High Dimensional Inference with Random Maximum A-Posteriori Perturbations

Citation: Hazan, Tamir et al. "High Dimensional Inference with Random Maximum A-Posteriori Perturbations." IEEE Transactions on Information Theory, 65, 10 (May 2019): 6539 - 6560 © 2019 The Author(s)

As Published: 10.1109/TIT.2019.2916805

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: https://hdl.handle.net/1721.1/129369

Version: Original manuscript: author's manuscript prior to formal peer review

High Dimensional Inference with Random Maximum A-Posteriori Perturbations

Tamir Hazan, Francesco Orabona, Anand D. Sarwate *Senior Member, IEEE,* Subhransu Maji and Tommi Jaakkola

Abstract

This paper presents a new approach, called perturb-max, for high-dimensional statistical inference that is based on applying random perturbations followed by optimization. This framework injects randomness to maximum a-posteriori (MAP) predictors by randomly perturbing the potential function for the input. A classic result from extreme value statistics asserts that perturb-max operations generate unbiased samples from the Gibbs distribution using high-dimensional perturbations. Unfortunately, the computational cost of generating so many high-dimensional random variables can be prohibitive. However, when the perturbations are of low dimension, sampling the perturb-max prediction is as efficient as MAP optimization. This paper shows that the expected value of perturb-max inference with low dimensional perturbations can be used sequentially to generate unbiased samples from the Gibbs distribution. Furthermore the expected value of the maximal perturbations is a natural bound on the entropy of such perturb-max models. A measure concentration result for perturb-max values shows that the deviation of their sampled average from its expectation decays exponentially in the number of samples, allowing effective approximation of the expectation.

Manuscript received February 10, 2016; revised November 4, 2016; accepted xxxxxxxxxx

T. Hazan is with the Faculty of Industrial Engineering & Management, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: tamir.hazan@technion.ac.il). F. Orabona is with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-2424. This work was done in part when he was with Yahoo Labs, 229 W 43rd St., New York, NY 10036, USA (e-mail: francesco@orabona.com). A.D. Sarwate is with the Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, 94 Brett Road, Piscataway NJ 08854, USA (e-mail: asarwate@ece.rutgers.edu). S. Maji is with the College of Information and Computer Sciences, University of Massachusetts Amherst, 140 Governors Drive, MA 01003-9264, USA (e-mail: smaji@cs.umass.edu). T. Jaakkola is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA (e-mail: tommi@csail.mit.edu).

Preliminary versions of these results appeared in conference proceedings [1]–[4].

Index Terms

Graphical models, MAP inference, Measure concentration, Markov Chain Monte Carlo

I. Introduction

Modern machine learning tasks in computer vision, natural language processing, and computational biology involve inference in high-dimensional models. Examples include scene understanding [5], parsing [6], and protein design [7]. In these settings, inference involves finding a likely assignment (or equivalently, structure) that fits the data: objects in images, parsers in sentences, or molecular configurations in proteins. Each structure corresponds to an assignment of values to random variables and the preference of a structure is based on defining potential functions that account for interactions over these variables. Given the observed data, these preferences yield a *posterior probability distribution* on assignments called the Gibbs distribution. The probability of an assignment is proportional to the exponential of the potential function value. High dimensional models that are commonly used in contemporary machine learning often incorporate local potential functions on the variables of the model that are derived from the data (signal) as well as higher order potential functions that account for interactions between the model variables and derived from domain-specific knowledge (coupling). The resulting posterior probability landscape is often "ragged"; in such landscapes Markov chain Monte Carlo (MCMC) approaches to sampling from the Gibbs distribution may become prohibitively expensive [8]–[10]. By contrast, when no data terms (local potential functions) exist, MCMC approaches can be quite successful. These methods include Gibbs sampling [11], Metropolis-Hastings [12], or Swendsen-Wang [13].

An alternative to sampling from the Gibbs distribution is to look for the *maximum a posteriori probability* (MAP) assignment. Substantial effort has gone into developing optimization algorithms for recovering MAP assignments by exploiting domain-specific structural restrictions [5], [14]–[18] or by linear programming relaxations [7], [19]–[22]. MAP inference is nevertheless limiting when there are a number of alternative likely assignments. Such alternatives arise either from inherent ambiguities (e.g., in image segmentation or text analysis) or due to the use of computationally/representationally limited potential functions (e.g., super-modularity) aliasing alternative assignments to have similar scores. For an example, see Figure 1.

Recently, several works have leveraged the current efficiency of MAP solvers to build (approximate) samplers for the Gibbs distribution, thereby avoiding the computational burden of MCMC methods [2], [23]–[33]. These works have shown that one can represent the Gibbs distribution by calculating the MAP

assignment of a *randomly perturbed potential function*, whenever the perturbations follow the Gumbel distribution [23], [24]. Unfortunately the total number of assignment (or structures), and consequently the total number of random perturbations, is exponential in the structure's dimension. We call this a *perturb-max* approach.

In this work, we perform high dimensional inference tasks using the expected value of perturb-max programs that are restricted to low dimensional perturbations. In this setting, the number of random perturbations is linear is the assignment's dimension and as a result statistical inference is as fast as computing the MAP assignment, as illustrated in Figure 1. We also provide measure concentration inequalities that show the expected perturb-max value can be estimated with high probability using only a few random samples. This work simplifies and extends our preliminary results [1]–[4].

We begin by introducing the setting of high dimensional inference as well as the necessary background in extreme value statistics in Section II. Subsequently, we develop high dimensional inference algorithms that rely on the expected MAP value of randomly perturbed potential functions, while using only low dimensional perturbations. In Section III-A we propose a novel sampling algorithm and in Section III-C we derive bounds on the entropy that may be of independent interest. Finally, we show that the expected value of the perturb-max value can be estimated efficiently despite the unboundedness of the perturbations. To show this we must prove new measure concentration results for the Gumbel distribution. In particular, in Section IV we prove new Poincaré and modified log-Sobolev inequalities for (non-strictly) log-concave distributions.

II. Inference and random perturbations

We first describe the high dimensional statistical inference problems that motivate this work. These involve defining the potential function, the Gibbs distribution, and its entropy. Further background can be found in standard texts on graphical models [34]. We will then describe the MAP inference problem and describe how to use extreme value statistics to perform statistical inference while recovering the maximal assignment of randomly perturbed potential functions [35] [36, pp.159–61]. To do this, we apply random perturbations to the potential function and use MAP solvers to produce a solution to the perturbed problem.

A. High dimensional models, inference and extreme value statistics

Statistical inference for high dimensional problems involves reasoning about the states of discrete variables whose configurations (assignments of values) describe discrete structures. Suppose that our

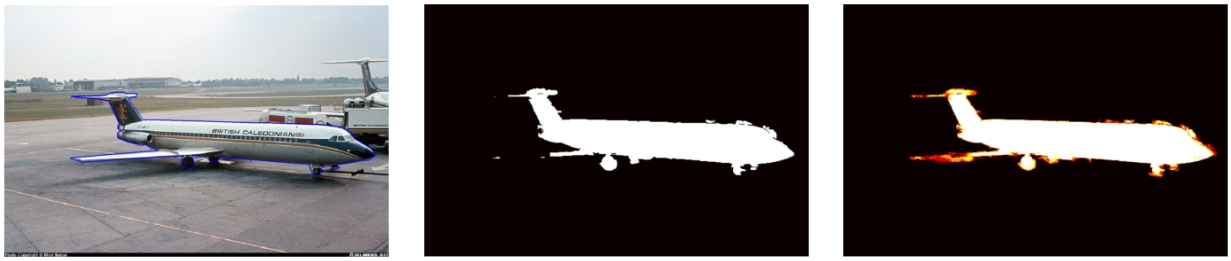


Fig. 1. Comparing MAP inference and perturbation models. A segmentation is modeled by $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where n is the number of pixels and $x_i \in \{0, 1\}$ is a discrete label relating a pixel to foreground ($x_i = 1$) or background ($x_i = 0$). $\theta(\mathbf{x})$ is the (super-modular) score of each segmentation. Left: original image along with the annotated boundary. Middle: the MAP segmentation $\operatorname{argmax}_{\mathbf{x}} \theta(\mathbf{x})$ recovered by the graph-cuts optimization algorithm using a region inside the boundary as seed pixels [15]. Note that the "optimal" solution is inaccurate because thin and long objects (wings) are labeled incorrectly. Right: The marginal probabilities of the perturb-max model estimated using 20 samples (random perturbations of $\theta(\mathbf{x})$ followed by executing graph-cuts). The information about the wings is recovered by these samples. Estimating the marginal probabilities of the corresponding Gibbs distribution by MCMC sampling is slow in practice and provably hard in theory [2], [9].

model has n variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where each x_i takes values in a discrete set $\mathcal{X}_i$. Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ so that $\mathbf{x} \in \mathcal{X}$. Let $\mathrm{Dom}(\theta) \subseteq \mathcal{X}$ be a subset of possible configurations and $\theta : \mathcal{X} \to \mathbb{R}$ be a potential function that gives a score to an assignment or structure $\mathbf{x}$. For convenience we define $\theta(\mathbf{x}) = -\infty$ for $\mathbf{x} \notin \mathrm{Dom}(\theta)$. The potential function induces a probability distribution on configurations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ via the Gibbs distribution:

$$p(\mathbf{x}) \triangleq \frac{1}{Z(\theta)} \exp(\theta(\mathbf{x})) \qquad \text{where} \qquad Z(\theta) \triangleq \sum_{\mathbf{x} \in \mathcal{X}} \exp(\theta(\mathbf{x})). \tag{1}$$

The normalization constant $Z(\theta)$ is called the partition function. Sampling from the Gibbs distribution is often difficult because the partition function involves exponentially many terms (equal to the number of discrete structures in $\mathcal{X}$). Computing the partition function is $\#P$-hard in general (e.g., Valiant [37]).

B. MAP inference

In practical inference tasks, the Gibbs distribution is constructed given observed data. Thus we call its maximizing assignment the maximum a-posteriori (MAP) prediction. We express the MAP inference problem as maximizing $p(\mathbf{x})$, which is defined in Equation (1). Since the exponent is a monotone function, maximizing $p(\mathbf{x})$ is equivalent to maximizing $\theta(\mathbf{x})$ and MAP prediction amounts to finding

$$\mathbf{x}^* = \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \theta(\mathbf{x}). \tag{2}$$

Methods for performing the optimization in (2) for high dimensional potential functions have been extensively researched in the last decade [5], [7], [15], [17], [18]. These have been useful in many cases of practical interest in computer vision, such as foreground-background image segmentation with supermodular potential functions (e.g., [38]), parsing and tagging (e.g., [6], [39]), branch and bound for scene understanding and pose estimation [40], [41], and dynamic programming predictions for outdoor scene understanding [42]. Although the run-time of these solvers can be exponential in the number of variables, they are often surprisingly effective in practice, [7], [19], [22], [43], [44].

C. Inference and extreme value statistics

Although MAP prediction is NP-hard in general, it is often simpler than sampling from the Gibbs distribution. Nevertheless, usually there are several values of $\mathbf{x}$ whose scores $\theta(\mathbf{x})$ are close to $\theta(\mathbf{x}^*)$ and we would like to sample these structures (see Figure 1). From such samples it is possible to estimate the amount of uncertainty in these models. A standard uncertainty measure is the entropy function:

$$H(p) = -\sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x}). \tag{3}$$

Sampling methods for posterior distributions often resort to MCMC algorithms that converge slowly in many practical settings [8]–[10].

An alternative approach to drawing unbiased samples from the Gibbs distribution is by randomly perturbing the potential function and solving the perturbed MAP problem. The "perturb-max" approach adds a random function $\gamma : \mathcal{X} \to \mathbb{R}$ to the potential function in (1) and solves the resulting MAP problem:

$$\mathbf{x}^* = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}, \tag{4}$$

where $\gamma(\mathbf{x})$ is a random function on $\mathcal{X}$. The simplest approach to designing a perturbation function is to associate an independent and identically distributed (i.i.d.) random variable $\gamma(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{X}$. In this case, the distribution of the perturb-max value $\theta(\mathbf{x})+\gamma(\mathbf{x})$ has an analytic form. To verify this observation we denote by $F(t) = \mathbb{P}(\gamma(\mathbf{x}) \leq t)$ the cumulative distribution function of $\gamma(\mathbf{x})$. The independence of $\gamma(\mathbf{x})$ across $\mathbf{x} \in \mathcal{X}$ implies that

$$\mathbb{P}_\gamma\left(\max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\} \leq t\right) = \mathbb{P}_\gamma\left(\forall \mathbf{x} \in \mathcal{X} : \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\} \leq t\right) \tag{5}$$

$$= \mathbb{P}_\gamma\left(\forall \mathbf{x} \in \mathcal{X} : \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\} \leq t\right) \tag{6}$$

$$= \prod_{\mathbf{x}\in\mathcal{X}} F(t - \theta(\mathbf{x})). \tag{7}$$

Unfortunately, the product of cumulative distribution functions is not usually a simple distribution.

The Gumbel, Fréchet, and Weibull distributions, used in extremal statistics, are max-stable distributions: the product $\prod_{\mathbf{x}\in\mathcal{X}} F(t-\theta(\mathbf{x}))$ can be described by their own cumulative distribution function $F(\cdot)$ [45]–[47]. In this work we focus on the Gumbel distribution with zero mean, which is described by a doubly exponential cumulative distribution function

$$G(t) = \exp(-\exp(-(t+c))), \tag{8}$$

where $c \approx 0.5772$ is the Euler-Mascheroni constant. Throughout our work we use the max-stability of the Gumbel distribution as described in the following theorem.

Theorem 1 (Max-stability of Gumbel perturbations [45]–[47]): Let $\gamma = \{\gamma(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ be a collection of i.i.d. Gumbel random variables whose cumulative distribution function is given by $G(t) = \mathbb{P}(\gamma(\mathbf{x}) \leq t) = \exp(-\exp(-(t+c)))$. Then the random variable $\max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}$ is distributed according to the Gumbel distribution whose mean is the log-partition function $\log Z(\theta)$.

Proof: The proof is known, but we include it for completeness. By the independence assumption,

$$\mathbb{P}_\gamma \left(\max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\} \leq t \right) = \prod_{x\in X} \mathbb{P}_{\gamma(\mathbf{x})} \left(\theta(\mathbf{x}) + \gamma(\mathbf{x}) \leq t\right).$$

The random variable $\theta(\mathbf{x}) + \gamma(\mathbf{x})$ follows the Gumbel distribution with mean $\theta(\mathbf{x})$. Therefore

$$\mathbb{P}_{\gamma(x)} \left(\theta(\mathbf{x}) + \gamma(\mathbf{x}) \leq t\right) = G(t - \theta(\mathbf{x})).$$

Lastly, the double exponential form of the Gumbel distribution yields the result:

$$\begin{aligned}
\prod_{x\in X} G(t-\theta(\mathbf{x})) &= \exp\left(-\sum_{\mathbf{x}\in\mathcal{X}} \exp\left(-(t-\theta(\mathbf{x})+c)\right)\right) \\
&= \exp\left(-\exp(-(t+c-\log Z(\theta))\right) \\
&= G(t - \log Z(\theta)).
\end{aligned}$$

■

We can use the log-partition function to recover the moments of the Gibbs distribution. Thus the log-partition function characterizes the stability of the randomized MAP predictor $\mathbf{x}^*$ in (4).

Corollary 1 (Sampling from perturb-max models [48]–[50]): Under the conditions of Theorem 1 the Gibbs distribution measures the stability of the perturb-max argument. That is, for all $\hat{\mathbf{x}}$,

$$\frac{\exp(\theta(\hat{\mathbf{x}}))}{Z(\theta)} = \mathbb{P}_\gamma \left(\hat{\mathbf{x}} = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\} \right), \tag{9}$$

Proof: From Theorem 1, we have $\log Z(\theta) = \mathbb{E}_\gamma[\max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}]$, so we can take the derivative with respect to some $\theta(\hat{\mathbf{x}})$. We note that by differentiating the left hand side we get the Gibbs

distribution:

$$\frac{\partial \log Z(\theta)}{\partial \theta(\hat{\mathbf{x}})} = \frac{\exp(\theta(\hat{\mathbf{x}}))}{Z(\theta)}.$$

Differentiating the right hand side is slightly more involved. First, we can differentiate under the integral sign (cf. [51]) so

$$\frac{\partial}{\partial \theta(\hat{\mathbf{x}})} \int_{\mathbb{R}^{|\mathcal{X}|}} \max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}\, d\gamma = \int_{\mathbb{R}^{|\mathcal{X}|}} \frac{\partial}{\partial \theta(\hat{\mathbf{x}})} \max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}\, d\gamma.$$

The (sub)gradient of the max-function is the indicator function (an application of Danskin's Theorem [52]):

$$\frac{\partial}{\partial \theta(\hat{\mathbf{x}})} \max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\} = \mathbf{1}\left(\hat{\mathbf{x}} = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}\right).$$

The corollary then follows by applying the expectation to both sides of the last equation. ■

An alternative proof of the preceding corollary can be given by considering the probability density function $g(t) = G'(t)$ of the Gumbel distribution. This proof consists of two steps. First, the probability that $\hat{\mathbf{x}}$ maximizes $\theta(\mathbf{x}) + \gamma(\mathbf{x})$ is $\int g(t - \theta(\hat{\mathbf{x}})) \prod_{\mathbf{x}\neq\hat{\mathbf{x}}} G(t - \theta(\mathbf{x}))dt$. Second, $g(t - \theta(\hat{\mathbf{x}})) = \exp(\theta(\hat{\mathbf{x}})) \cdot \exp(-(t+c))G(t-\theta(\hat{\mathbf{x}}))$. Thus, the probability that $\hat{\mathbf{x}}$ maximizes $\theta(\mathbf{x})+\gamma(\mathbf{x})$ is proportional to $\exp(\theta(\hat{\mathbf{x}}))$, i.e., it is the Gibbs distribution.

We can also use the random MAP perturbation to estimate the entropy of the Gibbs distribution.

Corollary 2: Let $p(x)$ be the Gibbs distribution, defined in (3) and let $\mathbf{x}^*$ be given by (4). Under the conditions of Theorem 1,

$$H(p) = \mathbb{E}_\gamma \left[\gamma(\mathbf{x}^*)\right].$$

Proof: The proof consists of evaluating the entropy in Equation (3) and using Theorem 1 to replace $\log Z(\theta)$ with $\mathbb{E}_\gamma[\theta(\mathbf{x}^*) + \gamma(\mathbf{x}^*)]$. Formally,

$$\begin{aligned} H(p) &= -\sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x})\theta(\mathbf{x}) + \mathbb{E}_\gamma[\theta(\mathbf{x}^*) + \gamma(\mathbf{x}^*)] \\ &= -\sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x})\theta(\mathbf{x}) + \sum_{\mathbf{x}\in\mathcal{X}} \theta(\mathbf{x})\mathbb{P}_\gamma(\mathbf{x}^* = \mathbf{x}) + \mathbb{E}_\gamma[\gamma(\mathbf{x}^*)] \\ &= \mathbb{E}_\gamma[\gamma(\mathbf{x}^*)], \end{aligned}$$

where in the last line we used Corollary 1, which says $\mathbb{P}_\gamma(\mathbf{x}^* = \mathbf{x}) = p(\mathbf{x})$. ■

A direct proof of the preceding corollary can be given by showing that $\mathbb{E}_\gamma\left[\gamma(\mathbf{x}^*) \cdot 1[\hat{\mathbf{x}} = \mathbf{x}^*]\right] = -p(\hat{\mathbf{x}}) \log p(\hat{\mathbf{x}})$ while the entropy is then attained by summing over all $\hat{\mathbf{x}}$, since $\sum_{\hat{\mathbf{x}}\in\mathcal{X}} 1[\hat{\mathbf{x}} = \mathbf{x}^*] = 1$.

To establish this equality we note that

$$\mathbb{E}_\gamma\left[\gamma(\mathbf{x}^*)\cdot 1[\hat{\mathbf{x}}=\mathbf{x}^*]\right]=\int (t-\theta(\hat{\mathbf{x}}))g(t-\theta(\hat{\mathbf{x}}))\prod_{\mathbf{x}\neq\hat{\mathbf{x}}}G(t-\theta(\mathbf{x}))dt.$$

Using the relation between $g(t)$ and $G(t)$ and the fact that $\prod_{x\in X}G(t-\theta(x))=G(t-\log Z(\theta))$ while changing the integration variable to $\hat{t}=t-\theta(\hat{\mathbf{x}})$ we can rephrase this quantity as $\int t\exp(-(c+t))G(t+\log p(\hat{\mathbf{x}}))dt$. Again by using the relation between $g(t+\log p(\hat{\mathbf{x}}))$ and $G(t+\log p(\hat{\mathbf{x}}))$ we derive that $\mathbb{E}_\gamma\left[\gamma(\mathbf{x}^*)\cdot 1[\hat{\mathbf{x}}=\mathbf{x}^*]\right]=p(\hat{\mathbf{x}})\int tg(t+\log p(\hat{\mathbf{x}}))dt$ while the integral is now the mean of a Gumbel random variable with expected value of $-\log p(\hat{\mathbf{x}})$.

The preceding derivations show that perturbing the potential function $\theta(\mathbf{x})$ and then finding the MAP estimate $\mathbf{x}^*$ of the perturbed Gibbs distribution allows us to perform many core tasks for high-dimensional statistical inference by using i.i.d. Gumbel perturbations. The distribution of $\mathbf{x}^*$ is $p(\mathbf{x})$, its expected maximum value is the log-partition function, and the expected maximizing perturbation is the entropy of $p(\mathbf{x})$. While theoretically appealing, these derivations are computationally intractable when dealing with high-dimensional structures. These derivations involve generating high-dimensional perturbations, namely $|\mathcal{X}|$ random variables in the image of $\gamma(\cdot)$, one for each assignment in $\mathcal{X}=\mathcal{X}_1\times\cdots\mathcal{X}_n$, which grows exponentially with n. The goal of this paper is to apply high-dimensional inference using max-solvers that involve a low-dimensional perturbation term. More specifically, we wish to involve only a linear (in n) number of random variables.

III. Low-dimensional perturbations

We now turn towards making the perturb-max framework more practical. The log-partition function $\log Z(\theta)$ (c.f. Theorem 2) is the key quantity to understand: its gradient is the Gibbs distribution and the entropy is its Fenchel dual. It is well-known that computing $Z(\theta)$ for high-dimensional models is challenging because of the exponential size of $\mathcal{X}$. This complexity carries over to the perturb-max approach to estimating the log-partition function, which also involves generating an exponential number of Gumbel random variables. In this section we show that the log-partition function can be computed using low-dimensional perturbations in a sequence of expected max-value computations. This will give us some insight on performing high dimensional inference using low dimensional perturbations. In what follows we will use the notation $\mathbf{x}_i^j$ to refer to the tuple $(x_i,x_{i+1},\ldots,x_j)$ for $i<j$, with $\mathbf{x}=\mathbf{x}_1^n$.

The partition function has a self-reducible form. That is, we can compute it iteratively while computing

partial partition functions of lower dimensions:

$$Z(\theta) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} \exp(\theta(x_1, x_2, \ldots, x_n)). \tag{10}$$

For example, the partition function is the sum, over x_1, of partial partition functions $\sum_{x_2,\ldots,x_n} \exp(\theta(\mathbf{x}))$. Fixing $x_1, x_2, \ldots, x_i$, the remaining summations are partial partition functions $\sum_{x_{i+1},\ldots,x_n} \exp(\theta(\mathbf{x}))$. With this in mind, we can compute each partial partition function using Theorem 1 but with low-dimensional perturbations for each partial partition.

Theorem 2: Let $\{\gamma_i(x_i)\}_{x_i \in \mathcal{X}_i, i=1,\ldots,n}$, be a collection of independent and identically distributed (i.i.d.) random variables following the Gumbel distribution, defined in Theorem 1. Define $\gamma_i = \{\gamma_i(x_i)\}_{x_i \in \mathcal{X}_i}$. Then

$$\log Z = \mathbb{E}_{\gamma_1} \max_{x_1} \cdots \mathbb{E}_{\gamma_n} \max_{x_n} \left\{ \theta(\mathbf{x}) + \sum_{i=1}^{n} \gamma_i(x_i) \right\}. \tag{11}$$

Proof: The result follows from applying Theorem 1 iteratively. Let $\theta_n(\mathbf{x}_1^n) = \theta(\mathbf{x}_1^n)$ and define

$$\theta_{i-1}(\mathbf{x}_1^{i-1}) = \mathbb{E}_{\gamma_i} \max_{x_i} \{\theta_i(\mathbf{x}_1^i) + \gamma_i(x_i)\} \qquad i = 2, 3, \ldots, n$$

If we think of $\mathbf{x}_1^{i-1}$ as fixed and apply Theorem 1 to $\theta_i(\mathbf{x}_1^{i-1}, x_i)$, we see that from (10),

$$\theta_{i-1}(\mathbf{x}_1^{i-1}) = \log \sum_{x_i} \exp(\theta_i(\mathbf{x}_1^i)).$$

Applying this for $i = n$ to $i = 2$, we obtain (11). ■

The computational complexity of the alternating procedure in (11) is still exponential in n. For example, the innermost iteration $\theta_{n-1}(\mathbf{x}_1^{n-1}) = \mathbb{E}_{\gamma_n} \max_{x_n} \{\theta_n(\mathbf{x}_1^n) + \gamma_n(x_n)\}$ needs to be estimated for every $\mathbf{x}_1^{n-1} = (x_1, x_2, \ldots, x_{n-1})$, which is growing exponentially with n. Thus from computational perspective the alternating formulation in Theorem 2 is just as inefficient as the formulation in Theorem 1. Nevertheless, this is the building block that enables inference in high-dimensional problems using low dimensional perturbations and max-solvers. Specifically, it provides the means for a new sampling algorithm from the Gibbs distribution and bounds on the log-partition and entropy functions.

A. Ideal Sampling

Sampling from the Gibbs distribution is inherently tied to estimating the partition function. If we could compute the partition function exactly, then we could sample from the Gibbs distribution sequentially: for dimension $i = 1, 2, \ldots, n$ sample x_i with probability which is proportional to $\sum_{\mathbf{x}_{i+1}^n} \exp(\theta(\mathbf{x}))$.

Unfortunately, directly computing the partition function is #P-hard. Instead, we construct a family of self-reducible upper bounds which imitate the partition function behavior, namely by bounding the summation over its exponentiations.

Corollary 3: Let $\{\gamma_i(x_i)\}_{x_i \in \mathcal{X}_i, i=1,2,\ldots,n}$ be a collection of i.i.d. random variables, each following the Gumbel distribution with zero mean. Set

$$\phi_j(\mathbf{x}_1^j) = \mathbb{E}_\gamma \left[\max_{x_{j+1}^n} \{\theta(\mathbf{x}) + \sum_{i=j+1}^{n} \gamma_i(x_i)\} \right]. \tag{12}$$

Then for every $j = 1, \ldots, n-1$ and every $\mathbf{x} = \mathbf{x}_1^n$ the following inequality holds:

$$\sum_{x_j} \exp\left(\phi_j(\mathbf{x}_1^j)\right) \leq \exp\left(\phi_{j-1}(\mathbf{x}_1^{j-1})\right). \tag{13}$$

In particular, for $j = n$ we have $\sum_{x_n} \exp(\theta(\mathbf{x}_1^n)) = \exp\left(\phi_{n-1}(\mathbf{x}_1^{n-1})\right)$.

Proof: The result is an application of the perturb-max interpretation of the partition function in Theorem 1. Intuitively, these bounds correspond to moving expectations outside the maximization operations in Theorem 2, each move resulting in a different bound. Formally, the left hand side can be expanded as

$$\mathbb{E}_{\gamma_j} \left[\max_{x_j} \mathbb{E}_{\gamma_{j+1},\ldots,\gamma_n} \left[\max_{\mathbf{x}_{j+1}^n} \left\{ \theta(\mathbf{x}_1^n) + \sum_{i=j}^{n} \gamma_i(x_i) \right\} \right] \right], \tag{14}$$

while the right hand side is attained by alternating the maximization with respect to x_j with the expectation of $\gamma_{j+1}, \ldots, \gamma_n$. The proof then follows by exponentiating both sides. ■

The above corollary is similar in nature to variational approaches that have been extensively developed to efficiently estimate the partition function in large-scale problems. These are often inner-bound methods where a simpler distribution is optimized as an approximation to the posterior in a KL-divergence sense (e.g., mean field) [53]. Variational upper bounds on the other hand are convex, usually derived by replacing the entropy term with a simpler surrogate function and relaxing constraints on sufficient statistics (see, e.g., [54]).

We use these upper bounds for every dimension $i = 1, \ldots, n$ to sample from a probability distribution that follows a summation over exponential functions, with a discrepancy that is described by the upper bound. This is formalized below in Algorithm 1. Note that $\mathbf{x} = (\mathbf{x}_1^{j-1}, x_j, \mathbf{x}_{j+1}^n)$.

This algorithm is forced to restart the entire sample if it samples the "reject" symbol r at any iteration. We say the algorithm accepts if it terminates with an output $\mathbf{x}$. The probability of accepting with particular $\mathbf{x}$ is the product of the probabilities of sampling x_j in round j for $j \in [n]$. Since these upper bounds are self-reducible, i.e., for every dimension i we are using the same quantities that were computed in

Algorithm 1 Unbiased sampling from Gibbs distribution

Require: Potential function $\theta(\mathbf{x})$, MAP solver

Initial step $j = 1$.

while $j < n$ **do**

For all $x \in \mathcal{X}_j$ compute

$$\phi_j(\mathbf{x}_1^{j-1}, x) = \mathbb{E}_\gamma \left[\max_{\mathbf{x}_{j+1}^n} \left\{ \theta(\mathbf{x}_1^{j-1}, x, \mathbf{x}_{j+1}^n) + \sum_{i=j+1}^{n} \gamma_i(x_i) \right\} \right]. \tag{15}$$

Define a distribution on $\mathcal{X}_j \cup \{r\}$:

$$p_j(x) = \frac{\exp\Big(\phi_j(\mathbf{x}_1^{j-1}, x)\Big)}{\exp\Big(\phi_{j-1}(\mathbf{x}_1^{j-1})\Big)}, \qquad x \in \mathcal{X}_j \tag{16}$$

$$p_j(r) = 1 - \sum_{x \in \mathcal{X}_j} p_j(x) \tag{17}$$

Sample x_j from $p_j(\cdot)$.

if $x_j = r$ **then**

Set $j = 1$ to restart sampler.

else $x_j \in \mathcal{X}_j$

Set $j \leftarrow j + 1$.

end if

end while

return $\mathbf{x} = (x_1, x_2, \ldots, x_n)$

the previous dimensions $1, 2, \ldots, i-1$, we are sampling an accepted configuration proportionally to $\exp(\theta(\mathbf{x}))$, the full Gibbs distribution. This is summarized in the following theorem.

Theorem 3: Let $p(\mathbf{x})$ be the Gibbs distribution defined in (1) and let $\{\gamma_i(x_i)\}$ be a collection of i.i.d. random variables following the Gumbel distribution with zero mean given in (8). Then

$$\mathbb{P}\left(\text{Algorithm 1 accepts}\right) = Z(\theta) \Big/ \exp\left(\mathbb{E}_\gamma \left[\max_{\mathbf{x}} \{\theta(\mathbf{x}) + \sum_{i=1}^{n} \gamma_i(x_i)\} \right] \right).$$

Moreover, if Algorithm 1 accepts then it produces a configuration $\mathbf{x} = (x_1, \ldots, x_n)$ according to the Gibbs distribution:

$$\mathbb{P}\left(\text{Algorithm 1 outputs } \mathbf{x} \;\middle|\; \text{Algorithm 1 accepts}\right) = \frac{\exp(\theta(\mathbf{x}))}{Z(\theta)}.$$

Proof: Set $\theta_j(\mathbf{x}_1^j)$ as in Corollary 3. The probability of sampling a configuration $\mathbf{x} = (x_1, \ldots, x_n)$ without rejecting is

$$\prod_{j=1}^{n} \frac{\exp\left(\phi_j(\mathbf{x}_1^j)\right)}{\exp\left(\phi_{j-1}(\mathbf{x}_1^{j-1})\right)} = \frac{\exp(\theta(\mathbf{x}))}{\exp\left(\mathbb{E}_\gamma\left[\max_{\mathbf{x}}\left\{\theta(\mathbf{x}) + \sum_{i=1}^n \gamma_i(x_i)\right\}\right]\right)}.$$

The probability of sampling without rejecting is thus the sum of this probability over all configurations, i.e.,

$$\mathbb{P}\left(\text{Algorithm 1 accepts}\right) = Z(\theta) \big/ \exp\left(\mathbb{E}_\gamma\left[\max_{\mathbf{x}}\left\{\theta(\mathbf{x}) + \sum_{i=1}^n \gamma_i(x_i)\right\}\right]\right).$$

Therefore conditioned on acceptance, the output configuration is produced according to the Gibbs distribution. ■

Since acceptance/rejection follows the geometric distribution, the sampling procedure rejects k times with probability $\left(1 - \mathbb{P}\left(\text{Algorithm 1 accepts}\right)\right)^k$. The running time of our Gibbs sampler is determined by the average number of rejections $1/\mathbb{P}(\text{Algorithm 1 accepts})$. The exponent of this error event is:

$$\log \frac{1}{\mathbb{P}\left(\text{Algorithm 1 accepts}\right)} = \mathbb{E}_\gamma\left[\max_{\mathbf{x}}\left\{\theta(\mathbf{x}) + \sum_{i=1}^n \gamma_i(x_i)\right\}\right] - \log Z(\theta).$$

To be able to estimate the number of steps the sampling algorithm requires, we construct an efficiently computable lower bound to the log-partition function that is based on perturb-max values.

B. Approximate Inference and Lower Bounds to the Partition Function

To be able to estimate the number of steps the sampling Algorithm 1 requires, we construct an efficiently computable lower bound to the log-partition function, that is based on perturb-max values. Let $\{M_i : i = 1, 2, \ldots, n\}$ be a collection of positive integers. For each $i = 1, 2, \ldots, n$ let $\tilde{\mathbf{x}}_i = \{x_{i,k_i} : k_i = 1, 2, \ldots, M_i\}$ be a tuple of M_i elements of $\mathcal{X}_i$. We define an extended potential function over a configuration space of $\sum_{i=1}^n M_i$ variables $\tilde{\mathbf{x}} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_n)$:

$$\hat{\theta}(\tilde{\mathbf{x}}) = \frac{1}{\prod_{i=1}^n M_i} \sum_{k_1=1}^{M_1} \sum_{k_2=1}^{M_2} \cdots \sum_{k_n=1}^{M_n} \theta(x_{1,k_1}, x_{2,k_2}, \ldots, x_{n,k_n}). \tag{18}$$

Now consider a collection of i.i.d. zero-mean Gumbel random variables $\{\tilde{\gamma}_{i,k_i}(x_{i,k_i})\}_{i=1,2,\ldots,n,k_i=1,2,\ldots,M_i}$ with distribution (8). Define the following perturbation for the extended model:

$$\tilde{\gamma}_i(\tilde{\mathbf{x}}_i) = \frac{1}{M_i} \sum_{k_i=1}^{M_i} \tilde{\gamma}_{i,k_i}(x_{i,k_i}). \tag{19}$$

Corollary 4: Let $\theta(\mathbf{x})$ be a potential function over $\mathbf{x} = (x_1, \ldots, x_n)$ and $\log Z$ be the log partition function for the corresponding Gibbs distribution. Then for any $\epsilon > 0$ we have

$$\mathbb{P}_{\tilde{\gamma}} \left(\log Z \geq \max_{\tilde{\mathbf{x}}} \left\{ \hat{\theta}(\tilde{\mathbf{x}}) + \sum_{i=1}^{n} \tilde{\gamma}_i(\tilde{\mathbf{x}}_i) \right\} - \epsilon n \right) \geq 1 - \sum_{i=1}^{n} \frac{\pi^2 \prod_{j=2}^{i} |\mathcal{X}_{j-1}|}{6 M_i \epsilon^2}. \tag{20}$$

Proof: The proof consists of three steps:

- developing a measure concentration analysis for Theorem 1, which states that a single max-evaluation is enough to lower bound the expected max-value with high probability;
- using the self-reducibility of the partition function in Theorem 2 to show the partition function can be computed by iteratively applying low-dimensional perturbations;
- proving that these lower dimensional partition functions can be lower bounded uniformly (i.e., all at once) with a single measure concentration statement.

We first provide a measure concentration analysis of Theorem 1. Specifically, we estimate the deviation of the random variable $F = \max_{\mathbf{x} \in \mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}$ from its expected value using Chebyshev's inequality. For this purpose we recall Theorem 1 which states that F is Gumbel-distributed and therefore its variance is $\pi^2/6$. Chebyshev's inequality then asserts that

$$\mathbb{P}_{\gamma} \left(|F - \mathbb{E}_{\gamma}[F]| \geq \epsilon \right) \leq \pi^2 / 6\epsilon^2. \tag{21}$$

Since we want this statement to hold with high probability for small epsilon we reduce the variance of the random variable while not changing its expectation by taking a sampled average of i.i.d. perturb-max values: Let $F(\gamma) = \max_{\mathbf{x}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}$. Suppose we sample M i.i.d. random variables $\gamma_1, \gamma_2, \ldots, \gamma_M$ with the same distribution as γ and generate the i.i.d. Gumbel-distributed values $F_j \stackrel{\Delta}{=} F(\gamma_j)$. We call $\gamma_1, \gamma_2, \ldots, \gamma_M$ "copies" of γ. Since[1] $\mathbb{E}_{\gamma}[F(\gamma)] = \log Z$, we can apply Chebyshev's inequality to the $\frac{1}{M} \sum_{i=1}^{M} F_j - \log Z$ to get

$$\mathbb{P} \left(\left| \frac{1}{M} \sum_{i=1}^{M} F_j - \log Z \right| \geq \epsilon \right) \leq \frac{\pi^2}{6M\epsilon^2}. \tag{22}$$

Using the explicit perturb-max notation and considering only the lower-side of the measure concentration bound, this shows that with probability at least $1 - \frac{\pi^2}{6M\epsilon^2}$ we have

$$\log Z \geq \frac{1}{M} \sum_{j=1}^{M} \max_{\mathbf{x} \in \mathcal{X}} \{\theta(\mathbf{x}) + \gamma_j(\mathbf{x})\} - \epsilon. \tag{23}$$

[1]Whenever θ is clear from the context we use the shorthand Z for $Z(\theta)$.

To complete the first step, we wish to compute the summation over M-maximum values using a single maximization. For this we form an extended model on $\mathcal{X}^M$ containing variables $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_M \in \mathcal{X}$ and note that

$$\sum_{j=1}^{M} \max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma_j(\mathbf{x})\} = \max_{\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2,\ldots,\tilde{\mathbf{x}}_M} \sum_{j=1}^{M} (\theta(\tilde{\mathbf{x}}_j) + \gamma_j(\tilde{\mathbf{x}}_j)). \tag{24}$$

For the remainder we use an argument by induction on n, the number of variables. Consider first the case $n = 2$ so that $\theta(\mathbf{x}) = \theta_{1,2}(x_1, x_2)$. The self-reducibility as described in Theorem 1 states that

$$\log Z = \log \left(\sum_{x_1} \exp \left[\log \left(\sum_{x_2} \exp(\theta_{1,2}(x_1, x_2)) \right) \right] \right). \tag{25}$$

As in the proof of Theorem 1, define $\theta_1(x_1) = \log(\sum_{x_2} \exp(\theta_{1,2}(x_1, x_2))$. Thus we have $\log Z = \log \left(\sum_{x_1} \exp(\theta_1(x_1)) \right)$, which is a partition function for a single-variable model.

We wish to uniformly approximate $\theta_1(x_1)$ over all $x_1 \in \mathcal{X}_1$. Fix $x_1 = a$ for some $a \in \mathcal{X}_1$ and consider the single-variable model $\theta_{1,2}(a, x_2)$ over x_2 which has $\theta_1(a)$ as its log-partition function. Then from Theorem 1, we have $\theta_1(a) = \mathbb{E}_{\gamma_2} [\max_{x_2} \{\theta(a, x_2) + \gamma_2(x_2)\}]$. Applying Chebyshev's inequality in (22) to M_2 "copies" of γ_2, we get

$$\mathbb{P} \left(\left| \frac{1}{M_2} \sum_{j=1}^{M_2} \max_{x_2} \{\theta(a, x_2) + \gamma_{2,j}(x_2)]\} - \theta_1(a) \right| \geq \epsilon \right) \leq \frac{\pi^2}{6 M_2 \epsilon^2}.$$

Taking a union bound over $a \in \mathcal{X}_1$ we have

$$\mathbb{P} \left(\left| \frac{1}{M_2} \sum_{j=1}^{M_2} \max_{x_2} \{\theta(x_1, x_2) + \gamma_{2,j}(x_2)]\} - \theta_1(x_1) \right| \leq \epsilon \ \ \forall x_1 \in \mathcal{X}_1 \right) \leq 1 - |\mathcal{X}_1| \frac{\pi^2}{6 M_2 \epsilon^2}.$$

This implies the following one-sided inequality with probability at least $1 - |\mathcal{X}_1| \frac{\pi^2}{6 M_2 \epsilon^2}$ uniformly over $x_1 \in \mathcal{X}_1$:

$$\theta_1(x_1) \geq \frac{1}{M_2} \sum_{j=1}^{M_2} \max_{x_2} \{\theta(x_1, x_2) + \gamma_{2,j}(x_2)]\} - \epsilon. \tag{26}$$

Now note that the overall log-partition function for the model $\theta(\mathbf{x}) = \theta_{1,2}(x_1, x_2)$ is a log-partition function for a single variable model with potential $\theta_1(x_1)$, so $\log Z = \log(\sum_{x_1} \exp(\theta_1(x_1)))$. Again using Theorem 1, we have $\log Z = \mathbb{E}_{\gamma_1} [\max_{x_1} \{\theta_1(x_1) + \gamma_1(x_1)\}]$, so we can apply Chebyshev's inequality to M_1 "copies" of γ_1 to get that with probability at least $1 - \frac{\pi^2}{6 M_1 \epsilon^2}$:

$$\log Z \geq \frac{1}{M_1} \sum_{k=1}^{M_1} \max_{x_1} \{\theta_1(x_1) + \gamma_{1,k}(x_1)\} - \epsilon. \tag{27}$$

Plugging in (26) into (27), we get that with probability at least $1 - \frac{\pi^2}{6M_1\epsilon^2} - |\mathcal{X}_1|\frac{\pi^2}{6M_2\epsilon^2}$:

$$\log Z \geq \frac{1}{M_1}\sum_{k=1}^{M_1}\max_{x_1}\left\{\left(\frac{1}{M_2}\sum_{j=1}^{M_2}\max_{x_2}\{\theta(x_1,x_2)+\gamma_{2,j}(x_2)\}\right)+\gamma_{1,k}(x_1)\right\} - 2\epsilon. \tag{28}$$

Now we pull the maximization outside the sum by introducing i.i.d. "copies" of the variables again: this time we have M_1 copies $\tilde{\mathbf{x}}_1$ and M_1M_2 copies $\tilde{\mathbf{x}}_2$ for $\tilde{\mathbf{x}}_2$ as in (24). Now,

$$\begin{aligned}
&\frac{1}{M_1}\sum_{k=1}^{M_1}\max_{x_1}\left\{\left(\frac{1}{M_2}\sum_{j=1}^{M_2}\max_{x_2}\{\theta(x_1,x_2)+\gamma_{2,j}(x_2)\}\right)+\gamma_{1,k}(x_1)\right\}\\
&= \frac{1}{M_1}\sum_{k=1}^{M_1}\max_{x_1}\left\{\left(\max_{\tilde{x}_{2,1},\ldots,\tilde{x}_{2,M_2}}\frac{1}{M_2}\sum_{j=1}^{M_2}\theta(x_1,\tilde{x}_{2,j})+\gamma_{2,j}(\tilde{x}_{2,j})\right)+\gamma_{1,k}(x_1)\right\}\\
&= \max_{\tilde{x}_{1,1},\ldots,\tilde{x}_{1,M_1}}\max_{\tilde{x}_{2,1},\ldots,\tilde{x}_{2,M_2}}\frac{1}{M_1M_2}\sum_{k=1}^{M_1}\sum_{j=1}^{M_2}\theta(\tilde{x}_{1,k},\tilde{x}_{2,j})+\gamma_{2,j}(\tilde{x}_{1,k},\tilde{x}_{2,j})+\gamma_{1,k}(\tilde{x}_{1,k}).
\end{aligned}$$

Note that in this bound we have to generate $|\mathcal{X}_1||\mathcal{X}_2|$ variables $\gamma_{2,j}(x_{1,k},x_{2,j})$, which will become inefficient as we add more variables. We can get an efficiently computable lower bound on this quantity by generating a smaller set of variables: we use the same perturbation realization $\gamma_{2,j}(x_{2,j})$ for every value of $x_{1,k}$. Thus we have the lower bound

$$\log Z \geq \max_{\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2}\frac{1}{M_1M_2}\sum_{k=1}^{M_1}\sum_{j=1}^{M_2}\left(\theta(x_{1,k},\tilde{x}_{2,j})+\gamma_{2,j}(\tilde{x}_{2,j})+\gamma_{1,k}(\tilde{x}_{1,k})\right) - 2\epsilon$$

with probability at least $1 - \frac{\pi^2}{6M_1\epsilon^2} - |\mathcal{X}_1|\frac{\pi^2}{6M_2\epsilon^2}$. Here we have abused notation slightly and used $\tilde{\mathbf{x}}_1 = \{\tilde{x}_{1,1},\tilde{x}_{1,2},\ldots,\tilde{x}_{1,M_1}\}$ and $\tilde{\mathbf{x}}_2 = \{\tilde{x}_{2,1},\tilde{x}_{2,2}\ldots,\tilde{x}_{2,M_2}\}$.

Now suppose the result holds for models on $n-1$ variables and consider the model $\theta(x_1,x_2,\ldots,x_n)$ on n variables. Consider the 2-variable model $\theta(x_1,\mathbf{x}_2^n)$ and define

$$\theta_1(x_1) = \log\left(\sum_{x_2^n}\exp(\theta(x_1,\mathbf{x}_2^n))\right). \tag{29}$$

From the analysis of the 2-variable case, as in (27), the following lower bound holds with probability at least $1 - \frac{\pi^2}{6M_1\epsilon^2}$:

$$\log Z \geq \frac{1}{M_1}\sum_{k_1=1}^{M_1}\max_{x_1}\{\theta_1(x_1)+\gamma_{1,k_1}(x_1)\} - \epsilon. \tag{30}$$

Now note that for each value of x_1, the function $\theta_1(x_1)$ is a log-partition function on the $n-1$ variables x_2^n. Applying the induction hypothesis to $\theta_1(x_1)$, we have with probability at least

$$1 - \frac{\pi^2}{6M_2\epsilon^2} - |\mathcal{X}_2|\frac{\pi^2}{6M_3\epsilon^2} - |\mathcal{X}_2||\mathcal{X}_3|\frac{\pi^2}{6M_4\epsilon^2} - \cdots - \prod_{j=2}^{n-1}|\mathcal{X}_j|\frac{\pi^2}{6M_n\epsilon^2}, \tag{31}$$

the following lower bound holds:

$$\theta_1(x_1) \geq \max_{\tilde{\mathbf{x}}_2^n} \left\{ \hat{\theta}(x_1, \tilde{\mathbf{x}}_2^n) + \sum_{i=2}^{n} \tilde{\gamma}_i(\tilde{\mathbf{x}}_i) \right\} - \epsilon(n-1). \tag{32}$$

Taking a union bound over all x_1, with probability at least

$$1 - \sum_{i=1}^{n} \left(\prod_{j=2}^{i} |\mathcal{X}_{j-1}| \right) \frac{\pi^2}{6M_n\epsilon^2} \tag{33}$$

we have

$$\begin{aligned} \log Z &\geq \frac{1}{M_1} \sum_{k_1=1}^{M_1} \max_{x_1} \left\{ \max_{\tilde{\mathbf{x}}_2^n} \left\{ \hat{\theta}(x_1, \tilde{\mathbf{x}}_2^n) + \sum_{i=2}^{n} \tilde{\gamma}_i(\tilde{\mathbf{x}}_i) \right\} + \gamma_{1,k_1}(x_1) \right\} - \epsilon n \\ &\geq \max_{\tilde{\mathbf{x}}} \hat{\theta}(\tilde{\mathbf{x}}) + \sum_{i=1}^{n} \tilde{\gamma}_i(\tilde{\mathbf{x}}_i) - \epsilon n, \end{aligned}$$

as desired. ■

The key to understand the efficiency of this lower bound is in analyzing the structure of the potential functions $\hat{\theta}(\tilde{\mathbf{x}})$ and $\theta(\mathbf{x})$. Although $\hat{\theta}(\tilde{\mathbf{x}})$ seems to consider exponentially many configurations, its order is the same as the original potential function $\theta(\mathbf{x})$. Particularly, if $\theta(\mathbf{x})$ is the sum of local and pairwise potential functions (as happens for the Ising model) then $\hat{\theta}(\tilde{\mathbf{x}})$ is also the sum of local and pairwise potential functions. Therefore, whenever the original model can be maximized efficiently, e.g., for super-modular functions, the inflated model $\hat{\theta}(\tilde{\mathbf{x}})$ can also be optimized efficiently. Moreover, while the theory requires M_i to be exponentially large (as a function of n), it turns out that in practice M_i may be very small to generate tight bounds (see Section V). Theoretically tighter bounds can be derived by our measure concentration results in Section IV but they do not fully capture the tightness of this lower bound.

C. Entropy bounds

We now show how to use perturb-max values to bound the entropy of high-dimensional models. Estimating the entropy is an important building block in many machine learning applications. Corollary 2 applies the interpretation of Gibbs distribution as a perturb-max model (see Corollary 1) in order to define the entropy of Gibbs distributions using the expected value of the maximal perturbation. Unfortunately, this procedure requires exponentially many independent perturbations $\gamma(\mathbf{x})$, for every $\mathbf{x} \in \mathcal{X}$.

We again use our low-dimensional perturbations to upper bound the entropy of perturb-max models. We need to extend our definition of perturb-max models as follows. Let $\mathcal{A}$ be a collection of subsets of

$\{1,2,\ldots,n\}$ such that $\bigcup_{\alpha\in\mathcal{A}}=\{1,2,\ldots,n\}$. For each $\alpha\in\mathcal{A}$ generate a Gumbel perturbation $\gamma_\alpha(\mathbf{x}_\alpha)$ where $\mathbf{x}_\alpha=(x_i)_{i\in\alpha}$. We define the perturb-max models as

$$p(\hat{\mathbf{x}};\theta)=\mathbb{P}_\gamma\left(\hat{\mathbf{x}}=\operatorname*{argmax}_{\mathbf{x}}\left\{\theta(\mathbf{x})+\sum_{\alpha\in\mathcal{A}}\gamma_\alpha(\mathbf{x}_\alpha)\right\}\right). \tag{34}$$

Our upper bound uses the duality between entropy and the log-partition function [34] and then upper bounds the log-partition function with perturb-max operations.

Upper bounds for the log-partition function using random perturbations can be derived from the refined upper bounds in Corollary 3. However, it is simpler to provide upper bounds that rely on Theorem 2. These bounds correspond to moving expectations outside the maximization operations.

Lemma 1: Let $\theta(\mathbf{x})$ be a potential function over $\mathbf{x}=(x_1,\ldots,x_n)$, and $\{\gamma_i(x_i)\}_{x_i\in\mathcal{X}_i,i=1,\ldots,n}$ be a collection of independent and identically distributed (i.i.d.) random variables following the Gumbel distribution. Then

$$\log Z(\theta)\leq\mathbb{E}_\gamma\left[\max_{\mathbf{x}=(x_1,x_2,\ldots,x_n)}\left\{\theta(\mathbf{x})+\sum_{i=1}^n\gamma_i(x_i)\right\}\right]. \tag{35}$$

Proof: The lemma follows from Theorem 2 that represents (11) as the log-partition as a sequence of alternating expectations and maximizations, namely

$$\log Z(\theta)=\mathbb{E}_{\gamma_1}\max_{x_1}\cdots\mathbb{E}_{\gamma_n}\max_{x_n}\left\{\theta(\mathbf{x})+\sum_{i=1}^n\gamma_i(x_i)\right\}. \tag{36}$$

The upper bound is attained from the right hand side of the above equation by Jensen's inequality (or equivalently, by moving all the expectations in front of the maximizations, yielding the following:

$$\mathbb{E}_{\gamma_1}\max_{x_1}\cdots\mathbb{E}_{\gamma_n}\max_{x_n}\left\{\theta(\mathbf{x})+\sum_{i=1}^n\gamma_i(x_i)\right\}\leq\mathbb{E}_{\gamma_1}\cdots\mathbb{E}_{\gamma_n}\max_{x_1}\cdots\max_{x_n}\left\{\theta(\mathbf{x})+\sum_{i=1}^n\gamma_i(x_i)\right\}. \tag{37}$$

■

In this case the bound is an average of MAP values corresponding to models with only single node perturbations $\gamma_i(x_i)$, for every $i=1,\ldots,n$ and $x_i\in\mathcal{X}_i$. If the maximization over $\theta(\mathbf{x})$ is feasible (e.g., due to supermodularity), it will typically be feasible after such perturbations as well. We generalize this basic result further below.

Corollary 5: Consider a family of subsets $\alpha\in\mathcal{A}$ such that $\bigcup_{\alpha\in\mathcal{A}}\alpha=\{1,\ldots,n\}$, and let $\mathbf{x}_\alpha=\{x_i:i\in\alpha\}$. Assume that the random variables $\gamma_\alpha(\mathbf{x}_\alpha)$ are i.i.d. according to the Gumbel distribution, for every $\alpha,\mathbf{x}_\alpha$. Then

$$\log Z(\theta)\leq\mathbb{E}_\gamma\left[\max_{\mathbf{x}}\left\{\theta(\mathbf{x})+\sum_{\alpha\in\mathcal{A}}\gamma_\alpha(\mathbf{x}_\alpha)\right\}\right].$$

Proof: If the subsets α are disjoint the upper bound is an application of Lemma 1 as follows: we consider the potential function $\theta(\mathbf{x})$ over the disjoint subsets of variables $\mathbf{x} = (\mathbf{x}_\alpha)_{\alpha \in \mathcal{A}}$ as well as the i.i.d. Gumbel random variables $\gamma_\alpha(\mathbf{x}_\alpha)$. Applying Lemma 1 yields the following upper bound:

$$\log Z(\theta) \leq \mathbb{E}_\gamma \left[\max_{\mathbf{x}=(\mathbf{x}_\alpha)_{\alpha \in \mathcal{A}}} \left\{ \theta(\mathbf{x}) + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}_\alpha) \right\} \right]. \tag{38}$$

In the general case, $\alpha, \beta \in \mathcal{A}$ may overlap. To follow the same argument, we lift the n-dimensional assignment $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ to an higher-dimensional assignment $a(\mathbf{x}) = (\mathbf{x}_\alpha)_{\alpha \in \mathcal{A}}$ which creates an independent perturbation for each $\alpha \in \mathcal{A}$. To complete the proof, we also construct a potential function $\theta'(\mathbf{x}')$ such that

$$\theta'(\mathbf{x}') = \begin{cases} \theta(\mathbf{x}) & \text{if } a(\mathbf{x}) = \mathbf{x}' \\ -\infty & \text{otherwise.} \end{cases} \tag{39}$$

Thus, $\log Z(\theta) = \log Z(\theta') = \sum_{\mathbf{x}'} \exp(\theta'(\mathbf{x}'))$ since inconsistent assignments (i.e., $\mathbf{x}'$ such that $a(\mathbf{x}) \neq \mathbf{x}'$ for any $\mathbf{x}$) receive zero weight. Moreover,

$$\max_{\mathbf{x}'} \left\{ \theta'(\mathbf{x}') + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}'_\alpha) \right\} = \max_{\mathbf{x}} \left\{ \theta(\mathbf{x}) + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}_\alpha) \right\}$$

for each realization of the perturbation. This equality holds after expectation over γ as well. Now, given that the perturbations are independent for each lifted coordinate, the basic result in (35) guarantees that

$$\log Z(\theta') \leq \mathbb{E}_\gamma \left[\max_{\mathbf{x}'} \left\{ \theta'(\mathbf{x}') + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}'_\alpha) \right\} \right],$$

from which the result follows since $\log Z(\theta) = \log Z(\theta')$ ■

Establishing bounds on the log-partition function allows us to derive bounds on the entropy. For this we use the conjugate duality between the (negative) entropy and the log-partition function [34]. The entropy bound then follows from the log-partition bound.

Theorem 4: Let $p(\mathbf{x}; \theta)$ be a perturb-max probability distribution in (34) and $\mathcal{A}$ be a collection of subsets of $\{1, 2, \ldots, n\}$. Let $\mathbf{x}^\gamma$ be the optimal perturb-max assignment using low dimensional perturbations:

$$\mathbf{x}^\gamma = \operatorname*{argmax}_{\mathbf{x}} \left\{ \theta(\mathbf{x}) + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}_\alpha) \right\}. \tag{40}$$

Then under the conditions of Corollary 5, we have the following upper bound:

$$H(p) \leq \mathbb{E}_\gamma \left[\sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}^\gamma_\alpha) \right].$$

Proof: We use the characterization of the log-partition function as the conjugate dual of the (negative) entropy function [34]:

$$H(p) = \min_{\hat{\theta}} \left\{ \log Z(\hat{\theta}) - \sum_{\mathbf{x}} p(\mathbf{x};\theta)\hat{\theta}(\mathbf{x}) \right\}.$$

The minimum is over all potential functions on $\mathcal{X}$. For a fixed score function $\hat{\theta}(\mathbf{x})$, let $W(\hat{\theta})$ be the expected value of the low-dimensional perturbation:

$$W(\hat{\theta}) = \mathbb{E}_\gamma \left[\max_{\mathbf{x}} \left\{ \hat{\theta}(\mathbf{x}) + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}_\alpha) \right\} \right].$$

Corollary 5 asserts that $\log Z(\hat{\theta}) \leq W(\hat{\theta})$. Thus we can upper bound $H(p)$ by replacing $\log Z(\hat{\theta})$ with $W(\hat{\theta})$ in the duality relation:

$$H(p) \leq \min_{\hat{\theta}} \left\{ W(\hat{\theta}) - \sum_{\mathbf{x}} p(\mathbf{x};\theta)\hat{\theta}(\mathbf{x}) \right\}.$$

The infimum of the right hand side is attained whenever the gradient vanishes, i.e., whenever $\nabla W(\hat{\theta}) = p(\mathbf{x};\theta)$. To compute $\nabla W(\hat{\theta})$ we differentiate under the integral sign:

$$\nabla W(\hat{\theta}) = \mathbb{E}_\gamma \left[\nabla \max_{\mathbf{x}} \left\{ \hat{\theta}(\mathbf{x}) + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}_\alpha) \right\} \right].$$

Since the (sub)gradient of the maximum-function is the indicator function, we deduce that $\nabla W(\hat{\theta})$ is the expected value of the events of $\mathbf{x}^\gamma$. Consequently, $\nabla W(\hat{\theta})$ is the vector of the probabilities of all these events, namely, the probability distribution $p(\mathbf{x};\hat{\theta})$. Since the derivatives of $W(\hat{\theta})$ are perturb-max models, and so is $p(\mathbf{x};\theta)$, then the the infimum is attained for $\hat{\theta} = \theta$. Therefore, recalling that $\mathbf{x}^\gamma$ has distribution $p(\mathbf{x};\theta)$ in (34):

$$\begin{aligned}
\min_{\hat{\theta}} \left\{ W(\hat{\theta}) - \sum_{\mathbf{x}} p(\mathbf{x};\theta)\theta(\mathbf{x}) \right\} &= W(\theta) - \sum_{\mathbf{x}} p(\mathbf{x};\theta)\theta(\mathbf{x}). \\
&= \mathbb{E}_\gamma \left[\max_{\mathbf{x}} \left\{ \theta(\mathbf{x}) + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}_\alpha) \right\} \right] - \mathbb{E}_\gamma \left[\theta(\mathbf{x}^\gamma) \right] \\
&= \mathbb{E}_\gamma \left[\hat{\theta}(\mathbf{x}^\gamma) + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}^\gamma_\alpha) \right] - \mathbb{E}_\gamma \left[\theta(\mathbf{x}^\gamma) \right] \\
&= \mathbb{E}_\gamma \left[\sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}^\gamma_\alpha) \right],
\end{aligned}$$

from which the result follows. ■

This entropy bound motivates the use of perturb-max posterior models. These models are appealing as they are uniquely built around prediction and as such they inherently have an efficient unbiased

sampler. The computation of this entropy bound relies on MAP solvers. Thus, computing these bounds is significantly faster than computing the entropy itself, whose computational complexity is generally exponential in n.

Using the linearity of expectation we may alternate summation and expectation. For simplicity, assume only local perturbations, i.e., $\gamma_i(x_i)$ for every dimension $i = 1, \ldots, n$. Then the preceding theorem bounds the entropy by summing the expected change of MAP perturbations $H(p) \leq \sum_i \mathbb{E}_\gamma[\gamma_i(x_i^\gamma)]$. This bound resembles to the independence bound for the entropy $H(p) \leq \sum_i H(p_i)$, where $p_i(x_i) = \sum_{\mathbf{x} \setminus x_i} p(\mathbf{x})$ are the marginal probabilities [55]. The independence bound is tight whenever the joint probability $p(\mathbf{x})$ is composed of independent systems, i.e., $p(\mathbf{x}) = \prod_i p_i(x_i)$. In the following we show that the same holds for perturbation bounds.

Corollary 6: Consider the setting in Theorem 4 with $\mathbf{x}^\gamma$ given by (40) and the independent probability distribution $p(\mathbf{x}) = \prod_i p_i(x_i)$. Let $\{\gamma_i(x_i)\}_{x_i \in \mathcal{X}_i, i=1,2,\ldots,n}$ be a collection of i.i.d. random variables, each following the Gumbel distribution with zero mean. Then there exists $\theta(\mathbf{x})$ for which

$$H(p) = \mathbb{E}_\gamma \left[\sum_{i=1}^n \gamma_i(x_i^\gamma) \right],$$

where

$$\mathbf{x}^\gamma = \operatorname*{argmax}_{\mathbf{x}} \left\{ \theta(\mathbf{x}) + \sum_{i=1}^n \gamma_i(x_i) \right\}.$$

Proof: Since the system is independent, $H(p) = \sum_i H(p_i)$. We first show that there exists $\theta_i(x_i)$ in each dimension for which $H(p_i) = \mathbb{E}_{\gamma_i}\left[\gamma_i(x_i^{\gamma_i})\right]$ and then complete the proof by constructing $\theta(\mathbf{x})$.

Set $\theta_i(x_i) = \log p_i(x_i)$. Since $\{\gamma_i(x_i)\}_{x_i \in \mathcal{X}_i}$ are independent, we may apply Corollary 2 to the $i-$th dimension and obtain $H(p_i) = \mathbb{E}_\gamma\left[\gamma_i(x_i^{\gamma_i})\right]$, where $x^{\gamma_i} = \operatorname{argmax}_{\mathbf{x}_i} \{\theta_i(\mathbf{x}_i) + \gamma_i(x_i)\}$.

To complete the proof, we set $\theta(\mathbf{x}) = \sum_i \theta_i(x_i)$. Since the system is independent, there holds $\mathbf{x}_i^\gamma = \mathbf{x}_i^{\gamma_i} \stackrel{def}{=} \operatorname{argmax}_{\mathbf{x}_i} \{\theta_i(\mathbf{x}_i) + \gamma_i(x_i)\}$. Therefore, $H(p) = \sum_i H(p_i) = \sum_i \mathbb{E}_{\gamma_i}\left[\gamma_i(x_i^{\gamma_i})\right] = \sum_i \mathbb{E}_\gamma\left[\gamma_i(x_i^\gamma)\right] = \mathbb{E}_\gamma\left[\sum_i \gamma_i(x_i^\gamma)\right]$.

■

There are two special cases for independent systems. First, the zero-one probability model, for which $p(\mathbf{x}) = 0$ except for a single configuration $p(\hat{\mathbf{x}}) = 1$. The entropy of such a probability distribution is 0 since the distribution is deterministic. In this case, the perturb-max entropy bound assigns $\mathbf{x}^\gamma = \hat{\mathbf{x}}$ for all random functions $\gamma = (\gamma_i(x_i))_{i,x_i}$. Since these random variables have zero mean, it follows that $\mathbb{E}_\gamma\left[\sum_i \gamma_i(\hat{x}_i)\right] = 0$. Another important case is for the uniform distribution with $p(\mathbf{x}) = 1/|\mathcal{X}|$ for every $\mathbf{x} \in \mathcal{X}$. The entropy of such a probability distribution is $\log |\mathcal{X}|$, as it has maximal uncertainty. Since our

entropy bounds equal the entropy for minimal uncertainty and maximal uncertainty cases, this suggests that the perturb-max bound can be used as an alternative uncertainty measure.

Corollary 7: Consider the setting of Theorem 4 with $\mathbf{x}^\gamma$ given by (40). Define the function $U(p)$ by

$$U(p) = \mathbb{E}_\gamma \left[\sum_{\alpha \in \mathcal{A}} \gamma_\alpha(\mathbf{x}_\alpha^\gamma) \right]. \tag{41}$$

Then $U(p)$ is non-negative and attains its minimal value for the deterministic distributions and its maximal value for the uniform distribution.

Proof: As argued above, $U(p)$ is 0 for deterministic p. Non-negativity follows from the requirement that the perturbation are zero-mean random variables: since $\sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha^\gamma) \geq \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha)$ for x, then $U(p) = \mathbb{E}_\gamma \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha^\gamma) \geq \mathbb{E}_\gamma \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha) = 0$. Lastly, we must show that the uniform distribution p_{uni} maximizes $U(\cdot)$, namely $U(p_{\mathsf{uni}}) \geq U(\cdot)$. The potential function for the uniform distribution is constant for all $\mathbf{x} \in \mathcal{X}$. This means that $U(p_{\mathsf{uni}}) = \mathbb{E}_\gamma \max_x \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha)$. On the other hand $U(\cdot)$ correspond to a potential function $\theta(\mathbf{x})$ and its corresponding $\mathbf{x}^\gamma$. Furthermore, we have $\max_{\mathbf{x}} \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha) \geq \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha^\gamma)$, and taking expectations on both sides shows $U(p_{\mathsf{uni}}) = \mathbb{E}_\gamma \max_x \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha) \geq \mathbb{E}_\gamma \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha^\gamma)$ for any other $\theta(\cdot)$ and its corresponding $\mathbf{x}^\gamma$. ■

The preceding result implies we can use $U(p)$ as a surrogate uncertainty measure instead of the entropy. Using efficiently computable uncertainty measures allows us to extend the applications of perturb-max models to Bayesian active learning [4]. The advantage of using the perturb-max uncertainty measure over the entropy function is that it does not require MCMC sampling procedures. Therefore, our approach fits well with contemporary techniques for using high-dimensional models that are popular in machine learning applications such as computer vision. Moreover, our perturb-max uncertainty measure is an upper bound on the entropy, so minimizing the upper bound can be a reasonable heuristic approach to reducing entropy.

IV. Measure concentration for log-concave perturbations

High dimensional inference with random perturbations relies on expected values of MAP predictions. In Section III-A we presented a Gibbs distribution sampler that involves calculating the expected value of randomized max-solvers $F(\gamma) = \max_{\mathbf{x}}\{\theta(\mathbf{x}) + \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha)\}$. In Section III-C the expected value of the perturbation themselves gave an upper bound on the entropy of the perturb-max model $F(\gamma) = \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha^\gamma)$. Practical application of this theory requires estimating these expectations; the simplest way to do this is by taking a sample average. We therefore turn to bounding the number of samples by proving concentration of measure results for our random perturbation framework. The key technical challenge

comes from the fact that the perturbations $\gamma_\alpha(\mathbf{x}_\alpha)$ are Gumbel random variables, which have support on the entire real line. Thus, many standard approaches for bounded random variables, such as McDiarmid's inequality, do not apply.

Nevertheless, the Gumbel distribution decays exponentially, thus one can expect that the distance between the perturbed MAP prediction and its expected value to decay exponentially as well. Our measure concentration bounds (in Section IV-E) show this; we bound the deviation of a general function $F(\gamma)$ of Gumbel variables via its moment generating function

$$\Lambda_F(\lambda) \triangleq \mathbb{E}\left[\exp(\lambda F)\right]. \tag{42}$$

For notational convenience we omit the subscript when the function we consider is clear from its context. The exponential decay follows from the Markov inequality: $\mathbb{P}\left(F(\gamma) \geq r\right) \leq \Lambda(\lambda)/\exp(\lambda r)$ for any $\lambda > 0$.

We derive bounds on the moment generating function $\Lambda(\lambda)$ (in Section IV-E) by looking at the expansion (i.e., gradient) of $F(\gamma)$. Since the max-value changes at most linearly with its perturbations, its expansion is bounded and so is $\Lambda(\lambda)$. Such bounds have gained popularity in the context of isoperimetric inequalities, and series of results have established measure concentration bounds for general families of distributions, including log-concave distributions [56]–[61]. The family of log-concave distribution includes the Gaussian, Laplace, logistic and Gumbel distributions, among many others. A one dimensional density function $q(t)$ is said to be log-concave if $q(t) = \exp(-Q(t))$ and $Q(t)$ is a convex function: log-concave distributions have log-concave densities. These probability density functions decay exponentially[2] with t. To see that we recall that for any convex function $Q(t) \geq Q(0) + tQ'(0)$ for any t. By exponentiating and rearranging we can see $q(t) \leq q(0)\exp(-tQ'(0))$.

A. A Poincaré inequality for log-concave distributions

A Poincaré inequality bounds the variance of a random variable by its expected expansion, i.e., the norm of its gradient. These results are general and apply to any (almost everywhere) smooth real-valued functions $f(t)$ for $t \in \mathbb{R}^m$. The variance of a random variable (or a function) is its square distance from its expectation, according to the measure μ:

$$\mathrm{Var}_\mu(f) \triangleq \int f^2(t)d\mu(t) - \left(\int f(t)d\mu(t)\right)^2. \tag{43}$$

[2]One may note that for the Gaussian distribution $Q'(0) = 0$ and that for Laplace distribution $Q'(0)$ is undefined. Thus to demonstrate the exponential decay one may verify that $Q(t) \geq Q(c) + tQ'(c)$ for any c thus $q(t) \leq q(c)\exp(-tQ'(c))$.

A Poincaré inequality is a bound of the form

$$\mathrm{Var}_\mu(f) \leq C \int \|\nabla f(t)\|^2 d\mu(t). \tag{44}$$

If this inequality holds for any function $f(t)$ we say that the measure μ satisfies the Poincaré inequality with a constant C. The optimal constant C is called the Poincaré constant. To establish a Poincaré inequality it suffices to derive an inequality for a one-dimensional function and extend it to the multivariate case by tensorization [59, Proposition 5.6].

Restricting to one-dimensional functions, $\mathrm{Var}_\mu(f) \leq \int_{-\infty}^{\infty} f(t)^2 q(t) dt$ and the one-dimensional Poincaré inequality takes the form:

$$\int_{-\infty}^{\infty} f(t)^2 q(t) dt \leq C \int_{-\infty}^{\infty} f'(t)^2 q(t) dt. \tag{45}$$

We provide an elementary proof that is based on the seminal work of Brascamp and Lieb [56]. We begin by considering a simpler setting, where $Q'(t) \neq 0$. This setting demonstrates the core idea of our general proof while avoiding technical complications.

Lemma 2: Let μ be a log-concave measure with density $q(t) = \exp(-Q(t))$, where $Q : \mathbb{R} \to \mathbb{R}$ is a convex function, twice continuously differentiable almost everywhere, satisfying $Q'(t) \neq 0 \ \forall t$, $\lim_{t\to+\infty} Q'(t) \geq 0$, and $\lim_{t\to-\infty} Q'(t) \leq 0$. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function in $L^2(\mu)$, differentiable almost everywhere and with $f' \in L^2(\mu)$. Then for any $\eta \in \left[\min_{t\in\mathbb{R}} -\frac{Q''(t)}{(Q'(t))^2}, 1\right]$, we have

$$\mathrm{Var}_\mu(f) \leq \frac{1}{1-\eta} \int_{\mathbb{R}} \frac{(f'(t))^2}{Q''(t) + \eta (Q'(t))^2} q(t) dt.$$

Proof: The variance of $f(t)$ is upper bounded by its second moment, so it suffices to prove that $\int_{-\infty}^{\infty} f^2(t) q(t) dt \leq \int_{-\infty}^{\infty} C(t) f'(t)^2 q(t) dt$. We define the function $\psi(t) = (h(t)g(t))'$ with $h(t) = f^2(t)$ and $g(t) = q(t)/Q'(t)$. Its integral is nonnegative since $\int \psi(t) dt = \int (h(t)g(t))' dt = \lim_{t\to\infty} f^2(t) q(t)/Q'(t) - \lim_{t\to-\infty} f^2(t) q(t)/Q'(t) \geq 0$ by the assumptions on the limits of $Q'(t)$.

The main challenge of the proof is to show the following bound on the function $\psi(t)$:

$$\psi(t) = \left(f^2(t) \cdot \frac{q(t)}{Q'(t)} \right)' \leq -q(t) f^2(t) + q(t) \eta f^2(t) + q(t) \frac{f'^2(t)}{Q''(t) + \eta Q'^2(t)}. \tag{46}$$

Assuming that (46) holds, the proof then follows by taking an integral over both sides, while noticing that the left hand side is nonnegative. To prove the inequality in (46) we first note that by differentiating the function $\psi(t)$, we get

$$\left[f^2(t) \cdot \frac{q(t)}{Q'(t)} \right]' = -q(t) f^2(t) - q(t) f^2(t) \frac{Q''(t)}{Q'^2(t)} + 2 f(t) f'(t) \frac{q(t)}{Q'(t)}. \tag{47}$$

Using the inequality $2ab \leq ca^2 + b^2/c$ for any $c \geq 0$ we derive the bound

$$2\frac{f'(t)}{Q'(t)}f(t) \leq \left(c(t)f^2(t) + \frac{f'^2(t)}{c(t)Q'^2(t)}\right). \tag{48}$$

Finally, we set $c(t) = \frac{Q''(t)}{Q'^2(t)} + \eta$ to satisfy $c(t) \geq 0$ and get the inequality in (46). ■

The above proof relies on the fact that $Q'(t) \neq 0$ to ensure that the function $\int (h(t)g(t))' dt$ is nonnegative. This holds, for example, for the Laplace distribution $q(t) = \exp(-|t|)/2$ where $Q'(t) \in \{-1, 1\}$. In particular, the Poincaré inequality for the Laplace distribution given by Ledoux [59] follows from our result by setting $\eta = 1/2$.

Unfortunately, the condition $Q'(t) \neq 0$ does not hold for all log-concave measures. For example, the Gaussian distribution has $Q'(0) = 0$ and for the Gumbel distribution $Q'(-c) = 0$. The proof above fails in these cases since the function $(h(t)g(t))'$ tends to infinity around the point of singularity, namely a for which $Q'(a) = 0$. To overcome the singularity $Q'(a) = 0$ we modify the function $f(t)$ such that it vanishes at a in such a way that the numerator (namely $f(a)$) approaches zero faster than the denominator (namely $Q'(a)$).

Theorem 5: Let μ be a log-concave measure with density $q(t) = \exp(-Q(t))$, where $Q : \mathbb{R} \to \mathbb{R}$ is a convex function that has a unique minimum in the point $t = a$. Also, assume $Q(t)$ is twice continuously differentiable excepts possibly at $t = a$, $\lim_{t\to a^{\pm}} Q'(t) \neq 0$ or $\lim_{t\to a^{\pm}} Q''(t) \neq 0$, $\lim_{t\to+\infty} Q'(t) \geq 0$, and $\lim_{t\to-\infty} Q'(t) \leq 0$. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function in $L^2(\mu)$, differentiable almost everywhere and with $f' \in L^2(\mu)$. Then for any $\eta \in \left[\min_{t\in\mathbb{R}\setminus\{a\}} -\frac{Q''(t)}{(Q'(t))^2}, 1\right]$, we have

$$\mathrm{Var}_\mu(f) \leq \frac{1}{1-\eta}\int_{\mathbb{R}} \frac{(f'(t))^2}{Q''(t) + \eta(Q'(t))^2} q(t)dt.$$

Proof: To ensure that $f(a) = 0$ we use the a different bound on the variance. Specifically, $\mathrm{Var}(f) \leq \int_{-\infty}^{\infty}(f(t) - K)^2 q(t)dt$ for any K. Thus we set $K = f(a)$ and follow the proof of Lemma 2 with $\hat{f}(t) = f(t) - f(a)$. Since $f'(t) = \hat{f}'(t)$ an inequality of the form

$$\int_{-\infty}^{\infty} \hat{f}(t)^2 q(t)dt \leq \frac{1}{1-\eta}\int_{\mathbb{R}} \frac{(\hat{f}'(t))^2}{Q''(t) + \eta(Q'(t))^2} q(t)dt \tag{49}$$

provides the desired Poincaré inequality.

To complete the proof, we show that $\int (h(t)g(t))' dt$ is nonnegative while $h(t) = \hat{f}^2(t)$ and $g(t) = q(t)/Q'(t)$. We do so by dividing it into two components with respect to the point a:

$$\int_{-\infty}^{\infty} \left(\hat{f}^2(t) \cdot \frac{q(t)}{Q'(t)}\right)' dt = \int_{-\infty}^{a} \left(\hat{f}^2(t) \cdot \frac{q(t)}{Q'(t)}\right)' dt + \int_{a}^{\infty} \left(\hat{f}^2(t) \cdot \frac{q(t)}{Q'(t)}\right)' dt. \tag{50}$$

Lastly,

$$\int_{-\infty}^{a} \left(\hat{f}^2(t) \cdot \frac{q(t)}{Q'(t)}\right)' dt = \lim_{t\to a} \hat{f}^2(t)q(t)/Q'(t) - \lim_{t\to-\infty} \hat{f}^2(t)q(t)/Q'(t). \tag{51}$$

As in the proof of Lemma 2, we have that $\lim_{t\to-\infty}\hat{f}^2(t)q(t)/Q'(t)\geq 0$. The treatment of the term $\lim_{t\to a}\hat{f}^2(t)q(t)/Q'(t)$ is slightly more involved since both $Q'(a)=0$ and $\hat{f}(a)=0$. Thus to evaluate the limit we use L'Hôpital's rule and differentiate both the numerator and denominator to obtain $2\hat{f}(t)\hat{f}'(a)q(a)/Q''(a)=0$ by the assumption that either $Q'(a)=0$ or $Q''(a)=0$ but not both. The same argument follows for the integral over the interval $[a,\infty]$. ■

Brascamp and Lieb [56] proved a Poincaré inequality for strongly log-concave measures, where $Q''(t)\geq c$. Their result may be obtained by our derivation when $\eta=0$. Their result was later extended by Bobkov [62] and more recently by Nguyen [61] to log-concave measures and to multivariate functions while restricting their content to the interval $\eta\in[1/2,1]$. In the one-dimensional setting, our bound is tighter, i.e., for $\eta\in\left[\min_{t\in\mathbb{R}\setminus\{a\}}-\frac{Q''(t)}{(Q'(t))^2},1\right]$.

B. Bounds for the Gumbel distribution

For the Gumbel distribution we get the following bound.

Corollary 8 (Poincaré inequality for the Gumbel distribution): Let μ be the measure corresponding to the Gumbel distribution. Let $f:\mathbb{R}^m\to\mathbb{R}$ be a multivariate function that satisfies the conditions in Lemma 2 for each dimension. Then

$$\mathrm{Var}_\mu(f)\leq 4\int_{\mathbb{R}}\|\nabla f(t)\|^2 d\mu(t). \tag{52}$$

Proof: First, we derive the Poincaré constant for a one dimensional Gumbel distribution. Then we derive the multivariate bound by tensorization. Following Theorem 5 it suffices to show that there exists η such that $\frac{1}{(1-\eta)(Q''(t)+\eta(Q'(t))^2)}\leq 4$ for any t.

For the Gumbel distribution,

$$Q(t)=t+c+\exp(-(t+c)) \tag{53}$$

$$Q''(t)+\eta(Q'(t))^2=e^{-(t+c)}+\eta(1-e^{-(t+c)})^2. \tag{54}$$

Simple calculus shows that t^* minimizing (54) is given by

$$0=-(t^*+c)e^{-(t^*+c)}-2\eta(t^*+c)(1-e^{-(t^*+c)})e^{-(t^*+c)} \tag{55}$$

or $e^{-(t^*+c)}=1-\frac{1}{2\eta}$. The lower bound is then $Q''(t)+\eta(Q'(t))^2\geq\frac{4\eta-1}{4\eta}$ whenever $1-\frac{1}{2\eta}$ is positive, or equivalently whenever $\eta>\frac{1}{2}$. For $\eta\leq\frac{1}{2}$, we note that $Q''(t)+\eta(Q'(t))^2=\eta+(1-2\eta)e^{-(t+c)}+\eta e^{-2(t+c)}\geq\eta$.

Combining these two cases, the Poincaré constant is at most $\min\left\{\frac{4\eta}{(4\eta-1)(1-\eta)},\frac{1}{\eta(1-\eta)}\right\}=4$ at $\eta=\frac{1}{2}$. By applying Theorem 2 we obtain the one-dimensional Poincaré inequality.

Finally, for $f : \mathbb{R}^m \to \mathbb{R}$ we denote by $\mathrm{Var}_i(f)$ the variance of i-th variable while fixing the rest of the $m-1$ variables in $f(t_1, t_2, \ldots, t_m)$. The one dimensional Poincaré inequality implies that

$$\mathrm{Var}_i(f) \leq 4 \int_{\mathbb{R}} |\partial f(t)/\partial t_i|^2 d\mu. \tag{56}$$

The proof then follows by a tensorization argument given by Ledoux [59, Proposition 5.6], which shows that

$$\mathrm{Var}_\mu(f) \leq \sum_{i=1}^{m} \mathbb{E}_{t \setminus t_i} \left[\mathrm{Var}_i(f) \right]. \tag{57}$$

■

Although the Poincaré inequality establishes a bound on the variance of a random variable, it may also be used to bound the moment generating function $\Lambda_f(\lambda) = \int \exp(\lambda f(t)) d\mu(t)$ [57], [58], [60]. For completeness we provide a proof specifically for the Gumbel distribution.

Corollary 9 (MGF bound for the Gumbel distribution): Let μ be the measure corresponding to the Gumbel distribution. Let $f : \mathbb{R}^m \to \mathbb{R}$ be a multivariate function that satisfies the conditions in Theorem 5 for each dimension and that $\|\nabla f(t)\| \leq a$. Then whenever $\lambda a \leq 1$ the moment generating function is bounded as

$$\Lambda(\lambda) \leq \frac{1+\lambda a}{1-\lambda a} \cdot \exp\left(\lambda \mathbb{E}\left[f\right]\right). \tag{58}$$

Proof: The proof is due to Bobkov and Ledoux [58]. Applying the Poincaré inequality with $g(t) = \exp(\lambda f(t)/2)$ implies

$$\Lambda(\lambda) - \Lambda(\lambda/2)^2 \leq 4 \int \frac{\lambda^2}{4} \exp(\lambda f(t)) \cdot \|\nabla f(t)\|^2 d\mu(t) \leq a^2 \lambda^2 \Lambda(\lambda). \tag{59}$$

Whenever $\lambda^2 a^2 \leq 1$ one can rearrange the terms to get the bound $\Lambda(\lambda) \leq \left(1 - \lambda^2 a^2\right)^{-1} \Lambda(\lambda/2)^2$. Applying this self-reducible bound recursively k times implies

$$\Lambda(\lambda) \leq \Lambda\left(\frac{\lambda}{2^k}\right)^{2^k} \prod_{i=0}^{k} \left(1 - \frac{\lambda^2 a^2}{4^i}\right)^{-2^i} \tag{60}$$

$$= \left(1 + \frac{\lambda \mathbb{E}\left[f\right]}{2^k} + o(2^k)\right)^{2^k} \prod_{i=0}^{k} \left(1 - \frac{\lambda^2 a^2}{4^i}\right)^{-2^i}, \tag{61}$$

where the last line follows from a Taylor expansion of (42). Taking $k \to \infty$ and noting that $(1+c/2^k)^{2^k} \to e^c$ we obtain the bound $\Lambda(\lambda) \leq \prod_{i=0}^{\infty} \left(1 - \frac{\lambda^2 a^2}{4^i}\right)^{-2^i} \exp(\lambda \mathbb{E}\left[f\right])$. Applying Lemma 4 (see the Appendix for a proof) shows that $\prod_{i=0}^{\infty} \left(1 - \frac{\lambda^2 a^2}{4^i}\right)^{-2^i} \leq \frac{1+\lambda a}{1-\lambda a}$, which completes the proof. ■

Bounds on the moment generating function generally imply (via the Markov inequality) that the deviation of a random variable from its mean decays exponentially in the number of samples. We apply

these inequalities in a setting in which the function f is random and we think of it as a random variable. With some abuse of notation then, we will call f a random variable in the following corollary.

Corollary 10 (Measure concentration for the Gumbel distribution): Consider a random function f that satisfies the same assumptions of Corollary 9 almost surely. Let $f_1, f_2, \ldots, f_M$ be M i.i.d. random variables with the same distribution as f. Then with probability at least $1-\delta$,

$$\frac{1}{M}\sum_{j=1}^{M} f_j - \mathbb{E}[f] \leq 2a\left(1+\sqrt{\frac{1}{2M}\log\frac{1}{\delta}}\right)^2.$$

Proof: From the independence assumption, using the Markov inequality, we have that

$$\mathbb{P}\left(\sum_{j=1}^{M} f_j \leq M\mathbb{E}[f] + Mr\right) \leq \exp(-\lambda M\mathbb{E}[f] - \lambda M r)\prod_{j=1}^{M}\mathbb{E}[\exp(\lambda f_j)].$$

Applying Corollary 9, we have, for any $\lambda \leq 1/a$,

$$\mathbb{P}\left(\frac{1}{M}\sum_{j=1}^{M} f_j \leq \mathbb{E}[f] + r\right) \leq \exp\Big(M(\log(1+\lambda a) - \log(1-\lambda a) - \lambda r)\Big).$$

Optimizing over positive λ subject to $\lambda \leq 1/a$ we obtain $\lambda = \frac{\sqrt{r-2a}}{a\sqrt{r}}$ for $r \geq 2a$. Hence, for $r \geq 2a$, the right side becomes

$$\exp\left(M\left(2\tanh^{-1}\left(\sqrt{1-\frac{2a}{r}}\right) - \frac{r\sqrt{1-\frac{2a}{r}}}{a}\right)\right) \tag{62}$$

$$\leq \exp\left(2M\sqrt{\frac{r}{2a}-1}\left(1-\sqrt{\frac{r}{2a}}\right)\right) \tag{63}$$

$$\leq \exp\left(-2M\left(\sqrt{\frac{r}{2a}}-1\right)^2\right), \tag{64}$$

where the first inequality can be easily verified comparing the derivatives. Equating the left side of the last inequality to δ and solving for r, we have the stated bound. ■

C. A modified log-Sobolev inequality for log-concave distributions

In this section we provide complementary measure concentration results that bound the moment generating function $\Lambda(\lambda) = \int \exp(\lambda f(t))d\mu(t)$ by its expansion (in terms of gradients). Such bounds are known as modified log-Sobolev bounds. We follow the same recipe as previous works [58], [59] and use the so-called Herbst argument. Consider the λ-scaled cumulant generating function of a random function with zero mean, i.e., $\mathbb{E}[f] = 0$:

$$K(\lambda) \triangleq \frac{1}{\lambda}\log\Lambda(\lambda). \tag{65}$$

First note that by L'Hôpital's rule $K(0) = \frac{\Lambda'(0)}{\Lambda(0)} = \mathbb{E}[f]$, so whenever $\mathbb{E}[f] = 0$ we may represent $K(\lambda)$ by integrating its derivative: $K(\lambda) = \int_0^\lambda K'(\hat{\lambda})d\hat{\lambda}$. Thus to bound the moment generating function it suffices to bound $K'(\lambda) \leq \alpha(\lambda)$ for some function $\alpha(\lambda)$. A direct computation of $K'(\lambda)$ translates this bound to

$$\lambda\Lambda'(\lambda) - \Lambda(\lambda)\log\Lambda(\lambda) \leq \lambda^2\Lambda(\lambda)\alpha(\lambda). \tag{66}$$

The left side of (66) turns out to be the so-called functional entropy Ent, which is not the same as the Shannon entropy [59]. We calculate the functional entropy of $\lambda f(t)$ with respect to a measure μ:

$$\mathrm{Ent}_\mu(\exp(f)) \stackrel{\Delta}{=} \int f(t)\cdot\exp(f(t))d\mu(t) - \left(\int \exp(f(t))d\mu(t)\right)\log\Big(\int \exp(f(t))d\mu(t)\Big).$$

In the following we derive a variation of the modified log-Sobolev inequality for log-concave distributions based on a Poincaré inequality for these distributions. This in turn provides a bound on the moment generating function. This result complements the exponential decay that appears in Section IV-A. Figure 2 compares these two approaches.

Lemma 3: Let μ be a measure that satisfies the Poincaré inequality with a constant C, i.e., $\mathrm{Var}_\mu(f) \leq C\int\|\nabla f(t)\|^2 d\mu(t)$ for any continuous and differentiable almost everywhere function $f(t)$. Assume that $\int f(t)d\mu(t) = 0$ and that $\|\nabla f(t)\| \leq a < 2/\sqrt{C}$. Then

$$\mathrm{Ent}_\mu(\exp(f)) \leq \frac{a^2C}{2}\left(\frac{2+a\sqrt{C}}{2-a\sqrt{C}}\right)^2\int\exp(f(t))d\mu(t). \tag{67}$$

Proof: First, $z\log z \geq z+1$. Setting $z = \int\exp(f)d\mu$ and applying this inequality results in the functional entropy bound $\mathrm{Ent}_\mu(\exp(f)) \leq \int f(t)\exp(f(t))d\mu(t) - \big(\int\exp(f(t))d\mu(t)\big) + 1$. Rearranging the terms, the right hand side is $\int\left(f(t)\cdot\exp(f(t)) - \exp(f(t)) + 1\right)d\mu(t)$. We proceed by using the identity (cf. [63], Equation 17.25.2) of the indefinite integral $\int s\exp(sc)ds = \frac{\exp(sc)}{c}(s - \frac{1}{c})$. Taking into account the limits $[0,1]$ and setting $c = f(t)$ we get the desired form: $f(t)^2\int_0^1 s\exp(sf(t))ds = f(t)\exp(f(t)) - \exp(f(t)) + 1$. Particularly,

$$\mathrm{Ent}_\mu(\exp(f)) \leq \int\left(\int_0^1 sf(t)^2\exp(sf(t))ds\right)d\mu(t) \tag{68}$$

$$= \lim_{\epsilon\to 0^+}\int_\epsilon^1\frac{1}{s}\left(\int s^2f(t)^2\exp(sf(t))d\mu(t)\right)ds. \tag{69}$$

The last equality holds by Fubini's theorem.

Next we use Proposition 3.3 from [58] that applies the Poincaré inequality to $g(t)\exp(g(t)/2)$ to show that for any function $g(t)$ with mean zero and $\|\nabla g(t)\| \leq a < 2/\sqrt{C}$, we have the inequality

$$\int g^2(t)\exp(g(t))d\mu(t) \leq \hat{C}\int\|\nabla g(t)\|^2\exp(g(t))d\mu(t), \tag{70}$$

where $\hat{C} = C((2 + a\sqrt{C})/(2 - a\sqrt{C}))^2$. Setting $g(t) = sf(t)$ in this inequality satisfies $\|\nabla g(t)\| = s\|\nabla f(t)\| \leq a < 2/\sqrt{C}$. This implies the inequality

$$\int s^2 f(t)^2 \exp(sf(t)) d\mu(t) \leq s^2 C \left(\frac{2 + a\sqrt{C}}{2 - a\sqrt{C}}\right)^2 \int \|\nabla f(t)\|^2 \exp(sf(t)) d\mu(t). \tag{71}$$

Using $\|\nabla f\| \leq a$ we obtain the bound

$$\mathrm{Ent}_\mu(\exp(f)) \leq a^2 C \left(\frac{2 + a\sqrt{C}}{2 - a\sqrt{C}}\right)^2 \int_0^1 s \left(\int \exp(sf(t)) d\mu(t)\right) ds. \tag{72}$$

The function $\phi(s) = \int \exp(sf(t)) d\mu(t)$ is convex in the interval $s \in [0, 1]$, so its maximum value is attained at $s = 0$ or $s = 1$. Also, $\phi(0) = 1$ and $\phi(1) = \int \exp(f(t)) d\mu(t)$. From Jensen's inequality, and the fact that $\int f(t) d\mu(t) = 0$, we have $\int \exp(f(t)) d\mu(t) \geq \exp(\int f(t) d\mu(t)) = 1$. Hence $\phi(1) \geq \phi(0)$. So, we have

$$\begin{aligned} \int_0^1 s \left(\int \exp(sf(t)) d\mu(t)\right) ds &\leq \int \exp(f(t)) d\mu(t) \cdot \int_0^1 s ds \\ &= \frac{1}{2} \int \exp(f(t)) d\mu(t) \end{aligned}$$

and the result follows. ■

The preceding lemma expresses an upper bound on the functional entropy in terms of the moment generating function. Applying this Lemma with the function λf, and assuming that $\|\nabla f\| \leq a$, we rephrase this upper bound as $\mathrm{Ent}_\mu(\exp(\lambda f)) \leq \Lambda(\lambda) \cdot \frac{\lambda^2 a^2 C}{2} \left(\frac{2 + \lambda a\sqrt{C}}{2 - \lambda a\sqrt{C}}\right)^2$, where $\Lambda(\lambda)$ is the moment generating function of f. Fitting it to (66) and (65) we deduce that

$$K'(\lambda) \leq \alpha(\lambda) = \frac{a^2 C}{2} \left(\frac{2 + \lambda a\sqrt{C}}{2 - \lambda a\sqrt{C}}\right)^2. \tag{73}$$

Since the Poincaré constant of the Gumbel distribution is at most 4 we obtain its corresponding bound $K'(\lambda) \leq 2a^2 \left(\frac{1 + \lambda a}{1 - \lambda a}\right)^2$. Applying the Herbst argument (cf. [59]), this translates to a bound on the moment generating function. This result is formalized as follows.

Corollary 11: Let μ denote the Gumbel measure on $\mathbb{R}$ and let $f : \mathbb{R}^m \to \mathbb{R}$ be a multivariate function that satisfies the conditions in Lemma 2 for each dimension. Also, assume that $\|\nabla f(t)\| \leq a$. Then whenever $\lambda a \leq 1$ the moment generating function is bounded as

$$\Lambda(\lambda) \leq \beta(\lambda) \, \exp\left(\lambda \mathbb{E}\left[f\right]\right),$$

where $\beta(\lambda) = \exp\left(2a^2\lambda^2 \frac{5 - \lambda a}{1 - \lambda a} + 8a\lambda \log(1 - \lambda a)\right)$.

Proof: We apply Lemma 3 to the function $\hat{f}(t) = f(t) - \mathbb{E}[f]$ which has zero mean. Thus

$$K'(\lambda) = \frac{\mathrm{Ent}_\mu(\exp(\lambda \hat{f}))}{\lambda^2 \Lambda(\lambda)} \leq 2a^2 \left(\frac{1 + \lambda a}{1 - \lambda a}\right)^2. \tag{74}$$

Recalling that $K(0) = 0$ we derive $K(\lambda) = \int_0^\lambda K'(\hat{\lambda})d\hat{\lambda}$. Using the bound on $K'(\lambda)$ we obtain

$$K(\lambda) \leq 2a^2 \int_0^\lambda \left(\frac{1+\hat{\lambda}a}{1-\hat{\lambda}a}\right)^2 d\hat{\lambda}. \tag{75}$$

A straight forward verification of the integral implies that

$$\begin{aligned} K(\lambda) &\leq 2a^2 \left[\frac{4\log(1-a\lambda)}{a} + \frac{4}{a(1-a\lambda)} + \lambda - \frac{4}{a}\right] \\ &= 2a^2 \left[\frac{4\log(1-a\lambda)}{a} + \frac{4\lambda}{1-a\lambda} + \lambda\right]. \end{aligned}$$

Now, from the definition of $K(\lambda)$ and the one of $\beta(\lambda)$, this implies $\log \mathbb{E}[\exp(\lambda \hat{f})] \leq \log \beta(\lambda)$. ■

D. Evaluating measure concentration

The above bound is tighter than our previous bound in Theorem 3 of [3]. In particular, the bound in Corollary 11 does not involve $\|f(t)\|_\infty$. It is interesting to compare $\beta(\lambda)$ in the above bound to the one that is attained directly from Poincaré inequality in Corollary 9, namely $\Lambda(\lambda) \leq \alpha(\lambda) \cdot \exp(\lambda \mathbb{E}[f])$ where $\alpha(\lambda) = \frac{1+\lambda a}{1-\lambda a}$. Both $\alpha(\lambda), \beta(\lambda)$ are finite in the interval $0 \leq \lambda < 1/a$ although they behave differently at the their limits. Particularly, $\alpha(0) = 1$ and $\beta(0) = 1$. On the other hand, $\alpha(\lambda) < \beta(\lambda)$ for $\lambda \to 1$. This is illustrated in Figure 2.

With this Lemma we can now upper bound the error in estimating the average $\mathbb{E}[f]$ of a function f of m i.i.d. Gumbel random variables by generating M independent samples of f and taking the sample mean. We again abuse notation to think of f as a random variable itself.

Corollary 12 (Measure concentration via log-Sobolev inequalities): Consider a random function f that satisfies the same assumptions as Corollary 11 almost surely. Let $f_1, f_2, \ldots, f_M$ be M i.i.d. random variables with the same distribution as f. Then with probability at least $1 - \delta$,

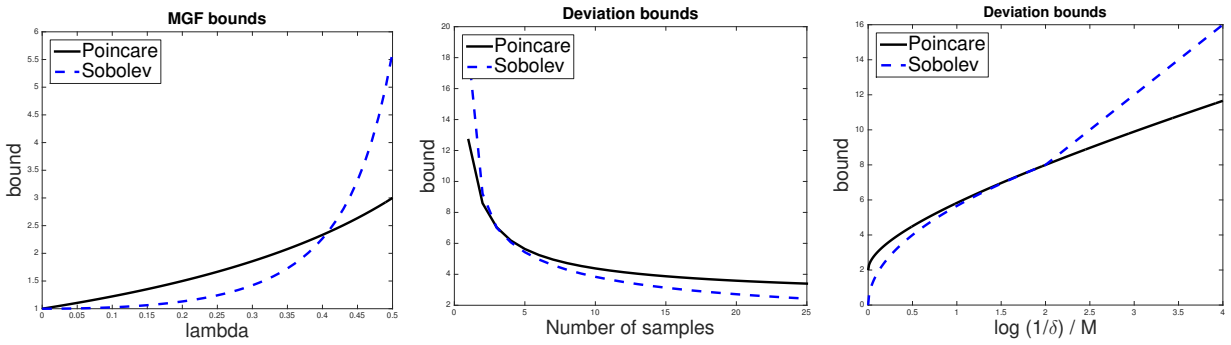
$$\frac{1}{M}\sum_{j=1}^M f_j - \mathbb{E}[f] \leq a \max\left(\frac{4}{M}\log\frac{1}{\delta}, \sqrt{\frac{32}{M}\log\frac{1}{\delta}}\right).$$

Proof: From the independence assumption, using the Markov inequality, we have that

$$\mathbb{P}\left(\sum_{j=1}^M f_j \leq M\mathbb{E}[f] + Mr\right) \leq \exp(-M\lambda\mathbb{E}[f] - Mr\lambda)\prod_{j=1}^M \mathbb{E}[\exp(\lambda f_j)].$$

We use the elementary inequality $\log(1-x) \leq \frac{-2x}{2-x}$ and Corollary 11, to have

$$\mathbb{P}\left(\frac{1}{M}\sum_{j=1}^M f_j \leq \mathbb{E}[f] + r\right) \leq \exp\left(M\left(2a^2\lambda^2\frac{a^2\lambda^2 + a\lambda + 2}{(1-a\lambda)(2-a\lambda)} - \lambda r\right)\right).$$



(a) (b) (c)

Fig. 2. Comparing the measure concentration bounds that are attained by the Poincaré and modified log-Sobolev inequalities for $a = 1$. Figure (2a): the moment generating functions bounds that are attained by the Poincaré inequality in Corollary 9 and the modified log-Sobolev inequality in Corollary 11. The plots show that respective functions $\alpha(\lambda) = \frac{1+\lambda a}{1-\lambda a}$ and $\beta(\lambda)$ that appear in these bounds. Figure (2b): the deviation bounds, of the sampled average from its mean, that are attained by the Poincaré inequality in Corollary 10 and the modified log-Sobolev inequality in Corollary 12 when $\delta = 0.1$ and the number of samples $M = 1, 2, \ldots, 100$. Figure (2c): the deviation bounds, of the sampled average from its mean, that are attained by the Poincaré inequality in Corollary 10 and the modified log-Sobolev inequality in Corollary 12, as a function for $\log(1/\delta)/M$ that ranges between $[0, 2]$.

For[3] any $|\lambda| \leq \frac{13}{25a}$, we have that $2\frac{a^2\lambda^2+a\lambda+2}{(1-a\lambda)(2-a\lambda)} \leq 8$. Hence, for any $|\lambda| \leq \frac{13}{25a}$, we have that

$$\mathbb{P}\left(\frac{1}{M}\sum_{j=1}^{M} f_j \leq \mathbb{E}[f] + r\right) \leq \exp\left(M\left(8a^2\lambda^2 - \lambda r\right)\right).$$

Optimizing over λ subject to $|\lambda| \leq \frac{13}{25a}$ we obtain

$$\exp\left(M\left(8a^2\lambda^2 - \lambda r\right)\right) \leq \exp\left(-M\min\left(\frac{r}{4a}, \frac{r^2}{32a^2}\right)\right).$$

Equating the left side of the last inequality to δ and solving for r, we have the stated bound. ■

E. Application to MAP perturbations

The derived bounds on the moment generating function imply the concentration of measure of our high-dimensional inference algorithms that we use both for sampling (in Section III-A) and to estimate prediction uncertainties or entropies (in Section III-C). The relevant quantities for our inference algorithms are the expectation of randomized max-solvers $F(\gamma) = \max_{\mathbf{x}}\{\theta(\mathbf{x}) + \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha)\}$ as well as the expectation of the maximizing perturbations $F(\gamma) = \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha^\gamma)$ for which $x^\gamma = \operatorname{argmax}_{\mathbf{x}}\{\theta(\mathbf{x}) + \sum_\alpha \gamma_\alpha(\mathbf{x}_\alpha)\}$, as in Theorem 4.

[3]The constants are found in order to have the junction of curve to approximately lie on the Poincare curve in Figure 2c.

To apply our measure concentration results to perturb-max inference we calculate the parameters in the bound given by the Corollary 9 and Corollary 11. The random functions $F(\gamma)$ above are functions of $m \triangleq \sum_{\alpha \in \mathcal{A}} |\mathcal{X}_\alpha|$ i.i.d. Gumbel random variables. The (sub)gradient of these functions is structured and points toward the $\gamma_\alpha(\mathbf{x}_\alpha)$ that corresponding to the maximizing assignment in $\mathbf{x}^*$, that is

$$\frac{\partial F(\gamma)}{\partial \gamma_\alpha(x_\alpha)} = \begin{cases} 1 & \text{if } x_\alpha \in \mathbf{x}^* \\ 0 & \text{otherwise} \end{cases}$$

Thus the gradient satisfies $\|\nabla F\|^2 = |\mathcal{A}|$ almost everywhere, so $a^2 = |\mathcal{A}|$. Suppose we sample M i.i.d. random variables $\gamma_1, \gamma_2, \ldots, \gamma_M$ with the same distribution as γ and denote their respective random values by $F_j \stackrel{def}{=} F(\gamma_j)$. We estimate their deviation from the expectation by $\frac{1}{M}\sum_{i=1}^{M} F_j - E[F]$. Applying Corollary 9 to both F and $-F$ we get the following double-sided bound with probability $1-\delta$:

$$\frac{1}{M}\sum_{j=1}^{M} F_j - \mathbb{E}[F] \leq 2\sqrt{|\mathcal{A}|}\left(1 + \sqrt{\frac{1}{2M}\log\frac{2}{\delta}}\right)^2.$$

Applying Corollary 11 to both F and $-F$ we get the following double-sided bound with probability $1-\delta$:

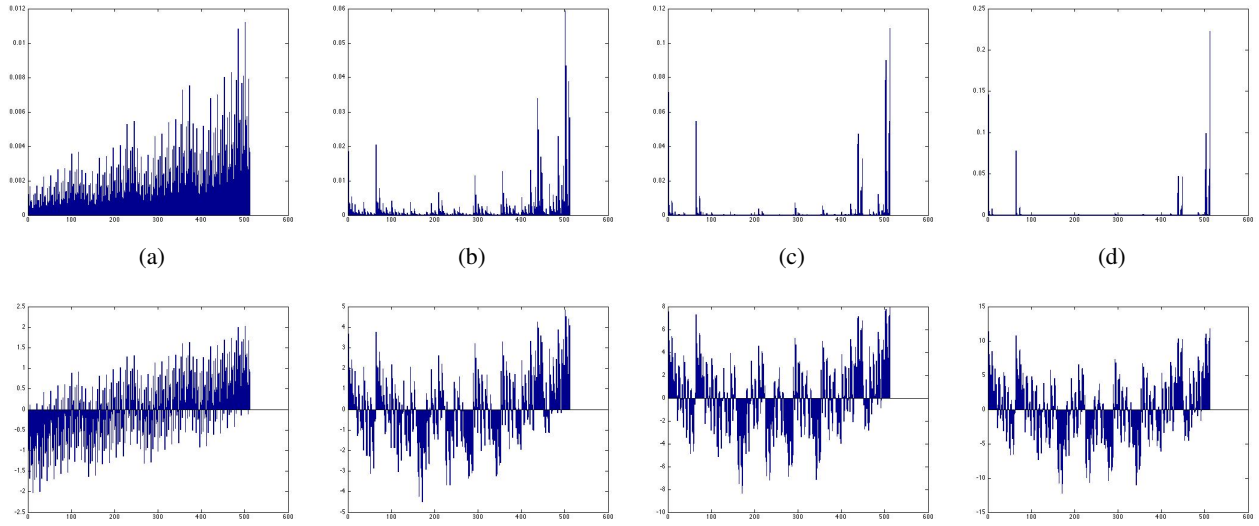
$$\frac{1}{M}\sum_{j=1}^{M} F_j - \mathbb{E}[F] \leq \sqrt{|\mathcal{A}|}\max\left(\frac{4}{M}\log\frac{2}{\delta}, \sqrt{\frac{32}{M}\log\frac{2}{\delta}}\right).$$

Clearly, the concentration of perturb-max inference is determined by the best concentration of these two bounds.

V. Empirical Evaluation

Statistical inference of high dimensional structures is closely related to estimating the partition function. Our proposed inference algorithms, both for sampling and inferring the entropy of high-dimensional structures, are derived from an alternative interpretation of the partition function as the expected value of the perturb-max value. We begin our empirical validation by computing the upper and lower bounds for the partition function computed as the expected value of a max-function. We then show empirically that the perturb-max algorithm for sampling from the Gibbs distribution has a sub-exponential computational complexity. Subsequently, we evaluate the properties of the perturb-max entropy bounds. Also, we explore the deviation of the sample mean of the perturb-max value from its expectation by deriving new measure concentration inequalities. Lastly, we use MAP perturbations inference as a sub-procedure for supervised learning binary image denoising (spin glass model) and demonstrate its power over structured-SVMs.

We evaluate our approach on spin glass models, where each variable x_i represents a spin, namely $x_i \in \{-1, 1\}$. Each spin has a local field parameter θ_i which correspond to the local potential function



(a) (b) (c) (d)

(e) $\theta_{i,j} = 0$ (f) $\theta_{i,j} \in [0, 1)$ (g) $\theta_{i,j} \in [0, 2)$ (h) $\theta_{i,j} \in [0, 3)$

Fig. 3. The probability (top row) and energy (bottom row) landscapes for all 512 configurations in a 3×3 spin glass system with strong local field, $\theta_i \in [-1, 1]$. When $\theta_{i,j} = 0$ the system is independent and one can observe the block pattern. As the coupling potentials get stronger the landscape get more ragged. By zooming one can see the ragged landscapes throughout the space, even for negligible configurations, which affect many local approaches. The random MAP perturbation directly targets the maximal configurations, thus performs well in these settings.

$\theta_i(x_i) = \theta_i x_i$. The parameter θ_i represents data signal, which in the spin model is the preference of a spin to be positive or negative. Adjacent spins interact with couplings $\theta_{i,j}(x_i x_j) = \theta_{i,j} x_i x_j$. Whenever the coupling parameters are positive the model is called attractive because adjacent variables give higher values to positively correlated configurations. The potential function of a spin glass model is then

$$\theta(x_1, x_2, \ldots, x_n) = \sum_{i \in V} \theta_i x_i + \sum_{(i,j) \in E} \theta_{i,j} x_i x_j. \tag{76}$$

In our experiments we consider adjacencies of a grid-shaped model. We used low dimensional random perturbations $\gamma_i(x_i)$ since such perturbations do not affect the complexity of the MAP solver.

Evaluating the partition function is challenging when considering strong local field potentials and coupling strengths. The corresponding energy landscape is ragged, and characterized by a relatively small set of dominating configurations. An example of these energy and probability landscapes are presented in Figure 3.

First, we compared our bounds to the partition function on 10×10 spin glass models. For such comparison we computed the partition function exactly using dynamic programming (the junction tree algorithm). The local field parameters θ_i were drawn uniformly at random from $[-f, f]$, where $f \in$

$\{0.1, 1\}$ reflects weak and strong data signal. The parameters $\theta_{i,j}$ were drawn uniformly from $[0, c]$ to obtain attractive coupling potentials. Attractive potentials are computationally favorable as their MAP value can be computed efficiently by the graph-cuts algorithm [15]. First, we evaluate an upper bound in (35) that holds in expectation with perturbations $\gamma_i(x_i)$. The expectation was computed using 100 random MAP perturbations, although very similar results were attained after only 10 perturbations, e.g., Figure 8 and Figure 9. We compared this upper bound with the sum-product form of tree re-weighted belief propagation with uniform distribution over the spanning trees [54]. We also evaluate our lower bound that holds in probability and requires only a single MAP prediction on an expanded model, as described in Corollary 4. We estimate our probable bound by expanding the model to 1000×1000 grids, setting the discrepancy ϵ in Corollary 4 to zero.[4] We compared this lower bound to the belief propagation algorithm, that provides the tightest lower bound on attractive models [64]–[66]. We computed the signed error (the difference between the bound and $\log Z$), averaged over 100 spin glass models, see Figure 9. One can see that the probabilistic lower bound is the tightest when considering the medium and high coupling domain, which is traditionally hard for all methods. Because the bound holds only with high probability probability it might generate a (random) estimate which is not a proper lower bound. We can see that on average this does not happen. Similarly, our perturb-max upper bound is better than the tree re-weighted upper bound in the medium and high coupling domain. In the attractive setting, both our bounds use the graph-cuts algorithm and were therefore considerably faster than the belief propagation variants. Finally, the sum-product belief propagation lower bound performs well on average, but from the plots one can observe that its variance is high. This demonstrates the typical behavior of belief propagation: it finds stationary points of the non-convex Bethe free energy and therefore works well on some instances but does not converge or attains bad local minima on others.

We also compared our bound in the mixed case, where the coupling potentials may either be attractive or repulsive, namely $\theta_{ij} \in [-c, c]$. Recovering the MAP solution in the mixed coupling domain is harder than the attractive domain. Therefore we could not test our lower bound in the mixed setting as it relies on expanding the model. We also omit the comparison to the sum-product belief propagation since it is no longer a lower bound in this setting. We evaluate the MAP perturbation value using MPLP [7]. One can verify that qualitatively the perturb-max upper bound is significantly better than the tree re-weighted upper bound. Nevertheless it is significantly slower as it relies on finding the MAP solution, a harder

[4]The empirical results show that even with $\epsilon = 0$, it is still a lower bound is tighter, with high probability, which may hint that there is a better analysis for this tight bound.

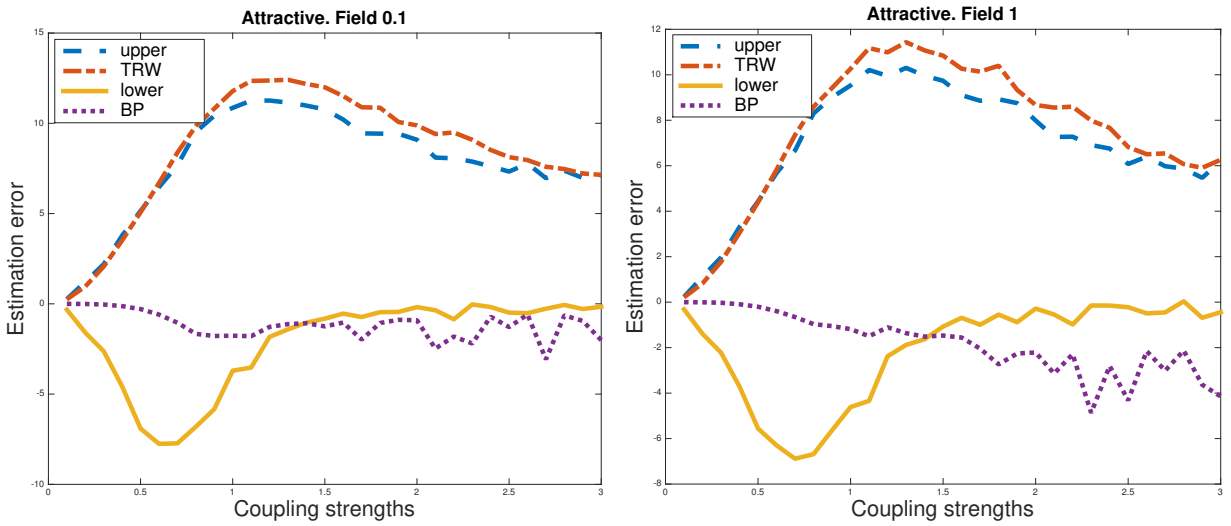


Fig. 4. The attractive case. The (signed) difference of the different bounds and the log-partition function. These experiments illustrate our bounds on 10×10 spin glass model with weak and strong local field potentials and attractive coupling potentials. The plots below zero are lower bounds and plots above zero are upper bounds. We compare our upper bound (35) with the tree re-weighted upper bound. We compare our lower bound (Corollary 4) with the belief propagation result, whose stationary points are known to be lower bounds to the log-partition function for attractive spin-glass models.

task in the presence of mixed coupling strengths.

Next, we evaluate the computational complexity of our sampling procedure. Section III-A describes an algorithm that generates unbiased samples from the full Gibbs distribution. For spin glass models with strong local field potentials, it is well-known that one cannot produce unbiased samples from the Gibbs distributions in polynomial time [8]–[10]. Theorem 3 connects the computational complexity of our unbiased sampling procedure to the gap between the log-partition function and its upper bound in (35). We use our probable lower bound to estimate this gap on large grids, for which we cannot compute the partition function exactly. Figure 6 suggests that in practice, the running time for this sampling procedure is sub-exponential.

Next we estimate our upper bounds for the entropy of perturb-max probability models that are described in Section III-C. We compare them to marginal entropy bounds $H(p) \leq \sum_i H(p_i)$, where $p_i(x_i) = \sum_{x \setminus x_i} p(x)$ are the marginal probabilities [55]. Unlike the log-partition case which relates to the entropy of Gibbs distributions, it is impossible to use dynamic programming to compute the entropy of perturb-max models. Therefore we restrict ourselves to a 4×4 spin glass model to compare these upper bounds as shown in Figure 7. One can see that the MAP perturbation upper bound is tighter than the marginalization

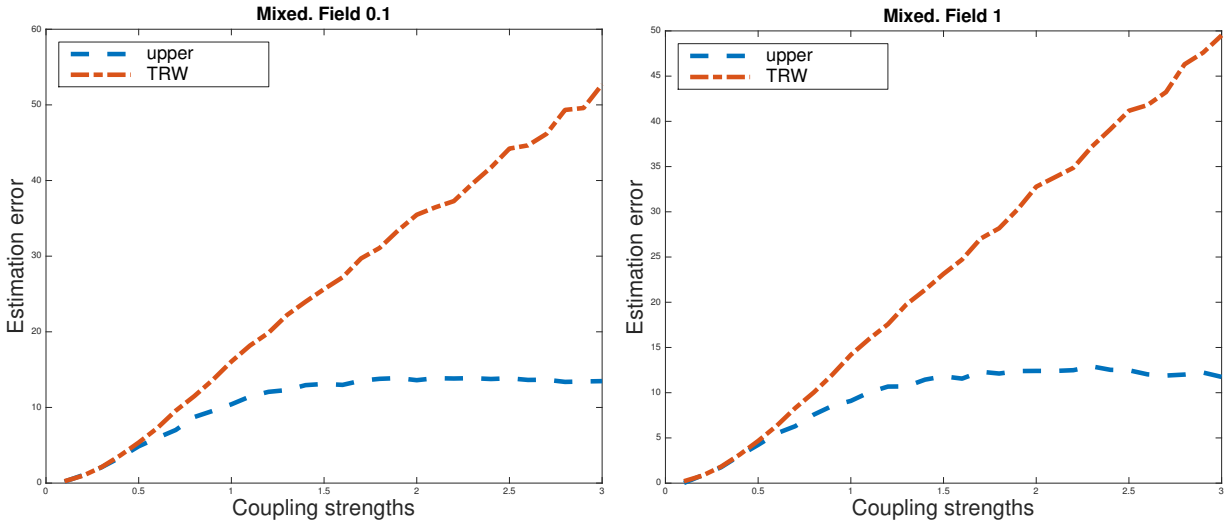


Fig. 5. The (signed) difference of the different bounds and the log-partition function. These experiments illustrate our bounds on 10×10 spin glass model with weak and strong local field potentials and mixed coupling potentials. We compare our upper bound (35) with the tree re-weighted upper bound.

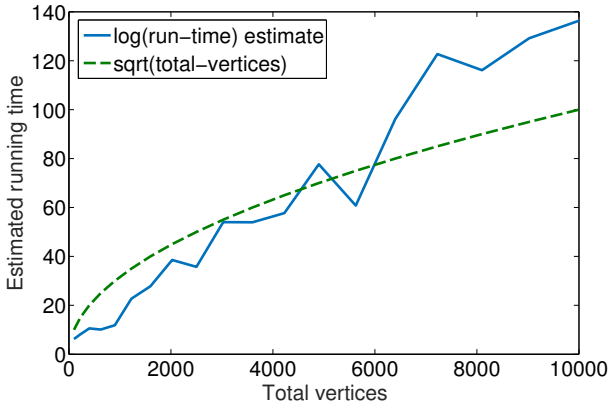


Fig. 6. Estimating our unbiased sampling procedure complexity on spin glass models of varying sizes, ranging from 10×10 spin glass models to 100×100 spin glass models. The running time is the difference between our upper bound in (35) and the log-partition function. Since the log-partition function cannot be computed for such a large scale model, we replaced it with its lower bound in Corollary 4.

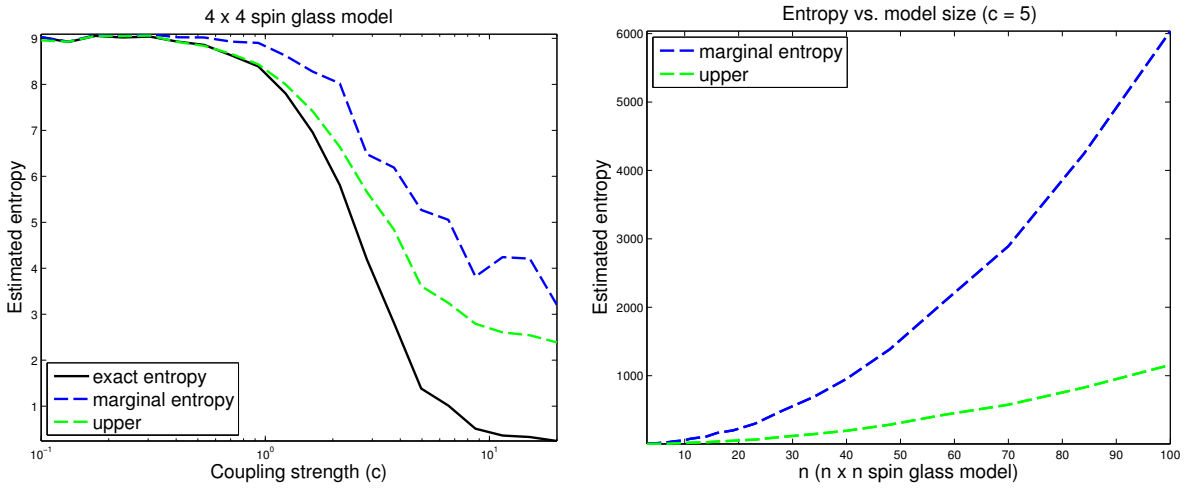


Fig. 7. Estimating our entropy bounds (in Section III-C) while comparing them to the true entropy and the marginal entropy bound. Left: comparison on small-scale spin models. Right: comparison on large-scale spin glass models.

upper bound in the medium and high coupling strengths. We can also compare the marginal entropy bounds and the perturb-max entropy bounds to arbitrary grid sizes without computing the true entropy. Figure 7 shows that the larger the model the better the perturb-max bound.

Both our log-partition bounds as well as our entropy bounds hold in expectation. Thus we evaluate their measure concentration properties, i.e., how many samples are required to converge to their expected value. We evaluate our approach on a 100×100 spin glass model with $n = 10^4$ variables. The local field parameters θ_i were drawn uniformly at random from $[-1, 1]$ to reflect high signal. To find the perturb-max assignment for such a large model we restrict ourselves to attractive coupling setting; the parameters $\theta_{i,j}$ were drawn uniformly from $[0, c]$, where $c \in [0, 4]$ to reflect weak, medium and strong coupling potentials. Throughout our experiments we evaluate the expected value of our bounds with 100 different samples. We note that both our log-partition and entropy upper bounds have the same gradient with respect to their random perturbations, so their measure concentration properties are the same. In the following we only report the concentration of our entropy bounds; the same concentration occurs for our log-partition bounds.

Figure 8 shows the error in the sample mean $\frac{1}{M}\sum_{j=1}^{M} F_j$ as described in Section IV-E. We do so for three different sample sizes $M = 1, 5, 10$, while $F(\gamma) = \sum_i \gamma_i(x_i^\gamma)$ is our entropy bound. The error reduces rapidly as M increases; only 10 samples are needed to estimate the expectation of the perturb-max random function that consist of 10^4 random variables $\gamma_i(x_i)$. To test our measure concentration

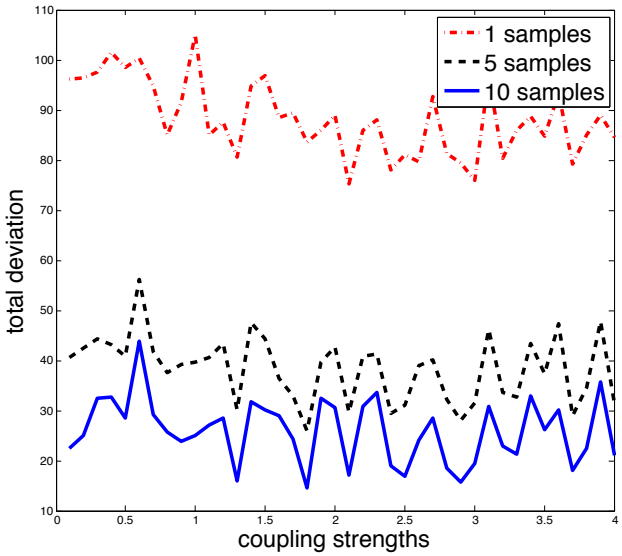


Fig. 8. Error of the sample mean versus coupling strength for 100×100 spin glass models. The local field parameter θ_i is chosen uniformly at random from $[-1, 1]$ to reflect high signal. With only 10 samples one can estimate the expectation well.

result, that ensures exponential decay, we measure the deviation of the sample mean from its expectation by using $M = 1, 5, 10$ samples. Figure 9 shows the histogram of the sample mean, i.e., the number of times that the sample mean has error more than r from the true mean. One can see that the decay is indeed exponential for every M, and that for larger M the decay is much faster. These show that by understanding the measure concentration properties of MAP perturbations, we can efficiently estimate the mean with high probability, even in very high dimensional spin-glass models.

Lastly, we demonstrate the effectiveness of MAP perturbations in supervised learning. We consider binary image denoising, which is equivalent to learning the parameters of a spin glass model. The training data was composed of ten 100×70 binary images, consisting of a man in silhouette corrupted by random binary noise, described in Figure 10. Each image x is described by binary local features $\phi_i(x, y_i)$ which equals 1 if the i-th pixel of image x is foreground and -1 otherwise. The pairwise features $\phi_{i,j}(y_i, y_i) = 1$ if $y_i = y_j$ and -1 otherwise. The goal is to estimate the parameters $\theta_i, \theta_{i,j}$ to de-noise the images. The parameters θ_i determine the importance of background-foreground observations in pixel i, and the parameters $\theta_{i,j}$ determine the coupling nature of pixels i, j, namely their attractive or repulsive strength.

Since we are considering 100×70 images, there are about $20,000$ parameters to estimate. Conditional

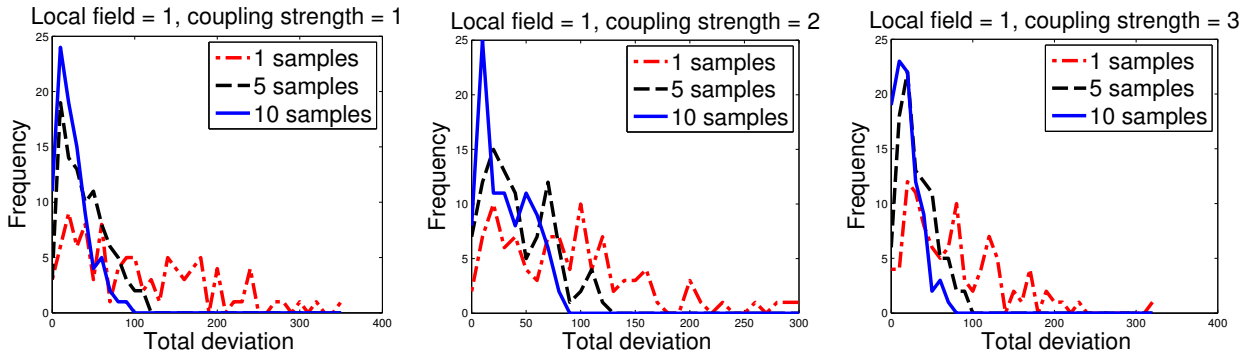


Fig. 9. Histogram that shows the decay of random MAP values, i.e., the number of times that the sample mean has error more than r from the true mean. These histograms are evaluated on 100×100 spin glass model with high signal $\theta_i \in [-1, 1]$ and various coupling strengths. One can see that the decay is indeed exponential for every M, and that for larger M the decay is faster.

random fields cannot be evaluated on this problem, as the partition function cannot be computed for graphs with many cycles. However, the MAP estimate can be efficiently approximated using MPLP.

Our learning objective function is $\min_\theta \sum_{(x,y)\in S} \frac{1}{|S|} \mathbb{E}_\gamma[\max\{\sum_i \theta_i \phi_i(x, y_i) + \sum_{i,j} \theta_{i,j} \phi_{i,j}(x, y_i, y_j) + \sum_i \gamma_i(y_i)\}] + \|\theta\|^2$, which serves as an upper bound to the Conditional random fields learning objective. When omitting the perturbations this learning objective is the structured-SVM objective (without label loss).

We estimated the expected max-perturbation value by evaluating 5 MAP predictions with random perturbation. We performed gradient decent and stopped either when the gradient step did not improve the objective (decreasing the learning rate by half for 10 times) or when the algorithm performed 30 iterations. We compared to structured-SVM by removing the perturbations from our learning algorithm (cf. [67]). Since structured-SVM is a non-smooth program, we used subgradient decent for $10,000$ iterations and learning rate of $10/\sqrt{T}$. For completeness, we note that in our previous evaluation, we ran structured-SVM for 30 iterations and the results were significantly worse [1]. The running time of the learning algorithms is dominated by the number of MAP evaluations, which is 150 for learning with random perturbations and $10,000$ with structured-SVM. To evaluate the two learning algorithms, we used MAP prediction on the test data.[5] When learning with MAP perturbations, the pixel based error on the test set was 1.8%. When learning without perturbations, i.e., with structured-SVMs, the pixel based error

[5]Shpakova and Bach have recently shown that one can improve the test performance by inferring the probabilities of MAP perturbations [68].

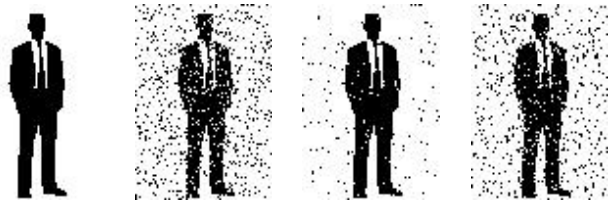


Fig. 10. *From left to right: (a) Binary* 100×70 *image. (b) A representative image in the training set and the test set, where* 10% *of the pixels are randomly flipped. (c) A de-noised test image with our method: The test set error is* 1.8%*. (d) A de-noised test image with SVM-struct: The pixel base error is* 2.5%*.*

on the test set was 2.5%.

VI. Conclusions

High dimensional inference is a key challenge for applying machine learning in real-life applications. While the Gibbs distribution is widely used throughout many areas of research, standard sampling algorithms may be too slow in many cases of interest. In the last few years many optimization algorithms were devised to avoid the computational burden of sampling and instead researchers predicted the most likely (MAP) solution. In this work we explore novel probability models that rely on MAP optimization as their core element. These models measure the robustness of MAP prediction to random shifts of the potential function. We show how to sample from the Gibbs distribution using the expected value of perturb-max operations. We also derive new entropy bounds for perturb-max models that use the expected value of the maximal perturbations. We complete our exploration by investigating the measure concentration of perturb-max value while showing it can be estimated with only a few perturb-max operations.

The results here can be extended in a number of different directions. In contrast to tree re-weighted or entropy covering bounds, the perturb-max bounds do not have a straightforward tightening scheme. Another direction is to consider the perturb-max model beyond the first moment (expectation). It remains open whether the variance or other related statistics of the perturb-max value can be beneficial for learning, e.g., learning the correlations between data measurements. Understanding the effect of approximate MAP solvers could extend the range of applicability [31], [69]. A natural extension of these methods is to consider high dimensional learning. Perturb-max models already appear implicitly in risk analysis [32] and online learning [33]. Novel approaches that consider perturb-max models explicitly [24], [31] may derive new learning paradigms for high-dimensional inference.

APPENDIX

Proof details for Corollary 9

The result in Corollary 9 follows by taking $C = 4$ in the following lemma.

Lemma 4: For any $a > 0$ and $C > 0$, for $\lambda \in [0, 2/a\sqrt{C}]$

$$\prod_{i=0}^{\infty}\left(1-\frac{\lambda^2 a^2 C}{4^{i+1}}\right)^{-2^i} \leq \frac{2+\lambda a\sqrt{C}}{2-\lambda a\sqrt{C}}. \tag{77}$$

Proof: To prove this inequality there are three simple steps. First, factor out the first term:

$$\prod_{i=0}^{\infty}\left(1-\frac{\lambda^2 a^a C}{4^{i+1}}\right)^{-2^i} = \left(1-\frac{\lambda^2 a^2 C}{4}\right)^{-1}\prod_{i=1}^{\infty}\left(1-\frac{\lambda^2 a^2 C}{4^{i+1}}\right)^{-2^i} \tag{78}$$

and define

$$V(\lambda) = \prod_{i=1}^{\infty}\left(1-\frac{\lambda^2 a^2 C}{4^{i+1}}\right)^{-2^i}. \tag{79}$$

Next, from the identity

$$\left(1-\frac{\lambda^2 a^2 C}{4}\right)^{-1} = \frac{4}{(2+\lambda a\sqrt{C})(2-\lambda a\sqrt{C})}. \tag{80}$$

If we can show that $\sqrt{V(\lambda)} < \frac{2+\lambda a\sqrt{C}}{2}$ then the result will follow.

We claim $\sqrt{V(\lambda)}$ is convex. Note that if $V(\lambda)$ log-convex, then $\sqrt{V(\lambda)}$ is also convex, so it is sufficient to show that $\log V(\lambda)$ is convex. Using the Taylor series expansion $\log(1-x) = -\sum_{j=1}^{\infty} x^j/j$ and switching the order of summation,

$$\log V(\lambda) = -\sum_{i=1}^{\infty} 2^i \log\left(1-\frac{\lambda^2 a^2 C}{4^{i+1}}\right) \tag{81}$$

$$= \sum_{i=1}^{\infty} 2^i \sum_{j=1}^{\infty} \frac{(\lambda^2 a^2 C)^j}{j\cdot 4^{ji+j}} \tag{82}$$

$$= \sum_{j=1}^{\infty} \frac{(\lambda^2 a^2 C)^j}{j4^j} \sum_{i=1}^{\infty} \frac{2^i}{2^{(2j-1)i}} \tag{83}$$

$$= \sum_{j=1}^{\infty} \frac{(\lambda^2 a^2 C)^j}{j4^j} \left(\frac{1}{1-2^{-(2j-1)}} - 1\right) \tag{84}$$

$$= \sum_{j=1}^{\infty} \frac{(\lambda^2 a^2 C)^j}{j4^j} \left(\frac{1}{2^{2j-1}-1}\right). \tag{85}$$

Note that the expansion holds only for $\frac{\lambda^2 a^2 C}{4^{i+1}} < 1$ and this bound is tightest for $i = 1$. This expansion is the sum of convex functions and hence convex. This means that for $\lambda < \frac{4}{a\sqrt{C}}$ the function $\sqrt{V(\lambda)}$ is convex.

At $\lambda = \frac{2}{a\sqrt{C}}$, we have

$$\log V\left(\frac{2}{a\sqrt{C}}\right) = \sum_{j=1}^{\infty} \frac{1}{j}\left(\frac{1}{2^{2j-1}-1}\right) \tag{86}$$

$$\leq 1 + \sum_{j=2}^{\infty} \frac{1}{j \cdot 2^{2j-2}} \tag{87}$$

$$= 1 + \sum_{j=2}^{\infty} \frac{4}{j}\frac{(1/4)^j}{j} \tag{88}$$

$$= 1 + 4\left(-\log\left(1-\frac{1}{4}\right) - \frac{1}{4}\right) \tag{89}$$

$$= 4\log\frac{4}{3} \tag{90}$$

$$< \log 4. \tag{91}$$

Therefore $V(2/a\sqrt{C}) < 4$.

Since $V(0) = 1$ and $V(2/a\sqrt{C}) < 4$, by convexity, for $\lambda \in [0, 2/a\sqrt{C}]$,

$$\sqrt{V(\lambda)} \leq \left(1 - \frac{\lambda a\sqrt{C}}{2}\right)\sqrt{V(\lambda)} + \frac{\lambda a\sqrt{C}}{2}\sqrt{V\left(\frac{2}{a\sqrt{C}}\right)} \tag{92}$$

$$< 1 + \frac{\lambda a\sqrt{C}}{2} \tag{93}$$

$$= \frac{2 + \lambda a\sqrt{C}}{2}. \tag{94}$$

Now, considering (78) and the terms in (79) and (80), we have

$$\prod_{i=0}^{\infty}\left(1 - \frac{\lambda^2 a^a C}{4^{i+1}}\right)^{-2^i} = \frac{4}{(2+\lambda a\sqrt{C})(2-\lambda a\sqrt{C})}V(\lambda) \tag{95}$$

$$< \frac{4}{(2+\lambda a\sqrt{C})(2-\lambda a\sqrt{C})} \cdot \left(\frac{2+\lambda a\sqrt{C}}{2}\right)^2 \tag{96}$$

$$= \frac{2+\lambda a\sqrt{C}}{2-\lambda a\sqrt{C}}, \tag{97}$$

as desired. ■

ACKNOWLEDGMENTS

The authors thank the reviewers for their detailed and helpful comments which helped considerably in clarifying the manuscript, Francis Bach and Tatiana Shpakova for helpful discussions, and Associate Editor Constantine Caramanis for his patience and understanding.

REFERENCES

[1] T. Hazan and T. Jaakkola, "On the partition function and random maximum a-posteriori perturbations," in *The 29th International Conference on Machine Learning (ICML 2012)*, 2012.

[2] T. Hazan, S. Maji, and T. Jaakkola, "On sampling from the Gibbs distribution with random maximum a-posteriori perturbations," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 1268–1276. [Online]. Available: http://papers.nips.cc/paper/4962-on-sampling-from-the-gibbs-distribution-with-random-maximum-a-posteriori-perturbations

[3] F. Orabona, T. Hazan, A. Sarwate, and T. Jaakkola, "On measure concentration of random maximum a-posteriori perturbations," in *Proceedings of The 31st International Conference on Machine Learning*, ser. JMLR: Workshop and Conference Proceedings, E. P. Xing and T. Jebara, Eds., vol. 32, 2014, p. 1. [Online]. Available: http://jmlr.csail.mit.edu/proceedings/papers/v32/orabona14.html

[4] S. Maji, T. Hazan, and T. Jaakkola, "Active boundary annotation using random MAP perturbations," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. JMLR: Workshop and Conference Proceedings, S. Kaski and J. Corander, Eds., vol. 33, 2014, pp. 604–613. [Online]. Available: http://jmlr.org/proceedings/papers/v33/maji14.html

[5] P. F. Felzenszwalb and R. Zabih, "Dynamic programming and graph algorithms in computer vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 4, pp. 721–740, April 2011. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2010.135

[6] T. Koo, A. Rush, M. Collins, T. Jaakkola, and D. Sontag, "Dual decomposition for parsing with non-projective head automata," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, 2010, pp. 1288–1298. [Online]. Available: http://dl.acm.org/citation.cfm?id=1870783

[7] D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss, "Tightening LP relaxations for MAP using message passing," in *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*. Corvallis, Oregon, USA: AUAI Press, 2008, pp. 503–510. [Online]. Available: https://dslpitt.org/papers/08/p503-sontag.pdf

[8] M. Jerrum and A. Sinclair, "Polynomial-time approximation algorithms for the Ising model," *SIAM Journal on computing*, vol. 22, no. 5, pp. 1087–1116, October 1993. [Online]. Available: http://dx.doi.org/10.1137/0222066

[9] L. A. Goldberg and M. Jerrum, "The complexity of ferromagnetic Ising with local fields," *Combinatorics Probability and Computing*, vol. 16, no. 1, p. 43, January 2007. [Online]. Available: http://dx.doi.org/10.1017/S096354830600767X

[10] ——, "Approximating the partition function of the ferromagnetic potts model," *Journal of the ACM (JACM)*, vol. 59, no. 5, p. 25, 2012.

[11] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, November 1984. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.1984.4767596

[12] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, April 1970. [Online]. Available: http://dx.doi.org/10.1093/biomet/57.1.97

[13] R. H. Swendsen and J.-S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," *Physical Review Letters*, vol. 58, no. 2, pp. 86–88, January 1987. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.58.86

[14] J. M. Eisner, "Three new probabilistic models for dependency parsing: an exploration," in *Proceedings of the 16th*

Conference on Computational Linguistics (COLING '96), vol. 1. Association for Computational Linguistics, 1996, pp. 340–345. [Online]. Available: http://dx.doi.org/10.3115/992628.992688

[15] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, November 2001. [Online]. Available: http://dx.doi.org/10.1109/34.969114

[16] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1568–1583, October 2006. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2006.200

[17] Gurobi Optimization. (2015) Gurobi optimizer documentation. [Online]. Available: http://www.gurobi.com/documentation/

[18] P. Swoboda, B. Savchynskyy, J. Kappes, and C. Schnörr, "Partial optimality via iterative pruning for the Potts model," in *Scale Space and Variational Methods in Computer Vision: 4th International Conference*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2013, vol. 7893, ch. 40, pp. 477–488. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-38267-3

[19] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "MAP estimation via agreement on trees: Message-passing and linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 3697–3717, November 2005. [Online]. Available: http://dx.doi.org/10.1109/TIT.2005.856938

[20] Y. Weiss, C. Yanover, and T. Meltzer, "MAP estimation, linear programming and belief propagation with convex free energies," in *Proceedings of the Twenty-Third Conference Conference on Uncertainty in Artificial Intelligence (2007)*. Corvallis, Oregon, USA: AUAI Press, 2007, pp. 416–425. [Online]. Available: https://dslpitt.org/papers/07/p416-weiss.pdf

[21] T. Werner, "High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (MAP-MRF)," in *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2008.4587355

[22] J. Peng, T. Hazan, N. Srebro, and J. Xu, "Approximate inference by intersecting semidefinite bound and local polytope," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, ser. JMLR: Workshop and Conference Proceedings, N. Lawrence and M. Girolami, Eds., vol. 22, 2012, pp. 868–876. [Online]. Available: http://jmlr.csail.mit.edu/proceedings/papers/v22/peng12.html

[23] G. Papandreou and A. Yuille, "Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models," in *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*, Barcelona, Spain, November 2011, pp. 193–200. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2011.6126242

[24] D. Tarlow, R. P. Adams, and R. S. Zemel, "Randomized optimum models for structured prediction," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, ser. JMLR: Workshop and Conference Proceedings, N. Lawrence and M. Girolami, Eds., vol. 22, 2012, pp. 1221–1229. [Online]. Available: http://jmlr.org/proceedings/papers/v22/tarlow12b.html

[25] S. Ermon, C. Gomes, A. Sabharwal, and B. Selman, "Taming the curse of dimensionality: Discrete integration by hashing and optimization," in *Proceedings of The 30th International Conference on Machine Learning*, ser. JMLR: Workshop and Conference Proceedings, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 2, 2013, pp. 334–342. [Online]. Available: http://jmlr.org/proceedings/papers/v28/ermon13.html

[26] ——, "Optimization with parity constraints: From binary codes to discrete integration," in *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*. Corvallis, Oregon: AUAI Press, 2013, pp. 202–211.

[27] 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 26 (NIPS 2013)*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2085–2093. [Online]. Available: http://papers.nips.cc/paper/4965-embed-and-project-discrete-sampling-with-universal-hashing.pdf

[28] ——, "Low-density parity constraints for hashing-based discrete integration," in *Proceedings of The 31st International Conference on Machine Learning*, ser. JMLR: Workshop and Conference Proceedings, E. P. Xing and T. Jebara, Eds., vol. 32, no. 1, 2014, pp. 271–279. [Online]. Available: http://jmlr.org/proceedings/papers/v32/ermon14.html

[29] C. Maddison, D. Tarlow, and T. Minka, "A* sampling," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2085–2093. [Online]. Available: http://papers.nips.cc/paper/5449-a-sampling

[30] G. Papandreou and A. Yuille, "Perturb-and-MAP random fields: Reducing random sampling to optimization, with applications in computer vision," in *Advanced Structured Prediction*, S. Nowozin, P. V. Gehler, J. Jancsary, and C. H. Lampert, Eds. Cambridge, MA, USA: MIT Press, 2014, ch. 7, pp. 159–186.

[31] A. Gane and T. S. J. Tamir Hazan, "Learning with maximum a-posteriori perturbation models," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. JMLR: Workshop and Conference Proceedings, S. Kaski and J. Corander, Eds., vol. 33, 2014, pp. 247—256. [Online]. Available: http://jmlr.org/proceedings/papers/v33/gane14.html

[32] J. Keshet, D. McAllester, and T. Hazan, "PAC-Bayesian approach for minimization of phoneme error rate," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2224–2227. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2011.5946923

[33] A. Kalai and S. Vempala, "Efficient algorithms for online decision problems," *Journal of Computer and System Sciences*, vol. 71, no. 3, pp. 291–307, October 2005. [Online]. Available: http://dx.doi.org/10.1016/j.jcss.2004.10.016

[34] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008. [Online]. Available: http://dx.doi.org/10.1561/2200000001

[35] S. Kotz and S. Nadarajah, *Extreme value distributions: theory and applications*. London, UK: Imperial College Press, 2000.

[36] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2003.

[37] L. G. Valiant, "The complexity of computing the permanent," *Theoretical Computer Science*, vol. 8, no. 2, pp. 189–201, 1979. [Online]. Available: http://dx.doi.org/10.1016/0304-3975(79)90044-6

[38] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, February 2004. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2004.1262177

[39] A. Rush, D. Sontag, M. Collins, and T. Jaakkola, "On dual decomposition and linear programming relaxations for natural language processing," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, 2010, pp. 1–11. [Online]. Available: http://dl.acm.org/citation.cfm?id=1870659

[40] A. G. Schwing and R. Urtasun, "Efficient exact inference for 3D indoor scene understanding," in *Computer Vision – ECCV 2012 : 12th European Conference on Computer Vision*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2012, vol. 7577, ch. 22, pp. 299–313. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33783-3

[41] M. Sun, M. Telaprolu, H. Lee, and S. Savarese, "An efficient branch-and-bound algorithm for optimal human pose

estimation," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, 2012, pp. 1616–1623. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2012.6247854

[42] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, September 2010. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2009.167

[43] A. Globerson and T. S. Jaakkola, "Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Curran Associates, Inc., 2007, vol. 21, pp. 553–560. [Online]. Available: http://papers.nips.cc/paper/3200-fixing-max-product-convergent-message-passing-algorithms-for-map-lp-relaxations.pdf

[44] D. Sontag and T. S. Jaakkola, "New outer bounds on the marginal polytope," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Curran Associates, Inc., 2008, pp. 1393–1400. [Online]. Available: http://papers.nips.cc/paper/3274-new-outer-bounds-on-the-marginal-polytope.pdf

[45] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, no. 02, pp. 180–190, April 1928. [Online]. Available: http://dx.doi.org/10.1017/S0305004100015681

[46] B. Gnedenko, "Sur la distribution limite du terme maximum d'une serie aleatoire," *Annals of Mathematics*, vol. 44, no. 3, pp. 423–453, July 1943. [Online]. Available: http://dx.doi.org/10.2307/1968974

[47] E. J. Gumbel, *Statistical theory of extreme values and some practical applications: a series of lectures*, ser. National Bureau of Standards Applied Mathematics Series. Washington, DC, USA: US Govt. Print. Office, 1954, no. 33.

[48] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*. New York, NY, USA: John Wiley and Sons, 1959.

[49] M. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA, USA: MIT press, 1985, vol. 9.

[50] D. McFadden, "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, P. Zarembka, Ed. New York, NY, USA: Academic Press, 1974, ch. 4, pp. 105–142.

[51] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed. New York, NY, USA: John Wiley & Sons, 2013.

[52] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Nashua, NH, USA: Athena Scientific, 2003.

[53] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[54] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "A new class of upper bounds on the log partition function," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2313–2335, July 2005. [Online]. Available: http://dx.doi.org/10.1109/TIT.2005.850091

[55] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: John Wiley & Sons, 2012.

[56] H. J. Brascamp and E. H. Lieb, "On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation," *Journal of Functional Analysis*, vol. 22, no. 4, pp. 366–389, August 1976. [Online]. Available: http://dx.doi.org/10.1016/0022-1236(76)90004-5

[57] S. Aida, T. Masuda, and I. Shigekawa, "Logarithmic Sobolev inequalities and exponential integrability," *Journal of Functional Analysis*, vol. 126, no. 1, pp. 83–101, November 1994. [Online]. Available: http://dx.doi.org/10.1006/jfan.1994.1142

[58] S. Bobkov and M. Ledoux, "Poincaré's inequalities and Talagrand's concentration phenomenon for the exponential distribution," *Probability Theory and Related Fields*, vol. 107, no. 3, pp. 383–400, March 1997. [Online]. Available: http://dx.doi.org/10.1007/s004400050090

[59] M. Ledoux, *The Concentration of Measure Phenomenon*, ser. Mathematical Surveys and Monographs. American Mathematical Society, 2001, vol. 89.

[60] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators*, ser. Grundlehren der mathematischen Wissenschaften. Switzerland: Springer International Publishing, 2014, vol. 348. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-00227-9

[61] V. H. Nguyen, "Dimensional variance inequalities of Brascamp–Lieb type and a local approach to dimensional Prékopas theorem," *Journal of Functional Analysis*, vol. 266, no. 2, pp. 931–955, January 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jfa.2013.11.003

[62] S. G. Bobkov and M. Ledoux, "Weighted Poincaré-type inequalities for Cauchy and other convex measures," *The Annals of Probability*, vol. 37, no. 2, pp. 403–427, 2009. [Online]. Available: http://dx.doi.org/10.1214/08-AOP407

[63] M. R. Spiegel, S. Lipschutz, and J. Liu, *Mathematical Handbook of Formulas and Tables*, 4th ed., ser. Schaum's Outlines. New York, NY, USA: McGraw-Hill Education, 2013.

[64] A. S. Willsky, E. B. Sudderth, and M. J. Wainwright, "Loop series and Bethe variational bounds in attractive graphical models," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Curran Associates, Inc., 2007, pp. 1425–1432. [Online]. Available: http://papers.nips.cc/paper/3354-loop-series-and-bethe-variational-bounds-in-attractive-graphical-models.pdf

[65] N. Ruozzi, "The Bethe partition function of log-supermodular graphical models," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 117–125. [Online]. Available: http://papers.nips.cc/paper/4649-the-bethe-partition-function-of-log-supermodular-graphical-models.pdf

[66] A. Weller and T. Jebara, "Clamping variables and approximate inference," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Cur, 2014, pp. 909–917. [Online]. Available: http://papers.nips.cc/paper/5529-clamping-variables-and-approximate-inference.pdf

[67] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, no. 2, p. 1453, 2006.

[68] T. Shpakova and F. Bach, "Parameter learning for log-supermodular distributions," *arXiv preprint arXiv:1608.05258*, 2016.

[69] T. Hazan, S. Maji, J. Keshet, and T. Jaakkola, "Learning efficient random maximum a-posteriori predictors with non-decomposable loss functions," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 1887–1895. [Online]. Available: http://papers.nips.cc/paper/4962-on-sampling-from-the-gibbs-distribution-with-random-maximum-a-posteriori-perturbations