Law of large numbers for matchings, extension and applications

Marc Lelarge
Summarized by Hang Zhou

Séminaire de Combinatoire Philippe Flajolet
Institut Henri Poincaré, March 28^{th}*, 2013*

Abstract

The fact that global properties of matchings can be read from local properties of the underlying graph has been rediscovered many times in statistical physics, combinatorics, group theory and computer science. I will present a probabilistic approach allowing to derive law of large numbers. I will show how it extends previous results in several directions and describe some algorithmic applications.

1 Introduction and main result

A h-uniform hypergraph $H = (V, E)$ is called (ℓ, k)-orientable if there exists an assignment of each hyperedge $e \in E$ to exactly ℓ of its vertices $v \in e$ such that no vertex is assigned more than k hyperedges. Let $H_{n,m,h}$ be a hypergraph, drawn uniformly at random from the set of all h-uniform hypergraphs with n vertices and m edges. In this work, we determine the threshold of the existence of a (ℓ, k)-orientation of $H_{n,m,h}$ for $k \geq 1$ and $h > \ell \geq 1$, extending recent results motivated by applications such as cuckoo hashing or load balancing with guaranteed maximum load. Our main result is in the following theorem (see [1]).

Theorem 1. *Let* $Q(x, y) = e^{-x} \sum_{j \geq y} \frac{x^j}{j!}$ *and* $Bin(n, p)$ *denote a binomial random variable with parameters* $n \in \mathbb{N}$ *and* $p \in [0, 1]$*, i.e.* $\mathbb{P}(Bin(n, p) = k) = \binom{n}{k} p^k (1-p)^{n-k}$*. For integers* $h > \ell \geq 1$*,* $k \geq 1$ *with* $\max(h - \ell, k) \geq 2$*, let* ξ^* *be the unique positive solution to*

$$hk = \xi^* \frac{\mathbb{E}\left[\max\left(\ell - Bin(h, 1 - Q(\xi^*, k)), 0\right)\right]}{Q(\xi^*, k+1)\mathbb{P}\left(Bin(h-1, 1 - Q(\xi^*, k)) < \ell\right)}.$$

Let

$$c^*_{h,\ell,k} = \frac{\xi^*}{h\mathbb{P}\left(Bin(h-1, 1 - Q(\xi^*, k)) < \ell\right)}.$$

Then

$$\lim_{n\to\infty} \mathbb{P}\left(H_{n,\lfloor cn \rfloor,h} \text{ is } (\ell,k)\text{-orientable}\right) = \begin{cases} 0 & \text{if } c > c^*_{h,\ell,k}, \\ 1 & \text{if } c < c^*_{h,\ell,k}. \end{cases}$$

Our proof combines the local weak convergence of sparse graphs and a careful analysis of a Gibbs measure on spanning subgraphs with degree constraints. It allows us to deal with a much broader class than the uniform hypergraphs.

The previous characterisation of the threshold $c^*_{h,k,\ell}$ in (for k sufficiently large) involves the solution of a differential equation system which is rather complicated and does not allow to get explicit values for $c^*_{h,k,\ell}$. We believe that our method of proof and the characterisation of the threshold is much simpler.

2 Approach and sketch of proof

2.1 Definitions

We consider a finite simple graph $G = (V, E)$ with a vector of $\mathbb{N}^V$ denoted by $\mathbf{w} = (w_v,\ v \in V)$ and called the vector of (degree) constraints. We are interested in spanning subgraphs (V, F) with degree constraints given by the vector $\mathbf{w}$. Each such subgraph is determined by its edge-set $F \subseteq E$ encoded by the vector $\mathbf{B} = (B_e,\ e \in E) \in \{0,1\}^E$ defined by $B_e = 1$ if and only if $e \in F$. We say that a spanning subgraph $\mathbf{B}$ satisfies the degree constraints or is admissible if for all $v \in V$, we have $\sum_{e\in\partial v} B_e \le w_v$, where ∂v denotes the set of incident edges in G to v. We introduce the family of probability distributions on the set of admissible spanning subgraphs parametrised by a parameter $z > 0$:

$$\mu_G^z(\mathbf{B}) = \frac{z^{\sum_e B_e}}{P_G(z)}, \tag{1}$$

where $P_G(z) = \sum_{\mathbf{B}} z^{\sum_e B_e} \prod_{v\in V} \mathbf{1}(\sum_{e\in\partial v} B_e \le w_v)$. We also define the size of the spanning subgraph by $|F| = \sum_e B_e$ and denote the maximum size by $M(G) = \max\{\sum_e B_e :\ \mathbf{B} \text{ admissible}\}$. Those spanning subgraphs which achieve this maximum are called maximum spanning subgraphs. For any finite graph, when z tends to infinity, the distribution μ_G^z converges to the uniform distribution over maximum spanning subgraphs. For an admissible spanning subgraph, the degree of v in the subgraph is simply $\sum_{e\in\partial v} B_e$. By linearity of expectation, the mean degree of v under the law μ_G^z is $D_v^z := \sum_{e\in\partial v} \mu_G^z\,(B_e = 1)$ so that we have

$$M(G) = \frac{1}{2}\sum_{v\in V} \lim_{z\to\infty} D_v^z. \tag{2}$$

We see hypergraph H as bipartite graph $G = (A \cup B, E)$. Then, an hypergraph is (ℓ, k)-orientable if and only if all vertices in A have degree ℓ in any maximum spanning subgraph of the corresponding bipartite graph with degree constraints (ℓ, k). Indeed in this case, we have for any $v \in A$, $\lim_{z\to\infty} D_v^z = \ell$ so that the size of a maximum spanning subgraph is $M(G) = \ell|A|$.

A fundamental ingredient of the proof is the fact that the bipartite graphs associated to h-uniform hypergraphs, i.e. with n vertices and cn hyperedges are locally tree-like: with high probability, there is no cycle in a ball (of fixed radius) around a vertex chosen at random. It is then instructive to study maximum spanning subgraphs when the underlying graph is a tree. Let us first study the Gibbs measures defined by (1) in the limit $z \to \infty$ in order to analyse maximum spanning subgraphs. When the underlying graph is a finite tree, we can use a more direct and algorithmic way that we now describe.

2.2 Message Passing Algorithm

To study the (ℓ, k)-orientability of the hypergraph H associated to the bipartite graph G, the vector of degree constraints $\mathbf{w}$ should be chosen such that $w_v = \ell$ for $v \in A$ and $w_v = k$ for $v \in B$. For simplicity, we assume here that the vector of degree constraints is constant so that all vertices have the same degree constraint say $w \ge 1$. Consider now the following message-passing algorithm forwarding messages in $\{0,1\}$ on the oriented edges of the underlying tree G as follows: at each round, each oriented edge forwards a message, hence two messages are sent on each edge (one in each direction) at each round. The message passed on the oriented edge $\overrightarrow{e} = (u, v)$ is 0 if the sum of the incoming messages to u from neighbours different from v in previous round is at least w and the message is 1 otherwise, i.e. if the sum of the incoming messages is strictly less than w. Let $\mathbf{I}_k \in \{0,1\}^{\overrightarrow{E}}$ be the vector describing the messages sent on the oriented edges in $\overrightarrow{E}$

at the k-th round of the algorithm. Denote by $\mathcal{P}_G$ the action of the algorithm on the messages in one round so that $\mathbf{I}_{k+1} = \mathcal{P}_G(\mathbf{I}_k)$. It is easy to see that the algorithm will converge on any finite tree after a number of steps equals to at most the diameter of the tree, whatever the initial condition. Hence the messages of the algorithm converge to a vector $\mathbf{I}^* = (I^*_{\overrightarrow{e}}, \overrightarrow{e} \in \overrightarrow{E})$ solving the fixed-point equation $\mathbf{I}^* = \mathcal{P}_G(\mathbf{I}^*)$ and the size of a maximum spanning subgraph is given by

$$\frac{1}{2}\sum_{v\in V}\left(2w\mathbf{1}\left(\sum_{\overrightarrow{e}\in\partial v} I^*_{\overrightarrow{e}} \geq w+1\right) + \mathbf{1}\left(\sum_{\overrightarrow{e}\in\partial v} I^*_{\overrightarrow{e}} \leq w\right)\sum_{\overrightarrow{e}\in\partial v} I^*_{\overrightarrow{e}}\right), \tag{3}$$

where ∂v is the set of oriented edges toward v.

Note that the correctness of the algorithm is ensured for trees only, but the definition of the algorithm does not require the graph to be a tree. It makes only local computations and can be used on any graph. Since the bipartite graphs associated to $H_{n,\lfloor cn\rfloor,h}$ are not trees but are locally tree like, it is tempting to use the algorithm directly on these graphs. It turns out that for low values of c, the algorithm will converge and will also be correct (with high probability). The algorithm allows to compute the size of a maximum spanning subgraph for values of c above $1/h$ but it breaks down at some higher value of c. From an algorithmic viewpoint, there is 'no correlation decay' and the computations made by the algorithm is not anymore local.

2.3 Bypassing corelation decay

In order to bypass this absence of 'correlation decay', we borrow ideas from statistical physics by introducing the Gibbs measures μ^z_G parametrised by a parameter $z > 0$ (usually called the activity or the fugacity). Informally, the introduction of this parameter z will allow us to capture sufficient additional information on our problem in order to identify the 'right' solution to the fixed-point equation $\mathbf{I} = \mathcal{P}_G(\mathbf{I})$, when we let z goes to infinity. Our first step in the analysis of these measures is to derive a message-passing algorithm allowing to compute the mean degree D^z_v of any vertex v in a spanning subgraph taken at random according to the probability distribution μ^z_G. We will proceed by first defining the local computations required at each node and we call them the local operators. We use these building blocks to define a message-passing algorithm which is valid on any finite tree. In particular, we show that as z tends to infinity, the dynamic of the algorithm becomes exactly the one described previously in this section. We shows that the message-passing algorithm converges to a unique fixed point for any $z < \infty$.

3 Extensions

The absence of phase transition has also algorithmic implications, such as sublinear algorithm to approximate the number of matchings (see [2]). A matching is a special case of a spanning subgraph with degree constraint being 1 for every node. For a graph G, let $Z(G,\lambda)$ be the partition function of the monomer-dimer system defined by: $Z(G,\lambda) = \sum_k m_k(G)\lambda^k$, where $m_k(G)$ is the number of matchings of cardinality k in G. We consider graphs of bounded degree and develop a sublinear-time algorithm for approximating $\log Z(G,\lambda)$ at an arbitrary value $\lambda > 0$ within additive error ϵn with high probability. The query complexity of our algorithm does not depend on the size of G and is polynomial in $1/\epsilon$, and we also provide a lower bound quadratic in $1/\epsilon$ for this problem. This is the first analysis of a sublinear-time approximation algorithm for a $\#P$-complete problem. Our approach is based on the correlation decay of the Gibbs distribution associated with $Z(G,\lambda)$. We show that our algorithm approximates the probability for a vertex to be covered by a matching, sampled according to this Gibbs distribution in a near-optimal sublinear-time. We extend our results to approximate the average size and the entropy of such a matching within an additive error with high probability, where again the query complexity is polynomial in $1/\epsilon$ and the lower bound is quadratic in $1/\epsilon$. This result can be extended to many other problems where the correlation decay is known to hold as for independent sets or the Ising model up to the critical activity.

References

[1] Marc Lelarge. A new approach to the orientation of random hypergraphs. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 251–264. SIAM, 2012.

[2] Marc Lelarge and Hang Zhou. Sublinear-time algorithms for monomer-dimer systems on bounded degree graphs. *arXiv preprint arXiv:1208.3629*, 2012.