On Sampling from the Gibbs Distribution with Random Maximum A-Posteriori Perturbations

Tamir Hazan (U of Haifa) Subhransu Maji (TTI Chicago) Tommi Jaakkola (MIT)

Motivation

- Sampling from the Gibbs distribution is provably hard in the data-knowledge domain of machine learning applications.
- Maximum A-Posteriori (MAP) is efficient but sub-optimal due to model inaccuracy.
- Random MAP perturbations generate unbiased samples efficiently.

Contribution: Relating Gibbs distributions to random MAP perturbations.

Background

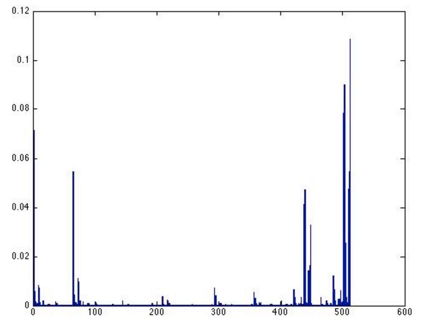
Gibbs distribution:

$$p(x_1,\dots,x_n) = \frac{1}{Z}\exp(\theta(x_1,\dots,x_n))$$

Data-knowledge domain: $\theta_i(x_i), \theta_{i,j}(x_i,x_j)$

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

Gibbs distribution landscape is ragged and samples are provably hard (Jerrum 1993):



Maximum A-Posterior (MAP):

$$\max_{x_1,\dots,x_n}\theta(x_1,\dots,x_n)$$

Random MAP Perturbations

(Papandreou et al. 11, Tarlow et al. 12, Hazan et al. 12)

MAP perturbations and Gibbs distributions: Add a random function γ:X→R with i.i.d. Gumbel random variables $\gamma(x)$

$$p(\hat{x}) = P_\gamma\Big[\hat{x}\in\arg\max_{x\in X}\{\theta(x)+\gamma(x)\}\Big].$$

$$\log Z = E_\gamma\Big[\max_{x\in X}\{\theta(x)+\gamma(x)\}\Big].$$

Proof: $F(t) \stackrel{def}{=} P[\gamma(x)\le t] = \exp(-\exp(-t))$

$$P_\gamma\Big[\max_{x\in X}\{\theta(x)+\gamma(x)\}\le t\Big] = \prod_{x\in X}F(t-\theta(x))$$

$$\exp\Big(-\sum_{x\in X}(-(t-\theta(x)))\Big) = F(t-\theta(x))$$

Theorem (approximate samples from Gibbs marginals with MAP perturbations)
If the graphical model has no cycles then with high probability

$$\Big|\log\Big(P_\gamma\Big[x_r,x_s\in\arg\max_{\hat{x}}\Big\{\hat{\theta}(x)+\sum_{i,j\in E}\hat{\gamma}_{i,j}(x_i,x_j)\Big\}\Big]\Big) - \log\Big(\sum_{x\backslash x_r,x_s}p(x)\Big)\Big| \le \epsilon n$$

Proof idea:

$$\theta(x) = \theta_{1,2}(x_1,x_2)+\theta_{2,3}(x_2,x_3)$$

$$\log\Big(\sum_{x_3}p(x_1,x_2,x_3)\Big) = \theta_{1,2}(x_1,x_2) + \log\Big(\sum_{x_3}\exp(\theta_{2,3}(x_2,x_3))\Big) - \log Z$$

$$\forall x_2\ \ Pr_\gamma\Big[\Big|\frac{1}{m_3}\sum_{j_3=1}^{m_3}\max_{x_3}\{\theta_{2,3}(x_2,x_3)+\gamma_{2,3,j_3}(x_2,x_3)\} - \log\Big(\sum_{x_3}\theta_{2,3}(x_2,x_3)\Big)\Big|\ge\epsilon\Big] \le \frac{\pi}{6m_3\epsilon^2}$$

$$\frac{1}{m_3}\sum_{j_3=1}^{m_3}\max_{x_3}\{\theta_{2,3}(x_2,x_3)+\gamma_{2,3,j_3}(x_2,x_3)\} = \max_{x_{3,j_3}}\Big\{\frac{1}{m_3}\sum_{j_3=1}^{m_3}(\theta_{2,3}(x_2,x_{3,j_3})+\gamma_{2,3,j_3}(x_2,x_{3,j_3}))\Big\}$$

Unbiased samples from Gibbs

Algorithm 1 Unbiased sampling from Gibbs distribution using randomized prediction

Iterate over $j=1,\dots,n$, while keeping fixed $x_1,\dots,x_{j-1}$. Set

1. $p_j(x_j) = \frac{\exp\left(E_\gamma\left[\max_{x_{j+1},\dots,x_n}\{\theta(x)+\sum_{i=j+1}^n\gamma_i(x_i)\}\right]\right)}{\exp\left(E_\gamma\left[\max_{x_j,\dots,x_n}\{\theta(x)+\sum_{i=j}^n\gamma_i(x_i)\}\right]\right)}$.
2. $p_j(r) = 1-\sum_{x_j}p(x_j)$
3. Sample an element 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: $x_1,\dots,x_n$

Low dimensional random functions γ_i:X_i→R with i.i.d. Gumbel random variables $\gamma_i(x_i)$ provide unbiased samples from Gibbs.

$$P\Big[\text{Algorithm 1 outputs x}\ \Big|\ \text{Algorithm 1 accepts}\Big] = p(x).$$

Proof: The probability of sampling $x_1,\dots,x_n$

$$\prod_{j=1}^n\frac{\exp\left(E_\gamma\left[\max_{x_{j+1},\dots,x_n}\{\theta(x)+\sum_{i=j+1}^n\gamma_i(x_i)\}\right]\right)}{\exp\left(E_\gamma\left[\max_{x_j,\dots,x_n}\{\theta(x)+\sum_{i=j}^n\gamma_i(x_i)\}\right]\right)} = \frac{\exp(\theta(x))}{\exp\left(E_\gamma\left[\max_{x_1,\dots,x_n}\{\theta(x)+\sum_{i=1}^n\gamma_i(x_i)\}\right]\right)}.$$

The probability the algorithm accepts is

$$\frac{Z}{\exp\left(E_\gamma\left[\max_{x_1,\dots,x_n}\{\theta(x)+\sum_{i=1}^n\gamma_i(x_i)\}\right]\right)}$$

Theorem (Self-reducing upper bounds):
Low-dimensional random functions γ_i:X_i→R with i.i.d. Gumbel random variables $\gamma_i(x_i)$ satisfy for every j=1,...,n and $x_1,\dots,x_{j-1}$:

$$\sum_{x_j}\exp\Big(E_\gamma\Big[\max_{x_{j+1},\dots,x_n}\{\theta(x)+\sum_{i=j+1}^n\gamma_i(x_i)\}\Big]\Big) \le \exp\Big(E_\gamma\Big[\max_{x_j,\dots,x_n}\{\theta(x)+\sum_{i=j}^n\gamma_i(x_i)\}\Big]\Big)$$

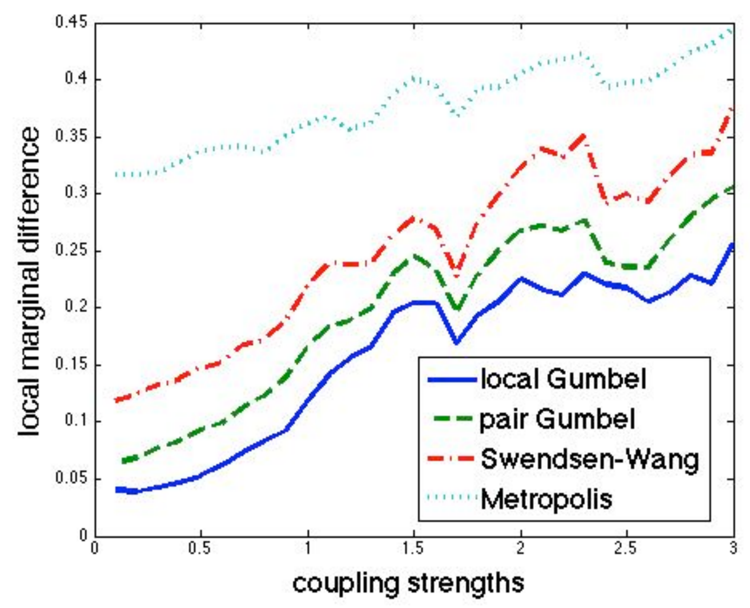
Proof: Taking logarithm on both sides

$$\text{LHS} = E_{\gamma_j}\Big[\max_{x_j}E_{\gamma_{j+1},\dots,\gamma_n}\Big[\max_{x_{j+1},\dots,x_n}\{\theta(x)+\sum_{i=j}^n\gamma_i(x_i)\}\Big]\Big]$$

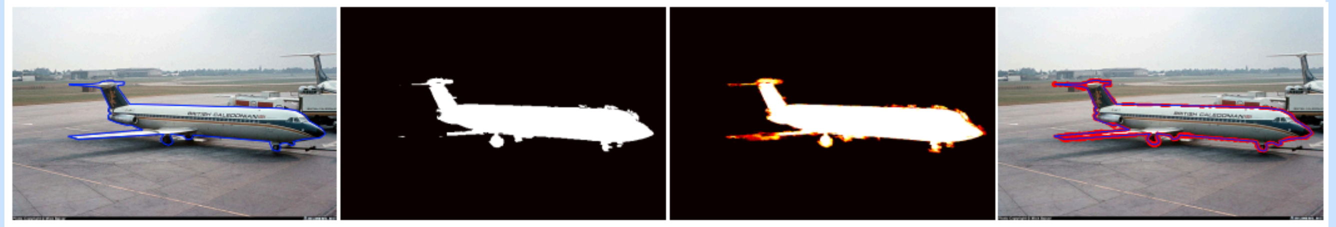
$$\text{RHS} = E_{\gamma_j}\Big[E_{\gamma_{j+1},\dots,\gamma_n}\Big[\max_{x_j}\max_{x_{j+1},\dots,x_n}\{\theta(x)+\sum_{i=j}^n\gamma_i(x_i)\}\Big]\Big]$$

Results

Approximating marginal probabilities:



The importance of probabilistic inference (MAP suffers from model inaccuracy)



Sampling allows computation of non-decomposable losses. *Example image with the boundary annotation (left) and the error estimates obtained using our method (right). Thin structures of the object are often lost in a single MAP solution (middle-left), which are recovered by averaging the samples (middle-right) and lead to better error estimates.*

The unbiased sampler is sub-exponential