \tilde{\mathbb{E}}[f(x)] \tag{14}$$

$$= \lambda_0 + \sum_{i=1}^{R} \lambda_i(f) \tilde{\mathbb{E}}[q_i] \tag{15}$$

$$\geq 0 \tag{16}$$

The last inequality follows because q_i are assumed to be nonnegative with $\mathcal{V}(q_i) \in \mathcal{S}$.

This implies that for all $f \in \mathcal{F}$:

$$s_f + \Delta \geq \tilde{\mathbb{E}}[f(x)] \tag{17}$$

Since this is true for all $\mathcal{S}$-local expectation functionals, it is also true for the one achieving the maximum. This implies the desired result.

□

We have just proved that among all relaxations having local constraints with variables belonging to $\mathcal{S}$, the $\mathcal{S}$-Sherali Adams relaxation is the one that achieves the smallest approximation error.

3.1 Minimal Set Systems

Given $f : \{-1, 1\}^n \to \mathbb{R}$, what is the minimal -ordered by containment- set system $\mathcal{S}$ that is needed so that $\mathbb{L}_{\mathcal{S}}$ achieve the same value as $\mathbb{L}_k$. For $k = 2, 3$ we have the following results:

Theorem 21 (Extension result for $\mathbb{L}_2$, $\mathbb{L}_3$)**.**

We conjecture that this result does not hold for $k = 2, 3$.

Slack variables / juntas perspective / duality perspective of the extension result.

4 Approximation preserving transformations

4.0.1 Technical Lemmas

In this section we develop some lemmas that will aide us.

Lemma 22. *Let f be a degree $\leq k$ polynomial over $\{-1, 1\}^n$ then $\tilde{\mathbb{E}}_k[f] \leq 0$ for all $\tilde{\mathbb{E}}_k$ iff $\mathbb{L}_k(f) \leq 0$ iff $-f$ is a sum of nonnegative $k-$juntas.*

Lemma 23. *Let $f : \{-1, 1\}^n \to \mathbb{R}$ be a degree $\leq k$ polynomial over $\{-1, 1\}^n$ and $f'(x_1, \cdots, x_{n-1}) = f(x_1, \cdots, x_{n-1}, 1)$. Then:*

$$\mathbb{L}_k(f') \leq \mathbb{L}_k(f) \tag{18}$$

The same result holds when $f'(x_1, \cdots, x_{n-1}) = f(x_1, \cdots, x_{n-1}, -1)$.

Corollary 24. *Let $f : \{-1, 1\}^n \to \mathbb{R}$ and $g : \{-1, 1\}^n \to \mathbb{R}$. $\tilde{\mathbb{E}}_k[f] \geq \tilde{\mathbb{E}}_k[g]$ for all $k-$ local pseudo distributions iff $f - g$ is the sum of nonnegative $k-$juntas.*

4.0.2 Clamping

Lemma 25. *Let $\mathcal{F}$ be a family of objective functions over $x_1, \cdots, x_n$ such that the $k-$th Sherali Adams has a Δ approximation error. Let $x^*_{n,f} \in \{-1, 1\}$ be the optimal assignment of x_n for f in a MAP solution. The fam-*

ily $F^{(1)} = \{g(x_1, \cdots, x_{n-1}) = f(x_1, \cdots, x_{n-1}, x^*_{n,f})\}$ *has a* Δ *approximation error as well by* $\mathbb{L}_k$.

Proof. Apply the characterization of Δ approximated families as a sum of nonnegative $k-$juntas.Setting a variable to a particular value yields a representation of the remaining function as a sum of nonnegative juntas as well. Since the MAP assignment for f' achieves the same score as that of f the upper bound s_f is still valid for f'. In particular if s_f equals the MAP value of f, this implies the approximation error is preserved. The statement might not hold if we set x_n to $-x^*_{n,f}$ because in this case the MAP value of f' might not equal that of f and therefore the approximation error might increase by an amount equal to the degradation in the MAP value of f vs that of f'. □

The lemma above proves that clamping variable x_n to its MAP value preserves the approximation error.

Now we provide sufficient conditions under which we can prove that $x^*_n = 1$ [The MAP value of x_n equals 1].

Let $f : \{-1, 1\}^n \to \mathbb{R}$ and $f'(x_1, \cdots, x_{n-1}) = f(x_1, \cdots, x_{n-1}, 1)$.

Lemma 26. *If* $f' - f$ *is a sum of nonnegative juntas (possibly a single size n junta) then there exists a MAP solution with* $x^*_n = 1$.

Proof. Assume the contrary, $\max_{x \in \{-1,1\}^n} f(x) > \max_{x \in \{-1,1\}^n, x_n = 1} f(x)$. Let x^*_{MAP} be a MAP assignment of f. Then $f(x^*_{MAP}) > f'(x^*_{MAP})$ which contradicts the assumption that $f' - f$ is a sum of nonnegative juntas. □

We can turn this lemma into an equivalent statement with an algorithmic flavor:

Lemma 27. *If* $\mathbb{L}_k(f - f') \leq 0$ *then there exists a MAP solution with* $x^*_n = 1$

Proof. By Lemma **??**, $f' - f$ is a sum of nonnegative $k-$juntas iff $\mathbb{L}_k(f - f') \leq 0$. The result follows. □

Equivalently, Lemma **??** states that in case $f' - f$ is tight for $\mathbb{L}_k$, there is a MAP assignment for f with optimal value o for variable x_i. Typically for the case of degree two objectives $f' - f$ will be a star graph and $\mathbb{L}_k$ will be tight for $k \geq 2$. This is intimately related to dual decomposition approaches to inference [Cite a bunch of stuff].

Algorithm 1: Clamping

Input : Objective function $f : \{-1, 1\}^n \to \mathbb{R}$, target variable x_i, objective value $o \in \{-1, 1\}$, Sherali Adams level k.

Set $f'(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) = f(x_1, \cdots, x_i = o, \cdots, x_n)$. Compute $\mathbb{L}_k(f - f')$.
if $\mathbb{L}_k(f - f') \leq 0$ **then**
  Clamp x_i to value o.
end
else
  No clamping is guaranteed.
end

4.0.3 Elimination algorithm

Assume the variable x_n appears in monomials $\{x_{S_i}\}$ for $n \in S_i$ with nonzero coefficients in f.

Define $\tilde{f} : \{-1, 1\}^{n-1} \to \mathbb{R}$ the function resulting from running the elimination algorithm on f and removing variable x_n. In other words:

$$\tilde{f}(x_1, \cdots, x_{n-1}) = \max_{x_n \in \{-1,1\}} f(x_1, \cdots, x_{n-1}, x_n)$$

For more details regarding how to derive the Fourier expansion of $\tilde{f}$ from the expansion of f consult the appendix.

We can split f into those monomials that involve n and those that do not.

$$f(x) = \underbrace{\left(\sum_{S \subset [n] | S \not\in \{S_i\}} \hat{f}(S) x_S \right)}_{\bar{f}_n} + \underbrace{\left(\sum_{S \in \{S_i\}} \hat{f}(S) x_S \right)}_{f_n}$$

Let $S(n) = (\cup_i S_i) \setminus \{n\}$ (the neighbors of x_n) where $S(n) = \{i_1, \cdots, i_{|S(n)|}\}$ then:

$$\tilde{f} = \bar{f}_n + \tilde{f}_n \tag{19}$$

Where abusing notation $\tilde{f}_n(x) = \tilde{f}_n(x_{i_1}, \cdots, x_{i_{|S(n)|}})$ and $f_n(x) = f_n(x_{i_1}, \cdots, x_{i_{|S(n)|}}, x_n)$. And $\tilde{f}_n(x_{i_1}, \cdots, x_{i_{|S(n)|}}) = f_n(x_{i_1}, \cdots, x_{i_{|S(n)|}}, 1)$ if $f_n(x_{i_1}, \cdots, x_{i_{|S(n)|}}, 1) > f_n(x_{i_1}, \cdots, x_{i_{|S(n)|}}, -1)$ and $\tilde{f}_n(x_{i_1}, \cdots, x_{i_{|S(n)|}}) = f_n(x_{i_1}, \cdots, x_{i_{|S(n)|}}, -1)$ otherwise.

After performing one round of elimination of a variable, the approximation error of the Sherali Adams relaxation can only get worse.

Lemma 28. *[Elimination degrades the relaxation error] If $\tilde{f}$ has approximation error Δ for $\mathbb{L}_{\mathcal{S}}$ then f has approximation error at most Δ as long as $S(n) \cup \{x_n\} \in \mathcal{S}$.*

Proof. The MAP score of f and the MAP score of $\tilde{f}$ coincide. Let $s_f = \max_{x \in \{-1,1\}^n} f(x)$ and $s_{\tilde{f}} = \max_{x \in \{-1,1\}^{n-1}} \tilde{f}(x)$. Then $s_f = s_{\tilde{f}}$.

$\tilde{f}$ has approximation error Δ iff $\Delta + s_f - \tilde{f}$ can be written as a sum of nonnegative $\mathcal{S}$-juntas.

In other words there exist $\lambda_0, \cdots, \lambda_R \geq 0$ and $q_1, \cdots, q_R$ nonnegative $\mathcal{S}-$juntas such that:

$$\Delta + s_f - \tilde{f} = \lambda_0 + \sum_{i=1}^{R} \lambda_i q_i \tag{20}$$

The function $\tilde{f}_n - f_n$ is a nonnegative $\mathcal{S}-$junta by definition since $S(n) \cup \{x_n\} \in \mathcal{S}$ and because $\tilde{f}_n - f_n$ is always ≥ 0.

Summing $\tilde{f}_n - f_n$ to both sides of the equation above yields:

$$\Delta + s_f - f = \lambda_0 + \sum_{i=1}^{R} \lambda_i q_i + \left(\tilde{f}_n - f_n\right) \tag{21}$$

Thus showing that $\Delta + s_f - f$ can be written as a sum of nonnegative $\mathcal{S}-$juntas. This in turn implies that the approximation error of $\mathbb{L}_{\mathcal{S}}$ on f is at most Δ. □

As a direct consequence of this lemma, if we can perform elimination and then bound the approximation error of the resulting model, we can obtain a bound for the original version. This yields a very simple proof of the following classical result cite[XXXXXXX]:

Theorem 29. *[Wainwright and Jordan] Any graph of treewidth $k-1$ is tight for $\mathbb{L}_k$*

Proof. Recall that for a graph G the following two statments are equivalent:

- G has treewidth at most $k-1$.
- G has a triangulation (chordal envelope) with maximum clique size of at most k.

In other words, there is an order of the elimination algorithm for which at all times, if v is beging eliminated v has at most $k-1$ neighbors.

Applying Lemma **??** to these successive reduced models, the approximation error must be increasing.

Since at the last step elimination for a treewidth $k-1$ graph is able to end up with a graph of size at most k, $\mathbb{L}_k$ is tight for the model resulting at this last step. Because the approximation error deteriorates and it equals zero at the end of this process it must have been zero all along.

□

We present another proof of Theorem **??** based on the juntas technology in the appendix.

In fact for $k = 2, 3$ it is possible to prove that elimination preserves the approximation error all along.

Theorem 30 (Elimination for $\mathbb{L}_2, \mathbb{L}_3$ preserves approximation)**.** *If f has approximation error Δ for $\mathbb{L}_{\mathcal{S}}$ then $\tilde{f}$ has approximation error at most Δ if $\mathcal{S} = \mathbb{B}_k$.*

We conjecture that for the case of $k \geq 4$ this is not true.

4.1 Minors

Let $\mathcal{G}_k(\Delta)$ be the set of graphs that achieve approximation error at most Δ under $\mathbb{L}_k$ for all node and edge potentials.

Theorem 31. *$\mathcal{G}_k(\Delta)$ is closed under taking minors (resp. signed minors).*

Proof. The same proof of minor closure as for the case of $\Delta = 0$ works here. □

4.2 Flippings

Lemma 32. *Flipping reparametrization preserve the approximation error for $\mathbb{L}_k$.*

Proof. $f : \{-1,1\}^n \to \mathbb{R}$ has approximation error at most Δ if and only if there are $\lambda_0, \cdots, \lambda_R \geq 0$ and $q_1, \cdots, q_R$ nonnegative $k-$juntas such that:

$$\Delta + s_f - f = \lambda_0 + \sum_{i=1}^{R} \lambda_i g_i \tag{22}$$

Where $s_f = \max_{x \in \{-1,1\}^n} f(x)$. Let $\hat{f}(x)$ be an objective function for which some variables are flipped with respect to f. Notice $s_{\hat{f}} = s_f$. By performing this flipping operation on the LHS and RHS of the equation above we obtain $\Delta + s_f - \hat{f}$ and flipped nonnegative $k-$juntas respectively, thus verifying that the approximation error of $\hat{f}$ is at most Δ as well. In fact this also shows that the approximation error is exactly the same for f and for $\hat{f}$. □

4.3 Balanced models

In [Adrian's paper on LOC tightness] it is shown that balanced models are tight for LOC where the objective function is over $\{0,1\}^n$:

$$f(z) = \sum_{i=1}^{n} \theta_i z_i + \sum_{(i,j)\in E} w_{i,j} z_i z_j \tag{23}$$

And the optimization problem equals:

$$\max_{z\in\{0,1\}^n} f(z) \tag{24}$$

Reparametrizing f so it is over variables in $x \in \{-1,1\}^n$ by the formula $x_i = 2z_i - 1$:

$$f(x) = \sum_{i=1}^{n} \theta_i \left(\frac{x_i+1}{2}\right) + \sum_{(i,j)\in E} w_{i,j} \left(\frac{x_i+1}{2}\cdot\frac{x_j+1}{2}\right) \tag{25}$$

The signs of the coefficients of $z_i z_j$ map exactly to the signs of coefficients for the $x_i x_j$ terms. In other words (abusing terminology) the model using the ensemble $\{z_i\}$ is almost balanced if and only if the omdel using the ensemble $\{x_i\}$ is. We can then provide a somewhat constructive proof of the following statement:

Lemma 33. *All balanced models $f(x) = \sum \theta_i x_i + \sum_{(i,j)} w_{i,j} x_i x_j$ for $x \in \{-1,1\}^n$ are tight for LOC.*

INCOMPLETE SO FAR - THE AISTATS IDEAS might work.
By flipping variables and Harari's theorem we can reduceit to prove tightness for attractive models. Wlog we can assume f has no constant (bias) term.

If $s_f = \max_{x\in\{-1,1\}^n} f(x)$, tightness of f for LOC holds iff $s_f - f$ can be written as a sum of nonnegative 2−juntas.

Let these juntas be $q_{i,j}$. We can aggregate all juntas involving exactly variables i and j into $q_{i,j}$. Under this condition, the only junta that can have a term of the form $x_i x_j$ is $q_{i,j}$. Therefore the coefficient of $x_i x_j$ in $q_{i,j}$ must be $-w_{i,j}$.

Write $q_{i,j}(x_i, x_j) = -\theta'_i x_i - \theta'_j x_j - w_{i,j} + c_{i,j}$ for some θ'_i, θ'_j. Since $q_{i,j}$ is nonnegative for all $x_i, x_j \in \{-1,1\}$ pairs, there is an assignment of x_i, x_j for which $-\theta_i^{(i,j)} x_i - \theta_j^{(i,j)} x_j - w_{i,j} \le 0$ implying that $c_{i,j} \ge 0$.

It must be the case that $\sum_j \theta_z^{(i,j)} = \theta_z$ for all z.

Notice that $s_f \ge 0$ because $w_{i,j} \ge 0$, either the all ones vector or the all -1 vector achieves a nonnegative score.

Let x^* be the optimal assignment for f.

We show that we can distribute the score s_f locally among the edges to ensure that all the juntas are non negative. □

4.4 Uprooting

The goal is to show an operator version of uprootings. Let $f(x) : \{-1,1\}^n \to \mathbb{R}$ be a polynomial made of monomials of degree d and degree $d-1$ only. Let's augment $f(x)$ by adding an extra variable x_{n+1} and write $f^{up}(x) : \{-1,1\}^{n+1} \to \mathbb{R}$ be the polynomial defined by multiplying all degree $d-1$ monomials of f by x_{n+1} and letting the degree d monomials intact. All monomials of f^{up} have degree d. Let $k \ge d$.

Lemma 34.

$$\mathbb{L}_k(f) = \max_{\tilde{E}_k | \tilde{E}_k[x_{n+1}]=1} \tilde{E}_k[f^{up}] \tag{26}$$

Proof. The two directions of the proof are easy to argue. □

Lemma 35. *When $d = 2$ and $k = 3$:*

$$\mathbb{L}_k(f) = \max_{\tilde{E}_k | \tilde{E}_k[x_{n+1}]=-1} \tilde{E}_k[f^{up}] \tag{27}$$

Proof. Write $\mathbb{L}_k(f) - f$ as a sum of nonnegative $k-$juntas. Collect all juntas that use exactly the same k variables.

The junta corresponding to (i_1, i_2, i_3) cannot have a nonzero coefficient on the monomial $x_{i_1} x_{i_2} x_{i_3}$ since by assumption $\mathbb{L}_k(f) - f$ doesn't contain that monomial. In other words, all juntas have only monomials of degree 1 and 2. Multiply the degree 1 monomials by x_{n+1}.

Notice that the range of the juntas hasn't changed after adding x_{n+1}.

Indeed any value in the range pre multiplying by x_{n+1} can be achieved by setting $x_{n+1} = 1$, any value achieved by setting $x_{n+1} = 1$ can be trivially achieved in the range before multiplying by x_{n+1}. Any value in the range obtained with $x_{n+1} = -1$ can be achieved in the range before multiplying by x_{n+1} by flipping the variables that are multiplied by x_{n+1}.

In other words, after multiplying by x_{n+1} the juntas are still nonnegative.

We have just written $\mathbb{L}_k(f) - f^{up}$ as a sum of nonnegative juntas (some of them possibly using $k+1$ variables). Since all the monomials in f^{up} have an even degree:

$f^{up}(x) = f^{up}(\bar{x})$ where $\bar{x}$ denotes the flipped version of $x \in \{-1,1\}^{n+1}$.

This implies that $\mathbb{L}_k(f) - f^{up}(x)|_{x_{n+1}=1}$ can achieve the same values as $\mathbb{L}_k(f) - f^{up}(x)|_{x_{n+1}=-1}$.

Since $\mathbb{L}_k(f) - f^{up}(x)|_{x_{n+1}=1} = \mathbb{L}_k(f) - f(x)$, this immediately yields a characterization of $\mathbb{L}_k(f) - f^{up}(x)|_{x_{n+1}=-1}$ as a sum of nonnegative $k-$juntas.

This concludes the result.

□

Lemma 36. *If $g : \{-1,1\}^m \to \mathbb{R}$ is a function such that $g(x) = g(\bar{x})$, where $\bar{x}$ is the flipped version of x, then $range(g|_{x_m=1}) = range(g|_{x_m=-1})$.*

Proof. A simple calculation shows this to be the case. □

Definition 37 (Generalized uprooting)**.** Let $f : \{-1,1\}^n \to \mathbb{R}$. Define the generalized uprooting of f to be a function $\hat{f} : \{-1,1\}^{n+1} \to \mathbb{R}$ built from f by multiplying all the odd degree monomials of f by x_{n+1}.

Definition 38 (Generalized rerooting)**.** Let $f : \{-1,1\}^{n+1} \to \mathbb{R}$. Define the generalized rerooting of f under variable v to be $\underline{\mathrm{f}} : \{-1,1\}^n \to \mathbb{R}$ built from f by setting x_v to 1.

Lemma 39. *Let $s_f = \max_{x\in\{-1,1\}^n} f(x)$. Let f be a polynomial of degree $d \le k$. If L_k be a Δ approximation for $\{f\}$. Then its generalized uprooting $\hat{f}$ satisfies that $\Delta + s_f - \hat{f}$ can be written as a sum of nonnegative $k+1$-juntas. Furthermore $\mathbb{L}_{k+1}$ is a Δ approximation for any rerooting of $\hat{f}$.*

Proof. By assumption $\Delta + s_f - f = \sum_i^R q_i$ where q_i are nonnegative $k-$juntas. Here we have coalesced the λ_i into the q_i.

Multiply all odd degree monomials of f by x_{n+1}. Do the same for the odd degree monomials of each junta q_i. The two sides of the quality should coincide after this procedure. All the resulting functions, $\hat{f}$ (since adding x_{n+1} is exactly the uprooting operation for f) and $\{\hat{q}_i\}$ (uprootings of q_i with variable x_{n+1}) are even. Furthermore, $\hat{q}_i$ are nonnegative. The later is true since by lemma **??**, $range(\hat{q}_i|_{x_{n+1}=-1}) = range(\hat{q}_i|_{x_{n+1}=1}) = range(q_i) \ge 0$. We obtain the following representation:

$$\Delta + s_f - \hat{f} = \sum_{i=1}^{R} \hat{q}_i \tag{28}$$

In othder words $\Delta + s_f - \hat{f}$ can be written as a sum of nonnegative $k+1$-juntas.

We now show the rerooting statement.

By Lemma **??**, setting any of the variables $x_1, \cdots, x_{n+1}$ to either 1 or -1 leaves the range of the function $\Delta + s_f - \hat{f}$ unchanged.

Therefore for any $v \in \{1, \cdots, n+1\}$, the generalized rerooting of $\hat{f}$ and the resulting $\Delta + s_f - \underline{\hat{\mathrm{f}}}$ can be written as:

$$\Delta + s_f - \underline{\hat{\mathrm{f}}} = \sum_{i=1}^{R} \underline{\hat{\mathrm{q}}}_i \tag{29}$$

Where $\underline{\hat{\mathrm{q}}}_i$ equals $\hat{q}_i(\cdot, x_v = 1)$ for all i. Notice that x_v might not affect $\hat{q}_i$ in which case, no variable in $\hat{q}_i$ will "cancel" in the rerooting.

Clearly $\underline{\hat{\mathrm{q}}}_i \ge 0$ and depends on at most $k+1$ variables for all i.

Let $s_{\underline{\hat{\mathrm{f}}}} = \max_{x\in\{-1,1\}^n} \underline{\hat{\mathrm{f}}}(x)$.

Notice that $s_{\underline{\hat{\mathrm{f}}}} = s_f$ because both quantities equal $\max_{x\in\{-1,1\}^{n+1}} \hat{f}(x)$- the max score of the uprooted model.

Since $s_{\underline{\hat{\mathrm{f}}}} = s_f$ it follows that $\Delta + s_{\underline{\hat{\mathrm{f}}}} - \underline{\hat{\mathrm{f}}}$ is a sum of $k+1$-nonnegative juntas and therefore $\mathbb{L}_{k+1}$ has an approximation error of at most Δ over the rerooted model $\underline{\hat{\mathrm{f}}}$.

□

4.4.1 Relationship with the existing Uprooting paper

If we identify in the paper 1 with 1 and -1 with the assignment value of 0, then we can identify even potentials with the polynomials:

$$\frac{\pi_{i\in U} x_i - 1}{2} \text{ if } |U| \text{ is even} \tag{30}$$

$$\frac{1 - \pi_{i\in U} x_i}{2} \text{ if } |U| \text{ is odd} \tag{31}$$

Clearly these polynomials for a basis of the space of polynomials of degree at most k as long as we define the polynomial corresponding to $U = \emptyset$ to be a constant equal to 1.

This is the equivalent of Proposition 12:

Following the above definition of uprootings and of even potentials, it immediately follows that any even -k-potential when k is even remains unmodified by the uprooting operation.

Theorem 17 is basically Lemma 36.

In order to build the counterexample that is shown for refuting the universal uprooting:

Let a distribution be defined over $x_{i_1}, \cdots, x_{i_k}$ by declaring that all configurations of the form $\{(-1, 1, \cdots, 1), (1, -1, \cdots, 1), \cdots, (1, 1, \cdots, -1)\}$ be equally likely with probability $\frac{1}{k}$.

Let $U \subseteq \{i_1, \cdots, i_k\}$:

$$P(x_U) = \begin{cases} 0 & \text{if } \exists\ a \neq b \text{ s.t. } x_{i_a} = x_{i_b} = -1 \\ \frac{1}{k} & \text{if } x_{i_a} = -1 \text{ for a single } a \\ \frac{k-|U|}{k} & \text{if } x_{i_a} = 1 \forall a \end{cases} \tag{32}$$

The resulting pseudoexpectation therefore satisfies:

$$\tilde{\mathbb{E}}[x_U] = \frac{k - 2|U|}{k} \tag{33}$$

In particular $\tilde{\mathbb{E}}[x_U] = -1$ whenever $|U| = k$. This tells us that for this objective f (sums of the indicators of even functions) this particular pseudoexpectation gives us a value of 0 for the case when k is even.

The same analysis that is done in the paper carries over.

The pseudoexpectation $\tilde{\mathbb{E}}$ can be written as a degree k polynomial over $\{-1, 1\}^{k+1}$ with coefficients for x_U equal to $\tilde{\mathbb{E}}[x_U]$.

4.4.2 Finding the $k-$junta core of a model

We present an algorithm such that given a real value B and $f : \{-1, 1\}^n \to \mathbb{R}$ a polynomial of degree d, finds a submodel $f' : \{-1, 1\}^n \to \mathbb{R}$ such that $\mathbb{L}_k(f') \leq B$, and the l_1 distance between f and f' is minimal.

Define the l_1 distance between two polynomials over $\{-1, 1\}^n$ as the sum of the absolute value of their coefficients' differences.

In order to find f' we propose to solve an LP based on the idea of making sure that $B - f'$ can be written as a sum of nonnegative $k-$juntas while at the same time ensuring that $|f - f'|$ be small.

Assume that $B - f'$ was a sum of nonnegative $k-$juntas. For every $S \subset [n]$ with $|S| = k$ and $S = \{i_1, \cdots, i_k\}$ let $g_S(x_{i_1}, \cdots, x_{i_k})$ be the corresponding $k-$junta for subset S in this decomposition. Let $W \subset [n]$, denote $g_S(W)$ the coefficient of x_W in the fourier expansion of g_S. Obviously $g_S(W) = 0$ if $W \not\subset S$.

If $\hat{f'}(U)$ is the fourier coefficient for $U \subset [n]$ the following relationship must hold $\forall U \subset [n]$ and $U \neq \emptyset$:

$$\hat{f'}(U) = -\sum_{S \subset [n] || S| = k} g_S(U) \tag{34}$$

And

$$B - \hat{f'}(\emptyset) = \sum_{S \subset [n] || S| = k} g_S(\emptyset) \tag{35}$$

A junta $g_S(x_{i_1}, \cdots, x_{i_k})$ is nonnegative if for all possible assignments of $x_{i_1}, \cdots, x_{i_k} \in \{-1, 1\}^k$ the function returns a nonnegative value.

The following LP solves for f' over variables $\{z_U\}_{U \subset [n] || U| \leq d}$, $\{g_S(U)\}_{S \subset [n] || S| = k, U \subset S}$:

$$\begin{aligned} \min \sum_{U \subset [n] || U| \leq d} z_U \\ z_U \geq \hat{f}(U) + \sum_{S \subset [n] || S| = k} g_S(U)\ \forall U \neq \emptyset, |U| \leq d \\ z_U \geq -\hat{f}(U) - \sum_{S \subset [n] || S| = k} g_S(U)\ \forall U \neq \emptyset, |U| \leq d \\ z_\emptyset \geq -B + \hat{f}(\emptyset) + \sum_{S \subset [n] || S| = k} g_S(\emptyset) \\ z_\emptyset \geq B - \hat{f}(\emptyset) - \sum_{S \subset [n] || S| = k} g_S(\emptyset)\ \forall U \neq \emptyset \\ \sum_{U \subset S} g_S(U) x_U \geq 0\ \forall (x_{i_1}, \cdots, x_{i_k}) \in \{-1, 1\}^k, \forall S \subset [n], |S| = k \end{aligned}$$

If this program achieves an objective value R, $|f' - f|_1 = R$ and $\mathbb{L}_k(f') \leq B$.

IDEA: Maybe we can impose LP constraints over the remaining coefficients that ensure tightness of the f' model. Maybe using some minors characterization?

IDEA: what about a minors characterization of higher order potentials programs by using the idea of Wainwright and Jordan or linearizing the objective.

4.4.3 Model Pasting with approximation errors

Here we revert to the notation pre Fourier analysis. We should be able to obtain a similar result with the new Fourier Technology, but for lack of time and in the spirit of completeness I append this version of the result.

Let A, B, C be three disjoint variable sets and consider $L(A, B)$ and $L(B, C)$ be relaxations of $M(A, B)$ and $M(B, C)$ the marginal polytopes over the set of variables $A \cup B$ and $B \cup C$.

We use the notation $\mu^{(1)}$ to denote points in $L(A, B)$ and $\mu^{(2)}$ to denote points in $L(B, C)$. We can partition $\mu^{(1)}$ and $\mu^{(2)}$ in two sets of dimensions as $\mu^{(1)} = [\mu_B, \mu_{A,B}]$ were $\mu_B \in L(B) \subset M(B)$ and

$\mu^{(2)} = [\mu_B, \mu_{B,C}]$ were $\mu_B \in L(B)$ as well, while $\mu_{A,B}$ and $\mu_{B,C}$ correspond to all dimensions where S is either entangled between variables in B and variables in A (respectively C) or only contains variables in A (respectively C).

Notice that $M(A, B, C)$ and $L(A, B, C)$ live in a $3^{|A|+|B|+|C|}$ dimensional space.

Embed $M(A, B)$, $L(A, B)$, $M(B, C)$ and $L(B, C)$ in the natural way as infinite polyhedron in the space where $M(A, B, C)$ and $L(A, B, C)$ live by padding the dimensions that are not constrained by taking a product with copies of $\mathbb{R}$, consider $M(A, B) \cap M(B, C)$ and $L(A, B) \cap L(B, C)$. Observe that $M(A, B, C) \subseteq M(A, B) \cap M(B, C)$ and $L(A, B, C) \subseteq L(A, B) \cap L(B, C)$.

Theorem 40 (Model pasting)**.** *Let $w \in \mathbb{R}^{3^{|A|+|B|+|C|}}$ such that all entries of w corresponding to dimensions not present in $M(A, B) \cap M(B, C) \subset L(A, B) \cap L(B, C)$ equal zero such that $w = [w^*_{A,B}, w_B, w^*_{B,C}, 0]$. Where $w^*_{A,B} \in \Omega_{A,B}$, $w^*_{B,C} \in \Omega_{B,C}$ and $w_B \in \mathbb{R}^{3^{|B|}}$.*

Assume that for any such w:

a) $\max_{\mu^{(1)} \in M(A,B)} \langle (w^*_{A,B}, w_B), \mu^{(1)} \rangle \leq \max_{\mu^{(1)} \in L(A,B)} \langle (w^*_{A,B}, w_B), \mu^{(1)} \rangle + c_{A,B}$ *for some constant $c_{A,B} \geq 0$.*

b) $\max_{\mu^{(2)} \in M(B,C)} \langle (w_B, w^*_{B,C}), \mu^{(2)} \rangle \leq \max_{\mu^{(2)} \in L(B,C)} \langle (w_B, w^*_{B,C}, w_B), \mu^{(2)} \rangle + c_{B,C}$ *for some constant $c_{B,C} \geq 0$.*

The following holds:

$$\max_{\mu \in M(A,B,C)} \langle w, \mu \rangle = \max_{\mu^* \in L(A,B) \cap L(B,C)} \langle w, \mu^* \rangle + c_{A,B} + c_{B,C} \tag{36}$$

References

A APPENDIX: SUPPLEMENTARY MATERIAL
Approximation Guarantees for Sherali Adams Relaxations

Theorem 41. *[From LP relaxations to polynomials and back] LP relaxation $\mathcal{L}(\mathcal{F})$ is a Δ approximation for $\mathcal{F}$ of size R with respect to $\{s_f\}$ if and only if there exist nonnegative functions $q_1, \cdots, q_R : \{-1,1\}^n \to \mathbb{R}_{\geq 0}$ such that for all $f \in \mathcal{F}$ with $\max_{x\in\{-1,1\}^n} f(x) \leq s_f$, ($s_f$ is an upper bound for the optimum of f, dependent on f) the function $s_f + \Delta - f(x)$ is a nonnegative combination of $q_1, \cdots, q_R$:*

$$s_f + \Delta - f(x) \in \{\lambda_0 + \sum_{i=1}^{R} \lambda_i q_i | \lambda_0, \cdots, \lambda_R \geq 0\} \tag{37}$$

Proof. Let $(\{f^D | f \in \mathcal{F}\}, \{y_v \in v \in \{-1,1\}^n)$ be the linearizations of $f \in \mathcal{F}$ and $\{-1,1\}^n$. Assume $\mathcal{L}(\mathcal{F})$ be a Δ approximation for $\{s_f\}$. Let P be the polytope corresponding to the relaxation $\mathcal{L}(\mathcal{F})$ all embedded in $\mathbb{R}^D$. Let P be specified by R linear inequalities $\langle A_i, y\rangle \leq b_i$ where $y \in \mathbb{R}^D$. Since every $y_v \in P$ for every $v \in \{-1,1\}^n$, it follows that $b_i - \langle A_i, y_v\rangle \geq 0 \ \forall v \in \{-1,1\}^n$. Define $q_i(x) = b_i - \langle A_i, y_x\rangle$. By definition $q_i : \{-1,1\}^n \to \mathbb{R}_{\geq 0}$.

Let $f \in \mathcal{F}$ such that $\max_{x\in\{-1,1\}^n} f(x) \leq s_f$. Since $\mathcal{L}(\mathcal{F})$ is assumed to be a Δ approximation we have that $\mathcal{L}(f) \leq s_f + \Delta$. In other words, $s_f + \Delta \geq \langle f^D, y\rangle \ \forall y \in P$. Farkas lemma tells us that any valid inequality over P can be written as a nonnegative combination over the inequalities $\{b_i \geq \langle A_i, y\rangle \geq 0 : i = 1, \cdots, R\}$ and the inequality $1 \geq 0$ (vector of all ones). This yields the existence of nonnegative numbers $\{\lambda_i(f)\}$ such that $\Delta + s_f - \langle f^D, y\rangle = \lambda_0(f) + \sum_{i=1}^{R} \lambda_i(f)(b_i - \langle A_i, y\rangle)$ holds for all $y \in P$.

In particular the later holds for all y_v for $v \in \{-1,1\}^n$. Since $\langle f^D, y_x\rangle = f(x)$, this finishes one direction of the proof.

For the return direction. Consider $\{q_i\}$ a set of functions satisfying the assumptions of the theorem. We will exhibit a $D = 2^n$ dimensional linear programming relaxation $\mathcal{L}(\mathcal{F})$ induced by these functions. We will then show this relaxation is a Δ approximation for $\mathcal{F}$.

We will embed the relaxation in a 2^n dimensional space. For all $x \in \{-1,1\}$ define y_x to be the 2^n dimensional vector $y_x(S) = x_S$ for all $S \subset [n]$.

By definition all the functions q_i belong to $L^2(\{-1,1\}^n)$ and therefore can be written as $q_i(x) = \sum_{S\subset[n]} \hat{q}_i(S) x_S$. Define the linearizations of q_i to be the vectors of their 2^n Fourier coefficients $\hat{q}_i$. For every $x \in \{-1,1\}^n$, $q_i(x) = \langle \hat{q}_i, y_x\rangle$. Recall that $q_i(x) \geq 0$ for all $x \in \{-1,1\}^n$ by assumption.

Let $P \subset \mathbb{R}^{2^n}$ defined by $P = \{y | \langle \hat{q}_i, y\rangle \geq 0 \forall i = 1, \cdots, R | y_\emptyset = 1\}$. By definition $y_x \in P$ for all $x \in \{-1,1\}^n$. Where $y_\emptyset$ denotes the coordinate of y corresponding to the empty set. This restriction encodes the constant terms of the functions q_i. By projecting down to the $2^n - 1$ dimensional space corresponding to all $S \subset [n]$ with $S \neq \emptyset$ we obtain a polytope $\hat{P}$ with at most R constraints.

Similarly embed every $f \in \mathcal{F}$ in $\mathbb{R}^{2^n}$ by mapping f to $\hat{f}$, the vector of f's Fourier coefficients.

It remains to see that P and these linearlizations define a Δ approximation of $\mathcal{F}$ with respect to $\{s_f\}$.

By assumption for every f there exist $\lambda_0(f), \cdots, \lambda_R(f) \geq 0$ such that $s_f + \Delta - \langle \hat{f}, y_x\rangle = \lambda_0(f) + \sum_{i=1}^{R} \lambda_i(f)\langle \hat{q}_i, y_x\rangle$ for all $x \in \{-1,1\}^n$. Let $1_\emptyset$ be the indicator vector for the coordinate corresponding to $\emptyset$. Then $\langle (s_f + \Delta) * 1_\emptyset, y_x\rangle_{L^2(\{-1,1\}^n)} = s_f + \Delta$.

Since this equality holds for all elements of a basis of the space (the set $\{y_x | x \in \{-1,1\}^n\}$), in fact a vector equality holds: $(s_f + \Delta) * 1_\emptyset - \hat{f} = \lambda_0(f) * 1_\emptyset + \sum_{i=1}^{R} \lambda_i(f)\hat{q}_i$.

This in turn implies that for all $y \in P$:

$$s_f + \Delta + \langle \hat{f}, y\rangle = \lambda_0(f) + \sum_{i=1}^{R} \lambda_i(f)\langle \hat{q}_i, y\rangle \tag{38}$$

Since for all $y \in P$ we have that $\langle \hat{q}_i, y\rangle \geq 0$, by Equation **??**, $\forall y \in P$:

$$\langle \hat{f}, y \rangle \leq \Delta + s_f \tag{39}$$

This implies that the resulting LP relaxation $L(\mathcal{F})$ is a δ approximation for $\mathcal{F}$.

□

Theorem 42. *Let $\mathcal{F}$ be a function class made of polynomials with maximum degree k. The $k-$Sherali relaxation is a Δ approximation for the objective class $\mathcal{F}$ if and only if $\Delta + s_f - f$ is a sum of nonnegative $k-$juntas for all $f \in \mathcal{F}$.*

Proof. Assume the $k-$Sherali Relaxation is a Δ approximation for the objective class $\mathcal{F}$. First we show that we can assume the optimum over the Sherali Adams polytope is a local expectation functional $\tilde{\mathbb{E}}$ with $\tilde{\mathbb{E}}[x_S] = 0$ for all $|S| > k$. This is true because for all objective functions the coefficients for those dimensions equal zero.

By Theorem **??**, if the $k-$Sherali Adams relaxation is a Δ approximation for $\mathcal{F}$ then for $R = \binom{n}{r}$ and every $f \in \mathcal{F}$ there exist $\lambda_0(f), \cdots, \lambda_R(f)$ and $q_1, \cdots, q_R$ such that:

$$s_f + \Delta - f = \lambda_0(f) + \sum_{i=1}^{R} \lambda_i(f) q_i \tag{40}$$

Where the q_i are the constraints of Sherali adams.

Since all the Sherali Adams constriaints are $k-$juntas, $s_f + \Delta - f$ can be written as a sum of nonnegative $k-$juntas, $\lambda_0(f), \lambda_1(f)q_1, \cdots, \lambda_R(f)q_R$.

If $s_f + \Delta - f$ is the sum of nonnegative $k-$juntas, then:

$$s_f + \Delta - f = \sum g_i \tag{41}$$

Where g_i are all nonnegative $k-$juntas.

This implies that for all $k-$Sherali Adams local expectation functionals $\tilde{\mathbb{E}}$ with $\tilde{\mathbb{E}}[x_S] = 0$ for all $|S| > k$ we have:

$$\begin{aligned} \tilde{\mathbb{E}}[s_f + \Delta - f] &= s_f + \Delta - \tilde{\mathbb{E}}[f] && (42)\\ &= \sum \tilde{\mathbb{E}}[g_i] && (43)\\ &\geq 0 && (44) \end{aligned}$$

$\tilde{\mathbb{E}}[f] \leq s_f + \Delta$ which implies that the $k-$Sherali Adams relaxation achieves an objective value of at most $s_f + \Delta$ and therefore that it is a Δ approximation for $\mathcal{F}$.

□

B Technical Lemmas

In this section we develop some lemmas that will aide us.

Lemma 43. *Let f be a degree $\leq k$ polynomial over $\{-1, 1\}^n$ then $\tilde{\mathbb{E}}_k[f] \leq 0$ for all $\tilde{\mathbb{E}}_k$ iff $\mathbb{L}_k(f) \leq 0$ iff $-f$ is a sum of nonnegative $k-$juntas.*

Proof. Let $s_f = \max_{x \in \{-1,1\}^n} f(x)$. If $\tilde{\mathbb{E}}_k[f] \leq 0$ for all $\tilde{\mathbb{E}}_k$ then $\mathbb{L}_k(f) \leq 0$ which implies that $s_f \leq \mathbb{L}_k(f) \leq 0$.

Therefore by Theorem **??**, $(0 - s_f) + s_f - f$ can be written as a sum of nonnegative $k-$juntas which implies the desired result.

For the return direction if $-f = \sum g_i$ with $g_i \geq 0$ a $k-$junta for all i then $\tilde{\mathbb{E}}_k[-f] = \sum \tilde{\mathbb{E}}_k[g_i] \geq 0$ for all $\tilde{\mathbb{E}}_k$ which in turn implies that $\mathbb{L}_k(f) \leq 0$ which implies the desired result.

□

Lemma 44. *Let $f : \{-1,1\}^n \to \mathbb{R}$ be a degree $\leq k$ polynomial over $\{-1,1\}^n$ and $f'(x_1, \cdots, x_{n-1}) = f(x_1, \cdots, x_{n-1}, 1)$. Then:*

$$\mathbb{L}_k(f') \leq \mathbb{L}_k(f) \tag{45}$$

The same result holds when $f'(x_1, \cdots, x_{n-1}) = f(x_1, \cdots, x_{n-1}, -1)$.

Proof. Every pseudo distribution $\tilde{\mathbb{E}}_k$ over variables $x_1, \cdots, x_{n-1}$ can be extended to a pseudo distribution $\tilde{\mathbb{E}}'_k$ over variables $x_1, \cdots, x_n$ by setting $\tilde{\mathbb{E}}'_k[x_n] = 1$ (or -1) which forces that for all $S \subset [n]$ with $|S| \leq k$ and $n \in S$, $\tilde{\mathbb{E}}'_k[x_S] = \tilde{\mathbb{E}}_k[x_{S \setminus n}]$ or $\tilde{\mathbb{E}}'_k[x_S] = -\tilde{\mathbb{E}}_k[x_{S \setminus n}]$ (if we set $\tilde{\mathbb{E}}'_k[x_n] = -1$).

We can then think of the set of pseudo distributions over which we are maximising to obtain $\mathbb{L}_k(f')$ is contained in the set of pseudo distributions we are maximising to obtain $\mathbb{L}_k(f)$, the inequality follows.

□

Corollary 45. *Let $f : \{-1,1\}^n \to \mathbb{R}$ and $g : \{-1,1\}^n \to \mathbb{R}$. $\tilde{\mathbb{E}}_k[f] \geq \tilde{\mathbb{E}}_k[g]$ for all $k-$ local pseudo distributions iff $f - g$ is the sum of nonnegative $k-$juntas.*

B.1 Wainwright Jordan

Here we prove the Wainwright Jordan result based on a different set of assumptions.

There is an alternative proof based on the following idea:

Let $s_f = \max_{x \in \{-1,1\}^n} f(x)$. The relaxation $\mathbb{L}_k$ is tight for f iff $s_f - f$ can be written as a sum of nonnegative $k-$juntas. It is possible to explicitly write this difference as a sum of nonnegative $k-$juntas if the graph has treewidth at most $k-1$. The construction is based on the elimination algorithm.

B.1.1 MAP Elimination algorithm in the fourier domain

We can produce a MAP elimination algorithm in the fourier domain by simply noting that to transform f into $\tilde{f}$ it is sufficient to multiply f_n by the following polynomial:

Define $u(x_{S(n)}) = \arg\max_{x_n \in \{-1,1\}} f_n(x_{S(n)}, x_n)$ the function that maps an ensemble of assignments for the neighbors of x_n to the maximizing value of x_n for f_n given this assignment.

$$g_n(x_{S(n)}) = \sum_{(a_{i_1}, \cdots, a_{i_{|S(n)|}}) \in \{-1,1\}^{|S(n)|}} u(a_{i_1}, \cdots, a_{i_{|S(n)|}}) x_n \prod_{j=1}^{|S(n)|} \left(\frac{a_{i_j} - x_{i_j}}{2} \right)^2 \tag{46}$$

It is easy to check that $f_n \cdot g_n = \tilde{f}_n$. This implies the algorithm. Recall that $x_i^2 = 1$ for all i.

Using these ideas we can switch between the value representation and the fourier representation when performing the elimination algorithm.

This procedure requires the value representation of the elimination step (I believe) so it might not really save time.

Working with the fourier basis has the virtue that it allows to decide what is the best sherali relaxation by just looking at the max degree of f, even while performing elimination whereas the value representation might indicate a larger k. This is equivalent to requiring the value representation be always written in terms of the even potentials of the uprootings paper.

Theorem 46 (Model pasting)**.** *Let $w \in \mathbb{R}^{3^{|A|+|B|+|C|}}$ such that all entries of w corresponding to dimensions not present in $M(A,B) \cap M(B,C) \subset L(A,B) \cap L(B,C)$ equal zero such that $w = [w^*_{A,B}, w_B, w^*_{B,C}, 0]$. Where $w^*_{A,B} \in \Omega_{A,B}$, $w^*_{B,C} \in \Omega_{B,C}$ and $w_B \in \mathbb{R}^{3^{|B|}}$.*

Assume that for any such w:

a) $\max_{\mu^{(1)} \in M(A,B)} \langle (w^*_{A,B}, w_B), \mu^{(1)} \rangle \leq \max_{\mu^{(1)} \in L(A,B)} \langle (w^*_{A,B}, w_B), \mu^{(1)} \rangle + c_{A,B}$ *for some constant* $c_{A,B} \geq 0$.

b) $\max_{\mu^{(2)} \in M(B,C)} \langle (w_B, w^*_{B,C}), \mu^{(2)} \rangle \leq \max_{\mu^{(2)} \in L(B,C)} \langle (w_B, w^*_{B,C}, w_B), \mu^{(2)} \rangle + c_{B,C}$ *for some constant* $c_{B,C} \geq 0$.

The following holds:

$$\max_{\mu \in M(A,B,C)} \langle w, \mu \rangle = \max_{\mu^* \in L(A,B) \cap L(B,C)} \langle w, \mu^* \rangle + c_{A,B} + c_{B,C} \tag{47}$$

Proof. Our proof strategy will be to show a dual witness. The following auxiliary observation will be crucial:

$$\max_{\mu \in M(\mathcal{X})} \langle w, \mu \rangle = \max_{y \in \{0,1\}^{|\mathcal{X}|}} \langle w, \mu_y \rangle \tag{48}$$

Where μ_y is defined as in **??**. This has the virtue of turning a continous definition into a discrete one that will prove easier to work with in the future.

Write $w = [w^*_{A,B}, w_B, w^*_{B,C}, 0]$

$$\begin{aligned}
\max_{\mu \in M(A,B,C)} \langle w, \mu \rangle &= \max_{y \in \{0,1\}^{|A|+|B|+|C|}} \langle w, \mu_y \rangle && (49)\\
&\leq \max_{\mu^* \in L(A,B) \cap L(B,C)} \langle w, \mu^* \rangle && (50)\\
&= \max_{\substack{\mu^{(1)} \in L(A,B) \\ \mu^{(2)} \in L(B,C) \\ \mu^{(1)}_B = \mu^{(2)}_B}} \langle w^*_{A,B}, \mu^{(1)}_{A,B} \rangle + \langle w^*_{B,C}, \mu^{(2)}_{B,C} \rangle + \langle \frac{w_B}{2}, \mu^{(1)}_B + \mu^{(2)}_B \rangle && (51)\\
&= \min_{\lambda_B \in \mathbb{R}^{3^{|B|}}} \Big(\max_{\substack{\mu^{(1)} \in M(A,B) \\ \mu^{(2)} \in M(B,C)}} \langle w^*_{A,B}, \mu^{(1)}_{A,B} \rangle + && (52)\\
&\langle w^*_{B,C}, \mu^{(2)}_{B,C} \rangle + \langle \frac{w_B}{2}, \mu^{(1)}_B + \mu^{(2)}_B \rangle + \langle \lambda_B, \mu^{(1)}_B - \mu^{(2)}_B \rangle \Big) + c_{A,B} + c_{B,C} && (53)
\end{aligned}$$

The jump between equation **??** and equation **??** follows by lagrange duality.

$$\max_{\mu^{(1)} \in L(A,B)} \langle w^*_{A,B}, \mu^{(1)}_{A,B} \rangle + \langle \frac{w_B}{2}, \mu^{(1)}_B \rangle + \langle \lambda_B, \mu^{(1)}_B \rangle = c_{A,B} + \max_{\mu^{(1)} \in M(A,B)} \langle w^*_{A,B}, \mu^{(1)}_{A,B} \rangle + \langle \frac{w_B}{2}, \mu^{(1)}_B \rangle + \langle \lambda_B, \mu^{(1)}_B \rangle \tag{54}$$

And similarly for $M(B,C)$ and $L(B,C)$. We can restrict w to specific tightness sets as long as any real number can go into the dimensions w_B of w)

Equation **??** holds by changing the optimization over the polytope into one over its vertices. The first inequality **??** holds because $M(A,B,C) \subseteq L(A,B) \cap L(B,C)$. The second equality **??** holds by definition. The third equality **??** holds by duality. Furthermore, the last expression equals:

$$\begin{aligned}
&\min_{\lambda_B \in \mathbb{R}^{3^{|B|}}} \Big(\max_{\mu^{(1)} \in M(A,B)} \langle w^*_{A,B}, \mu^{(1)}_{A,B} \rangle + \langle \frac{w_B}{2}, \mu^{(1)}_B \rangle + \langle \lambda_B, \mu^{(1)}_B \rangle && (55)\\
&+ \max_{\mu^{(2)} \in M(B,C)} \langle w^*_{B,C}, \mu^{(2)}_{B,C} \rangle + \langle \frac{w_B}{2}, \mu^{(2)}_B \rangle + \langle -\lambda_B, \mu^{(2)}_B \rangle \Big) && (56)
\end{aligned}$$

Similar to equation **??** we can restrict the maximisation over $M(A,B)$ and $M(B,C)$ to their vertices. Equation **??** becomes:

$$\min_{\lambda_B \in \mathbb{R}^{3^{|B|}}} \Big(\max_{a \in \{0,1\}^{|A|}, b \in \{0,1\}^{|B|}} \langle w^*_{A,B}, \mu^{(1)}_{a,b} \rangle + \langle \frac{w_B}{2}, \mu^{(1)}_b \rangle + \langle \lambda_B, \mu^{(1)}_b \rangle + \tag{57}$$

$$\max_{b' \in \{0,1\}^{|B|}, c \in \{0,1\}^{|C|}} \langle w^*_{B,C}, \mu^{(2)}_{b',c} \rangle + \langle \frac{w_B}{2}, \mu^{(2)}_{b'} \rangle + \langle -\lambda_B, \mu^{(2)}_{b'} \rangle \Big) \tag{58}$$

Where we define μ_b, $\mu_{b'}$ as in Equation **??**. $\mu_{a,b}$ and $\mu_{b',c}$ are the remaining dimensions to complete vectors in $\mu_{a \odot b}$ and $\mu_{b' \odot c}$, where $\odot$ denotes concatenation, so that $a \odot b \in \{0,1\}^{|A|+|B|}$ and $b' \odot c \in \{0,1\}^{|B|+|C|}$ are vertices of the marginal polytopes $M(A,B)$ and $M(B,C)$ as defined in Equation **??**.

Now we show that there exists λ_B such that when plugged into this equation it achieves the same value as the left hand side of equation **??**.

Recall that the entries of λ_B are indexed by tuples of the form (S,y) with $S \subseteq B$ and $y \in \{0,1\}^{|S|}$. We will pick a special λ_B, call it $\lambda^{(o)}_B$.

For $b \in \{0,1\}^B$ define:

$$\lambda^{(o)}_B(B,b) = \frac{1}{2} \left(\max_{c \in \{0,1\}^{|C|}} \langle w^*_{B,C}, \mu^{(2)}_{b,c} \rangle - \max_{a \in \{0,1\}^{|A|}} \langle w^*_{A,B}, \mu^{(1)}_{a,b} \rangle \right) \tag{59}$$

And let $\lambda^{(0)}_B(S,y) = 0$ whenever $S \neq B$.

Observe that for any vertex μ of $M(\mathcal{X})$ exactly one entry $\mu(S,y)$ is nonzero when $S = \mathcal{X}$ and y iterates over all assignments in $\{0,1\}^{|X|}$.

This implies that (using the notation above) for $b \in \{0,1\}^{|B|}$, $\langle \lambda^{(o)}_B, \mu^{(1)}_b \rangle = \lambda^{(o)}_B(B,b)$. The same holds for $b' \in \{0,1\}^{|B|}$ and $\mu^{(2)}_{b'}$, $\langle \lambda^{(o)}_B, \mu^{(2)}_{b'} \rangle = \lambda^{(o)}_B(B,b')$

Plugging $\lambda^{(o)}_B$ into the spot reserved for λ_B within the bracketed part of expression **??**, and using the previous observation yields exactly the right hand side of **??**.

This concludes the proof.

□