Barrier Frank-Wolfe for Marginal Inference

Rahul G. Krishnan
Courant Institute
New York University

Simon Lacoste-Julien
INRIA - Sierra Project-Team
École Normale Supérieure, Paris

David Sontag
Courant Institute
New York University

Abstract

We introduce a globally-convergent algorithm for optimizing the tree-reweighted (TRW) variational objective over the marginal polytope. The algorithm is based on the conditional gradient method (Frank-Wolfe) and moves pseudomarginals within the marginal polytope through repeated maximum a posteriori (MAP) calls. This modular structure enables us to leverage black-box MAP solvers (both exact and approximate) for variational inference, and obtains more accurate results than tree-reweighted algorithms that optimize over the local consistency relaxation. Theoretically, we bound the sub-optimality for the proposed algorithm despite the TRW objective having unbounded gradients at the boundary of the marginal polytope. Empirically, we demonstrate the increased quality of results found by tightening the relaxation over the marginal polytope as well as the spanning tree polytope on synthetic and real-world instances.

1 Introduction

Markov random fields (MRFs) are used in many areas of computer science such as vision and speech. Inference in these undirected graphical models is generally intractable. Our work focuses on performing approximate marginal inference by optimizing the Tree Re-Weighted (TRW) objective (Wainwright et al., 2005). The TRW objective is concave, is exact for tree-structured MRFs, and provides an upper bound on the log-partition function.

Fast combinatorial solvers for the TRW objective exist, including Tree-Reweighted Belief Propagation (TRBP) (Wainwright et al., 2005), convergent message-passing based on geometric programming (Globerson and Jaakkola, 2007), and dual decomposition (Jancsary and Matz, 2011). These methods optimize over the set of pairwise consistency constraints, also called the local polytope. Sontag and Jaakkola (2007) showed that significantly better results could be obtained by optimizing over tighter relaxations of the marginal polytope. However, deriving a message-passing algorithm for the TRW objective over tighter relaxations of the marginal polytope is challenging. Instead, Sontag and Jaakkola (2007) use the conditional gradient method (also called Frank-Wolfe) and off-the-shelf linear programming solvers to optimize TRW over the cycle consistency relaxation. Rather than optimizing over the cycle relaxation, Belanger et al. (2013) optimize the TRW objective over the exact marginal polytope. Then, using Frank-Wolfe, the linear minimization performed in the inner loop can be shown to correspond to MAP inference.

The Frank-Wolfe optimization algorithm has seen increasing use in machine learning, thanks in part to its efficient handling of complex constraint sets appearing with structured data (Jaggi, 2013; Lacoste-Julien and Jaggi, 2015). However, applying Frank-Wolfe to variational inference presents challenges that were never resolved in previous work. First, the linear minimization performed in the inner loop is computationally expensive, either requiring repeatedly solving a large linear program, as in Sontag and Jaakkola (2007), or performing MAP inference, as in Belanger et al. (2013). Second, the TRW objective involves entropy terms whose gradients go to infinity near the boundary of the feasible set, therefore existing convergence guarantees for Frank-Wolfe do not apply. Third, variational inference using TRW involves both an outer and inner loop of Frank-Wolfe, where the outer loop optimizes the edge appearance probabilities in the TRW entropy bound to tighten it.

Neither Sontag and Jaakkola (2007) nor Belanger et al. (2013) explore the effect of optimizing over the edge appearance probabilities.

Although MAP inference is in general NP hard (Shimony, 1994), it is often possible to find exact solutions to large real-world instances within reasonable running times (Sontag et al., 2008; Allouche et al., 2010; Kappes et al., 2013). Moreover, as we show in our experiments, even approximate MAP solvers can be successfully used within our variational inference algorithm. As MAP solvers improve in their runtime and performance, their iterative use could become feasible and as a byproduct enable more efficient and accurate marginal inference. Our work provides a fast deterministic alternative to recently proposed Perturb-and-MAP algorithms (Papandreou and Yuille, 2011; Hazan and Jaakkola, 2012; Ermon et al., 2013).

Contributions. This paper makes several theoretical and practical innovations. We propose a modification to the Frank-Wolfe algorithm that optimizes over adaptively chosen contractions of the domain and prove its rate of convergence for functions whose gradients can be unbounded at the boundary. Our algorithm does not require a different oracle than standard Frank-Wolfe and could be useful for other convex optimization problems where the gradient is ill-behaved at the boundary.

We instantiate the algorithm for approximate marginal inference over the marginal polytope with the TRW objective. With an exact MAP oracle, we obtain the first provably convergent algorithm for the optimization of the TRW objective over the marginal polytope, which had remained an open problem to the best of our knowledge. Traditional proof techniques of convergence for first order methods fail as the gradient of the TRW objective is not Lipschitz continuous.

We develop several heuristics to make the algorithm practical: a fully-corrective variant of Frank-Wolfe that reuses previously found integer assignments thereby reducing the need for new (approximate) MAP calls, the use of local search between MAP calls, and significant re-use of computations between subsequent steps of optimizing over the spanning tree polytope. We perform an extensive experimental evaluation on both synthetic and real-world inference tasks.

2 Background

Markov Random Fields: MRFs are undirected probabilistic graphical models where the probability distribution factorizes over cliques in the graph. We consider marginal inference on pairwise MRFs with N random variables $X_1, X_2, \ldots, X_N$ where each variable takes discrete states $x_i \in \text{VAL}_i$. Let $G = (V, E)$ be the Markov network with an undirected edge $\{i, j\} \in E$ for every two variables X_i and X_j that are connected together. Let $\mathcal{N}(i)$ refer to the set of neighbors of variable X_i. We organize the edge log-potentials $\theta_{ij}(x_i, x_j)$ for all possible values of $x_i \in \text{VAL}_i$, $x_j \in \text{VAL}_j$ in the vector $\boldsymbol{\theta}_{ij}$, and similarly for the node log-potential vector $\boldsymbol{\theta}_i$. We regroup these in the overall vector $\vec{\boldsymbol{\theta}}$. We introduce a similar grouping for the marginal vector $\vec{\boldsymbol{\mu}}$: for example, $\mu_i(x_i)$ gives the coordinate of the marginal vector corresponding to the assignment x_i to variable X_i.

Tree Re-weighted Objective (Wainwright et al., 2005): Let $Z(\vec{\boldsymbol{\theta}})$ be the partition function for the MRF and $\mathcal{M}$ be the set of all valid marginal vectors (the marginal polytope). The maximization of the TRW objective gives the following upper bound on the log partition function:

$$\log Z(\vec{\boldsymbol{\theta}}) \leq \min_{\boldsymbol{\rho} \in \mathbb{T}} \max_{\vec{\boldsymbol{\mu}} \in \mathcal{M}} \underbrace{\langle \vec{\boldsymbol{\theta}}, \vec{\boldsymbol{\mu}} \rangle + H(\vec{\boldsymbol{\mu}}; \boldsymbol{\rho})}_{\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})}, \tag{1}$$

where the TRW entropy is:

$$H(\vec{\boldsymbol{\mu}}; \boldsymbol{\rho}) := \sum_{i \in V} (1 - \sum_{j \in \mathcal{N}(i)} \rho_{ij}) H(\boldsymbol{\mu}_i) + \sum_{(ij) \in E} \rho_{ij} H(\boldsymbol{\mu}_{ij}), \quad H(\boldsymbol{\mu}_i) := - \sum_{x_i} \mu_i(x_i) \log \mu_i(x_i). \tag{2}$$

$\mathbb{T}$ is the spanning tree polytope, the convex hull of edge indicator vectors of all possible spanning trees of the graph. Elements of $\boldsymbol{\rho} \in \mathbb{T}$ specify the probability of an edge being present under a specific distribution over spanning trees. $\mathcal{M}$ is difficult to optimize over, and most TRW algorithms optimize over a relaxation called the local consistency polytope $\mathbb{L} \supseteq \mathcal{M}$:

$$\mathbb{L} := \left\{ \vec{\boldsymbol{\mu}} \geq \mathbf{0},\ \textstyle\sum_{x_i} \mu_i(x_i) = 1\ \forall i \in V,\ \sum_{x_i} \mu_{ij}(x_i, x_j) = \mu_j(x_j), \sum_{x_j} \mu_{ij}(x_i, x_j) = \mu_i(x_i)\ \ \forall \{i, j\} \in E \right\}.$$

The TRW objective $\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ is a globally concave function of $\vec{\boldsymbol{\mu}}$ over $\mathbb{L}$, assuming that $\boldsymbol{\rho}$ is obtained from a valid distribution over spanning trees of the graph (i.e. $\boldsymbol{\rho} \in \mathbb{T}$).

Frank-Wolfe (FW) Algorithm: In recent years, the Frank-Wolfe (aka conditional gradient) algorithm has gained popularity in machine learning (Jaggi, 2013) for the optimization of convex

functions over compact domains (denoted $\mathcal{D}$). The algorithm is used to solve $\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$ by iteratively finding a good descent vertex by solving the linear subproblem:

$$\boldsymbol{s}^{(k)} = \arg\min_{\boldsymbol{s}\in\mathcal{D}} \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{s}\rangle \qquad \text{(FW oracle)}, \tag{3}$$

and then taking a convex step towards this vertex: $\boldsymbol{x}^{(k+1)} = (1-\gamma)\boldsymbol{x}^{(k)} + \gamma \boldsymbol{s}^{(k)}$ for a suitably chosen step-size $\gamma \in [0,1]$. The algorithm remains within the feasible set (is projection free), is invariant to affine transformations of the domain, and can be implemented in a memory efficient manner. Moreover, the FW gap $g(\boldsymbol{x}^{(k)}) := \langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{s}^{(k)} - \boldsymbol{x}^{(k)}\rangle$ provides an upper bound on the suboptimality of the iterate $\boldsymbol{x}^{(k)}$. The primal convergence of the Frank-Wolfe algorithm is given by Thm. 1 in Jaggi (2013), restated here for convenience: for $k \geq 1$, the iterates $\boldsymbol{x}^{(k)}$ satisfy:

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \leq \frac{2C_f}{k+2}, \tag{4}$$

where C_f is called the "curvature constant". Under the assumption that ∇f is L-Lipschitz continuous[1] on $\mathcal{D}$, we can bound it as $C_f \leq L \operatorname{diam}_{||.||}(\mathcal{D})^2$.

Marginal Inference with Frank-Wolfe: To optimize $\max_{\vec{\boldsymbol{\mu}}\in\mathcal{M}} \text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ with Frank-Wolfe, the linear subproblem (3) becomes $\arg\max_{\vec{\boldsymbol{\mu}}\in\mathcal{M}} \langle \tilde{\boldsymbol{\theta}}, \vec{\boldsymbol{\mu}}\rangle$, where the perturbed potentials $\tilde{\boldsymbol{\theta}}$ correspond to the gradient of $\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ with respect to $\vec{\boldsymbol{\mu}}$. Elements of $\tilde{\boldsymbol{\theta}}$ are of the form $\theta_c(x_c) + K_c(1 + \log \mu_c(x_c))$, evaluated at the pseudomarginals' current location in $\mathcal{M}$, where K_c is the coefficient of the entropy for the node/edge term in (2). The FW linear subproblem here is thus equivalent to performing MAP inference in a graphical model with potentials $\tilde{\boldsymbol{\theta}}$ (Belanger et al., 2013), as the vertices of the marginal polytope are in 1-1 correspondence with valid joint assignments to the random variables of the MRF, and the solution of a linear program is always achieved at a vertex of the polytope. The TRW objective does not have a Lipschitz continuous gradient over $\mathcal{M}$, and so standard convergence proofs for Frank-Wolfe do not hold.

3 Optimizing over Contractions of the Marginal Polytope

Motivation: We wish to (1) use the fewest possible MAP calls, and (2) avoid regions near the boundary where the unbounded curvature of the function slows down convergence. A viable option to address (1) is through the use of *correction steps*, where after a Frank-Wolfe step, one optimizes over the polytope defined by previously visited vertices of $\mathcal{M}$ (called the fully-corrective Frank-Wolfe (FCFW) algorithm and proven to be linearly convergence for strongly convex objectives (Lacoste-Julien and Jaggi, 2015)). This does not require additional MAP calls. However, we found (see Sec. 5) that when optimizing the TRW objective over $\mathcal{M}$, performing correction steps can surprisingly *hurt* performance. This leaves us in a dilemma: correction steps enable decreasing the objective without additional MAP calls, but they can also slow global progress since iterates after correction sometimes lie close to the boundary of the polytope (where the FW directions become less informative). In a manner akin to barrier methods and to Garber and Hazan (2013)'s local linear oracle, our proposed solution maintains the iterates within a contraction of the polytope. This gives us most of the mileage obtained from performing the correction steps *without* suffering the consequences of venturing too close to the boundary of the polytope. We prove a global convergence rate for the iterates with respect to the true solution over the full polytope.

We describe convergent algorithms to optimize $\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ for $\vec{\boldsymbol{\mu}} \in \mathcal{M}$. The approach we adopt to deal with the issue of unbounded gradients at the boundary is to perform Frank-Wolfe within a contraction of the marginal polytope given by $\mathcal{M}_\delta$ for $\delta \in [0,1]$, with either a fixed δ or an adaptive δ.

Definition 3.1 (Contraction polytope)**.** $\mathcal{M}_\delta := (1-\delta)\mathcal{M} + \delta\, \boldsymbol{u}_0$*, where* $\boldsymbol{u}_0 \in \mathcal{M}$ *is the vector representing the uniform distribution.*

Marginal vectors that lie within $\mathcal{M}_\delta$ are bounded away from zero as all the components of $\boldsymbol{u}_0$ are strictly positive. Denoting $\mathcal{V}^{(\delta)}$ as the set of vertices of $\mathcal{M}_\delta$, $\mathcal{V}$ as the set of vertices of $\mathcal{M}$ and $f(\vec{\boldsymbol{\mu}}) := -\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$, the key insight that enables our novel approach is that:

$$\underbrace{\arg\min_{\boldsymbol{v}^{(\delta)}\in\mathcal{V}^{(\delta)}} \left\langle \nabla f, \boldsymbol{v}^{(\delta)} \right\rangle}_{\textit{(Linear Minimization over } \mathcal{M}_\delta\textit{)}} \equiv \arg\min_{\boldsymbol{v}\in\mathcal{V}} \underbrace{\langle \nabla f, (1-\delta)\boldsymbol{v} + \delta \boldsymbol{u}_0\rangle}_{\textit{(Definition of } \boldsymbol{v}^{(\delta)}\textit{)}} \equiv \underbrace{(1-\delta) \arg\min_{\boldsymbol{v}\in\mathcal{V}} \langle \nabla f, \boldsymbol{v}\rangle + \delta \boldsymbol{u}_0}_{\textit{(Run MAP solver and shift vertex)}}.$$

[1] I.e. $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\|_* \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|$ for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$. Notice that the dual norm $\|\cdot\|_*$ is needed here.

Algorithm 1: Updates to δ after a MAP call (Adaptive δ variant)

1: At iteration k. Assuming $\boldsymbol{x}^{(k)}, \boldsymbol{u}_0, \delta^{(k-1)}, f$ are defined and $\boldsymbol{s}^{(k)}$ has been computed
2: Compute $g(\boldsymbol{x}^{(k)}) = \langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{s}^{(k)} - \boldsymbol{x}^{(k)} \rangle$ *(Compute FW gap)*
3: Compute $g_u(\boldsymbol{x}^{(k)}) = \langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{u}_0 - \boldsymbol{x}^{(k)} \rangle$ *(Compute "uniform gap")*
4: **if** $g_u(\boldsymbol{x}^{(k)}) < 0$ **then**
5: Let $\tilde{\delta} = \frac{g(\boldsymbol{x}^{(k)})}{-4 g_u(\boldsymbol{x}^{(k)})}$ *(Compute new proposal for δ)*
6: **if** $\tilde{\delta} < \delta^{(k-1)}$ **then**
7: $\delta^{(k)} = \min\left(\tilde{\delta}, \frac{\delta^{(k-1)}}{2}\right)$ *(Shrink by at least a factor of two if proposal is smaller)*
8: **end if**
9: **end if** *(and set $\delta^{(k)} = \delta^{(k-1)}$ if it was not updated)*

Therefore, to solve the FW subproblem (3) over $\mathcal{M}_\delta$, we can run as usual a MAP solver and simply shift the resulting vertex of $\mathcal{M}$ towards $\boldsymbol{u}_0$ to obtain a vertex of $\mathcal{M}_\delta$. Our solution to optimize over restrictions of the polytope is more broadly applicable to the optimization problem defined below, with f satisfying Prop. 3.3 (satisfied by the TRW objective) in order to get convergence rates.

Problem 3.2. *Solve $\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$ where $\mathcal{D}$ is a compact convex set and f is convex and continuously differentiable on the relative interior of $\mathcal{D}$.*

Property 3.3. *(Controlled growth of Lipschitz constant over $\mathcal{D}_\delta$). We define $\mathcal{D}_\delta := (1-\delta)\mathcal{D} + \delta\boldsymbol{u}_0$ for a fixed $\boldsymbol{u}_0$ in the relative interior of $\mathcal{D}$. We suppose that there exists a fixed $p \geq 0$ and L such that for any $\delta > 0$, $\nabla f(\boldsymbol{x})$ has a bounded Lipschitz constant $L_\delta \leq L\delta^{-p}$ $\forall \boldsymbol{x} \in \mathcal{D}_\delta$.*

Fixed δ: The first algorithm fixes a value for δ a-priori and performs the optimization over $\mathcal{D}_\delta$. The following theorem bounds the sub-optimality of the iterates with respect to the optimum over $\mathcal{D}$.

Theorem 3.4 (Suboptimality bound for fixed-δ algorithm)**.** *Let f satisfy the properties in Prob. 3.2 and Prop. 3.3, and suppose further that f is finite on the boundary of $\mathcal{D}$. Then the use of Frank-Wolfe for $\min_{\boldsymbol{x}\in\mathcal{D}_\delta} f(\boldsymbol{x})$ realizes a sub-optimality over $\mathcal{D}$ bounded as:*

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \leq \frac{2C_\delta}{(k+2)} + \omega\left(\delta \operatorname{diam}(\mathcal{D})\right),$$

where $\boldsymbol{x}^*$ is the optimal solution in $\mathcal{D}$, $C_\delta \leq L_\delta \operatorname{diam}_{||.||}(\mathcal{D}_\delta)^2$, and ω is the modulus of continuity function of the (uniformly) continuous f (in particular, $\omega(\delta) \downarrow 0$ as $\delta \downarrow 0$).

The full proof is given in App. C. The first term of the bound comes from the standard Frank-Wolfe convergence analysis of the sub-optimality of $\boldsymbol{x}^{(k)}$ relative to $\boldsymbol{x}^*_{(\delta)}$, the optimum over $\mathcal{D}_\delta$, as in (4) and using Prop. 3.3. The second term arises by bounding $f(\boldsymbol{x}^*_{(\delta)}) - f(\boldsymbol{x}^*) \leq f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x}^*)$ with a cleverly chosen $\tilde{\boldsymbol{x}} \in \mathcal{D}_\delta$ (as $\boldsymbol{x}^*_{(\delta)}$ is optimal in $\mathcal{D}_\delta$). We pick $\tilde{\boldsymbol{x}} := (1-\delta)\boldsymbol{x}^* + \delta\boldsymbol{u}_0$ and note that $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\| \leq \delta \operatorname{diam}(\mathcal{D})$. As f is continuous on a compact set, it is uniformly continuous and we thus have $f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq \omega(\delta \operatorname{diam}(\mathcal{D}))$ with ω its modulus of continuity function.

Adaptive δ: The second variant to solve $\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$ iteratively perform FW steps over $\mathcal{D}_\delta$, but also decreases δ adaptively. The update schedule for δ is given in Alg. 1 and is motivated by the convergence proof. The idea is to ensure that the FW gap over $\mathcal{D}_\delta$ is always at least half the FW gap over $\mathcal{D}$, relating the progress over $\mathcal{D}_\delta$ with the one over $\mathcal{D}$. It turns out that FW-gap-$\mathcal{D}_\delta$ $=$ $(1-\delta)$FW-gap-$\mathcal{D} + \delta \cdot g_u(\boldsymbol{x}^{(k)})$, where the "uniform gap" $g_u(\boldsymbol{x}^{(k)})$ quantifies the decrease of the function when contracting towards $\boldsymbol{u}_0$. When $g_u(\boldsymbol{x}^{(k)})$ is negative and large compared to the FW gap, we need to shrink δ (see step 5 in Alg. 1) to ensure that the δ-modified direction is a sufficient descent direction. We can show that the algorithm converges to the global solution as follows:

Theorem 3.5 (Global convergence for adaptive-δ variant over $\mathcal{D}$)**.** *For a function f satisfying the properties in Prob. 3.2 and Prop. 3.3, the sub-optimality of the iterates obtained by running the FW updates over $\mathcal{D}_\delta$ with δ updated according to Alg. 1 is bounded as:*

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \leq O\left(k^{-\frac{1}{p+1}}\right).$$

A full proof with a precise rate and constants is given in App. D. The sub-optimality $h_k := f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*)$ traverses three stages with an overall rate as above. The updates to $\delta^{(k)}$ as in Alg. 1 enable us

Algorithm 2: Approximate marginal inference over $\mathcal{M}$ (solving (1)). Here f is the negative TRW objective.

1: Function **TRW-Barrier-FW**($\boldsymbol{\rho}^{(0)}, \epsilon, \delta^{(\text{init})}, \boldsymbol{u}_0$):
2: **Inputs:** Edge-appearance probabilities $\boldsymbol{\rho}^{(0)}$, $\delta^{(\text{init})} \leq \frac{1}{4}$ initial contraction of polytope, inner loop stopping criterion ϵ, fixed reference point $\boldsymbol{u}_0$ in the interior of $\mathcal{M}$. Let $\delta^{(-1)} = \delta^{(\text{init})}$.
3: Let $V := \{\boldsymbol{u}_0\}$ (visited vertices), $\boldsymbol{x}^{(0)} = \boldsymbol{u}_0$ (Initialize the algorithm at the uniform distribution)
4: **for** $i = 0 \ldots$ **MAX_RHO_ITS do** {*FW outer loop to optimize $\boldsymbol{\rho}$ over $\mathbb{T}$*}
5: **for** $k = 0 \ldots$ **MAXITS do** {*FCFW inner loop to optimize $\boldsymbol{x}$ over $\mathcal{M}$*}
6: Let $\tilde{\theta} = \nabla f(\boldsymbol{x}^{(k)}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho}^{(i)})$ *(Compute gradient)*
7: Let $\boldsymbol{s}^{(k)} \in \arg\min_{\boldsymbol{v} \in \mathcal{M}} \langle \tilde{\theta}, \boldsymbol{v} \rangle$ *(Run MAP solver to compute FW vertex)*
8: Compute $g(\boldsymbol{x}^{(k)}) = \langle -\tilde{\theta}, \boldsymbol{s}^{(k)} - \boldsymbol{x}^{(k)} \rangle$ *(Inner loop FW duality gap)*
9: **if** $g(\boldsymbol{x}^{(k)}) \leq \epsilon$ **then**
10: **break** FCFW inner loop *($\boldsymbol{x}^{(k)}$ is ϵ-optimal)*
11: **end if**
12: $\delta^{(k)} = \delta^{(k-1)}$ *(For Adaptive-δ: Run Alg. 1 to modify δ)*
13: Let $\boldsymbol{s}^{(k)}_{(\delta)} = (1 - \delta^{(k)})\boldsymbol{s}^{(k)} + \delta^{(k)}\boldsymbol{u}_0$ and $\boldsymbol{d}^{(k)}_{(\delta)} = \boldsymbol{s}^{(k)}_{(\delta)} - \boldsymbol{x}^{(k)}$ *(δ-contracted quantities)*
14: $\boldsymbol{x}^{(k+1)} = \arg\min\{f(\boldsymbol{x}^{(k)} + \gamma\, \boldsymbol{d}^{(k)}_{(\delta)}) : \gamma \in [0, 1]\}$ *(FW step with line search)*
15: Update correction polytope: $V := V \cup \{\boldsymbol{s}^{(k)}\}$
16: $\boldsymbol{x}^{(k+1)} :=$ **CORRECTION**($\boldsymbol{x}^{(k+1)}, V, \delta^{(k)}, \boldsymbol{\rho}^{(i)}$) *(optional: correction step)*
17: $\boldsymbol{x}^{(k+1)}, V_{\text{search}} :=$ **LOCALSEARCH**($\boldsymbol{x}^{(k+1)}, \boldsymbol{s}^{(k)}, \delta^{(k)}, \boldsymbol{\rho}^{(i)}$) *(optional: fast MAP solver)*
18: Update correction polytope (with vertices from **LOCALSEARCH**): $V := V \cup \{V_{\text{search}}\}$
19: **end for**
20: $\boldsymbol{\rho}^{v} \leftarrow$ **minSpanTree**(**edgesMI**($\boldsymbol{x}^{(k)}$)) *(FW vertex of the spanning tree polytope)*
21: $\boldsymbol{\rho}^{(i+1)} \leftarrow \boldsymbol{\rho}^{(i)} + (\frac{i}{i+2})(\boldsymbol{\rho}^{v} - \boldsymbol{\rho}^{(i)})$ *(Fixed step-size schedule FW update for $\boldsymbol{\rho}$ kept in $\text{relint}(\mathbb{T})$)*
22: $\boldsymbol{x}^{(0)} \leftarrow \boldsymbol{x}^{(k)}$, $\delta^{(-1)} \leftarrow \delta^{(k-1)}$ *(Re-initialize for FCFW inner loop)*
23: If $i <$ **MAX_RHO_ITS** then $\boldsymbol{x}^{(0)} =$ **CORRECTION**($\boldsymbol{x}^{(0)}, V, \delta^{(-1)}, \boldsymbol{\rho}^{(i+1)}$)
24: **end for**
25: **return** $\boldsymbol{x}^{(0)}$ and $\boldsymbol{\rho}^{(i)}$

to (1) upper bound the duality gap over $\mathcal{D}$ as a function of the duality gap in $\mathcal{D}_\delta$ and (2) lower bound the value of $\delta^{(k)}$ as a function of h_k. Applying the standard Descent Lemma with the Lipschitz constant on the gradient of the form $L\delta^{-p}$ (Prop. 3.3), and replacing $\delta^{(k)}$ by its bound in h_k, we get the recurrence: $h_{k+1} \leq h_k - Ch_k^{p+2}$. Solving this gives us the desired bound.

Application to the TRW Objective: $\min_{\vec{\boldsymbol{\mu}} \in \mathcal{M}} -\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ is akin to $\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$ and the (strong) convexity of $-\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ has been previously shown (Wainwright et al., 2005; London et al., 2015). The gradient of the TRW objective is Lipschitz continuous over $\mathcal{M}_\delta$ since all marginals are strictly positive. Its growth for Prop. 3.3 can be bounded with $p = 1$ as we show in App. E.1. This gives a rate of convergence of $O(k^{-1/2})$ for the adaptive-δ variant, which interestingly is a typical rate for non-smooth convex optimization. The hidden constant is of the order $O(\|\theta\| \cdot |V|)$. The modulus of continuity ω for the TRW objective is close to linear (it is almost a Lipschitz function), and its constant is instead of the order $O(\|\theta\| + |V|)$.

4 Algorithm

Alg. 2 describes the pseudocode for our proposed algorithm to do marginal inference with $\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$. **minSpanTree** finds the minimum spanning tree of a weighted graph, and **edgesMI**($\vec{\boldsymbol{\mu}}$) computes the mutual information of edges of G from the pseudomarginals in $\vec{\boldsymbol{\mu}}$[2] (to perform FW updates over $\boldsymbol{\rho}$ as in Alg. 2 in Wainwright et al. (2005)). It is worthwhile to note that our approach uses three levels of Frank-Wolfe: (1) for the (tightening) optimization of $\boldsymbol{\rho}$ over $\mathbb{T}$, (2) to perform approximate marginal inference, i.e for the optimization of $\vec{\boldsymbol{\mu}}$ over $\mathcal{M}$, and (3) to perform the correction steps (lines 16 and 23). We detail a few heuristics that aid practicality.

Fast Local Search: Fast methods for MAP inference such as Iterated Conditional Modes (Besag, 1986) offer a cheap, low cost alternative to a more expensive combinatorial MAP solver. We

[2]The component ij has value $H(\boldsymbol{\mu}_i) + H(\boldsymbol{\mu}_j) - H(\boldsymbol{\mu}_{ij})$.

warm start the ICM solver with the last found vertex $\boldsymbol{s}^{(k)}$ of the marginal polytope. The subroutine **LOCALSEARCH** (Alg. 6 in Appendix) performs a fixed number of FW updates to the pseudo-marginals using ICM as the (approximate) MAP solver.

Re-optimizing over the Vertices of $\mathcal{M}$ (FCFW algorithm): As the iterations of FW progress, we keep track of the vertices of the marginal polytope found by Alg. 2 in the set V. We make use of these vertices in the **CORRECTION** subroutine (Alg. 5 in Appendix) which re-optimizes the objective function over (a contraction of) the convex hull of the elements of V (called the correction polytope). $\boldsymbol{x}^{(0)}$ in Alg. 2 is initialized to the uniform distribution which is guaranteed to be in $\mathcal{M}$ (and $\mathcal{M}_\delta$). After updating $\boldsymbol{\rho}$, we set $\boldsymbol{x}^{(0)}$ to the approximate minimizer in the correction polytope. The intuition is that changing $\boldsymbol{\rho}$ by a small amount may not substantially modify the optimal $\boldsymbol{x}^*$ (for the new $\boldsymbol{\rho}$) and that the new optimum might be in the convex hull of the vertices found thus far. If so, **CORRECTION** will be able to find it without resorting to any additional MAP calls. This encourages the MAP solver to search for new, unique vertices instead of rediscovering old ones.

Approximate MAP Solvers: We can swap out the exact MAP solver with an approximate MAP solver. The primal objective plus the (approximate) duality gap may no longer be an upper bound on the log-partition function (black-box MAP solvers could be considered to optimize over an inner bound to the marginal polytope). Furthermore, the gap over $\mathcal{D}$ may be negative if the approximate MAP solver fails to find a direction of descent. Since adaptive-δ requires that the gap be positive in Alg. 1, we take the max over the last gap obtained over the correction polytope (which is always non-negative) and the computed gap over $\mathcal{D}$ as a heuristic.

Theoretically, one could get similar convergence rates as in Thm. 3.4 and 3.5 using an approximate MAP solver that has a multiplicative guarantee on the gap (line 8 of Alg. 2), as was done previously for FW-like algorithms (see, e.g., Thm. C.1 in Lacoste-Julien et al. (2013)). With an ϵ-additive error guarantee on the MAP solution, one can prove similar rates up to a suboptimality error of ϵ. Even if the approximate MAP solver does not provide an approximation guarantee, if it returns an *upper bound* on the value of the MAP assignment (as do branch-and-cut solvers for integer linear programs, or Sontag et al. (2008)), one can use this to obtain an upper bound on $\log Z$ (see App. J).

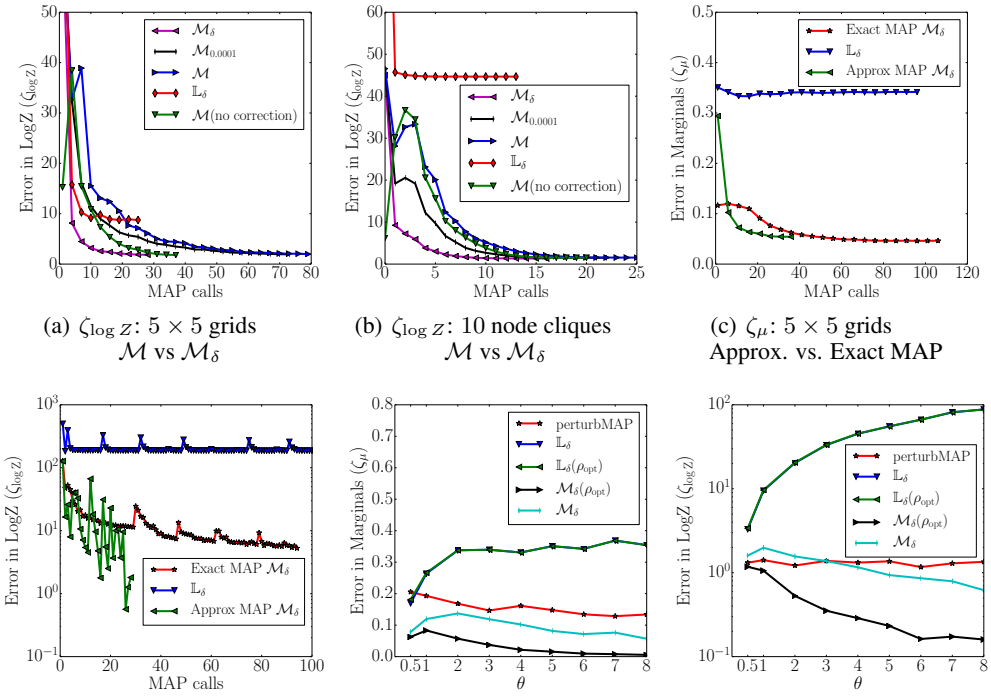
5 Experimental Results

Setup: The L1 error in marginals is computed as: $\zeta_\mu := \frac{1}{N}\sum_{i=1}^{N}|\mu_i(1) - \mu_i^*(1)|$. When using exact MAP inference, the error in $\log Z$ (denoted $\zeta_{\log Z}$) is computed by adding the duality gap to the primal (since this guarantees us an upper bound). For approximate MAP inference, we plot the primal objective. We use a non-uniform initialization of $\boldsymbol{\rho}$ computed with the Matrix Tree Theorem (Sontag and Jaakkola, 2007; Koo et al., 2007). We perform 10 updates to $\boldsymbol{\rho}$, optimize $\vec{\boldsymbol{\mu}}$ to a duality gap of 0.5 on $\mathcal{M}$, and always perform correction steps. We use **LOCALSEARCH** only for the real-world instances. We use the implementation of TRBP and the Junction Tree Algorithm (to compute exact marginals) in libDAI (Mooij, 2010). Unless specified, we compute marginals by optimizing the TRW objective using the adaptive-δ variant of the algorithm (denoted in the figures as M_δ).

MAP Solvers: For approximate MAP, we run three solvers in parallel: QPBO (Kolmogorov and Rother, 2007; Boykov and Kolmogorov, 2004), TRW-S (Kolmogorov, 2006) and ICM (Besag, 1986) using OpenGM (Andres et al., 2012) and use the result that realizes the highest energy. For exact inference, we use Gurobi Optimization (2015) or toulbar2 (Allouche et al., 2010).

Test Cases: All of our test cases are on binary pairwise MRFs. (1) *Synthetic 10 nodes cliques*: Same setup as Sontag and Jaakkola (2007, Fig. 2), with 9 sets of 100 instances each with coupling strength drawn from $\mathcal{U}[-\theta, \theta]$ for $\theta \in \{0.5, 1, 2, \ldots, 8\}$. (2) *Synthetic Grids*: 15 trials with 5×5 grids. We sample $\theta_i \sim \mathcal{U}[-1, 1]$ and $\theta_{ij} \in [-4, 4]$ for nodes and edges. The potentials were $(-\theta_i, \theta_i)$ for nodes and $(\theta_{ij}, -\theta_{ij}; -\theta_{ij}, \theta_{ij})$ for edges. (3) *Restricted Boltzmann Machines (RBMs)*: From the Probabilistic Inference Challenge 2011.[3] (4) *Horses*: Large ($N \approx 12000$) MRFs representing images from the Weizmann Horse Data (Borenstein and Ullman, 2002) with potentials learned by Domke (2013). (5) *Chinese Characters*: An image completion task from the KAIST Hanja2 database, compiled in OpenGM by Andres et al. (2012). The potentials were learned using Decision Tree Fields (Nowozin et al., 2011). The MRF is not a grid due to skip edges that tie nodes at various offsets. The potentials are a combination of submodular and supermodular and therefore a harder task for inference algorithms.

[3] http://www.cs.huji.ac.il/project/PASCAL/index.php



(a) $\zeta_{\log Z}$: 5×5 grids $\mathcal{M}$ vs $\mathcal{M}_\delta$

(b) $\zeta_{\log Z}$: 10 node cliques $\mathcal{M}$ vs $\mathcal{M}_\delta$

(c) ζ_μ: 5×5 grids Approx. vs. Exact MAP

(d) $\zeta_{\log Z}$: 40 node RBM Approx. vs. Exact MAP

(e) ζ_μ: 10 node cliques Optimization over $\mathbb{T}$

(f) $\zeta_{\log Z}$: 10 node cliques Optimization over $\mathbb{T}$

Figure 1: Synthetic Experiments: In Fig. 1(c) & 1(d), we unravel MAP calls across updates to $\boldsymbol{\rho}$. Fig. 1(d) corresponds to a single RBM (not an aggregate over trials) where for "Approx MAP" we plot the absolute error between the primal objective and $\log Z$ (not guaranteed to be an upper bound).

On the Optimization of $\mathcal{M}$ versus $\mathcal{M}_\delta$

We compare the performance of Alg. 2 on optimizing over $\mathcal{M}$ (with and without correction), optimizing over $\mathcal{M}_\delta$ with fixed-$\delta = 0.0001$ (denoted $M_{0.0001}$) and optimizing over $\mathcal{M}_\delta$ using the adaptive-δ variant. These plots are averaged across all the trials for the *first* iteration of optimizing over $\mathbb{T}$. We show error as a function of the number of MAP calls since this is the bottleneck for large MRFs. Fig. 1(a), 1(b) depict the results of this optimization aggregated across trials. We find that all variants settle on the same average error. The adaptive δ variant converges faster on average followed by the fixed δ variant. Despite relatively quick convergence for $\mathcal{M}$ with no correction on the grids, we found that correction was crucial to reducing the number of MAP calls in subsequent steps of inference after updates to $\boldsymbol{\rho}$. As highlighted earlier, correction steps on $\mathcal{M}$ (in blue) worsen convergence, an effect brought about by iterates wandering too close to the boundary of $\mathcal{M}$.

On the Applicability of Approximate MAP Solvers

Synthetic Grids: Fig. 1(c) depicts the accuracy of approximate MAP solvers versus exact MAP solvers aggregated across trials for 5×5 grids. The results using approximate MAP inference are competitive with those of exact inference, even as the optimization is tightened over $\mathbb{T}$. This is an encouraging and non-intuitive result since it indicates that one can achieve high quality marginals through the use of relatively cheaper approximate MAP oracles.

RBMs: As in Salakhutdinov (2008), we observe for RBMs that the bound provided by $\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ over $\mathbb{L}_\delta$ is loose and does not get better when optimizing over $\mathbb{T}$. As Fig. 1(d) depicts for a single RBM, optimizing over $\mathcal{M}_\delta$ realizes significant gains in the upper bound on $\log Z$ which improves with updates to $\boldsymbol{\rho}$. The gains are preserved with the use of the approximate MAP solvers. Note that there are also fast approximate MAP solvers specifically for RBMs (Wang et al., 2014).

Horses: See Fig. 2 (right). The models are close to submodular and the local relaxation is a good approximation to the marginal polytope. Our marginals are visually similar to those obtained by TRBP and our algorithm is able to scale to large instances by using approximate MAP solvers.

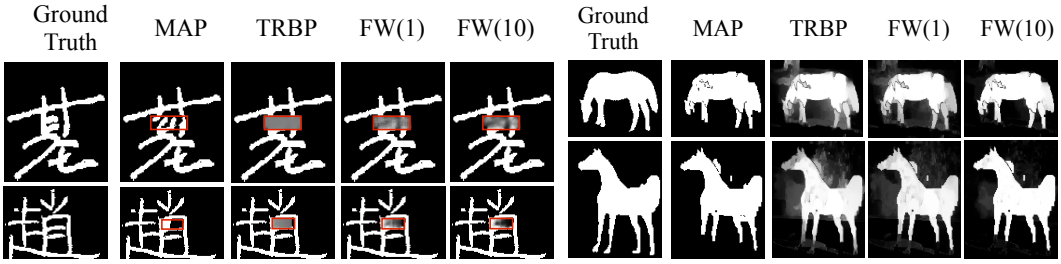


Figure 2: Results on real world test cases. FW(i) corresponds to the final marginals at the ith iteration of optimizing $\boldsymbol{\rho}$. The area highlighted on the Chinese Characters depicts the region of uncertainty.

On the Importance of Optimizing over $\mathbb{T}$

Synthetic Cliques: In Fig. 1(e), 1(f), we study the effect of tightening over $\mathbb{T}$ against coupling strength θ. We consider the ζ_μ and $\zeta_{\log Z}$ obtained for the final marginals before updating $\boldsymbol{\rho}$ (step 19) and compare to the values obtained after optimizing over $\mathbb{T}$ (marked with $\boldsymbol{\rho}_{opt}$). The optimization over $\mathbb{T}$ has little effect on TRW optimized over $\mathbb{L}_\delta$. For optimization over $\mathcal{M}_\delta$, updating $\boldsymbol{\rho}$ realizes better marginals and bound on $\log Z$ (over and above those obtained in Sontag and Jaakkola (2007)).

Chinese Characters: Fig. 2 (left) displays marginals across iterations of optimizing over $\mathbb{T}$. The submodular and supermodular potentials lead to frustrated models for which $\mathbb{L}_\delta$ is very loose, which results in TRBP obtaining poor results.[4] Our method produces reasonable marginals even before the first update to $\boldsymbol{\rho}$, and these improve with tightening over $\mathbb{T}$.

Related Work for Marginal Inference with MAP Calls

Hazan and Jaakkola (2012) estimate $\log Z$ by averaging MAP estimates obtained on randomly perturbed inflated graphs. Our implementation of the method performed well in approximating $\log Z$ but the marginals (estimated by fixing the value of each random variable and estimating $\log Z$ for the resulting graph) were less accurate than our method (Fig. 1(e), 1(f)).

6 Discussion

We introduce the first provably convergent algorithm for the TRW objective over the marginal polytope, under the assumption of exact MAP oracles. We quantify the gains obtained both from marginal inference over $\mathcal{M}$ *and* from tightening over the spanning tree polytope. We give heuristics that improve the scalability of Frank-Wolfe when used for marginal inference. The runtime cost of iterative MAP calls (a reasonable rule of thumb is to assume an approximate MAP call takes roughly the same time as a run of TRBP) is worthwhile particularly in cases such as the Chinese Characters where $\mathbb{L}$ is loose. Specifically, our algorithm is appropriate for domains where marginal inference is hard but there exist efficient MAP solvers capable of handling non-submodular potentials. Code is available at `https://github.com/clinicalml/fw-inference`.

Our work creates a flexible, modular framework for optimizing a broad class of variational objectives, not simply TRW, with guarantees of convergence. We hope that this will encourage more research on building better entropy approximations. The framework we adopt is more generally applicable to optimizing functions whose gradients tend to infinity at the boundary of the domain.

Our method to deal with gradients that diverge at the boundary bears resemblance to barrier functions used in interior point methods insofar as they bound the solution away from the constraints. Iteratively decreasing δ in our framework can be compared to decreasing the strength of the barrier, enabling the iterates to get closer to the facets of the polytope, although its worthwhile to note that we have an *adaptive* method of doing so.

Acknowledgements

RK and DS gratefully acknowledge the support of the Defense Advanced Research Projects Agency (DARPA) Probabilistic Programming for Advancing Machine Learning (PPAML) Program under

[4] We run TRBP for 1000 iterations using damping = 0.9; the algorithm converges with a max norm difference between consecutive iterates of 0.002. Tightening over $\mathbb{T}$ did not significantly change the results of TRBP.

Air Force Research Laboratory (AFRL) prime contract no. FA8750-14-C-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA, AFRL, or the US government.

References

D. Allouche, S. de Givry, and T. Schiex. Toulbar2, an open source exact cost function network solver, 2010.

B. Andres, B. T., and J. H. Kappes. Opengm: A c++ library for discrete graphical models, June 2012.

D. Belanger, D. Sheldon, and A. McCallum. Marginal inference in MRFs using Frank-Wolfe. *NIPS Workshop on Greedy Optimization, Frank-Wolfe and Friends*, 2013.

J. Besag. On the statistical analysis of dirty pictures. *J R Stat Soc Series B*, 1986.

E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*, 2002.

Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *TPAMI*, 2004.

J. Domke. Learning graphical model parameters with approximate marginal inference. *TPAMI*, 2013.

S. Ermon, C. P. Gomes, A. Sabharwal, and B. Selman. Taming the curse of dimensionality: Discrete integration by hashing and optimization. In *ICML*, 2013.

D. Garber and E. Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *arXiv preprint arXiv:1301.4666*, 2013.

A. Globerson and T. Jaakkola. Convergent propagation algorithms via oriented trees. In *UAI*, 2007.

I. Gurobi Optimization. Gurobi optimizer reference manual, 2015.

T. Hazan and T. Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *ICML*, 2012.

M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

J. Jancsary and G. Matz. Convergent decomposition solvers for tree-reweighted free energies. In *AISTATS*, 2011.

J. Kappes et al. A comparative study of modern inference techniques for discrete energy minimization problems. In *CVPR*, 2013.

V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *TPAMI*, 2006.

V. Kolmogorov and C. Rother. Minimizing nonsubmodular functions with graph cuts-a review. *TPAMI*, 2007.

T. Koo, A. Globerson, X. Carreras, and M. Collins. Structured prediction models via the matrix-tree theorem. In *EMNLP-CoNLL*, 2007.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *NIPS*, 2015.

S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, 2013.

B. London, B. Huang, and L. Getoor. The benefits of learning with strongly convex approximate inference. In *ICML*, 2015.

J. M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *JMLR*, 2010.

S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. In *ICCV*, 2011.

G. Papandreou and A. Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *ICCV*, 2011.

R. Salakhutdinov. Learning and evaluating boltzmann machines. Technical report, 2008.

S. Shimony. Finding MAPs for Belief Networks is NP-hard. *Artificial Intelligence*, 1994.

D. Sontag and T. Jaakkola. New outer bounds on the marginal polytope. In *NIPS*, 2007.

D. Sontag, T. Meltzer, A. Globerson, Y. Weiss, and T. Jaakkola. Tightening LP relaxations for MAP using message-passing. In *UAI*, 2008.

M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 2005.

S. Wang, R. Frostig, P. Liang, and C. Manning. Relaxations for inference in restricted Boltzmann machines. In *ICLR Workshop*, 2014.

A Preliminaries

A.1 Summary of Supplementary Material

The supplementary material is divided into two parts:

(1) The first part is dedicated to the exposition of the theoretical results presented in the main paper. Section B details the variants of the Frank-Wolfe algorithm that we used and analyzed. Section C gives the proof to Theorem 3.4 (fixed δ) while Section D gives the proof to Theorem 3.5 (adaptive δ). Finally, Section E applies the convergence theorem to the TRW objective and investigates the relevant constants.

(2) The remainder of the supplementary material provides more information about the experimental setup as well as additional experimental results.

A.2 Descent Lemma

The following descent lemma is proved in Bertsekas (1999) (Prop. A24) and is standard for any convergence proof of first order methods. We provide a proof here for completeness. It also highlights the origin of the requirement that we use dual norm pairings between $\boldsymbol{x}$ and the gradient of $f(\boldsymbol{x})$ (because of the generalized Cauchy-Schwartz inequality).

Lemma A.1. ***Descent Lemma***

Let $\boldsymbol{x}_\gamma := \boldsymbol{x} + \gamma \boldsymbol{d}$ and suppose that f is continuously differentiable on the line segment from $\boldsymbol{x}$ to $\boldsymbol{x}_{\gamma_{\max}}$ for some $\gamma_{\max} > 0$. Suppose that $L = \sup_{\alpha \in]0, \gamma_{\max}]} \frac{||\nabla f(\boldsymbol{x}+\alpha \boldsymbol{d}) - \nabla f(\boldsymbol{x})||_*}{||\alpha \boldsymbol{d}||}$ is finite, then we have:

$$f(\boldsymbol{x}_\gamma) \leq f(\boldsymbol{x}) + \gamma \langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + \frac{\gamma^2}{2} L ||\boldsymbol{d}||^2, \quad \forall \gamma \in [0, \gamma_{\max}]. \tag{5}$$

Proof. Let $0 < \gamma \leq \gamma_{\max}$. Denoting $l(\alpha) = f(\boldsymbol{x} + \alpha \boldsymbol{d})$, we have that:

$$\begin{aligned}
f(\boldsymbol{x}_\gamma) - f(\boldsymbol{x}) &= l(\gamma) - l(0) \\
&= \int_0^\gamma \nabla_\alpha l(\alpha) d\alpha \\
&= \int_0^\gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x} + \alpha \boldsymbol{d}) \rangle d\alpha \\
&= \int_0^\gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle d\alpha + \int_0^\gamma \boldsymbol{d}^T (\nabla f(\boldsymbol{x} + \alpha \boldsymbol{d}) - \nabla f(\boldsymbol{x})) d\alpha \\
&\leq \int_0^\gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle d\alpha + \left| \int_0^\gamma \boldsymbol{d}^T (\nabla f(\boldsymbol{x} + \alpha \boldsymbol{d}) - \nabla f(\boldsymbol{x})) \right| d\alpha \\
&\leq \int_0^\gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle d\alpha + \int_0^\gamma ||\boldsymbol{d}|| \;\; ||\nabla f(\boldsymbol{x} + \alpha \boldsymbol{d}) - \nabla f(\boldsymbol{x})||_* \, d\alpha \\
&= \gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle + \int_0^\gamma ||\boldsymbol{d}|| \;\; \frac{||\nabla f(\boldsymbol{x} + \alpha \boldsymbol{d}) - \nabla f(\boldsymbol{x})||_*}{\alpha ||\boldsymbol{d}||} \alpha ||\boldsymbol{d}|| \, d\alpha \\
&\leq \gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle + \int_0^\gamma ||\boldsymbol{d}|| \;\; L ||\boldsymbol{d}|| \alpha \, d\alpha \\
&= \gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle + \frac{L}{2} \gamma^2 ||\boldsymbol{d}||^2
\end{aligned}$$

Rearranging terms, we get the desired bound. □

B Frank-Wolfe Algorithms

In this section, we present the various algorithms that we use to do fully corrective Frank-Wolfe (FCFW) with adaptive contractions over the domain $\mathcal{D}$, as was done in our experiments.

B.1 Overview of the Modified Frank-Wolfe Algorithm (FW with Away Steps)

To implement the approximate correction steps in the fully corrective Frank-Wolfe (FCFW) algorithm, we use the Frank-Wolfe algorithm with away steps (Wolfe, 1970), also known as the modified Frank-Wolfe (MFW) algorithm (Guélat and Marcotte, 1986). We give pseudo-code for MFW in Algorithm 3 (taken from (Lacoste-Julien and Jaggi, 2015)). This variant of Frank-Wolfe adds the possibility to do an "away step" (see step 5 in Algorithm 3) in order to avoid the zig zagging phenomenon that slows down Frank-Wolfe when the solution is close to the boundary of the polytope. For a strongly convex objective (with Lipschitz continuous gradient), the MFW was known to have asymptotic linear convergence (Guélat and Marcotte, 1986) and its global linear convergence rate was shown recently (Lacoste-Julien and Jaggi, 2015), accelerating the slow general sublinear rate of Frank-Wolfe. When performing a correction over the convex hull over a (somewhat small) set of vertices of $\mathcal{D}_\delta$, this convergence difference was quite significant in our experiments (MFW converging in a small number of iterations to do an approximate correction vs. FW taking hundreds of iterations to reach a similar level of accuracy). We note that the TRW objective is strongly convex when all the edge probabilities are non-zero (Wainwright et al., 2005); and that it has Lipschitz gradient over $\mathcal{D}_\delta$ (but not $\mathcal{D}$).

The gap computed in step 6 of Algorithm 3 is non-standard; it is a sufficient condition to ensure the global linear convergence of the outer FCFW algorithm when using Algorithm 3 as a subroutine to implement the approximate correction step. See Lacoste-Julien and Jaggi (2015) for more details.

The MFW algorithm requires more bookkeeping than standard FW: in addition to the current iterate $\boldsymbol{x}^{(k)}$, it also maintains both the active set $\mathcal{S}^{(k)}$ (to search for the "away vertex") as well as the barycentric coordinates $\boldsymbol{\alpha}^{(k)}$ (to know what are the away step-sizes that ensure feasibility – see step 13) i.e. $\boldsymbol{x}^{(k)} = \sum_{\boldsymbol{v} \in \mathcal{S}^{(k)}} \alpha_{\boldsymbol{v}}^{(k)} \boldsymbol{v}$.

Algorithm 3: Modified Frank-Wolfe algorithm (FW with Away Steps) – used for approximate correction

1: Function **MFW**$(\boldsymbol{x}^{(0)}, \boldsymbol{\alpha}^{(0)}, \mathcal{V}, \epsilon)$ to optimize over $\text{conv}(\mathcal{V})$:
2: **Inputs:** Set of atoms $\mathcal{V}$, starting point $\boldsymbol{x}^{(0)} = \sum_{\boldsymbol{v} \in \mathcal{S}^{(0)}} \alpha_{\boldsymbol{v}}^{(0)} \boldsymbol{v}$ where $\mathcal{S}^{(0)}$ is active set and $\boldsymbol{\alpha}^{(0)}$ the active coordinates, stopping criterion ϵ.
3: **for** $k = 0 \dots K$ **do**
4: Let $\boldsymbol{s}_k \in \arg\min_{\boldsymbol{v} \in \mathcal{V}} \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{v} \rangle$ and $\boldsymbol{d}_k^{\text{FW}} := \boldsymbol{s}_k - \boldsymbol{x}^{(k)}$ *(the FW direction)*
5: Let $\boldsymbol{v}_k \in \arg\max_{\boldsymbol{v} \in \mathcal{S}^{(k)}} \left\langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{v} \right\rangle$ and $\boldsymbol{d}_k^{\text{A}} := \boldsymbol{x}^{(k)} - \boldsymbol{v}_k$ *(the away direction)*
6: $g_k^{\text{pFW}} := \left\langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{d}_k^{\text{FW}} + \boldsymbol{d}_k^{\text{A}} \right\rangle$ *(stringent gap is FW + away gap to work better for FCFW)*
7: **if** $g_k^{\text{pFW}} \leq \epsilon$ **then**
8: **return** $\boldsymbol{x}^{(k)}, \boldsymbol{\alpha}^{(k)}, \mathcal{S}^{(k)}$.
9: **else**
10: **if** $\left\langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{d}_k^{\text{FW}} \right\rangle \geq \left\langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{d}_k^{\text{A}} \right\rangle$ **then**
11: $\boldsymbol{d}_k := \boldsymbol{d}_k^{\text{FW}}$, and $\gamma_{\max} := 1$ *(choose the FW direction)*
12: **else**
13: $\boldsymbol{d}_k := \boldsymbol{d}_k^{\text{A}}$, and $\gamma_{\max} := \frac{\alpha_{\boldsymbol{v}_k}}{(1 - \alpha_{\boldsymbol{v}_k})}$ *(choose away direction; maximum feasible step-size)*
14: **end if**
15: Line-search: $\gamma_k \in \arg\min_{\gamma \in [0, \gamma_{\max}]} f\left(\boldsymbol{x}^{(k)} + \gamma \boldsymbol{d}_k\right)$
16: Update $\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} + \gamma_k \boldsymbol{d}_k$
17: Update coordinates $\boldsymbol{\alpha}^{(k+1)}$ accordingly (see Lacoste-Julien and Jaggi (2015)).
18: Update $\mathcal{S}^{(k+1)} := \{\boldsymbol{v} \; s.t. \; \alpha_{\boldsymbol{v}}^{(k+1)} > 0\}$
19: **end if**
20: **end for**

B.2 Fully Corrective Frank-Wolfe (FCFW) with Adaptive-δ

We give in Algorithm 4 the pseudo-code to perform fully corrective Frank-Wolfe optimization over $\mathcal{D}$ by iteratively optimizing over $\mathcal{D}_\delta$ with adaptive-δ updates. If δ is kept constant (skipping step 10), then Algorithm 4 implements the fixed δ variant over $\mathcal{D}_\delta$. We describe the algorithm as maintaining the correction set of atoms $V^{(k+1)}$ over $\mathcal{D}$ (rather than $\mathcal{D}_\delta$), as δ is constantly changing. One can easily move back and forth between $V^{(k+1)}$ and its contraction $V_\delta = (1 - \delta^{(k)})V^{(k+1)} + \delta^{(k)}\boldsymbol{u}_0$, and so we note that an efficient implementation might work with either representation cheaply (for example, by storing only $V^{(k+1)}$ and δ, not the perturbed version of the correction polytope). The approximate correction over V_δ is implemented using the MFW algorithm described in Algorithm 3, which requires a barycentric representation $\boldsymbol{\alpha}^{(k)}$ of the current iter-

ate $\boldsymbol{x}^{(k)}$ over the correction polytope V_δ. Our notation in Algorithm 4 uses the elements of $\mathcal{V}$ as indices, rather than their contracted version; that is, we maintain the property that $\boldsymbol{x}^{(k)} = \sum_{\boldsymbol{v}\in\mathcal{V}} \alpha_{\boldsymbol{v}}^{(k)}[(1-\delta^{(k)})\boldsymbol{v} + \delta^{(k)}\boldsymbol{u}_0]$. As V_δ changes when δ changes, we need to update the barycentric representation of $\boldsymbol{x}^{(k)}$ accordingly – this is done in step 11 with the following equation. Suppose that we decrease δ to δ'. Then the old coordinates $\boldsymbol{\alpha}$ can be updated to new coordinates $\boldsymbol{\alpha}'$ for the new contraction polytope as follows:

$$\begin{aligned} \alpha'_{\boldsymbol{v}} &= \alpha_{\boldsymbol{v}} \frac{1-\delta}{1-\delta'} \qquad \text{for} \quad \boldsymbol{v} \in \mathcal{V} \setminus \{\boldsymbol{u}_0\}, \\ \alpha'_{\boldsymbol{u}_0} &= 1 - \sum_{\boldsymbol{v}\neq\boldsymbol{u}_0} \alpha'_{\boldsymbol{v}}. \end{aligned} \tag{6}$$

This ensures that $\sum_{\boldsymbol{v}} \alpha_{\boldsymbol{v}} \boldsymbol{v}_{(\delta)} = \sum_{\boldsymbol{v}} \alpha'_{\boldsymbol{v}} \boldsymbol{v}_{(\delta')}$, where $\boldsymbol{v}_{(\delta)} := (1-\delta)\boldsymbol{v} + \delta\boldsymbol{u}_0$, and that the coordinates form a valid convex combination (assuming that $\delta' \leq \delta$), as can be readily verified.

Algorithm 4: Optimizing f over $\mathcal{D}$ using Fully Corrective Frank-Wolfe (FCFW) with Adaptive-δ Algorithm.

1: **FCFW**$(\boldsymbol{x}^{(0)}, \mathcal{V}, \epsilon, \delta^{(\text{init})})$
2: **Inputs:** Set of atoms $\mathcal{V}$ so that $\mathcal{D} = \text{conv}(\mathcal{V})$, active set $\mathcal{S}^{(0)}$, starting point $\boldsymbol{x}^{(0)} = \sum_{\boldsymbol{v}\in\mathcal{S}^{(0)}} \alpha_{\boldsymbol{v}}^{(0)}[(1-\delta^{(\text{init})})\boldsymbol{v} + \delta^{(\text{init})}\boldsymbol{u}_0]$ where $\boldsymbol{\alpha}^{(0)}$ are the active coordinates, $\delta^{(\text{init})} \leq \frac{1}{4}$ describes the initial contraction of the polytope, stopping criterion ϵ, $\boldsymbol{u}_0$ is a fixed reference point in the relative interior of $\mathcal{D}$.
3: Let $V^{(0)} := \mathcal{S}^{(0)}$ (optionally, a bigger $V^{(0)}$ could be passed as argument for a warm start), $\delta^{(-1)} := \delta^{(\text{init})}$
4: **for** $k = 0 \dots K$ **do**
5: Let $\boldsymbol{s}^{(k)} \in \arg\min_{\boldsymbol{v}\in\mathcal{V}} \left\langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{v} \right\rangle$ *(the FW vertex)*
6: Compute $g(\boldsymbol{x}^{(k)}) = \langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{s}^{(k)} - \boldsymbol{x}^{(k)} \rangle$ *(FW gap)*
7: **if** $g(\boldsymbol{x}^{(k)}) \leq \epsilon$ **then**
8: **return** $\boldsymbol{x}^{(k)}$
9: **end if**
10: Let $\delta^{(k)}$ be $\delta^{(k-1)}$ updated according to Algorithm 1.
11: Update $\boldsymbol{\alpha}^{(k)}$ accordingly (using (6))
12: Let $\boldsymbol{s}_{(\delta)}^{(k)} := (1-\delta^{(k)})\boldsymbol{s}^{(k)} + \delta^{(k)}\boldsymbol{u}_0$
13: Let $\boldsymbol{d}_k^{\text{FW}} := \boldsymbol{s}_{(\delta)}^{(k)} - \boldsymbol{x}^{(k)}$
14: Line-search: $\gamma_k \in \arg\min_{\gamma\in[0,1]} f\left(\boldsymbol{x}^{(k)} + \gamma \boldsymbol{d}_k^{\text{FW}}\right)$
15: Set $\boldsymbol{x}^{(\text{temp})} := \boldsymbol{x}^{(k)} + \gamma_k \boldsymbol{d}_k^{\text{FW}}$ *(initialize correction to the update after a FW step with line search)*
16: $\boldsymbol{\alpha}^{(\text{temp})} = (1-\gamma_k)\boldsymbol{\alpha}^{(k)}$
17: $\alpha_{\boldsymbol{s}^{(k)}}^{(\text{temp})} \leftarrow \alpha_{\boldsymbol{s}^{(k)}}^{(\text{temp})} + \gamma_k$ *(update coordinates according to the FW step)*
18: Update (non-contracted) correction polytope: $V^{(k+1)} := V^{(k)} \cup \{\boldsymbol{s}^{(k)}\}$
19: Let $V_\delta = (1-\delta^{(k)})V^{(k+1)} + \delta^{(k)}\boldsymbol{u}_0$ *(contracted correction polytope)*
20: $\boldsymbol{x}^{(k+1)}, \boldsymbol{\alpha}^{(k+1)} := \textbf{MFW}(\boldsymbol{x}^{(\text{temp})}, \boldsymbol{\alpha}^{(\text{temp})}, V_\delta, \epsilon)$ *(approximate correction step on V_δ using MFW)*
21: **end for**

C Bounding the Sub-optimality for Fixed δ Variant

The pseudocode for optimizing over $\mathcal{D}_\delta$ for a fixed δ is given in Algorithm 4 (by ignoring the step 10 which updates δ). It is stated with a stopping criterion ϵ, but it can alternatively be run for a fixed number of K iterations. The following theorem bounds the suboptimality of the iterates with respect to the true optimum $\boldsymbol{x}^*$ over $\mathcal{D}$. If one can compute the constants in the theorem, one can choose a target contraction amount δ to guarantee a specific suboptimality of ϵ'; otherwise, one can choose δ using heuristics. Note that unlike the adaptive-δ variant, this algorithm does not converge to the true solution as $K \to \infty$ unless $\boldsymbol{x}^*$ happens to belong to $\mathcal{D}_\delta$. But the error can be controlled by choosing δ small enough.

Theorem C.1 (Suboptimality bound for fixed-δ algorithm)**.** *Let f satisfy the properties in Problem 3.2 and suppose its gradient is Lipschitz continuous on the contractions $\mathcal{D}_\delta$ as in Property 3.3. Suppose further that f is finite on the boundary of $\mathcal{D}$.*

Then f is uniformly continuous on $\mathcal{D}$ and has a modulus of continuity *function ω quantifying its level of continuity, i.e. $|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq \omega(\|\boldsymbol{x} - \boldsymbol{x}'\|)$ $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$, with $\omega(\sigma) \downarrow 0$ as $\sigma \downarrow 0$.*

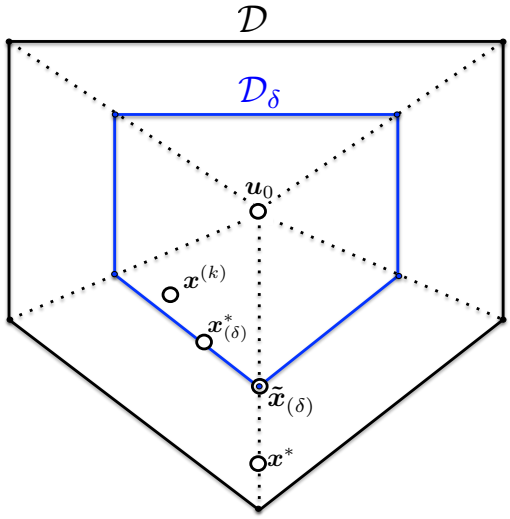


Figure 3: Illustration of the four points considered for the error analysis of the fixed-δ variant

Let $\boldsymbol{x}^*$ be an optimal point of f over $\mathcal{D}$. The iterates $\boldsymbol{x}^{(k)} \in \mathcal{D}_\delta$ of the FCFW algorithm as described in Algorithm 4 for a fixed $\delta > 0$ has sub-optimality over $\mathcal{D}$ bounded as:

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \leq \frac{2C_\delta}{(k+2)} + \omega\left(\delta \operatorname{diam}(\mathcal{D})\right), \tag{7}$$

where $C_\delta \leq \operatorname{diam}(\mathcal{D}_\delta)^2 L_\delta$. Note that different norms can be used in the definition of $\omega(\cdot)$ and C_δ.

Proof. Let $\boldsymbol{x}^*_{(\delta)}$ be an optimal point of f over $\mathcal{D}_\delta$. As f has a Lipschitz continuous gradient over $\mathcal{D}_\delta$, we can use any standard convergence result of the Frank-Wolfe algorithm to bound the suboptimality of the iterate $\boldsymbol{x}^{(k)}$ over $\mathcal{D}_\delta$. Algorithm 4 (with a fixed δ) describes the FCFW algorithm which guarantees at least as much progress as the standard FW algorithm (by step 15 and 20a), and thus we can use the convergence result from Jaggi (2013) as already stated in (4): $f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*_{(\delta)}) \leq \frac{2C_\delta}{(k+2)}$ with $C_\delta \leq \operatorname{diam}(\mathcal{D}_\delta)^2 L_\delta$, where L_δ comes from Property 3.3. This gives the first term in (7). Note that if the function f is *strongly* convex, then the FCFW algorithm has also a linear convergence rate (Lacoste-Julien and Jaggi, 2015), though we do not cover this here.

We now need to bound the difference $f(\boldsymbol{x}^*_{(\delta)}) - f(\boldsymbol{x}^*)$ coming from the fact that we are not optimizing over the full domain, and giving the second term in (7). We let $\tilde{\boldsymbol{x}}_{(\delta)}$ be the contraction of $\boldsymbol{x}^*$ on $\mathcal{D}_\delta$ towards $\boldsymbol{u}_0$, i.e. $\tilde{\boldsymbol{x}}_{(\delta)} := (1-\delta)\boldsymbol{x}^* + \delta \boldsymbol{u}_0$.[5] Note that $\|\tilde{\boldsymbol{x}}_{(\delta)} - \boldsymbol{x}^*\| = \delta \|\boldsymbol{x}^* - \boldsymbol{u}_0\| \leq \delta \operatorname{diam}(\mathcal{D})$, and thus can be made arbitrarily small by letting $\delta \downarrow 0$. Because $\tilde{\boldsymbol{x}}_{(\delta)} \in \mathcal{D}_\delta$, we have that $f(\tilde{\boldsymbol{x}}_{(\delta)}) \geq f(\boldsymbol{x}^*_{(\delta)})$ as $\boldsymbol{x}^*_{(\delta)}$ is optimal over $\mathcal{D}_\delta$. Thus $f(\boldsymbol{x}^*_{(\delta)}) - f(\boldsymbol{x}^*) \leq f(\tilde{\boldsymbol{x}}_{(\delta)}) - f(\boldsymbol{x}^*) \leq \omega(\|\tilde{\boldsymbol{x}}_{(\delta)} - \boldsymbol{x}^*\|)$ by the uniform continuity of f (that we explain below). Since ω is an increasing function, we have $\omega(\|\tilde{\boldsymbol{x}}_{(\delta)} - \boldsymbol{x}^*\|) \leq \omega(\delta \operatorname{diam}(\mathcal{D}))$, giving us the control on the second term of (7). See Figure 3 for an illustration of the four points considered in this proof.

Finally, we explain why f is uniformly continuous. As f is a (lower semi-continuous) convex function, it is continuous at every point where it is finite. As f is said to be finite at its boundary (and it is obviously finite in the relative interior of $\mathcal{D}$ as it is continuously differentiable there), then f is continuous over the whole of $\mathcal{D}$. As $\mathcal{D}$ is compact, this means that f is also uniformly continuous over $\mathcal{D}$. □

We note that the modulus of continuity function ω quantifies the level of continuity of f. For a Lipschitz continuous function, we have $\omega(\sigma) \leq L\sigma$. If instead we have $\omega(\sigma) \leq C\sigma^\alpha$ for some $\alpha \in [0,1]$, then f is actually α-Hölder continuous. We will see in Section E.2 that the TRW objective is not Lipschitz continuous, but it is α-Hölder continuous for any $\alpha < 1$, and so is "almost" Lipschitz continuous. From the theorem, we see that to get an accuracy of the order ϵ, we would need $(\delta \operatorname{diam}(\mathcal{D}))^\alpha < \epsilon$, and thus a contraction of $\delta < \frac{\epsilon^{(1/\alpha)}}{\operatorname{diam}(\mathcal{D})}$.

[5] Note that without a strong convexity assumption on f, the optimum over $\mathcal{D}_\delta$, $\boldsymbol{x}^*_{(\delta)}$, could be quite far from the optimum over $\mathcal{D}$, $\boldsymbol{x}^*$, which is why we need to construct this alternative close point to $\boldsymbol{x}^*$.

D Convergence with Adaptive-δ

In this section, we show the convergence of the adaptive-δ FW algorithm to optimize a function f satisfying the properties in Problem 3.2 and Property 3.3 (Lipschitz gradient over $\mathcal{D}_\delta$ with bounded growth).

The adaptive update for δ (given in Algorithm 1) can be used with the standard Frank-Wolfe optimization algorithm or also the fully corrective Frank-Wolfe (FCFW) variant. In FCFW, we ensure that every update makes more progress than a standard FW step with line-search, and thus we will show the convergence result in this section for standard FW (which also applies to FCFW). We describe the FCFW variant with approximate correction steps in Algorithm 4, as this is what we used in our experiments.

We first list a few definitions and lemmas that will be used for the main convergence convergence result given in Theorem D.6. We begin with the definitions of duality gaps that we use throughout this section. The Frank-Wolfe gap is our primary criterion for halting and measuring the progress of the optimization over $\mathcal{D}$. The uniform gap is a measure of the decrease obtainable from moving towards the uniform distribution.

Definition D.1. *We define the following gaps:*

1. *The Frank-Wolfe (FW) gap is defined as:* $g(\boldsymbol{x}^{(k)}) := \langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{s}^{(k)} - \boldsymbol{x}^{(k)} \rangle$.
2. *The uniform gap is defined as:* $g_u(\boldsymbol{x}^{(k)}) := \langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{u}_0 - \boldsymbol{x}^{(k)} \rangle$.
3. *The FW gap over $\mathcal{D}_\delta$ is:* $g_{(\delta^{(k)})}(\boldsymbol{x}^{(k)}) := \langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{s}_{(\delta)}^{(k)} - \boldsymbol{x}^{(k)} \rangle$.

The name for the uniform gap comes from the fact that the FW gap over $\mathcal{D}_\delta$ can be expressed as a convex combination of the FW gap over $\mathcal{D}$ and the uniform gap:

$$\begin{aligned} g_{(\delta^{(k)})}(\boldsymbol{x}^{(k)}) &= \langle -\nabla f(\boldsymbol{x}^{(k)}),\ (1-\delta^{(k)})\boldsymbol{s}^{(k)} + \delta^{(k)}\boldsymbol{u}_0 - \boldsymbol{x}^{(k)} \rangle \\ &= (1-\delta^{(k)}) g(\boldsymbol{x}^{(k)}) + \delta^{(k)} g_u(\boldsymbol{x}^{(k)}). \end{aligned} \tag{8}$$

The uniform gap represents the negative directional derivative of f at $\boldsymbol{x}^{(k)}$ in the direction $\boldsymbol{u}_0 - \boldsymbol{x}^{(k)}$. When the uniform gap is negative (thus f is increasing when moving towards $\boldsymbol{u}_0$ from $\boldsymbol{x}^{(k)}$), then the contraction is hurting progress, which explains the type of adaptive update for δ given by Algorithm 1 where we consider shrinking δ in this case. This enables us to crucially relate the FW gap over $\mathcal{D}_\delta$ with the one over $\mathcal{D}$, as given in the following lemma, using the assumption that $\delta^{(\text{init})} \leq \frac{1}{4}$.

Lemma D.2 (Gaps relationship)**.** *For iterates progressing as in Algorithm 4 with adaptive update on δ as given in Algorithm 1, the gap over $\mathcal{D}_\delta$ and $\mathcal{D}$ are related as :* $g_{(\delta^{(k)})}(\boldsymbol{x}^{(k)}) \geq \frac{g(\boldsymbol{x}^{(k)})}{2}$.

Proof. The duality gaps $g(\boldsymbol{x}^{(k)})$ and $g_{(\delta^{(k)})}(\boldsymbol{x}^{(k)})$ computed as defined in (D.1) during Algorithm 4 are related by equation (8).

We analyze two cases separately:

(1) When $g_u(\boldsymbol{x}^{(k)}) \geq 0$, for $\delta^{(\text{init})} \leq \frac{1}{4}$, we have $g_{(\delta^{(k)})}(\boldsymbol{x}^{(k)}) \geq \frac{3}{4} g(\boldsymbol{x}^{(k)})$ as $\delta^{(k)} \leq \delta^{(\text{init})}$.

(2) When $g_u(\boldsymbol{x}^{(k)}) < 0$, from the update rule in lines 5 to 7 in Algorithm 1, we have $\delta^{(k)} \leq \frac{g(\boldsymbol{x}^{(k)})}{-4 g_u(\boldsymbol{x}^{(k)})} \implies \delta^{(k)} g_u(\boldsymbol{x}^{(k)}) \geq -\frac{g(\boldsymbol{x}^{(k)})}{4}$. Therefore, $g_{(\delta^{(k)})}(\boldsymbol{x}^{(k)}) \geq \frac{3}{4} g(\boldsymbol{x}^{(k)}) - \frac{g(\boldsymbol{x}^{(k)})}{4} = \frac{g(\boldsymbol{x}^{(k)})}{2}$.

Therefore, the gap over $\mathcal{D}_\delta$ and $\mathcal{D}$ are related as : $g_{(\delta^{(k)})}(\boldsymbol{x}^{(k)}) \geq \frac{g(\boldsymbol{x}^{(k)})}{2}$. □

Another property that we will use in the convergence proof is that $-g_u$ is upper bounded for any convex function f:[6]

Lemma D.3 (Bounded negative uniform gap)**.** *Let f be a continuously differentiable convex function on the relative interior of $\mathcal{D}$. Then for any fixed $\boldsymbol{u}_0$ in the relative interior of $\mathcal{D}$, $\exists B$ s.t.*

$$\forall \boldsymbol{x} \in \mathcal{D},\quad -g_u(\boldsymbol{x}) = \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}_0 - \boldsymbol{x} \rangle \leq B. \tag{9}$$

In particular, we can take the finite value:

$$B := \|\nabla f(\boldsymbol{u}_0)\|_* \text{diam}_{\|\cdot\|}(\mathcal{D}) \tag{10}$$

[6] Note that on the other hand, $g_u(\boldsymbol{x})$ might go to infinity as $\boldsymbol{x}$ gets close to the boundary of $\mathcal{D}$ as the gradient of f is allowed to be unbounded. Fortunately, we only need an upper bound on $-g_u$, not a lower bound.

Proof. As f is convex, its directional derivative is a monotone increasing function in any direction. Let $\boldsymbol{u}_0$ and $\boldsymbol{x}$ be points in the relative interior of $\mathcal{D}$; then their gradient exists and we have by the monotonicity property:

$$\begin{aligned} &\langle \nabla f(\boldsymbol{u}_0) - \nabla f(\boldsymbol{x}), \boldsymbol{u}_0 - \boldsymbol{x} \rangle \geq 0 \\ \implies &\langle \nabla f(\boldsymbol{u}_0), \boldsymbol{u}_0 - \boldsymbol{x} \rangle \geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}_0 - \boldsymbol{x} \rangle . \end{aligned}$$

This inequality is valid for all $\boldsymbol{x}$ in the relative interior of $\mathcal{D}$, and can be extended to the boundary by taking limits (with potentially the RHS become minus infinity, but this is not a problem). Finally, by the definition of the dual norm (generalized Cauchy-Schwartz), we have $\langle \nabla f(\boldsymbol{u}_0), \boldsymbol{u}_0 - \boldsymbol{x} \rangle \leq \|\nabla f(\boldsymbol{u}_0)\|_* \|\boldsymbol{u}_0 - \boldsymbol{x}\| \leq \|\nabla f(\boldsymbol{u}_0)\|_* \mathrm{diam}_{\|\cdot\|}(\mathcal{D})$. □

Finally, we need a last property of Algorithm 4 that allows us to bound the amount of perturbation $\delta^{(k)}$ of the polytope at every iteration as a function of the sub-optimality over $\mathcal{D}$.

Lemma D.4 (Lower bound on perturbation)**.** *Let B be a bound such that $-8 g_u(\boldsymbol{x}) \leq B$ for all $\boldsymbol{x} \in \mathcal{D}$ (given by Lemma D.3). Then at every stage of Algorithm 4, we have that:*

$$\delta^{(\mathrm{init})} \geq \delta^{(k)} \geq \min \left\{ \frac{h_k}{B}, \delta^{(\mathrm{init})} \right\},$$

where $\delta^{(\mathrm{init})}$ is the initial value of δ and $h_k := f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*)$ is the sub-optimality of the iterate.

Proof. When defining $\delta^{(k)}$ in step 10 of Algorithm 4, we either preserve the value of $\delta^{(k-1)}$ or if we update it, then by the lines 6 and 7 of Algorithm 1, we have $\delta^{(k)} \geq \frac{\tilde{\delta}}{2} = \frac{1}{2} \frac{g(\boldsymbol{x}^{(k)})}{-4 g_u(\boldsymbol{x}^{(k)})} \geq \frac{g(\boldsymbol{x}^{(k)})}{B}$ (by using $g_u(\boldsymbol{x}^{(k)}) < 0$ in this case). Since $g(\boldsymbol{x}^{(k)}) \geq h_k$ (the FW gap always upper bounds the suboptimality), we conclude $\delta^{(k)} \geq \min\{\frac{h_k}{B}, \delta^{(k-1)}\}$. Unrolling this recurrence, we thus get:

$$\delta^{(k)} \geq \min \left\{ \min_{0 \leq l \leq k} \frac{h_l}{B}, \ \delta^{(\mathrm{init})} \right\} = \min \left\{ \frac{h_k}{B}, \delta^{(\mathrm{init})} \right\}.$$

For the last equality, we used the fact that h_k is non-increasing since Algorithm 4 decreases the objective at every iteration (using the line-search in step 14). □

We now bound the generalization of a standard recurrence that will arise in the proof of convergence. This is a generalization of the technique used in Teo et al. (2007) (also used in the context of Frank-Wolfe in the proof of Theorem C.4 in Lacoste-Julien et al. (2013)). The basic idea is that one can bound a recurrence inequality by the solution to a differential equation. We provide a detailed proof of the bound for completeness here.

Lemma D.5 (Recurrence inequality solution)**.** *Let $1 < a \leq b$. Suppose that h_k is any non-negative sequence that satisfies the recurrence inequality:*

$$h_{k+1} \leq h_k - \frac{1}{bC_0} (h_k)^a \qquad \textit{with initial condition} \quad h_0^{a-1} \leq C_0.$$

Then h_k is strictly decreasing (unless it equals zero) and can be bounded for $k \geq 0$ as:

$$h_k \leq \left(\frac{C_0}{(\frac{a-1}{b})k + 1} \right)^{\frac{1}{a-1}}$$

Proof. Taking the continuous time analog of the recurrence inequality, we consider the differential equation:

$$\frac{\mathrm{d}h}{\mathrm{d}t} = \frac{-h^a}{bC_0} \qquad \text{with initial condition} \quad h(0) = C_0^{\frac{1}{a-1}}.$$

Solving it:

$$\begin{aligned}
&\frac{\mathrm{d}h}{\mathrm{d}t} = \frac{-h^a}{bC_0} \\
\implies &\int \frac{\mathrm{d}h}{h^a} = \int \frac{-\mathrm{d}t}{bC_0} \\
\implies &\left[\frac{-h^{1-a}}{a-1}\right]_{h(0)}^{h(t)} = -\frac{t-0}{bC_0} \\
&\textit{(Using the initial conditions:)} \\
\implies &\frac{-1}{h(t)^{a-1}} + \frac{1}{C_0} = \frac{-t(a-1)}{bC_0} \\
\implies &\frac{1}{h(t)^{a-1}} = \left((\frac{a-1}{b})t + 1\right)\frac{1}{C_0} \\
\implies &h(t) = \left(\frac{C_0}{(\frac{a-1}{b})t+1}\right)^{\frac{1}{a-1}}.
\end{aligned}$$

We now denote the solution to the differential equation as $\tilde{h}(t)$. Note that it is a strictly decreasing convex function (which could also be directly implied from the differential equation as: $\frac{\mathrm{d}^2 h}{\mathrm{d}t^2} = -a \underbrace{\frac{h^{a-1}}{bC_0}}_{>0} \underbrace{h'(t)}_{<0} > 0$).

Our strategy will be to show by induction that if $h_k \leq \tilde{h}(k)$, then $h_{k+1} \leq \tilde{h}(k+1)$. This allows us to bound the recurrence by the solution to the differential equation.

Assume that $h_k \leq \tilde{h}(k)$. The base case is $h_0 \leq \tilde{h}(0) = C_0^{\frac{1}{a-1}}$, which is true by the initial condition on h_0.

Consider the utility function $l(h) := h - \frac{h^a}{bC_0}$ which is maximized at $\bar{h} := \left(\frac{bC_0}{a}\right)^{\frac{1}{a-1}}$. This function can be verified to be strictly concave for $a > 1$ and therefore is increasing for $h \leq \bar{h}$. Note that the recurrence inequality can be written as $h_{k+1} \leq l(h_k)$. Since $\tilde{h}$ is decreasing and that $\tilde{h}(0)) = C_0^{\frac{1}{a-1}} \leq \left(\frac{bC_0}{a}\right)^{\frac{1}{a-1}} = \bar{h}$ (the last inequality holds since $b \geq a$), we have $\tilde{h}(t) \leq \bar{h}$ for all $t \geq 0$, and so $\tilde{h}(t)$ is always in the monotone increasing region of l.

From the induction hypothesis and the monotonicity of l, we thus get that $l(h_k) \leq l(\tilde{h}(k))$.

Now the convexity of $\tilde{h}(t)$ gives us $\tilde{h}(k+1) \geq \tilde{h}(k) + \tilde{h}'(k) = \tilde{h}(k) - \frac{\tilde{h}(k)^a}{bC_0} = l(\tilde{h}(k))$. Combining these two facts with the recurrence inequality $h_{k+1} \leq l(h_k)$, we get: $h_{k+1} \leq l(h_k) \leq l(\tilde{h}(k)) \leq \tilde{h}(k+1)$, completing the induction step and the main part of the proof.

Finally, whenever $h_k > 0$, we have that $h_{k+1} < h_k$ from the recurrence inequality, and so h_k is strictly decreasing as claimed. □

Given these elements, we are now ready to state the main convergence result for Algorithm 4. The convergence rate goes through three stages with increasingly slower rate. The level of suboptimality h_k determines the stage. We first give the high level intuition behind these stages. Recall that by Lemma D.4, h_k lower bounds the amount of perturbation $\delta^{(k)}$, and thus when h_k is big, the function f is well-behaved by Property 3.3. In the first stage, the suboptimality is bigger than some target constant (which implies that the FW gap is big), yielding a geometric rate of decrease of error (as is standard for FW with line-search in the first few steps). In the second stage, the suboptimality is in an intermediate regime: it is smaller than the target constant, but big enough compared to the initial δ^{init} so that f is still well-behaved on $\mathcal{D}_{\delta^{(k)}}$. We get there the usual $O(1/k)$ rate as in standard FW. Finally, in the third stage, we get the slower $O(k^{-\frac{1}{p+1}})$ rate where the growth in $O(\delta^{-p})$ of the Lipschitz constant of f over $\mathcal{D}_\delta$ comes into play.

Theorem D.6 (Global convergence for adaptive-δ variant over $\mathcal{D}$)**.** *Consider the optimization of f satisfying the properties in Problem 3.2 and Property 3.3. Let $\widetilde{C} := L \operatorname{diam}_{\|\cdot\|}(\mathcal{D})^2$, where L is from Property 3.3. Let B be the upper bound on the negative uniform gap: $-8g_u(\boldsymbol{x}) \leq B$ for all $\boldsymbol{x} \in \mathcal{D}$, as used in Lemma D.4 (arising from Lemma D.3). Then the iterates $\boldsymbol{x}^{(k)}$ obtained by running the Frank-Wolfe updates over $\mathcal{D}_\delta$ with line-search with δ updated according to Algorithm 1 (or as summarized in a FCFW variant in Algorithm 4), have suboptimality h_k upper bounded as:*

1. $h_k \leq \left(\frac{1}{2}\right)^k h_0 + \frac{\widetilde{C}}{\delta_0^p}$ *for k such that* $h_k \geq \max\{B\delta_0, \frac{2\widetilde{C}}{\delta_0^p}\}$,

2. $h_k \leq \frac{2\widetilde{C}}{\delta_0^p}\left[\frac{1}{\frac{1}{4}(k-k_0)+1}\right]$ *for* k *such that* $B\delta_0 \leq h_k \leq \frac{2\widetilde{C}}{\delta_0^p}$,

3. $h_k \leq \left[\frac{\max(\widetilde{C}, B\delta_0^{p+1})B^p}{\frac{p+1}{\max(8,p+2)}(k-k_1)+1}\right]^{\frac{1}{p+1}} = O(k^{-\frac{1}{p+1}})$ *for* k *such that* $h_k \leq B\delta_0$,

where $\delta_0 = \delta^{(\text{init})}$, h_0 *is the initial suboptimality, and* k_0 *and* k_1 *are the number of steps to reach stage 2 and 3 respectively which are bounded as:* $k_0 \leq \max(0, \lceil \log_{\frac{1}{2}} \frac{\widetilde{C}}{h_0 \delta_0^p} \rceil)$, $k_0 \leq k_1 \leq k_0 + \max\left(0, \lceil \frac{8\widetilde{C}}{B\delta_0^{p+1}} \rceil - 4\right)$.

Proof. Let $\boldsymbol{x}_\gamma := \boldsymbol{x}^{(k)} + \gamma \boldsymbol{d}_k^{\text{FW}}$ with $\boldsymbol{d}_k^{\text{FW}}$ defined in step 12 in Algorithm 4. Note that $\boldsymbol{x}_\gamma \in \mathcal{D}_\delta$ with $\delta = \delta^{(k)}$ for all $\gamma \in [0,1]$. We apply the Descent Lemma A.1 on this update to get:

$$f(\boldsymbol{x}_\gamma) \leq f(\boldsymbol{x}^{(k)}) + \gamma \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{d}_k^{\text{FW}} \rangle + \gamma^2 \frac{L\|\boldsymbol{d}_k^{\text{FW}}\|^2}{2\,(\delta^{(k)})^p} \qquad \forall \gamma \in [0,1].$$

We have $L\|\boldsymbol{d}_k^{\text{FW}}\|^2 \leq \widetilde{C}$ by assumption and $\langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{d}_k^{\text{FW}} \rangle = -g_{(\delta)}(\boldsymbol{x}^{(k)})$ by definition. Moreover, $\boldsymbol{x}^{(k+1)}$ is defined to make at least as much progress than the line-search result $\min_{\gamma \in [0,1]} f(\boldsymbol{x}_\gamma)$ (line 14 and 15), and so we have:

$$\begin{aligned} f(\boldsymbol{x}^{(k+1)}) &\leq f(\boldsymbol{x}^{(k)}) - \gamma g_{(\delta)}(\boldsymbol{x}^{(k)}) + \gamma^2 \frac{\widetilde{C}}{2\,(\delta^{(k)})^p} \quad \forall \gamma \in [0,1] \\ &\leq f(\boldsymbol{x}^{(k)}) - \frac{\gamma}{2} g(\boldsymbol{x}^{(k)}) + \gamma^2 \frac{\widetilde{C}}{2\,(\delta^{(k)})^p} \quad \forall \gamma \in [0,1]. \end{aligned}$$

For the final inequality, we used Lemma D.2 which relates the gap over $\mathcal{D}_\delta$ to the gap over $\mathcal{D}$.

Subtracting $f(\boldsymbol{x}^*)$ from both sides and using $g(\boldsymbol{x}^{(k)}) \geq h_k$ by convexity, we get:

$$h_{k+1} \leq h_k - \frac{\gamma h_k}{2} + \frac{\gamma^2 \widetilde{C}}{2\,(\delta^{(k)})^p}.$$

Now, using Lemma D.4, we have that $\delta^{(k)} \geq \min(\frac{h_k}{B}, \delta^{(\text{init})})$:

$$h_{k+1} \leq h_k - \gamma \frac{h_k}{2} + \frac{\gamma^2}{2} \frac{\widetilde{C}}{\left(\min(\frac{h_k}{B}, \delta^{(\text{init})})\right)^p} \quad \forall \gamma \in [0,1]. \tag{11}$$

We refer to (11) as the master inequality. Since we no longer have a dependance on $\delta^{(k)}$, we refer to $\delta^{(\text{init})}$ as δ_0. We now follow a similar form of analysis as in the proof of Theorem C.4 in Lacoste-Julien et al. (2013). To solve this and bound the suboptimality, we consider three stages:

1. **Stage 1:** The min in the denominator is δ_0 and h_k is big: $h_k \geq \max\{B\delta_0, \frac{2\widetilde{C}}{\delta_0^p}\}$.
2. **Stage 2:** The min in the denominator is δ_0 and h_k is small: $B\delta_0 \leq h_k \leq \frac{2\widetilde{C}}{\delta_0^p}$.
3. **Stage 3:** The min in the denominator is $\frac{h_k}{B}$, i.e.: $h_k \leq B\delta_0$.

Since h_k is decreasing, once we leave a stage, we no longer re-enter it. The overall strategy for each stage is as follows. For each recurrence that we get, we select a γ^* that realizes the tightest upper bound on it.

Since we are restricted that $\gamma^* \in [0,1]$, we have to consider when $\gamma^* > 1$ and $\gamma^* \leq 1$. For the former, we bound the recurrence obtained by substituting $\gamma = 1$ into (11). For the latter, we substitute the form of γ^* into the recurrence and bound the result.

Stage 1

We consider the case where $h_k \geq B\delta_0$. This yields:

$$h_{k+1} \leq h_k - \frac{\gamma h_k}{2} + \frac{\gamma^2 \widetilde{C}}{2\,(\delta_0)^p} \tag{12}$$

The bound is minimized by setting $\gamma^* = \frac{h_k \delta_0^p}{2\widetilde{C}}$. On the other hand, the bound is only valid for $\gamma \in [0, 1]$, and thus if $\gamma^* > 1$, i.e. $h_k > \frac{2\widetilde{C}}{\delta_0^p}$ (stage 1), then $\gamma = 1$ will yield the minimum feasible value for the bound. Unrolling the recursion (12) for $\gamma = 1$ during this stage (where $h_l > \frac{2\widetilde{C}}{\delta_0^p}$ for $l < k$ as h_k is decreasing), we get:

$$\begin{aligned} h_{k+1} &\leq \frac{h_k}{2} + \frac{\widetilde{C}}{2\delta_0^p} \\ &\leq \frac{1}{2}\left(\frac{h_{k-1}}{2} + \frac{\widetilde{C}}{2\delta_0^p}\right) + \frac{\widetilde{C}}{2\delta_0^p} \\ &\leq \left(\frac{1}{2}\right)^{k+1} h_0 + \frac{\widetilde{C}}{2\delta_0^p} \underbrace{\sum_{l=0}^{k} \left(\frac{1}{2}\right)^l}_{\leq \sum_{l=0}^{\infty}\left(\frac{1}{2}\right)^l = 2} \\ \text{thus} \quad h_k &\leq \left(\frac{1}{2}\right)^k h_0 + \frac{\widetilde{C}}{\delta_0^p}, \end{aligned} \tag{13}$$

giving the bound for the iterates in the first stage.

We can compute an upper bound on the number of steps it takes to reach a suboptimality of $\frac{2\widetilde{C}}{\delta_0^p}$ by looking at the minimum k which ensures that the bound in (13) becomes smaller than $\frac{2\widetilde{C}}{\delta_0^p}$, yielding $k_{\max} = \max(0, \lceil \log_{\frac{1}{2}} \frac{\widetilde{C}}{h_0 \delta_0^p} \rceil)$. Therefore, let $k_0 \leq k_{\max}$ be the first k such that $h_k \leq \frac{2\widetilde{C}}{\delta_0^p}$.

Stage 2

For this case analysis, we refer to k as being the iterations *after* k_0 steps have elapsed. I.e. if $k_{\text{new}} := k - k_0$, then we refer to k_{new} as k moving forward.

In stage 2, we suppose that $B\delta_0 \leq h_k \leq \frac{2\widetilde{C}}{\delta_0^p}$. This means that $\gamma^* = \frac{h_k \delta_0^p}{2\widetilde{C}} \leq 1$.

Substituting $\gamma = \gamma^*$ into (12) yields: $h_{k+1} \leq h_k - h_k^2 \frac{\delta_0^p}{8\widetilde{C}}$.

Using the result of Lemma D.5 with $a = 2$, $b = 4$ and $C_0 = \frac{2\widetilde{C}}{\delta_0^p}$, we get the bound:

$$h_k \leq \frac{\frac{2\widetilde{C}}{\delta_0^p}}{\frac{k-k_0}{4} + 1}.$$

It is worthwhile to point out at this juncture that the bound obtained for stage 2 is the same as the one for regular Frank-Wolfe, but with a factor of 4 worse due to the factor of $\frac{1}{2}$ in front of the FW gap which appeared due to Lemma D.2.

Stage 3

Here, we suppose $h_k \leq B\delta_0$. We can compute a bound on the number of steps k_1 needed get to stage 3 by looking at the number of steps it takes for the bound in stage 2 to becomes less than $B\delta_0$:

$$\begin{aligned} \frac{2\widetilde{C}}{\delta_0^p}\left[\frac{4}{k_1 - k_0 + 4}\right] &\leq B\delta_0 \\ \left[\frac{1}{k_1 - k_0 + 4}\right] &\leq \frac{B\delta_0^{p+1}}{8\widetilde{C}} \\ k_1 &\geq k_0 + \lceil \frac{8\widetilde{C}}{B\delta_0^{p+1}} \rceil - 4. \end{aligned}$$

As before, moving forward, our notation on k represents the number of steps taken after k_1 steps.

Then, the master inequality (11) becomes:

$$h_{k+1} \leq h_k - \frac{\gamma}{2} h_k + \frac{\gamma^2 \widetilde{C} B^p}{2 h_k^p}.$$

To simplify the rest of the analysis, we replace $\widetilde{C}B^p$ with $F := \max(B\delta_0^{p+1}, \widetilde{C})B^p$. We then get the bound:

$$h_{k+1} \leq h_k - \frac{\gamma}{2} h_k + \frac{\gamma^2 F}{2 h_k}, \tag{14}$$

which is minimized by setting $\gamma^* := \frac{h_k^{p+1}}{2F}$. Since $F \geq B^{p+1}\delta_0^{p+1}$ (by construction) and $h_k^{p+1} \leq (B\delta_0)^{p+1}$ (by the condition to be in stage 3), we necessarily have that $\gamma^* \leq 1$. We chose the value of F to avoid having to consider the possibility $\gamma^* > 1$ as we did in the distinction between stage 1 and stage 2.

Hence, substituting $\gamma = \gamma^*$ in (14), we get:

$$h_{k+1} \leq h_k - \frac{h_k^{p+2}}{8F}.$$

Using the result of Lemma D.5 with $a = p+2$, $b = \max(8, p+2)$ and $C_0 = F$, we get the bound:

$$h_k \leq \left[\frac{\max(\widetilde{C}, B\delta_0^{p+1})B^p}{\frac{p+1}{\max(8,p+2)}(k-k_1)+1} \right]^{\frac{1}{p+1}} = O(k^{-\frac{1}{p+1}}),$$

concluding the proof. □

Interestingly, the obtained rate of $O(1/\sqrt{k})$ for $p = 1$ (for the TRW objective e.g.) is the standard rate that one would get for the optimization of a general non-smooth convex function with the projected subgradient method (and it is even a lower bound for some class of first-order methods; see e.g. Section 3.2 in Nesterov (2004)). The fact that our function f does not have Lipschitz continuous gradient on the whole domain brings us back to the realm of non-smooth optimization. It is an open question whether Algorithm 4 has an optimal rate for the class of functions defined in the assumptions of Theorem D.6.

E Properties of the TRW Objective

In this section, we explicitly compute bounds for the constants appearing in the convergence statements for our fixed-δ and adaptive-δ algorithms for the optimization problem given by:

$$\min_{\vec{\boldsymbol{\mu}} \in \mathcal{M}} -\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho}).$$

In particular, we compute the Lipschitz constant for its gradient over $\mathcal{M}_\delta$ (Property 3.3), we give a form for its modulus of continuity function $\omega(\cdot)$ (used in Theorem 3.4), and we compute B, the upper bound on the negative uniform gap (as used in Lemma D.3).

E.1 Property 3.3 : Controlled Growth of Lipschitz Constant over $\mathcal{M}_\delta$

We first motivate our choice of norm over $\mathcal{M}$. Recall that $\vec{\boldsymbol{\mu}}$ can be decomposed into $|V|+|E|$ blocks, with one pseudo-marginal vector $\boldsymbol{\mu}_i \in \Delta_{\text{VAL}_i}$ for each node $i \in V$, and one vector $\boldsymbol{\mu}_{ij} \in \Delta_{\text{VAL}_i \text{VAL}_j}$ per edge $\{i,j\} \in E$, where Δ_d is the probability simplex over d values. We let c be the cliques in the graph (either nodes or edges). From its definition in (2), $f(\vec{\boldsymbol{\mu}}) := -\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ decomposes as a separable sum of functions of each block only:

$$f(\vec{\boldsymbol{\mu}}) := -\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho}) = -\sum_c \left(K_c H(\boldsymbol{\mu}_c) + \langle \boldsymbol{\theta}_c, \boldsymbol{\mu}_c \rangle \right) =: \sum_c g_c(\boldsymbol{\mu}_c), \tag{15}$$

where K_c is $(1 - \sum_{j \in \mathcal{N}(i)} \rho_{ij})$ if $c = i$ and ρ_{ij} if $c = \{i,j\}$. The function g_c also decomposes as a separable sum:

$$g_c(\boldsymbol{\mu}_c) := \sum_{x_c} K_c \mu_c(x_c) \log(\mu_c(x_c)) - \theta_c(x_c)\mu_c(x_c) =: \sum_{x_c} g_{c,x_c}(\mu_c(x_c)). \tag{16}$$

As $\mathcal{M}$ is included in a product of probability simplices, we will use the natural ℓ_∞/ℓ_1 block-norm, i.e. $\|\vec{\boldsymbol{\mu}}\|_{\infty,1} := \max_c \|\boldsymbol{\mu}_c\|_1$. The diameter of $\mathcal{M}$ in this norm is particularly small: $\text{diam}_{\|\cdot\|_{\infty,1}}(\mathcal{M}) \leq 2$. The dual norm of the ℓ_∞/ℓ_1 block-norm is the ℓ_1/ℓ_∞ block-norm, which is what we will need to measure the Lipschitz constant of the gradient (because of the dual norm pairing requirement from the Descent Lemma A.1).

Lemma E.1. *Consider the ℓ_∞/ℓ_1 norm on $\mathcal{M}$ and its dual norm ℓ_1/ℓ_∞ to measure the gradient. Then $\nabla TRW(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ is Lipschitz continuous over $\mathcal{M}_\delta$ with respect to these norms with Lipschitz constant $L_\delta \leq \frac{L}{\delta}$ with:*

$$L \leq 4|V| \max_{ij \in E} (VAL_i VAL_j). \tag{17}$$

Proof. We first consider one scalar component of the separable $g_c(\boldsymbol{\mu}_c)$ function given in (16) (i.e. for one $\mu_c(x_c)$ coordinate). Its derivative is $K_c(1 + \log(\mu_c(x_c)) - \theta_c(x_c)$ with second derivative $\frac{K_c}{\mu_c(x_c)}$. If $\vec{\boldsymbol{\mu}} \in \mathcal{M}_\delta$, then we have $\mu_c(x_c) \geq \delta u_0(x_c) = \frac{\delta}{n_c}$, where n_c is the number of possible values that the assignment variable x_c can take. Thus for $\vec{\boldsymbol{\mu}} \in \mathcal{M}_\delta$, we have that the x_c-component of g_c is Lipschitz continuous with constant $|K_c|n_c/\delta$. We thus have:

$$\begin{aligned}\|\nabla g_c(\boldsymbol{\mu}_c) - \nabla g_c(\boldsymbol{\mu}'_c)\|_\infty &= \max_{x_c} |g'_{c,x_c}(\mu(x_c)) - g'_{c,x_c}(\mu'(x_c))| \\ &\leq \frac{|K_c|n_c}{\delta}\|\boldsymbol{\mu}_c - \boldsymbol{\mu}'_c\|_\infty \leq \frac{|K_c|n_c}{\delta}\|\boldsymbol{\mu}_c - \boldsymbol{\mu}'_c\|_1.\end{aligned}$$

Considering now the ℓ_1-sum over blocks, we have:

$$\begin{aligned}\|\nabla f(\vec{\boldsymbol{\mu}}) - \nabla f(\vec{\boldsymbol{\mu}}')\|_{1,\infty} &= \sum_c \|\nabla g_c(\boldsymbol{\mu}_c) - \nabla g_c(\boldsymbol{\mu}'_c)\|_\infty \\ &\leq \sum_c \frac{K_c n_c}{\delta}\|\boldsymbol{\mu}_c - \boldsymbol{\mu}'_c\|_1 \leq \frac{1}{\delta}\left(\sum_c K_c n_c\right)\|\vec{\boldsymbol{\mu}} - \vec{\boldsymbol{\mu}}'\|_{\infty,1}.\end{aligned}$$

The Lipschitz constant is thus indeed $\frac{L}{\delta}$ with $L := \sum_c |K_c|n_c$. Let us first consider the sum for $c \in V$; we have $K_i = 1 - \sum_{j \in \mathcal{N}(i)} \rho_{ij}$. Thus:

$$\begin{aligned}\sum_i |K_i| &\leq |V| + \sum_i \sum_{j \in \mathcal{N}(i)} \rho_{ij} \\ &= |V| + 2\sum_{ij \in E} \rho_{ij} = |V| + 2(|V| - 1) \leq 3|V|.\end{aligned}$$

Here we used the fact that ρ_{ij} came from the marginal probability of edges of spanning trees (and so with $|V|-1$ edges). Similarly, we have $\sum_{ij \in E} |K_{ij}| \leq |V|$. Combining these we get:

$$L = \sum_c |K_c|n_c \leq (\max_c n_c) \sum_c |K_c| \leq \max_{ij \in E} \text{VAL}_i \text{VAL}_j 4|V|. \tag{18}$$

□

Remark 1. The important quantity in the convergence of Frank-Wolfe type algorithms is $\tilde{C} = L \operatorname{diam}(\mathcal{M})^2$. We are free to take any dual norm pairs to compute this quantity, but some norms are better aligned with the problem than others. Our choice of norm in Lemma E.1 gives $\tilde{C} \leq 16|V|k^2$ where k is the maximum number of possible values a random variable can take. It is interesting that $|E|$ does not appear in the constant. If instead we had used the ℓ_2/ℓ_1 block-norm on $\mathcal{M}$, we get that $\operatorname{diam}_{\ell_2/\ell_1}(\mathcal{M})^2 = 4(|V| + |E|)$, while the constant L with dual norm ℓ_2/ℓ_∞ would be instead $\max_c |K_c|n_c$ which is bigger than $\max_c n_c = k^2$, thus giving a worse bound.

E.2 Modulus of Continuity Function

We begin by computing a modulus of continuity function for $-x \log x$ with an additive linear term.

Lemma E.2. *Let $g(x) := -Kx \log x + \theta x$. Consider $x, x' \in [0, 1]$ such that $|x - x'| \leq \sigma$, then:*

$$|g(x') - g(x)| \leq \sigma|\theta| + 2\sigma|K| \max\{-\log(2\sigma), 1\} =: \omega_g(\sigma). \tag{19}$$

Proof. Without loss of generality assume $x' > x$, then we have two cases:

Case i. If $x > \sigma$, then we have that the Lipschitz constant of $g(x)$ is $L_\sigma = |\theta| + |K||(1 + \log \sigma)|$ (obtained by taking the supremum of its derivative). Therefore, we have that $|g(x') - g(x)| \leq L_\sigma \sigma$. Note that $L_\sigma \sigma \to 0$ when $\sigma \to 0$ even if $L_\sigma \to \infty$, since L_σ grows logarithmically.

Case ii. If $x \leq \sigma$, then $x' \leq x + \sigma \leq 2\sigma$. Therefore:

$$|g(x') - g(x)| \leq |K||x \log x - x' \log x'| + |\theta||x' - x|. \tag{20}$$

Now, we have that $-x \log x$ is non-negative for $x \in [0,1]$. Furthermore, we have that $-x \log x$ is increasing when $x < \exp(-1)$ and decreasing afterwards. First suppose that $2\sigma \leq \exp(-1)$; then $-x' \log x' \geq -x \log x \geq 0$ which implies:

$$|x \log x - x' \log x'| \leq -x' \log x' \leq -2\sigma \log(2\sigma).$$

In the case $2\sigma > \exp(-1)$, then we have:

$$|x \log x - x' \log x'| \leq \max_{y \in [0,1]} \{-y \log y\} = \exp(-1) \leq 2\sigma.$$

Combining these two possibilities, we get:

$$|x \log x - x' \log x'| \leq 2\sigma \max\{-\log(2\sigma), 1\}.$$

The inequality (20) thus becomes:

$$|g(x') - g(x)| \leq |K| 2\sigma \max\{-\log(2\sigma), 1\} + |\theta|\sigma,$$

which is what we wanted to prove. □

For small σ, the dominant term of the function $\omega_g(\sigma)$ in Lemma E.2 is of the form $C \cdot -\sigma \log \sigma$ for a constant C. If we require that this be smaller than some small $\xi > 0$, then we can choose an approximate σ by solving for x in $-Ax \log x = \xi$ yielding $x = \exp(W_{-1} \frac{\xi}{A})$ where W_{-1} is the negative branch of the Lambert W-function. This is almost linear and yields approximately $x = O(\xi)$ for small ξ. In fact, we have that $\omega_g(\sigma) \leq C' \sigma^\alpha$ for any $\alpha < 1$, and thus g is "almost" Lipschitz continuous.

Lemma E.3. *The following function is a modulus of continuity function for the TRW$(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ objective over $\mathcal{M}$ with respect to the ℓ_∞ norm:*

$$\omega(\sigma) := \sigma \|\theta\|_1 + 2\sigma \tilde{K} \max\{-\log(2\sigma), 1\}, \tag{21}$$

where $\tilde{K} := 4|V| \max_{ij \in E} VAL_i VAL_j$.

That is, for $\vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\mu}}' \in \mathcal{M}$ with $\|\vec{\boldsymbol{\mu}}' - \vec{\boldsymbol{\mu}}\|_\infty \leq \sigma$, we have:

$$|TRW(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho}) - TRW(\vec{\boldsymbol{\mu}}'; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})| \leq \omega(\sigma).$$

Proof. TRW$(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ can be decomposed into functions of the form $-Kx \log x + \theta x$ (see (15) and (16)) and so we apply the Lemma E.2 element-wise. Let c index the clique component in the marginal vector.

$$\begin{aligned}
|\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho}) - \text{TRW}(\vec{\boldsymbol{\mu}}'; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})| &= \sum_c \sum_{x_c} |g_{c,x_c}(\mu_c(x_c)) - g_{c,x_c}(\mu'_c(x_c))| \\
&\quad \textit{(Using Lemma E.2 and } \|\vec{\boldsymbol{\mu}}' - \vec{\boldsymbol{\mu}}\|_\infty \leq \sigma\textit{)} \\
&\leq \sum_c \sum_{x_c} (|K_c| 2\sigma \max\{-\log(2\sigma), 1\} + |\theta(x_c)|\sigma) \\
&= 2\sigma \max\{-\log(2\sigma), 1\} \sum_c |K_c| n_c + \|\theta\|_1 \sigma,
\end{aligned}$$

where we recall n_c is the number of values that x_c can take. By re-using the bound on $\sum_c |K_c| n_c$ from (18), we get the result. □

E.3 Bounded Negative Uniform Gap

Lemma E.4 (Bound for the negative uniform gap of TRW objective)**.** *For the negative TRW objective $f(\vec{\boldsymbol{\mu}}) := -TRW(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$, the bound B on the negative uniform gap as given in Lemma D.3 for $\boldsymbol{u}_0$ being the uniform distribution can be taken as:*

$$B = 2 \sum_c \max_{x_c} |\theta_c(x_c)| =: 2\|\vec{\boldsymbol{\theta}}\|_{1,\infty} \tag{22}$$

Proof. From Lemma D.3, we want to bound $\|\nabla f(\boldsymbol{u}_0)\|_* = \|\vec{\boldsymbol{\theta}} + \nabla_{\vec{\boldsymbol{\mu}}} H(\boldsymbol{u}_0; \boldsymbol{\rho}))\|_*$. The clique entropy terms $H(\boldsymbol{\mu}_c)$ are maximized by the uniform distribution, and thus $\boldsymbol{u}_0$ is a stationary point of the TRW entropy function with zero gradient. We can thus simply take $B = \|\vec{\boldsymbol{\theta}}\|_* \text{diam}_{\|\cdot\|}(\mathcal{M})$. By taking the ℓ_∞ / ℓ_1 norm on $\mathcal{M}$, we get a diameter of 2, giving the given bound. □

E.4 Summary

We now give the details of suboptimality guarantees for our suggested algorithm to optimize $f(\vec{\boldsymbol{\mu}}) := -\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ over $\mathcal{M}$. The (strong) convexity of the negative TRW objective is shown in (Wainwright et al., 2005; London et al., 2015). $\mathcal{M}$ is the convex hull of a finite number of vectors representing assignments to random variables and therefore a compact convex set. The entropy function is continously differentiable on the relative interior of the probability simplex, and thus the TRW objective has the same property on the relative interior of $\mathcal{M}$. Thus $-\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ satisfies the properties laid out in Problem 3.2.

Lemma E.5 (Suboptimality bound for optimizing $-\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ with the fixed-δ algorithm)**.** *For the optimization of $-TRW(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ over $\mathcal{M}_\delta$ with $\delta \in (0, 1]$, the suboptimality is bounded as:*

$$TRW(\vec{\boldsymbol{\mu}}^*; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho}) - TRW(\vec{\boldsymbol{\mu}}^{(k)}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho}) \leq \frac{2\mathcal{C}_\delta}{(k+2)} + \omega(2\delta), \tag{23}$$

with $\vec{\boldsymbol{\mu}}^*$ the optimizer of $TRW(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ in $\mathcal{M}$, where $C_\delta \leq 16\frac{|V|\max_{(ij)\in E} VAL_i VAL_j}{\delta}$, and $\omega(\sigma) = \sigma\|\vec{\boldsymbol{\theta}}\|_1 + 2\sigma\tilde{K}\max\{-\log(2\sigma), 1\}$, where $\tilde{K} := 4|V|\max_{ij\in E} VAL_i VAL_j$.

Proof. Using $\text{diam}_{\|\cdot\|_{\infty,1}}(\mathcal{M}) \leq 2$, and L_δ from Lemma E.1, we can compute $C_\delta \leq \text{diam}(\mathcal{M})^2 L_\delta$. Lemma E.3 computes the modulus of continuity $\omega(\sigma)$. The rate then follows directly from Theorem C.1. □

Lemma E.6 (Global convergence rate for optimizing $-\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ with the adaptive-δ algorithm)**.** *Consider the optimization of $-TRW(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ over $\mathcal{M}$ with the optimum given by $\vec{\boldsymbol{\mu}}^*$. The iterates $\vec{\boldsymbol{\mu}}^{(k)}$ obtained by running the Frank-Wolfe updates over $\mathcal{M}_\delta$ using line-search with δ updated according to Algorithm 1 (or as summarized in a FCFW variant in Algorithm 4), have suboptimality $h_k = TRW(\vec{\boldsymbol{\mu}}^*; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho}) - TRW(\vec{\boldsymbol{\mu}}^{(k)}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ upper bounded as:*

1. $h_k \leq \left(\frac{1}{2}\right)^k h_0 + \frac{\widetilde{C}}{\delta_0}$ *for k such that* $h_k \geq \max\{B\delta_0, \frac{2\widetilde{C}}{\delta_0}\}$,
2. $h_k \leq \frac{2\widetilde{C}}{\delta_0}\left[\frac{1}{\frac{1}{4}(k-k_0)+1}\right]$ *for k such that* $B\delta_0 \leq h_k \leq \frac{2\widetilde{C}}{\delta_0}$,
3. $h_k \leq \left[\frac{\max(\widetilde{C}, B\delta_0^2)B}{\frac{1}{4}(k-k_1)+1}\right]^{\frac{1}{2}} = O(k^{-\frac{1}{2}})$ *for k such that* $h_k \leq B\delta_0$,

where

- $\delta_0 = \delta^{(\text{init})} \leq \frac{1}{4}$
- $\widetilde{C} := 16|V|\max_{(ij)\in E}(VAL_i VAL_j)$
- $B = 16\|\vec{\boldsymbol{\theta}}\|_{1,\infty}$
- h_0 *is the initial suboptimality*
- k_0 *and* k_1 *are the number of steps to reach stage 2 and 3 respectively which are bounded as:* $k_0 \leq \max(0, \lceil \log_{\frac{1}{2}} \frac{\widetilde{C}}{h_0\delta_0} \rceil)$ $k_0 \leq k_1 \leq k_0 + \max\left(0, \lceil \frac{8\widetilde{C}}{B\delta_0^2} \rceil - 4\right)$

Proof. Using $\text{diam}_{\|\cdot\|_{\infty,1}}(\mathcal{M}) \leq 2$, we bound $\widetilde{C} \leq L\,\text{diam}_{\|\cdot\|_{\infty,1}}(\mathcal{M})^2$ with L (from Property 3.3) derived in Lemma E.1. We bound $-8g_u(\vec{\boldsymbol{\mu}}^{(k)})$ (the upper bound on the negative uniform gap) using the value derived in Lemma E.4. The rate then follows directly from Theorem D.6 using $p = 1$ (see Lemma E.1 where $L_\delta \leq \frac{L}{\delta}$). □

The dominant term in Lemma E.6 is $\widetilde{C}B\,k^{-\frac{1}{2}}$, with $\widetilde{C}B = O(\|\vec{\boldsymbol{\theta}}\|_{1,\infty}|V|)$. We thus find that both bounds depend on norms of $\vec{\boldsymbol{\theta}}$. This is unsurprising since large potentials drive the solution of the marginal inference problem away from the centre of $\mathcal{M}$, corresponding to regions of high entropy, and towards the boundary of the polytope (lower entropy). Regions of low entropy correspond to smaller components of the marginal vector,

which in turn result in larger and poorly behaved gradients of $-\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$, which slows down the resulting optimization.

F Correction and Local Search Steps in Algorithm 2

Algorithm 5 details the **CORRECTION** procedure used in line 16 of Algorithm 2 to implement the correction step of the FCFW algorithm. It uses the modified Frank-Wolfe algorithm (FW with away steps), as detailed in Algorithm 3. Algorithm 6 depicts the **LOCALSEARCH** procedure used in line 17 of Algorithm 2. The local search is performing FW over $\mathcal{M}_\delta$ for a fixed δ using the iterated conditional mode algorithm as an approximate FW oracle. This enables the finding in a cheap of way of more vertices to augment the correction polytope V.

Algorithm 5: Re-Optimizing over correction polytope V using MFW, f is the negative TRW objective

CORRECTION$(\boldsymbol{x}^{(0)}, V, \delta, \boldsymbol{\rho})$
Let $f(\cdot) :=$ -TRW$(\cdot; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$; we use MFW to optimize over the contracted correction polytope $\text{conv}(V_\delta)$ where $V_\delta := (1-\delta)V + \delta \boldsymbol{u}_0$.
Let ϵ be the desired accuracy of the approximate correction.
Let $\boldsymbol{\alpha}^{(0)}$ be such that $\boldsymbol{x}^{(0)} = \sum_{\boldsymbol{v} \in V_\delta} \alpha_{\boldsymbol{v}}^{(0)} \boldsymbol{v}$.
$\boldsymbol{x}^{(\text{new})} \leftarrow$ **MFW**$(\boldsymbol{x}^{(0)}, \boldsymbol{\alpha}^{(0)}, V_\delta, \epsilon)$ *(see Algorithm 3)*
return $\boldsymbol{x}^{(\text{new})}$

Algorithm 6: Local Search using Iterated Conditional Modes, f is the negative TRW objective

LOCALSEARCH$(\boldsymbol{x}^{(0)}, \boldsymbol{v}_{\text{init}}, \delta, \boldsymbol{\rho})$
$\boldsymbol{s}^{(0)} \leftarrow \boldsymbol{v}_{\text{init}}$
$V \leftarrow \emptyset$
for $k = 0 \ldots$ **MAXITS do**
  $\tilde{\theta} = \nabla f(\boldsymbol{x}^{(k)}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$
  $\boldsymbol{s}^{(k+1)} \leftarrow$ICM$(-\tilde{\theta}, \boldsymbol{s}^{(k)})$ *(Approximate FW search using ICM; we initialize ICM at previously found vertex $\boldsymbol{s}^{(k)}$)*
  $\boldsymbol{s}_{(\delta)}^{(k+1)} \leftarrow (1-\delta)\boldsymbol{s}^{(k+1)} + \delta \boldsymbol{u}_0$
  $V \leftarrow V \cup \{\boldsymbol{s}^{(k+1)}\}$
  $\boldsymbol{d}_{(\delta)}^{(k)} \leftarrow \boldsymbol{s}_{(\delta)}^{(k+1)} - \boldsymbol{x}^{(k)}$
  Line-search: $\gamma_k \in \arg\min_{\gamma \in [0,1]} f\left(\boldsymbol{x}^{(k)} + \gamma \boldsymbol{d}_{(\delta)}^{(k)}\right)$
  Update $\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} + \gamma_k \boldsymbol{d}_{(\delta)}^{(k)}$ *(FW update)*
end for
return $\boldsymbol{x}^{(k+1)}, V$

G Comparison to perturbAndMAP

Perturb & MAP. We compared the performance between our method and perturb & MAP for inference on 10 node Synthetic cliques. We expand on the method we used to evaluate perturbAndMAP in Figure 1(e) and 1(f). We re-implemented the algorithm to estimate the partition function in Python (as described in Hazan and Jaakkola (2012), Section 4.1) and used toulbar2 (Allouche et al., 2010) to perform MAP inference over an inflated graph where every variable maps to five new variables. The log partition function is estimated as the mean energy of 10 exact MAP calls on the expanded graph where the single node potentials are perturbed by draws from the Gumbel distribution. To extract marginals, we fix the value of a variable to every assignment, estimate the log partition function of the conditioned graph and compute beliefs based on averaging the results of adding the unary potentials to the conditioned values of the log partition function.

H Correction Steps for Frank-Wolfe over $\mathcal{M}$

Recall that the correction step is done over the correction polytope, the set of all vertices of $\mathcal{M}$ encountered thus far in the algorithm. On experiments conducted over $\mathcal{M}$, we found that using a better correction algorithm often *hurt* performance. This potentially arises in other constrained optimization problems where the gradients are

Figure 4: Chinese Characters : Additional Experiments. TRBP (opt) denotes our implementation of tightening over $\mathbb{T}$ using a wrapper over libDAI (Mooij, 2010)

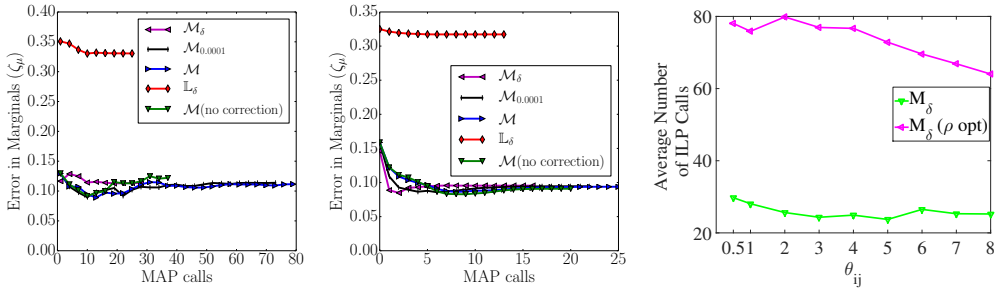
unbounded at the boundaries of the polytope. We found that better correction steps over the correction polytope (the convex hull of the vertices explored by the MAP solver, denoted V in Algorithm 2), often resulted in a solution at or near a boundary of the marginal polytope (shared with the correction polytope). This resulted in the iterates becoming too small. We know that the Hessian of $\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ is ill conditioned near the boundaries of the marginal polytope. Therefore, we hypothesize that this is because the gradient directions obtained when the iterates became too small are simply less informative. Consequently, the optimization over $\mathcal{M}$ suffered. We found that the duality gap over $\mathcal{M}$ would often increase after a correction step when this phenomenon occurred. The variant of our algorithm based on $\mathcal{M}_\delta$ is less sensitive to this issue since the restriction of the polytope bounds the smallest marginal and therefore also controls the quality of the gradients obtained.

I Additional Experiments

For experiments on the 10 node synthetic cliques, we can also track the average number of ILP calls required to converge to a fixed duality gap for any θ. This is depicted in Figure 5(c). Optimizing over $\mathbb{T}$ realized three to four times as many MAP calls as the first iteration of inference.

Figure 4 depicts additional examples from the Chinese Characters test set. Here, we also visualize results from a wrapper around TRBPs implementation in libDAI (Mooij, 2010) that performs tightening over $\mathbb{T}$. Here too we find few gains over optimizing over $\mathbb{L}$.

Figure 5(a), 5(b) depicts the comparison of convergence of algorithm variants over $\mathcal{M}$ and $\mathcal{M}_\delta$ (same setup as Figure 1(a), 1(b). Here, we plot ζ_μ.



(a) ζ_μ: 5×5 grid, $\mathcal{M}$ vs $\mathcal{M}_\delta$ (b) ζ_μ: 10 node clique, $\mathcal{M}$ vs $\mathcal{M}_\delta$ (c) Average ILP calls versus θ: 10 node clique

Figure 5: Figure 5(a), 5(b) depict ζ_μ corresponding to the experimental setup in Figure 1(a), 1(b) respectively. Figure 5(c) explores the average number of ILP calls taken to convergence with and without optimizing over $\boldsymbol{\rho}$

J Bounding $\log Z$ with Approximate MAP Solvers

Suppose that we use an approximate MAP solver for line 7 of Algorithm 2. We show in this section that if the solver returns an *upper bound* on the value of the MAP assignment (as do branch-and-cut solvers for integer linear programs), we can use this to get an upper bound on $\log Z$. For notational consistency, we consider using Algorithm 2 for $\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$, where $f(\boldsymbol{x}) = -\text{TRW}(\vec{\boldsymbol{\mu}}; \vec{\boldsymbol{\theta}}, \boldsymbol{\rho})$ is convex, $\boldsymbol{x} = \vec{\boldsymbol{\mu}}$, and $\mathcal{D} = \mathcal{M}$.

The property that the duality gap may be used as a certificate of optimality (Jaggi, 2013) gives us:

$$f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}^{(k)}) - g(\boldsymbol{x}^{(k)}) \implies -f(\boldsymbol{x}^*) \leq -f(\boldsymbol{x}^{(k)}) + g(\boldsymbol{x}^{(k)}). \tag{24}$$

Adding the gap onto the TRW objective yields an upper bound on the optimum (which from Equation 1 is an upper bound on $\log Z$), i.e. $\log Z \leq -f(\boldsymbol{x}^*)$. From our definition of the duality gap $g(\boldsymbol{x}^{(k)})$ (line 8 in Algorithm 2) and (24), we have:

$$\begin{aligned}\log Z \leq -f(\boldsymbol{x}^*) &\leq -f(\boldsymbol{x}^{(k)}) + \left\langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{s}^{(k)} - \boldsymbol{x}^{(k)} \right\rangle \\ &= -f(\boldsymbol{x}^{(k)}) + \underbrace{\left\langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{s}^{(k)} \right\rangle}_{\text{MAP call}} - \underbrace{\left\langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x}^{(k)} \right\rangle}_{\text{Can be computed efficiently}},\end{aligned}$$

where $\boldsymbol{s}^{(k)} = \arg\min_{\boldsymbol{v}\in\mathcal{D}} \left\langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{v} \right\rangle = \arg\max_{\boldsymbol{v}\in\mathcal{D}} \left\langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{v} \right\rangle$ (line 7 in Algorithm 2). Thus, if the approximate MAP solver returns an upper bound κ such that $\max_{\boldsymbol{v}\in\mathcal{D}} \left\langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{v} \right\rangle \leq \kappa$, then we get the following upper bound on the log-partition function:

$$\log Z \leq -f(\boldsymbol{x}^{(k)}) + \kappa - \left\langle -\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x}^{(k)} \right\rangle. \tag{25}$$

For example, we could use a linear programming relaxation or a message-passing algorithm based on dual decomposition such as Sontag et al. (2008) to obtain the upper bound κ. There is a subtle but important point to note about this approach. Despite the fact that we may use a relaxation of $\mathcal{M}$ such as $\mathbb{L}$ or the cycle relaxation to compute the upper bound, we evaluate it at $\vec{\boldsymbol{\mu}}^{(k)}$ that is *guaranteed* to be within $\mathcal{M}$. This should be contrasted to instead optimizing over a relaxation such as $\mathbb{L}$ directly with Algorithm 2. In the latter setting, the moment we move towards a fractional vertex (in line 14) we would immediately take $\vec{\boldsymbol{\mu}}^{(k+1)}$ out of $\mathcal{M}$. Because of this difference, we expect that this approach will typically result in significantly tighter upper bounds on $\log Z$.

Supplementary References

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1999.

J. Guélat and P. Marcotte. Some comments on Wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, 1986.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, Norwell, MA, 2004.

C. Teo, A. Smola, S. Vishwanathan, and Q. Le. A scalable modular convex solver for regularized risk minimization. In *KDD*, 2007.

P. Wolfe. Convergence Theory in Nonlinear Programming. In J. Abadie, editor, *Integer and Nonlinear Programming*, pages 1–23. North-Holland, 1970.