On Measure Concentration of Random Maximum A-Posteriori Perturbations

Francesco Orabona[*] ORABONA@TTIC.EDU
Toyota Technological Institute at Chicago, 6045 S. Kenwood Ave, Chicago, IL 60637

Tamir Hazan[*] TAMIR@CS.HAIFA.AC.IL
Dept. of Computer Science, University of Haifa, 31905 Haifa, Israel

Anand D. Sarwate[*] ASARWATE@ECE.RUTGERS.EDU
Rutgers University, Dept. of Electrical and Computer Engineering, 94 Brett Road, Piscataway, NJ 08854

Tommi S. Jaakkola TOMMI@CSAIL.MIT.EDU
MIT CSAIL, Stata Center, Bldg 32-G470, 77 Mass Ave. Cambridge, MA 02139

Abstract

The maximum a-posteriori (MAP) perturbation framework has emerged as a useful approach for inference and learning in high dimensional complex models. By maximizing a randomly perturbed potential function, MAP perturbations generate unbiased samples from the Gibbs distribution. Unfortunately, the computational cost of generating so many high-dimensional random variables can be prohibitive. More efficient algorithms use sequential sampling strategies based on the expected value of low dimensional MAP perturbations. This paper develops new measure concentration inequalities that bound the number of samples needed to estimate such expected values. Applying the general result to MAP perturbations can yield a more efficient algorithm to approximate sampling from the Gibbs distribution. The measure concentration result is of general interest and may be applicable to other areas involving Monte Carlo estimation of expectations.

1. Introduction

Modern machine learning tasks in computer vision, natural language processing, and computational biology involve inference in high-dimensional complex models. Examples include scene understanding (Felzenszwalb & Zabih, 2011), parsing (Koo et al., 2010), and protein design (Sontag et al., 2008). In these settings inference involves finding likely structures that fit 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 likelihood of an assignment is based on defining potential functions that account for interactions over these variables. Given the observed data, these likelihoods yield a *posterior probability distribution* on assignments known as the Gibbs distribution. Contemporary posterior probabilities that are used in machine learning incorporate local potential functions on the variables of the model that are derived from the data (high signal) as well as higher order potential functions that account for interactions between the model variables and derived from domain-specific knowledge (high coupling). The resulting posterior probability landscape is often "ragged," and in such landscapes Markov chain Monte Carlo (MCMC) approaches to sampling from the Gibbs distribution may become prohibitively expensive. By contrast, when no data terms (local potential functions) exist, MCMC approaches can be quite successful (e.g., Jerrum et al. (2004); Huber (2003)).

An alternative to sampling from the Gibbs distribution is to look for the *maximum a posteriori probability* (MAP) structure. Substantial effort has gone into developing algorithms for recovering MAP assignments by exploiting domain-specific structural restrictions such as super-modularity (Kolmogorov, 2006) or by linear programming relaxations such as cutting-

[*] *FO, TH and ADS contributed equally to this paper.*

Proceedings of the 31^{st} *International Conference on Machine Learning*, Copyright 2014 by the author(s).

planes (Sontag et al., 2008; Werner, 2008). However, in many contemporary applications, the complex potential functions on a large number of variables yield several likely structures. We want to find these "highly probable" assignments as well, and unfortunately MAP inference returns only a single assignment.

Recent work leverages the current efficiency of MAP solvers to build samplers for the Gibbs distribution, thereby avoiding the computational burden of MCMC methods. These works calculate the MAP structure of a *randomly perturbed potential function.* Such an approach effectively ignores the raggedness of the landscape that hinders MCMC. Papandreou & Yuille (2011) and Tarlow et al. (2012) have shown that randomly perturbing the potential of each structure with an independent random variable that follows the Gumbel distribution and finding the MAP assignment of the perturbed potential function provides an unbiased sample from the Gibbs distribution. Unfortunately the total number of structures, and consequently the total number of random perturbations, is exponential in the structure's dimension. Alternatively, Hazan et al. (2013b) use expectation bounds on the partition function (Hazan & Jaakkola, 2012) to build a sampler for Gibbs distribution using MAP solvers on low dimensional perturbations; the complexity is linear in the dimension of the structures.

The low dimensional samplers require calculating expectations of the value of the MAP solution after perturbing the posterior distribution. In this paper we study the distribution of this perturbed MAP solution. In particular, we prove new measure concentration inequalities that show the expected perturbed MAP value can be estimated with high probability using only a few random samples. This is an important ingredient to construct an alternative to MCMC in the data-knowledge domain that relies on MAP solvers. The key technical challenge comes from the fact that the perturbations are Gumbel random variables, which have support on the entire real line. Thus, standard approaches for bounded random variables, such as McDiarmid's inequality, do not apply. Instead, we derive a new Poincaré inequality for the Gumbel distribution, as well as a modified logarithmic Sobolev inequality using the approach suggested by Bobkov & Ledoux (1997); Ledoux (2001). These results, which are of independent interest, guarantee that the deviation between the sample mean of random MAP perturbations and the expectation has an exponential decay.

2. Problem statement

Notation: Boldface will denote tuples or vectors and calligraphic script sets. For a tuple $\mathbf{x} = (x_1, x_2, \dots, x_n)$, let $\mathbf{x}_{j:k} = (x_j, x_{j+1}, \dots, x_k)$.

2.1. The MAP perturbation framework

Statistical inference problems involve reasoning about the states of discrete variables whose configurations (assignments of values) specify the discrete structures of interest. Suppose that our model has n variables $\mathbf{x} = (x_1, x_2, \dots, x_n)$ where each x_i taking 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 $\mathbf{x}$, where $\theta(\mathbf{x}) = -\infty$ for $\mathbf{x} \notin \mathrm{Dom}(\theta)$. The potential function induces a probability distribution on configurations $\mathbf{x}$ via the Gibbs distribution:

$$p(\mathbf{x}) \triangleq \frac{1}{Z} \exp(\theta(\mathbf{x})), \tag{1}$$

$$Z \triangleq \sum_{\mathbf{x} \in \mathcal{X}} \exp(\theta(\mathbf{x})). \tag{2}$$

The normalization constant Z is called the *partition function.* Sampling from (1) is often difficult because the sum in (2) involves an exponentially large number of terms (equal to the number of discrete structures). In many cases, computing the partition function is in the complexity class $\#P$ (e.g., Valiant (1979)).

Finding the most likely assignment of values to variables is easier. As the Gibbs distribution is typically constructed given observed data, we call this the maximum a-posteriori (MAP) prediction. Maximizing (1):

$$\hat{\mathbf{x}}_{\mathsf{MAP}} = \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \theta(\mathbf{x}). \tag{3}$$

There are many good optimization algorithms for solving (3) in cases of practical interest. Although MAP prediction is still NP-hard in general, it is often simpler than sampling from the Gibbs distribution.

Due to modeling inaccuracies, there are often several meaningful structures $\mathbf{x}$ whose scores $\theta(\mathbf{x})$ are close to $\theta(\hat{\mathbf{x}}_{\mathsf{MAP}})$, and we would like to recover those as well. As an alternative to MCMC methods for sampling from the Gibbs distribution in (1), we can draw samples by perturbing the potential function and solving the resulting MAP problem. The MAP perturbation approach adds a random function $\gamma : \mathcal{X} \to \mathbb{R}$ to the potential function in (1) and solves the resulting MAP problem:

$$\hat{\mathbf{x}}_{\mathsf{R-MAP}} = \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \left\{ \theta(\mathbf{x}) + \gamma(\mathbf{x}) \right\}. \tag{4}$$

The random function $\gamma(\cdot)$ associates a random variable to each $\mathbf{x} \in \mathcal{X}$. 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 $\mathbf{x} \in \mathcal{X}$. We can find the distribution of the randomized MAP predictor in (4) when $\{\gamma(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ are i.i.d.; in particular, suppose each $\gamma(\mathbf{x})$ is a Gumbel random variable with zero mean, variance $\pi^2/6$, and cumulative distribution function

$$G(y) = \exp(-\exp(-(y + c))), \tag{5}$$

where $c \approx 0.5772$ is the Euler-Mascheroni constant. The following result characterizes the distribution of the randomized predictor $\hat{\mathbf{x}}_{\mathsf{R-MAP}}$ in (4).

Theorem 1. *(Gumbel & Lieblein, 1954) Let* $\mathbf{\Gamma} = \{\gamma(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ *be a collection of i.i.d. Gumbel random variables whose distribution is given by* (5)*. Then*

$$\mathbb{P}_{\mathbf{\Gamma}}\left(\hat{\mathbf{x}} = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}\right) = \frac{\exp(\theta(\hat{\mathbf{x}}))}{Z}, \tag{6}$$

$$\mathbb{E}_{\mathbf{\Gamma}}\left[\max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}\right] = \log Z,$$

$$\mathrm{Var}_{\mathbf{\Gamma}}\left[\max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}\right] = \pi^2/6.$$

The max-stability of the Gumbel distribution provides a straightforward approach to generate unbiased samples from the Gibbs distribution – simply generate the perturbations in $\mathbf{\Gamma}$ and solve the problem in (4). These solution may also be used to approximate the partition function in (2), as random samples concentrate around their expectation according to their variance. $F(\mathbf{\Gamma}) = [\max_{\mathbf{x}\in\mathcal{X}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}]$ then

$$\mathbb{P}_{\mathbf{\Gamma}}\left[\Big|F(\mathbf{\Gamma}) - \mathbb{E}[F(\mathbf{\Gamma})]\Big| \geq \epsilon\right] \leq \frac{\mathrm{Var}\, F(\mathbf{\Gamma})}{\epsilon^2}. \tag{7}$$

However, because $\mathbf{\Gamma}$ contains $|\mathcal{X}|$ i.i.d. random variables, this approach to inference has complexity which is exponential in n.

2.2. Sampling from the Gibbs distribution using low dimensional perturbations

Sampling from the Gibbs distribution is inherently tied to estimating the partition function in (2). If we could compute Z exactly, then we could sample x_1 with probability proportional to $\sum_{x_2,\ldots,x_n} \exp(\theta(\mathbf{x}))$, and for each subsequent dimension i, sample x_i with probability proportional to $\sum_{x_{i+1},\ldots,x_n} \exp(\theta(\mathbf{x}))$, yielding a Gibbs sampler. However, this involves computing the partition function (2), which is hard. Instead, Hazan et al. (2013b) use the representation in (6) to derive a family of self-reducible upper bounds on Z and then use these upper bounds in an iterative algorithm that samples from the Gibbs distribution using low dimensional random MAP perturbations. This gives a method which has complexity linear in n.

Algorithm 1 Sampling with low-dimensional random MAP perturbations from the Gibbs distribution (Hazan et al., 2013b)

```
Iterate over $j = 1, ..., n$, while keeping fixed $\mathbf{x}_{1:(j-1)}$

  1. For each $x_j \in \mathcal{X}_j$, set $p_j(x_j) = \frac{\exp(\mathbb{E}_{\mathbf{\Gamma}}[V_{j+1}])}{\exp(\mathbb{E}_{\mathbf{\Gamma}}[V_j])}$, where
     $V_j$ is given by (8)

  2. Set $p_j(r) = 1 - \sum_{x_j \in \mathcal{X}_j} p(x_j)$

  3. Sample an element in $\mathcal{X}_j \cup \{r\}$ according to $p_j(\cdot)$.
     If $r$ is sampled then reject and restart with $j = 1$.
     Otherwise, fix the sampled element $x_j$ and continue the iterations

Output: $\mathbf{x} = (x_1, ..., x_n)$
```

In the following, instead of the $|\mathcal{X}|$ independent random variables in (4), we define the random function $\gamma(\mathbf{x})$ in (4) as the sum of independent random variables for each coordinate x_i of $\mathbf{x}$:

$$\gamma(\mathbf{x}) = \sum_{i=1}^{n} \gamma_i(x_i).$$

This function involves generating $\sum_{i=1}^{n} |\mathcal{X}_i|$ random variables for each i and $x_i \in \mathcal{X}_i$. Let

$$\mathbf{\Gamma} = \bigcup_{i=1}^{n} \{\gamma_i(x_i) : x_i \in \mathcal{X}_i\}$$

be a collection of $\sum_i |\mathcal{X}_i|$ i.i.d. Gumbel random variables with distribution (5). The sampling algorithm in Algorithm 1 uses these random perturbations to draw unbiased samples from the Gibbs distribution. For a fixed $\mathbf{x}_{1:(j-1)} = (x_1, \ldots, x_{j-1})$, define

$$V_j = \max_{\mathbf{x}_{j:n}} \left\{ \theta(\mathbf{x}) + \sum_{i=j}^{n} \gamma_i(x_i) \right\}. \tag{8}$$

The sampler proceeds sequentially – for each j it constructs a distribution $p_j(\cdot)$ on $\mathcal{X}_j \cup \{r\}$, where r indicates a “restart” and attempts to draw an assignment for x_j. If it draws r then it starts over again from $j = 1$, and if it draws an element in $\mathcal{X}_j$ it fixes x_j to that element and proceeds to $j + 1$.

Implementing Algorithm 1 requires estimating the expectations $\mathbb{E}_{\mathbf{\Gamma}}[V_j]$ in (8). In this paper we show how to estimate $\mathbb{E}_{\mathbf{\Gamma}}[V_j]$ and bound the error with high probability by taking the sample mean of M i.i.d. copies of V_j. Specifically, we show that the estimation error decays exponentially with M by proving measure concentration via a modified logarithmic Sobolev inequality for the product of Gumbel random variables. To do so we derive a more general result – a Poincaré inequality for log-concave distributions that may not be log-strongly concave, i.e., for which the second derivative of the exponent is not bounded away from zero.

2.3. Measure concentration

We can think of the maximum value of the perturbed MAP problem as a function of the associated perturbation variables $\mathbf{\Gamma} = \{\gamma_i(x_i) : i \in [n], x_i \in \mathcal{X}_i\}$. There are $m \triangleq |\mathcal{X}_1| + |\mathcal{X}_2| + \cdots + |\mathcal{X}_n|$ i.i.d. random variables in $\mathbf{\Gamma}$. For practical purposes, e.g., to estimate the quality of the sampling algorithm in Algorithm 1, it is important to evaluate the deviation of its sampled mean from its expectation. For notational simplicity we would only describe the deviation of the maximum value of the perturbed MAP from its expectation, namely

$$F(\mathbf{\Gamma}) = V_1 - \mathbb{E}\left[V_1\right]. \tag{9}$$

Since the expectation is a linear function, $\mathbb{E}\left[F\right] = \int F(\mathbf{\Gamma})d\mu(\mathbf{\Gamma}) = 0$ for any measure μ on $\mathbf{\Gamma}$. The Chebyshev's inequality in (7) shows that the deviation of $F(\mathbf{\Gamma})$ is dominated by its variance. Although the variance of a Gumbel random variable is $\pi^2/6$, the variance of low dimensional MAP perturbations does not have an analytic form. However, we may bound it as follows: $\mathrm{Var}[F(\mathbf{\Gamma})] \leq (\max_{\mathbf{x}}\{\theta(\mathbf{x}) + \sum_{i=1}^n \gamma_i(x_i)\})^2 \leq \max_{\mathbf{x}}\{\theta(\mathbf{x})^2\} + n\pi^2/6$. However, this bound on the variance is loose and does not reveal the concentration of measure phenomena of MAP perturbations.

To account for the offsets $\theta(\mathbf{x})$ seamlessly, it is natural to consider measure concentration results that rely on the variation $F(\mathbf{\Gamma})$. For example, the random function $F(\mathbf{\Gamma})$ is a Lipschitz functions, and thus it is concentrated around its mean whenever the random variables in $\mathbf{\Gamma}$ are sampled i.i.d from a (sub-)Gaussian distribution (Pisier, 1986; Ledoux & Talagrand, 1991). Unfortunately, such results do not hold for the Gumbel distribution, for which the density function decays on the positive reals $\mathbb{R}^+$ more slowly than the Gaussian distribution (see Fig. 1). An alternative is to apply measure concentration for Lipschitz random functions of the Laplace distribution (cf. Ledoux (2001)), since the Laplace distribution and the Gumbel distribution decay similarly on the positive real numbers. However, representing $F(\mathbf{\Gamma})$ over Gumbel random variables $\mathbf{\Gamma}$ using the logarithm of an exponential random variables $\hat{\mathbf{\Gamma}}$ results in a compound function over exponential random variables $F(-\log\hat{\mathbf{\Gamma}})$ that is no longer Lipschitz.

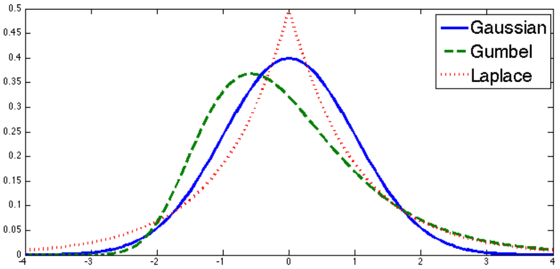

Figure 1. Comparing the decay of Gaussian, Gumbel and Laplace random variables. The decay of the Gaussian random variable is significantly faster and its moments $\Lambda(\lambda)$ always exist. The moments of the Gumbel and Laplace random variables do not always exist (e.g., $\Lambda(1)$).

Although many measure concentration results (such as McDiarmid's inequality) use bounds on the variation of $F(\mathbf{\Gamma})$, MAP perturbations are unbounded, so we derive our result using a different approach. We bound the deviation of $F(\mathbf{\Gamma})$ via its moment generating function

$$\Lambda(\lambda) \triangleq \mathbb{E}\left[\exp(\lambda F)\right]. \tag{10}$$

To wit, for every $\lambda > 0$,

$$\mathbb{P}\left(F(\boldsymbol{\gamma}) \geq r\right) \leq \Lambda(\lambda)/\exp(-\lambda r).$$

More specifically, we construct a differential bound on the $\lambda-$scaled cumulant generating function:

$$H(\lambda) \triangleq \frac{1}{\lambda}\log\Lambda(\lambda). \tag{11}$$

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

$$\lambda\Lambda'(\lambda) - \Lambda(\lambda)\log\Lambda(\lambda) \leq \lambda^2\Lambda(\lambda)\alpha(\lambda). \tag{12}$$

The left side of (12) turns out to be the so-called functional entropy (Ledoux, 2001) of the function $h = \exp(\lambda F)$ with respect to a measure μ:

$$\mathrm{Ent}_\mu(h) \triangleq \int h\log h d\mu - \left(\int h d\mu\right)\log\int h d\mu.$$

Unlike McDiarmid's inequality, this approach provides measure concentration for unbounded functions such those arising from MAP perturbations.

Log-Sobolev inequalities upper-bound the entropy $\mathrm{Ent}_\mu(h)$ in terms of an integral involving $\|\nabla F\|^2$. They are appealing to derive measure concentration results in product spaces, i.e., for functions of subsets of variables $\mathbf{\Gamma}$, because it is sufficient to prove a log-Sobolev inequality on a single variable function f. Given such a scalar result, the additivity property of the entropy (e.g., (Boucheron et al., 2004)) extends the inequality to functions F of many variables. In this work we derive a log-Sobolev inequality for the Gumbel distribution, by bounding the variance of a function by its derivative:

$$\mathrm{Var}_\mu(f) \triangleq \int f^2 d\mu - \left(\int f d\mu \right)^2 \leq C \int |f'|^2 d\mu. \tag{13}$$

This is called a Poincaré inequality, proven originally for the Gaussian case. We prove such an inequality for the Gumbel distribution, which then implies the log-Sobolev inequality and hence measure concentration. We then apply the result to the MAP perturbation framework.

2.4. Related work

We are interested in efficient sampling from the Gibbs distribution in (1) when n is large and the model is complex due to the amount of data and the domain-specific modeling. MCMC approaches to sampling(cf. Koller & Friedman (2009)) may be computationally intractable in such ragged probability landscapes. MAP perturbations use efficient MAP solvers as black box, but the statistical properties of the solutions, beyond Theorem 1, are still being studied. Papandreou & Yuille (2011) consider probability models that are defined by the maximal argument of randomly perturbed potential function, while Tarlow et al. (2012) considers sampling techniques for such models. Keshet et al. (2011) and Hazan et al. (2013a) study the generalization bounds for such models. Rather than focus on the statistics of the solution (the $\mathrm{argmax}_{\mathbf{x}}$) we study statistical properties of the MAP value (the $\max_{\mathbf{x}}$) of the estimate in (4). Other strategies for sampling from the Gibbs distribution using MAP solvers include randomized hashing of Ermon et al. (2013).

Hazan & Jaakkola (2012) used the random MAP perturbation framework to derive upper bounds on the partition function in (2), and Hazan et al. (2013b) derived the unbiased sampler in Algorithm 1. Both of these approaches involve computing an expectation over the distribution of the MAP perturbation – this can be estimated by sample averages. This paper derives new measure concentration results that bound the error of this estimate in terms of the number of samples, making Algorithm 1 practical.

Measure concentration has appeared in many machine learning analyses, most commonly to bound the rate of convergence for risk minimization, either via empirical risk minimization (ERM) (e.g., Bartlett & Mendelson (2003)) or in PAC-Bayesian approaches (e.g., McAllester (2003)). In these applications the function for which we want to show concentration is "well-behaved" in the sense that the underlying random variables are bounded or the function satisfies some bounded-difference or self-bounded conditions conditions, so measure concentration follows from inequalities such as Bernstein (1946), Azuma-Hoeffding (Azuma, 1967; Hoeffding, 1963; McDiarmid, 1989), or Bousquet (2003). However, in our setting, the Gumbel random variables are not bounded, and random perturbations may result in unbounded changes of the perturbed MAP value.

There are several results on measure concentration for Lipschitz functions of *Gaussian* random variables (c.f. the result of Maurey (1979) in Pisier (1986)). In this work we use logarithmic Sobolev inequalities (Ledoux, 2001) and prove a new measure concentration result for *Gumbel* random variables. To do this we generalize a classic result of Brascamp & Lieb (1976) on Poincaré inequalities to non-strongly log-concave distributions, and also recover the concentration result of Bobkov & Ledoux (1997) for functions of Laplace random variables.

3. Concentration of measure

In this section we prove the main technical results of this paper – a new Poincaré inequality for log concave distributions and the corresponding measure concentration result. We will then specialize our result to the Gumbel distribution and apply it to the MAP perturbation framework. Because of the tensorization property of the functional entropy, it is sufficient for our case to prove an inequality like (13) for functions f of a single random variable with measure μ. Proofs are deferred to the supplement due to space considerations.

3.1. A Poincaré inequality for log-concave distributions

Our Theorem 2 in this section generalizes a celebrated result of Brascamp & Lieb (1976, Theorem 4.1) to a wider family of log-concave distributions and strictly improves their result. For an appropriately scaled convex function Q on $\mathbb{R}$, the function $q(y) = \exp(-Q(y))$ defines a density on $\mathbb{R}$ corresponding to a log concave measure μ. Unfortunately, their result is restricted to distributions for which $Q(y)$ is strongly convex. The Gumbel distribution with CDF (5) has density

$$g(y) = \exp\left(-\left(y + c + \exp(-(y + c))\right)\right), \tag{14}$$

and the second derivative of $y+c+\exp(-(y+c))$ cannot be lower bounded by any constant greater than 0, so it is not log-strongly convex.

Theorem 2. *Let μ be a log-concave measure with density $q(y) = \exp(-Q(y))$, where $Q : \mathbb{R} \to \mathbb{R}$ is a convex function satisfying the following conditions:*

- *Q has a unique minimum in a point $y = a$,*
- *Q is twice continuously differentiable in each point of his domain, except possibly in $y = a$,*
- *$Q'(y) \neq 0$ for any $y \neq a$,*
- *$\lim_{y\to a^\pm} Q'(y) \neq 0$ or $\lim_{y\to a^\pm} Q''(y) \neq 0$.*

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function, differentiable almost everywhere, such that

$$\lim_{y\to\pm\infty} f(y)q(y) = 0. \tag{15}$$

Then for any $0 \leq \eta < 1$ such that $\frac{Q''(y)}{|Q'(y)|} + \eta|Q'(y)| \neq 0$ for all $y \in \mathbb{R} \setminus \{a\}$, we have

$$\mathrm{Var}_\mu(f) \leq \frac{1}{1-\eta} \int_{\mathbb{R}} \frac{(f'(y))^2}{Q''(y) + \eta(Q'(y))^2} q(y)dy.$$

Proof. The proof is based on the one in Brascamp & Lieb (1976), but it uses a different strategy in the final critical steps. We first observe that for any $K \in \mathbb{R}$,

$$\mathrm{Var}_\mu(f) \leq \int_{\mathbb{R}} (f(y) - K)^2 d\mu, \tag{16}$$

so we will focus on bounding the right-hand side of (16) for the particular choice of $K = h(a)$.

Let $\tilde{f}(y) \triangleq f(y) - f(a)$ and $U(y) \triangleq \frac{\tilde{f}(y)^2 q(y)}{Q'(y)}$. Note that $d\mu = q(y)dy$. We have that

$$U'(y) = \frac{2\tilde{f}'(y)\tilde{f}(y)q(y)}{Q'(y)} - \tilde{f}(y)^2 q(y)\left(\frac{Q''(y)}{(Q'(y))^2} + 1\right).$$

Rearranging terms and integrating, we see that

$$\int \tilde{f}(y)^2 q(y)dy$$
$$= \int \left(\frac{2\tilde{f}'(y)\tilde{f}(y)}{Q'(y)} - \frac{\tilde{f}(y)^2 Q''(y)}{(Q'(y))^2}\right) q(y)dy - U(y).$$

We now consider the integral between $-\infty$ and a (analogous reasoning holds for the one between a and $+\infty$). We claim that $\lim_{y\to a^-} U(y) = 0$. There are two possible cases: $Q'(a) \neq 0$ and $Q'(a) = 0$. In the first case the claim is obvious, in the second case we have $\lim_{y\to a^-} \frac{\tilde{f}(y)^2}{Q'(y)} = \lim_{y\to a^-} \frac{2f'(y)\tilde{f}(y)}{Q''(y)} = 0$, and anagously for the limit from the left. Using (15) too, we have

$$\int_{-\infty}^{a} \tilde{f}(y)^2 q(y)dy$$
$$= \lim_{\epsilon\to 0^-} \int_{-\infty}^{a+\epsilon} \left(\frac{2\tilde{f}'(y)\tilde{f}(y)}{Q'(y)} - \frac{\tilde{f}(y)^2 Q''(y)}{(Q'(y))^2}\right) q(y)dy$$
$$\leq \lim_{\epsilon\to 0^-} \int_{-\infty}^{a+\epsilon} \left(\frac{2|\tilde{f}'(y)||\tilde{f}(y)|}{|Q'(y)|} - \frac{\tilde{f}(y)^2 Q''(y)}{(Q'(y))^2}\right) q(y)dy$$
$$\leq \lim_{\epsilon\to 0^-} \int_{-\infty}^{a+\epsilon} \left(\frac{\tilde{f}'(y)^2}{Q''(y) + \eta(Q'(y))^2} + \eta\tilde{f}(y)^2\right) q(y)dy,$$

where in the second inequality we used $2\alpha\beta \leq \frac{\alpha^2}{\zeta} + \beta^2\zeta$, for any $\alpha, \zeta \in \mathbb{R}$ and $\zeta > 0$, with $\alpha = |\tilde{f}'(y)|$, $\beta = |\tilde{f}(x)|$, and $\zeta = \frac{Q''(y)}{|Q'(y)|} + \eta|Q'(y)|$. Reasoning in the same way for the interval $[a, +\infty)$, reordering the terms, and using (16), we have the result. □

The main difference between Theorem 2 and the result of Brascamp & Lieb (1976, Theorem 4.1) is that the latter requires the function Q to be strongly convex. Our result holds for non-strongly concave functions including the Laplace and Gumbel distributions. If we take $\eta = 0$ in Theorem 2 we recover the original result of Brascamp & Lieb (1976, Theorem 4.1). For the case $\eta = 1/2$, Theorem 2 yields the Poincaré inequality for the Laplace distribution given in Ledoux (2001). Like the Gumbel distribution, the Laplace distribution is not strongly log-concave and previously required an alternative technique to prove measure concentration (Ledoux, 2001). The following gives a Poincaré inequality for the Gumbel distribution.

Corollary 1. *Let μ be the measure corresponding to the Gumbel distribution and $q(y) = g(y)$ in (14). For any function f that satisfies the conditions in Theorem 2, we have*

$$\mathrm{Var}_\mu(f) \leq 4 \int_{\mathbb{R}} (f'(y))^2 d\mu. \tag{17}$$

Proof. For the Gumbel distribution we have $Q(y) = y + c + \exp(-(y + c))$ in Theorem 2, so

$$Q''(y) + \eta(Q'(y))^2 = e^{-(y+c)} + \eta(1 - e^{-(y+c)})^2.$$

We want an lower bound for all y. Minimizing,

$$e^{-(y+c)} = 2\eta(1 - e^{-(y+c)})e^{-(y+c)}$$

or $e^{-(y+c)} = 1 - \frac{1}{2\eta}$, so the lower bound is $1 - \frac{1}{2\eta} + \frac{1}{4\eta}$ or $\frac{4\eta-1}{4\eta}$ for $\eta > \frac{1}{2}$. For $\eta \leq \frac{1}{2}$,

$$\eta + (1 - 2\eta)e^{-(y+c)} + e^{-2(y+c)} \geq \eta.$$

So $\min\left\{\frac{4\eta}{(4\eta-1)(1-\eta)}, \frac{1}{\eta(1-\eta)}\right\} = 4$ at $\eta = \frac{1}{2}$. Applying Theorem 2 we obtain (17). □

3.2. Measure concentration for the Gumbel distribution

MAP perturbations such as those in (8) are a function of many random variables. We now derive a result based on the Corollary 1 to bound the moment generating function for random variables defined as a function of m random variables. This gives a measure concentration inequality for the product measure μ^m of μ on $\mathbb{R}^m$, where μ corresponds to a scalar Gumbel random variable.

Theorem 3. *Let μ denote the Gumbel measure on $\mathbb{R}$ and let $F : \mathbb{R}^m \to \mathbb{R}$ be a function such that μ^m-almost everywhere we have $\|\nabla F\|^2 \leq a^2$ and $\|\nabla F\|_\infty \leq b$. Furthermore, suppose that for $\mathbf{y} = (y_1, \ldots, y_m)$,*

$$\lim_{y_i \to \pm\infty} F(y_1, \ldots, y_m) \prod_{i=1}^{m} g(y_i) = 0,$$

where $g(\cdot)$ is given by (14). Then, for any $r \geq 0$ and any $|\lambda| \leq \frac{1}{10b}$, we have

$$\mathbb{E}[\exp(\lambda(F - \mathbb{E}[F]))] \leq \exp(5a^2\lambda^2).$$

Proof. For each $i = 1, 2 \ldots, m$, we can think of F as a scalar function f_i of its i-th argument for $i = 1, \ldots, m$. Using Theorem 5.14 of Ledoux (2001) and Corollary 1, for any $|\lambda|b \leq \rho \leq 1$,

$$\begin{aligned}&\mathrm{Ent}_{\mu_i}(\exp(\lambda f_i))\\ &\quad\leq 2\lambda^2 \left(\frac{1+\rho}{1-\rho}\right)^2 \exp(2\sqrt{5}\rho) \int |\partial_i F|^2 d\mu_i.\end{aligned}$$

We now use Proposition 5.13 in Ledoux (2001) to tensorize the entropy by summing over $i = 1$ to m:

$$\begin{aligned}&\mathrm{Ent}_{\mu^m}(\exp(\lambda f_i))\\ &\leq 2\lambda^2 \left(\frac{1+\rho}{1-\rho}\right)^2 \exp(2\sqrt{5}\rho) \int \sum_{i=1}^{m} |\partial_i F|^2 \exp(\lambda F) d\mu^m\\ &\leq 2\lambda^2 \left(\frac{1+\rho}{1-\rho}\right)^2 \exp(2\sqrt{5}\rho) a^2 \int \exp(\lambda f) d\mu^m.\end{aligned}$$

Hence, choosing $\rho = \frac{1}{10}$, we obtain, for any $|\lambda| \leq \frac{1}{10b}$

$$\mathrm{Ent}_{\mu^m}(\exp(\lambda F)) \leq 5a^2\lambda^2 \mathbb{E}_{\mu^m}[\exp(\lambda F)]. \tag{18}$$

Recall the moment generating function in (10) and $\lambda-$scaled cumulant generating function in (11), and note that $H(0) = \mathbb{E}[F]$. We now use Herbst's argument (Ledoux, 2001). Using (18) we have

$$H'(\lambda) = \frac{\mathrm{Ent}_{\mu^m}(\exp(\lambda F))}{\lambda^2 \Lambda(\lambda)} \leq 5a^2. \tag{19}$$

Integrating (19) we get

$$H(\lambda) \leq H(0) + 5a^2\lambda = \mathbb{E}[F] + 5a^2\lambda,$$

Now, from the definition of $H(\lambda)$, this implies

$$\log \mathbb{E}[\exp(\lambda F)] \leq \lambda \mathbb{E}[F] + 5a^2\lambda^2 \ .$$ □

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.

Corollary 2. *Consider the same assumptions of Theorem 3. Let $\eta_1, \eta_2, \ldots, \eta_M$ be M i.i.d. random variables with the same distribution as F. Then with probability at least $1 - \delta$,*

$$\frac{1}{M}\sum_{j=1}^{M} \eta_j - \mathbb{E}[F] \leq \max\left(\frac{20b}{M}\log\frac{1}{\delta}, \sqrt{\frac{20a^2}{M}\log\frac{1}{\delta}}\right).$$

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

$$\begin{aligned}&\mathbb{P}\left(\sum_{j=1}^{M} \eta_j \leq M\mathbb{E}[F] + Mr\right)\\ &\qquad\leq \exp(-M\mathbb{E}[F] - Mr) \prod_{j=1}^{M} \mathbb{E}[\exp(\lambda\eta_j)].\end{aligned}$$

Applying Theorem 3, we have, for any $|\lambda| \leq \frac{1}{10b}$,

$$\mathbb{P}\left(\frac{1}{M}\sum_{j=1}^{M} \eta_j \leq \mathbb{E}[F] + r\right) \leq \exp(M(5a^2\lambda^2 - \lambda r)).$$

Optimizing over λ subject to $|\lambda| \leq \frac{1}{10b}$ we obtain

$$\exp(M(5a^2\lambda^2 - \lambda r)) \leq \exp\left(-\frac{M}{20}\min\left(\frac{r}{b}, \frac{r^2}{a^2}\right)\right).$$

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

3.3. Application to MAP perturbations

To apply these results to the MAP perturbation problem we must calculate the parameters in the bound given by the Corollary 2. Let $F(\mathbf{\Gamma})$ be the random MAP perturbation as defined in (9). This is a function of $m \triangleq \sum_{i=1}^n |\mathcal{X}_i|$ i.i.d. Gumbel random variables. The (sub)gradient of this function is structured and points toward the $\gamma_i(x_i)$ corresponding to the maximizing assignment in $\hat{\mathbf{x}}_{\mathsf{R-MAP}}$ defined in (4), when $\gamma(\mathbf{x}) = \sum_{i=1}^n \gamma_i(x_i)$. That is,

$$\frac{\partial F(\mathbf{\Gamma})}{\partial \gamma_i(x_i)} = \begin{cases} 1 & \text{if } x_i \in \hat{\mathbf{x}}_{\mathsf{R-MAP}} \\ 0 & \text{otherwise} \end{cases} .$$

We have $\|\nabla F\|^2 = n$ and $\|\nabla F\|_\infty = 1$ almost everywhere, so $a^2 = n$ and $b = 1$. Suppose we sample M i.i.d. copies $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \ldots, \mathbf{\Gamma}_M$ copies of $\mathbf{\Gamma}$ and estimate the deviation from the expectation by $\frac{1}{M}\sum_{i=1}^M F(\mathbf{\Gamma}_i)$. We can apply Corollary 2 to both F and $-F$ to get the following double-sided bound with probability $1-\delta$:

$$\left| \frac{1}{M} \sum_{i=1}^{M} F(\mathbf{\Gamma}_i) \right| \leq \max\left(\frac{20}{M} \log\frac{2}{\delta}, \sqrt{\frac{20n}{M}\log\frac{2}{\delta}} \right) .$$

Thus this result gives an estimate of the MAP perturbation $\mathbb{E}\left[\max_{\mathbf{x}} \{\theta(\mathbf{x}) + \sum_{i=1}^n \gamma_i(x_i)\}\right]$ that holds in high probability.

This result can also be applied to estimate the quality of Algorithm 1, which samples from the Gibbs distribution using MAP solvers. To do so, let F equal V_j from (8). This is a function of $m_j \triangleq \sum_{i=j}^n |\mathcal{X}_i|$ i.i.d. Gumbel random variables whose gradient satisfies $\|\nabla V_j\|^2 = n - j + 1$ and $\|\nabla V_j\|_\infty = 1$ almost everywhere, so $a^2 = n - j + 1$ and $b = 1$. Suppose $U = V_j - \mathbb{E}[V_j]$ is a random variable that measures the deviation of V_j from its expectation, and assume we sample M_j i.i.d. random variable $U_1, U_2, \ldots, U_{M_j}$. We then estimate this deviation by the sample mean $\frac{1}{M_j}\sum_{i=1}^{M_j} U_i$. Applying Corollary 2 to both V_j and $-V_j$ to get the following bound with probability $1-\delta$:

$$\left| \frac{1}{M_j} \sum_{i=1}^{M_j} U_i \right| \leq \max\left(\frac{20}{M_j} \log\frac{2}{\delta}, \sqrt{\frac{20(n-j+1)}{M_j}\log\frac{2}{\delta}} \right) . \quad (20)$$

For each j in Algorithm 1, we must estimate $|\mathcal{X}_j|$ expectations $\mathbb{E}_{\mathbf{\Gamma}}[V_{j+1}]$, for a total at most m expectation estimates. For any $\epsilon > 0$ we can choose $\{M_j : j = 1, \ldots, n\}$ so that the right side of (20) is at most ϵ for each j with probability $1 - n\delta$. Let $\hat{p}_j(x_j)$ be the ratio estimated in the first step of Algorithm 1, and $\delta' = n\delta$. Then with probability $1 - \delta'$, for all $j = 1, 2, \ldots, n$, $\frac{\exp(\mathbb{E}[V_{j+1}]-\epsilon)}{\exp(\mathbb{E}[V_j]+\epsilon)} \leq \hat{p}_j(x_j) \leq \frac{\exp(\mathbb{E}[V_{j+1}]+\epsilon)}{\exp(\mathbb{E}[V_j]-\epsilon)}$, or

$$\exp(-2\epsilon) \leq \frac{\hat{p}_j(x_j)}{p_j(x_j)} \leq \exp(2\epsilon).$$

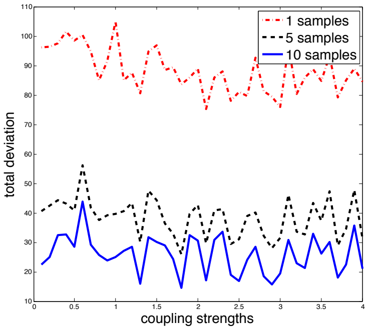

Figure 2. Error of the sample mean versus coupling strength. With only 10 samples one can estimate the expectation well.

4. Experiments

We evaluated our approach on a 100×100 spin glass model with $n = 10^4$ variables, for which

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

where $x_i \in \{-1, 1\}$. Each spin has a local field parameter $\theta_i(x_i) = \theta_i x_i$ and interacts in a grid shaped graphical structure 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 since adjacent variables give higher values to positively correlated configurations. We used low dimensional random perturbations $\gamma(\mathbf{x}) = \sum_{i=1}^n \gamma_i(x_i)$.

The local field parameters θ_i were drawn uniformly at random from $[-1, 1]$ to reflect high signal. The parameters $\theta_{i,j}$ were drawn uniformly from $[0, c]$, where $c \in [0, 4]$ to reflect weak, medium and strong coupling potentials. As these spin glass models are attractive, we are able to use the graph-cuts algorithm (Kolmogorov (2006)) to compute the MAP perturbations efficiently. Throughout our experiments we evaluated the expected value of $F(\mathbf{\Gamma})$ with 100 different samples of $\mathbf{\Gamma}$. We note that we have two random variables $\gamma_i(x_i)$ for each of the spins in the 100×100 model, thus $\mathbf{\Gamma}$ consists of $m = 2 * 10^4$ random variables.

Figure 2 shows the error in the sample mean $\frac{1}{M}\sum_{k=1}^M F(\mathbf{\Gamma}_k)$ versus the coupling strength for three

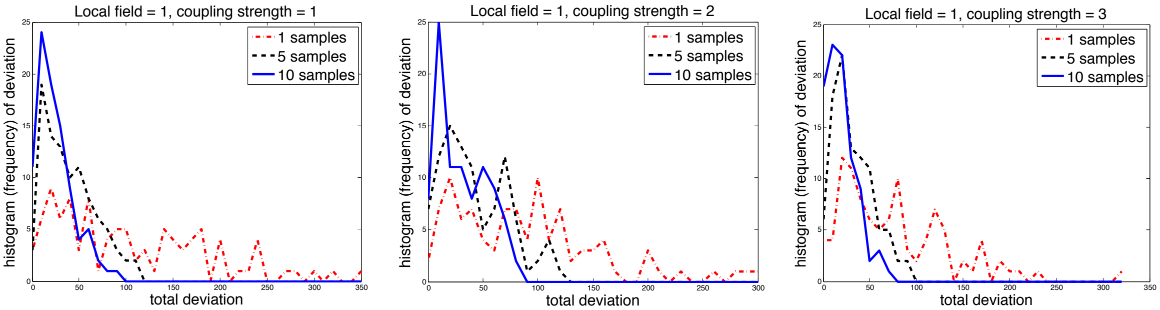

Figure 3. Histogram of MAP values for the 100 × 100 spin glass model.

different sample sizes $M = 1, 5, 10$. The error reduces rapidly as M increases; only 10 samples are needed to estimate the expected value of a random MAP perturbation with 10^4 variables. To test our measure concentration result, that ensures exponential decay, we measure the deviation of the sample mean from its expectation by using $M = 1, 5, 10$ samples. Figure 3 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.

5. Conclusion

Sampling from the Gibbs distribution is important because it helps find near-maxima in the "ragged" posterior probability landscapes typically encountered in the high dimensional complex models. MCMC approaches are inefficient in such settings due to domain-specific modeling (coupling) and the influence of data (signal). However, sampling based on MAP perturbations ignores the ragged landscape. In this paper we characterized the statistics of MAP perturbations.

The low-dimensional MAP perturbation technique requires estimating the expected value of the quantities V_j under the perturbations. We derived high-probability estimates of these expectations that allow estimation with arbitrary precision. This followed from more general results on measure concentration for functions of Gumbel random variables and a Poincaré inequality for non-strongly log-concave distributions. These results may be of use in other applications.

Our results can be extended in several ways. The connection between MAP perturbations and PAC-Bayesian generalization bounds suggests that we may be able to show PAC-Bayesian bounds for unbounded loss functions. Such loss functions may exclude certain configurations and are already used implicitly in vision applications such as interactive segmentation. More generally, Poincaré inequalities relate the variance of a function and its derivatives. Our result may suggest new stochastic gradient methods that control variance via controlling gradients. This connection between variance and gradients may be useful in the analysis of other learning algorithms and applications.

References

Azuma, K. Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal*, 19(3):357–367, 1967.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2003.

Bernstein, S. *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow, 1946.

Bobkov, S. and Ledoux, M. Poincaré's inequalities and Talagrand's concentration phenomenon for the exponential measure. *Probability Theory and Related Fields*, 107(3): 383–400, March 1997.

Boucheron, S., Lugosi, G., and Bousquet, O. Concentration inequalities. In *Advanced Lectures on Machine Learning*, pp. 208–240. Springer, 2004.

Bousquet, O. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, pp. 213–247. Springer, 2003.

Brascamp, H. J. and Lieb, E. H. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *J. Func. Analysis*, 22(4):366 – 389, August 1976.

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

Felzenszwalb, P.F. and Zabih, R. Dynamic programming and graph algorithms in computer vision. *IEEE Trans. PAMI*, 33(4):721–740, 2011.

Gumbel, E. J. and Lieblein, J. *Statistical theory of extreme values and some practical applications: a series of lectures*. Number 33 in National Bureau of Standards Applied Mathematics Series. US Govt. Print. Office, Washington, DC, 1954.

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

Hazan, T., Maji, S., J., Keshet, and Jaakkola, T. Learning efficient random maximum a-posteriori predictors with non-decomposable loss functions. *Advances in Neural Information Processing Systems*, 2013a.

Hazan, T., Maji, S., and Jaakkola, T. On sampling from the gibbs distribution with random maximum a-posteriori perturbations. *Advances in Neural Information Processing Systems*, 2013b.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *JASA*, 58(301):13–30, March 1963.

Huber, M. A bounding chain for swendsen-wang. *Random Structures & Algorithms*, 22(1):43–59, 2003.

Jerrum, M., Sinclair, A., and Vigoda, E. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *JACM*, 51(4):671–697, 2004.

Keshet, J., McAllester, D., and Hazan, T. PAC-Bayesian approach for minimization of phoneme error rate. In *ICASSP*, 2011.

Koller, D. and Friedman, N. *Probabilistic graphical models*. MIT press, 2009.

Kolmogorov, V. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 28(10), 2006.

Koo, T., Rush, A.M., Collins, M., Jaakkola, T., and Sontag, D. Dual decomposition for parsing with non-projective head automata. In *EMNLP*, 2010.

Ledoux, M. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

Maurey, B. Construction de suites symétriques. *Comptes Rendus de l'Académie des Sciences, Paris, Série A-B*, 288:A679–681, 1979.

McAllester, D. Simplified PAC-Bayesian margin bounds. *Learning Theory and Kernel Machines*, pp. 203–215, 2003.

McDiarmid, C. On the method of bounded differences. In *Surveys in Combinatorics*, number 141 in London Mathematical Society Lecture Note Series, pp. 148–188. Cambridge University Press, Cambridge, 1989.

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

Pisier, G. Probabilistic methods in the geometry of Banach spaces. In *Probabilty and Analysis, Varenna (Italy) 1985*, volume 1206 of *Lecture Notes in Mathematics*. Springer, Berlin, 1986.

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

Tarlow, D., Adams, R.P., and Zemel, R.S. Randomized optimum models for structured prediction. In *AISTATS*, pp. 21–23, 2012.

Valiant, L.G. The complexity of computing the permanent. *Theoretical computer science*, 8(2):189–201, 1979.

Werner, T. High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (MAP-MRF). In *CVPR*, 2008.