Information Theoretic Interpretation of
Szemerédi’s Regularity Lemma

A paper in
Mathematics

by
Mohamed Nafea

Submitted in Partial Fulfillment
Of the Requirements
For the Degree of
Masters of Arts

Advisor: Prof. Jan Reimann

August 2017
Contents

1 Introduction 2

2 Definitions from Graph Theory 4

3 Szemerédi’s Regularity Lemma 7

4 Definitions from Probability and Information Theories 9

5 An Information Theoretic Variant of the Regularity Lemma 13
Abstract

Szemerédi’s regularity lemma is a fundamental tool in the theory of very large and dense graphs. In particular, it can be viewed as a structure theorem for arbitrary dense graphs which decomposes such graphs into a large number of parts such that the sub-graphs between these parts are random-like (simple-structured) on each pair, except for a small number of pairs. The regularity lemma has several variations and generalizations, of varying degrees of strength, and does have multiple applications in the fields of discrete geometry, combinatorics, and hypergraphs. In this paper, we present a different view of the regularity lemma as a structure theorem of random variables and events in a product probability space. Particularly, we focus on an interpretation of the regularity lemma using information theoretic terms. This alternative perspective of the lemma enables discovering stronger and more flexible versions of the lemma lying beneath the standard version. The key idea here is that this change of perspective allows for employing tools from information and probability theories, which in turn yield more precise characterization (bounds) on the error events than the classical combinatorial tools.
1 Introduction

Back in 1926, van der Waerden in [1] has proved that if the set of integers \( \mathbb{Z} \) is arbitrarily divided into two subclasses, then at least one of these two classes contains an arbitrarily long arithmetic progression (i.e., a sequence of numbers in which each differs from the preceding by a constant quantity, e.g., 3, 6, 9, 12, \( \cdots \)). A finite form of van der Waerden’s theorem can be stated as follows: For each integer \( n > 0 \), there exists a least integer \( f(n) \) such that if the integers 1, 2, \( \cdots \), \( f(n) \) are partitioned arbitrarily into two classes, then at least one class contains an arithmetic progression of \( n \)-terms. In an effort to derive an upper bound on \( f(n) \), Erdős and Turán in [2] have investigated the quantity \( r_k(n) \), defined as the greatest integer \( l \) for which there exists a sequence of integers \( 0 < a_1 < a_2 < \cdots < a_l \leq n \) which does not contain an arithmetic progression of \( k \) terms. Observing that \( r_k(m + n) \leq r_k(m) + r_k(n) \), Erdős and Turán showed that \( \lim_{n \to \infty} \frac{r_k(n)}{n} = c_k \) exists, and conjectured that this constant is always equal to zero, i.e., \( c_k = 0 \) for all \( k \). In other words, they conjectured that \( r_k(n) = o(n) \) for all \( k \), which means that any sequence of integers of positive density contains an arbitrarily long arithmetic progression.

In order to prove Erdős and Turán’s conjecture, Szemerédi in his seminal paper [3] has proved a lemma on bipartite graphs, which is nowadays well known as Szemerédi’s regularity lemma. Since then, Szemerédi’s regularity lemma has emerged as a fundamental tool in the theory of very large, dense graphs. The lemma has many applications in this extremal graph theory as well as in the areas of computer science and the combinatorics, see for example [4]. Roughly speaking, the lemma asserts that given any such large dense graph \( G \), and for a given error tolerance \( 0 < \epsilon \ll 1 \), the node set of \( G \) can be partitioned into \( k > 1/\epsilon \) parts such that the sub-graphs between pairs of these parts are “random-like”, or \( \epsilon \)-regular in particular, on most pairs. The lemma can hence be considered as a structural tool for large dense graphs which approximates such graphs to any specified accuracy by objects whose complexity is bounded independently of the number of vertices in the original graph.

Depending on the measure of the approximation, the lemma can take several forms, some equivalent to the original form in [3] and some are not. The lemma has also been extended to sparse graphs in [5] and had a number of generalizations to hypergraphs, of varying degrees of strength, see for examples [6–9].
In order to obtain a hypergraph regularity lemma with sufficient strength for applications to Szemerédi-type theorems\(^2\), Tao in [9] has utilized a different view of Szemerédi’s regularity lemma. In particular, Tao in [9] viewed the lemma as a structure theorem for events and random variables in a product probability space, instead of the canonical view of the lemma as a structure theorem on large dense graphs. As was mentioned by Tao in his note [10], this change of perspective allows for employing tools from probability theory to further clarify the regularity lemma, or in other words, to obtain a stronger and more flexible version of the lemma which hide underneath the standard version. In fact, the probability theoretic view of the regularity lemma is a generalization of the original graph theoretic version, where the graph theoretic version can be deduced from the probability theoretic one [10].

The probabilistic view of the regularity lemma has led Tao to further investigate an information theoretic interpretation of the lemma in [10], which is the focus of this master’s paper. The information-theoretic version of the lemma states that, roughly speaking, given any two high-entropy random variables and any event \(E\) which is fully determined by these two random variables, one can always find two low-entropy random variables determined by the original high-entropy variables such that they, in some sense, approximate the event \(E\). In other words, a high entropy random variables can be divided into multiple low-entropy variables each of which approximately determines a certain event.

The remainder of the paper is organized as follows. In Section 2, we review the needed notations and definitions in graph theory. Section 3 presents and discusses the original form of Szemerédi’s regularity lemma. In Section 4, we review some notations and definitions from probability theory and information theory. Section 5 presents the probabilistic and information theoretic generalizations of the lemma.

**Notation:** We use \(|A|\) to denote the cardinality of any finite set \(A\). \(1_A\) denotes the indicator function of the event \(A\). For and integer \(k \in \mathbb{Z}\), we use \([k]\) to denote the set of integers \(\{1, 2, \cdots, k\}\). We use \(f(X) \triangleq O(X)\) if \(f(X) \leq CX\) for some absolute constant \(C > 0\). More generally, we use \(f(x, \epsilon) \triangleq O_\epsilon(X)\) if \(f(x, \epsilon) \leq C(\epsilon)X\), where \(C(\epsilon) > 0\) depends on the parameter \(\epsilon\). We use \(f(X) \triangleq o(X)\) if \(\lim_{X \to \infty} \frac{f(X)}{X} = 0\).

---

\(^2\)Szemerédí-type theorems are structure theorems similar in spirit to the theorem that any subset of the integers of positive density contains arbitrarily long arithmetic progressions, which was proved in his celebrated paper [3].
2 Definitions from Graph Theory

In this section, we review some definitions in graph theory [11]. We begin by a formal definition of a graph:

Definition 1. A Graph: A graph \( G = (V, E) \) consists of a set of vertices (nodes) \( V \) and a set of edges \( E \subseteq V \times V \), where each edge connects two vertices.

The vertex (node) set of \( G \) is denoted by \( V(G) \) and the edge set of \( G \) is denoted by \( E(G) \). The order of a graph is the number of vertices in this graph, for example, a graph of order \( n \) has \( n \) vertices. For \( X, Y \subseteq V \), \( E(X, Y) \) denotes the set of edges with one end-node in \( X \) and the other end-node in \( Y \). For \( v \in V \), \( E(v) \) denotes the set of edges such that \( v \) is an end-node of each of these edges. \( G' = (V', E') \), with \( V' \subseteq V \) and \( E' \subseteq E \), is a sub-graph of \( G = (V, E) \), written as \( G' \subseteq G \). If \( G' \subseteq G \) contains all edges \((x, y) \in E \) with \( x, y \in V' \), then \( G' \) is an induced sub-graph. For \( U \subseteq V \), \( G(U) \) denotes the induced sub-graph over \( U \). Two vertices in a graph \( G \) are called adjacent if there is an edge \( e \in E \) which connects them. The degree of a vertex \( v \in V \), denoted by \( d_G(v) \), is the number of adjacent vertices to \( v \).

A graph \( G \) is simple if it is unweighted (no weights on its nodes or edges), undirected, and contains no loops or multiple edges between the nodes. Note that for a simple graph, the degree of the node \( v \), \( d_G(v) \), is equal to the number of edges \( E(v) \) incident at this node. A graph \( G = (V, E) \) is complete if \( E = V \times V \). Next, we define the notion of a bipartite graph as follows:

Definition 2. Bipartite graph: A bipartite graph \( G \) is a triplet \((V_1, V_2, E)\), where \( V_1 \) and \( V_2 \) are two disjoint non-empty node sets and the edge set \( E \subseteq V_1 \times V_2 \).

The edge density of a graph \( G = (V, E) \) is defined as follows:

Definition 3. Edge density: The edge density of a graph \( G = (V, E) \) is the number of edges per node, which is given by

\[
d_G(V) = \frac{|E(G)|}{|V(G)|}. \tag{1}
\]

Now, let \( G = (V, E) \) be a graph, and let \( X, Y \subseteq V \) be two disjoint subsets of the vertex set \( V \). Let \( e_G(X, Y) = |E(X, Y)| \) denote the number of edges with one end-node in \( X \) and the other end-node in \( Y \). Let \( G[X, Y] \) be the bipartite graph on the two node sets \( X \) and \( Y \) obtained by removing all edges with both end-nodes either in \( X \) or \( Y \). The edge density of the bipartite graph \( G[X, Y] \) is given by

\[
d_G(X, Y) = \frac{e_G(X, Y)}{|X||Y|}. \tag{2}
\]
We then define the notion of a quotient graph as follows:

**Definition 4.** Quotient graph: Let $G = (V, E)$ be a graph, and $\mathcal{P} \triangleq \{V_1, V_2, \cdots, V_k\}$ be a partition of the node set $V$. Let $R$ be the equivalence relation induced by the partition $\mathcal{P}$, i.e., for any $x, y \in V$, $R(x, y)$ is true if and only if $x, y \in V_i$ for some $i \in [k]$. The quotient graph $G/\mathcal{P}$ is the graph with the node set $[k]$ and the edge set $\{(\lceil u \rceil_R, \lceil v \rceil_R) : (u, v) \in E\}$, where $\lceil \cdot \rceil_R$ is the equivalence class with respect to $R$. The weighted quotient graph is defined on $[k]$ as well with node weight $|V_i|/|V|$ for node $i$ and edge weight $d_G(V_i, V_j)$ for edge $(i, j)$.

Next, we discuss the important definition of a random graph. A random graph $G$ is a general term to refer to a probability distribution over a graph $G$. There are several models which give a precise definition of a random graph, the most famous among which, and the one which is usually referred to as the random graph, is the Erdős-Rényi random graph model $G(n, \frac{1}{2})$ which is defined as follows:

**Definition 5.** Erdős-Rényi random graph: Erdős-Rényi random graph model $G(n, \frac{1}{2})$ is constructed by first considering $n$ isolated vertices, and then connecting the vertices with edges independently and uniformly at random (i.e., with probability $\frac{1}{2}$). This is equivalent to the graph $G$ of $n$ nodes which is selected uniformly at random from the set of all possible graphs on $n$ nodes.

The random graph $G(n, p)$ is defined in the same manner, with the edges are added independently with probability $p$, instead of $1/2$. Throughout this manuscript, we use the term “random graph” to denote the Erdős-Rényi random graph model $G(n, \frac{1}{2})$.

The next definition we present is the pseudo-random graph, which is, once again, a general concept that describes a graph that behaves like a random graph of the same edge density. For instance, this might represent the case that the graph has roughly the same edge density between any two large sets or that it contains roughly the same number of copies of every small graph as one would expect to find in a random graph. One common notion of pseudo-randomness is due to Andrew Thomason [12] which is the $(\rho, \beta)$-jumbled graphs:

**Definition 6.** Jumbled graphs: A graph $G = (V, E)$ is a $(\rho, \beta)$-jumbled graph if for all $X, Y \subseteq V$, we have

$$|e_G(X, Y) - \rho|X||Y|| \leq \beta \sqrt{|X||Y|}.$$  \hspace{1cm} (3)

Two relevant concepts (in fact special cases) to the jumbled graphs are the $\epsilon$-homogeneous and $\epsilon$-regular graphs, which are defined as follows:
Definition 7. Homogeneous and regular graphs: Let $G = (V_1, V_2, E)$ be a bipartite graph. The graph $G$ is said to be an $\epsilon$-homogeneous bipartite graph if the following condition holds for every $X \subseteq V_1$ and $Y \subseteq V_2$ and some $\epsilon > 0$:

$$|e_G(X, Y) - d_G(V_1, V_2)||X||Y| \leq \epsilon|V_1||V_2|.$$  \hspace{1cm} (4)

In addition, the graph $G$ is called an $\epsilon$-regular bipartite graph if the stronger condition

$$|e_G(X, Y) - d_G(V_1, V_2)||X||Y| \leq \epsilon|X||Y|$$  \hspace{1cm} (5)

holds for every $X \subseteq V_1$ and $Y \subseteq V_2$ such that $|X| > \epsilon|V_1|$ and $|Y| > \epsilon|V_2|$.

Note that the condition for an $\epsilon$-homogeneous bipartite graph in (4), while being stated for all subsets $X, Y$ of $V_1, V_2$, is non-trivial only for large subsets $X, Y$. The condition in (5) is stronger than the condition in (4), i.e., every $\epsilon$-regular bipartite graph is also $\epsilon$-homogeneous. However, for a graph to be $\epsilon$-regular, condition (5) is required to hold for large subsets $X$ and $Y$ (cf. $|X| > \epsilon|V_1|$ and $|Y| > \epsilon|V_2|$). This results in every $\epsilon$-homogeneous graph being an $\epsilon^3$-regular graph. That is, the properties of $\epsilon$-homogeneity and $\epsilon$-regularity are essentially equivalent.
Szemerédi’s Regularity Lemma

In this section, we present the original graph theoretic form of Szemerédi’s regularity lemma [3]. Roughly speaking, the regularity lemma states that the node set of every large graph $G$ has an equitable partition $P$ into a large number of classes such that the weighted quotient graph $G/P$ approximates the original graph $G$. The approximation metric utilized results in various forms of the lemma. We now state the original graph theoretic form and provide a few remarks about the lemma:

**Theorem 1.** Szemerédi’s regularity lemma for graphs [13, Lemma 9.2]:
For any $\epsilon > 0$, there exists a positive integer $S(\epsilon) \in \mathbb{N}$ such that any graph $G = (V, E)$ has an equitable partition $\{V_1, V_2, \cdots, V_k\}$, with $1/\epsilon \leq k \leq S(\epsilon)$ and for all but $\epsilon k^2$ pairs of indices $1 \leq i < j \leq k$, the bipartite graph $G[V_i, V_j]$ is $\epsilon$-regular.

**Remark 1.** In Theorem 1, the upper bound on the size of the partition, $k \leq S(\epsilon)$, is independent from the graph $G$.

**Remark 2.** In Theorem 1, the set of exceptional $\epsilon k^2$ pairs, which can be considered as error terms in the approximation of the graph $G$, contains at most $\epsilon k^2 \left(\frac{n}{k}\right)^2 = \epsilon n^2$ edges. The lower bound on the size of the partition, $k \geq 1/\epsilon$, is set to ensure that the maximum possible number of edges inside the classes of the partition, which is given by $k(n/k)^2$, is upper bounded by $\epsilon n^2$. Hence, all these edges can be considered as error terms as well.

Now, we restate the regularity lemma in a form which is more suited for our discussion. In particular, we state the lemma for a bipartite graph $G = (V_1, V_2, E)$ and using the big-$O$ notation$^4$ as follows:

**Theorem 2.** [10, Theorem 2.3]:
Let $0 < \epsilon \ll 1$ and let $G = (V_1, V_2, E)$ be a large bipartite graph $(|V_1|, |V_2| \geq O_\epsilon(1))$. Then, there exists a positive integer $S = O_\epsilon(1)$ and partitions $\mathcal{P}_1 = \{V_{1,0}, V_{1,1}, V_{1,2}, \cdots, V_{1,S}\}$ and $\mathcal{P}_2 = \{V_{2,0}, V_{2,1}, V_{2,2}, \cdots, V_{2,S}\}$ of $V_1$ and $V_2$ such that:

1. $|V_{1,0}| = O_\epsilon(1)$ and $|V_{2,0}| = O_\epsilon(1)$,
2. $\mathcal{P}_1$ and $\mathcal{P}_2$ are equitable partitions of $V_1$ and $V_2$.

$^3$A partition $\mathcal{P} = \{V_1, V_2, \cdots, V_k\}$ of $V$ is called an equitable partition if $\left\lfloor \frac{|V_i|}{k} \right\rfloor \leq |V_i| \leq \left\lceil \frac{|V_i|}{k} \right\rceil$ for all $i \in [k]$.

$^4$Recall that $O_\epsilon(X)$ denotes a function $f(X, \epsilon)$ which satisfies $f(X, \epsilon) \leq C(\epsilon)X$, where $C(\epsilon) > 0$ depends on the parameter $\epsilon$. 

7
3. The induced bipartite graph \((V_{1,i_1}, V_{2,i_2}, E(V_{1,i_1}, V_{2,i_2}))\) is \(\epsilon\)-regular for all but \(O(\epsilon S^2)\) pairs of indices \(i_1, i_2 \in [S]\).

Note that the number of atoms in each partition, \(S = O_\epsilon(1) \leq C(\epsilon)\), while being a function of \(\epsilon\), is in fact very large (depending on \(1/\epsilon\)). For instance, in [14], the size of the partition, \(S\), grows as an exponential tower of a height equals to some power of \(1/\epsilon\). Furthermore, much like the previous form, \(S\) does not depend on the number of nodes in the graph \(G\).
4 Definitions from Probability and Information Theories

Before presenting the probabilistic and information theoretic generalizations of Szemerédi’s regularity lemma in Section 5, we recall here some standard notations and definitions from probability and information theories. Let us begin with definitions from probability theory [15]:

Definition 8. **Probability space:** A probability space is a triple \((\Omega, \mathcal{F}, P)\) where \(\Omega\) is the sample space, \(\mathcal{F}\) is the sigma algebra of all possible subsets of \(\Omega\), and \(P\) is the probability measure on \(\mathcal{F}\).

Definition 9. **Random variable:** Let \((\Omega, \mathcal{F}, P)\) be a probability space. A random variable \(X\) is an \(\mathcal{F}\)-measurable function from \(\Omega\) to the real line \(\mathbb{R}\), i.e., \(X : \Omega \mapsto \mathbb{R}\).

Definition 10. **Sigma algebra generated by a random variable:** Let \((\Omega, \mathcal{F}, P)\) be a probability space. The sigma algebra generated by a random variable \(X\), denoted as \(\mathcal{F}_X\), is the smallest sigma algebra which contains the set \(\{X^{-1}(B) : \forall B \in \mathcal{B}\\}\), where \(X^{-1}(B) \triangleq \{\omega \in \Omega : X(\omega) \in B\}\).

Let \(L^1(\mathcal{F})\) denote the space of real-valued, \(\mathcal{F}\)-measurable, absolutely integrable functions. For a random variable \(X \in L^1(\mathcal{F})\), we use \(\mathbb{E}(X)\) to denote its expectation. For the event \(A \in \mathcal{F}\), we have \(\mathbb{E}(1_A) = P(A)\). We use \(\text{Supp}(X)\) to denote the support set of a discrete random variable \(X\), i.e., \(\text{Supp}(X) \triangleq \{x \in \mathcal{X} : P(X = x) > 0\}\).

Next, we present an important definition in probability theory which is the conditional expectation:

Definition 11. **Conditional expectation:** Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \(\mathcal{F}'\) be a sub-sigma algebra of \(\mathcal{F}\). Let \(L^2(\mathcal{F}')\) be the Hilbert space of \(\mathcal{F}'\)-measurable, real-valued, and square-integrable random variables, defined under the norm \(||X||_{L^2(\mathcal{F}')} = \sqrt{\mathbb{E}(|X|^2)}\). We thus have \(L^2(\mathcal{F}')\) is a closed subspace\(^5\) of \(L^2(\mathcal{F})\). Define the map \(X \mapsto \mathbb{E}(X|\mathcal{F}')\) as the orthogonal projection from \(L^2(\mathcal{F})\) to \(L^2(\mathcal{F}')\). That is, for a square-integrable \(\mathcal{F}\)-measurable random variable \(X \in L^2(\mathcal{F})\), the conditional expectation \(\mathbb{E}(X|\mathcal{F}')\) is a square-integrable \(\mathcal{F}'\)-measurable random variable.

When the sub-sigma algebra \(\mathcal{F}'\) is generated by a finite number of disjoint events \(\{A_1, A_2, \ldots, A_m\}\), the conditional expectation of the random variable \(X \in L^2(\mathcal{F})\), \(\mathbb{E}(X|\mathcal{F}')\), on each event \(A_i\), is almost surely equal to

\[
\mathbb{E}(X|A_i) = \frac{1}{P(A_i)}\mathbb{E}(X1_{A_i}), \quad (6)
\]

\(^5\)Since \(\mathcal{F}' \subseteq \mathcal{F}\), an \(\mathcal{F}'\)-measurable function is also an \(\mathcal{F}\)-measurable function.
where the “almost-surely” in this statement is in order to avoid the events with zero probability.

Next, let us define the complexity of a finite sigma algebra as follows:

**Definition 12. Complexity:** Let $\mathcal{F}$ be a sigma algebra which is generated by a finite set of events in a probability space $(\Omega, \mathcal{F}, P)$. The complexity of $\mathcal{F}$, $\text{complex}(\mathcal{F})$, is defined as the least number of events needed to generate $\mathcal{F}$.

That is, if $\text{complex}(\mathcal{F}) = M$, then $\mathcal{F}$ can be generated by $M$ events, i.e., contains at most $2^M$ atoms. For two finite sigma algebras $\mathcal{F}, \mathcal{F}'$, we have the sub-additivity property for the complexity function, i.e., $\text{complex}(\mathcal{F} \cup \mathcal{F}') \leq \text{complex}(\mathcal{F}) + \text{complex}(\mathcal{F}')$, where $\mathcal{F} \cup \mathcal{F}'$ is the smallest finite sigma algebra which contains both $\mathcal{F}$ and $\mathcal{F}'$.

Next, we proceed with basic definitions from information theory [16]. Let us first fix a probability space $(\Omega, \mathcal{F}, P)$. Unless otherwise stated, all logarithms are taken to the base-2, which sets the quantities defined below to be measured in bits. In addition, we use the convention $0 \log \frac{1}{0} = 0$. We begin with the definition of “Shannon entropy” as follows:

**Definition 13. Shannon entropy:** Let $\mathcal{F}'$ be a sub-sigma algebra of $\mathcal{F}$ which is generated by a finite set of events in the probability space $(\Omega, \mathcal{F}, P)$. Shannon entropy of $\mathcal{F}'$, denoted by $H(\mathcal{F}')$, is defined as

$$H(\mathcal{F}') \triangleq \sum_{A \in \mathcal{F}'} P(A) \log \frac{1}{P(A)},$$

(7)

where the sum is over all events $A$ (atoms of $\mathcal{F}'$). For a discrete random variable $X$ (which takes finitely many values), Shannon entropy of $X$, denoted by $H(X) = H(\mathcal{F}_X)$, where $\mathcal{F}_X$ is the sigma algebra generated by $X$, can be defined as

$$H(X) = \sum_{x \in \text{Supp}(X)} P(X = x) \log \frac{1}{P(X = x)}.$$  

(8)

Shannon entropy of a random variable $X$ measures the amount of uncertainty about (or the information contained in) this random variable. In other words, how many bits are needed to describe the random variable $X$.

Next, we define the conditional entropy of a random variable $X$ given another random variable $Y$ as follows:

**Definition 14. Conditional entropy:** Let $X$ and $Y$ be two discrete random variables. The conditional entropy of $X$ given $Y$, denoted by $H(X|Y)$, is defined
as
\begin{align}
H(X|Y) &= \sum_{y \in \text{Supp}(Y)} P(Y = y) H(X|Y = y) \\
&= \sum_{y \in \text{Supp}(Y)} P(Y = y) \sum_{x \in \text{Supp}(X)} P(X = x|Y = y) \log \frac{1}{P(X = x|Y = y)}.
\end{align}
(9)

That is, the conditional entropy \( H(X|Y) \) describes the uncertainty about (or the information contained in) the random variable \( X \) when the random variable \( Y \) is known, averaged over the probability distribution of \( Y \). Note that, \textit{conditioning does not increase entropy}. That is, \( H(X|Y) \leq H(X) \) is always true, where the equality holds if and only if \( X \) and \( Y \) are independent random variables. In addition, \( H(X|X) = 0 \). Using Baye’s rule, we can write the conditional entropy as
\begin{align}
H(X|Y) &= H(X,Y) - H(Y).
\end{align}
(10)

For two discrete random variables \( X, Y \), the \textit{mutual information} between \( X \) and \( Y \), denoted by \( I(X;Y) \), is defined as
\begin{align}
I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y).
\end{align}
(11)

(12)

Roughly speaking, the mutual information between \( X \) and \( Y \) represents the averaged reduction in the amount of uncertainty about one random variable when the other variable is given. Using Jensen’s inequality, one can easily show that, for any two random variables \( X \) and \( Y \), we have \( I(X;Y) \geq 0 \) where the equality holds if and only if \( X \) and \( Y \) are independent. Furthermore, we have \( I(X;X) = H(X) \).

In addition, for three discrete random variables \( X, Y, Z \), the \textit{conditional mutual information} between \( X \) and \( Y \) given \( Z \), denoted by \( I(X;Y|Z) \), is defined as
\begin{align}
I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) = H(Y|Z) - H(Y|X,Z) \\
&= H(X|Z) + H(Y|Z) - H(X,Y|Z).
\end{align}
(13)

(14)

Since \( I(X;Y|Z) \geq 0 \), we have \( H(X,Y|Z) \leq H(X|Z) + H(Y|Z) \), and by adding \( 2H(Z) \) to both sides, we have the \textit{sub-modularity} inequality for entropy as follows
\begin{align}
H(X,Y,Z) + H(Z) \leq H(X,Z) + H(Y,Z).
\end{align}
(15)

If \( X \) and \( Y \) are conditionally independent given \( Z \), i.e., \( X, Y, Z \) form the Markov
chain \( X - Z - Y \), we have that \( I(X;Y|Z) = 0 \).

Finally, we provide the following notation: For two random variables \( X \) and \( Y \), we use \( X \mapsto Y \) to denote that the sigma algebra generated by \( Y \) is a sub-sigma algebra of the sigma algebra generated by \( X \), i.e., \( \mathcal{F}_Y \subseteq \mathcal{F}_X \). For discrete random variables, this is equivalent to the existence of a deterministic functional mapping from \( X \) to \( Y \), i.e., \( Y = g(X) \) for some deterministic function \( g \). In addition, for discrete random variables, \( X \mapsto Y \) is equivalent to \( H(Y|X) = 0 \). When \( X \mapsto Y \) is true, we say that \( Y \) is determined by \( X \).
5 An Information Theoretic Variant of the Regularity Lemma

Tao in [9] have utilized a different view, which is in fact a generalization, of the graph theoretic Szemerédi’s regularity lemma in [3], in order to obtain a stronger version of the hypergraph extension of the lemma than the version that follows by a direct extension of the standard argument as in [6, 7]. In this new (general) perspective, the lemma is generalized to probability spaces. In particular, the vertices and edges in a graph are replaced with random variables (or more precisely, the sigma-algebras generated by these random variables) and events in a product probability space. Later in [10], Tao have further investigated a variation of the probabilistic generalization of the lemma in information theoretic terms. We proceed by presenting the probabilistic generalization of the lemma and then the information theoretic version. Before presenting these, we provide the following informal discussion [10] which shows that the regularity lemma can be viewed in information theoretic terms instead of graph theoretic terms.

Let \( G = (V_1, V_2, E) \) be a bipartite graph. Let \( x_1, x_2 \) be two vertices drawn independently and uniformly at random from \( V_1, V_2 \), respectively. That is, \( x_1, x_2 \) are two independent uniform random variables over the sets \( V_1, V_2 \).

Define the probabilistic event \( E \triangleq \{ (x_1, x_2) \in E \} \). Now, one can see that the probabilistic event \( E \) carries many of the statistics about the edge set \( E \). For example, the edge density of \( E \) is equal to the probability of \( E \) (or more precisely, \( d_G(V_1, V_2) = \mathbb{E}(1_E) \)).

Now, the event \( E \) is determined by the random variables \( x_1, x_2 \). More precisely, the event \( E \) lies in the sigma algebra generated by \( x_1, x_2 \). That is, knowing the values of \( x_1, x_2 \) determines whether \( E \) is true (\( 1_E = 1 \)) or false (\( 1_E = 0 \)). The converse, however, is far from being correct. Knowing \( E \) is true or false does not even begin to determine the values of \( x_1, x_2 \). For \( |V_1|, |V_2| \) very large, the amount of information contained in the random variables \( x_1, x_2 \) is much more than the information in \( E \).

Let us now consider two extreme cases. The first case is that when the graph \( G \) is random. For this case, the event \( E \), while being determined by full knowledge of \( x_1, x_2 \), does behave as if it is independent of these random variables when we restrict our knowledge about \( x_1, x_2 \). For example, define the events \( A_1 \triangleq \{ x_1 \in A_1, A_1 \subseteq V_1 \} \) and \( A_2 \triangleq \{ x_2 \in A_2, A_2 \subseteq V_2 \} \). When \( A_1, A_2 \) are two large subsets\(^6\) of \( V_1, V_2 \) and \( G \) is random, \( d_G(A_1, A_2) \approx d_G(V_1, V_2) \), and hence, \( A_1, A_2 \) are almost uncorrelated with \( E \). In information theoretic terms, if we know all the information about \( x_1, x_2 \) \( H(x_1) = \log_2 |V_1| \) bits of \( x_1 \) and \( H(x_2) = \log_2 |V_2| \) bits

\(^6\)Note that if \( A_1, A_2 \) are small subsets of \( V_1, V_2 \), then the correlation of \( A_1, A_2 \) with \( E \) is trivially very small.
of \( x_2 \), then the single-bit event \( \mathcal{E} \) is fully determined. However, if \( G \) is random and we are only given a few bits about \( x_1 \) and a few bits about \( x_2 \) (which may tell that \( x_1, x_2 \) are lying in certain sets \( A_1, A_2 \)), then we obtain almost no information about \( \mathcal{E} \). To summarize, roughly speaking, \( \mathcal{E} \) is independent from \( x_1, x_2 \) at a coarse-scale, i.e., when only a few bits about \( x_1, x_2 \), are known. On the contrary, \( \mathcal{E} \) is completely determined by \( x_1, x_2 \) at a fine-scale, i.e., when all the information about \( x_1, x_2 \) are known.

The second extreme case is when the sub-graph \( G[A_1, A_2] \) is a complete bipartite graph which connects all nodes in \( A_1 \) to all nodes in \( A_2 \), and all the remaining vertices in \( G \setminus G[A_1, A_2] \) are not connected. Then \( \mathcal{E} \) is completely determined with the single (or a few) bits of \( x_1 \) and \( x_2 \) which specify that \( x_1 \in A_1 \) and \( x_2 \in A_2 \).

A question which arises here is that what if \( G \) is a mixture of the two extreme cases above? For instance, think of \( G \) as a pseudo-random graph over the complete bipartite graph which connects \( A_1 \) and \( A_2 \). In this case, \( \mathcal{E} \) is no longer determined by one bit from \( x_1 \) and one bit from \( x_2 \) which tell that \( x_1 \in A_1 \) and \( x_2 \in A_2 \). However, conditioning on \( A_1, A_2 \), \( \mathcal{E} \) is almost independent from \( x_1, x_2 \) at a coarse scale. In particular, if we know that \( A_1, A_2 \) are true or false, then \( \mathcal{E} \) is almost independent from single-bit (or few-bits) events arising from \( x_1, x_2 \).

The information-theoretic version of the Szemerédi’s regularity lemma says that every event \( \mathcal{E} \) consists of a mixture of the two extremes in the sense described above. Roughly speaking, given any two high-entropy random variables \( x_1 \) and \( x_2 \), and given any event \( \mathcal{E} \), it is possible to find some low-entropy random variable \( z_1 \) determined by \( x_1 \) and some low-entropy random variable \( z_2 \) determined by \( x_2 \) such that, conditioned on \( z_1 \) and \( z_2 \), the event \( \mathcal{E} \) is almost independent of \( x_1 \) and \( x_2 \). That is, there is a small number of bits from \( x_1 \) and \( x_2 \) which correlate with \( \mathcal{E} \) such that no further bits from \( x_1 \) and \( x_2 \) have much of a correlation with \( \mathcal{E} \).

In summary, we can see that these two random variables \( z_1, z_2 \) are low-entropy “approximations” of the event \( \mathcal{E} \).

Let us now express the aforementioned discussion in more rigorous probability theoretic terms. For a probability space \((\Omega, \mathcal{F}, P)\), we can take \( \Omega = V_1 \times V_2, \mathcal{F} \) to be the sigma algebra of all subsets of \( \Omega \) (i.e., the power set of \( \Omega, \mathcal{F} = 2^\Omega \)), and \( P \) to be the uniform probability measure over \( \mathcal{F} \). This is equivalent to the random variables \( x_1, x_2 \) being independent and uniform over the sets \( V_1 \) and \( V_2 \).

Furthermore, let \( \{\mathcal{F}_1, \mathcal{F}_2\} \) be a partition of the sigma algebra \( \mathcal{F} \) such that \( \mathcal{F}_1' \subset \mathcal{F}_1 \subset \mathcal{F} \) and \( \mathcal{F}_2' \subset \mathcal{F}_2 \subset \mathcal{F} \), where \( \{\mathcal{F}_1', \mathcal{F}_2'\} \) are the sub-sigma algebras generated by partitioning the probability space \( \Omega \) into \( \{V_{1,1} \times V_2, V_{1,2} \times V_2, \ldots, V_{1,S} \times V_2\} \) and \( \{V_1 \times V_{2,1}, V_1 \times V_{2,2}, \ldots, V_1 \times V_{2,S}\} \), respectively. Thus, the joint sigma algebra \( \mathcal{F}_1' \cup \mathcal{F}_2' \) represents a partition of the probability space into the pairs \((V_{1,i_1}, V_{2,i_2})\). In information theoretic terms, the sub-sigma algebras \( \mathcal{F}_1', \mathcal{F}_2' \) describe the information about which cell of the partitions of \( V_1 \) and
V_2 do x_1 and x_2 belong to, whereas, in similar terms, the partition \{F_1, F_2\} of the sigma algebra \( \mathcal{F} \) represents the exact information about \( x_1 \) and \( x_2 \). Also note that \( \text{complex}(F'_1) = \text{complex}(F'_2) = \log S \). Now, for a random variable \( X \in L^2(\mathcal{F}) : V_1 \times V_2 \mapsto \mathbb{R} \), the conditional expectation \( \mathbb{E}(X | F'_1 \cup F'_2) \) represents an approximation of the random variable \( X \) using the sub-sigma algebra \( F'_1 \cup F'_2 \).

We are now ready to present the probabilistic generalization of the regularity lemma as follows:

**Theorem 3.** Probabilistic generalization of Szemerédi’s regularity lemma [10, Theorem 2.11]:

Let \((\Omega, \mathcal{F}, P)\) be a probability space, and let \( \{F_i\}_{i \in I} \) be a finite collection of sub-sigma algebras of \( \mathcal{F} \), where \( I \) is a finite set. Let \( X \in L^2(\mathcal{F}) \) be a random variable satisfying \( \|X\|_{L^2(\mathcal{F})} \leq 1 \). Let \( 0 < \epsilon \ll 1, m \geq 0, \) and \( F : \mathbb{R}^+ \mapsto \mathbb{R}^+ \) be an arbitrary monotonically increasing function. Then, there exist a positive real number \( M \) and the following chain of finite sub-sigma algebras \( F''_i \subseteq F'_i \subseteq F_i \) for each \( i \in I \), such that

1. \( M = O_{\epsilon,F,m}(1) \) and \( M \geq m \),

2. \( \text{complex}(F''_i) \leq M \), for all \( i \in I \),

3. We have
   \[
   \left\| \mathbb{E}\left( X \bigg\rvert \bigcup_{i \in I} F'_i \right) - \mathbb{E}\left( X \bigg\rvert \bigcup_{i \in I} F''_i \right) \right\|_{L^2(\mathcal{F})} \leq \epsilon, \tag{16}
   \]

4. For any collection of events \( \{A_i\}_{i \in I} \) such that \( A_i \in F_i \) for all \( i \in I \), we have
   \[
   \left| \mathbb{E}\left( X - \mathbb{E}\left( X \bigg\rvert \bigcup_{i \in I} F'_i \right) \right) \prod_{i \in I} 1_{A_i} \right| \leq \frac{1}{F(M)}. \tag{17}
   \]

Much like in the discussion we have given earlier in this section, \( \{F_i\}_{i \in I} \) in Theorem 3 represents a partition\(^7\) of the original sigma algebra \( \mathcal{F} \). In addition, the sub-sigma algebras \( \{F'_i\}_{i \in I} \) represent a fine approximation of the partition \( \{F_i\}_{i \in I} \) (and hence of the overall sigma algebra \( \mathcal{F} \)), while \( \{F''_i\}_{i \in I} \) represent a coarse approximation of the partition.

Having \( X \) as the random variable which represents the overall random experiment (i.e., the sigma algebra generated by \( X \) is equal to \( \mathcal{F} \)), the third item in the

\(^7\)In the example illustrated earlier, this partition was into \( \{F_1, F_2\} \), as we were attempting to map this probabilistic interpretation of the lemma to a bipartite graph.
result in Theorem 3 states, roughly speaking, that the coarse approximation of \(X\), 
\[ \mathbb{E}\left(X \mid \bigcup_{i \in I} F_i''\right), \]
is close to the fine approximation of \(X\), 
\[ \mathbb{E}\left(X \mid \bigcup_{i \in I} F_i'\right), \]
in the \(L^2\)-norm distance sense. However, the second item in the result states that only the 
complexity of the coarse approximation depends on \(M\) bits of information from 
each \(F_i\) in the partition, for which we have a bound on its size in the first item. 
That is, while having the bound only on the complexity of the coarse approximation, both approximations are close. The last item of the result states that the fine 
approximation is very accurate in the sense that adding an additional bit of infor-
mation from each of the sigma algebras of the partition \(\{F_i\}\) can only enhance the 
approximation (cause an additional correlation between \(X\) and the fine approxi-
mation \(\mathbb{E}\left(X \mid \bigcup_{i \in I} F_i'\right)\)) at most with 
\[ \frac{1}{F(M)}, \]
where \(F(M)\) is any monotonically increasing function of \(M\).

Using non-straightforward analysis, Tao showed that the graph theoretic ver-
sion of the regularity lemma in Theorem 2 can be deduced from the probability 
theoretic version in Theorem 3 \[10\]. In addition, Theorem 3 can be generalized 
to the case of a random vector \(X = \{X_1, X_2, \ldots, X_n\}\), instead of one random 
variable \(X\), which corresponds, in the graph theoretic perspective, to regularizing 
\(n\) graphs simultaneously using a single partitioning of the vertex classes.

In the following, we provide the proof for Theorem 3:

**Proof:** Fix the probability space \((\Omega, \mathcal{F}, P)\), the collection of sub-sigma algebras 
\(\{F_i\}_{i \in I}\), the random variable \(X, 0 < \epsilon \ll 1, m \geq 0\), and the monotonically 
increasing function \(F\). An important concept of the proof is the energy of a sub-sigma algebra, which is defined in terms of a conditional expectation as follows:

**Definition 15. Energy of a sub-sigma algebra:** Let \((\Omega, \mathcal{F}, P)\) be a probability 
space, and let \(X \in L^2(\mathcal{F})\) be a random variable. Let \(\mathcal{F}' \subseteq \mathcal{F}\) be a sub-sigma 
algebra. The energy of \(\mathcal{F}'\), denoted by \(\mathcal{E}(\mathcal{F}')\), is defined as

\[ \mathcal{E}(\mathcal{F}') \equiv \|\mathbb{E}(X \mid \mathcal{F}')\|_{L^2(\mathcal{F})}^2. \]  

That is, the energy of \(\mathcal{F}'\) measures the approximation of \(X\) by the subspace 
\(L^2(\mathcal{F}')\) of the Hilbert space \(L^2(\mathcal{F})\).

By using the hypothesis \(\|X\|_{L^2(\mathcal{F})} \leq 1\) of the theorem, and since \(X \mapsto \mathbb{E}(X \mid \mathcal{F}')\) is an orthonormal projection by definition of conditional expectation, it follows that

\[ 0 \leq \mathcal{E}(\mathcal{F}') \leq 1. \]
Using Pythagoras theorem, we have that, for $F'' \subseteq F'$,

$$
E(F') = E(F'') + \|\mathbb{E}(X|F') - \mathbb{E}(X|F'')\|^2_{L^2(F)},
$$

(20)

where both $X \mapsto \mathbb{E}(X|F')$ and $X \mapsto \mathbb{E}(X|F'')$ are orthogonal projections from the Hilbert space $L^2(F)$. That is, finer sub-sigma algebras have higher energy.

Now, the proof for Theorem 3 utilizes an “energy incrementation” argument, which is described as follows: We start with some sub-sigma algebras $F'_i$ and $F''_i$ for each $i \in I$. If these sigma algebras do not satisfy the conditions of the theorem, we replace them with finer sub-sigma algebra of slightly higher complexity and larger energy. Finally, the properties of the energy in (19) and (20) are utilized to show that this process can not run indefinitely, and when it halts, the properties of the lemma are satisfied. We first prove the following lemma:

**Lemma 1.** Let $\{F'_i\}_{i \in I}$ be a finite collection of finite sub-sigma algebras and $\{A_i\}_{i \in I}$ be a finite collection of events such that $F'_i \subset F_i$, $A_i \in F_i$, for each $i$, and for some $M > 0$, we have

$$
\left| \mathbb{E}\left( X - \mathbb{E}\left( X \big| \bigcup_{i \in I} F'_i \right) \bigg) \prod_{i \in I} 1_{A_i} \right) \right| > \frac{1}{F(M)}. \tag{21}
$$

Then, for each $i$, by setting the sub-sigma algebra $\tilde{F}_i$ to be the sigma algebra generated by $F'_i$ and $A_i$ (precisely, $\tilde{F}_i = F'_i \cup \{\Phi, \Omega, A_i, A_i^c\}$), then, for each $i \in I$, we have the following complexity increment

$$
\text{complex}(\tilde{F}_i) \leq \text{complex}(F'_i) + 1, \tag{22}
$$

and the following energy increment

$$
\mathcal{E}\left( \bigcup_{i \in I} \tilde{F}_i \right) \geq \mathcal{E}\left( \bigcup_{i \in I} F'_i \right) + \frac{1}{F(M)^2}. \tag{23}
$$

**Proof:** The complexity increment follows directly from the construction of $\tilde{F}_i$ and definition of complexity. For the energy increment, we have, by the hypothesis in the lemma, that

$$
\frac{1}{F(M)^2} < \left| \mathbb{E}\left( X - \mathbb{E}\left( X \big| \bigcup_{i \in I} F'_i \right) \bigg) \prod_{i \in I} 1_{A_i} \right|^2
$$
\[ \mathbb{E} \left( \mathbb{E} \left( X \mid \bigcup_{i \in I} \tilde{F}_i \right) - \mathbb{E} \left( X \mid \bigcup_{i \in I} \mathcal{F}'_i \right) \right) \prod_{i \in I} \mathbb{1}_{\mathcal{A}_i} \right)^2 \leq \left\| \mathbb{E} \left( \mathbb{E} \left( X \mid \bigcup_{i \in I} \tilde{F}_i \right) - \mathbb{E} \left( X \mid \bigcup_{i \in I} \mathcal{F}'_i \right) \right) \right\|_{L^2(\mathcal{F})}^2 \times \mathbb{E} \left( \prod_{i \in I} \mathbb{1}_{\mathcal{A}_i} \right)^2 \] (25)

\[ \leq \left\| \mathbb{E} \left( \mathbb{E} \left( X \mid \bigcup_{i \in I} \tilde{F}_i \right) - \mathbb{E} \left( X \mid \bigcup_{i \in I} \mathcal{F}'_i \right) \right) \right\|_{L^2(\mathcal{F})}^2 \] (26)

\[ = \mathbb{E} \left( \bigcup_{i \in I} \tilde{F}_i \right) - \mathbb{E} \left( \bigcup_{i \in I} \mathcal{F}'_i \right). \] (27)

The equality in (24) follows because \( \prod_{i \in I} \mathbb{1}_{\mathcal{A}_i} \) is \( \left( \bigcup_{i \in I} \tilde{F}_i \right) \)-measurable, which in turn follows from the construction of \( \tilde{F}_i \) in the lemma; \( \tilde{F}_i = \mathcal{F}'_i \cup \{ \Phi, \Omega \} \cup \{ A_i, A_i^c \} \) for each \( i \in I \). Thus, we have that \( \mathbb{E} \left( \mathbb{E} \left( X \mid \bigcup_{i \in I} \tilde{F}_i \right) \prod_{i \in I} \mathbb{1}_{\mathcal{A}_i} \right) = \mathbb{E} \left( X \prod_{i \in I} \mathbb{1}_{\mathcal{A}_i} \right) \).

The inequality in (25) follows from the Cauchy-Schwarz inequality. Equation (27) follows from applying (20) to (26).

We now proceed with the proof for Theorem 3. The energy incrementation algorithm can be described as follows:

**Step 0: Initialization:** For each \( i \in I \), initialize \( \mathcal{F}'_i'' = \mathcal{F}'_i = \{ \Phi, \Omega \} \), i.e., the trivial sigma algebra.

**Step 1:** Set \( M = \max \{ m, \max_{i \in I} \text{complex}(\mathcal{F}'_i) \} \).

**Step 2:** If (17) is satisfied, terminate the algorithm. Otherwise, by Lemma 1, there exist \( \tilde{F}_i \subseteq \mathcal{F}'_i \subseteq \mathcal{F}_i \), for each \( i \in I \), which satisfy the complexity and energy increment properties in the lemma.

**Step 3:** If we have

\[ \mathbb{E} \left( \bigcup_{i \in I} \tilde{F}_i \right) \leq \mathbb{E} \left( \bigcup_{i \in I} \mathcal{F}'_i'' \right) + \epsilon^2, \] (28)

then, for each \( i \in I \), set \( \mathcal{F}_i' = \tilde{F}_i \) and return to step 2. Otherwise, for each \( i \in I \), set \( \mathcal{F}_i' = \mathcal{F}_i'' = \tilde{F}_i \) and return to step 1.

Now, we show that when the algorithm halts, the properties of the theorem are satisfied. From step 3, we have, at each stage of the algorithm, that

\[ \mathbb{E} \left( \bigcup_{i \in I} \mathcal{F}_i' \right) \leq \mathbb{E} \left( \bigcup_{i \in I} \mathcal{F}_i'' \right) + \epsilon^2. \] (29)

18
Thus, using (20), the third item of the theorem is satisfied:

$$\left\| \mathbb{E} \left( X \big| \bigcup_{i \in I} \mathcal{F}_i' \right) - \mathbb{E} \left( X \big| \bigcup_{i \in I} \mathcal{F}_i'' \right) \right\|_{L^2(\mathcal{F})}^2 = \mathcal{E} \left( \bigcup_{i \in I} \mathcal{F}_i' \right) - \mathcal{E} \left( \bigcup_{i \in I} \mathcal{F}_i'' \right) \leq \epsilon^2.$$ 

Note that the algorithm halts if and only if the fourth item of the theorem is satisfied. In addition, from step 1 of the algorithm, we have that $m \leq M$ and $\text{complex}(\mathcal{F}_i'') \leq M$ for all $i \in I$, at each stage of the algorithm. Thus, it remains to show that $M = O_{\epsilon,F,m}(1)$ and that the algorithm indeed halts in a finite number of iterations.

The algorithm described above has two loops. An inner loop which is moving from step 3 back to step 2 and an outer loop which is moving from step 3 back to step 1. When the algorithm goes to step 2 for the first time (i.e., from step 1), we have that $\mathcal{F}_i' = \mathcal{F}_i''$. While running in the inner loop, i.e., from step 3 back to step 2, the energy $\mathcal{E} \left( \bigcup_{i \in I} \mathcal{F}_i' \right)$ increases at least by $\frac{1}{F(M)^2}$ and the complexities $\{\text{complex}(\mathcal{F}_i)\}_{i \in I}$ increase at most by 1, due to Lemma 1. The inner loop ends and the outer loop begins when the energy $\mathcal{E} \left( \bigcup_{i \in I} \mathcal{F}_i' \right)$ increases by more than $\epsilon^2$. That is, for each fixed iteration of the outer loop, the inner loop runs at most for $F(M)^2/\epsilon^2 + 1$ iterations, and the complexity increases at most by $F(M)^2/\epsilon^2 + 1$. Thus, the inner loop indeed runs for a finite number of iterations.

Now, for the outer loop, after each iteration of this loop, the sigma algebra $\mathcal{F}_i''$ is replaced by the sigma algebra $\tilde{\mathcal{F}}_i$ whose complexity is at most $F(M)^2/\epsilon^2 + 1$ more than $\mathcal{F}_i''$. More specifically, the new value of $\mathcal{F}_i''$ has a complexity at most $M + F(M)^2/\epsilon^2 + 1$, and hence the updated $M$, due to step 1, is bounded by $M + F(M)^2/\epsilon^2 + 1$. Also, after each iteration of the outer loop, the energy $\mathcal{E} \left( \bigcup_{i \in I} \mathcal{F}_i'' \right)$ increases at least by $\epsilon^2$, and hence, due to (19), the outer loops runs for at most $1/\epsilon^2$ iterations. Thus, the algorithm terminates in a finite time and the final value of $M$ is bounded by the quantity obtained by applying $\frac{1}{\epsilon}$ iterations of the map $M \mapsto M + F(M)^2/\epsilon^2 + 1$, i.e., $M = O_{\epsilon,F,m}(1)$, which completes the proof for Theorem 3.

Next, we present the information theoretic analogue of Theorem 3 as follows:

**Theorem 4.** Information theoretic variant of Szemerédi’s regularity lemma [10, Lemma 4.3]:

Let $X_1, X_2, Y$ be discrete random variables such that $H(Y) \leq m$ for some $m \geq 0$. Let $0 < \epsilon \ll 1$, and $F : \mathbb{R}^+ \mapsto \mathbb{R}^+$ be an arbitrary monotonically increasing function. Then, there exist discrete random variables $Z_1, Z_2$ (coarse approximation) and $Z'_1, Z'_2$ (fine approximation) such that

1. We have the deterministic relations $X_1 \mapsto Z'_1 \mapsto Z_1$ and $X_2 \mapsto Z'_2 \mapsto Z_2$,
2. \( H(Z_1, Z_2) \leq H(Z'_1, Z'_2) = O_{\epsilon,F,m}(1) \) (i.e., the coarse approximation has a bounded entropy),

3. We have
\[
I(Y; Z_1, Z_2| Z'_1, Z'_2) \leq \epsilon.
\] (30)

4. For any random variables \( W_1, W_2 \) such that the deterministic relations \( X_1 \mapsto W_1 \) and \( X_2 \mapsto W_2 \) hold, we have
\[
I(Y; W_1, W_2| Z'_1, Z'_2) \leq \frac{H(W_1, W_2)}{F(H(Z_1, Z_2))}.
\] (31)

Much like the probability theoretic generalization of the regularity lemma in Theorem 3, in Theorem 4, the random variables \( \{X_1, X_2\} \) represent a partition of the random variable \( Y \), and the random variables \( Z_1, Z_2 \) represent a coarse approximation of \( X_1, X_2 \) (and hence of \( Y \)), while the random variables \( Z'_1, Z'_2 \) represent a fine approximation of \( X_1, X_2 \).

The first item of the result above states that the coarse approximation \( Z_1, Z_2 \) is determined by (a deterministic function of) the fine approximation \( Z'_1, Z'_2 \), while the fine approximation is a deterministic function of of \( X_1, X_2 \). We thus have the Markov chain \((Z_1, Z_2) - (Z'_1, Z'_2) - Y\), and hence, the mutual information in the third item of the result is equal to \( H(Y|Z'_1, Z'_2) - H(Y_1|Z_1, Z_2) \). That is, the third item of the result states that \( H(Y|Z_1, Z_2) \) is at most \( \epsilon \)-far from \( H(Y|Z'_1, Z'_2) \), i.e., the fine and coarse approximations are close. Finally, the last item of the result states, once again, that the fine approximation is very accurate in the sense that any two random variables \( W_1, W_2 \) that are determined by \( X_1, X_2 \) satisfy that \( H(Y|Z'_1, Z'_2, W_1, W_2) \) is at most \( \frac{H(W_1, W_2)}{F(H(Z_1, Z_2))} \) far from \( H(Y|Z'_1, Z'_2) \).

In the following, we present the proof of Theorem 4:

**Proof:** The key idea of the proof is to follow an “entropy incrementation” algorithm in order to construct the random variables \( Z_1, Z_2, Z'_1, Z'_2 \) (the coarse and fine approximations), which is similar to the energy incrementation argument utilized in the proof for Theorem 3.

Let us first fix the function \( F \) and \( \epsilon > 0 \). The “entropy incrementation” algorithm can be described as follows:

**Step 0- Initialization:** Start by initializing \( Z_1, Z_2 \) to any deterministic values (e.g., set \( Z_1, Z_2 = 0 \)).

**Step 1- Minimization:** Choose \( Z'_1, Z'_2 \) to be the random variables which mini-
mimize

\[ H(Y|Z'_1, Z'_2) + \frac{H(Z'_1, Z'_2)}{F(H(Z_1, Z_2))}, \]

such that the deterministic relations \( X_1 \mapsto Z'_1 \mapsto Z_1 \) and \( X_2 \mapsto Z'_2 \mapsto Z_2 \) hold; If there exist several minimizers, we arbitrarily choose one of them.

**Step 2 - Entropy incrementation:** If \( H(Y|Z_1, Z_2) - H(Y|Z'_1, Z'_2) > \epsilon \), replace \( Z_1, Z_2 \) with \( Z'_1, Z'_2 \) and return to step 1. Otherwise, terminate the algorithm.

The existence of at least one minimizer in Step 1 in the algorithm follows since \( X_1, X_2 \) are discrete random variables which take finitely many values, and hence, the space of possible random variables \( Z'_1, Z'_2 \) that we minimize over is finite as well.

Note that, in step 1, \( Z'_1, Z'_2 \) are chosen to carry as much information as possible about \( Y \), while being determined by \( X_1, X_2 \). However, if we choose \( Z'_1, Z'_2 \) to minimize \( H(Y|Z'_1, Z'_2) \) only, the algorithm will end up by \( Z'_1 = X_1 \) and \( Z'_2 = X_2 \), which is not what we want. That’s why the penalty term \( \frac{H(Z'_1, Z'_2)}{F(H(Z_1, Z_2))} \) is added to the minimization, which is designed to control the entropy of the random variables \( Z'_1, Z'_2 \).

In step 2, the criterion to return to step 1 is chosen such that \( H(Y|Z_1, Z_2) \), which measures the remaining uncertainty about \( Y \) when \( Z_1, Z_2 \) are known, decreases at least by \( \epsilon \). Knowing that \( 0 \leq H(Y|Z_1, Z_2) \leq H(Y) \leq m \), the algorithm must halt after at most \( m/\epsilon \) iterations.

So far, it is clear that the first and third items in Theorem 4 are guaranteed to be satisfied when the algorithm halts, by the construction of steps 1 and 2. Next, we show that the second and fourth items are satisfied as well when the algorithm halts.

Let \( W_1, W_2 \) be any two random variables which are determined by \( X_1, X_2 \), respectively. The minimizer \( Z'_1, Z'_2 \) resulting from step 1 of the algorithm must satisfy

\[
H(Y|Z'_1, Z'_2) + \frac{H(Z'_1, Z'_2)}{F(H(Z_1, Z_2))} \leq H(Y|Z'_1, Z'_2, W_1, W_2) + \frac{H(Z'_1, Z'_2, W_1, W_2)}{F(H(Z_1, Z_2))}.
\]

(32)

Otherwise, the minimizer will be \((Z'_1, W_1), (Z'_2, W_2)\) which of course satisfies the deterministic relations \( X_1 \mapsto (Z'_1, W_1) \mapsto Z_1 \) and \( X_2 \mapsto (Z'_2, W_2) \mapsto Z_2 \) since both \( Z'_1, W_1 \) are determined by \( X_1 \) and both \( Z'_2, W_2 \) are determined by \( X_2 \).

The condition in (32) can be re-written as

\[
I(Y; W_1, W_2|Z'_1, Z'_2) = H(Y|Z'_1, Z'_2) - H(Y|Z'_1, Z'_2, W_1, W_2)
\]
\[
\begin{align*}
&\leq \frac{H(Z'_1, Z'_2, W_1, W_2) - H(Z'_1, Z'_2)}{F(H(Z_1, Z_2))} \\
&\leq \frac{H(W_1, W_2)}{F(H(Z_1, Z_2))},
\end{align*}
\]

where (34) follows since \(H(H(Z'_1, Z'_2, W_1, W_2)) \leq H(Z'_1, Z'_2) + H(W_1, W_2)\).

Thus, the fourth item of Theorem 4 is satisfied.

Finally, since \(Z_1, Z_2\) is a feasible choice in the minimization step (step 1), the minimizer \(Z'_1, Z'_2\) must satisfy

\[
\begin{align*}
H(Y|Z'_1, Z'_2) + \frac{H(Z'_1, Z'_2)}{F(H(Z_1, Z_2))} &\leq H(Y|Z_1, Z_2) + \frac{H(Z_1, Z_2)}{F(H(Z_1, Z_2))} \\
\Leftrightarrow \frac{H(Z'_1, Z'_2) - H(Z_1, Z_2)}{F(H(Z_1, Z_2))} &\leq H(Y|Z_1, Z_2) - H(Y|Z'_1, Z'_2) \\
\Leftrightarrow H(Z'_1, Z'_2) - H(Z_1, Z_2) &\leq F(H(Z_1, Z_2)) \left( H(Y|Z_1, Z_2) - H(Y|Z'_1, Z'_2) \right) \\
&\leq mF(H(Z_1, Z_2))
\end{align*}
\]

where (37) follows since \(H(Y|Z_1, Z_2) \leq H(Y) \leq m\) and \(H(Y|Z'_1, Z'_2) \geq 0\).

Thus, we have

\[
H(Z'_1, Z'_2) \leq H(Z_1, Z_2) + mF(H(Z_1, Z_2))
\]

That is, moving from step 2 to step 1 in the algorithm, where we replace \(Z_1, Z_2\) by \(Z'_1, Z'_2\), increases \(H(Z_1, Z_2)\) at most by \(mF(H(Z_1, Z_2))\). The algorithm begin initialized with deterministic \(Z_1, Z_2\), and knowing that it goes for at most \(m/\epsilon\) iterations guarantees that \(H(Z_1, Z_2) = O_{\epsilon,F,m}(1)\), which completes the proof of Theorem 4.

\[\blacksquare\]
References


