Convergence of multivariate belief propagation, with applications to cuckoo hashing and load balancing.

Mathieu Leconte, Marc Lelarge, Laurent Massoulié

Convergence of Multivariate Belief Propagation, with Applications to Cuckoo Hashing and Load Balancing

M. Leconte
Technicolor - INRIA
mathieu.leconte@inria.fr

M. Lelarge
INRIA - École Normale Supérieure
marc.lelarge@ens.fr

L. Massoulié
INRIA, Microsoft Research - INRIA Joint Center*
laurent.massoulie@inria.fr

Abstract

This paper is motivated by two applications, namely i) generalizations of cuckoo hashing, a computationally simple approach to assigning keys to objects, and ii) load balancing in content distribution networks, where one is interested in determining the impact of content replication on performance. These two problems admit a common abstraction: in both scenarios, performance is characterized by the maximum weight of a generalization of a matching in a bipartite graph, featuring node and edge capacities. Our main result is a law of large numbers characterizing the asymptotic maximum weight matching in the limit of large bipartite random graphs, when the graphs admit a *local weak limit* that is a tree. This result specializes to the two application scenarios, yielding new results in both contexts. In contrast with previous results, the key novelty is the ability to handle edge capacities with arbitrary integer values. An analysis of belief propagation algorithms (BP) with multivariate belief vectors underlies the proof. In particular, we show convergence of the corresponding BP by exploiting monotonicity of the belief vectors with respect to the so-called *upshifted likelihood ratio* stochastic order. This auxiliary result can be of independent interest, providing a new set of structural conditions which ensure convergence of BP.

1 Introduction

Belief Propagation (BP) is a popular message-passing algorithm for determining approximate marginal distributions in Bayesian networks [25] and statistical physics [22] or for decoding LDPC codes [26]. The popularity of BP stems from its successful application to very diverse contexts where it has been observed to converge quickly to meaningful limits [30], [21]. In contrast, relatively few theoretical results are available to prove rigorously its convergence and uniqueness of its fixed points when the underlying graph is not a tree [4].

In conjunction with the local weak convergence [2], BP has also been used as an analytical tool to study combinatorial optimization problems on random graphs: through a study of its fixed points, one can determine so-called Recursive Distributional Equations (RDE) associated with specific combinatorial problems. In turn, these RDEs determine the asymptotic behaviour of solutions to the associated combinatorial problems in the limit of large instances. Representative results in this vein concern matchings [5], spanning subgraphs with degree constraints [27] and orientability of random hypergraphs [19].

All these problems can be encoded with binary values on the edges of the underlying graph and these contexts involve BP with scalar messages. A key step in these results consists in showing monotonicity of the BP message-passing routine with respect to the input messages. As an auxiliary result, the analyses of [27] and [19] provide structural monotonicity properties under which BP is guaranteed to converge (when messages are scalar).

The present work is in line with [27], [19] and contributes to a rigorous formalization of the cavity method, originating from statistical physics [23], [16], and applied here to a generalized matching problem [20]. The initial motivation is the analysis of generalized matching problems in bipartite graphs with both edge and node capacities. This generic problem has several applications. In particular, it accurately models the service capacity of distributed content delivery networks under various content encoding scenarios, by letting nodes of the bipartite graph represent either contents or servers. It also models problem instances of cuckoo hashing, where in that context nodes represent either

[*] This work was conducted while the author was with Technicolor.

objects or keys to be matched.

Previous studies of these two problems [19, 17] essentially required unit edge capacities, which in turn ensured that the underlying BP involved only scalar messages. It is however necessary to go beyond such unit edge capacities to accurately model general server capacities and various content coding schemes in the distributed content delivery network case. The extension to general edge capacities is also interesting in the context of cuckoo hashing when keys can represent sets of addresses to be matched to objects (see Section 3.1).

Our main contribution is Theorem 2.1, a law of large numbers characterizing the asymptotic size of maximum size generalized matchings in random bipartite graphs in terms of RDEs. It is stated in Section 2. It is then applied to cuckoo hashing and distributed content delivery networks in Section 3, providing generalizations of the results in [19] and [17] respectively.

Besides obtaining these new laws of large numbers, our results also have algorithmic implications. Indeed to prove Theorem 2.1, in Section 4 we state Proposition 4.2, giving simple continuity and monotonicity conditions on the message-passing routine of BP which guarantee its convergence to a unique fixed-point. This result is shown to apply in the present context for the so-called upshifted likelihood ratio stochastic order. Beyond its application to the present matching problem, this structural result might hold under other contexts, and with stochastic orders possibly distinct from the upshifted likelihood ratio order, to establish convergence of BP in the case of multivariate messages.

2 Main result

Let $G = (V, E)$ be a finite graph, with additionally an integer vertex-constraint b_v attached to each vertex $v \in V$ and an integer edge-constraint c_e attached to each edge $e \in E$.

A vector $\mathbf{x} = (x_e)_{e \in E} \in \mathbb{N}^E$ is called an *allocation* of G (or a c-capacitated b-matching, in [28]) if

$$\forall e \in E,\ 0 \le x_e \le c_e \text{ and } \forall v \in V,\ \sum_{e \in \partial v} x_e \le b_v,$$

where ∂v is the set of edges incident to v in G. We also write $u \sim v$ when $uv \in E$.

For an allocation $\mathbf{x}$ of G, we define the size $|\mathbf{x}|$ of $\mathbf{x}$ as $|\mathbf{x}| := \sum_{e \in E} x_e$, and we denote by $M(G)$ the maximum size of an allocation of G. Our aim is to characterize the behaviour of $M(G)/|V|$ for large graphs G in the form of a law of large numbers as $|V|$ goes to infinity.

We focus mainly on sequences of graphs $(G_n)_{n \in \mathbb{N}}$ which converge locally weakly towards Galton-Watson trees G. In short (we will explain more in detail later), what this convergence means is that, if we let R_n be a vertex chosen uniformly at random in G_n, what R_n sees within any finite graph distance k looks more and more like the k-hop neighborhood of the root of a Galton-Watson tree as $n \to \infty$. Such a tree is characterized by a joint law $\Phi \sim (D, W, \{C_i\}_{i=1}^D)$ for respectively the degree, vertex-constraint and adjacent edge-constraints (counted with multiplicity) of the vertices of G. We always assume that the graphs are locally finite, i.e. $D < \infty$ a.s.

To sample a Galton-Watson tree G, we first draw a sample from Φ for the root. Then we construct at each dangling edge the missing vertex and its other adjacent edges (therefore maybe creating new dangling edges), until no dangling edge remains. Independently for each dangling edge and conditionally on its capacity c_0, we draw a sample $(\widetilde{D}, \widetilde{W}, \{\widetilde{C}_i\}_{i=1}^{\widetilde{D}} | c_0) \sim \widetilde{\Phi}(\cdot | c_0)$ for the number of other adjacent edges (not counting the dangling edge), the capacity of the vertex, and the other adjacent edge-constraints. Specifically, the distribution $\widetilde{\Phi}$ is given by

$$\widetilde{\Phi}(\tilde{d} - 1, \tilde{b}, \{\tilde{c}_1, \ldots, \tilde{c}_{\tilde{d}-1}\} | c_0) = \frac{\Phi(\tilde{d}, \tilde{b}, \{c_0, \tilde{c}_1, \ldots, \tilde{c}_{\tilde{d}-1}\})(1 + \sum_{i=1}^{\tilde{d}-1} \mathbf{1}(\tilde{c}_i = c_0))}{\sum\limits_{(d,b,\{c_1,\ldots,c_{d-1}\})} \Phi(d, b, \{c_0, \ldots, c_{d-1}\})(1 + \sum_{i=1}^{d-1} \mathbf{1}(c_i = c_0))}.$$

The construction above can be extended to bipartite graphs $G = (A \cup B, E)$. In that case, there are two laws Φ^A and Φ^B for the characteristics $(D^A, W^A, \{C_i^A\}_{i=1}^{D^A})$ and $(D^B, W^B, \{C_i^B\}_{i=1}^{D^B})$ of vertices in A and B respectively. These satisfy the consistency relation for all edge capacities c:

$$\frac{1}{\mathbb{E}[D^A]} \mathbb{E} \sum_{i=1}^{D^A} \mathbf{1}(C_i^A = c) = \frac{1}{\mathbb{E}[D^B]} \mathbb{E} \sum_{i=1}^{D^B} \mathbf{1}(C_i^B = c).$$

The construction then alternates between $\widetilde{\Phi}^A$ and $\widetilde{\Phi}^B$ for vertices at even and odd distances from the root.

We define $[z]_x^y = \max\{x, \min\{y, z\}\}$ and $(z)^+ = \max\{z, 0\}$. Our main result allows to compute the limit $\mathcal{M}(\Phi^A, \Phi^B)$ of $M(G_n)/|A_n|$ when $(G_n)_{n \in \mathbb{N}}$ converges locally weakly towards a bipartite Galton-Watson tree $G = (A \cup B, E)$ defined by Φ^A and Φ^B:

Theorem 2.1 (Maximum allocation for bipartite Galton-Watson limits) *Provided $\mathbb{E}[W^A]$ and $\mathbb{E}[W^B]$ are finite, the limit $\mathcal{M}(\Phi^A, \Phi^B) = \lim_{n \to \infty} M(G_n)/|A_n|$ exists and equals*

$$\mathcal{M}(\Phi^A, \Phi^B) = \inf \Bigg\{ \mathbb{E}\left[\min\left\{W^A, \sum_{i=1}^{D^A} X_i(C_i^A)\right\}\right] + \frac{\mathbb{E}[D^A]}{\mathbb{E}[D^B]} \mathbb{E}\Bigg[\left(W^B - \sum_{i=1}^{D^B} \left[W^B - \sum_{j \neq i} Y_j(C_j^B) \right]_0^{C_i^B} \right)^+ \mathbf{1}\left(W^B < \sum_{i=1}^{D^B} C_i^B \right) \Bigg] \Bigg\}$$

where for all i, $(X_i(c), Y_i(c))_{c\in\mathbb{N}}$ *is an independent copy of* $(X(c), Y(c))_{c\in\mathbb{N}}$*, and the infimum is taken over distributions for* $(X(c), Y(c))_{c\in\mathbb{N}}$ *satisfying the RDE*

$$\begin{aligned} Y(c) &= \left\{ \left[\widetilde{W}^A - \sum_{i=1}^{\widetilde{D}^A} X_i(\widetilde{C}_i^A) \right]_0^c \middle| C_0^A = c \right\}; \\ X(c) &= \left\{ \left[\widetilde{W}^B - \sum_{i=1}^{\widetilde{D}^B} Y_i(\widetilde{C}_i^B) \right]_0^c \middle| C_0^B = c \right\}. \end{aligned}$$

Remark 2.1. *A similar result holds when the graphs are not bipartite; the limiting tree is then simply a Galton-Watson tree described by a joint distribution* Φ*. We set* $\Phi^A = \Phi^B = \Phi$*, and the formula in Theorem 2.1 then computes* $\lim_{n\to\infty} \frac{2M(G_n)}{|V_n|} = \mathcal{M}(\Phi, \Phi)$.

3 Applications

We now apply Theorem 2.1 to performance analysis of generalized cuckoo hashing and distributed content-delivery networks.

3.1 Cuckoo hashing and hypergraph orientability Cuckoo hashing is a simple approach for assigning keys (hashes) to items. Given an initial collection of n keys, each item is proposed upon arrival two keys chosen at random and must select one of them. Depending on the number m of items and the random choices offered to each item, it may or may not be possible to find such an assignement of items to keys. In the basic scenario, it turns out that such an assignement will be possible with probability tending to 1 as $m, n \to \infty$ for all $m = \lfloor \tau n \rfloor$ with $\tau < \frac{1}{2}$.

The basic problem can be extended in the following meaningful ways:

- each item can choose among $h \geq 2$ random keys [9, 12, 13];
- each key can hold a maximum of k items [10, 6, 11];
- each item must be replicated at least l times [14, 19];
- each (item,key) pair can be used a maximum of r times (not covered previously)

the basic setup corresponding to $(h, k, l, r) = (2, 1, 1, 1)$. We let $\tau^*_{h,k,l,r}$ be the associated threshold, i.e. if $m = \lfloor \tau n \rfloor$ with $\tau < \tau^*_{h,k,l,r}$ then an assignement of items to keys satisfying the conditions above will exist with probability tending to 1 as $m, n \to \infty$; on the contrary, if $\tau > \tau^*_{h,k,l,r}$, then the probability that such an assignement exists will tend to 0 as $m, n \to \infty$.

An alternative description of the present setup consists in the following hypergraph orientation problem. For $h \in \mathbb{N}^*$, a h-uniform hypergraph is a hypergraph whose hyperedges all have size h. We assign marks in $\{0, \dots, r\}$ to each of the endpoints of a hyperedge. For $l < h$ in $\mathbb{N}^*$, a hyperedge is said to be (l, r)-oriented if the sum of the marks at its endpoints is equal to l. The in-degree of a vertex of the hypergraph is the sum of the marks assigned to it in all its adjacent hyperedges. For a positive integer k, a (k, l, r)-orientation of an h-uniform hypergraph is an assignement of marks to all endpoints of all hyperedges such that every hyperedge is (l, r)-oriented and every vertex has in-degree at most k; if such a (k, l, r)-orientation exists, we say that the hypergraph is (k, l, r)-orientable. We now consider the probability space $\mathcal{H}_{n,m,h}$ of the set of all h-uniform hypergraphs with n vertices and m hyperedges, and we denote by $H_{n,m,h}$ a random sample from $\mathcal{H}_{n,m,h}$. In this context, we can interpret Theorem 2.1 as follows:

Theorem 3.1 (Threshold for (k, l, r)-orientability of h-uniform hypergraphs) *Let* h, k, l, r *be positive integers such that* $k, l \geq r$, $(h-1)r \geq l$ *and* $k + (h-2)r - l > 0$ *(i.e. at least one of the inequalities among* $k \geq r$ *and* $(h-1)r \geq l$ *is strict). We define* Φ^A *and* Φ^B_τ *by* $(h, l, \{r\}) \sim \Phi^A$ *and* $(\mathrm{Poi}(\tau h), k, \{r\}) \sim \Phi^B_\tau$*, and*

$$\tau^*_{h,k,l,r} = \sup\left\{\tau : \mathcal{M}(\Phi^A, \Phi^B_\tau) < l\right\}.$$

Then,

$$\lim_{n\to\infty} \mathbb{P}\left(H_{n,\lfloor \tau n\rfloor,h} \textit{ is } (k,l,r)\textit{-orientable}\right) = \begin{cases} 1 \textit{ if } \tau < \tau^*_{h,k,l,r} \\ 0 \textit{ if } \tau > \tau^*_{h,k,l,r} \end{cases}$$

This result extends those from [19], where the value of the threshold $\tau^*_{h,k,l,1}$ was computed. The proof can be found in the appendix.

3.2 Distributed content delivery network Consider a content delivery network (CDN) in which service can be given either from a powerful but costly data center, or from a large number of small, inexpensive servers. Content requests are then served if possible by the small servers and otherwise redirected to the datacenter. One is then interested in determining the fraction of load that can be absorbed by the small servers. A natural asymptotic to consider is that of large number m of small servers with fixed storage and service capacity and large collection n of content items.

The precise model we consider follows the statistical assumptions from [17]. It is described by a bipartite graph $G = (A \cup B, E)$, where A is the set of servers and B the set of contents, $|A| \sim |B|\tau$. An edge in E

between a server s in A and a content c in B indicates that server s stores a copy of content c and is thus able to serve requests for it.

An assignement of servers to requests corresponds exactly to an allocation of G provided the vertex-constraint at server s is its upload capacity, the vertex-constraint at content c is its number of requests ω_c, and the edge-constraint is ∞. Thus, $M(G)$ is the maximum number of requests absorbed by the small servers. Assuming Φ^A is the distribution of storage and upload capacity of the servers and Φ^B the distribution of number of replicas and requests of the contents, then $\mathcal{M}(\Phi^A, \Phi^B)$ computed from Theorem 2.1 is the asymptotic maximum load absorbed by the servers (in number of requests per server). This represents a generalization of the results in [17] which handled only servers with unit service capacity, while our result applies to any capacity distribution with finite mean.

Furthermore, the addition of edge capacities also allows us to model more complex cases. Suppose that all contents may have unequal sizes, say the size of a randomly chosen content is a random variable L, and that each content is fragmented into segments of constant unit size. The storage and upload capacity of the servers is then measured in terms of size rather than number of contents, and the servers now choose which content and also which segment they store.

Assume further that when a server chooses to cache a segment from content c, instead of storing the raw segment it instead stores a random linear combination of all the l_c segments corresponding to content c. Then, when a user requests content c it needs only download a coded segment from any l_c servers storing segments from c, as any l_c coded segments are sufficient to recover the content c. An assignement of servers to requests still corresponds to an allocation of G, with the vertex-constraints at servers unchanged, the vertex-constraints at content c equal to $\omega_c l_c$ and the edge-constraints linked to a content c equal to ω_c. Indeed a given encoded segment can be used only once per request of the corresponding content. Then, letting Φ^A and Φ^B be the appropriate joint laws, $\mathcal{M}(\Phi^A, \Phi^B)$ is the asymptotic maximum absorbed load (in number of fragments per server).

One could then follow the same path as in [17] and determine the replication ratios of contents based on a priori knowledge about their number of requests so as to maximize the load asymptotically absorbed by the server pool; this is beyond the scope of the present paper.

4 Main Proof Elements

We start with a high level description of this section. The proof strategy uses a detour, by introducing a finite activity parameter $\lambda > 0$ playing the role of an inverse temperature. For a given finite graph G, a Gibbs distribution μ_G^λ is defined on edge occupancy parameters $\mathbf{x}$ (Section 4.1) such that an average under μ_G^λ approaches the quantity of interest $M(G)/|V|$ as λ tends to infinity. Instead of considering directly the limit of this parameter over a series of converging graphs G_n, we take an indirect route, changing the order of limits over λ and n.

We thus first determine for fixed λ the asymptotics in n of averages under $\mu_{G_n}^\lambda$. This is where BP comes into play. We characterize the behaviour of BP associated with μ_G^λ on finite G (Section 4.2), establishing its convergence to a unique fixed point thanks to structural properties of monotonicity for the upshifted likelihood ratio order, and of log-concavity of messages (Sections 4.3 and 4.4). This allows to show that limits over n of averages under $\mu_{G_n}^\lambda$ are characterized by fixed point relations à la BP. Taking limits over $\lambda \to \infty$, one derives from these fixed points the RDEs appearing in the statement of Theorem 1. It then remains to justify interchange of limits in λ and n. These last three steps are handled similarly to [19] (see [18]).

Before we proceed we introduce some necessary notation. Letters or symbols in bold such as $\mathbf{x}$ denote collections of objects $(x_i)_{i\in I}$ for some set I. For a subset S of I, $\mathbf{x}_S$ is the sub-collection $(x_i)_{i\in S}$ and $|\mathbf{x}_S| := \sum_{i\in S} x_i$ is the L_1-norm of $\mathbf{x}_S$. Inequalities between collections of items should be understood componentwise, thus $\mathbf{x} \leq \mathbf{c}$ means $x_i \leq c_i$ for all $i \in I$. For distributions m_i, we let $\mathbf{m}_S(\mathbf{x}) := \prod_{i\in S} m_i(x_i)$. When summing such terms as in $\sum_{\mathbf{x}\in\mathbb{N}^S:|\mathbf{x}|\leq b,\, \mathbf{x}\leq\mathbf{c}} \mathbf{m}_S(\mathbf{x})$, we shall omit the constraint $\mathbf{x} \in \mathbb{N}^S$. Similarly, we let $*_S\mathbf{m} = *_{i\in S} m_i$, where $*$ is the convolution of two vectors (will be defined in Section 4.3).

4.1 Gibbs measure

Let $G = (V, E)$ be a finite graph, with collections of vertex- and edge-constraints $\mathbf{b} = (b_v)_{v\in V}$ and $\mathbf{c} = (c_e)_{e\in E}$. The Gibbs measure at activity parameter $\lambda \in \mathbb{R}_+$ on the set of all vectors in $\mathbb{N}^E$ is then defined, for $\mathbf{x} \in \mathbb{N}^E$, as

$$\begin{aligned}\mu_G^\lambda(\mathbf{x}) &= \frac{\lambda^{|\mathbf{x}|}}{Z_G(\lambda)} \mathbf{1}(\mathbf{x} \text{ allocation of } G) \\ &= \frac{\lambda^{|\mathbf{x}|}}{Z_G(\lambda)} \prod_{v\in V} \mathbf{1}\Big(\sum_{e\in\partial v} x_e \leq b_v\Big) \prod_{e\in E} \mathbf{1}(x_e \leq c_e),\end{aligned}$$

where $Z_G(\lambda)$ is a normalization factor.

When $\lambda \to \infty$, μ_G^λ tends to the uniform probability measure on the set of all allocations of G of maximum size. Thus, $\lim_{\lambda\to\infty} \mu_G(|\mathbf{X}|) = M(G)$, where $\mu_G^\lambda(|\mathbf{X}|)$ is the expected size of a random allocation $\mathbf{X}$ drawn according to μ_G^λ. Hence, we can compute $M(G)/|V|$ as follows:

$$\begin{aligned} \frac{M(G)}{|V|} &= \lim_{\lambda\to\infty} \mu_G^\lambda \left(\sum_{v\in V} \frac{1}{|V|} \frac{\sum_{e\in\partial v} X_e}{2} \right) \\ &= \frac{1}{2} \lim_{\lambda\to\infty} \mathbb{E}\left[\mu_G^\lambda \left(\sum_{e\in\partial R} X_e \right) \right], \qquad (4.1) \end{aligned}$$

where R is a root-vertex chosen uniformly at random among all vertices in V, and the first expectation is with respect to the choice of R.

4.2 Associated BP message passing

We introduce the set $\overrightarrow{E}$ of directed edges of G comprising two directed edges $\overrightarrow{uv}$ and $\overrightarrow{vu}$ for each undirected edge $uv \in E$. We also define $\overrightarrow{\partial v}$ as the set of edges directed towards vertex $v \in V$, $\overleftarrow{\partial v}$ as the set of edges directed outwards from v, and $\partial\overrightarrow{e} := (\overrightarrow{wv})_{w\in\partial v\setminus u}$ if $\overrightarrow{e}$ is the directed edge $\overrightarrow{vu}$.

An allocation puts an integer weight on each edge of the graph. Accordingly the messages to be sent along each edge are distributions over the integers. We let $\mathcal{P}$ be the set of all probability distributions on integers with bounded support, i.e.

$$\begin{aligned} \mathcal{P} \;=\; &\Big\{ p \in [0,1]^{\mathbb{N}}; \sum_{i\in\mathbb{N}} p(i) = 1 \\ &\text{and } \exists k \in \mathbb{N} \text{ such that } p(i) = 0, \forall i > k \Big\}, \end{aligned}$$

and $\widetilde{\mathcal{P}}$ the set of distributions in $\mathcal{P}$ whose support is an interval containing 0.

A message on directed edge $\overrightarrow{e}$ with capacity c_e is a distribution in $\mathcal{P}$ with support in $\{0,\ldots,c_e\}$. The message to send on edge $\overrightarrow{e}$ outgoing from vertex v is computed from the messages incoming to v on the other edges via

$$\mathcal{R}^{(\lambda)}_{\overrightarrow{e}}[\mathbf{m}](x) = \frac{\lambda^x \mathbf{1}(x \le c_{vu}) \sum_{|\mathbf{y}|\le b_v - x} \mathbf{m}_{\partial\overrightarrow{vu}}(\mathbf{y})}{\sum_{t\le c_{vu}} \lambda^t \sum_{|\mathbf{y}|\le b_v - t} \mathbf{m}_{\partial\overrightarrow{vu}}(\mathbf{y})},$$

where we introduced the operator $\mathcal{R}^{(\lambda)}_{\overrightarrow{e}} : \widetilde{\mathcal{P}}^{\partial\overrightarrow{e}} \to \widetilde{\mathcal{P}}$. For notational convenience, we write $\mathcal{R}^{(\lambda)}_{\overrightarrow{e}}[\mathbf{m}]$ instead of $\mathcal{R}^{(\lambda)}_{\overrightarrow{e}}[\mathbf{m}_{\partial\overrightarrow{e}}]$. We also introduce $\mathcal{R}_{\overrightarrow{e}}$ for $\mathcal{R}^{(1)}_{\overrightarrow{e}}$. The two operators are linked via the relationship

$$\mathcal{R}^{(\lambda)}_{\overrightarrow{e}}[\mathbf{m}](x) = \frac{\lambda^x \mathcal{R}_{\overrightarrow{e}}[\mathbf{m}](x)}{\sum_{t\ge 0} \lambda^t \mathcal{R}_{\overrightarrow{e}}[\mathbf{m}](t)}.$$

We also define an operator $\mathcal{D}_v : \widetilde{\mathcal{P}}^{\overrightarrow{\partial v}} \to \mathbb{R}^+$ meant to approximate the average occupancy at a vertex v under μ_G^λ from the messages incoming to v:

$$\mathcal{D}_v[\mathbf{m}] = \frac{\sum_{|\mathbf{x}|\le b_v} |\mathbf{x}| \mathbf{m}_{\overrightarrow{\partial v}}(\mathbf{x})}{\sum_{|\mathbf{x}|\le b_v} \mathbf{m}_{\overrightarrow{\partial v}}(\mathbf{x})}.$$

Finally we denote by $\mathcal{R}^{(\lambda)}_G$ the operator that performs the action of all the $\mathcal{R}^{(\lambda)}_{\overrightarrow{e}}$ for all $\overrightarrow{e}$ simultaneously, i.e. $\mathcal{R}^{(\lambda)}_G[\mathbf{m}] = \left(\mathcal{R}^{(\lambda)}_{\overrightarrow{e}}[\mathbf{m}]\right)_{\overrightarrow{e}\in\overrightarrow{E}}$ (the same type of notation will be used for other operators). It is well known that belief propagation converges and is exact on finite trees [22]:

Proposition 4.1. *In a finite tree G, the fixed point equation $\mathbf{m} = \mathcal{R}^{(\lambda)}_G[\mathbf{m}]$ admits a unique solution $\mathbf{m}^{(\lambda)} \in \widetilde{\mathcal{P}}^{\overrightarrow{E}}$, and it satisfies for every vertex v:*

$$\mu_G^\lambda \left(\sum_{e\in\partial v} X_e \right) = \mathcal{D}_v[\mathbf{m}^{(\lambda)}].$$

However, to be able to take the limit as the activity parameter λ goes to infinity as well as to deal with cases when G is not a tree anymore, we need to study further the operators $\mathcal{R}_{\overrightarrow{e}}$ and $\mathcal{D}_v$, which we term the *local* operators.

4.3 Structural properties of local operators

In this section, we focus on the one-hop neighborhood of a vertex v of a graph G, i.e. on vertex v and its set ∂v of incident edges. We thus only consider the directed edges in $\overrightarrow{\partial v} \cup \overleftarrow{\partial v}$. We let b_v be the vertex-constraint at v and $\mathbf{c} = (c_e)_{e\in\partial v}$ be the collection of the edge-constraints on the edges in ∂v.

Among the many stochastic orders studied for comparing distributions (see e.g. [24]), the one best adapted to the structure of operators $\mathcal{R}_{\overrightarrow{e}}$ and $\mathcal{D}_v$ is the so-called *upshifted likelihood-ratio* stochastic order (abbreviated lr ↑). For two distributions m and m' in $\mathcal{P}$, we say that m is smaller than m' (for the lr ↑ stochastic order) and we write $m \le_{\mathrm{lr}\uparrow} m'$ if

$$m(i+k+l)m'(i) \le m(i+l)m'(i+k), \forall i,k,l \in \mathbb{N}.$$

In particular, if m and m' have the same interval as support, we have $m \le_{\mathrm{lr}\uparrow} m' \Leftrightarrow \frac{m(i+1)}{m(i)} \le \frac{m'(i+1)}{m'(i)}$, for all i for which the denominators are non-zero. In this paper, we will always use the lr ↑-order when comparing distributions.

We shall also need the following definition. A distribution $(p_j)_{j\ge 0}$ is **log-concave** if its support is an interval and $p_i p_{i+2} \le p_{i+1}^2$, for all $i \in \mathbb{N}$. This property has strong ties with the lr ↑-order. In particular one can note that p is log-concave if and only if $p \le_{\mathrm{lr}\uparrow} p$. We let

$\mathcal{P}_{\rm lc} \subset \mathcal{P}$ be the set of all log-concave distributions over integers with finite support, and $\widetilde{\mathcal{P}}_{\rm lc} = \widetilde{\mathcal{P}} \cap \mathcal{P}_{\rm lc}$:

$$\mathcal{P}_{\rm lc} \;=\; \Big\{p \in [0,1]^{\mathbb{N}}; \sum_{i\in\mathbb{N}} p(i) = 1,\ p \text{ is log-concave,} \\ \text{and } \exists k \in \mathbb{N} \text{ such that } p(i) = 0, \forall i > k\Big\}.$$

The key result of this Section is then the following:

Proposition 4.2 (Monotonicity of the local operators for the lr ↑-order) *The operator $\mathcal{R}^{(\lambda)}_{\vec{e}}$ is non-increasing; furthermore, if the inputs of $\mathcal{R}^{(\lambda)}_{\vec{e}}$ are log-concave, then the output is also log-concave. The operator $\mathcal{D}_v$ is non-decreasing, and strictly increasing if all its inputs are log-concave with* 0 *in their support.*

The proof will rely on the following lemma from [29] establishing stablity of lr ↑-order w.r.t. convolution $*$, where $m * m'(x) = \sum_y m(y)m'(x-y)$:

Lemma 4.1. *For a set $\overrightarrow{S}$ of directed edges, if $\mathbf{m}^1_{\overrightarrow{S}} \le_{lr\uparrow} \mathbf{m}^2_{\overrightarrow{S}}$ in $\mathcal{P}^{\overrightarrow{S}}$, then $*_{\overrightarrow{S}}\mathbf{m}^1 \le_{lr\uparrow} *_{\overrightarrow{S}}\mathbf{m}^2$.*

We shall also need the following notions:

- the *reweighting* of a vector m by a vector p is defined by $m \bullet p(x) := \frac{m(x)p(x)}{\sum_{y\in\mathbb{N}} m(y)p(y)}$ for $x \in \mathbb{N}$, for p and m with non-disjoint supports and $|p| < \infty$ or $|m| < \infty$. If p or m is in $\mathcal{P}$, then $m \bullet p \in \mathcal{P}$. Note that $\mathcal{R}^{(\lambda)}_{\vec{e}}[\mathbf{m}] = \lambda^{\mathbb{N}} \bullet \mathcal{R}_{\vec{e}}[\mathbf{m}]$, where $\lambda^{\mathbb{N}} = (\lambda^x)_{x\in\mathbb{N}}$.
- the *shifted reversal* of a vector p is defined by $p^R(x) = p(b_v - x)\mathbf{1}(x \le b_v)$ for $x \in \mathbb{N}$; if $p \in \mathcal{P}$ and its support is included in $[0, b_v]$, then $p^R \in \mathcal{P}$ as well.

It is straightforward to check that

Lemma 4.2. *Reweighting preserves the lr ↑-order; shifted reversal reverses the lr ↑-order.*

Note that by the previous lemma it suffices to prove the results of Proposition 4.2 for $\mathcal{R}_{\vec{e}}$ and they will then extend to $\mathcal{R}^{(\lambda)}_{\vec{e}}$. For space reasons, we prove here only the part of the statement concerning $\mathcal{R}_{\vec{e}}$, and only for inputs in $\widetilde{\mathcal{P}}$. The rest of the proof can be found in [18].

Proof. Let $\vec{e}$ be an edge outgoing from vertex v, and $\mathbf{m}^1_{\partial\vec{e}}, \mathbf{m}^2_{\partial\vec{e}} \in \widetilde{\mathcal{P}}^{\partial\vec{e}}$ such that $\mathbf{m}^1_{\partial\vec{e}} \le_{\rm lr\uparrow} \mathbf{m}^2_{\partial\vec{e}}$. Let $\delta_{[0,b_v]}(x) = \mathbf{1}(0 \le x \le b_v)$; we have $\delta_{[0,b_v]} *_{\partial\vec{e}} \mathbf{m}^i(x) = \sum_{x-b_v\le|\mathbf{y}|\le x} \mathbf{m}^i_{\partial\vec{e}}(\mathbf{y})$. Clearly, $\delta_{[0,b_v]}$ is log-concave, so $\delta_{[0,b_v]} \le_{\rm lr\uparrow} \delta_{[0,b_v]}$ and Lemma 4.1 then implies $\delta_{[0,b_v]} *_{\partial\vec{e}} \mathbf{m}^1 \le_{\rm lr\uparrow} \delta_{[0,b_v]} *_{\partial\vec{e}} \mathbf{m}^2$. Lemma 4.2 then says $\left(\delta_{[0,b_v]} *_{\partial\vec{e}} \mathbf{m}^1\right)^R \ge_{\rm lr\uparrow} \left(\delta_{[0,b_v]} *_{\partial\vec{e}} \mathbf{m}^2\right)^R$. It is easy to check that

$$\mathcal{R}_{\vec{e}}[\mathbf{m}^i] = \delta_{[0,c_e]} \bullet \left(\delta_{[0,b_v]} *_{\partial\vec{e}} \mathbf{m}^i\right)^R; \tag{4.2}$$

and as $\left(\delta_{[0,b_v]} *_{\partial\vec{e}} \mathbf{m}^i\right)^R(0) > 0$ Lemma 4.2 again implies that $\mathcal{R}_{\vec{e}}[\mathbf{m}^1] \ge_{\rm lr\uparrow} \mathcal{R}_{\vec{e}}[\mathbf{m}^2]$.

If now $\mathbf{m}_{\partial\vec{e}} \in \widetilde{\mathcal{P}}^{\partial\vec{e}}_{\rm lc}$, then $\mathbf{m}_{\partial\vec{e}} \le_{\rm lr\uparrow} \mathbf{m}_{\partial\vec{e}}$ and $\mathcal{R}_{\vec{e}}[\mathbf{m}] \ge_{\rm lr\uparrow} \mathcal{R}_{\vec{e}}[\mathbf{m}]$, hence $\mathcal{R}_{\vec{e}}[\mathbf{m}] \in \widetilde{\mathcal{P}}_{\rm lc}$.

To pave the way for the analysis of the limit $\lambda \to \infty$, we distinguish between two collections of messages $\mathbf{m}_{\overrightarrow{\partial v}}$ and $\mathbf{n}_{\overrightarrow{\partial v}}$ in $\widetilde{\mathcal{P}}^{\overrightarrow{\partial v}}$ and introduce additional operators. For an edge $\vec{e}$ outgoing from v we define the operator $\mathcal{Q}^{(\lambda)}_{\vec{e}} : \widetilde{\mathcal{P}}^{\partial\vec{e}} \to \widetilde{\mathcal{P}}$ by $\mathcal{Q}^{(\lambda)}_{\vec{e}}[\mathbf{n}] = \mathcal{R}^{(\lambda)}_{\vec{e}}[\lambda^{\mathbb{N}} \bullet \mathbf{n}]$, where $\lambda^{\mathbb{N}} \bullet \mathbf{n} = \left(\lambda^{\mathbb{N}} \bullet n_{\vec{e}}\right)_{\vec{e}\in\vec{E}}$. As reweighting preserves the lr ↑-order, the operator $\mathcal{Q}^{(\lambda)}_{\vec{e}}$ is non-increasing. It also verifies the following useful monotonicity property with respect to λ, proven in [18]:

Proposition 4.3 (Monotonicity in λ) *For $\mathbf{n}_{\partial\vec{e}} \in \widetilde{\mathcal{P}}^{\partial\vec{e}}$, the mapping $\lambda \mapsto \mathcal{Q}^{(\lambda)}_{\vec{e}}[\mathbf{n}]$ is non-decreasing.*

As $\lambda \to \infty$, limiting messages may not have 0 in their support. We thus define $\alpha_{\vec{e}}$ as the infimum of the support of $m_{\vec{e}} \in \mathcal{P}$, i.e. $\alpha_{\vec{e}} = \min\{x \in \mathbb{N} : m_{\vec{e}}(x) > 0\}$, and $\beta_{\vec{e}}$ as the supremum of the support of $n_{\vec{e}} \in \widetilde{\mathcal{P}}$, i.e. $\beta_{\vec{e}} = \max\{x \in \mathbb{N} : n_{\vec{e}}(x) > 0\}$. When there may be confusion, we will write $\alpha(m_{\vec{e}})$ and $\beta(m_{\vec{e}})$ for the infimum and the supremum of the support of $m_{\vec{e}}$. We also extend the definition of the local operators given previously so that they allow inputs with arbitrary supports in $\mathbb{N}$: for an edge $\vec{e}$ outgoing from vertex v, we define $\mathcal{R}_{\vec{e}} : \mathcal{P}^{\partial\vec{e}} \to \widetilde{\mathcal{P}}$, $\mathcal{D}_v : \mathcal{P}^{\overrightarrow{\partial v}} \to \mathbb{R}^+$, $\mathcal{Q}_{\vec{e}} : \widetilde{\mathcal{P}}^{\partial\vec{e}} \to \mathcal{P}$ and $\mathcal{S}_{\vec{e}} : \mathbb{N}^{\partial\vec{e}} \to \mathbb{N}$ as

$$\mathcal{R}_{\vec{e}}[\mathbf{m}](x) = \begin{cases} \frac{\mathbf{1}(x\le c_e)\sum_{|\mathbf{y}|\le b_v - x} \mathbf{m}_{\partial\vec{e}}(\mathbf{y})}{\sum_{t\le c_e}\sum_{|\mathbf{y}|\le b_v - t} \mathbf{m}_{\partial\vec{e}}(\mathbf{y})} & \text{if } |\boldsymbol{\alpha}_{\partial\vec{e}}| \le b_v \\ \delta_0(x) & \text{otherwise} \end{cases} \tag{4.3}$$

$$\mathcal{D}_v[\mathbf{m}] = \begin{cases} \frac{\sum_{|\mathbf{x}|\le b_v} |\mathbf{x}|\mathbf{m}_{\overrightarrow{\partial v}}(\mathbf{x})}{\sum_{|\mathbf{x}|\le b_v} \mathbf{m}_{\overrightarrow{\partial v}}(\mathbf{x})} & \text{if } |\boldsymbol{\alpha}_{\overrightarrow{\partial v}}| \le b_v \\ b_v & \text{otherwise} \end{cases} \tag{4.4}$$

$$\mathcal{Q}_{\vec{e}}[\mathbf{n}](x) = \begin{cases} \frac{\mathbf{1}(x\le c_e)\sum_{|\mathbf{y}|= b_v - x} \mathbf{n}_{\partial\vec{e}}(\mathbf{y})}{\sum_{t\le c_e}\sum_{|\mathbf{y}|= b_v - t} \mathbf{n}_{\partial\vec{e}}(\mathbf{y})} & \text{if } |\boldsymbol{\beta}_{\partial\vec{e}}| \ge b_v - c_e \\ \delta_{c_e}(x) & \text{otherwise} \end{cases} \tag{4.5}$$

$$\mathcal{S}_{\vec{e}}(\mathbf{x}) = [b_v - |\mathbf{x}_{\partial\vec{e}}|]_0^{c_e}. \tag{4.6}$$

Note that the support of $\mathcal{R}_{\vec{e}}[\mathbf{m}]$ is $\{0, \ldots, \mathcal{S}_{\vec{e}}(\boldsymbol{\alpha})\}$ and that of $\mathcal{Q}_{\vec{e}}[\mathbf{n}]$ is $\{\mathcal{S}_{\vec{e}}(\boldsymbol{\beta}), \ldots, c_e\}$. The following result is established in [18]:

Proposition 4.4 (Continuity for log-concave inputs and limiting operators) *The operators $\mathcal{R}_{\overrightarrow{e}}$ and $\mathcal{D}_v$ given by equations (4.3),(4.4) are continuous for the L_1 norm for inputs in $\widetilde{\mathcal{P}}_{lc}$. Also, $\mathcal{Q}_{\overrightarrow{e}}$ defined in equation (4.5) satisfies $\mathcal{Q}_{\overrightarrow{e}}[\mathbf{n}] = \lim\uparrow_{\lambda\to\infty} \mathcal{Q}^{(\lambda)}_{\overrightarrow{e}}[\mathbf{n}]$ for any $\mathbf{n}_{\partial\overrightarrow{e}} \in \widetilde{\mathcal{P}}^{\partial\overrightarrow{e}}$.*

It follows naturally that $\mathcal{Q}_{\overrightarrow{e}}$ is non-increasing. Moreover, we can extend the results of Proposition 4.2 to the extended operators, i.e. $\mathcal{R}_{\overrightarrow{e}}$ is still non-increasing and $\mathcal{D}_v$ non-decreasing.

4.4 Convergence of BP on finite graphs

The main result of this section is the following

Proposition 4.5 (Convergence of BP to a unique fixed point) *Synchronous BP message updates according to $\mathbf{m}^{t+1} = \mathcal{R}^{(\lambda)}_G[\mathbf{m}^t]$ for $t \geq 0$ converge to the unique solution $\mathbf{m}^{(\lambda)}$ of the fixed point equation $\mathbf{m} = \mathcal{R}^{(\lambda)}_G[\mathbf{m}]$.*

Proof. For all $\overrightarrow{e} \in \overrightarrow{E}$ initialize the message on $\overrightarrow{e}$ at $m^0_{\overrightarrow{e}} = \delta_0 \in \widetilde{\mathcal{P}}_{\mathrm{lc}}$. As $\mathcal{R}^{(\lambda)}_G$ is non-increasing and δ_0 is a smallest element for the lr $\uparrow$ order, it can readily be shown that the following inequalities hold for all $t \geq 0$:

$$\mathbf{m}^{2t} \leq_{\mathrm{lr}\uparrow} \mathbf{m}^{2t+2} \leq_{\mathrm{lr}\uparrow} \mathbf{m}^{2t+3} \leq_{\mathrm{lr}\uparrow} \mathbf{m}^{2t+1}.$$

In other words the two series $(\mathbf{m}^{2t})_{t\geq 0}$ and $(\mathbf{m}^{2t+1})_{t\geq 0}$ are *adjacent* and hence converge to respective limits $\mathbf{m}^-$, $\mathbf{m}^+$ such that $\mathbf{m}^- \leq_{\mathrm{lr}\uparrow} \mathbf{m}^+$. Continuity of $\mathcal{R}^{(\lambda)}_G$ further guarantees that $\mathbf{m}^+ = \mathcal{R}^{(\lambda)}_G(\mathbf{m}^-)$ and $\mathbf{m}^- = \mathcal{R}^{(\lambda)}_G(\mathbf{m}^+)$. Moreover, considering any other sequence of vectors of messages $(\mathbf{m}'^t)_{t\geq 0}$ with an arbitrary initialization, since $\mathbf{m}^0 \leq_{\mathrm{lr}\uparrow} \mathbf{m}'^0$, monotonicity of $\mathcal{R}^{(\lambda)}_G$ ensures that for all $t \geq 0$, one has

$$\mathbf{m}^{2t} \leq_{\mathrm{lr}\uparrow} \mathbf{m}'^{2t}, \mathbf{m}'^{2t+1} \leq_{\mathrm{lr}\uparrow} \mathbf{m}^{2t+1}.$$

The result will then follow if we can show that $\mathbf{m}^+ = \mathbf{m}^-$.

We establish this by exploiting the fact that $\mathcal{D}_v$ is strictly increasing for inputs in $\widetilde{\mathcal{P}}_{\mathrm{lc}}$. As $\mathbf{m}^- \leq_{\mathrm{lr}\uparrow} \mathbf{m}^+$ and $\mathcal{D}_v$ is non-decreasing for the lr $\uparrow$-order for all $v \in V$, it follows $\mathcal{D}_v[\mathbf{m}^-] \leq \mathcal{D}_v[\mathbf{m}^+]$ for all $v \in V$. Then, summing over all vertices of G, we get

$$\begin{aligned}
\sum_{v\in V} \mathcal{D}_v[\mathbf{m}^-] &= \sum_{v\in V}\sum_{u\sim v} \frac{\sum_{x\in\mathbb{N}} x m^-_{\overrightarrow{uv}}(x)\mathcal{R}_{\overrightarrow{vu}}[\mathbf{m}^-](x)}{\sum_{x\in\mathbb{N}} m^-_{\overrightarrow{uv}}(x)\mathcal{R}_{\overrightarrow{vu}}[\mathbf{m}^-](x)} \\
&= \sum_{v\in V}\sum_{u\sim v} \frac{\sum_{x\in\mathbb{N}} x \mathcal{R}_{\overrightarrow{uv}}[\mathbf{m}^+](x) m^+_{\overrightarrow{vu}}(x)}{\sum_{x\in\mathbb{N}} \mathcal{R}_{\overrightarrow{uv}}[\mathbf{m}^+](x) m^+_{\overrightarrow{vu}}(x)} \\
&= \sum_{u\in V}\sum_{v\sim u} \frac{\sum_{x\in\mathbb{N}} x \mathcal{R}_{\overrightarrow{uv}}[\mathbf{m}^+](x) m^+_{\overrightarrow{vu}}(x)}{\sum_{x\in\mathbb{N}} \mathcal{R}_{\overrightarrow{uv}}[\mathbf{m}^+](x) m^+_{\overrightarrow{vu}}(x)} \\
&= \sum_{u\in V} \mathcal{D}_u[\mathbf{m}^+].
\end{aligned}$$

Hence, in fact, $\mathcal{D}_v[\mathbf{m}^-] = \mathcal{D}_v[\mathbf{m}^+]$ for all $v \in V$. As $\mathcal{D}_v$ is strictly increasing for these inputs, $\mathbf{m}^- = \mathbf{m}^+ = \mathbf{m}^{(\lambda)}$ follows.

We finish this Section by stating results on the limiting behaviour of the fixed point of BP on a fixed finite graph G as $\lambda \to \infty$. First, in [7] Chertkov builds upon the fact an associated linear programming relaxation is gapless due to the total unimodularity of the adjacency matrix of bipartite graphs. Even though [7] only deals with unitary capacities, we simply mention here without proof (as it is not a necessary step towards our main theorems) that the same argument can be used in our setup and yields the following:

Proposition 4.6 (Correctness for finite bipartite graphs) *In finite bipartite graphs,*

$$\frac{1}{2} \lim\uparrow_{\lambda\to\infty} \sum_{v\in V} \mathcal{D}_v[\mathbf{m}^{(\lambda)}] = M(G).$$

Remark 4.1. *In view of this proposition, BP can be used as an algorithm to compute the maximum size of allocations in finite bipartite graphs to any accuracy needed, by running the algorithm at finite but large enough activity parameter λ and computing $\mathcal{D}_v[\mathbf{m}^{(\lambda)}]$ for all v from the fixed-point messages.*

In the non-bipartite case, the fixed-point $\mathbf{m}^{(\lambda)}$ at finite λ admits a limit $\mathbf{m}^{(\infty)}$, and the value of $\sum_v \mathcal{D}_v[\mathbf{m}^{(\infty)}]$ is equal to $\sum_v F_v(\boldsymbol{\alpha}^{(\infty)})$, where F_v is defined in the propositions below (which proofs are in [18]). This sum is computed from the infimum $\boldsymbol{\alpha}^{(\infty)}$ of the support of $\mathbf{m}^{(\infty)}$. Furthermore, $\boldsymbol{\alpha}^{(\infty)}$ can also be obtained from a fixed-point equation, of which it is the solution that gives the lowest value of $\sum_v F_v$.

Proposition 4.7 (Limit of $\lambda \to \infty$) *$\mathbf{m}^{(\lambda)}$ is non-decreasing in λ for the lr$\uparrow$-order, and $\mathbf{m}^{(\infty)} = \lim\uparrow_{\lambda\to\infty} \mathbf{m}^{(\lambda)} \in \mathcal{P}^{\overrightarrow{E}}_{lc}$ is the minimal solution (for the lr$\uparrow$-order) of $\mathbf{m}^{(\infty)} = \mathcal{Q}_G \circ \mathcal{R}_G[\mathbf{m}^{(\infty)}]$.*

Proposition 4.8 (BP estimate in finite graphs) *In a finite graph G, we have*

$$\begin{aligned}
\lim\uparrow_{\lambda\to\infty} \sum_{v\in V} \mathcal{D}_v[\mathbf{m}^{(\lambda)}] &= \sum_{v\in V} \mathcal{D}_v[\mathbf{m}^{(\infty)}] \\
&= \sum_{v\in V} F_v(\boldsymbol{\alpha}^{(\infty)}) \\
&= \inf_{\boldsymbol{\alpha}=\mathcal{S}_G\circ\mathcal{S}_G(\boldsymbol{\alpha})} \sum_{v\in V} F_v(\boldsymbol{\alpha}),
\end{aligned}$$

where $F_v(\boldsymbol{\alpha}) = \min(b_v, |\boldsymbol{\alpha}_{\overrightarrow{\partial v}}|) + (b_v - |\boldsymbol{\alpha}_{\overleftarrow{\partial v}}|)^+$.

Remark 4.2. *In a finite tree, there is only one possible value for* $\alpha_{\overrightarrow{e}} = \mathcal{S}_{\overrightarrow{e}} \circ \mathcal{S}_{\partial\overrightarrow{e}}[\boldsymbol{\alpha}]$*, where* $\circ$ *is the composition operation, when* $\overrightarrow{e}$ *is an edge outgoing from a leaf* v*: it is* $\alpha_{\overrightarrow{e}} = \min\{b_v, c_e\}$*. It is then possible to compute the whole, unique fixed-point vector* $\boldsymbol{\alpha} = \mathcal{S}_G \circ \mathcal{S}_G(\boldsymbol{\alpha})$ *in an iterative manner, starting from the leaves of the tree and climbing up. This gives a simple, iterative way to compute the maximum size of allocations in finite trees, which is the natural extension of the leaf-removal algorithm for matchings.*

4.5 Infinite unimodular graphs

This section extends the results obtained so far for finite graphs to infinite graphs. As in [27, 19], we use for this the framework of [1]. We still denote by $G = (V, E)$ a possibly infinite graph with vertex set V and undirected edge set E (and directed edge set $\overrightarrow{E}$). We always assume that the degrees are finite, i.e. the graph is locally finite. A network is a graph G together with a complete separable metric space Ξ called the mark space, and maps from V and $\overrightarrow{E}$ to Ξ. Images in Ξ are called marks. A rooted network (G, r) is a network with a distinguished vertex r of V called the root. A rooted isomorphism of rooted networks is an isomorphism of the underlying networks that takes the root of one to the root of the other. We do not distinguish between a rooted network and its isomorphism class denoted by $[G, r]$. Indeed, it is shown in [1] how to define a canonical representative of a rooted isomorphism class.

Let $\mathcal{G}_*$ denote the set of rooted isomorphism classes of rooted connected locally finite networks. Define a metric on $\mathcal{G}_*$ by letting the distance between $[G_1, r_1]$ and $[G_2, r_2]$ be $1/(1+\delta)$ where δ is the supremum of those $d \geq 0$ such that there is some rooted isomorphism of the balls of graph-distance radius $\lfloor d \rfloor$ around the roots of G_i such that each pair of corresponding marks has distance less than $1/d$. $\mathcal{G}_*$ is separable and complete in this metric [1].

Similarly to the space $\mathcal{G}_*$, we define the space $\mathcal{G}_{**}$ of isomorphism classes of locally finite connected networks with an ordered pair of distinguished vertices and the natural topology thereon.

Definition 4.1. *Let* ρ *be a probability measure on* $\mathcal{G}_*$*. We call* ρ *unimodular if it obeys the Mass-Transport Principle (MTP): for Borel* $f : \mathcal{G}_{**} \to [0, \infty]$*, we have*

$$\int \sum_{v \in V} f(G, r, v) d\rho([G, r]) = \int \sum_{v \in V} f(G, v, r) d\rho([G, r])$$

Let $\mathcal{U}$ denote the set of unimodular Borel probability measures on $\mathcal{G}_*$. For $\rho \in \mathcal{U}$, we write $\overline{b}(\rho)$ for the expectation of the capacity constraint of the root with respect to ρ. Our first result (whose proof can be found in [18]) is that the BP updates admit a unique fixed-point at finite activity parameter λ:

Proposition 4.9. *Let* $\rho \in \mathcal{U}$ *with* $\overline{b}(\rho) < \infty$*. Then, the fixed point equation* $\mathbf{m} = \mathcal{R}^{(\lambda)}[\mathbf{m}]$ *admits a unique solution* $\boldsymbol{\alpha}^{(\lambda)}$ *for any* $\lambda \in \mathbb{R}^+$ *for* ρ*-almost every marked graph* G*.*

The proof differs from that in the finite graph case in that we cannot sum $\mathcal{D}_v$ over all the vertices $v \in V$ anymore. Instead, we use the MTP for

$$f(G, r, v) = \frac{\sum_{x \in \mathbb{N}} x m^-_{\overrightarrow{vr}}(x) \mathcal{R}_{\overrightarrow{rv}}[\mathbf{m}^-](x)}{\sum_{x \in \mathbb{N}} m^-_{\overrightarrow{vr}}(x) \mathcal{R}_{\overrightarrow{rv}}[\mathbf{m}^-](x)}.$$

The rest of the reasoning goes as in the finite graph case and the proofs can be found in [18] (using the MTP again, instead of summing over all directed edges): Proposition 4.7 is still valid and the following proposition is analogous to Proposition 4.8:

Proposition 4.10 (BP estimate in unimodular random graphs) *Let* $\rho \in \mathcal{U}$ *with* $\overline{b}(\rho) < \infty$*,*

$$\begin{aligned} \lim \uparrow_{\lambda \to \infty} \quad & \int \mathcal{D}_r[\mathbf{m}^{(\lambda)}] d\rho([G, r]) \\ = \quad & \int \mathcal{D}_r[\mathbf{m}^{(\infty)}] d\rho([G, r]) \\ = \quad & \int F_r(\boldsymbol{\alpha}^{(\infty)}) d\rho([G, r]) \\ = \quad & \inf_{\boldsymbol{\alpha} = \mathcal{S}_G \circ \mathcal{S}_G(\boldsymbol{\alpha})} \int F_r(\boldsymbol{\alpha}) d\rho([G, r]), \end{aligned}$$

where $F_v(\boldsymbol{\alpha}) = \min(b_v, |\boldsymbol{\alpha}_{\overrightarrow{\partial v}}|) + (b_v - |\boldsymbol{\alpha}_{\overleftarrow{\partial v}}|)^+$*.*

4.6 From finite graphs to unimodular trees

Once Proposition 4.10 holds, the end of the proof for sequences of (sparse) random graphs is quite systematic and follows the same steps as in [5], [27] and [19]. We first need to show that we can invert the limits in n and λ (see Proposition 6 in [19]):

Proposition 4.11 (Asymptotic correctness for large, sparse random graphs) *Let* $G_n = (V_n, E_n)_n$ *be a sequence of finite marked graphs with random weak limit* ρ *concentrated on unimodular trees, with* $\overline{b}(\rho) < \infty$*. Then,*

$$\begin{aligned} \lim_{n \to \infty} \frac{2M_n}{|V_n|} \quad = \quad & \int \mathcal{D}_r[\mathbf{m}^{(\infty)}] d\rho([G, r]) \\ = \quad & \inf_{\boldsymbol{\alpha} = \mathcal{S}_G \circ \mathcal{S}_G(\boldsymbol{\alpha})} \int F_r(\boldsymbol{\alpha}) d\rho([G, r]). \end{aligned}$$

The second step uses the Markovian nature of the limiting Galton-Watson tree to simplify the infinite recursions $\boldsymbol{\alpha} = \mathcal{S}_G \circ \mathcal{S}_G(\boldsymbol{\alpha})$ into recursive distributional equations as described in Theorem 2.1. Finally, the fact that the sequence of graphs considered in the introduction converges locally weakly to unimodular Galton-Watson trees follows from standard results in the random graphs literature (see [15] for random hypergraphs or [8] for graphs with fixed degree sequence). The details of the rest of the proof of Theorem 2.1 can be found in the appendix.

References

[1] D. Aldous and R. Lyons. Processes on unimodular random networks. *Electron. J. Probab.*, 12:no. 54, 1454–1508, 2007.

[2] D. Aldous and J. M. Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 1–72. Springer, Berlin, 2004.

[3] J. Aldous and A. Bandyopadhyay. A survey of max-type recursive distributional equations. *Annals of Applied Probability 15*, 15:1047–1110, 2005.

[4] M. Bayati, C. Borgs, J. Chayes, and R. Zecchina. On the exactness of the cavity method for weighted b-matchings on arbitrary graphs and its relation to linear programs. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(06):L06001, 2008.

[5] C. Bordenave, M. Lelarge, and J. Salez. Matchings on infinite graphs. *Arxiv preprint arXiv:1102.0712*, 2011.

[6] J. A. Cain, P. Sanders, and N. Wormald. The random graph threshold for k-orientiability and a fast algorithm for optimal multiple-choice allocation. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 469–476, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[7] M. Chertkov. Exactness of belief propagation for some graphical models with loops. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10016, 2008.

[8] A. Dembo and A. Montanari. Gibbs measures and phase transitions on sparse random graphs. *Braz. J. Probab. Stat.*, 24(2):137–211, 2010.

[9] M. Dietzfelbinger, A. Goerdt, M. Mitzenmacher, A. Montanari, R. Pagh, and M. Rink. Tight thresholds for cuckoo hashing via xorsat. In *Proceedings of the 37th international colloquium conference on Automata, languages and programming*, ICALP'10, pages 213–225, Berlin, Heidelberg, 2010. Springer-Verlag.

[10] M. Dietzfelbinger and C. Weidling. Balanced allocation and dictionaries with tightly packed constant size bins. *Theoretical Computer Science*, 380(12):47 – 68, 2007.

[11] D. Fernholz and V. Ramachandran. The k-orientability thresholds for gn, p. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 459–468, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[12] N. Fountoulakis, M. Khosla, and K. Panagiotou. The multiple-orientability thresholds for random hypergraphs. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, pages 1222–1236. SIAM, 2011.

[13] A. Frieze and P. Melsted. Maximum matchings in random bipartite graphs and the space utilization of cuckoo hash tables. *Random Structures & Algorithms*, 41(3):334–364, 2012.

[14] P. Gao and N. C. Wormald. Load balancing and orientability thresholds for random hypergraphs. In *Proceedings of the 42nd ACM symposium on Theory of computing*, STOC '10, pages 97–104, New York, NY, USA, 2010. ACM.

[15] J. H. Kim. Poisson Cloning Model for Random Graphs. *ArXiv e-prints*, May 2008.

[16] F. Krząkała, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proc. Natl. Acad. Sci. USA*, 104(25):10318–10323 (electronic), 2007.

[17] M. Leconte, M. Lelarge, and L. Massoulié. Bipartite graph structures for efficient balancing of heterogeneous loads. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, pages 41–52, New York, NY, USA, 2012. ACM.

[18] M. Leconte, M. Lelarge, and L. Massoulié. Convergence of multivariate belief propagation, with applications to cuckoo hashing and load balancing. *Arxiv preprint arXiv:1207.1659*, 2012.

[19] M. Lelarge. A new approach to the orientation of random hypergraphs. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 251–264. SIAM, 2012.

[20] L. Lovász and M. D. Plummer. *Matching theory.* AMS Chelsea Publishing, Providence, RI, 2009. Corrected reprint of the 1986 original [MR0859549].

[21] E. Maneva, E. Mossel, and M. J. Wainwright. A new look at survey propagation and its generalizations. *J. ACM*, 54(4), July 2007.

[22] M. Mezard and A. Montanari. *Information, Physics, and Computation.* Oxford University Press, Inc., New York, NY, USA, 2009.

[23] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond*, volume 9 of *World Scientific Lecture Notes in Physics.* World Scientific Publishing Co. Inc., Teaneck, NJ, 1987.

[24] A. Müller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks.* Wiley, 2009.

[25] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* The Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, San Mateo, CA, 1988.

[26] T. Richardson and R. Urbanke. *Modern coding theory.*

Cambridge University Press, Cambridge, 2008.

[27] J. Salez. Weighted enumeration of spanning subgraphs in locally tree-like graphs. *Random Structures & Algorithms*, 2012.

[28] A. Schrijver. *Combinatorial Optimization : Polyhedra and Efficiency (Algorithms and Combinatorics)*. Springer, July 2004.

[29] J. G. Shanthikumar and D. D. Yao. The preservation of likelihood ratio ordering under convolution. *Stochastic Processes and their Applications*, 23(2):259–267, 1986.

[30] J. Yedidia, W. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282 – 2312, july 2005.

5 Appendix:

The main theorem follows quite straightforwardly from Propositions 4.10 and 4.11. The missing steps are standard and can be found in [19]; they resemble much the computation done in the proof of Proposition 4.10.

Proof. (Proof of Theorem 2.1) Propositions 4.10 and 4.11 together give that

$$\begin{aligned}\lim_{n\to\infty}\frac{2M(G_n)}{|A_n|+|B_n|} &= \inf_{\boldsymbol{\alpha}=\mathcal{S}_G\circ\mathcal{S}_G(\boldsymbol{\alpha})}\int F_r(\boldsymbol{\alpha})d\rho([G,r]) \\ &= \int \mathcal{D}_r[\mathbf{m}^{(\infty)}]d\rho([G,r])\end{aligned}$$

We introduce the probability measures ρ^A and ρ^B on $\mathcal{U}$ by conditioning on the root being in A or B:

$$\rho^A([G,r]) = \rho([G,r])\mathbf{1}(r\in A)\frac{\mathbb{E}[D^A]+\mathbb{E}[D^B]}{\mathbb{E}[D^B]},$$

and similarly for ρ^B.

For $\lambda\in\mathbb{R}^+$, applying the MTP to $\rho\in\mathcal{U}$ with

$$f^A(G,r,v) = \frac{\sum_{x\in\mathbb{N}} x m^{(\lambda)}_{\overrightarrow{vr}}(x)\mathcal{R}_{\overrightarrow{rv}}[\mathbf{m}^{(\lambda)}](x)}{\sum_{x\in\mathbb{N}} m^{(\lambda)}_{\overrightarrow{vr}}(x)\mathcal{R}_{\overrightarrow{rv}}[\mathbf{m}^{(\lambda)}](x)}\mathbf{1}(r\in A),$$

we obtain

$$\begin{aligned}&\int \mathcal{D}_r[\mathbf{m}^{(\lambda)}]d\rho^A([G,r]) \\ =\ & \frac{\mathbb{E}[D^A]+\mathbb{E}[D^B]}{\mathbb{E}[D^B]}\int\sum_v f^A(G,r,v)d\rho([G,r]) \\ =\ & \frac{\mathbb{E}[D^A]+\mathbb{E}[D^B]}{\mathbb{E}[D^B]}\int\sum_v f^A(G,v,r)d\rho([G,r]) \\ =\ & \frac{\mathbb{E}[D^A]+\mathbb{E}[D^B]}{\mathbb{E}[D^B]}\int\sum_v f^B(G,r,v)d\rho([G,r]) \\ =\ & \frac{\mathbb{E}[D^A]}{\mathbb{E}[D^B]}\int \mathcal{D}_r[\mathbf{m}^{(\lambda)}]d\rho^B([G,r])\end{aligned}$$

Letting $\lambda\to\infty$ yields in turn

$$\begin{aligned}\int \mathcal{D}_r[\mathbf{m}^{(\infty)}]d\rho^A([G,r]) &= \frac{\mathbb{E}[D^A]}{\mathbb{E}[D^B]}\int \mathcal{D}_r[\mathbf{m}^{(\infty)}]d\rho^B([G,r]) \\ &= \lim_{n\to\infty}\frac{M(G_n)}{|A_n|}.\end{aligned}$$

We then follow exactly the steps in the proof of Proposition 4.10 for ρ^A instead of ρ. This gives

$$\begin{aligned}&\int \mathcal{D}_r[\mathbf{m}^{(\infty)}]d\rho^A([G,r]) \\ =\ & \inf_{\boldsymbol{\alpha}=\mathcal{S}_G\circ\mathcal{S}_G(\boldsymbol{\alpha})}\left\{\int \min(b_r,|\boldsymbol{\alpha}_{\overrightarrow{\partial r}}|)d\rho^A([G,r])\right. \\ +\ & \left.\frac{\mathbb{E}[D^A]}{\mathbb{E}[D^B]}\int (b_r-|\boldsymbol{\alpha}_{\overleftarrow{\partial r}}|)^+d\rho^B([G,r])\right\}\end{aligned}$$

As G is an unimodular tree, for any vertex $v\in V$, all the components of $\boldsymbol{\alpha}_{\overrightarrow{\partial v}}$ can be chosen independently (as they are independent in $\boldsymbol{\alpha}^{(\infty)}_{\overrightarrow{\partial v}}$, which achieves the infimum). Then, for $\overrightarrow{e}$ incoming to v, $\alpha_{\overrightarrow{e}}$ is determined only from the subtree stemming from the tail of $\overrightarrow{e}$; furthermore it satisfies $\alpha_{\overrightarrow{e}} = \mathcal{S}_{\overrightarrow{e}}\circ\mathcal{S}_{\partial\overrightarrow{e}}[\boldsymbol{\alpha}]$. However, the distribution of the subtree at the tail of an $\overrightarrow{e}'$ which is an input to $\mathcal{S}_{\partial\overrightarrow{e}}$ is the same as that of the subtree at the tail of $\overrightarrow{e}$, by the two-step branching property of the bipartite Galton-Watson tree G. This implies that, for $\overrightarrow{e}$ incoming to a root $r\in A$, $\alpha_{\overrightarrow{e}}$ is solution of the two-step RDE given in the statement of the theorem. As detailed in Lemma 6 of [3], there is actually a one-to-one mapping between the solutions of $\boldsymbol{\alpha} = \mathcal{S}_G\circ\mathcal{S}_G[\boldsymbol{\alpha}]$ on a Galton-Watson tree G and the solutions of the RDE considered here. This completes the proof.

We now show how Theorem 3.1 follows from Theorem 2.1. The proof follows the same lines as that in [19].

Proof. (Proof of Theorem 3.1) For any h-uniform hypergraph H_n on n vertices, we let $G_n = (A_n\cup B_n, E_n)$ be the associated bipartite graph, where B_n contains the vertices of H_n and A_n the hyperedges of H_n. Let $|B_n| = n$, and $|A_n| = m = \lfloor\tau n\rfloor$ for some τ. First-of-all, it is clear by coupling that $\tau\mapsto\mathcal{M}(\Phi^A,\Phi^B_\tau)$ as defined in Theorem 2.1, is a non-decreasing function. Let then $\tau > \tau^*_{h,k,l,r}$. Then, by Theorem 2.1, we have

$$\lim_{n\to\infty}\frac{M(G_n)}{|A_n|} < l,$$

which immediately implies that G_n is a.a.s. not (k,l,r)-orientable.

Let now $\tau < \tau^*_{h,k,l,r}$. According to Theorem 2.1 again, we have $\lim_{n\to\infty}\frac{M(G_n)}{|A_n|} = l$ but there may still exist $\mathrm{o}(n)$ hyperedges which are not (l, r)-oriented. We will then rely on specific properties of $H_{n,m,h}$ to show that a.a.s. all hyperedges are (l, r)-oriented. We follow here a similar path as in [14, 19]. It is easier to work with a different model of hypergraphs, that we call $H_{n,p,h}$, and that is essentially equivalent to the $H_{n,\lfloor \tau n \rfloor,h}$ model [15]: each possible h-hyperedge is included independently with probability p, with $p = \tau h / \binom{n-1}{h-1}$.

We let $\tilde{\tau}$ be such that $\tau < \tilde{\tau} < \tau^*_{h,k,l,r}$, and consider the bipartite graph $\tilde{G}_n = (\tilde{A}_n \cup B_n, \tilde{E}_n)$ obtained from $H_{n,\tilde{p},h}$ with $\tilde{p} = \tilde{\tau} h / \binom{n-1}{h-1}$. Consider a maximum allocation $\tilde{\mathbf{x}} \in \mathbb{N}^{\tilde{E}_n}$ of $\tilde{G}_n$. We say that a vertex of $w \in \tilde{A}_n$ (resp. a vertex $w \in B_n$) is covered if $\sum_{e \in \partial w} \tilde{x}_e = l$ (resp. $\sum_{e \in \partial w} \tilde{x}_e = k$); we also say that an edge $e \in \tilde{E}_n$ is saturated if $\tilde{x}_e = c$.

Let v be a vertex in $\tilde{A}_n$ that is not covered. We define $K(v)$ as the minimum subgraph of $\tilde{G}_n$ such that:

- v belongs to $K(v)$;
- all the unsaturated edges adjacent to a vertex in $\tilde{A}_n \cap K(v)$ belong to $K(v)$ (and thus their endpoints in B_n also belongs to $K(v)$);
- all the edges e for which $\tilde{x}_e > 0$ and that are adjacent to a vertex in $B_n \cap K(v)$ belong to $K(v)$ (and so do their endpoints in $\tilde{A}_n$).

The subgraph $K(v)$ defined in this way is in fact constitued of v and all the paths starting from v and alternating between unsaturated edges and edges e with $\tilde{x}_e > 0$ (we call such a path an alternating path). It is then easy to see that all the vertices in $B_n \cap K(v)$ must be covered, otherwise we could obtain a strictly larger allocation by applying the following change: take the path $(e_1, \dots, e_{2t+1})$ between v and an unsaturated vertex in $B_n \cap K(v)$; add 1 to each $\tilde{x}_{e_i}$ for i odd, and remove 1 from each $\tilde{x}_{e_i}$ for i even; all these changes are possible due to the way the edges in $K(v)$ have been chosen, and the resulting allocation has size larger by 1 than $|\tilde{\mathbf{x}}|$.

We will now show that the subgraph $K(v)$ is dense, in the sense that the average induced degree of its vertices is strictly larger than 2. We first show that all the vertices in $K(v)$ have degree at least 2. We have $(h-1)r \geq l$ and v is not covered, hence v has at least two adjacent edges in $\tilde{G}_n$ which are not saturated, thus the degree of v in $K(v)$, written $\deg_{K(v)} v$, is at least 2. Let w be a vertex in $B_n \cap K(v)$. By definition, there is an edge $e \in \partial w \cap K(v)$ through which w is reached from v in an alternating path, and $\tilde{x}_e < r$. Then, because $\sum_{e \in \partial w} \tilde{x}_e = k$ and $k \geq r$ there must be another edge e' adjacent to w such that $\tilde{x}_{e'} > 0$; such an edge belongs to $K(v)$ and thus w is at least of degree 2 in $K(v)$. Let now w be a vertex in $\tilde{A}_n \cap K(v)$, $w \neq v$. By definition, there must exist an edge $e \in \partial w \cap K(v)$ such that $\tilde{x}_e > 0$. Because $(h-1)r \geq l$ and $\tilde{x}_e > 0$ there must be another edge e' adjacent to w such that $\tilde{x}_{e'} < r$; e' belongs to $K(v)$ and thus $\deg_{K(v)} w \geq 2$.

Consider a path $(e_1 = (v_1 v_2), \dots, e_t = (v_t v_{t+1}))$ in $K(v)$ such that $v_1 \in \tilde{A}_n \cap K(v)$ and any two consecutive edges in the path are distinct. We will show that at least one vertex out of $2r$ consecutive vertices along this path must have degree at least 3 in $K(v)$, by showing that $\tilde{x}_{e_{2(i+1)+1}} < \tilde{x}_{e_{2i+1}}$ provided $v_{2(i+1)}$ and $v_{2(i+1)+1}$ have degree 2 in $K(v)$ for all i. $v_{2(i+1)} \in B_n \cap K(v)$ must be covered, so if $\deg_{K(v)} v_{2(i+1)} = 2$ we must have $\tilde{x}_{e_{2(i+1)}} = k - \tilde{x}_{e_{2i+1}}$. Then, if $\deg_{K(v)} v_{2(i+1)+1} = 2$, all the edges adjacent to $v_{2(i+1)+1}$ except $e_{2(i+1)}$ and $e_{2(i+1)+1}$ must be saturated, thus we must also have $(h-2)r + \tilde{x}_{e_{2(i+1)}} + \tilde{x}_{e_{2(i+1)+1}} \leq l$. This immediately yield $\tilde{x}_{e_{2(i+1)+1}} + \{k + (h-2)r - l\} \leq \tilde{x}_{e_{2i+1}}$, and thus $\tilde{x}_{e_{2(i+1)+1}} < \tilde{x}_{e_{2i+1}}$ as claimed. But $\tilde{x}_{e_{2i+1}} < r$ and so $\tilde{x}_{e_{2i+2r+1}} \leq -1$ if the hypothesis that all the vertices encountered meanwhile have degree 2 in $K(v)$ is correct, which is thus not possible. Note that we did not need to assume that the path considered was vertex-disjoint, hence it is not possible that $K(v)$ is reduced to a single cycle.

We will now count vertices and edges of $K(v)$ in a way that clearly shows that the number of edges in $K(v)$ is at least γ times its number of vertices, with $\gamma > 1$. We can always see $K(v)$ as a collection P of edge-disjoint paths, with all vertices interior to a path of degree 2 in $K(v)$ and the extremal vertices of a path having degree at least 3 in $K(v)$. To form $K(v)$ we would simply need to merge the extremal vertices of some of these paths. We have shown before that each path in P has at most $2r$ vertices. Let $p = (e_1 = (v_1 v_2), \dots, e_t = (v_t v_{t+1}))$ be a path in P, we let $\theta_E(p) = t$ be the number of edges in p and $\theta_V(p) = \sum_{e_i \in p} \frac{1}{\deg_{K(v)} v_i} + \frac{1}{\deg_{K(v)} v_{i+1}}$ be a partial count of the vertices in p (all the interior vertices are counted as 1 but the extremal vertices are only partially counted in $\theta_V(p)$, as they belong to many different paths). We have $\theta_V(p) = t - 1 + \frac{1}{\deg_{K(v)} v_1} + \frac{1}{\deg_{K(v)} v_{t+1}} \leq t - 1 + \frac{2}{3}$. Hence,

$$\frac{\theta_E(p)}{\theta_V(p)} \geq \frac{t}{t - 1 + \frac{2}{3}} \geq \frac{1}{1 - \frac{1}{6r}} > 1.$$

Furthermore, it is easy to see that

$$\begin{aligned} \sum_{p \in P} \theta_E(p) &= \text{number of edges in } K(v), \\ \sum_{p \in P} \theta_V(p) &= \text{number of vertices in } K(v), \end{aligned}$$

which shows that the number of edges in $K(v)$ is at least $\gamma = \frac{1}{1-\frac{1}{6r}} > 1$ times the number of vertices in $K(v)$.

Now, it is classical that any subgraph of a sparse random graph like $\tilde{G}_n$ with a number of edges equal to at least $\gamma > 1$ times its number of vertices must contain at least a fraction $\epsilon > 0$ of the vertices of $\tilde{G}_n$, with probability tending to 1 as $n \to \infty$ (see [15, 14]). Therefore, $K(v)$ contains at least a fraction $\epsilon' > 0$ of the vertices in $\tilde{A}_n$.

There exists a natural coupling between $H_{n,p,h}$ and $H_{n,\tilde{p},h}$: we can obtain $H_{n,p,h}$ from $H_{n,\tilde{p},h}$ by removing independently each hyperedge with probability $\tilde{p} - p > 0$. This is equivalent to removing independently with probability $\tilde{p} - p$ each vertex in $\tilde{A}_n$. We let $\text{gap}_n = l|\tilde{A}_n| - M(\tilde{G}_n) = \text{o}(n)$. For any uncovered vertex v in $\tilde{A}_n$ we can construct a subgraph $K(v)$ as above. If we remove a vertex w in $\tilde{A}_n \cap K(v)$ for such a v, then either this vertex w is itself uncovered, and then gap_n is decreased by at least 1, or w is covered and then it must belong to an alternating path starting from v and we can construct a new allocation with size equal to that of $\tilde{x}$ and in which w is uncovered and there is one more unit of weight on one of the edges adjacent to v, hence removing w will also reduce gap_n by 1. We proceed as follows: we attach independently to each hyperedge a of $H_{n,\tilde{p},h}$ a uniform $[0,1]$ random variable U_a. To obtain $H_{n,p,h}$ we remove all hyperedges a such that $U_a \leq \tilde{p} - p$. This can be done sequentlially by removing at each step the hyperedge corresponding to the lowest remaining U_a. Then, at each step, assuming there are still uncovered vertices v in $\tilde{A}_n$ we can consider the union K of the subgraphs $K(v)$, which has size at least $\epsilon' \tau n$. Hence, with positive probability the hyperedge removed will decrease the value of gap_n. By Chernoff's bound, the number of hyperedges removed is at least $\tau n \frac{\tilde{p}-p}{2}$ with high probability as $n \to \infty$, therefore gap_n will reach 0 with high probability as $n \to \infty$ before we remove all the hyperedges that should be removed. Hence, $H_{n,p,h}$ (and thus $H_{n,\lfloor \tau m \rfloor,h}$) is (k,l,r)-orientable a.a.s.