Statistics and Probability Letters 115 (2016) 8–15

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On some properties of the low-dimensional Gumbel perturbations in the Perturb-and-MAP model

Jakub M. Tomczak*

Department of Computer Science, Faculty of Computer Science and Management, Wrocław University of Science and Technology, wybrzeże Wyspiańskiego 27, 50-370, Wrocław, Poland

ARTICLE INFO

Article history:
Received 11 March 2015
Received in revised form 29 March 2016
Accepted 29 March 2016
Available online 7 April 2016

Keywords:
Gibbs distribution
Gumbel distribution
Perturbation
Entropy
Kullback–Leibler divergence

ABSTRACT

The full-order Perturb-and-MAP model is equivalent to the Gibbs distribution, but its applicability is limited. Empirically it is shown that low-dimensional Gumbel perturbations allow to approximate the Gibbs distribution but their theoretical properties remain unclear. In this note we fill this gap.

1. Introduction

Many modern applications of statistical inference, *e.g.*, in computer vision or natural language processing, involve reasoning in high-dimensional spaces. Typically, the high-dimensional phenomena are modeled using the Gibbs distribution in which a configuration is scored by a potential function. The potential of the configuration can be calculated reasonably easily, however, computation of the partition function requires summation over an exponential number of possible configurations. That is why it is infeasible to reason from the distribution without knowing the normalizing constant.

Recently, in order to overcome the problem of calculating the partition function a new framework has been proposed. The idea of the approach relies on injecting noise to the potential function and then finding the most probable solution (maximum-a-posteriori, MAP) of the perturbed objective. This approach is known as the Perturb-and-MAP (PM) model and it has been shown that injecting full-order Gumbel perturbations is equivalent to the Gibbs distribution (Papandreou and Yuille, 2011). The PM model has been used in formulating sampling procedure (Hazan et al., 2013a) and learning algorithm (Gane et al., 2014). Nevertheless, the application of the full-order perturbations is impractical and it was advocated to use low-dimensional (or 1-order) perturbations (Papandreou and Yuille, 2011) for which some interesting properties have been shown, *e.g.*, an upper bound on the log-partition function (Hazan and Jaakkola, 2012), PAC-Bayes bounds (Hazan et al., 2013b), or a concentration measure for the Gumbel perturbations (Orabona et al., 2014). However, there is no formal justification that indeed low-dimensional perturbations approximate the Gibbs distribution. In this note we fill this gap.

The contribution of the paper is fourfold. First, we show that the maximum of potentials with low-dimensional perturbations approximately follows the Gumbel distribution. Second, we prove that the distribution of the MAP solution of

* Corresponding author. Tel.: +48 71 320 44 53.
E-mail address: jakub.tomczak@pwr.edu.pl.

http://dx.doi.org/10.1016/j.spl.2016.03.019
0167-7152/© 2016 Elsevier B.V. All rights reserved.

the potentials with the 1-order perturbations is approximately equivalent to the Gibbs distribution. Third, we quantify the difference between the full-order and low-dimensional perturbations expressed in the average number of additional bits, *i.e.*, in the sense of the Kullback–Leibler divergence. Fourth, we present an upper bound on the entropy originally proposed in Maji et al. (2014) and provide its slightly simplified proof using Fenchel's inequality.

2. Background

Throughout the paper we assume a real-valued potential function $\theta : \mathcal{X} \to \mathbb{R}$ defined over a discrete product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \mathcal{X}_n$, and such that $\theta(\mathbf{x}) = -\infty$ whenever $\mathbf{x} \notin dom(\theta)$.

The Gibbs distribution maps the real-valued potentials to the probability:

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

where

$$Z(\theta) = \sum_{\mathbf{x}} \exp(\theta(\mathbf{x})) \tag{2}$$

is known as the *partition function*.

The applicability of the Gibbs distribution is fully dependent on the ability of calculating the partition function. The problem of evaluating Z requires summation over all possible configurations of $\mathbf{x}$ which in general belongs to the complexity class #P.

The problem of finding the most probable configuration (the maximum-a-posteriori inference problem) is defined as follows:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \theta(\mathbf{x}). \tag{3}$$

The MAP problem is NP-hard in general (Shimony, 1994), however, it is still simpler than calculating the partition function.

Typically, there are often several meaningful configurations $\mathbf{x}$ whose potentials $\theta(\mathbf{x})$ are close to $\theta(\mathbf{x}^*)$. In order to discover them all, one should apply some sort of sampling procedure. In the context of the Gibbs distribution, the applicability of the widely used Markov Chain Monte Carlo techniques is proved to be limited (Goldberg and Jerrum, 2007). However, we can draw a sample from the Gibbs distribution by perturbing the potential function and further solving the resulting MAP problem. This procedure yields the Perturb-and-MAP model which is based on addition of a random function $\gamma : \mathcal{X} \to \mathbb{R}$ to the potential function in (3) and solving the resulting inference problem:

$$\mathbf{x}^*_{\gamma} = \arg\max_{\mathbf{x}} \{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}. \tag{4}$$

The full-order perturbation γ associates a random variable to each configuration $\mathbf{x}$. Typically, perturbations are i.i.d. In the context of the Gibbs distribution it turns out that the appropriate perturbation is a standard Gumbel random variable (Papandreou and Yuille, 2011). The cumulative distribution function (cdf) for Gumbel random variable is the following (Coles, 2001):

$$G(t; \mu, \lambda) = \exp\left(-\exp\left(-\frac{1}{\lambda}(t-\mu)\right)\right), \tag{5}$$

where μ is the location parameter, and λ is the scale parameter, and the probability density function (pdf) is as follows:

$$g(t; \mu, \lambda) = \frac{1}{\lambda} \exp\left(-\frac{1}{\lambda}(t-\mu)\right) G(t; \mu, \lambda). \tag{6}$$

The Gumbel distribution with $\mu = 0$ and $\lambda = 1$ is called the standard Gumbel distribution.

The Gumbel perturbations fit perfectly to the PM framework for the Gibbs distribution because of the following two theorems, see Papandreou and Yuille (2011) and Orabona et al. (2014).

Theorem 1. *Let $\gamma(\mathbf{x})$ be i.i.d. perturbations with the standard Gumbel distribution. Then the distribution of maximum of perturbed potentials* $\max_{\mathbf{x}}\{\theta(\mathbf{x})+\gamma(\mathbf{x})\}$ *follows the Gumbel distribution with location parameter* $\log Z(\theta)$ *and scale parameter* 1, *that is,*

$$\mathbb{P}[\max_{\mathbf{x}}\{\theta(\mathbf{x}) + \gamma(\mathbf{x})\}] = G(t; \log Z(\theta), 1). \tag{7}$$

Theorem 2. *Let $\gamma(\mathbf{x})$ be i.i.d. perturbations with the standard Gumbel distribution. Then the distribution of MAP solutions of* (4) *is the Gibbs distribution:*

$$\mathbb{P}[\mathbf{x} = \arg\max_{\hat{\mathbf{x}}}\{\theta(\hat{\mathbf{x}}) + \gamma(\hat{\mathbf{x}})\}] = \frac{\exp(\theta(\mathbf{x}))}{Z(\theta)}. \tag{8}$$

The proofs of Theorems 1 and 2 can be found in the Supplementary Material of Papandreou and Yuille (2011).