Contributions to Discrete Mathematics

SZEMERÉDI'S REGULARITY LEMMA REVISITED

TERENCE TAO

ABSTRACT. Szemerédi's regularity lemma is a basic tool in graph theory, and also plays an important role in additive combinatorics, most notably in proving Szemerédi's theorem on arithmetic progressions [19], [18]. In this note we revisit this lemma from the perspective of probability theory and information theory instead of graph theory, and observe a slightly stronger variant of this lemma, related to similar strengthenings of that lemma in [1]. This stronger version of the regularity lemma was extended in [21] to reprove the analogous regularity lemma for hypergraphs.

1. INTRODUCTION

Szemerédi's regularity lemma, introduced by Szemerédi in [19], is a fundamental tool in graph theory, and more precisely in the theory of very large, dense graphs. Roughly speaking, it asserts that given any such large dense graph G, and given an error tolerance $0 < \varepsilon \ll 1$, one can approximate G by a much simpler object, namely a partition of the vertex set into $O_\varepsilon(1)$ classes, together with some edge densities between atoms of this partition, such that the approximation is "ε-regular" on most pairs of this partition; we will formalize these notations shortly. This lemma can thus be viewed as a structure theorem for large dense graphs, approximating such graphs to any specified accuracy by objects whose complexity is bounded independently of the number of vertices in the original graph.

The regularity lemma has had many applications in graph theory, computer science, discrete geometry and in additive combinatorics, see [10] for a survey. In particular, this lemma and its variants play an important role in Szemerédi's celebrated theorem [19] that any subset of the integers of positive density contain arbitrarily long arithmetic progressions. A variant of this structure theorem (also borrowing heavily from ideas in ergodic theory) was also crucial in showing in [11] that the primes contained arbitrarily long arithmetic progressions. The lemma has also had a number of generalizations to hypergraphs of varying degrees of strength, see [3],

---

[4], [5], [13], [14], [15], [9], [21]. The more recent formulations of the hypergraph lemma are in fact strong enough to rather easily imply Szemerédi's theorem on arithmetic progressions, as well as a multidimensional version due to Furstenberg and Katznelson [7]. They were also used in the recent paper [22] establishing infinitely many constellations of any given shape in the Gaussian primes.

The proof of Szemerédi's lemma is now standard in the literature. However, this standard proof is difficult to extend to the hypergraph case; a direct application of the argument does give fairly easily a regularity lemma for hypergraphs (see [3], [4]), but that lemma does not seem to be strong enough for applications such as Szemerédi's theorem or the Furstenberg-Katznelson theorem[1], except when concerning progressions or constellations consisting of at most three points (see [17]).

In this paper we shall present a slightly different way of looking at Szemerédi's regularity lemma, which we used in [21] to obtain a hypergraph regularity lemma with sufficient strength for applications to Szemerédi-type theorems. In this new perspective, one views the regularity lemma not as a structure theorem for large dense graphs, but rather as a structure theorem for events or random variables in a product probability space. This change of perspective is analogous to Furstenberg's highly successful approach to Szemerédi's theorem in [6], in which the purely combinatorial result of Szemerédi was recast as a statement about recurrence for arbitrary events or random variables in a probability-preserving system. Just as Furstenberg's change of perspective allowed the powerful techniques of ergodic theory to be brought to bear on the problem, the change of perspective here allows one to employ tools from probability theory and information theory to clarify the regularity lemma. In particular we will use three very useful concepts from those theories, namely *σ-algebras* (partitions), *conditional expectation* (relative density), and *entropy* (complexity). As the parenthetical comments suggest, each of these concepts has a combinatorial analogue, however the author believes that there is some conceptual advantage to be gained by using a probabilistic and information-theoretic perspective rather than a graph-theoretic one[2]. One byproduct of this new

---

[1]The difficulty is that in the hypergraph situation, there are several levels of regularity or discrepancy that need to be controlled in order to yield a useful bound for arithmetic progressions or similar structures, and the lemma in [3] or [4] controls only one of these discrepancies. Later regularity lemmas control all of the relevant discrepancies, but there are some non-trivial technical issues concerning the relative sizes of the error estimates, as certain losses coming from one level of approximation must be compensated for by gains from the discrepancy bounds in other levels of approximation.

[2]The situation is somewhat analogous to that of the probabilistic method in combinatorics. While every probabilistic argument could, in principle, be written in a deterministic way (replacing expectations by averages, etc.), it is undeniable that there are significant conceptual benefits in using a "probabilistic way of thinking" to approach combinatorial problems.

perspective is that one discovers a stronger and more flexible version of the regularity lemma hiding underneath the standard one. This stronger version is difficult to state here without the requisite notational setup, but let us just say for now that it is closely related to a similar improvement of the regularity lemma discovered recently[3] in [14], in which it was observed that the regularity of the large dense graph G relative to the partition given by that lemma can be vastly improved after adding or removing a small number of edges from G. This strengthened version of the regularity lemma turns out to be quite amenable for iterating, and thus gives a relatively painless proof of the hypergraph regularity lemma; see [21].

We will turn to the details in later sections, but for now let us just give an informal discussion which already shows that the regularity lemma can be viewed in information theoretic terms rather than graph theoretic terms. It will be convenient to work with bipartite graphs. Let $G = (V_1, V_2, E)$ be a large dense bipartite graph. Let x_1 and x_2 be two vertices selected independently and uniformly at random from V_1 and V_2 respectively; thus x_1 and x_2 are independent random variables, taking values in V_1 and V_2 respectively. The edge set E can now be re-interpreted as a probabilistic event, namely the event that the pair (x_1, x_2) lies in E. We shall abuse notation and refer to this event also as E, thus E is now some event determined by the random variables x_1, x_2 (or more precisely, it lies in the σ-algebra generated by the random variables x_1 and x_2). Many of the important statistics about the edge set E can now be recast in terms of the event E; for instance, the edge density of the edge set E is equal to the probability of the event E, or equivalently the expectation of the indicator random variable 1_E. Similarly one can view relative edge densities of E as conditional expectations of 1_E.

We have already observed that E is, in principle, determined by x_1 and x_2. However, from an information-theoretic perspective this determinism relationship can be very "high-complexity" or "fine-scaled", in a sense we shall describe shortly. If the vertex sets V_1, V_2 have N elements, then the random variables x_1 and x_2 have a Shannon entropy of $\log_2 N$ (they can be described by roughly $\log_2 N$ bits each). On the other hand, the event E (or the Boolean function 1_E) has a Shannon entropy of at most $\log_2 2 = 1$ (it can be described by one bit). If N is very large, we thus see that there is much more information contained in the random variables x_1 and x_2 than is contained in the event E. To put it another way, knowing that the event E is true or false (i.e., that the pair (x_1, x_2) is an edge in G or not) does not even begin to let one determine the exact values of x_1 and x_2. Indeed, in the extreme case when the graph G is a random (or pseudorandom) graph, the event E behaves almost as if it were *independent* of the random variables

---

[3]Note added in proof: a closely related version of this lemma was recently introduced in [1], [2]. See also [12] for yet another perspective on the regularity lemma, this time from functional analysis.

x_1 and x_2, despite being actually determined by these variables. More precisely, if A_1 is any event determined by x_1 (thus A_1 can be thought of as the event that x_1 lies in a fixed subset of V_1, which by abuse of notation we shall also call A_1), and A_2 is any event determined by x_2, then in the random or pseudorandom case the event E will be almost completely uncorrelated with the events A_1, A_2. This corresponds to the well-known fact that when G is a random or pseudorandom graphs, the relative edge density between two large sets A_1, A_2 in V_1, V_2 will, with high probability, be very close to the global edge density of G. (Note that if A_1 and A_2 were small sets, i.e. events of very low probability, then the correlation, or more precisely the *mutual information*, with E would automatically be small.)

Let us summarize the above discussion in information-theoretic terms. If one is given all $\log_2 N$ bits of x_1, and all $\log_2 N$ bits of x_2, then the single-bit event E is completely determined. But if G is random or pseudorandom, and one is only given one bit of x_1 (specifically, whether x_1 lies in a fixed set A_1) and one bit of x_2, one learns almost no information about the bit E. Let us informally describe this by saying that E is approximately independent of x_1 and x_2 at "coarse scales" - when only a few bits of x_1 and x_2 are known, even though E is determined by x_1 and x_2 at "fine scales" - when most or all of the bits of x_1 and x_2 are known.

Of course, if G is not pseudorandom, then E can be highly correlated with a few special bits of x_1 and x_2. To take an extreme opposite case to the pseudorandom case, suppose that G is a complete bipartite graph connecting all the vertices of a set $A_1 \subseteq V_1$ to that of a set $A_2 \subseteq V_2$, and not connecting any other pairs of vertices. Then the event E is completely determined by one bit of x_1 (namely, whether it lies in A_1) and one bit of x_2 (namely, whether it lies in A_2).

Furthermore, it is possible for G to be a hybrid between these two extremes. Suppose now that G is a pseudorandom subgraph of the complete bipartite graph connecting A_1 to A_2. Then E is no longer determined by the one special bit of x_1 associated to A_1, and the one special bit of x_2 associated to A_2. However, it is now approximately independent at coarse scales of x_1 and x_2, *conditioning on A_1 and A_2*. In other words, once the events A_1 and A_2 are known to be true or false, the event E is then approximately independent to any further bits of information arising from x_1 and x_2. In graph theory terms, this means that when restricting V_1 to A_1 or its complement, and restricting V_2 to A_2 or its complement, the induced subgraph of G behaves pseudorandomly (with some edge density depending on which sets were being restricted to).

The information-theoretic version of the Szemerédi regularity lemma is an assertion, roughly speaking, that every event E is a hybrid of the two extremes in the sense given above. Very informally, given any two high-entropy random variables x_1 and x_2, and given any event E, it is possible to find some low-entropy random variable Z_1 determined by x_1, and a low-entropy random variable Z_2 determined by x_2, such that E is approximately

independent of x_1 and x_2 conditioning on Z_1 and Z_2. Again being very informal, this means that there exist a small number of bits from x_1 and x_2 which correlate with E, and such that no further bits from x_1 and x_2 have much of a correlation with E. Interestingly, this formulation of the regularity lemma requires no independence properties of x_1 and x_2, and also does not require E to be determined by x_1 and x_2; but we do not know any applications of this more general version.

One can view the low-entropy random variables Z_1, Z_2 discussed above as "approximations" to the event E, where the approximation is in some coarse information-theoretic sense. It turns out that the proof of the regularity lemma (see Lemma 4.3 below) in fact yields *two* such approximations, a "coarse approximation" Z_1, Z_2 and a "fine approximation" Z'_1, Z'_2. The coarse approximation has low entropy. The fine approximation has significantly higher entropy, but it is an exceedingly accurate approximation to E; in particular, any error arising from this approximation can exceed any losses coming from the entropy of the coarse approximation, in a way which can be made precise using a "growth function" $F : \mathbf{R}^+ \to \mathbf{R}^+$. Finally, the coarse and fine approximations will be close to each other, both in an L^2 sense, and also in an information theoretic sense. We will make these statements more precise later, however we remark for now that the presence of the new parameter F, used to compare the accuracy of the fine approximation against the entropy of the coarse approximation, is very suitable for iteration purposes, and allows one to extend the regularity lemma to the hypergraph setting, in which one has multiple random variables $x_1, \dots, x_d$ instead of just two, and furthermore one is interested in low-entropy approximations to an event which arise not only from individual random variables x_i, but also from joint random variables such as (x_i, x_j) (and the approximations coming from the joint random variables should themselves be approximated by other, lower-order random variables). See [21]. A closely related regularity lemma, which also involves an arbitrary growth function F, has also recently appeared in [1] in applications to property testing.

2. A PROBABILISTIC FORMULATION

Before we give the rigorous information-theoretic version of the Szemerédi regularity lemma, let us first give a standard formulation of the lemma, and also a probabilistic formulation which can be viewed as bridging the graph-theoretic version and the information-theoretic version[4] of

---

[4]We say a formulation is "probabilistic" if it involves such concepts as probability spaces, σ-algebras, random variables, (conditional) expectation, and correlation. We say a formulation is "information-theoretic" if it involves such concepts as probability spaces, σ-algebras, random variables, (conditional) entropy, and mutual information. Clearly these two perspectives share much in common, for instance the concept of independence is important in both.

the lemma. We begin with the graph-theory version; again, it is convenient to restrict ones attention to bipartite graphs.

We use $O(X)$ to denote any quantity bounded in magnitude by CX for some absolute constant $C > 0$, and more generally we use $O_{a_1,\ldots,a_k}(X)$ to denote any quantity bounded in magnitude by $C(a_1,\ldots,a_k)X$, where $C(a_1,\ldots,a_k) > 0$ depends on the parameters $a_1,\ldots,a_k$. We also use $|A|$ to denote the cardinality of a finite set A.

Definition 2.1. A *bipartite graph* is a triplet (V_1, V_2, E) where V_1, V_2 are two finite non-empty sets, and $E \subset V_1 \times V_2$. If $\varepsilon > 0$, we say that a bipartite graph (V_1, V_2) is *ε-regular* if we have

$$(1) \qquad |E \cap (A_1 \times A_2)| = \frac{|A_1 \times A_2|}{|V_1 \times V_2|}|E| + O(\varepsilon|V_1 \times V_2|)$$

for all $A_1 \subseteq V_1$ and $A_2 \subseteq V_2$.

Remark 2.2. While we assert that (1) holds for all subsets A_1, A_2 of V_1, V_2, this condition is only non-trivial for large subsets; it holds trivially when $|A_1 \times A_2| = O(\varepsilon|V_1 \times V_2|)$. Thus this definition of ε-regularity is essentially equivalent to other formulations of regularity in the literature in which a lower bound is imposed on the size of A_1 and A_2.

Theorem 2.3 (Szemerédi regularity lemma, graph-theoretic version). *Let (V_1, V_2, E) be a bipartite graph, and let $0 < \varepsilon \leq 1$. Assume that V_1 and V_2 are large depending on ε, thus $|V_1|, |V_2| \geq O_\varepsilon(1)$. Then there exists a positive integer $J = O_\varepsilon(1)$ and decompositions*

$$V_i = V_{i,0} \cup V_{i,1} \cup \ldots \cup V_{i,J}$$

for $i = 1, 2$ with the following properties:
- *(Exceptional set) For all $i = 1, 2$, we have $|V_{i,0}| = O(\varepsilon|V_i|)$.*
- *(Uniform partition) For all $i = 1, 2$ and $1 \leq j \leq J$ we have $|V_{i,j}| = |V_{i,j'}|$.*
- *(Regularity) The induced bipartite graph $(V_{1,j_1}, V_{2,j_2}, E \cap (V_{1,j_1} \times V_{2,j_2}))$ is ε-regular for all but $O(\varepsilon J^2)$ of the pairs $1 \leq j_1 \leq M, 1 \leq j_2 \leq J$.*

Remark 2.4. The bound $J = O_\varepsilon(1)$ is a little deceptive, as it conceals the fact that J can in fact be extremely large depending on $1/\varepsilon$, indeed there are examples where J grows like an exponential tower of height equal to some power of $1/\varepsilon$ (see [8]). However, the key point is that the bound on J does not depend on the cardinality of V_1 or V_2. Indeed we shall shortly give a probabilistic formulation in which V_1 and V_2 could be infinite (cf. [12]).

We now give a probabilistic generalization of the above regularity lemma. We first recall some standard notation from probability theory.

Definition 2.5 (Probability space). A *probability space* is a triple $(\Omega, \mathcal{B}_{\max}, \mathbf{P})$, where Ω is a set (called the *sample space*), $\mathcal{B}_{\max}$ is a σ-algebra[5] of sets of Ω

---

[5]A σ-algebra is a collection $\mathcal{B}$ of sets in the probability space Ω which is closed under (countable) unions, intersections, and complements, and contains the empty set and Ω. In

(the elements of $\mathcal{B}_{\max}$ being the *events*), and $\mathbf{P}$ is a probability measure on $\mathcal{B}_{\max}$ (thus it is non-negative and has total mass one). A *random variable* is any measurable map $X : \Omega \to K$ to some space K (which will typically either be a finite set, or the real line). We let $L^1(\mathcal{B}_{\max})$ denote the space of real-valued, absolutely integrable random variables; as is customary we identify two random variables if they agree outside of an event of zero probability. If $X \in L^1(\mathcal{B}_{\max})$, we let $\mathbf{E}(X)$ denote the expectation of X. In particular, if E is an event, then $\mathbf{E}(1_E) = \mathbf{P}(E)$.

Remark 2.6. For application to the regularity lemma, Ω will be a finite set, and $\mathcal{B}_{\max}$ will be the algebra of all subsets of Ω, so there will be no issues as to whether a random variable is measurable or integrable. However, it is interesting to note that the arguments we give below extend with no difficulty whatsoever to the case of infinite probability spaces.

Example 2.7. Our primary application will be to bipartite graphs, say between two vertex classes V_1 and V_2. In this case we can take $\Omega = V_1 \times V_2$, $\mathcal{B}_{\max}$ to be the power set of Ω (thus all subsets of Ω are measurable events), and $\mathbf{P}$ to be the uniform probability measure on Ω; this corresponds to the operation of sampling two vertices x_1 and x_2 uniformly and independently at random from V_1 and V_2 respectively. In this case, all functions $X : V_1 \times V_2 \to \mathbf{R}$ are measurable, and the expectation is just the average value on $V_1 \times V_2$.

A crucial concept from probability theory is that of *conditional expectation*.

Definition 2.8 (Conditional expectation). Let $(\Omega, \mathcal{B}_{\max}, \mathbf{P})$ be a probability space, and let $\mathcal{B}$ be a sub-σ-algebra of $\mathcal{B}_{\max}$. If we let $L^2(\mathcal{B})$ be the Hilbert space of $\mathcal{B}$-measurable, square-integrable real-valued random variables, with the usual norm $\|X\|_{L^2(\mathcal{B})} := \mathbf{E}(|X|^2)^{1/2}$, then $L^2(\mathcal{B})$ is a closed subspace of $L^2(\mathcal{B}_{\max})$, and we let $X \mapsto \mathbf{E}(X|\mathcal{B})$ be the associated orthogonal projection map from $L^2(\mathcal{B}_{\max})$ to $L^2(\mathcal{B})$; thus for any square-integrable random variable $X \in L^2(\mathcal{B}_{\max})$, $\mathbf{E}(X|\mathcal{B})$ will be a square-integrable $\mathcal{B}$-measurable random variable.

The conditional expectation can be defined explicitly in the case when $\mathcal{B}$ is finite, which is in fact the only case we will need in this paper. In such a case, the σ-algebra $\mathcal{B}$ is generated by a finite number of disjoint events $A_1, \ldots, A_n$ of positive probability, possibly together with some additional events of zero probability which we can safely ignore. If $X \in L^2(\mathcal{B}_{\max})$, the conditional expectation $\mathbf{E}(X|\mathcal{B})$ will be equal (almost surely) to $\mathbf{E}(X|A_i) := \frac{1}{\mathbf{P}(A_i)} \mathbf{E}(X 1_{A_i})$ on each event A_i.

---

the our applications $\mathcal{B}$ will typically be finite, in which case it can be identified with a finite partition $\Omega = \Omega_1 \cup \ldots \cup \Omega_M$ of the underlying probability space. Indeed, the cells of this partition are the atoms (minimal non-empty elements) of $\mathcal{B}$, while $\mathcal{B}$ itself consists of all the sets which are unions of zero or more atoms in the partition.

Next, we define the complexity of a σ-algebra, which is a simplified version of the Shannon entropy.

Definition 2.9 (Complexity). Let $\mathcal{B}$ be a finite σ-algebra in a probability space $(\Omega, \mathcal{B}_{\max}, \mathbf{P})$. Then the *complexity* $\mathrm{complex}(\mathcal{B})$ of $\mathcal{B}$ is defined as the least number of events needed to generate $\mathcal{B}$ as a σ-algebra.

Informally, a finite σ-algebra of complexity M can be described using M bits of information (equivalently, it contains at most 2^M atoms).

If $\mathcal{B}, \mathcal{B}'$ are two sub-σ-algebras of $\mathcal{B}_{\max}$, we let $\mathcal{B} \vee \mathcal{B}'$ denote the smallest σ-algebra which contains both $\mathcal{B}$ and $\mathcal{B}'$. Note that if $\mathcal{B}$ and $\mathcal{B}'$ are finite, then $\mathcal{B} \vee \mathcal{B}'$ is also finite, with the sub-additivity property

$$\mathrm{complex}(\mathcal{B} \vee \mathcal{B}') \leq \mathrm{complex}(\mathcal{B}) + \mathrm{complex}(\mathcal{B}').$$

Example 2.10. We continue the running example in Example 2.7. Any partition $V_1 = V_{1,1} \cup \ldots \cup V_{1,M}$ of the first vertex class induces a partition $V_1 \times V_2 = (V_{1,1} \times V_2) \cup \ldots \cup (V_{1,M} \times V_2)$ of the probability space Ω and hence creates a sub-σ-algebra $\mathcal{B}_1$ of $\mathcal{B}_{\max}$, which in information-theoretic terms captures the information of which cell of the partition the first vertex x_1 belongs to. The complexity of $\mathcal{B}_1$ is essentially $\log_2 M$. If we have another partition $V_2 = V_{2,1} \cup \ldots \cup V_{2,M}$ of the second vertex class we can form another σ-algebra $\mathcal{B}_2$, and thence create the joint σ-algebra $\mathcal{B}_1 \vee \mathcal{B}_2$, whose atoms are pairs $V_{1,i} \times V_{2,j}$ and whose complexity is essentially $2\log_2 M$ (assuming for sake of discussion that all the cells in the partitions are non-empty). If $X : V_1 \times V_2 \to \mathbf{R}$ is any random variable (which one can think of as a weight function assigning a number to each putative edge (x_1, x_2)), the conditional expectation $\mathbf{E}(X | \mathcal{B}_1 \vee \mathcal{B}_2)$ is then the function which on each pair of cells $V_{1,i} \times V_{2,j}$ is equal to the relative density $\frac{1}{|V_{1,i}||V_{2,j}|} \sum_{x_1 \in V_{1,i}} \sum_{x_2 \in V_{2,j}} X(x_1, x_2)$ of X on this pair of cells. We remark that when X is the indicator function $X = 1_E$ of a graph, the L^2 norm of this conditional expectation (which we shall refer to here as the *energy*) is a familiar concept in the standard treatment of the regularity lemma and is usually referred to as the *index* of the partitions $\mathcal{B}_1, \mathcal{B}_2$.

We now give a probabilistic Szemerédi regularity lemma, which we state in considerably more generality than we need to establish Theorem 2.3.

Theorem 2.11 (Szemerédi regularity lemma, probabilistic version).
Let $(\Omega, \mathcal{B}_{\max}, \mathbf{P})$ be a probability space, let $(\mathcal{B}_{i,\max})_{i \in I}$ be a finite collection of sub-σ-algebras of $\mathcal{B}_{\max}$, and let $X \in L^2(\mathcal{B}_{\max})$ be a random variable with $\|X\|_{L^2(\mathcal{B}_{\max})} \leq 1$. Let $\varepsilon > 0$ be a number, let $m \geq 0$, and let $F : \mathbf{R}^+ \to \mathbf{R}^+$ be an arbitrary monotone increasing function. Then there exists finite sub-σ-algebras $\mathcal{B}_i \subseteq \mathcal{B}'_i \subseteq \mathcal{B}_{i,\max}$ for each $i \in I$, and a non-negative real number[6] M, obeying the following bounds:

---

[6]It may be helpful to the reader to think of M as simply being the quantity $\max_{i \in I}(m, \mathrm{complex}(\mathcal{B}_i))$. Thus the upper bound on M translates to an upper bound on the complexity of the coarse partitions $\mathcal{B}_i$, while the estimate (3) asserts, roughly speaking,

- *(Size of M) We have $M \geq m$ and $M = O_{\varepsilon,F,m}(1)$.*
- *(Complexity bound) We have $\mathrm{complex}(\mathcal{B}_i) \leq M$ for all $i \in I$.*
- *(Coarse and fine approximations are close) We have*

$$
(2) \qquad \left\| \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i') - \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i) \right\|_{L^2(\mathcal{B}_{\max})} \leq \varepsilon.
$$

- *(Fine approximation is extremely accurate) For any collection $(A_i)_{i \in I}$ of events with $A_i \in \mathcal{B}_{i,max}$ for all $i \in I$, we have*

$$
(3) \qquad \left| \mathbf{E} \left((X - \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i')) \prod_{i \in I} 1_{A_i} \right) \right| \leq \frac{1}{F(M)}.
$$

Remark 2.12. In the application to Theorem 2.3, we will only need this theorem in the special case when $X = 1_E$ is an indicator function, when $I = \{1, 2\}$, when $\mathcal{B}_{1,max}$, $\mathcal{B}_{2,\max}$ are finite and independent, with each atom having equal probability, F is essentially the exponential function, and $\mathcal{B}_{\max} = \mathcal{B}_{1,\max} \vee \mathcal{B}_{2,\max}$. However the more general version above is no harder to prove than this special case. One can also generalize to the case when $X = (X_1, \ldots, X_n)$ is vector-valued, taking values in $\mathbf{R}^n$; on the graph level, this would correspond to regularizing n graphs simultaneously using a single partitioning of the vertex classes. This vector-valued generalization is useful for iteration purposes, in order to easily obtain the corresponding hypergraph regularity lemma; this generalization is implicit in [21].

Remark 2.13. Informally, this theorem starts with a square-integrable random variable X, and some reference σ-algebras $\mathcal{B}_{i,\max}$. It then creates two approximations to X, namely a coarse approximation $\mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i)$ and a fine approximation $\mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i')$. The coarse approximation depends on only M "bits" of information from each of the $\mathcal{B}_{i,\max}$, where M is a quantity for which we have some bounds. The fine approximation is rather close to the coarse approximation in $L^2(\mathcal{B}_{\max})$ norm. Finally, the fine approximation is extremely accurate, in the sense that adding an additional bit of information from each of the $\mathcal{B}_{i,\max}$ can only create an additional correlation of at most $1/F(M)$, where $F(M)$ is a function of M which can be specified in advance to be as rapidly growing as one pleases. (Of course, there is a price to pay in selecting a function F which grows too rapidly, which is that the upper bound on M will deteriorate.) Somewhat remarkably, no independence or dependence assumptions between X and the $\mathcal{B}_{i,\max}$ need to be made in order for this theorem to be applicable.

We will prove Theorem 2.11 in the next section. For the remainder of this section, we show how Theorem 2.11 implies Theorem 2.3.

---

that the accuracy of the fine partitions exceeds the complexity of the coarse partitions (and also exceeds any specified constant m) by an arbitrary growth function F.

Proof of Theorem 2.3 assuming Theorem 2.11. Let $G = (V_1, V_2, E)$ be a bipartite graph, thus E can be viewed as a subset of $V_1 \times V_2$. We then define a probability space by setting the sample space $\Omega := V_1 \times V_2$, setting the σ-algebra $\mathcal{B}_{\max} = 2^{\Omega}$ be the space of all subsets of Ω, and setting $\mathbf{P}$ be the uniform probability measure on Ω. In particular, E is now an event in $\mathcal{B}_{\max}$. As mentioned in the introduction, this probability space corresponds to the space generated by selecting vertices x_1, x_2 from V_1, V_2 independently and uniformly. We then set $I := \{1, 2\}$, and set $\mathcal{B}_{1,\max} := \{A_1 \times V_2 : A_1 \subseteq V_1\}$ and $\mathcal{B}_{2,\max} := \{V_1 \times A_2 : A_2 \subseteq V_2\}$, thus $\mathcal{B}_{1,\max}$ and $\mathcal{B}_{2,\max}$ are the σ-algebras generated by the random variables x_1 and x_2 respectively. We set $X := 1_E$; clearly $\|X\|_{L^2(\mathcal{B}_{\max})} \leq 1$.

We now apply Theorem 2.11, with the growth function $F : \mathbf{R}^+ \to \mathbf{R}^+$ to be chosen later, and ε replaced by $\varepsilon^{3/2}$. This gives us some σ-algebras $\mathcal{B}_1 \subseteq \mathcal{B}_1' \subseteq \mathcal{B}_{1,\max}$ and $\mathcal{B}_2 \subseteq \mathcal{B}_2' \subseteq \mathcal{B}_{2,\max}$ and a non-negative quantity $M = O_{F,\varepsilon}(1)$ such that

(4) $$\mathrm{complex}(\mathcal{B}_1), \mathrm{complex}(\mathcal{B}_2) \leq M$$

(5) $$\|\mathbf{E}(1_E | \mathcal{B}_1' \vee \mathcal{B}_2') - \mathbf{E}(1_E | \mathcal{B}_1 \vee \mathcal{B}_2)\|_{L^2(\mathcal{B}_{\max})} \leq \varepsilon^{3/2}$$

(6) $$\left| \mathbf{E}\left((1_E - \mathbf{E}(1_E | \mathcal{B}_1' \vee \mathcal{B}_2')) 1_{A_1 \times A_2} \right) \right| \leq \frac{1}{F(M)}$$

for all $A_1 \subseteq V_1, A_2 \subseteq V_2$.

Now let J be a large integer to be chosen later; we will eventually show $J = O_\varepsilon(1)$. By hypothesis we may take $|V_1|, |V_2| > J$. For each $i \in \{1, 2\}$, the finite σ-algebras $\mathcal{B}_i$ consists of at most 2^M atoms, thanks to (4). Then we can subdivide each of these atoms arbitrarily into sets of size $\lfloor \frac{|V_i|}{(1+O(\varepsilon))J} \rfloor$, plus an error of size $O(|V_i|/J)$. Combining all of the errors into a single exceptional set $V_{i,0}$, we obtain a partition

$$V_i = V_{i,0} \cup V_{i,1} \cup \ldots V_{i,J},$$

where the sets $V_{i,1}, \ldots, V_{i,J}$ all have the same cardinality (comparable to $|V_i|/J$), and each lies in an atom of $\mathcal{B}_i$, and the exceptional set $V_{i,0}$ obeys the bounds

$$|V_{i,0}| = O(\varepsilon|V_i|) + O(2^M |V_i|/J).$$

Thus, if we choose J to be the nearest integer to $2^M/\varepsilon$, we obtain $|V_{i,0}| = O(\varepsilon|V_i|)$ as desired. Also we observe that since $M = O_{F,\varepsilon}(1)$, we have $J = O_{F,\varepsilon}(1)$.

Now consider an induced bipartite graph $G_{j_1,j_2} := (V_{1,j_1}, V_{2,j_2}, E \cap (V_{1,j_1} \times V_{2,j_2}))$ where $1 \leq j_1, j_2 \leq J$. Suppose we wish to show that G_{j_1,j_2} is ε-regular, thus

$$|E \cap (A_1 \times A_2)| = \frac{|E \cap (V_{1,j_1} \times V_{2,j_2})|}{|V_{1,j_1} \times V_{2,j_2}|} |A_1 \times A_2| + O(\varepsilon|V_{1,j_1}||V_{2,j_2}|)$$

whenever $A_1 \subseteq V_{1,j_1}$ and $A_2 \subseteq V_{2,j_2}$. By the triangle inequality (and by specializing the estimate below to the case $A_1 = V_{1,j_1}, A_2 = V_{2,j_2}$), it suffices

to find a quantity d which is independent of A_1, A_2 (but which depends on E, V_{1,j_1}, V_{2,j_2}) such that

$$|E \cap (A_1 \times A_2)| = d|A_1 \times A_2| + O(\varepsilon|V_{1,j_1}||V_{2,j_2}|)$$

whenever $A_1 \subseteq V_{1,j_1}$ and $A_2 \subseteq V_{2,j_2}$. Dividing by $|V_1||V_2|$, we can rewrite this as

$$\mathbf{E}((1_E - d)1_{A_1 \times A_2}) = O(\varepsilon/J^2).$$

Observe that $A_1 \times A_2$ is contained in a single atom of $\mathcal{B}_1 \vee \mathcal{B}_2$. Thus we may take $d := \mathbf{E}(1_E|\mathcal{B}_1 \vee \mathcal{B}_2)$ on this atom. Our task is thus to establish

$$\mathbf{E}((1_E - \mathbf{E}(1_E|\mathcal{B}_1 \vee \mathcal{B}_2))1_{A_1 \times A_2}) = O(\varepsilon/J^2).$$

From (6) we have

$$\mathbf{E}((1_E - \mathbf{E}(1_E|\mathcal{B}_1' \vee \mathcal{B}_2'))1_{A_1 \times A_2}) = O(1/F(M))$$

and so if we choose $F(M) := 2^{2M}/\varepsilon^3$ then we have

$$\mathbf{E}((1_E - \mathbf{E}(1_E|\mathcal{B}_1' \vee \mathcal{B}_2'))1_{A_1 \times A_2}) = O(\varepsilon^3/2^{2M}) = O(\varepsilon/J^2).$$

Note that we now have $J = O_{F,\varepsilon}(1) = O_\varepsilon(1)$ as desired. Thus, in order to establish ε-regularity of G_{j_1,j_2}, it suffices by the triangle inequality to establish that

$$\mathbf{E}(|\mathbf{E}(1_E|\mathcal{B}_1' \vee \mathcal{B}_2') - \mathbf{E}(1_E|\mathcal{B}_1 \vee \mathcal{B}_2)|1_{V_{1,j_1} \times V_{2,j_2}}) = O(\varepsilon/J^2).$$

Note that $\mathbf{E}(1_{V_{1,j_1} \times V_{2,j_2}}) = O(1/J^2)$. Thus by Cauchy-Schwarz, it would thus suffice to show that

$$(7) \qquad \mathbf{E}(|\mathbf{E}(1_E|\mathcal{B}_1' \vee \mathcal{B}_2') - \mathbf{E}(1_E|\mathcal{B}_1 \vee \mathcal{B}_2)|^2 1_{V_{1,j_1} \times V_{2,j_2}}) = O(\varepsilon^2/J^2).$$

On the other hand, from (5) we have

$$\mathbf{E}(|\mathbf{E}(1_E|\mathcal{B}_1' \vee \mathcal{B}_2') - \mathbf{E}(1_E|\mathcal{B}_1 \vee \mathcal{B}_2)|^2) = O(\varepsilon^3).$$

Thus there are at most $O(\varepsilon J^2)$ pairs (j_1, j_2) for which (7) fails. Thus we have ε-regularity for all but at most $O(\varepsilon J^2)$ pairs, as desired. $\qquad \square$

Remark 2.14. It is clear from the argument that we can enforce a lower bound on the number J of partitions, simply by setting the parameter m equal to a large number rather than equal to zero, since this will give a lower bound for M and hence for J. Of course, this will also increase the lower bound required for $|V_1|, |V_2|$, although in applications the cases when $|V_1|$ or $|V_2|$ are small tend to be fairly easy (and the regularity lemma is of little use in such situations anyway). Also, by considering multiple vertex sets $(V_i)_{i \in I}$ instead of just two, one can prove a version of hypergraph regularity lemma (similar to the early hypergraph lemma in [3]) by a similar argument to the one given above; we omit the details. However to obtain the stronger and more modern versions of the hypergraph regularity lemma one needs to apply results such as the one above repeatedly; see [21] for more details.

3. PROOF OF THEOREM 2.11

We now give the proof of Theorem 2.11. Let us fix $(\Omega, \mathcal{B}_{\max}, \mathbf{P})$, $(\mathcal{B}_{i,\max})_{i \in I}$, X, ε, m, F. A crucial concept in the proof (as in the standard proof of the regularity lemma) will be that of the *energy* (or index) of a σ-algebra (or partition). This energy has a particularly simple description in the language of conditional expectation:

Definition 3.1. For any σ-algebra $\mathcal{B} \subseteq \mathcal{B}_{\max}$, we define the *energy* $\mathcal{E}(\mathcal{B})$ of $\mathcal{B}$ to be the quantity

$$\mathcal{E}(\mathcal{B}) := \|\mathbf{E}(X|\mathcal{B})\|^2_{L^2(\mathcal{B}_{\max})}.$$

Informally, $\mathcal{E}(\mathcal{B})$ measures how close the subspace $L^2(\mathcal{B})$ of the Hilbert space $L^2(\mathcal{B}_{\max})$ gets to containing the vector X.

Remark 3.2. In the running example of Example 2.10, with X the indicator function of a graph and $\mathcal{B} = \mathcal{B}_1 \vee \mathcal{B}_2$, the energy corresponds to the *index* of the partitions associated to $\mathcal{B}_1, \mathcal{B}_2$, as used for instance in [19].

From the hypothesis $\|X\|_{L^2(\mathcal{B}_{\max})} \leq 1$, and the fact that $X \mapsto \mathbf{E}(X|\mathcal{B})$ is an orthonormal projection we observe the estimate

$$(8) \qquad\qquad 0 \leq \mathcal{E}(\mathcal{B}) \leq 1.$$

Also, if $\mathcal{B} \subseteq \mathcal{B}'$, then a simple application of Pythagoras's theorem yields

$$(9) \qquad \mathcal{E}(\mathcal{B}') = \mathcal{E}(\mathcal{B}) + \|\mathbf{E}(X|\mathcal{B}') - \mathbf{E}(X|\mathcal{B})\|^2_{L^2(\mathcal{B}_{\max})}.$$

In particular, finer σ-algebras have higher energy.

We shall prove the regularity lemma via an *energy incrementation argument*. We shall take some σ-algebras $\mathcal{B}_i, \mathcal{B}'_i$ and see if they verify the required properties of the lemma. If they do not, we will be able to replace some of these σ-algebras by finer σ-algebras with slightly higher complexity and somewhat larger energy. The bounds (8), (9) will be used to show that this energy incrementation cannot continue indefinitely, and when it does stop, we will establish the theorem.

The key step in the argument is the following.

Lemma 3.3 (Lack of regularity implies energy increment). *Suppose we have finite σ-algebras $\mathcal{B}'_i \subseteq \mathcal{B}_{i,\max}$ and events $A_i \in \mathcal{B}_{i,\max}$ for each $i \in I$ such that*

$$\left| \mathbf{E}\left(\left(X - \mathbf{E}(X| \bigvee_{i \in I} \mathcal{B}'_i)\right) \prod_{i \in I} 1_{A_i} \right) \right| > \frac{1}{F(M)}$$

for some $M > 0$. Then if we set

$$\mathcal{B}''_i := \mathcal{B}'_i \vee \{\emptyset, A_i, \Omega \backslash A_i, \Omega\} \text{ for all } i \in I$$

(thus $\mathcal{B}''_i$ is the σ-algebra generated by $\mathcal{B}'_i$ and A_i), then we have the complexity increment

$$(10) \qquad \text{complex}(\mathcal{B}''_i) \leq \text{complex}(\mathcal{B}'_i) + 1 \text{ for all } i \in I$$

and the energy increment

$$(11) \qquad \mathcal{E}(\bigvee_{i \in I} \mathcal{B}_i'') \geq \mathcal{E}(\bigvee_{i \in I} \mathcal{B}_i') + \frac{1}{F(M)^2}.$$

Proof. The complexity increment is immediate from the definition of complexity. As for the energy increment, observe that $\prod_{i \in I} 1_{A_i}$ is measurable in $\bigvee_{i \in I} \mathcal{B}_i''$. Thus we have

$$\mathbf{E}\left((X - \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i')) \prod_{i \in I} 1_{A_i} \right) = \mathbf{E}\left((\mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i'') - \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i')) \prod_{i \in I} 1_{A_i} \right).$$

On the other hand, we clearly have $\mathbf{E}((\prod_{i \in I} 1_{A_i})^2) \leq 1$. Applying Cauchy-Schwarz, we conclude

$$\left| \mathbf{E}\left((X - \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i')) \prod_{i \in I} 1_{A_i} \right) \right|^2 \leq \left\| \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i'') - \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i') \right\|^2_{L^2(\mathcal{B}_{\max})}.$$

By hypothesis, we thus have

$$\left\| \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i'') - \mathbf{E}(X | \bigvee_{i \in I} \mathcal{B}_i') \right\|^2_{L^2(\mathcal{B}_{\max})} \geq \frac{1}{F(M)^2}.$$

The claim now follows from (9). □

We can now quickly prove Theorem 2.11. We shall run the following double-loop algorithm to generate $\mathcal{B}_i$, $\mathcal{B}_i'$, and M.

- Step 0: Initialize $\mathcal{B}_i = \mathcal{B}_i' = \{\emptyset, \Omega\}$ to be the trivial σ-algebra for each $i \in I$.
- Step 1: Set M to be the quantity

$$M := \max\left(m, \max_{i \in I} \text{complex}(\mathcal{B}_i) \right).$$

Thus, for instance, the initial value of M will be m.

- Step 2: If (3) holds, then we halt the algorithm. Otherwise, we can apply Lemma 3.3 to locate σ-algebras $\mathcal{B}_i' \subseteq \mathcal{B}_i'' \subseteq \mathcal{B}_{i,\max}$ for $i \in I$ obeying (10) and (11).
- Step 3: If we have

$$\mathcal{E}(\bigvee_{i \in I} \mathcal{B}_i'') \leq \mathcal{E}(\bigvee_{i \in I} \mathcal{B}_i) + \varepsilon^2$$

then we set $\mathcal{B}_i'$ equal to $\mathcal{B}_i''$ for each $i \in I$, and return to Step 2. Otherwise, we set $\mathcal{B}_i$ and $\mathcal{B}_i'$ *both* equal to $\mathcal{B}_i''$ for each $i \in I$, and return to Step 1.

The following observations about the above algorithm are easily verified by induction:

- At every stage of the algorithm, we have $\mathcal{B}_i \subseteq \mathcal{B}_i' \subseteq \mathcal{B}_{i,\max}$ for all $i \in I$.

- At every stage of the algorithm, we have

$$\mathcal{E}(\bigvee_{i \in I} \mathcal{B}_i') \leq \mathcal{E}(\bigvee_{i \in I} \mathcal{B}_i) + \varepsilon^2$$

  and hence by (9) we have (2).
- At every stage of the algorithm we have $m \leq M$ and $\text{complex}(\mathcal{B}_i) \leq M$ for all $i \in I$.

Thus, if the algorithm does halt (so that (3) holds), then we will have achieved every objective of Theorem 2.11, except possibly for the upper bound $M = O_{F,\varepsilon,E}(1)$ on M. Hence the only remaining task is to show that the algorithm does indeed halt in finite time with the required bound on M.

Let us first analyze the inner loop of the algorithm, which loops between Step 2 and Step 3. At the start of this inner loop (i.e. when one enters Step 2 from Step 1), the $\mathcal{B}_i'$ are equal to $\mathcal{B}_i$. At each execution of this inner loop, the energy $\mathcal{E}(\bigvee_{i \in I} \mathcal{B}_i')$ increases by at least $\frac{1}{F(M)^2}$, thanks to (11), while the complexities $\text{complex}(\mathcal{B}_i')$ increase by at most 1, thanks to (10). On the other hand, if the energy $\mathcal{E}(\bigvee_{i \in I} \mathcal{B}_i')$ ever increases by more than ε^2, then we will end the inner loop and instead trigger the outer loop (returning from Step 3 to Step 1). Thus for any fixed iteration of the outer loop, the inner loop can run for at most $F(M)^2/\varepsilon^2 + 1$ iterations, and the complexity of the σ-algebras $\mathcal{B}_i'$ increase by at most $F(M)^2/\varepsilon^2 + 1$ when doing so. In particular, the inner loop always terminates in finite time.

Now we can analyze the outer loop. At the beginning of this loop, the $\mathcal{B}_i$ are equal to the trivial algebra, and M is equal to m. After each iteration of this outer loop, each $\mathcal{B}_i$ is replaced by a σ-algebra $\mathcal{B}_i''$ whose complexity is at most $F(M)^2/\varepsilon^2 + 1$ more than the complexity of $\mathcal{B}_i$. In particular, the complexity of the new value of $\mathcal{B}_i$ is at most $M + F(M)^2/\varepsilon^2 + 1$, which causes the new value of M to be bounded by $M + F(M)^2/\varepsilon^2 + 1$. Also, the energy $\mathcal{E}(\bigvee_{i \in I} \mathcal{B}_i)$ of $\mathcal{B}_i$ will increase by at least ε^2. From (8) we thus see that the outer loop can execute at most $\lfloor 1/\varepsilon^2 \rfloor$. Thus the algorithm terminates in finite time, and the final value of M is bounded by the quantity obtained by applying $\lfloor 1/\varepsilon^2 \rfloor$ iterations of the map $M \mapsto M + F(M)^2/\varepsilon^2 + 1$ to m, so in particular $M = O_{F,\varepsilon,m}(1)$. This completes the proof of Theorem 2.11. □

Remark 3.4. The doubly-iterated nature of the argument, combined with the desire for the growth function F to be exponential for the application to Theorem 2.3, causes the final bounds on M (and hence on J) to be tower-exponential in $1/\varepsilon^C$ for some absolute constant C. As discussed in [8], this tower exponential bound cannot be significantly improved. However, by lowering F to linear or polynomial growth one can obtain a somewhat weaker regularity lemma, but with better bounds; see [10] for some further discussion on how one can adjust the strength of the regularity lemma to suit one's application. In the converse direction, we will need to increase F further, to tower-exponential or even faster, when we iterate this lemma

to obtain hypergraph regularity lemmas[7]. The flexibility afforded by this additional parameter F, which is not present in the usual formulation of the regularity lemma, may hopefully be useful for other applications also.

4. AN ENTROPY VARIANT OF THE REGULARITY LEMMA

One can also give a variant of the above arguments, in which the L^2 norm is replaced by the Shannon entropy. In particular, the energy incrementation argument is replaced by an entropy incrementation argument, which gives the lemma a much more information-theoretic flavour than before. As always we fix an ambient probability space $(\Omega, \mathcal{B}_{\max}, \mathbf{P})$.

Definition 4.1 (Entropy). If $\mathcal{B} \subset \mathcal{B}_{\max}$ is a finite σ-algebra, we define the *Shannon entropy* $\mathbf{H}(\mathcal{B})$ to be the quantity

$$\mathbf{H}(\mathcal{B}) := \sum_A \mathbf{P}(A) \log_2 \frac{1}{\mathbf{P}(A)}$$

where A ranges over all the atoms of $\mathcal{B}$ and we adopt the convention that $0 \log \frac{1}{0} = 0$. If X is a random variable taking only finitely many values, we define $\mathbf{H}(X) := \mathbf{H}(\mathcal{B}_X)$, where $\mathcal{B}_X$ is the σ-algebra generated by X. In other words

$$\mathbf{H}(X) := \sum_x \mathbf{P}(X = x) \log_2 \frac{1}{\mathbf{P}(X = x)}.$$

It is easy to verify that if X is a Boolean variable (only taking the values 0 and 1), then $\mathbf{H}(X)$ can be at most 1. More generally, we have the inequality

$$\mathbf{H}(\mathcal{B}) \leq \text{complex}(\mathcal{B})$$

for any finite σ-algebra $\mathcal{B}$. The quantity $\mathbf{H}(X)$ measures, roughly speaking, how much information one could learn from X. It can be viewed as a more refined version of the complexity, which is less sensitive to exceptional events of small probability than the complexity is.

In the probabilistic formulation of the regularity lemma, conditional expectation played a prominent role. In the entropy formulation, the analogous concept is *conditional entropy*.

---

[7]Basically, to obtain a satisfactory regularity control on hypergraphs, say 3-uniform hypergraphs, one has to first apply a result such as Theorem 2.11 with some growth function F^{fast} to approximate some 3-uniform object by a collection of 2-uniform σ-algebras (i.e. partitions of complete graphs into incomplete graphs). One then applies Theorem 2.11 again with another growth function F to approximate the atoms of those 2-uniform σ-algebras by some 1-uniform objects (vertex partitions). In order for the error terms to be manageable, it turns out that F^{fast} has to grow much faster than F, in fact it must essentially be an iterated version of F. See [21] for further discussion.

Definition 4.2 (Conditional entropy). If X, Y are random variables taking finitely many values, we define the conditional entropy $\mathbf{H}(X|Y)$ by the formula

$$\mathbf{H}(X|Y) := \sum_y \mathbf{P}(Y = y)\mathbf{H}(X|Y = y)$$

$$= \sum_y \mathbf{P}(Y = y) \sum_x \mathbf{P}(X = x|Y = y) \log_2 \frac{1}{\mathbf{P}(X = x|Y = y)}.$$

An equivalent definition is given by the Bayes identity

$$\mathbf{H}(X|Y) = \mathbf{H}(X, Y) - \mathbf{H}(Y).$$

The quantity $\mathbf{H}(X|Y)$ measures, roughly speaking, how much new information one could still learn from X if one already knew the value of Y (thus for instance $\mathbf{H}(X|X)$ is always zero).

Another key quantity we need is the *conditional mutual information* $\mathbf{I}(X : Y|Z)$ of three random variables X, Y, Z taking finitely many values, defined by

$$\mathbf{I}(X : Y|Z) := \mathbf{H}(X|Z) - \mathbf{H}(X|Y, Z) = \mathbf{H}(Y|Z) - \mathbf{H}(Y|X, Z);$$

informally, it measures how much knowing Y would tell one about X, or vice versa, assuming that Z is already known. A handy (and intuitive) fact is that the conditional mutual information is always non-negative; this is equivalent to the *submodularity inequality*

$$\mathbf{H}(X, Y, Z) + \mathbf{H}(Z) \leq \mathbf{H}(X, Z) + \mathbf{H}(Y, Z)$$

for entropy, and can be proven via Jensen's inequality. A more quantitative assertion of this fact is given in Lemma 4.4 below.

If X and Y are random variables, we write $X \mapsto Y$, and say that Y is *determined* by X, if $\mathcal{B}_Y \subseteq \mathcal{B}_X$. If X and Y take only finite values, then $X \mapsto Y$ is equivalent to the existence of a functional relationship $Y = f(X)$ for some deterministic function f, and is also equivalent (up to events of probability zero) to the conditional entropy $\mathbf{H}(Y|X)$ vanishing.

We now give the information-theoretic analogue of Theorem 2.11. To simplify the notation a little bit we will restrict to the case $I = \{1, 2\}$, although the generalization to more than two reference σ-algebras is not difficult.

Lemma 4.3 (Information-theoretic regularity lemma). *Let X_1, X_2, Y be random variables taking finitely many values such that $\mathbf{H}(Y) \leq m$ for some $m \geq 0$. Let $F : \mathbf{R}^+ \to \mathbf{R}^+$ be an arbitrary function, and $\varepsilon > 0$. Then there exists random variables Z_1, Z_2 (the "coarse approximation") and Z_1', Z_2' (the "fine approximation"), also taking finitely many values, with the following properties.*

- *(Determinism) We have the determinism relations*

(12) $$X_1 \mapsto Z_1' \mapsto Z_1; \quad X_2 \mapsto Z_2' \mapsto Z_2.$$

- *(Coarse approximation has bounded entropy) We have*

(13) $\mathbf{H}(Z_1, Z_2) \le \mathbf{H}(Z_1', Z_2') = O_{F,\varepsilon,m}(1).$

- *(Coarse and fine approximations are close) We have*

(14) $\mathbf{I}(Y : Z_1', Z_2' | Z_1, Z_2) \le \varepsilon.$

- *(Fine approximation is nearly optimal) For any random variables W_1, W_2 with $X_1 \mapsto W_1$ and $X_2 \mapsto W_2$ we have*

(15) $\mathbf{I}(Y : W_1, W_2 | Z_1', Z_2') \le \dfrac{\mathbf{H}(W_1, W_2)}{F(\mathbf{H}(Z_1, Z_2))}.$

Proof. To construct Z_1, Z_2, Z_1', Z_2' we perform the following "entropy incrementation" algorithm, which is closely analogous to the energy incrementation algorithm used in the proof of Theorem 2.11.

- Step 0. Initialize $Z_1 = Z_2 = 0$ (one can of course replace 0 by any other deterministic random variable).
- Step 1. Let Z_1', Z_2' be random variables which minimize the quantity

(16) $\mathbf{H}(Y | Z_1', Z_2') + \dfrac{\mathbf{H}(Z_1', Z_2')}{F(\mathbf{H}(Z_1, Z_2))}$

    subject to the constraints $X_1 \mapsto Z_1' \mapsto Z_1$ and $X_2 \mapsto Z_2' \mapsto Z_2$. (If there are several such minimizers, we select among them arbitrarily.)
- Step 2. If we have

$$\mathbf{H}(Y | Z_1, Z_2) - \mathbf{H}(Y | Z_1', Z_2') > \varepsilon$$

    then we replace Z_1, Z_2 with Z_1', Z_2' respectively, and return to Step 1. Otherwise, we terminate the algorithm.

We remark that because X_1, X_2 take only finitely many values, the number of possibilities for the random variables Z_1', Z_2' is finite up to equivalence. Hence a minimizer to the quantity (16) always exists. Intuitively, Z_1', Z_2' is constructed to capture as much information about Y as is possible while remaining determined by X_1, X_2; the slight penalty term in (16) is designed to keep some control of the entropy of Z_1', Z_2' (otherwise it would be as large as that of X_1, X_2, for which we have no bounds). Observe that every time we return from Step 2 to Step 1, the quantity $\mathbf{H}(Y | Z_1, Z_2)$ (which measures the amount of information in Y that remains to be captured by Z_1, Z_2) decreases by at least ε. On the other hand, from Jensen's inequality one can verify that

$$0 \le \mathbf{H}(Y | Z_1, Z_2) \le \mathbf{H}(Y) \le m.$$

Thus the above algorithm must halt after at most m/ε iterations. It is also clear that the random variables Z_1, Z_2, Z_1', Z_2' generated by this algorithm will obey the determinism relationships (12) and (14).

Also, if W_1, W_2 are any random variables determined by X_1, X_2 respectively, then by comparing the minimizer Z_1', Z_2' against the competitor (Z_1', W_1), (Z_2', W_2) (which obeys the required constraints), we have

$$\mathbf{H}(Y|Z_1', Z_2') + \frac{\mathbf{H}(Z_1', Z_2')}{F(\mathbf{H}(Z_1, Z_2))} \le \mathbf{H}(Y|Z_1', Z_2', W_1, W_2) + \frac{\mathbf{H}(Z_1', Z_2', W_1, W_2)}{F(\mathbf{H}(Z_1, Z_2))}.$$

Since $\mathbf{H}(Y|Z_1', Z_2') - \mathbf{H}(Y|Z_1', Z_2', W_1, W_2) = \mathbf{I}(Y : W_1, W_2|Z_1', Z_2')$ and $\mathbf{H}(Z_1', Z_2', W_1, W_2) \le \mathbf{H}(Z_1', Z_2') + \mathbf{H}(W_1, W_2)$, we obtain (15) as desired after some algebra.

Now we compare the entropies of Z_1, Z_2 and Z_1', Z_2'. Since Z_1, Z_2 obeys the constraints in the minimization problem (16), we have

$$\mathbf{H}(Y|Z_1', Z_2') + \frac{\mathbf{H}(Z_1', Z_2')}{F(\mathbf{H}(Z_1, Z_2))} \le \mathbf{H}(Y|Z_1, Z_2) + \frac{\mathbf{H}(Z_1, Z_2)}{F(\mathbf{H}(Z_1, Z_2))}.$$

As observed earlier, the first summand on either side ranges between 0 and m. Thus we have (after some rearranging)

$$\mathbf{H}(Z_1', Z_2') \le \mathbf{H}(Z_1, Z_2) + mF(\mathbf{H}(Z_1, Z_2)).$$

In particular, every time we return from Step 2 to Step 1, the quantity $\mathbf{H}(Z_1, Z_2)$ increases by at most $mF(\mathbf{H}(Z_1, Z_2))$. From Step 0, the initial value of $\mathbf{H}(Z_1, Z_2)$ is 0. Since the number of iterations is bounded by m/ε, we see that the final value of $\mathbf{H}(Z_1, Z_2)$ is bounded by a finite (but extremely large) quantity $O_{m,F,\varepsilon}(1)$ or more explicitly the value obtained after m/ε iterations of the map $M \mapsto M + mF(M)$ applied to 0. $\square$

To pass from an entropy formulation to an expectation formulation, we need a way to pass from control of entropy to control of expectations. A clue to how to do this is provided by the following observation: if $Y \mapsto Y'$ and $\mathbf{I}(X : Y|Y') = 0$, then X and Y are independent conditionally on Y'. In particular, if X takes values in a vector space, this implies that $\mathbf{E}(X|Y) = \mathbf{E}(X|Y')$. In other words, whenever $\mathbf{I}(X : Y|Y') = \mathbf{H}(X|Y') - \mathbf{H}(X|Y)$ is zero, so is $\mathbf{E}(X|Y') - \mathbf{E}(X|Y)$. This may help motivate the following lemma, which is a perturbative version of the above observation.

Lemma 4.4 (Relation between entropy and expectation). *Let X, Y, Y' be discrete random variables with $Y \mapsto Y'$, and with X taking values in the unit interval $\{-1 \le x \le 1\}$. Then we have*

$$\mathbf{E}\left(\left|\mathbf{E}(X|Y') - \mathbf{E}(X|Y)\right|\right) \le 2\mathbf{I}(X : Y|Y')^{1/2}.$$

More informally, this lemma asserts that approximate conditional independence in the entropy sense implies approximate conditional independence in an expectation sense. The bound $2\mathbf{I}(X : Y|Y')^{1/2}$ is not best possible, but any bound which decays to zero as $\mathbf{I}(X : Y|Y') \to 0$ will be sufficient for our purposes.

Proof. The basic idea is to exploit the observation that the function $x \log \frac{1}{x}$ is not only concave but also strictly concave on $[0, 1]$. Let us first verify

the lemma in the special case when Y' is deterministic (so the hypothesis $Y \mapsto Y'$ is vacuous), thus we wish to prove

$$\mathbf{E}(|\mathbf{E}(X) - \mathbf{E}(X|Y)|) \leq 2\mathbf{I}(X:Y)^{1/2}.$$

Let $1 \leq x_1, \ldots, x_n \leq -1$ be the essential range of X, and let $y_1, \ldots, y_m$ be the essential range of Y. For any $1 \leq i \leq n$ and $1 \leq j \leq m$, define the probabilities

$$p_{ij} := \mathbf{P}(X = x_i | Y = y_j)$$
$$q_j := \mathbf{P}(Y = y_j)$$
$$\overline{p_i} := \sum_{j=1}^{m} q_j p_{ij} = \mathbf{P}(X = x_i)$$

Then we observe that $0 \leq p_{ij}, q_j \leq 1$ and that $\sum_{j=1}^{m} q_j = 1$. If we define $f : [0,1] \to \mathbf{R}$ to be the function $f(x) := -x \log x$ (with the convention $f(0) := 0$), we thus have

$$\mathbf{I}(X:Y) = \mathbf{H}(X) - \mathbf{H}(X|Y)$$
$$= \sum_{i=1}^{n} (f(\overline{p_i}) - \sum_{j=1}^{m} q_j f(p_{ij})).$$

Now observe that f is concave, indeed we have $f''(x) = -1/x$ for all $x \in (0, 1]$. Thus by Taylor's theorem with remainder,

$$f(p_{ij}) \leq f(\overline{p_i}) + f'(\overline{p_i})(p_{ij} - \overline{p_i}) - \frac{1}{2}(p_{ij} - \overline{p_i})^2 / p_{ij}^*$$

where p_{ij}^* is a quantity between p_{ij} and $\overline{p_i}$. Inserting this into the preceding estimate and noting that $\sum_{j=1}^{m} q_j (p_{ij} - \overline{p_i}) = 0$, we conclude that

$$\sum_{j=1}^{m} q_j \sum_{i=1}^{n} (p_{ij} - \overline{p_i})^2 / p_{ij}^* \leq 2\mathbf{I}(X:Y).$$

Now we compute using the boundedness of x_i and Cauchy-Schwarz, as well as the crude estimate $p_{ij}^* \leq \overline{p}_i + p_{ij}$,

$$
\begin{aligned}
\mathbf{E}(|\mathbf{E}(X) - \mathbf{E}(X|Y)|) &= \sum_{j=1}^{m} q_j |\mathbf{E}(X) - \mathbf{E}(X|Y = y_j)| \\
&= \sum_{j=1}^{m} q_j |\sum_{i=1}^{n} x_i (\overline{p}_i - p_{ij})| \\
&\leq \sum_{j=1}^{m} q_j \sum_{i=1}^{n} |\overline{p}_i - p_{ij}| \\
&\leq (\sum_{j=1}^{m} q_j \sum_{i=1}^{n} |\overline{p}_i - p_{ij}|^2 / p_{ij}^*)^{1/2} (\sum_{j=1}^{m} q_j \sum_{i=1}^{n} p_{ij}^*)^{1/2} \\
&\leq [2\mathbf{I}(X : Y) \sum_{j=1}^{m} q_j \sum_{i=1}^{n} \overline{p}_i + p_{ij}]^{1/2} \\
&= 2\mathbf{I}(X : Y)^{1/2}.
\end{aligned}
$$

Now we consider the general case when Y' is not deterministic. In that case we write

$$
\mathbf{E}(|\mathbf{E}(X|Y') - \mathbf{E}(X|Y)|) = \sum_{y'} \mathbf{P}(Y' = y') \mathbf{E}(|\mathbf{E}(X|Y' = y') - \mathbf{E}(X; Y'; Y' = y')|).
$$

(Here we have taken advantage of the hypothesis $Y \mapsto Y'$.) Applying the preceding computation, we conclude

$$
\mathbf{E}(|\mathbf{E}(X|Y') - \mathbf{E}(X|Y)|) \leq \sum_{y'} \mathbf{P}(Y' = y') 2\mathbf{I}(X : Y|Y' = y')^{1/2}.
$$

Applying Cauchy-Schwarz again we conclude

$$
\mathbf{E}(|\mathbf{E}(X|Y') - \mathbf{E}(X|Y)|) \leq 2\sqrt{\sum_{y'} \mathbf{P}(Y' = y') \mathbf{I}(X : Y|Y' = y')}
$$
$$
= 2\mathbf{I}(X : Y|Y')^{1/2}
$$

as desired. □

By combining this with Lemma 4.3 it is possible to give a statement closely resembling Theorem 2.11, and which is also sufficient to imply Theorem 2.3. We omit the details.

References

[1] N. Alon, E. Fischer, M. Krivelevich, M. Szegedy, Efficient testing of large graphs, *Combinatorica* **20** (2000), 451–476.

[2] N. Alon, A. Shapira, Every monotone graph property is testable, preprint.

[3] F. Chung, Regularity lemmas for hypergraphs and quasi-randomness, *Random Struct. Alg.* **2** (1991), 241–252.

[4] P. Frankl, V. Rödl, The uniformity lemma for hypergraphs, *Graphs Combinat.* **8**(4) (1992), 309–312.

[5] P. Frankl, V. Rödl, Extremal problems on set systems, *Random Struct. Algorithms* **20** (2002), no. 2, 131-164.

[6] H. Furstenberg, Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions, *J. Analyse Math.* **31** (1977), 204–256.

[7] H. Furstenberg, Y. Katznelson, An ergodic Szemerédi theorem for commuting transformations, *J. Analyse Math.* **34** (1978), 275–291.

[8] T. Gowers, Lower bounds of tower type for Szemerédi's uniformity lemma, *Geom. Func. Anal.* **7** (1997), 322–337.

[9] T. Gowers, Hypergraph regularity and the multidimensional Szemerédi theorem, preprint.

[10] J. Komlós, M. Simonovits, Szemerédi's regularity lemma and its applications in graph theory, *Combinatorics, Paul Erdös is eighty*, Vol. 2 (Keszthely, 1993), 295–352, Bolyai Soc. Math. Stud., 2, János Bolyai Math. Soc., Budapest, 1996.

[11] B. Green, T. Tao, The primes contain arbitrarily long proper arithmetic progressions, preprint.

[12] L. Lovász, B. Szegedy, Szemerédi's regularity lemma for the analyst, preprint.

[13] B. Nagle, V. Rödl, M. Schacht, The counting lemma for regular k-uniform hypergraphs, to appear, *Random Structures and Algorithms*.

[14] V. Rödl, M. Schacht, Regular partitions of hypergraphs, preprint.

[15] V. Rödl, J. Skokan, Regularity lemma for k-uniform hypergraphs, to appear, *Random Structures and Algorithms*.

[16] V. Rödl, J. Skokan, Applications of the regularity lemma for uniform hypergraphs, preprint.

[17] J. Solymosi, Note on a generalization of Roth's theorem, Discrete and computational geometry, 825–827, *Algorithms Combin.* **25**, Springer Verlag, 2003.

[18] E. Szemerédi, On sets of integers containing no four elements in arithmetic progression, *Acta Math. Acad. Sci. Hungar.* **20** (1969), 89–104.

[19] E. Szemerédi, On sets of integers containing no k elements in arithmetic progression, *Acta Arith.* **27** (1975), 299–345.

[20] E. Szemerédi, Regular partitions of graphs, in *"Proc. Colloque Inter. CNRS"* (J.-C. Bermond, J.-C. Fournier, M. Las Vergnas, D. Sotteau, eds.) (1978), 399–401.

[21] T. Tao, A variant of the hypergraph removal lemma, preprint.

[22] T. Tao, The Gaussian primes contain arbitrarily shaped constellations, preprint.

DEPARTMENT OF MATHEMATICS, UCLA, LOS ANGELES CA 90095-1555
E-mail address: tao@@math.ucla.edu