Tight Variational Bounds via Random Projections and I-Projections

Lun-Kai Hsu
Stanford University
luffykai@stanford.edu

Tudor Achim
Stanford University
tachim@cs.stanford.edu

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Abstract

Information projections are the key building block of variational inference algorithms and are used to approximate a target probabilistic model by projecting it onto a family of tractable distributions. In general, there is no guarantee on the quality of the approximation obtained. To overcome this issue, we introduce a new class of random projections to reduce the dimensionality and hence the complexity of the original model. In the spirit of random projections, the projection preserves (with high probability) key properties of the target distribution. We show that information projections can be combined with random projections to obtain provable guarantees on the quality of the approximation obtained, regardless of the complexity of the original model. We demonstrate empirically that augmenting mean field with a random projection step dramatically improves partition function and marginal probability estimates, both on synthetic and real world data.

1 Introduction

Probabilistic inference is a core problem in machine learning, physics, and statistics [Koller and Friedman, 2009]. Probabilistic inference methods are needed for training, evaluating, and making predictions with probabilistic models [Murphy, 2012]. Developing scalable and accurate inference techniques is the key computational bottleneck towards deploying large-scale statistical models, but exact inference is known to be computationally intractable. The root cause is the curse of dimensionality – the number of possible scenarios to consider grows exponentially in the number of variables, and in continuous domains,

the volume grows exponentially in the number of dimensions [Bellman, 1961]. Approximate techniques are therefore almost always used in practice [Murphy, 2012].

Sampling-based techniques and variational approaches are the two main paradigms for approximate inference [Andrieu et al., 2003, Wainwright and Jordan, 2008]. Sampling-based approaches attempt to approximate intractable, high-dimensional distributions using a (small) collection of representative samples [Gogate and Dechter, 2011, Ermon et al., 2013a, Maddison et al., 2014]. Unfortunately, it is usually no easier to obtain such samples than it is to solve the original probabilistic inference problem [Jerrum and Sinclair, 1997]. Variational approaches, on the other hand, approximate an intractable distribution using a family of tractable ones. Finding the best approximation, also known as computing an *I-Projection* onto the family, is the key ingredient in all variational inference algorithms. In general, there is no guarantee on the quality of the approximation obtained [Globerson and Jaakkola, 2007, Ruozzi, 2013, Weller et al., 2014]. Intuitively, if the target model is too complex with respect to the family used, then the approximation will be poor.

To overcome this issue, we introduce a new class of *random* projections [Vadhan, 2011, Goldreich, 2011, Ermon et al., 2013b]. These projections take as input a probabilistic model and randomly perturb it, reducing its degrees of freedom. The projections can be computed efficiently and they reduce the effective dimensionality and complexity of the target model. Our key result is that the randomly projected model can then be approximated with I-projections onto simple families of distributions such as mean field with provable accuracy guarantees, regardless of the complexity of the original model. Crucially, in the spirit of random projections for dimensionality reduction, the random projections affect key properties such as the partition function in a highly predictable way. The I-projection of the projected model can therefore be used to accurately recover properties of the original model with high probability.

We demonstrate the effectiveness of our approach by using mean field augmented with random projections to estimate marginals and the log partition function on models of synthetic and real-world data, empirically showing large

improvements on both tasks.

2 Preliminaries

Let $p(x) = \frac{1}{Z} \prod_\alpha \psi_\alpha(\{x\}_\alpha)$ be a probability distribution over n binary variables $x \in \{0,1\}^n$ specified by an undirected graphical model[1], where α ranges over the factors in the model and $\{x\}_\alpha$ are the variables to which factor ψ_α is connected [Koller and Friedman, 2009]. We further assume that $p(x)$ is a member of an exponential family of distributions parameterized by $\theta \in \mathbb{R}^d$ and with sufficient statistics $\phi(x)$ [Wainwright and Jordan, 2008], i.e., $p(x) = \exp(\theta'\phi(x))/Z$. The constant $Z = \sum_{x \in \{0,1\}^n} \exp(\theta'\phi(x))$ is known as the partition function and ensures that the probability distribution is properly normalized. Computing the partition function is needed to evaluate likelihoods and to compare competing models of data. This computation is known to be intractable (#-P hard) in the worst-case. Intuitively, the issue is that the sum is defined over an exponentially large number of terms, and is therefore hard to evaluate unless there is special structure [Valiant, 1979, Koller and Friedman, 2009]. While several strategies for approximating the partition function are possible, we focus on variational approaches.

2.1 Variational Inference and I-projections

The key idea of variational inference is to approximate the intractable probability distribution $p(x)$ with one that is more tractable. The approach is to define a family $\mathcal{Q}$ of tractable distributions and then to find a distribution in this family that minimizes a notion of divergence from p. Typically, the Kullback-Leibler divergence $D_{KL}(q||p)$ is used, which is defined as follows

$$D_{KL}(q||p) = \sum_x q(x) \log \frac{q(x)}{p(x)} \tag{1}$$
$$= \sum_x q(x) \log q(x) - \theta \cdot \sum_x q(x)\phi(x) + \log Z$$

A distribution $q^* \in \mathcal{Q}$ that minimizes this divergence, $q^* = \arg\min_{q \in Q} D_{KL}(q||p)$, is called an information projection (*I-projection*) onto $\mathcal{Q}$. Intuitively, q^* is the "closest" distribution to p among all the distributions in $\mathcal{Q}$. Typically, one chooses $\mathcal{Q}$ to be a family of distributions for which inference is tractable so that (1) can be evaluated efficiently. The simplest choice, which removes all conditional dependencies, is to let $\mathcal{Q}$ be the set of fully factored probability distributions over $\mathcal{X}$, namely $\mathcal{Q}_{MF} = \{q(x)|q(x) = \prod_i q_i(x_i)\}$. This is known as the mean field approximation. Even when $\mathcal{Q}$ is tractable, computing an I-projection

---

[1] We restrict ourselves to binary variables for the ease of exposition. Our approach applies more generally to discrete graphical models.

is a non-convex optimization problem which can be difficult to solve.

Since the KL-divergence is non-negative, equation (1) shows that any distribution $q \in \mathcal{Q}$ provides a lower bound on the value of the partition function

$$\log Z \geq \max_{q \in \mathcal{Q}} \left\{ -\sum_x q(x) \log q(x) + \theta \cdot \sum_x q(x)\phi(x) \right\} \tag{2}$$

The distribution q^* that minimizes $D_{KL}(q||p)$ is also the distribution that provides the tightest lower bound on the partition function by maximizing the RHS of equation (2). The larger the set $\mathcal{Q}$ is, the better q^* can approximate p and the tighter the bound becomes. If $\mathcal{Q}$ is so large that $p \in \mathcal{Q}$, then $\min_{q \in Q} D_{KL}(q||p) = 0$, because when $q^* = p$, $D_{KL}(q^*||p) = 0$. In general, however, there is no guarantee on the tightness of bound (2) even if the optimization can be solved exactly.

2.2 Random Projections

We introduce a different class of *random* projections that we will use for probabilistic inference. Let $\mathcal{P}$ be the set of all probability distributions over $\{0,1\}^n$. We introduce a family of operators $\mathcal{R}_{A,b}^m : \mathcal{P} \to \mathcal{P}$, where $m \in [0,n]$, $A \in \{0,1\}^{m \times n}$, and $b \in \{0,1\}^m$. $R_{A,b}^m \in \mathcal{R}$ maps $p(x) = \frac{1}{Z} \exp(\theta'\phi(x))$ to a new probability distribution $R_{A,b}^m(p)$ whose support is restricted to $\{x : Ax = b \mod 2\}$ and whose probability mass function is proportional to p. Formally,

$$R_{A,b}^m(p)(x) = \frac{1}{Z(A,b)} \exp(\theta'\phi(x)) \tag{3}$$

with

$$Z(A,b) = \sum_{x|Ax=b \mod 2} \exp(\theta'\phi(x)) \tag{4}$$

These operators are clearly idempotent and can thus be interpreted as projections on $\mathcal{P}$.

When the parameters A, b are chosen randomly, the operator $R_{A,b}^m$ can be seen as a random projection. We consider random projections obtained by choosing $A \in \{0,1\}^{m \times n}$ and $b \in \{0,1\}^m$ independently and uniformly at random, i.e., choosing each entry by sampling an independent unbiased Bernoulli random variable. This can be shown to implement a strongly universal hash function [Vadhan, 2011, Goldreich, 2011]. Intuitively, the projection randomly subsamples the original space, selecting configurations $x \in \{0,1\}^n$ pairwise independently with probability 2^{-m}. It can be shown that

$$\mathbb{E}[Z(A,b)] = 2^{-m} Z \tag{5}$$

where the expectation is over the random choices of A, b, and that $Var[Z(A,b)] =$

$\frac{1}{2^m}\left(1 - \frac{1}{2^m}\right)\sum_x \exp(\theta'\phi(x))^2$ [Ermon et al., 2013b, 2014]. As we will formalize later, this random projection simplifies the model without losing too much information because it affects the partition function in a highly predictable way (controlling the expectation and the variance is sufficient to achieve high probability bounds).

To gain some intuition on the effect of the random projection, we can rewrite the linear system $Ax = b \bmod 2$ in reduced row-echelon form [Ermon et al., 2013b]. Assuming A is full-rank, we perform row reduction on A and b simultaneously to obtain the row reduced $C = [I_m|A']$ and b'. Here I_m is the $m \times m$ identity matrix, A' is the $m \times n - m$ sub-matrix of A that remains after the row reduction procedure, and b' is the result of performing the same operations on b as they are performed on A. Thus the system $Ax = b$ is equivalent to $Cx = b'$. For notational simplicity, however, we continue to use b instead of b'. We can equivalently rewrite the constraints $Ax = b \bmod 2$ as the following set of constraints

$$x_1 = \bigoplus_{i=m+1}^{n} c_{1i}x_i \oplus b_1, \cdots, x_m = \bigoplus_{i=m+1}^{n} c_{mi}x_i \oplus b_m$$

where $\oplus$ denotes the exclusive-or (XOR) operator. Thus, the random projection reduces the degrees of freedom of the model by m, as the first m variables are completely determined by the last $n - m$. For later development it will also be convenient to rewrite these linear equations modulo 2 as polynomial equations by changing variables from $\{0, 1\}$ to $\{-1, 1\}$:

$$(1 - 2x_k) = \prod_{i=m+1}^{n} (1 - 2C_{ki}x_i)(1 - 2b_k) \qquad (6)$$

for $k = 1, \cdots, m$.

3 Combining Random Projections with I-Projections

Given an intractable target distribution p and a candidate set of tractable distributions $\mathcal{Q}$, there are two main issues with variational approximation techniques: (i) p can be far from the approximating family $\mathcal{Q}$ in the sense that even the optimal $q^* = \arg\min_{q \in \mathcal{Q}} D_{KL}(q\|p)$ can have a large divergence $D_{KL}(q^*\|p)$ and therefore yield a poor lower bound in Eq. (2), and (ii) the variational problem in Eq. (2) is non-convex and thus difficult to solve exactly in high dimensions. Our key idea is to address (i) by using the random projections introduced in the previous section to "simplify" p, producing a projection $R^m_{A,b}(p)$ that provably is closer to $\mathcal{Q}$. Crucially, because of the statistical properties of the random projection used, variational inferences on the randomly projected model $R^m_{A,b}(p)$ reveal useful information about the original distribution p. Randomization plays a key role in our approach, boosting the power of variational inference. A pictorial representation is given in Figure 1.

A similar approach combining variational inference with random projections has been recently introduced by Zhu and Ermon [2015]. They analyze the first and second moments of (4), and provide a family of probabilistic bounds on the partition function based on ideas from variational inference. These bounds, however, are not guaranteed to be tight, i.e., they suffer from the same limitations of traditional variational bounds. Further, these bounds hold only in expectation, and the variance can be too high to guarantee concentration around the mean using a small number of samples. In contrast, we introduce a novel analysis based on information geometry (KL-divergences, see Figure 1) which provides new estimators with provably tight bounds. These bounds are valid with high probability and can be computed with a small number of samples. Further, our information geometric approach leads to new algorithmic ideas to take advantage of the reduced dimensionality space introduced by the random projection.

3.1 Provably Tight Variational Bounds on the Partition Function

Let $\mathcal{D} = \{\delta x_0 | x_0 \in X\}$ denote the set of degenerate probability distributions over $\{0, 1\}^n$, i.e. probability distributions that place all the probability mass on a single configuration. There are 2^n such probability distributions and the entropy of each is zero. Given any probability distribution p, its projection on $\mathcal{D}$, i.e., $\arg\min_{q \in \mathcal{D}} D_{KL}(q\|p)$, is given by a distribution that places all the probability on $\arg\max_{x \in \mathcal{X}} \log p(x)$. Thus computing the I-projection on $\mathcal{D}$ is equivalent to solving a Most Probable Explanation query which is NP-hard in the worst-case [Koller and Friedman, 2009].

Let $\mathcal{Q}$ be a family of probability distributions that contains $\mathcal{D}$. Our key result is that we can get *provably tight bounds* on the partition function Z by taking an I-projection onto $\mathcal{Q}$ *after* a suitable random projection. The proof relies on Theorem 2, which states that p can be approximated using a small number of configurations obtained by randomly projecting p and solving a Most Probable Explanation query Ermon et al. [2013c]. Interpreted variationally, p can be approximated using a small number of degenerate distributions that can be computed by taking random projections followed by I-projections onto $\mathcal{D}$. Observing that the family of degenerate distributions $\mathcal{D}$ is contained in $\mathcal{Q}$, we can show that I-projecting onto $\mathcal{Q}$ after suitable random projections yields a provably accurate approximation of the target distribution p.

Theorem 1. *Let $A^{i,t} \in \{0,1\}^{i \times n} \overset{iid}{\sim}$ Bernoulli($\frac{1}{2}$) and $b^{i,t} \in \{0,1\}^i \overset{iid}{\sim}$ Bernoulli($\frac{1}{2}$) for $i \in [0, n]$ and $t \in [1, T]$. Let $\mathcal{Q}$ be a family of distributions such that $\mathcal{D} \subseteq \mathcal{Q}$. Let*
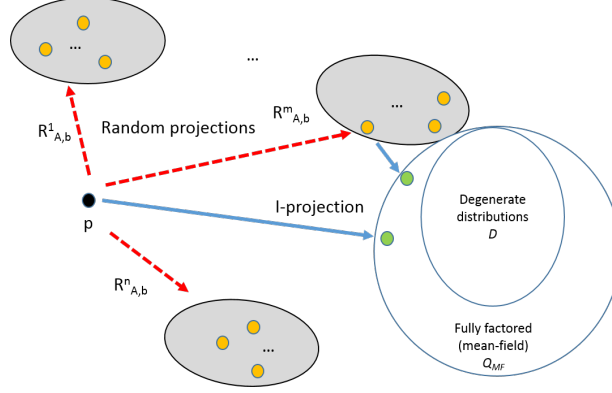
Figure 1: Pictorial representation of the approach. p is the target intractable distribution. Each gray oval labeled $R_{A,b}^m$ represents the family of distributions induced by projecting p onto m parity constraints. The small circles inside the ovals represent the projections for particular choices of parity constraint matrices A, b; these circles are themselves probability distributions. The I-projection from p to Q_{MF} (bottom solid arrow) can yield inaccurate estimates of the partition function of p. Theorem 1 states that there exists an m such that by projecting p onto $R_{A,b}^m$ (dashed red lines) we obtain a distribution that can be well-approximated by some distribution in $\mathcal{D}$ (and therefore in $\mathcal{Q}_{MF}$ as well). Since the random projection alters the partition function of p in a predictable way, we can combine these two projections to yield a provably tighter approximation of the original distribution.

$$\gamma^{i,t} = \exp\left(\max_{q \in \mathcal{Q}} \ \theta \cdot \sum_{x : A^{i,t}x = b^{i,t}} q(x)\phi(x) - \sum_{x : A^{i,t}x = b^{i,t}} q(x)\log q(x)\right) \quad (7)$$

be the optimal solutions for the projected variational inference problems. Then for all $i \in [0, n]$ and for all $T \geq 1$ the rescaled variational solution is a lower bound for Z in expectation

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\gamma^{i,t}2^i\right] = \mathbb{E}[\gamma^{i,t}2^i] \leq Z$$

There also exists an m such that for any $\Delta > 0$ and positive constant $\alpha \leq 0.0042$, if $T \geq \frac{1}{\alpha}\left(\log(n/\Delta)\right)$ then with probability at least $(1 - 2\Delta)$

$$\frac{1}{T}\sum_{t=1}^{T}\gamma^{m,t}2^m \geq \frac{Z}{64(n+1)} \quad (8)$$

$$4Z \geq Median\left(\gamma^{m,1}, \cdots, \gamma^{m,T}\right)2^m \geq \frac{Z}{32(n+1)} \quad (9)$$

Proof. We present a sketch of the proof; see the Appendix for the formal derivation. For the first part of the theorem, notice that $\gamma^{i,t}$ is the standard variational lower bound on $Z(A^{i,t}, b^{i,t})$, the partition function of the randomly projected model. The inequality $Z(A^{i,t}, b^{i,t}) \geq \gamma^{i,t}$ holds for any realization of the random matrices $A^{i,t}, b^{i,t}$; therefore,

if we consider the expected value and use (5) we obtain the desired result $\mathbb{E}[\gamma^{i,t}2^i] \leq Z$.

For the second part, the upper bound $4Z \geq$ Median $\left(\gamma^{m,1}, \cdots, \gamma^{m,T}\right)2^m$ is an immediate consequence of Markov's inequality, given that $\mathbb{E}[\gamma^{i,t}2^i] \leq Z$. For the lower bound, notice that by assumption the conditions of Theorem 2 are satisfied. Therefore, we know that equation (10) holds with probability at least $1 - \Delta$. Since the terms are all nonnegative the maximum element is at least $1/(n+1)$ times the sum, so we apply (10) to get

$$\max_i \exp\left(\text{Median} - \min_{q \in \mathcal{D}} D_{KL}(q||R_{A^{i,t},b^{i,t}}^i(p))\right.$$
$$\left. + \log Z(A^{i,t}, b^{i,t})\right)2^{i-1} \geq \frac{1}{32}Z\frac{1}{n+1}$$

Equality holds if all terms in the sum are equal. Therefore there exists m such that

$$\text{Median} - D_{KL}(q^*||R_{A^{m,t},b^{m,t}}^m(p)) + \log Z(A^{m,t}, b^{m,t})$$
$$+ (m-1)\log 2 \geq \log Z - \log 32(n+1)$$

where q^* is the I-projection of $R_{A^{m,t},b^{m,t}}^m(p)$ onto $\mathcal{D}$. This is the key result: one can always find a small number of T degenerate distributions such that with proper rescaling they account for at least a $32(n+1)$ fraction of the value of Z with high probability.

We rely on the fact that $\mathcal{Q}$ contains the degenerate family $\mathcal{D}$ which implies that the lower bound can only improve if we optimize over the larger family:

$$\min_{q \in \mathcal{Q}} D_{KL}(q||R_{A,b}(p)) \leq \min_{q \in \mathcal{D}} D_{KL}(q||R_{A,b}(p))$$

The final high probability bound then follows by Chernoff's inequality. $\qquad\square$

This proves that appropriately rescaled variational lower bounds obtained on the randomly projected models (aggregated through median as in Equation (9)) are within a factor n of the true value Z, where n is the number of variables in the model. This is an improvement on prior variational approximations which can either be unboundedly suboptimal or provide guarantees that hold only in expectation [Zhu and Ermon, 2015]; in contrast, our bounds are tight and require a relatively small number of samples proportional to $\log n/\Delta$. The proof, reported in the appendix for space reasons, relies on the following technical result which can be seen as a variational interpretation of Theorem 1 from [Ermon et al., 2013c] and is of independent interest:

Theorem 2. *For any* $\Delta > 0$*, and positive constant* $\alpha \leq 0.0042$*, let* $T \geq \frac{1}{\alpha}\left(\log(n/\Delta)\right)$*. Let* $A^{i,t} \in \{0,1\}^{i \times n} \overset{iid}{\sim}$ *Bernoulli*$(\frac{1}{2})$ *and* $b^{i,t} \in \{0,1\}^i \overset{iid}{\sim}$ *Bernoulli*$(\frac{1}{2})$ *for* $i \in [0,n]$ *and* $t \in [1,T]$*. Let*

$$\delta^{i,t} = \min_{q \in \mathcal{D}} D_{KL}(q||R^i_{A^{i,t},b^{i,t}}(p))$$

Then with probability at least $(1-\Delta)$

$$\sum_{i=0}^n \exp\left(\underset{t\in[T]}{Median}\left(-\delta^{i,t} + \log Z(A^{i,t},b^{i,t})\right)\right)2^{i-1} \quad (10)$$

is a 32-approximation to Z *(that is, at least* $Z/32$ *and at most* $32Z$*).*

Intuitively, Theorem 2 states that one can always find a small number of degenerate distributions (which can be equivalently thought of as special states or samples that can be discovered through random projections and KL-divergence minimization) that are with high probability representative of the original probabilistic model, regardless of how complex the model is. Theorem 1 extends this idea to more general families of distributions such as mean field.

Proof of Theorem 2. For ease of notation we define $X_{A,b} = \{x \in \{0,1\}^n : Ax = b \mod 2\}$. Thus

$$\min_{q \in \mathcal{D}} D_{KL}(q||R^i_{A,b}(p))$$

$$= \min_{\delta x_0 \in \mathcal{D}} H(\delta x_0) - \sum_{x \in X_{A,b}} \delta x_0(x)\theta'\phi(x) + \log Z(A,b)$$

$$= \min_{x \in X_{A,b}} -\theta'\phi(x_0) + \log Z(A,b)$$

since $H(\delta x_0) = 0$ for any x_0. Therefore

$$-\min_{q \in \mathcal{D}} D_{KL}(q||R^i_{A^{i,t},b^{i,t}}(p)) + \log Z(A^{i,t},b^{i,t})$$

$$= \max_{x \in X_{A^{i,t},b^{i,t}}} \theta'\phi(x)$$

We can substitute this into Eq. (10) to rewrite it as

$$\sum_{i=0}^n \exp\left(\underset{t\in[T]}{Median}\left(\max_{x \in X_{A^{i,t},b^{i,t}}} \theta'\phi(x)\right)\right)2^{i-1}$$

$$= \sum_{i=0}^n \left(\underset{t\in[T]}{Median}\exp\left(\max_{x \in X_{A^{i,t},b^{i,t}}} \theta'\phi(x)\right)\right)2^{i-1}$$

$$= \sum_{i=0}^n \left(\underset{t\in[T]}{Median}\left(\max_{x \in X_{A^{i,t},b^{i,t}}} \exp(\theta'\phi(x))\right)\right)2^{i-1}$$

The result then follows directly from Theorem 1 from [Ermon et al., 2013c]. $\qquad\square$

3.2 Solving Randomly Projected Variational Inference Problems

To apply the results from Theorem 1 we must choose a tractable approximating family $\mathcal{D} \subseteq \mathcal{Q}$ for the I-projection part and incorporate our random projections into the optimization procedure. We focus on mean field ($\mathcal{Q} = \mathcal{Q}_{MF}$) as our approximating family, but the results can be easily extended to structured mean field [Bouchard-Côté and Jordan, 2009]. For simplicity of exposition we consider only probabilistic models p with unary and binary factors (e.g. Ising models, restricted Boltzmann machines). That is, $p(x) = \exp(\theta \cdot \phi(x))/Z$, where $\phi(x)$ are single node and pairwise edge indicator variables.

Recall that our projection $R^m_{A,b}(p)$ constrains the distribution p to $\{x|Ax = b \mod 2\}$. The projected variational optimization problem (7) is therefore

$$\log Z(A,b) \geq \max_q \theta \cdot \sum_{x|Ax=b \mod 2} q(x)\phi(x)$$

$$- \sum_{x|Ax=b \mod 2} q(x)\log q(x)$$

Or, equivalently,

$$\log Z(A,b) \geq \max_\mu \ \theta \cdot \mu + \sum_{i=m+1}^n H(\mu_i) \quad (11)$$

where μ is the vector of singleton and pairwise marginals of $q(x)$ and $H(\mu_i)$ is the entropy of a Bernoulli random variable with parameter μ_i. To solve this optimization problem efficiently we need a clever way to take into account the parity constraints, for running traditional mean field with message passing as in [Zhu and Ermon, 2015] would fail in the normalization step because of the presence of hard parity constraints. The key idea is to consider the equivalent row-reduced representation of the constraints from (6) and

define

$$q(x_1, \cdots, x_n) = \prod_{i=m+1}^{n} q_i(x_i) \cdot$$

$$\prod_{k=1}^{m} \mathbb{I}\left\{(1 - 2x_k) = \prod_{i=m+1}^{n} (1 - 2C_{ki}x_i)(1 - 2b_k)\right\}$$

where we have a set of independent "free variables" (wlog., the last $n - m$) and a set of "constrained variables" (the first m) that are always set so as to satisfy the parity constraints. Since the variables $x_1, \ldots, x_m$ are fully determined by $x_{m+1}, \ldots, x_n$, we see that the marginals $\mu_1, \ldots, \mu_m$ are also determined by $\mu_{m+1}, \ldots, \mu_n$, as shown by the following proposition:

Proposition 1. *The singleton and pairwise marginals in (11) can be computed as follows:*

Singleton marginals: for $k \in [m + 1, n]$, $\mu_k = E_q[x_k] = q_k(1)$. For $k \in [1, m]$,

$$\mu_k = \left(1 - (1 - 2b_k) \prod_{i=m+1}^{n} (1 - 2C_{ki}\mu_i)\right)/2$$

Pairwise marginals: for $k, \ell \in [m + 1, n]$, $\mu_{kl} = E_q[x_k x_\ell] = \mu_k \mu_\ell$. For $k \in [m + 1, n]$, $\ell \in [1, m]$,

$$\mu_{kl} = \begin{cases} \mu_k \frac{1}{2}(1 + (1 - 2b_l) \prod_{i \neq k, i=m+1}^{n} (1 - 2C_{li}\mu_i)) \\ \qquad \text{if } C_{lk} = 1 \\ \mu_k \mu_l \qquad \text{otherwise} \end{cases}$$

For $k \in [1, m]$, $\ell \in [1, m]$,

$$\mu_{kl} = \frac{1}{4}\left(1 - (1 - 2b_k) \prod_{i=m+1}^{n} (1 - 2C_{ki}\mu_i) + \right.$$

$$(1 - 2b_k)(1 - 2b_l) \prod_{i=m+1}^{n} (1 - \mu_i(2C_{ki} + 2C_{li} - 4C_{ki}C_{li}))$$

$$\left. -(1 - 2b_l) \prod_{i=m+1}^{n} (1 - 2C_{li}\mu_i)\right)$$

The derivation is found in the appendix. We can therefore maximize the lower bound in (11) by optimizing only over the "free marginals" $\mu_{m+1}, \ldots, \mu_n$, as the remaining one are completely determined per Proposition 1. Compared to a traditional mean field variational approximation, we have a problem with a smaller number of variables, but with additional non-convex constraints.

4 Algorithm: Mean Field with Random Projections

Theorem 1 guarantees that the approximation to Z has a tight lower bound only if we are able to find globally optimal solutions for (11). However, nontrivial variational inference problems (2) are non-convex in general even without any random projections and even when $\mathcal{Q}$ is simple,

e.g., $\mathcal{Q} = \mathcal{Q}_{MF}$. We do not explicitly handle this nonconvexity, but nevertheless we show empirically that we can vastly improve on mean field lower bounds. The key insight for our optimization procedure is that the objective function is still coordinate-wise concave, like in a traditional mean-field approximation:

Proposition 2. *The objective function $\theta \cdot \mu + \sum_{i=m+1}^{n} H(\mu_i)$ in (11) is concave with respect to any particular free marginal $\mu_{m+1}, \ldots, \mu_n$.*

Proof. By inspection, all the marginals in Proposition 1 are linear with respect to any specific free marginal $\mu_{m+1}, \ldots, \mu_n$. Since the entropy term is concave, the RHS in (11) is concave in each free marginal $\mu_{m+1}, \ldots, \mu_n$. $\square$

Since (11) is concave in each variable we devise a coordinate-wise ascent algorithm, called Mean Field with Random Projections (MFRP), for maximizing the lower bound in (11) over the free marginals defined in Proposition 1. The pseudocode is reported as Algorithm 1. It takes as input an (intractable) discrete probability distribution $p(x)$ and a number of parity constraints m. It returns a probabilistic lower bound for the partition function of $p(x)$.

The algorithm proceeds as follows. Starting from a random initialization, we iterate over each free marginal μ_k and maximize (11) with the rest of the free marginals fixed by setting the gradient with respect to μ_k equal to 0 and solving for μ_k. Because the overall optimization problem is not concave the algorithm may converge at a local maximum; therefore, we use J random initializations and use the best lower bound found across the J runs of the ascent algorithm. For a given m, we repeat this procedure T times and return the median across the runs. Each coordinate ascent step for free marginal μ_i takes $O(m + n + |E_{cc}|(n - m))$ steps in expectation where E_{cc} is the number of variables co-occurring in a parity constraint. Recomputing the constrained marginals takes $O(m(n - m))$ steps.

The algorithm returns the maximum of $\text{MFRP}(p(x), m)$ over $m \in [0, n]$. If MFRP finds a global optimum, then Theorem 1 guarantees it is a tight lower bound for $\log Z$ with high probability. Since MFRP uses coordinate-wise ascent we cannot certify global optimality; however, our experiments show large improvements in the lower bound when compared to existing variational methods.

5 Experiments

We investigate MFRP's empirical performance on Ising models and on Restricted Boltzmann Machines. In particular, we are interested in the log partition function estimates and in the quality of the marginal estimates. Although our theoretical results only apply to the partition function, it is believed that better partition function estimates lead to better marginal estimates. This relationship is however gener-

---

Algorithm 1 $\text{MFRP}(p(x) \propto \exp(\theta'\phi(x)), m)$

---

for $t = 1, \cdots, T$ **do**   ▷ Do T random projections

     Generate parity bits $b^{(t)} \overset{iid}{\sim} \text{Bernoulli}(\frac{1}{2})^m$ ▷ Generate random projection $R^m_{A^{(t)}, b^{(t)}}$

     Generate matrix $A^{(t)} \overset{iid}{\sim} \text{Bernoulli}(\frac{1}{2})^{m \times n}$

     Row reduce $A^{(t)}, b^{(t)}$ to yield $C = [I|A^{(t)'}]$ and b ▷ Compute constraints

     $\widetilde{Z}^{(t)} \leftarrow 0$

     for $j = 1, \cdots, J$ **do**   ▷ Try different initializations

         Initialize $\mu^{(j,t)} \overset{iid}{\sim} \text{Unif}(0,1)^n$

         for $l = 1, \cdots, m$ **do**   ▷ Compute constrained marginals

            $\mu^{(j,t)} \leftarrow \left(1 - \prod_{i=m+1}^{n}(1 - 2C_{li}\mu_i^{(j,t)})(1 - 2b_l) \right)/2$

         end for

         while not converged **do**   ▷ Stop when the increment is small or timeout

            for $k = m+1, \cdots, n$ **do**   ▷ Coordinate ascent over free marginals

                $\mu_k^{(j,t)} \leftarrow \arg\max_{\mu_k} \theta \cdot \mu^{(j,t)} + \sum_{i=m+1}^{n} H(\mu_i^{(j,t)})$

                for $l = 1, \cdots, m$ **do**

                   $\mu_k^{(j,t)} \leftarrow \left(1 - \prod_{i=m+1}^{n}(1 - 2C_{li}\mu_i^{(j,t)})(1 - 2b_l) \right)/2$ ▷ Update constrained marginals

                end for

            end for

         end while

     end for

     $\widetilde{Z}^{(t)} \leftarrow \max_j \exp(\theta \cdot \mu^{(j,t)} + \sum_{i=m+1}^{n} H(\mu_i^{(j,t)}))$ ▷ Pick best over initializations

end for

Return $2^m Median\left(\widetilde{Z}^{(1)}, \ldots, \widetilde{Z}^{(T)} \right)$ ▷ Aggregate across projections

---
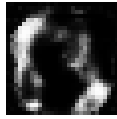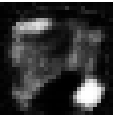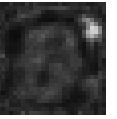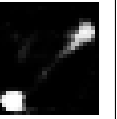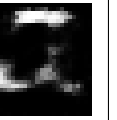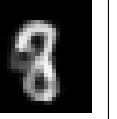
No. Hidden Nodes	100	100	100	200	200	200
CD-k	1	5	15	5	15	25
MF $\log Z$	501	348	297	203	293	279
MFRP $\log Z$	501	**433**	**342**	**380**	**313**	**295**
MF μ						
MFRP μ						

Table 1: Log partition function ($\log Z$) and marginals for the visible units (μ) estimates across two RBMs trained with contrastive divergence (CD) using different sampling parameters. MFRP provides improved lower bounds on the partition function and qualitatively better marginal estimates.

ally poorly understood, even for standard variational methods. Where applicable, exact ground truth estimates are obtained with the libDAI implementation of Junction Tree [Mooij, 2010]. Upper bounds are calculated with Tree-Reweighted Belief Propagation (TRW-BP) [Wainwright, 2003], also implemented in libDAI. All methods are compared to mean field (MF) optimized with coordinate-wise ascent and random restarts.

5.1 Ising Models

We consider $n \times n$ binary grid Ising models with variables $x_i \in \{-1, 1\}$ and potentials $\psi_{ij}(x_i, x_j) = \exp(w_{ij}x_ix_j + f_ix_i + f_jx_j)$. In particular, we look at mixed models where the w_{ij}'s are drawn uniformly from $[-10, 10]$ and the f_i's uniformly from $[-1, 1]$.

Figure 2 compares the log partition function estimates from mean field, junction tree, MFRP, and TRW-BP. For each grid size, we generated five different grids and computed
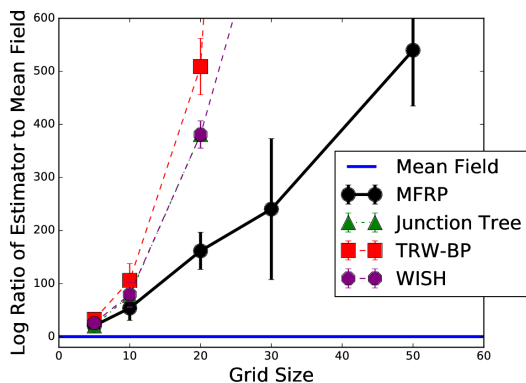
Figure 2: Ising grids: for each size, we plot the ratios of the estimates from each method to the mean field estimate, with standard error bars based on 5 runs. The mean field line is 0 and flat because results are reported as a log ratio over the mean field estimate.
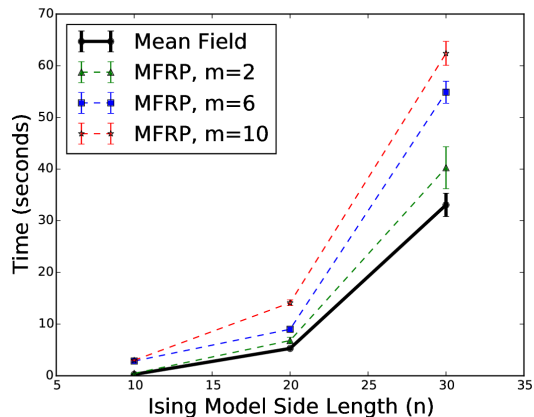

Figure 3: Runtimes for MFRP across different grid sizes and constraints. Empirically, both methods scale linearly in the number of variables in the model (n^2 in the Ising model case).

the mean field estimate for each as a baseline lower bound. For each of the five grids we also computed the best MFRP lower bound over $m \in [0, 20]$ with $T = 5$ trials each. For comparison we include the exact log partition calculation from Junction Tree up to $n = 20$ and the TRW-BP upper bounds for all n. We plot the mean and standard error bars of the log ratio of each estimate over mean field for each method over the five grids. All results are reported as log ratios with respect to the MF estimate, thus the MF line is zero and constant. Note that for large grid sizes, the lower bound provided by MFRP is hundreds of orders of magnitude better than than those found by mean field. We also report results from solving the discrete problem *to optimality* using the optimization package CPLEX as the WISH line in Figure 2, following [Ermon et al., 2013c]. The accuracy of this method suggests that the approximation gap of our algorithm is due to the difficulty of optimizing over the highly nonconvex energy landscape with randomly-restarted coordinate ascent.

Finally, we consider the empirical runtime of the method for varying grid sizes n and number of constraints m in Figure 3. As expected, the runtime for mean field grows linearly in the number of variables in the graph (quadratically with the side length n) and there is a linear slowdown as constraints are added to the optimization problem solved by MFRP. These experiments show that MFRP scales well (with a small overhead with respect to mean field) and can potentially be applied in a wide range of application domains where traditional mean field is used.

5.2 Restricted Boltzmann Machines

We train Restricted Boltzmann Machines (RBMs) [Hinton et al., 2006] using Contrastive Divergence (CD) [Welling

and Hinton, 2002, Carreira-Perpinan and Hinton, 2005] on the MNIST hand-written digits dataset. In an RBM there is a layer of n_h hidden binary variables $h = h_1, \cdots, h_{n_h}$ and a layer of n_v binary visible units $v = v_1, \cdots, v_{n_v}$. The joint probability distribution is given by $P(h, v) = \frac{1}{Z} \exp(b'v + c'h + h'Wv)$. We use $n_h \in \{100, 200\}$ hidden units and $n_v = 28 \times 28 = 784$ visible units. We learn the parameters b, c, W using CD-k for $k \in \{1, 5, 15\}$, where k denotes the number of Gibbs sampling steps used in the inference phase, with 15 training epochs and minibatches of size 20.

We then use MF and MFRP to estimate the log partition function and also consider the aggregate marginals of the visible units. Results are reported in Table 1. For most of the cases we see a clear improvement in both the log partition lower bounds and in the marginals, with the marginal for $h = 100$, CD-$k = 15$ similar visually to an average over all digits in the dataset.

6 Conclusions

We introduced a new, general approach to variational inference that combines random projections with I-projections to obtain provably tight lower bounds for the log partition function. Our approach is the first to leverage universal hash functions and their properties in a variational sense. We demonstrated the effectiveness of this idea by extending mean field with random projections and empirically showed a large improvement in the partition function lower bounds and marginals obtained on both synthetic and real world data. Natural extensions to the approach include applications to other variational methods, like the Bethe approximation, and the use of better optimization techniques.

References

C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50 (1-2):5–43, 2003.

R. Bellman. *Adaptive control processes: A guided tour*. Princeton University Press, Princeton, NJ, 1961.

A. Bouchard-Côté and M. I. Jordan. Optimization of structured mean field objectives. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 67–74. AUAI Press, 2009.

M. Carreira-Perpinan and G. Hinton. On contrastive divergence learning. In *Proc. of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, page 17, 2005.

S. Ermon, C. P. Gomes, A. Sabharwal, and B. Selman. Embed and project: Discrete sampling with universal hashing. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2085–2093, 2013a.

S. Ermon, C. P. Gomes, A. Sabharwal, and B. Selman. Optimization with parity constraints: From binary codes to discrete integration. In *Proc. of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013b.

S. Ermon, C. P. Gomes, A. Sabharwal, and B. Selman. Taming the curse of dimensionality: Discrete integration by hashing and optimization. In *Proc. of the 30th International Conference on Machine Learning (ICML)*, 2013c.

S. Ermon, C. P. Gomes, A. Sabharwal, and B. Selman. Low-density parity constraints for hashing-based discrete integration. In *Proc. of the 31st International Conference on Machine Learning (ICML)*, pages 271–279, 2014.

A. Globerson and T. Jaakkola. Approximate inference using conditional entropy decompositions. In *International Conference on Artificial Intelligence and Statistics*, pages 130–138, 2007.

V. Gogate and R. Dechter. SampleSearch: Importance sampling in presence of determinism. *Artificial Intelligence*, 175(2):694–729, 2011.

O. Goldreich. Randomized methods in computation. *Lecture Notes*, 2011.

G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

M. Jerrum and A. Sinclair. The markov chain monte carlo method: An approach to approximate counting and integration. In *Approximation Algorithms for NP-hard Problems*, pages 482–520. PWS Publishing, Boston, MA, 1997.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

C. J. Maddison, D. Tarlow, and T. Minka. A* sampling. In *Advances in Neural Information Processing Systems*, pages 3086–3094, 2014.

J. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, 2010.

K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

N. Ruozzi. Beyond log-supermodularity: Lower bounds and the bethe partition function. In *Uncertainty in Artificial Intelligence*, volume 4, page 546. Citeseer, 2013.

S. Vadhan. Pseudorandomness. *Foundations and Trends in Theoretical Computer Science*, 2011.

L. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

M. J. Wainwright. Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. In *Proc. of the 8th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2003.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

A. Weller, K. Tang, D. Sontag, and T. Jebara. Understanding the bethe approximation: When and how can it go wrong. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

M. Welling and G. Hinton. A new learning algorithm for mean field Boltzmann machines. In *Proc. of the 12th International Conference on Artificial Neural Networks (ICANN)*, 2002.

M. Zhu and S. Ermon. A hybrid approach for probabilistic inference using random projections. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2039–2047, 2015.