We introduce a novel technique for computing the permanent of a matrix, which generalizes a technique introduced by Ermon et al. (2013). Our method allows a (16-approximation) factor of of the permanent of a matrix. The permanent of a matrix is a canonical example of a problem defined by eq. (1).

\[ \text{Perm}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} A_{i,\sigma(i)} \]

where \(S_n\) is the symmetric group of \(n\) elements and \(A_{i,j}\) is the \((i,j)\)-th element of \(A\). Clearly, here \(S_n\) is playing the role of \(\Omega\), and \(w(\sigma) = \prod_{i=1}^{n} A_{i,\sigma(i)}\). Therefore computing permanent of a nonnegative matrix is a canonical example of a problem defined by eq. (1).

Similar counting problems arise when one wants to compute the partition functions of the well-known probabilistic generative models of statistical physics, such as the Ising model, or more generally the Ferromagnetic Potts Model (Potts, 1952). Given a graph \(G(V, E)\), and a label-space \(Q \equiv \{0, 1, 2, \ldots, q-1\}\), the partition function \(Z(G)\) of the Potts model is given by:

\[ \sum_{\sigma \in Q^V} \exp \left( -\xi \sum_{(u,v) \in E} \delta(\sigma(u), \sigma(v)) + H \sum_{u \in V} \delta(\sigma(u), 0) \right) \]
where $\zeta$, $J$ and $H$ are system-constants (representing the temperature, spin-coupling and external force respectively), $\delta(x, y)$ is the delta-function that is 1 if and only if $x = y$ and otherwise 0, and $\sigma$ represents a label-vector, where $\sigma(u)$ is the label of vertex $u$.

It has been shown that, under the availability of an NP-oracle, every problem in $\#P$ can be approximated within a factor of $(1 + \epsilon), \epsilon > 0$, with high probability via a randomized algorithm (Stockmeyer, 1985). This result says $\#P$ can be approximated by BP$^{\text{NP}}$ and the power of an NP-oracle and randomization is sufficient. However, depending on the weight function $w(\cdot)$, eq. (1) may not be in $\#P$. There are related approaches to count the number of models of propositional formulas based on SAT-solvers, such as (Birnbaum and Lozinskii, 1999; Jr. and Pehoushek, 2000; Wei and Selman, 2005; Pesant, 2005; Chakraborty et al., 2014; Chakraborty et al., 2016) among others.

The standard techniques to evaluate eq. (1) include the very influential fast variational methods (Wainwright et al., 2008), and Markov-Chain-Monte-Carlo based sampling schemes (Jerrum and Sinclair, 1996). In practice, except for limited number of cases, these approaches are mostly used in a heuristic manner without nonasymptotic qualitative guarantees. Recently, Ermon et al. proposed an alternative approach (that they call WISH - Weighted-Integrals-And-Sums-By-Hashing) to solve these counting problems (Ermon et al., 2013a; Ermon et al., 2013b) by breaking them into multiple optimization problems. Namely, they use families of hash functions $h: \Omega \rightarrow \hat{\Omega}, |\hat{\Omega}| < |\Omega|$, and use a (possibly NP) oracle that can return the correct solution of the optimization problem: $\max_{\sigma, h(\sigma) = a} w(\sigma)$, for any $a \in \hat{\Omega}$. We call this oracle a MAX-oracle. In particular, when $\Omega = \{0, 1\}^n$, and $h(\cdot)$ is a random hash function, assuming the availability of a MAX-oracle, Ermon et al. (Ermon et al., 2013a) propose a randomized algorithm that approximates the discrete sum within a factor of sixteen (a 16-approximation) with high probability. Ermon et al. use simple linear sketches over $\mathbb{F}_2$ (the finite field of size 2), i.e., the hash function $h_{A, b} : \mathbb{F}_2^n \rightarrow \mathbb{F}_2^n, A \in \mathbb{F}_2^{n \times n}, b \in \mathbb{F}_2^n$ is defined to be

$$h_{A, b}(x) = Ax + b,$$

(4)

where the arithmetic operations are over $\mathbb{F}_2$. The matrix $A$ and the vector $b$ are randomly and uniformly chosen from the respective sample spaces. The MAX-oracle in this case simply provides solutions to the optimization problem: $\max_{\sigma \in \mathbb{F}_2^n : A\sigma = b} w(\sigma)$.

The constraint space $\{\sigma \in \mathbb{F}_2^n : A\sigma = b\}$ is nice since it is a coset of the nullspace of $A$, and experimental results showed them to be manageable by optimization softwares/SAT solvers. In particular it was observed that being Integer Programming constraints, real-world instances are often solved in reasonable time. Since the implementation of the hash function heavily affects the runtime, it makes sense to keep constraints of the MAX-oracle as an affine space as above. These constraints are also called parity constraints. The idea of using such constraints to show reduction among class of problems appeared in several papers before, including (Sipser, 1983; Valiant and Vazirani, 1986; Gomes et al., 2006; Thurley, 2011; Gomes et al., 2007) among others. The key property that the hash functions $\{h_{A, b}\}$ satisfy is that they are pairwise independent. This property can be relaxed somewhat - and in a subsequent paper Ermon et al. show that a hash family would work even if the matrix $A$ is sparse and random, thus effectively reducing the randomness as well as making the problem more tractable empirically (Ermon et al., 2014). Subsequently, Achlioptas and Jiang (Achlioptas and Jiang, 2014) have shown another way of achieving similar guarantees. Instead of arriving at the set $\{\sigma \in \mathbb{F}_2^n : A\sigma = b\}$ as a solution of a system of linear equations (over $\mathbb{F}_2$), they view the set as the image of a lower-dimensional space. This is akin to the generator matrix view of a linear error-correcting code as opposed to the parity-check matrix view. This viewpoint allows their MAX-oracle to solve just an unconstrained optimization problem.

**Drawbacks of obvious extensions of (Ermon et al., 2013a) to large alphabets.** Note that, some crucial counting problems, such as computing the partition function of the Ferromagnetic Potts model of Eq. (3), naturally have $\Omega = \{0, 1, \ldots, q - 1\}^n, q > 2$, i.e., a hypergrid. It is worth noting that while there exists polynomial time approximation (FPRAS) for the Ising model ($q = 2$), FPRAS for general Potts model ($q > 2$) is significantly more challenging (and likely impossible (Goldberg and Jerrum, 2012)). There are a few possible obvious extensions of Ermon et al. (Ermon et al., 2013a) to larger alphabets.

- **(The straightforward extension).** The method of (Ermon et al., 2013a) can be used for $q$-ary in stead of binary. However, the drawback is that it provides a $q^2$-approximation at best that is particularly bad if $q$ is large or growing with $n$.

- **(Convert $q$-ary to binary).** To use the binary-domain algorithm of (Ermon et al., 2013a) for any $\Omega = \{0, 1, \ldots, q - 1\}^n$, we need to use a look-up table to map $q$-ary numbers to binary. In this process the number of variables (and also the number of constraints) increases by a factor of $\log q$. This makes the MAX-oracle significantly slower, especially when $q$ is large. Also, for the permanent problem, where $|\Omega| = \exp(n \log n)$, this creates a computational bottleneck. It would be useful to extend the
method of (Ermon et al., 2013a) for $\Omega = \mathbb{F}_q^n$ without increasing the number of variables.

Furthermore, when $q$ is not a power of 2, by converting $q$-ary configurations to binary, we introduce exponentially many invalid configurations. To account for these, the MAX-oracle must be adjusted accordingly which is a difficult task. This motivates us to keep the problem in its original domain and not convert the domain to binary.

- For the binary setting, it has been noted in (Ermon et al., 2013a, section 5.3) that the approximation ratio can be improved to any $a > 1$ by increasing the number of variables, which extends to this $q$-ary setting. However this also results in an increase in number of variables by a factor of $\log_a(q^2)$ which is undesirable.

**Our contributions.** Our first contribution in this paper is to provide a new and improved algorithm to handle counting problems over nonbinary domains. For any hypergrid $\Omega = \{0, 1, \ldots, q - 1\}^n$, $q$ is a power of prime, our algorithm provides a $4(1 + \frac{1}{q})^2$-approximation, when $q$ is odd, and $4(1 + \frac{1}{q^2})^2$-approximation, when $q > 2$ is even, to the optimization problem of (1) assuming availability of the MAX-oracle. Our algorithm utilizes an idea of using optimization over multiple bins of the hash function that can be easily implemented via inequality constraints. The constraint space of the MAX-oracle remains an affine space and still can be represented as a modular integer linear program (ILP). Our multi-bin technique can also be used to extend the generator-matrix based algorithm of Achlioptas and Jiang (Achlioptas and Jiang, 2015). As a result, we need the MAX-oracle to only perform unconstrained maximization, as opposed to constrained. This lead to significant speed-up in the system, while resulting in the same approximation guarantees.

Finally, we show the performance of our algorithms to compute the partition function of the ferromagnetic Potts model by running experiments on both synthetic datasets and real-worlds datasets. While in this paper we concentrate on theoretical results, the experiments serve as good ‘proof of concepts’ for applications. We also use our algorithm to compute the Total Variation (TV) distance between two joint probability distributions over a large number of variables. In addition to comparing with the straightforward generalization of Ermon et al.’s method (Ermon et al., 2013a), we also show comparisons with the popular Markov-Chain-Monte-Carlo (MCMC) method and the belief propagation method for discrete integration. All the experiments exhibit good performance guarantees.

**Organization.** In Section 2 we describe the technique by (Ermon et al., 2013a) called the WISH algorithm, and then elaborate our new ideas and main results. In Section 3, we provide the main technical results that lead to an improved approximation. We provide an algorithm with unconstrained optimization oracle (similar to (Achlioptas and Jiang, 2015)) and its analysis in Section 4. The experimental results on computation of partition functions and total variation distance are provided in Section 5. Most of the proofs and some experimental results are delegated to the appendix in the supplementary material.

While only of auxiliary interest here, we note that it is possible to derandomize the hash families based on parity-constraints to the optimal extent while maintaining the essential properties necessary for their performance. Namely, it can be ensured that the hash family can still be represented as $\{x \mapsto Ax + b\}$ while using information theoretically optimal memory to generate them. We discuss this in Appendix C in the supplement.

It turns out that, by using our technique and some modifications to the MAX-oracle, it is possible to obtain close-to-4-approximation to the problem of computing permanent of nonnegative matrices (assuming existence of NP-oracles). The NP-oracle still is amenable to be implemented in a commercial optimization solver. The idea of optimization over multiple bins is crucial here, since the straightforward generalization of Ermon et al.’s result would have given an approximation factor of $\Omega(n^2)$. Since there exists polynomial time randomized approximation scheme $(1 + \varepsilon$-approximation) of permanent of a nonnegative matrix (Jerrum et al., 2004), the point of this exercise is to show that our method extends to find permanent of a matrix (albeit not with the best guarantees). We discuss this in Appendix D in the supplementary material.

### 2 BACKGROUND AND OUR TECHNIQUES

In this section we describe the main ideas developed by (Ermon et al., 2013a) and provide an overview of the techniques that we use to arrive at our new results.

Let the elements in $\Omega$ be $\sigma_1, \sigma_2, \ldots, \sigma_{|\Omega|}$ arranged according to a decreasing order of their weight, i.e., $w(\sigma_1) \geq w(\sigma_2) \geq \cdots \geq w(\sigma_{|\Omega|})$. Let $\beta_i = w(\sigma_i)$, for $i = 0, 1, \ldots, n'$, where $n'$ is the smallest integer such that $q^{n'} \geq |\Omega|$. When $q^n > |\Omega|$ we set $\beta_0 = 0$.

Clearly $\beta_0 \geq \beta_1 \geq \cdots \geq \beta_{n'}$. As we have not made any assumption on the values of the weight function, $\beta_i$ and $\beta_{i+1}$ can be far from each other. On the other hand we can try to bound the sum $S_{\Omega}(w)$ by bounding the area of the slice between $\beta_i$ and $\beta_{i+1}$. This area is at least $q^i(\beta_i - \beta_{i+1})$ and at most $q^{i+1}(\beta_i - \beta_{i+1})$. Therefore: $\sum_{i=0}^{n'-1} q^i(\beta_i - \beta_{i+1}) + q^{n'}\beta_{n'} \leq S_{\Omega}(w) \leq \sum_{i=0}^{n'-1} q^{i+1}(\beta_i - \beta_{i+1}) + q^{n'}\beta_{n'}$
which implies

$$\beta_0 + (q - 1) \sum_{i=1}^{n'} q^{-i} \beta_i \leq S_\Omega(w) \leq \beta_0 + (q - 1) \sum_{i=1}^{n'} q^i \beta_i.$$  

(5)

Figure 1: The $T(u)$ vs. $u$ curve and the illustration of $\beta_i$s.

Hence $\beta_0 + (q - 1) \sum_{i=1}^{n'} q^{-i} \beta_i$ is a $q$-factor approximation of $S_\Omega(w)$ and if we are able to find a $k$-approximation of each value of $\beta_i$ we will be able to obtain a $kq$-factor approximation of $S_\Omega(w)$. In (Ermon et al., 2013a), subsequently the main idea is to estimate the coefficients $\{\beta_i, 0 \leq i \leq n'\}$. Now note that, $q^i = |\{\sigma \in \Omega : w(\sigma) \geq \beta_i\}|$, for $i = 0, 1, \ldots, n' - 1$. This also hold for $i = n'$ unless $q^{n'} > |\Omega|$ in which case $\beta_{n'} = 0$. Suppose, using a random hash function $h : \Omega \to \{0, 1, \ldots, q^i - 1\}$ we compute hashes of all elements in $\Omega$. The pre-image of an entry in $\{0, 1, \ldots, q^i - 1\}$ is called the bin corresponding to that value, i.e., $\{\sigma \in \Omega : h(\sigma) = x\}$ is the bin corresponding to the value $x \in \{0, 1, \ldots, q^i - 1\}$. In every bin for the hash function, there is on average one element $\sigma$ such that $w(\sigma) \geq \beta_i$. So for a randomly and arbitrarily chosen bin $x \in \{0, 1, \ldots, q^i - 1\}$, if $w^* = \max_{\sigma : h(\sigma) = x} w(\sigma)$, then $w^*$ is a ‘good’ approximation of $\beta_i$ (this will be made rigorous later). Indeed, suppose one performs this random hashing $\ell = O(\log n')$ times and then take the aggregate (in this case the median) value of $w^*$’s. That is, let $w^* = \text{median}(w^*_1, \ldots, w^*_\ell)$; then by using the independence of the hash functions, it can be shown that the aggregate is an upper bound on $\beta_i$ with high probability. In (Ermon et al., 2013a), $\Omega = \mathbb{F}_q^n$ and if the hash family is pairwise independent, then by using the Chebyshev inequality it was shown that $w^* \in [\beta_{i+2}, \beta_{i-2}]$ with high probability. The WISH algorithm proposed by (Ermon et al., 2013a) makes use of the above analysis and provides a $2^{2/3} = 16$-approximation of $S_\omega(\Omega)$. If we naively extend this algorithm for $S_\omega(\Omega) = \mathbb{F}_q^n, q > 2$, then it can be shown that $w^* \in [\beta_{i+1}, \beta_{i-1}]$ with high probability. This gives an approximation factor of $q^2$. E.g., for a ternary alphabet, $\Omega = \mathbb{F}_q^n$, we have a 9-approximation to $S_\omega(\Omega)$.

Instead of using a straightforward analysis for the $q$-ary case, in this paper we use a MAX-oracle that can optimize over multiple bins of the hash function. Using this oracle we proposed a modified WISH algorithm and call it MB-WISH (Multi-Bin WISH). Just as in the case of (Ermon et al., 2013a; Ermon et al., 2014), the MAX-oracle constraints can be integer linear programming constraints and commercial softwares such as CPLEX can be used. The main intuition of using an optimization over multiple bins is that it boosts the probability that the $w^*$ we are getting above is close to $\beta_i$. To be precise, we redefine $\beta_i = w(\sigma[i:j])$ for $i = 1, 2, \ldots, n' \equiv [n \log_q r]$.

If we define $T(u) \equiv |\{\sigma \in \Omega : w(\sigma) \geq u\}|$, then Figure 1 illustrate the $T(u)$ vs. $u$ curve and locates $\beta_i$s therein. Note that, we would like to find the area under the $T(u)$ vs. $u$ curve, for which we use the sum of the vertical slices. Now to estimate the new $\beta_i$, we choose a hash function as before, and optimize over $r^i$ bins of the hash function. These steps are made rigorous in Section 3. However if we restrict ourselves to the binary alphabet then (as will be clear later) there is no immediate way to represent such multiple bins in a compact way in the MAX-oracle. For the non-binary case, it is possible to represent multiple bins of the hash function as simple inequality constraints.

This idea leads to an improvement in the approximation factor of $S_\omega(\Omega)$ to $4 + \epsilon$, where $\epsilon$ decays to 0 proportional to $q^{-1}$. Note that we need to choose $q$ to be a power of prime so that $\mathbb{F}_q$ is a field.

In (Achlioptas and Jiang, 2015), the bins (as described above) are produced as images of some function, and not as pre-images of hashes. Since we want the number of bins to be $q^i$, this can be achieved by looking at images of $q : \mathbb{F}_q^n \to \Omega$ where $|\{\sigma \in \mathbb{F}_q^n : \sigma[i,j] = x\}| \leq q^{n-1}$. The rest of the analysis of (Achlioptas and Jiang, 2015) is almost same as above. The benefit of this approach is that the MAX-oracle just has to solve an unconstrained optimization here. Implementing our multi-bin idea for this perspective of (Achlioptas and Jiang, 2015) is not straightforward as we can no longer use inequality constraints for this. However, as we show later, we found a way to combine bins here in a succinct way generalizing the design of $g$. As a result, we get the same approximation guarantee as in MB-WISH, with the oracle load heavily reduced (this algorithm, that we call Unconstrained MB-WISH, can be found in Section 4).

3 THE MB-WISH ALGORITHM AND ANALYSIS

Let us assume $\Omega = \mathbb{F}_q^n$ where $q$ is a prime-power. Let us also fix an ordering among the elements of $\mathbb{F}_q^n \equiv \{a_0, a_1, \ldots, a_{q-1}\}$ and write $a_0 < a_1 < \cdots < a_{q-1}$. In
this section, the symbol ‘<’ just signifies a fixed ordering and has no real meaning over the finite field. Extending this notation, for any two distinct vectors \(x, y \in \mathbb{F}_q^m\), we will say \(x < y\) if and only if the \(i\)th coordinates of \(x\) and \(y\) satisfy \(x_i < y_i\) for all \(i = 1, \ldots, m\). Below 1 denotes an all-one vector of a dimension that would be clear from context. Also, for any event \(E\) let \(1[E]\) denote the indicator for the event \(E\).

The MAX-oracle for \(\text{MB-WISH}\) performs the following optimization, given \(A \in \mathbb{F}_q^{m \times n}, b \in \mathbb{F}_q^m\):

\[
\max_{\sigma \in \mathbb{F}_q^m \setminus Aa + b < s} w(\sigma).
\]

The modified \(\text{WISH}\) algorithm is presented as Algorithm 1. The main result of this section is below.

**Algorithm 1** \(\text{MB-WISH}\) algorithm for \(\Omega = \mathbb{F}_q^n\), a weight function \(w\) and an input parameter \(r \leq \lfloor \frac{q-1}{2} \rfloor\)

**Initialize:**
\[ y = \frac{q}{3r} \left( \frac{1}{2} - \frac{q}{2} \right), \quad \ell = \left\lfloor \frac{1}{r} \ln \frac{2n}{\delta} \right\rfloor, \quad n' = \left\lceil \frac{n \log_q q/r}{\Delta} \right\rceil, \quad M_0 = \max_{\sigma \in \mathbb{F}_q^m} w(\sigma)\]

for \(j \in \{1, 2, \ldots, n'\}\) do

for \(k \in \{1, \ldots, \ell\}\) do

Sample hash functions \(h_{j,k} \equiv h_{A^j,b}^i\), uniformly at random from \(\mathcal{H}_{n,R}\) as defined in (7)

\[ w_{j,k}^i = \max_{A^j,b} w(\sigma) \]

end for

\[ M_{i,j} = \text{Median}(w_{j,1}^{(i)}, w_{j,2}^{(i)}, \ldots, w_{j,\ell}^{(i)}) \]

end for

Return \(M_0 + \left( \frac{q}{2} - 1 \right) \sum_{i=0}^{n'} M_{i+1} \left( \frac{q}{2} \right)^i\)

**Theorem 1.** Suppose \(q > 2\) is a prime power, \(\Omega = \mathbb{F}_q^n\) and \(r \leq \lfloor \frac{q-1}{2} \rfloor\) is a positive integer. For any \(\delta > 0\), Algorithm 1 makes \(\Theta(n \log \frac{q}{r})\) calls to the MAX-oracle, and with probability \(\geq 1 - \delta\) outputs a \((\frac{q}{2})^2\)-approximation of \(S_{\alpha}(\Omega)\).

By setting \(r = \lfloor \frac{q-1}{2} \rfloor\), our algorithm provides a \(4(1 + \frac{1}{q-1})^2\)-approximation, when \(q\) is odd, and \(4(1 + \frac{2}{q-2})^2\)-approximation, when \(q > 2\) is even.

The constant in the big-O term in the number of calls to the oracle is a function of \(q\) and \(r\). In particular, when \(r = \lfloor \frac{q-1}{2} \rfloor\) and \(q\) odd, this constant varies as \(q^2 \log q\). We can tune the value of \(r\) to reduce the number of calls to the oracle at the expense of the approximation factor.

The theorem will be proved by a series of lemmas. The key trick that we are using is to ask the MAX-oracle to solve an optimization problem over not a single bin, but multiple bins of the hash function. This is going to boost the probability that our estimates of \(\beta_i\)'s are good. In particular we will solve the optimization over \(r^m\) bins of the hash function. The hash family is defined in the following way. We have \(h_{A,b} : \mathbb{F}_q^m \rightarrow \mathbb{F}_q^m : x \mapsto Ax + b\), the operations are over \(\mathbb{F}_q\). Let

\[
\mathcal{H}_{m,n} = \{h_{A,b} : A \in \mathbb{F}_q^{m \times n}, b \in \mathbb{F}_q^m\}.
\]

For readers familiar with coding theory, the basis behind our technique is simple. The set of configurations \(\{\sigma \in \mathbb{F}_q^m : A\sigma = 0\}\) forms a linear code of dimension \(n - m\). The bins of the hash function define the cosets of this linear code. We would like to choose \(q^r\) cosets of a random linear code and find the optimum value of \(w\) over the configurations of these cosets as the MAX-oracle. To choose a hash function uniformly and randomly from \(\mathcal{H}\), we can just choose the entries of \(A\) and \(b\) uniformly at random from \(\mathbb{F}_q\) independently.

Note that, the hash family \(\mathcal{H}_{m,n}\) as defined in (7) is uniform and pairwise independent.

**Lemma 1.** Let us define \(Z_\sigma\) to be the indicator random variable denoting \(A\sigma + b < \sigma\cdot1\) for some \(r \in \{0, \ldots, q-1\}\) and \(A, b\) randomly and uniformly sampled from \(\mathcal{H}_{m,n}\). Then \(\Pr(Z_\sigma = 1) = \left(\frac{q}{2}\right)^m\) and for any two distinct configurations \(\sigma_1, \sigma_2 \in \mathbb{F}_q^m\) the random variables \(Z_{\sigma_1}\) and \(Z_{\sigma_2}\) are independent.

Fix an ordering of the configurations \(\{\sigma_i, 1 \leq i \leq q^n\}\) such that \(1 \leq j \leq q^r, w(\sigma_j) \geq w(\sigma_{j+1})\). For \(i \in \{0, 1, 2, \ldots, q^n\}\), define \(\beta_i = \frac{w(\sigma_i)}{w(\sigma_{i+1})}\), where \(t = \frac{q^n}{2}\). We take \(w(\sigma_k) = 0\) for \(k > q^n\). See Figure 1 for an illustration.

To prove Thm. 1 we need the following crucial lemma.

**Lemma 2.** Let \(M_i = \text{Median}(w_{j,1}^{(i)}, \ldots, w_{j,\ell}^{(i)})\) be defined as in the Algorithm 1. Then for \(\gamma = \frac{q}{3r} \left( \frac{1}{2} - \frac{q}{2} \right)^2\), we have

\[
\Pr\left( M_i \in [\beta_{\min(i+1,n'), \beta_{\max(i-1,0)}]} \right) \geq 1 - 2\exp(-\gamma\ell).
\]

From Lemma 2, the output of the algorithm lies in the range \([L', U']\) with probability at least \(1 - \delta\) where \(L' = \beta_0 + (t - 1) \sum_{i=0}^{n'-1} \beta_{\min(i+2,n')} t^i\) and \(U' = \beta_0 + (t - 1) \sum_{i=0}^{n'-1} \beta_i t^i\). \(L'\) and \(U'\) are a factor of \(t^i\) apart. Now, following an argument similar to (5), we can show \(L' \leq S_{\alpha}(\Omega) \leq U'\). Therefore Algorithm 1 provides a \(t^2\)-approximation to \(S_{\alpha}(\Omega)\) and the total number of calls to the MAX-oracle is \(n't' + 1 = O(n \log(n/\delta))\). The full proof of Theorem 1 is deferred to the Appendix A in the supplementary material.

To exemplify this result, suppose \(q = 3\). In this case the algorithm provides a 9-approximation. Later, in the experimental section, we have used a ferromagnetic Potts model with \(q = 5\). \(\text{MB-WISH}\) provides a \(\frac{3}{5} = 0.625\)-approximation in that case. Note that, for a 5-ary Potts model, it is only natural to use our algorithm instead of
4 MB–WISH WITH UNCONSTRAINED OPTIMIZATION ORACLE

In this section, we modify and generalize the results of Achlioptas and Jiang (Achlioptas and Jiang, 2015) to formulate a version of MB-WISH that can use unconstrained optimizers as the MAX-oracle. We call this algorithm Unconstrained MB-WISH. Let us assume $\Omega = \mathbb{F}_q^m$ where $q$ is a prime-power. As before, let us fix an ordering among the elements of $\mathbb{F}_q \equiv \{a_0, a_1, \ldots, a_{q-1}\}$ and write $a_0 < a_1 < \cdots < a_{q-1}$. Recall that, here the symbol $<$ signifies a fixed ordering and has no real meaning over the finite field.

The MAX-oracle for Unconstrained MB-WISH performs an unconstrained optimization of the following form, given $A \in \mathbb{F}_q^{n \times m}, b \in \mathbb{F}_q^n$ and a set $B \subseteq \mathbb{F}_q^m$:

$$\max_{\sigma \in B} w(A \sigma + b). \quad (8)$$

The aim is to carefully design $B$ so that all the desirable statistical properties are satisfied. This part is quite different from the hashing-based analysis and not an immediate extension of (Achlioptas and Jiang, 2015). We provide the algorithm (Unconstrained MB-WISH) and its analysis in the next section.

The Unconstrained MB-WISH algorithm is presented as Algorithm 2. The main result of this section is the following.

**Theorem 2.** Suppose $q > 2$ is a power of a prime and a positive integer $r \leq \lfloor \frac{n}{2} \rfloor$. Let $\Omega = \mathbb{F}_q^n$. For any $\delta > 0$, Algorithm 2 makes $\Theta(n \log \frac{2}{\delta})$ calls to the MAX-oracle (cf. (8)), and with probability at least $1 - \delta$ outputs a $(\frac{2}{\delta})^2$-approximation of $S_w(\Omega)$.

To prove this theorem we borrow some ideas from coding theory. We define a linear $q$-ary code $C$ of dimension $n - m$ and length $n$ as the set of vectors $\{Ax : x \in \mathbb{F}_q^{n-m}\}$ where $A$ is a full-rank matrix of size $n \times n - m$ and rank $n - m$. For a vector $a \in \mathbb{F}_q^n$, we define the set $\{a + C\}$ as a coset of $C$. It is well known that $\mathbb{F}_q^n$ is partitioned by the $q^m$ distinct cosets, each of size $q^{n-m}$. The main technique behind our algorithm is that for a random linear code $C$ of size $q^{n-m}$, we randomly sample $r^m$ distinct cosets of $C$. Subsequently, we find the maximum value $w(x)$ of an element among those $r^m$ cosets.

Let $E \in \mathbb{F}_q^{n \times n}$ be an $n \times n$ full rank matrix randomly and uniformly chosen from the set of all $n \times n$ rank-$n$ matrices over $\mathbb{F}_q$. One can choose such a matrix via rejection sampling: independently and uniformly sample the entries of the matrix from $\mathbb{F}_q$ and then reject the matrix and resample it if it is not full rank. Let $A$ denote the random matrix formed by the first $n - m$ columns of $E$ as columns and let $R$ be the random matrix formed by the remaining $m$ columns of $E$ as columns. Also let $b$ be a vector sampled randomly and uniformly from $\mathbb{F}_q^n$. The MAX-oracle for Unconstrained MB-WISH is going to perform the following optimization when $m \leq n$:

$$\max_{\sigma_1 \in \mathbb{F}_q^{n-m}, \sigma_2 \in \{a_0, a_1, \ldots, a_{q-1}\}^m} w(A\sigma_1 + R\sigma_2 + b). \quad (9)$$

Analogous to Theorem 1, here we are creating union of $r^m$ distinct random bins. If we can prove that, for any element of $\mathbb{F}_q^{n-m}$, the probability that it belongs to one of these bins is $\left(\frac{q}{q^m}\right)^m$ and for any pair of different elements from $\mathbb{F}_q^n$, whether they belong to one of these bins are independent (pairwise independence), the rest of the proof of Theorem 2 will just follow that of Theorem 1.

In particular, we just have to prove the lemma that is

---

**Algorithm 2** Unconstrained MB-WISH algorithm for $\Omega = \mathbb{F}_q^n$ and a weight function $w$

Initialize: $\ell \rightarrow \lceil \frac{1}{2} \ln \frac{2n}{\delta} \rceil, r, n' = \lceil n \log_q r, q \rceil$

$M_0 \equiv \max_{\sigma \in \mathbb{F}_q^n} w(\sigma)$

for $i \in \{1, 2, \ldots, n\}$ do

for $k \in \{1, \ldots, \ell\}$ do

Sample a full rank matrix uniformly at random from the set of all full rank $n \times n$ matrices in $\mathbb{F}_q^{n \times n}$ and construct matrices $A$ and $R$ by taking the first $n - i$ columns and the last $i$ columns respectively.

Sample $b \in \mathbb{F}_q^n$ uniformly at random

$w_i(k) = \max_{x \in \mathbb{F}_q^i} w(Ax + Ry + b)$

end for

$M_i = \text{Median}(w_i(1), w_i(2), \ldots, w_i(\ell))$

end for

for $i \in \{n + 1, \ldots, n'\}$ do

for $k \in \{1, \ldots, \ell\}$ do

Sample full rank matrix $A \in \mathbb{F}_q^{n \times n}, b \in \mathbb{F}_q^n$ uniformly at random. Set $S_i$ as defined in Equation (10)

$w_i(k) = \max_{y \in S_i} w(Ay + b)$

end for

$M_i = \text{Median}(w_i(1), w_i(2), \ldots, w_i(\ell))$

end for

Return $M_0 + (\frac{q}{2} - 1) \sum_{i=0}^{n'} M_i + (\frac{q}{2})^i$
analogous to Lemma 1. Define a set

\[ S_{A,R,b} \equiv \{ Ax + b + Ry \mid x \in \mathbb{F}_q^{m-n}, y \in \{a_0, a_1, \ldots, a_r\}^m \}. \]

For each configuration \( \sigma \in \mathbb{F}_q^n \), associate an indicator random variable \( Z_\sigma \) denoting whether \( \sigma \in S_{A,R,b} \).

**Lemma 3.** For each configuration \( \sigma \in \mathbb{F}_q^n \), we must have \( \Pr(Z_\sigma = 1) = \left( \frac{q}{r} \right)^m \) and moreover for any two distinct configurations \( \sigma_1, \sigma_2 \in \mathbb{F}_q^n \), we must have \( \Pr(Z_{\sigma_1} = 1 \land Z_{\sigma_2} = 1) \leq (\Pr(Z_\sigma = 1))^2 \).

Although the two random variables \( Z_{\sigma_i} \) and \( Z_{\sigma_j} \) defined above are not independent, we show that they are negatively correlated. Note that, the pairwise independence was then subsequently used in computing a variance for the Chebyshev’s inequality (see Lemma 2). However, the negative correlation is sufficient to obtain an upper bound on the variance. From Algorithm 2 it is clear that Lemma 3 allows us to obtain the values of \( M_i \) for \( i \in \{1, 2, \ldots, n\} \). Indeed, the MAX-oracle is not well defined when \( m > n \). In order to obtain the values of \( M_i \) for \( i \in \{n + 1, \ldots, n^‘\} \), we propose the following technique.

Recall that the elements of \( \mathbb{F}_q^n \) can be represented as \( n \) dimensional vectors where each element belongs to \( \mathbb{F}_q \). Moreover we defined an ordering over the elements of the finite field \( \mathbb{F}_q \equiv \{0, 1, \ldots, q-1\} \) so that \( a_i < a_j \) for \( i < j \). Consider the lexicographic ordering of the elements (vectors) of \( \mathbb{F}_q^m \). Let \( s_m \) be the \( \lfloor \frac{m}{q^{n-m}} \rfloor \) th element in this ordering of \( \mathbb{F}_q^n \). Define the set

\[ S_m = \{ x \in \mathbb{F}_q^n \mid x < s_m \} \quad (10) \]

for all \( m > n \). Now, let \( A \in \mathbb{F}_q^{n \times n} \) be an \( n \times n \) full rank matrix randomly and uniformly chosen from the set of all \( n \times n \) rank-\( n \) matrices over \( \mathbb{F}_q \), which can be generated by rejection sampling as before. Let \( b \in \mathbb{F}_q^n \) be a uniform random vector. Subsequently, the MAX-Oracle for Unconstrained MB-WISH solves the following optimization problem for \( m > n \):

\[ \max_{y \in S_m} w(Ay + b). \]

In order to analyze the statistical properties of this oracle, define the random set \( T_{A,b,m} \equiv \{ Ay + b \mid y \in S_m \} \). Again, for each configuration \( \sigma \in \mathbb{F}_q^n \), associate an indicator random variable \( Z_\sigma \) denoting \( \sigma \in T_{A,b,m} \).

**Lemma 4.** For each configuration \( \sigma \in \mathbb{F}_q^n \), we must have \( \left( \frac{q}{r} \right)^m - 1 \leq \Pr(Z_{\sigma} = 1) \leq \left( \frac{q}{r} \right)^m \) and moreover for any two distinct configurations \( \sigma_1, \sigma_2 \in \mathbb{F}_q^n \), \( \Pr(Z_{\sigma_1} = 1 \land Z_{\sigma_2} = 1) \leq (\Pr(Z_\sigma = 1))^2 \).

Given the two lemmas, the remainder of the proof of Theorem 2 follows that of Theorem 1 straightforwardly.

## 5 EXPERIMENTAL RESULTS

All the experiments were performed in a shared parallel computing environment that is equipped with 50 compute nodes with 28 cores Xeon E5-2680 v4 2.40GHz processors with 128GB RAM. Further experiments on estimating the TV distance is reported in Appendix B.

**Experiments on simulated Potts model (regular degree graph).** We implemented our algorithm to estimate the partition function of Potts Model. Recall that the partition function of the Potts model on a graph \( G = (V, E) \) is given in Eq. (3). First of all, we computed partition functions for small graphs where a brute-force algorithm can also be used to compute the ground truth function values. For our simulation, we have randomly generated the graph \( G \) with number of nodes \( n \equiv |V| \) varying in 4, 5, 6, 7, 8, 9, and corresponding regular degree \( d = 2, 4, 4, 4, 4 \), using a python library `networkx`. We took the number of states of the Potts model \( q = 5 \), the external force \( H \) and the spin-coupling \( J \) to be 0.1 and then varied the values of \( \zeta \). The partition functions for different cases are calculated using both brute force and our algorithm (MB-WISH). We have used a python module `constraint` to handle the constrained optimization for MAX-oracle. The obtained approximation factors for different \( \zeta \) are listed in Table 1. The worst approximation factor observed in all these trials is 5.442. Note that the theoretical guarantee on the approximation ratio for this setting obtained from Theorem 1 is 6.25. This experiment shows that, for small graphs the partition functions computed by MB-WISH are good approximations to the actual values.

For graphs with larger number of vertices, it is not possible to compute the ground truth partition function of Potts Model by brute force. Therefore, we compare the partition function computed by Unconstrained MB-WISH (\( \bar{Z} \)) with two standard techniques: Belief propagation (BP) (Koller and Friedman, 2009) and Markov-Chain-Monte-Carlo (MCMC) (Jerrum and Sinclair, 1996). It is known that BP provides exact result when the underlying graph is cycle-free (Koller and Friedman, 2009). To implement this we use the PGMPY library in python (PGM, ). For MCMC, we employ the popular Metropolis-Hastings (MH) algorithm (Kroese et al., 2011) to sample random points from \( \Omega \), where we evaluate the function \( w: \Omega \rightarrow \mathbb{R} \) and take a scaled-sum to estimate the discrete integration problem. We have calculated the average of the partition function over 10 different trials of the MH algorithm, and each trial was given the same time as that of Unconstrained MB-WISH.

Again, for our simulation, we have randomly generated the graph \( G \) with number of nodes \( n \equiv |V| \) vary-
where there are exactly two nodes of degree \( n \), and every other node has degree 2. We perform the experiment with the number of nodes \( n = \lfloor V \rfloor \) varying in 20, \ldots, 50 on a path-graph such that the number of states \( q = 31 \) and the external parameters \( J = 0.1, H = 0.5 \) and \( \zeta = -5 \). For two different values of \( r \), respectively 1 (single-bin) and 15 (multi-bin) we compute the estimates of the partition function. We have plotted the ratio of the estimates with the corresponding ones computed by \( \text{BP} \) (which is exact), in Figure 2. It is clear from the figure and the table that the \text{Unconstrained MB-WISH} performs much better than its single-bin counterpart. The timeout for each call to the oracle is chosen to be \( n/10 \) where \( n \) is the number of nodes in the graph.

### Real-world constraint satisfaction problem (CSPs)

Many instances of real-world graphical models are available in [http://www.cs.huji.ac.il/project/PASCAL/showExample.php](http://www.cs.huji.ac.il/project/PASCAL/showExample.php). Notably, some of them (e.g., image alignment, protein folding) are defined on non-Boolean domains, which justify the use of \text{MB-WISH}. We have computed the partition functions for several of them.

The dataset \text{Network.uai} is a Markov network with 120 nodes each having a binary value. A configuration here is a binary sequence of length 120. To calculate the partition function, we need to find the sum of weights for \( 2^{120} \) different configurations. In order to use \text{Unconstrained MB-WISH}, we view each configura-

### Table 1: Log-partition function computed by \text{MB-WISH} \((r = 2)\) and the actual value calculated by brute force: \( \hat{Z}_r \).

<table>
<thead>
<tr>
<th>( \zeta )</th>
<th>0.976</th>
<th>1.220</th>
<th>0.610</th>
<th>1.907</th>
<th>0.953</th>
<th>1.192</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.580</td>
<td>0.708</td>
<td>1.639</td>
<td>0.755</td>
<td>0.630</td>
<td>0.599</td>
</tr>
<tr>
<td>10</td>
<td>0.747</td>
<td>1.191</td>
<td>3.271</td>
<td>0.989</td>
<td>1.875</td>
<td>1.25</td>
</tr>
<tr>
<td>15</td>
<td>1.430</td>
<td>1.036</td>
<td>1.013</td>
<td>1.224</td>
<td>1.399</td>
<td>1.692</td>
</tr>
<tr>
<td>20</td>
<td>1.032</td>
<td>1.590</td>
<td>1.141</td>
<td>1.173</td>
<td>1.365</td>
<td>1.491</td>
</tr>
<tr>
<td>25</td>
<td>0.839</td>
<td>1.118</td>
<td>1.339</td>
<td>1.035</td>
<td>1.429</td>
<td>1.326</td>
</tr>
<tr>
<td>30</td>
<td>0.510</td>
<td>4.056</td>
<td>2.226</td>
<td>1.060</td>
<td>0.690</td>
<td>2.122</td>
</tr>
<tr>
<td>35</td>
<td>1.073</td>
<td>5.442</td>
<td>0.489</td>
<td>2.871</td>
<td>1.639</td>
<td>1.263</td>
</tr>
<tr>
<td>40</td>
<td>1.210</td>
<td>2.434</td>
<td>0.980</td>
<td>0.582</td>
<td>0.666</td>
<td>0.969</td>
</tr>
<tr>
<td>45</td>
<td>1.127</td>
<td>4.640</td>
<td>2.348</td>
<td>1.336</td>
<td>0.3673</td>
<td>1.341</td>
</tr>
<tr>
<td>50</td>
<td>1.152</td>
<td>1.025</td>
<td>2.511</td>
<td>3.4307</td>
<td>1.1522</td>
<td>2.636</td>
</tr>
</tbody>
</table>

### Table 2: Log-partition function computed by \text{unconstrained MB-WISH}, \text{Belief Propagation (BP)} and Markov Chain Monte Carlo (MCMC) respectively for the cases of \( \zeta = 1, 2 \) and 5.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \zeta = 1 )</th>
<th>( \zeta = 2 )</th>
<th>( \zeta = 5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MB-WISH</td>
<td>BP</td>
<td>MCMC</td>
</tr>
<tr>
<td>10</td>
<td>15.16</td>
<td>15.51</td>
<td>12.60</td>
</tr>
<tr>
<td>15</td>
<td>23.10</td>
<td>23.27</td>
<td>20.51</td>
</tr>
<tr>
<td>20</td>
<td>31.04</td>
<td>31.03</td>
<td>28.69</td>
</tr>
<tr>
<td>25</td>
<td>38.28</td>
<td>38.79</td>
<td>36.63</td>
</tr>
<tr>
<td>30</td>
<td>46.23</td>
<td>46.55</td>
<td>44.49</td>
</tr>
<tr>
<td>40</td>
<td>61.88</td>
<td>62.06</td>
<td>59.75</td>
</tr>
<tr>
<td>50</td>
<td>77.31</td>
<td>77.58</td>
<td>75.28</td>
</tr>
</tbody>
</table>
Figure 2: Comparison of approximation ratios obtained by using Unconstrained MB-WISH (red) and single-bin (Ermon et al.’s method) (blue). A ratio closer to 1 is better.

...tion as a 16-ary string of length 30. Our results for the log-partition came out to be 156.00 with one hour time out for each call to the MAX-oracle. The benchmark for the log-partition function is provided to be 163.204.

The Object detection dataset comprised of 60 nodes each having a 11-ary value and by Unconstrained MB-WISH we found the log-partition function to be −38.9334. The CSP dataset is a Markov network with 30 node having a ternary value: we found the log partition function to be −39.9933. For these datasets there were no baselines available for comparison. The purpose of these experiments were to establish the scalability of MB-WISH.

6 CONCLUSION

Large scale counting problems (or discrete integrations of nonnegative weight functions) are often computationally intractable, but come up frequently in variety of inference tasks, most prominently as evaluations of partition functions. In this paper we extend a recent technique of hashing and optimization due to Ermon et al. for discrete integration over hypercube \( \{0, 1\}^n \) to that over hypergrids \( \{0, 1, \ldots, q - 1\}^n \). The trivial generalization results in an approximation factor that rapidly becomes worse as \( q \) increases. We remedy the situation by providing constant factor approximation algorithms for all \( q \).

The main drawback of this approach of discrete integration is the delegation of a hard combinatorial optimization to an oracle. In this line of work, an open problem is to come up with hash functions that maintain the essential properties (such as pairwise independence), but make the oracle optimization amenable. While in general this is not possible, for certain classes of weight functions this may be a plausible task and requires further exploration.

Acknowledgements: This research is supported in parts by NSF awards CCF 1642658, CCF 1618512, CCF 1909046 and CCF 1934846.

References


