Lost Relatives of the Gumbel Trick

Matej Balog [1 2] **Nilesh Tripuraneni** [3] **Zoubin Ghahramani** [1 4] **Adrian Weller** [1 5]

Abstract

The Gumbel trick is a method to sample from a discrete probability distribution, or to estimate its normalizing partition function. The method relies on repeatedly applying a random perturbation to the distribution in a particular way, each time solving for the most likely configuration. We derive an entire family of related methods, of which the Gumbel trick is one member, and show that the new methods have superior properties in several settings with minimal additional computational cost. In particular, for the Gumbel trick to yield computational benefits for discrete graphical models, Gumbel perturbations on all configurations are typically replaced with so-called low-rank perturbations. We show how a subfamily of our new methods adapts to this setting, proving new upper and lower bounds on the log partition function and deriving a family of sequential samplers for the Gibbs distribution. Finally, we balance the discussion by showing how the simpler analytical form of the Gumbel trick enables additional theoretical results.

1. Introduction

In this work we are concerned with the fundamental problem of sampling from a discrete probability distribution and evaluating its normalizing constant. A probability distribution p on a discrete sample space $\mathcal{X}$ is provided in terms of its potential function $\phi : \mathcal{X} \to [-\infty, \infty)$, corresponding to log-unnormalized probabilities via $p(\mathbf{x}) = e^{\phi(\mathbf{x})}/Z$, where the normalizing constant Z is the *partition function*. In this context, p is the *Gibbs distribution* on $\mathcal{X}$ associated with the potential function ϕ. The challenges of sampling from such a discrete probability distribution and estimating the partition function are fundamental problems with ubiquitous applications in machine learning, classical statistics and statistical physics (see, e.g., Lauritzen, 1996).

Perturb-and-MAP methods (Papandreou & Yuille, 2010) constitute a class of randomized algorithms for estimating partition functions and sampling from Gibbs distributions, which operate by randomly perturbing the corresponding potential functions and employing maximum a posteriori (MAP) solvers on the perturbed models to find a maximum probability configuration. This MAP problem is NP-hard in general; however, substantial research effort has led to the development of solvers which can efficiently compute or estimate the MAP solution on many problems that occur in practice (e.g., Boykov et al., 2001; Kolmogorov, 2006; Darbon, 2009). Evaluating the partition function is a harder problem, containing for instance #P-hard counting problems. The general aim of perturb-and-MAP methods is to reduce the problem of partition function evaluation, or the problem of sampling from the Gibbs distribution, to repeated instances of the MAP problem (where each instance is on a different random perturbation of the original model).

The Gumbel trick (Papandreou & Yuille, 2011) relies on adding Gumbel-distributed noise to each configuration's potential $\phi(\mathbf{x})$. We derive a wider family of perturb-and-MAP methods that can be seen as perturbing the model in different ways – in particular using the Weibull and Fréchet distributions alongside the Gumbel. We show that the new methods can be implemented with essentially no additional computational cost by simply averaging existing Gumbel MAP perturbations in different spaces, and that they can lead to more accurate estimators of the partition function.

Evaluating or perturbing each configuration's potential with i.i.d. Gumbel noise can be computationally expensive. One way to mitigate this is to cleverly prune computation in regions where the maximum perturbed potential is unlikely to be found (Maddison et al., 2014; Chen & Ghahramani, 2016). Another approach exploits the product structure of the sample space in discrete graphical models, replacing i.i.d. Gumbel noise with a "low-rank" approximation. Hazan & Jaakkola (2012); Hazan et al. (2013) showed that from such an approximation, upper and lower bounds on the partition function and a sequential sampler for the Gibbs distribution can still be recovered. We show that a subfamily of our new methods, consisting of Fréchet, Exponential and Weibull tricks, can also be used with low-

[1]University of Cambridge, UK [2]MPI-IS, Tübingen, Germany [3]UC Berkeley, USA [4]Uber AI Labs, USA [5]Alan Turing Institute, UK. Correspondence to: Matej Balog <first.last@gmail.com>. Code: https://github.com/matejbalog/gumbel-relatives.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

rank perturbations, and use these tricks to derive new upper and lower bounds on the partition function, and to construct new sequential samplers for the Gibbs distribution.

Our main contributions are as follows:

1. A family of tricks that can be implemented by simply averaging Gumbel perturbations in different spaces, and which can lead to more accurate or more sample efficient estimators of Z (Section 2).
2. New upper and lower bounds on the partition function of a discrete graphical model computable using low-rank perturbations, and a corresponding family of sequential samplers for the Gibbs distribution (Section 3).
3. Discussion of advantages of the simpler analytical form of the Gumbel trick including new links between the errors of estimating Z, sampling, and entropy estimation using low-rank Gumbel perturbations (Section 4).

Background and Related work The idea of perturbing the potential function of a discrete graphical model in order to sample from its associated Gibbs distribution was introduced by Papandreou & Yuille (2011), inspired by their previous work on reducing the sampling problem for Gaussian Markov random fields to the problem of finding the mean, using independent local perturbations of each Gaussian factor (Papandreou & Yuille, 2010). Tarlow et al. (2012) extended this perturb-and-MAP approach to sampling, in particular by considering more general structured prediction problems. Hazan & Jaakkola (2012) pointed out that MAP perturbations are useful not only for sampling the Gibbs distribution (considering the argmax of the perturbed model), but also for bounding and approximating the partition function (by considering the value of the max).

Afterwards, Hazan et al. (2013) derived new lower bounds on the partition function and proposed a new sampler for the Gibbs distribution that samples variables of a discrete graphical model sequentially, using expected values of low-rank MAP perturbations to construct the conditional probabilities. Due to the low-rank approximation, this algorithm has the option to reject a sample. Orabona et al. (2014) and Hazan et al. (2016) subsequently derived measure concentration results for the Gumbel distribution that can be used to control the rejection probability. Maji et al. (2014) derived an uncertainty measure from random MAP perturbations, using it within a Bayesian active learning framework for interactive image boundary annotation.

Perturb-and-MAP was famously generalized to continuous spaces by Maddison et al. (2014), replacing the Gumbel distribution with a Gumbel process and calling the resulting algorithm *A* sampling*. Maddison (2016) cast this work into a unified framework together with adaptive rejection sampling techniques, based on the notion of exponential races. This recent view generally brings together perturb-and-MAP and accept-reject samplers, exploiting the connection between the Gumbel distribution and competing exponential clocks that we also discuss in Section 2.1.

Inspired by A* sampling, Kim et al. (2016) proposed an exact sampler for discrete graphical models based on lazily-instantiated random perturbations, which uses linear programming relaxations to prune the optimization space. Further recent applications of perturb-and-MAP include structured prediction in computer vision (Bertasius et al., 2017) and turning the discrete sampling problem into an optimization task that can be cast as a multi-armed bandit problem (Chen & Ghahramani, 2016), see Section 5.2 below.

In addition to perturb-and-MAP methods, we are aware of three other approaches to estimate the partition function of a discrete graphical model via MAP solver calls. The WISH method (weighted-integrals-and-sums-by-hashing, Ermon et al., 2013) relies on repeated MAP inference calls applied to the model after subjecting it to random hash constraints. The Frank-Wolfe method may be applied by iteratively updating marginals using a constrained MAP solver and line search (Belanger et al., 2013; Krishnan et al., 2015). Weller & Jebara (2014a) instead use just one MAP call over a discretized mesh of marginals to approximate the Bethe partition function, which itself is an estimate (which often performs well) of the true partition function.

2. Relatives of the Gumbel Trick

In this section, we review the Gumbel trick and state the mechanism by which it can be generalized into an entire family of tricks. We show how these tricks can equivalently be viewed as averaging standard Gumbel perturbations in different spaces, instantiate several examples, and compare the various tricks' properties.

Notation Throughout this paper, let $\mathcal{X}$ be a finite sample space of size $N := |\mathcal{X}|$. Let $\tilde{p} : \mathcal{X} \to [0, \infty)$ be an unnormalized mass function over $\mathcal{X}$ and let $Z := \sum_{x \in \mathcal{X}} \tilde{p}(x)$ be its normalizing partition function. Write $p(x) := \tilde{p}(x)/Z$ for the normalized version of $\tilde{p}$, and $\phi(x) := \ln \tilde{p}(x)$ for the log-unnormalized probabilities, i.e. the potential function.

We write $\mathrm{Exp}(\lambda)$ for the exponential distribution with rate (inverse mean) λ and $\mathrm{Gumbel}(\mu)$ for the Gumbel distribution with location μ and scale 1. The latter has mean $\mu + c$, where $c \approx 0.5772$ is the Euler-Mascheroni constant.

2.1. The Gumbel Trick

Similarly to the connection between the Gumbel trick and the Poisson process established by Maddison (2016), we introduce the Gumbel trick for discrete probability distributions using a simple and elegant construction via *competing exponential clocks*. Consider N independent clocks,

Table 1: New tricks for constructing unbiased estimators of different transformations $f(Z)$ of the partition function.

Trick	$g(x)$	Mean $f(Z)$	Variance of $g(T)$	Asymptotic var. of $\hat{Z}$
Gumbel	$-\ln x - c$	$\ln Z$	$\frac{\pi^2}{6}$	$\frac{\pi^2}{6}Z^2$
Exponential	x	$\frac{1}{Z}$	$\frac{1}{Z^2}$	Z^2
Weibull α	$x^\alpha, \alpha > 0$	$Z^{-\alpha}\Gamma(1+\alpha)$	$\frac{\Gamma(1+2\alpha)-\Gamma(1+\alpha)^2}{Z^{2\alpha}}$	$\frac{1}{\alpha^2}\left(\frac{\Gamma(1+2\alpha)}{\Gamma(1+\alpha)^2}-1\right)Z^2$
Fréchet α	$x^\alpha, \alpha \in (-1,0)$	$Z^{-\alpha}\Gamma(1+\alpha)$	$\frac{\Gamma(1+2\alpha)-\Gamma(1+\alpha)^2}{Z^{2\alpha}}$ for $\alpha > -\frac{1}{2}$	$\frac{1}{\alpha^2}\left(\frac{\Gamma(1+2\alpha)}{\Gamma(1+\alpha)^2}-1\right)Z^2$
Pareto	e^x	$\frac{Z}{Z-1}$ for $Z > 1$	$a\frac{Z}{(Z-1)^2(Z-2)}$ for $Z > 2$	$\frac{Z^2}{(Z-2)^2}$
Tail t	$\mathbb{1}_{\{x>t\}}$	e^{-tZ}	$e^{-tZ}(1-e^{-tZ})$	$\frac{(1-e^{-tZ})^2}{t^2}$

started simultaneously, such that the j-th clock rings after a random time $T_j \sim \mathrm{Exp}(\lambda_j)$. Then it is easy to show that (1) the time until some clock rings has $\mathrm{Exp}(\sum_{j=1}^N \lambda_j)$ distribution, and (2) the probability of the j-th clock ringing first is proportional to its rate λ_j. These properties are also widely used in survival analysis (Cox & Oakes, 1984).

Consider N competing exponential clocks $\{T_x\}_{x\in\mathcal{X}}$, indexed by elements of $\mathcal{X}$, with respective rates $\lambda_x = \tilde{p}(x)$. Property (1) of competing exponential clocks tells us that

$$\min_{x\in\mathcal{X}}\{T_x\} \sim \mathrm{Exp}(Z). \tag{1}$$

Property (2) says that the random variable $\mathrm{argmin}_x T_x$, taking values in $\mathcal{X}$, is distributed according to p:

$$\operatorname*{argmin}_{x\in\mathcal{X}}\{T_x\} \sim p. \tag{2}$$

The Gumbel trick is obtained by applying the function $g(x) = -\ln x - c$ to the equalities in distribution (1) and (2). When g is applied to an $\mathrm{Exp}(\lambda)$ random variable, the result follows the $\mathrm{Gumbel}(-c+\ln\lambda)$ distribution, which can also be represented as $\ln\lambda + \gamma$, where $\gamma \sim \mathrm{Gumbel}(-c)$. Defining $\{\gamma(x)\}_{x\in\mathcal{X}} \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(-c)$ and noting that g is strictly decreasing, applying the function g to equalities in distribution (1) and (2), we obtain:

$$\max_{x\in\mathcal{X}}\{\phi(x)+\gamma(x)\} \sim \mathrm{Gumbel}(-c+\ln Z), \tag{1'}$$

$$\operatorname*{argmax}_{x\in\mathcal{X}}\{\phi(x)+\gamma(x)\} \sim p, \tag{2'}$$

where we have recalled that $\phi(x) = \ln\lambda_x = \ln\tilde{p}(x)$. The distribution $\mathrm{Gumbel}(-c+\ln Z)$ has mean $\ln Z$, and thus the log partition function can be estimated by averaging samples (Hazan & Jaakkola, 2012).

2.2. Constructing New Tricks

Given the equality in distribution (1), we can treat the problem of estimating the partition function Z as a parameter estimation problem for the exponential distribution. Applying the function $g(x) = -\ln x - c$ as in the Gumbel trick to obtain a $\mathrm{Gumbel}(-c+\ln Z)$ random variable, and estimating its mean to obtain an unbiased estimator of $\ln Z$, is just one way of inferring information about Z.

We consider applying different functions g to (1); particularly those functions g that transform the exponential distribution to another distribution with known mean. As the original exponential distribution has rate Z, the transformed distribution will have mean $f(Z)$, where f will in general no longer be the logarithm function. Since we often are interested in estimating various transformations $f(Z)$ of Z, this provides us a with a collection of unbiased estimators from which to choose. Moreover, further transforming these estimators yields a collection of (biased) estimators for other transformations of Z, including Z itself.

Example 1 (Weibull tricks)**.** For any $\alpha > 0$, applying the function $g(x) = x^\alpha$ to an $\mathrm{Exp}(\lambda)$ random variable yields a random variable with the $\mathrm{Weibull}(\lambda^{-\alpha}, \alpha^{-1})$ distribution with scale $\lambda^{-\alpha}$ and shape α^{-1}, which has mean $\lambda^{-\alpha}\Gamma(1+\alpha)$ and can be also represented as $\lambda^{-\alpha}W$, where $W \sim \mathrm{Weibull}(1, \alpha^{-1})$. Defining $\{W(x)\}_{x\in\mathcal{X}} \overset{\text{i.i.d.}}{\sim} \mathrm{Weibull}(1, \alpha^{-1})$ and noting that g is increasing, applying g to the equality in distribution (1) gives

$$\min_{x\in\mathcal{X}}\{\tilde{p}^{-\alpha}W(x)\} \sim \mathrm{Weibull}(Z^{-\alpha}, \alpha^{-1}). \tag{1"}$$

Estimating the mean of $\mathrm{Weibull}(Z^{-\alpha}, \alpha^{-1})$ yields an unbiased estimator of $Z^{-\alpha}\Gamma(1+\alpha)$. The special case $\alpha = 1$ corresponds to the identity function $g(x) = x$; we call the resulting trick the *Exponential trick*. □

Table 1 lists several examples of tricks derived this way. As Example 1 shows, these tricks may not involve additive perturbation of the potential function $\phi(x)$; the Weibull tricks multiplicatively perturb exponentiated unnormalized probabilities $\tilde{p}^{-\alpha}$ with Weibull noise. As models of interest are often specified in terms of potential functions, to be able to reuse existing MAP solvers in a black-box manner with the new tricks, we seek an equivalent formulation in terms of the potential function. The following Proposition shows that by not passing the function g through the minimization in equation (1), the new tricks can be equivalently formulated as averaging additive Gumbel perturbations of the potential function in different spaces.

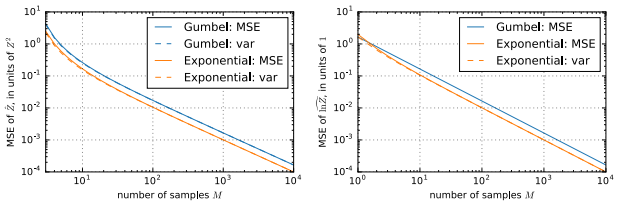


Figure 1: Analytically computed MSE and variance of Gumbel and Exponential trick estimators of Z (left) and $\ln Z$ (right). The MSEs are dominated by the variance, so the dashed and solid lines mostly overlap. See Section 2.3.2 for details.

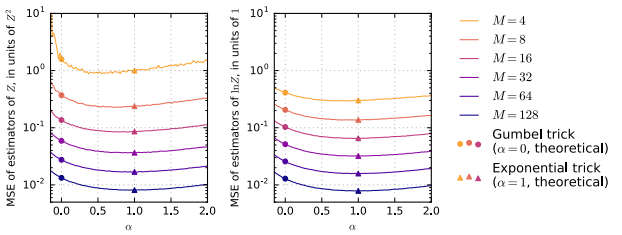


Figure 2: MSE of estimators of Z (left) and $\ln Z$ (right) stemming from Fréchet ($-\frac{1}{2} < \alpha < 0$), Gumbel ($\alpha = 0$) and Weibull tricks ($\alpha > 0$). See Section 2.3.2 for details.

Proposition 2. *For any function $g : [0, \infty) \to \mathbb{R}$ such that $f(Z) = \mathbb{E}_{T\sim\mathrm{Exp}(Z)}[g(T)]$ exists, we have*

$$f(Z) = \mathbb{E}_\gamma \left[g\left(e^{-c} \exp\left(- \max_{x\in\mathcal{X}} \{\phi(x) + \gamma(x)\} \right) \right) \right],$$

where $\{\gamma(x)\}_{x\in\mathcal{X}} \overset{i.i.d.}{\sim} \mathrm{Gumbel}(-c)$.

Proof. As $\max_x\{\phi(x) + \gamma(x)\} \sim \mathrm{Gumbel}(-c + \ln Z)$, we have $e^{-c}\exp(\max_x\{\phi(x)+\gamma(x)\}) \sim \mathrm{Exp}(Z)$ and the result follows by the assumption relating f and g. □

Proposition 2 shows that the new tricks can be implemented by solving the same MAP problems $\max_x\{\phi(x)+\gamma(x)\}$ as in the Gumbel trick, and then merely passing the solutions through the function $x \mapsto g(e^{-c}\exp(x))$ before averaging them to approximate the expectation.

2.3. Comparing Tricks

2.3.1. Asymptotic efficiency

The Delta method (Casella & Berger, 2002) is a simple technique for assessing the asymptotic variance of estimators that are obtained by a differentiable transformation of an estimator with known variance. The last column in Table 1 lists asymptotic variances of corresponding tricks when unbiased estimators of $f(Z)$ are passed through the function f^{-1} to yield (biased, but consistent and non-negative) estimators of Z itself. It is interesting to examine the constants that multiply Z^2 in some of the obtained asymptotic variance expressions for the different tricks. For example, it can be shown using Gurland's ratio (Gurland, 1956) that this constant is at least 1 for the Weibull and Fréchet tricks, which is precisely the value achieved by the Exponential trick (which corresponds to $\alpha = 1$). Moreover, the Gumbel trick constant $\pi^2/6$ can be shown to be the limit as $\alpha \to 0$ of the Weibull and Fréchet trick constants. In particular, the constant of the Exponential trick is strictly better than that of the standard Gumbel trick: $1 < \pi^2/6 \approx 1.65$. This motivates us to compare the Gumbel and Exponential tricks in more detail.

2.3.2. Mean squared error (MSE)

For estimators Y, their $\mathrm{MSE}(Y) = \mathrm{var}(Y) + \mathrm{bias}(Y)^2$ is a commonly used comparison metric. When the Gumbel or Exponential tricks are used to estimate either Z or $\ln Z$, the biases, variances, and MSEs of the estimators can be computed analytically using standard methods (Appendix A).

For example, the unbiased estimator of $\ln Z$ from the Gumbel trick can be turned into a consistent non-negative estimator of Z by exponentiation: $Y = \exp(\frac{1}{M}\sum_{m=1}^{M} X_m)$, where $X_1, \ldots, X_M \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(-c + \ln Z)$ are obtained using equation (1'). The bias and variance of Y can be computed using independence and the moment generating functions of the X_m's, see Appendix A for details.

Perhaps surprisingly, all estimator properties only depend on the true value of Z and not on the structure of the model (distribution p), since the estimators rely only on i.i.d. samples of a $\mathrm{Gumbel}(-c + \ln Z)$ random variable. Figure 1 shows the analytically computed estimator variances and MSEs. For estimating Z itself (left), the Exponential trick outperforms the Gumbel trick in terms of MSE for all sample sizes $M \geq 3$ (for $M \in \{1, 2\}$, both estimators have infinite variance and MSE). The ratio of MSEs quickly approaches $\pi^2/6$, and in this regime the Exponential trick requires $1 - 6/\pi^2 \approx 39\%$ fewer samples than the Gumbel trick to reach the same MSE. Also, for estimating $\ln Z$, (Figure 1, right), the Exponential trick provides a lower MSE estimator for sample sizes $M \geq 2$; only for $M = 1$ the Gumbel trick provides a better estimator.

Note that as biases are available analytically, the estimators can be easily debiased (by subtracting their bias). One then obtain estimators with MSEs equal to the variances of the original estimators, shown dashed in Figure 1. The Exponential trick would then always outperform the Gumbel trick when estimating $\ln Z$, even with sample size $M = 1$.

For Weibull tricks with $\alpha \neq 1$ and Fréchet tricks, we estimated the biases and variances of estimators of Z and $\ln Z$ by constructing $K = 100,000$ estimators in each case and evaluating their bias and variance. Figure 2 shows the results for varying α and several sample sizes M. We plot the

analytically computed value for the Gumbel trick at $\alpha = 0$, as we observe that the Weibull trick interpolates between the Gumbel trick and the Exponential trick as α increases from 0 to 1. We note that the minimum MSE estimator is obtained by choosing a value of α that is close to 1, i.e. the Exponential trick. This agrees with the finding from Section 2.3.1 that $\alpha = 1$ is optimal as $M \to \infty$.

2.4. Bayesian Perspective

A Bayesian approach exposes two choices when constructing estimators of Z, or of its transformations $f(Z)$:

1. A choice of prior distribution $p_0(Z)$, encoding prior beliefs about the value of Z before any observations.
2. A choice of how to summarize the posterior distribution $p_M(Z|X_1, \dots, X_M)$ given M samples.

Taking the Jeffrey's prior $p_0(Z) \propto Z^{-1}$, an improper prior that it is invariant under reparametrization, observing M samples $X_1, \dots, X_M \overset{\text{i.i.d.}}{\sim} \text{Exp}(Z)$ yields the posterior:

$$p_M(Z|X_1, \dots, X_M) \propto Z^{M-1} e^{-Z \sum_{m=1}^{M} X_m}.$$

Recognizing the density of a $\text{Gamma}(M, \sum_{m=1}^{M} X_m)$ random variable, the posterior mean is

$$\mathbb{E}[Z|X_1, \dots, X_M] = \frac{M}{\sum_{m=1}^{M} X_m} = \left(\frac{1}{M} \sum_{m=1}^{M} X_m \right)^{-1},$$

coinciding with the Exponential trick estimator of Z.

3. Low-rank Perturbations

One way of exploiting perturb-and-MAP to yield computational savings is to replace independent perturbations of each configuration's potential with an approximation. Such approximations are available e.g. in discrete graphical models, where the sampling space $\mathcal{X}$ has a product space structure $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, with $\mathcal{X}_i$ the state space of the i-th variable.

Definition 3 ((Hazan & Jaakkola, 2012)). The *sum-unary perturbation MAP value* is the random variable

$$U := \max_{\mathbf{x} \in \mathcal{X}} \left\{ \phi(\mathbf{x}) + \sum_{i=1}^{n} \gamma_i(x_i) \right\},$$

where $\{\gamma_i(x_i) \mid x_i \in \mathcal{X}_i, 1 \le i \le n\} \overset{\text{i.i.d}}{\sim} \text{Gumbel}(-c)$.

This definition involves $|\mathcal{X}_1| + \cdots + |\mathcal{X}_n|$ i.i.d. Gumbel random variables, rather than $|\mathcal{X}|$. (With $n = 1$ this coincides with full-rank perturbations and $U \sim \text{Gumbel}(-c + \ln Z)$.) For $n > 2$ the distribution of U is not available analytically. One can similarly define the *pairwise* (or higher-order) *perturbations*, where independent Gumbel noise is added to each pairwise (or higher-order) potential.

Unary perturbations provide the upper bound $\ln Z \le \mathbb{E}[U]$ on the log partition function (Hazan & Jaakkola, 2012), can be used to construct a sequential sampler for the Gibbs distribution (Hazan et al., 2013), and, if the perturbations are scaled down by a factor of n, a lower bound on $\ln Z$ can also be recovered (Hazan et al., 2013). In this section we show that a subfamily of tricks introduced in Section 2, consisting of Fréchet and Weibull (and Exponential) tricks, is applicable in the low-rank perturbation setting and use them to derive new families of upper and lower bounds on $\ln Z$ and sequential samplers for the Gibbs distribution. Please note full proofs are deferred to Appendix B and C.

3.1. Upper Bounds on the Partition Function

The following family of upper bounds on $\ln Z$ can be derived from the Fréchet and Weibull tricks.

Proposition 4. *For any $\alpha \in (-1, 0) \cup (0, \infty)$, the upper bound $\ln Z \le \mathcal{U}(\alpha)$ holds with*

$$\mathcal{U}(\alpha) := n \frac{\ln \Gamma(1 + \alpha)}{\alpha} + nc - \frac{1}{\alpha} \ln \mathbb{E}_\gamma \left[e^{-\alpha U} \right].$$

Proof. (Sketch.) By induction on n, with the induction step provided by our Clamping Lemma (Lemma 7) below. □

To evaluate these bounds in practice, $\mathbb{E}[e^{-\alpha U}]$ is estimated using samples of U. Corollary 9 of Hazan et al. (2016) can be used to show that $\text{var}(e^{-\alpha U})$ is finite for $\alpha > -\frac{1}{2\sqrt{n}}$, and so then the estimation is well-behaved.

A natural question is how these new bounds relate to the Gumbel trick upper bound $\ln Z \le \mathbb{E}[U]$ by Hazan & Jaakkola (2012). The following result aims to answers this:

Proposition 5. *The limit of $\mathcal{U}(\alpha)$ as $\alpha \to 0$ exists and equals $\mathcal{U}(0) := \mathbb{E}[U]$, i.e. the Gumbel trick upper bound.*

The question remains: When is it advantageous to use a value $\alpha \neq 0$ to obtain a tighter bound on $\ln Z$ than the Gumbel trick bound? The next result can provide guidance:

Proposition 6. *The function $\mathcal{U}(\alpha)$ is differentiable at $\alpha = 0$ and the derivative equals*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \mathcal{U}(\alpha) \Big|_{\alpha=0} = \frac{1}{2} \left(n \frac{\pi^2}{6} - \text{var}(U) \right).$$

While the variance of U is generally not tractable, in practice one obtains samples from U to estimate the expectation in $\mathcal{U}(\alpha)$ and these samples can be reused to assess $\text{var}(U)$. Interestingly, $\text{var}(U)$ equals $n\pi^2/6$ for both the uniform distribution and the distribution concentrated on a single configuration, and in our empirical investigations always $\text{var}(U) \le n\pi^2/6$. Then the derivative at 0 is non-negative and Fréchet tricks provide tighter bounds on $\ln Z$. However, as $\mathcal{U}(\alpha)$ is estimated with samples, the question of

estimator variance arises. We investigate the trade-off between tightness of the bound $\ln Z \leq \mathcal{U}(\alpha)$ and the variance incurred in estimating $\mathcal{U}(\alpha)$ empirically in Section 5.3.

3.2. Clamping

Consider the *partial sum-unary perturbation MAP values*, where the values of the first $j-1$ variables have been fixed, and only the rest are perturbed:

$$U_j(x_1, \ldots, x_{j-1}) := \max_{x_j, \ldots, x_n} \left\{ \phi(\mathbf{x}) + \sum_{i=j}^{n} \gamma_i(x_i) \right\}.$$

The following lemma involving the U_j's serves three purposes: (I.) it provides the induction step for Proposition 4, (II.) it shows that clamping never hurts partition function estimation with Fréchet and Weibull tricks, and (III.) it will be used to show that a sequential sampler constructed in Section 3.3 below is well-defined.

Lemma 7 (Clamping Lemma)**.** *For any $j \in \{1, \ldots, n\}$ and $(x_1, \ldots, x_{j-1}) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{j-1}$, the following inequality holds with any $\alpha \in (-1, 0) \cup (0, \infty)$:*

$$\sum_{x_j \in \mathcal{X}_j} \mathbb{E}_\gamma \left[e^{-(n-j)\ln\Gamma(1+\alpha) - \alpha(n-j)c} e^{-\alpha U_{j+1}} \right]^{-1/\alpha}$$
$$\leq \mathbb{E}_\gamma \left[e^{-(n-(j-1))\ln\Gamma(1+\alpha) - \alpha(n-(j-1))c} e^{-\alpha U_j} \right]^{-1/\alpha}$$

Proof. This follows directly from the Fréchet trick ($\alpha \in (-1, 0)$) or the Weibull trick ($\alpha > 0$) and representing the Fréchet resp. Weibull random variables in terms of Gumbel random variables. See Appendix B.1 for more details. □

Corollary 8. *Clamping never hurts $\ln Z$ estimation using any of the Fréchet or Weibull upper bounds $\mathcal{U}(\alpha)$.*

Proof. Applying the function $x \mapsto \ln(x)$ to both sides of the Clamping Lemma 7 with $j = 1$, the right-hand side equals $\mathcal{U}(\alpha)$, while the left-hand side is the estimate of $\ln Z$ after clamping variable x_1. □

This was shown previously in restricted settings (Hazan et al., 2013; Zhao et al., 2016). Similar results showing that clamping improves partition function estimation have been obtained for the mean field and TRW approximations (Weller & Domke, 2016), and in certain settings for the Bethe approximation (Weller & Jebara, 2014b) and L-FIELD (Zhao et al., 2016).

3.3. Sequential Sampling

Hazan et al. (2013) derived a sequential sampling procedure for the Gibbs distribution by exploiting the $\mathcal{U}(0)$ Gumbel trick upper bound on $\ln Z$. In the same spirit, one can derive sequential sampling procedures from the Fréchet and Weibull tricks, leading to the following algorithm.

Algorithm 1 Sequential sampler for Gibbs distribution

Input: $\alpha \in (-1, 0) \cup (0, \infty)$, potential function ϕ on $\mathcal{X}$
Output: a sample $\mathbf{x}$ from the Gibbs distribution $\propto e^{\phi(\mathbf{x})}$

1: **for** $j = 1$ **to** n **do**
2: **for** $x_j \in \mathcal{X}_j$ **do**
3: $p_j(x_j) \leftarrow \frac{e^{-c}}{\Gamma(1+\alpha)^{1/\alpha}} \frac{\mathbb{E}_\gamma\left[e^{-\alpha U_{j+1}(x_1,\ldots,x_j)}\right]^{-1/\alpha}}{\mathbb{E}_\gamma\left[e^{-\alpha U_j(x_1,\ldots,x_{j-1})}\right]^{-1/\alpha}}$
4: $p_j(\text{reject}) \leftarrow 1 - \sum_{x_j \in \mathcal{X}_j} p_j(x_j)$
5: $x_j \leftarrow$ sample according to p_j
6: **if** $x_j ==$ reject **then**
7: RESTART (goto 1)

This algorithm is well-defined if $p_j(\text{reject}) \geq 0$ for all j, which can be shown by canceling terms in the Clamping Lemma 7. We discuss correctness in Appendix B.2. As for the Gumbel sequential sampler of Hazan et al. (2013), the expected number of restarts (and hence the running time) only depend on the quality of the upper bound ($\mathcal{U}(\alpha) - \ln Z$), and not on the ordering of variables.

3.4. Lower Bounds on the Partition Function

Similarly as in the Gumbel trick case (Hazan et al., 2013), one can derive lower bounds on $\ln Z$ by perturbing an arbitrary subset S of variables.

Proposition 9. *Let $\mathcal{X} = \mathcal{X}_1 \times \cdots \mathcal{X}_n$ be a product space and ϕ a potential function on $\mathcal{X}$. Let $\alpha \in (-1, 0) \cup (0, \infty)$. For any subset $S \subseteq \{1, \ldots, n\}$ of the variables $x_1, \ldots, x_n$ we have $\ln Z \geq$*

$$c + \frac{\ln\Gamma(1+\alpha)}{\alpha} - \frac{1}{\alpha} \ln \mathbb{E}\left[e^{-\alpha \max_{\mathbf{x}} \{\phi(\mathbf{x}) + \gamma_S(\mathbf{x}_S)\}} \right],$$

where $\mathbf{x}_S := \{x_i : i \in S\}$ and $\gamma_S(\mathbf{x}_S) \sim \text{Gumbel}(-c)$ independently for each setting of $\mathbf{x}_S$.

By averaging n such lower bounds corresponding to singleton sets $S = \{i\}$ together, we obtain a lower bound on $\ln Z$ that involves the *average-unary perturbation MAP value*

$$L := \max_{\mathbf{x} \in \mathcal{X}} \left\{ \phi(\mathbf{x}) + \frac{1}{n} \sum_{i=1}^{n} \gamma_i(x_i) \right\}.$$

Corollary 10. *For any $\alpha \in (-1, 0) \cup (0, \infty)$, we have the lower bound $\ln Z \geq \mathcal{L}(\alpha)$, where*

$$\mathcal{L}(\alpha) := c + \frac{\ln\Gamma(1+\alpha)}{\alpha} - \frac{1}{n\alpha} \ln \mathbb{E}\left[\exp\left(-n\alpha L\right)\right].$$

Again, $\mathcal{L}(0) := \mathbb{E}[L]$ can be defined by continuity, where $\mathbb{E}[L] \leq \ln Z$ is the Gumbel trick lower bound by Hazan et al. (2013).

4. Advantages of the Gumbel Trick

We have seen how the Gumbel trick can be embedded into a continuous family of tricks, consisting of Fréchet, Exponential, and Weibull tricks. We showed that the new tricks can provide more efficient estimators of the partition function in the full-rank perturbation setting (Section 2), and in the low-rank perturbation setting lead to sequential samplers and new bounds on $\ln Z$, which can be also more efficient, as we investigate in Section 5.3. To balance the discussion of merits of different tricks, in this section we briefly highlight advantages of the Gumbel trick that stem from its simpler analytical form.

First, by consulting Table 1 we see that the function $g(x) = -\ln x - c$ has the property that the variance of the resulting estimator (of $\ln Z$) does not depend on the value of Z; the function g is a variance stabilizing transformation for the Exponential distribution.

Second, exploiting the fact that the logarithm function leads to additive perturbations, Maji et al. (2014) showed that the entropy of x^*, the configuration with maximum potential after sum-unary perturbation in the sense of Definition 3, can be bounded as $H(x^*) \leq B(p) := \sum_{i=1}^{n} \mathbb{E}_{\gamma_i}[\gamma_i(x_i^*)]$. We extend this result to show how the errors of bounding $\ln Z$, sampling, and entropy estimation are related:

Proposition 11. *Writing p for the Gibbs distribution and $B(p) := \mathbb{E}_{\gamma_i}[\gamma_i(x_i^*)]$ for the entropy bound, we have*

$$\underbrace{(\mathcal{U}(0) - \ln Z)}_{\text{error in } \ln Z \text{ bound}} + \underbrace{\mathrm{KL}(x^* \parallel p)}_{\text{sampling error}} = \underbrace{B(p) - H(x^*)}_{\text{error in entropy estimation}} .$$

Third, the additive character of the Gumbel perturbations can also be used to derive a new result relating the error of the lower bound $\mathcal{L}(0)$ and of sampling x^{**} as the configuration achieving the maximum average-unary perturbation value L, instead of sampling from the Gibbs distribution p:

Proposition 12. *Writing p for the Gibbs distribution,*

$$\underbrace{\ln Z - \mathcal{L}(0)}_{\text{error in } \ln Z \text{ bound}} \geq \underbrace{\mathrm{KL}(x^{**} \parallel p)}_{\text{sampling error}} \geq 0.$$

Remark. While we knew from Hazan et al. (2013) that $\ln Z - \mathcal{L}(0) \geq 0$, this is a stronger result showing that the size of the gap is an upper bound on the KL divergence between the approximate sampling distribution of x^{**} and the Gibbs distribution p.

Proofs of the new results appear in Appendix B.3 and C.2.

Fourth, viewed as a function of the Gumbel perturbations γ, the random variable U has a bounded gradient, allowing earlier measure concentration results (Orabona et al., 2014; Hazan et al., 2016). Proving similar measure concentration results for the expectations $\mathbb{E}[e^{-\alpha U}]$ appearing in $\mathcal{U}(\alpha)$ for $\alpha \neq 0$ may be more challenging.

5. Experiments

We conducted experiments with the following aims:

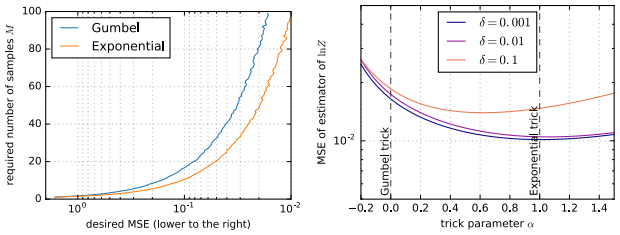
1. To show that the higher efficiency of the Exponential trick in the full-rank perturbation setting is useful in practice, we compared it to the Gumbel trick in A* sampling (Maddison et al., 2014) (Section 5.1) and in the large-scale discrete sampling setting of Chen & Ghahramani (2016) (Section 5.2).
2. To show that non-zero values of α can lead to better estimators of $\ln Z$ in the low-rank perturbation setting as well, we compare the Fréchet and Weibull trick bounds $\mathcal{U}(\alpha)$ to the Gumbel trick bound $\mathcal{U}(0)$ on a common discrete graphical model with different coupling strengths; see Section 5.3.

5.1. A* Sampling

A* sampling (Maddison et al., 2014) is a sampling algorithm for continuous distributions that perturbs the log-unnormalized density ϕ with a continuous generalization of the Gumbel trick, called the Gumbel process, and uses a variant of A* search to find the location of the maximum of the perturbed ϕ. Returning the location yields an exact sample from the original distribution, as in the discrete Gumbel trick. Moreover, the corresponding maximum value also has the $\mathrm{Gumbel}(-c + \ln Z)$ distribution (Maddison et al., 2014). Our analysis in Section 2.3 tells us that the Exponential trick yields an estimator with lower MSE than the Gumbel trick; we briefly verified this on the Robust Bayesian Regression experiment of Maddison et al. (2014). We constructed estimators of $\ln Z$ from the Gumbel and Exponential tricks (debiased version, see Section 2.3.2), and assessed their variances by constructing each estimator $K = 1000$ times and looking at the sample variance. Figure 3a shows that the Exponential trick requires up to 40% fewer samples to reach a given MSE.

5.2. Scalable Partition Function Estimation

Chen & Ghahramani (2016) considered sampling from a discrete distribution of the form $p(x) \propto f_0(x) \prod_{s=1}^{S} f_s(x)$ when the number of factors S is large relative to the sample space size $|\mathcal{X}|$. Computing i.i.d. Gumbel perturbations $\gamma(x)$ for each $x \in \mathcal{X}$ is then relatively cheap compared to evaluating all potentials $\phi(x) = f_0(x) + \sum_{s=1}^{S} \ln f_s(x)$. Chen & Ghahramani (2016) observed that each (perturbed) potential can be estimated by subsampling the factors, and potentials that appear unlikely to yield the MAP value can be pruned off from the search early on. The authors formalized the problem as a Multi-armed bandit problem with a finite reward population and derived approximate algorithms for efficiently finding the maximum perturbed potential with a probabilistic guarantee.



(a) (b)

Figure 3: (a) Sample size M required to reach a given MSE using Gumbel and Exponential trick estimators of $\ln Z$, using samples from A^* *sampling* (see Section 5.1) on a Robust Bayesian Regression task. The Exponential trick is more efficient, requiring up to 40% fewer samples to reach a given MSE. (b) MSE of $\ln Z$ estimators for different values of α, using $M = 100$ samples from the approximate MAP algorithm discussed in Section 5.2, with different error bounds δ. For small δ, the Exponential trick is close to optimal, matching the analysis of Section 2.3.2. For larger δ, the Weibull trick interpolation between the Gumbel and Exponential tricks can provide an estimator with lower MSE.

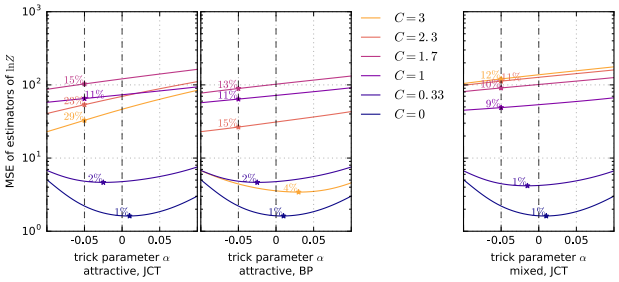


Figure 4: MSEs of $\mathcal{U}(\alpha)$ as estimators of $\ln Z$ on 10×10 attractive (left, middle) and mixed (right) spin glass model with different coupling strengths C (see Section 5.3). We also show the percentage of samples saved by using the best α in place of the Gumbel trick estimator $\mathcal{U}(0)$, assuming the asymptotic regime. For this we only considered $\alpha > -1/(2\sqrt{n}) = -0.05$, where variance is provably finite, see Section 3.1. The MAP problems were solved using the exact junction tree algorithm (JCT, left and right), or approximate belief propagation (BP, middle). In all cases, when coupling is very low, α close to 0 is optimal. This also holds for BP when coupling is high. In other regimes, upper bounds for the Fréchet trick, i.e. $\alpha < 0$, provide more accurate estimators.

While Chen & Ghahramani (2016) considered sampling, by modifying their procedure to return the value of the maximum perturbed potential rather than the argmax (cf equations (1) and (2)), we can estimate the partition function instead. However, the approximate algorithm only guarantees to find the MAP configuration with a probability $1 - \delta$. Figure 3b shows the results of running the Racing-Normal algorithm of Chen & Ghahramani (2016) on the synthetic dataset considered by the authors with the "very hard" noise setting $\sigma = 0.1$. For low error bounds δ the Exponential trick remained close to optimal, but for a larger error bound the Weibull trick interpolation between the Gumbel and Exponential tricks proved useful to provide an estimator with lower MSE.

5.3. Low-rank Perturbation Bounds on $\ln Z$

Hazan & Jaakkola (2012) evaluated tightness of the Gumbel trick upper bound $\mathcal{U}(0) \geq \ln Z$ on 10×10 binary spin glass models. We show one can obtain more accurate estimates of $\ln Z$ on such models by choosing $\alpha \neq 0$. To account for the fact that in practice an expectation in $\mathcal{U}(\alpha)$ is replaced with a sample average, we treat $\mathcal{U}(\alpha)$ as an estimator of $\ln Z$ with asymptotic bias equal to the bound gap $(\mathcal{U}(\alpha) - \ln Z)$, and estimate its MSE.

Figure 4 shows the MSEs of $\mathcal{U}(\alpha)$ as estimators of $\ln Z$ on 10×10 ($n = 100$) binary pairwise grid models with unary potentials sampled uniformly from $[-1, 1]$ and pairwise potentials from $[0, C]$ (*attractive* models) or from $[-C, C]$ (*mixed* models), for varying coupling strengths C. We replaced the expectations in $U(\alpha)$'s with sample averages of size $M = 100$, using libDAI (Mooij, 2010) to solve the MAP problems yielding these samples. We constructed each estimator 1000 times to assess its variance.

6. Discussion

By casting partition function evaluation as a parameter estimation problem for the exponential distribution, we derived a family of methods of which the Gumbel trick is a special case. These methods can be equivalently seen as (1) perturbing models using different distributions, or as (2) averaging standard Gumbel perturbations in different spaces, allowing implementations with little additional cost.

We showed that in the full-rank perturbation setting, the new Exponential trick provides an estimator with lower MSE, or instead allows using up to 40% fewer samples than the Gumbel trick estimator to reach the same MSE.

In the low-rank perturbation setting, we used our Fréchet, Exponential and Weibull tricks to derive new bounds on $\ln Z$ and sequential samplers for the Gibbs distribution, and showed that these can also behave better than the corresponding Gumbel trick results. However, the optimal trick to use (as specified by α) depends on the model, sample size, and MAP solver used (if approximate). Since in practice the dominant computational cost is carried by solving repeated instances of the MAP problem, one can try and assess different values of α on the problem at hand. That said, we believe that investigating when different tricks yield better results is an interesting avenue for future work.

Finally, we balanced the discussion by pointing out that the Gumbel trick has a simpler analytical form which can be exploited to derive more interesting theoretical statements in the low-rank perturbation setting. Beyond existing results, we derived new connections between errors of different procedures using low-rank Gumbel perturbations.

Acknowledgements

The authors thank Tamir Hazan for helpful discussions, and Mark Rowland, Maria Lomeli, and the anonymous reviewers for helpful comments. AW acknowledges support by the Alan Turing Institute under EPSRC grant EP/N510129/1, and by the Leverhulme Trust via the CFI.

References

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

Bertasius, G., Liu, Q., Torresani, L., and Shi, J. Local Perturb-and-MAP for Structured Prediction. In *AISTATS*, 2017.

Boykov, Y., Veksler, O., and Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.

Casella, G. and Berger, R. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Chen, Y. and Ghahramani, Z. Scalable discrete sampling as a multi-armed bandit problem. In *ICML*, 2016.

Cox, D. and Oakes, D. *Analysis of survival data*, volume 21. CRC Press, 1984.

Darbon, J. Global optimization for first order Markov random fields with submodular priors. *Discrete Applied Mathematics*, 157(16):3412 – 3423, 2009.

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

Gurland, J. An inequality satisfied by the Gamma function. *Scandinavian Actuarial Journal*, 1956(2):171–172, 1956.

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

Hazan, T., Maji, S., and Jaakkola, T. On sampling from the Gibbs distribution with random maximum a-posteriori perturbations. In *NIPS*. 2013.

Hazan, T., Orabona, F., Sarwate, A., Maji, S., and Jaakkola, T. High dimensional inference with random maximum a-posteriori perturbations. *CoRR*, abs/1602.03571, 2016.

Kim, C., Sabharwal, A., and Ermon, S. Exact sampling with integer linear programs and random perturbations. In *AAAI*, pp. 3248–3254, 2016.

Kolmogorov, V. Convergent tree-reweighted message passing for energy minimization. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1568–1583, 2006.

Krishnan, Rahul G, Lacoste-Julien, Simon, and Sontag, David. Barrier Frank-Wolfe for Marginal Inference. In *NIPS*. 2015.

Lauritzen, S. *Graphical models*. Oxford statistical science series. Clarendon Press, Oxford, 1996. Autre tirage : 1998.

Maddison, C. A Poisson process model for Monte Carlo. In Hazan, T., Papandreou, G., and Tarlow, D. (eds.), *Perturbation, Optimization, and Statistics*. MIT Press, 2016.

Maddison, C., Tarlow, D., and Minka, T. A* sampling. In *NIPS*. 2014.

Maji, S., Hazan, T., and Jaakkola, T. Active boundary annotation using random MAP perturbations. In *AISTATS*, 2014.

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

Orabona, F., Hazan, T., Sarwate, A., and Jaakkola, T. On measure concentration of random maximum a-posteriori perturbations. In *ICML*, 2014.

Papandreou, G. and Yuille, A. Gaussian sampling by local perturbations. In *NIPS*. 2010.

Papandreou, G. and Yuille, A. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 193–200, November 2011.

Tarlow, D., Adams, R., and Zemel, R. Randomized optimum models for structured prediction. In *AISTATS*, 2012.

Wainwright, M. and Jordan, M. Graphical Models, Exponential Families, and Variational Inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January 2008.

Weller, A. and Domke, J. Clamping improves TRW and mean field approximations. In *AISTATS*, 2016.

Weller, A. and Jebara, T. Approximating the Bethe partition function. In *UAI*, 2014a.

Weller, A. and Jebara, T. Clamping variables and approximate inference. In *NIPS*, 2014b.

Wright, S. and Nocedal, J. Numerical optimization. *Springer Science*, 35:67–68, 1999.

Zhao, J., Djolonga, J., Tschiatschek, S., and Krause, A. Variable clamping for optimization-based inference. In *NIPS Workshop on Advances in Approximate Bayesian Inference*, December 2016.