36-781 Lecture:
Limits of (dense) graph sequences

Lecturer: Cosma Shalizi
Scribe: Nicolás Kim

1 November 2016

Contents

1 Warm-up: convergence in distribution

1.1 Starting case: empirical distirbution on k categories

Represent a distribution on k categories by a vector $p_n \in \mathbb{R}^k_{\geq 0}$, where $\sum_{i=1}^k p_n = 1$. To be more precise, then, p_n lives in the k-simplex S_k:

So, p_n is any point on the k-simplex with rational coefficients. In other words, each rational point on the k-simplex represents a *distribution* on k categories. Figure 1 depicts the simplices for $k = 2$ and $k = 3$.

So, as $n \to \infty$, $p_n \to \rho$ even if ρ has irrational coefficients. This is the notion of convergence in distribution for categorical distributions.

1.2 Convergence of distributions on $\mathbb{R}^1$

Definition 1. *At any finite n, define the* empirical cdf *to be the function*

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i \leq x).$$

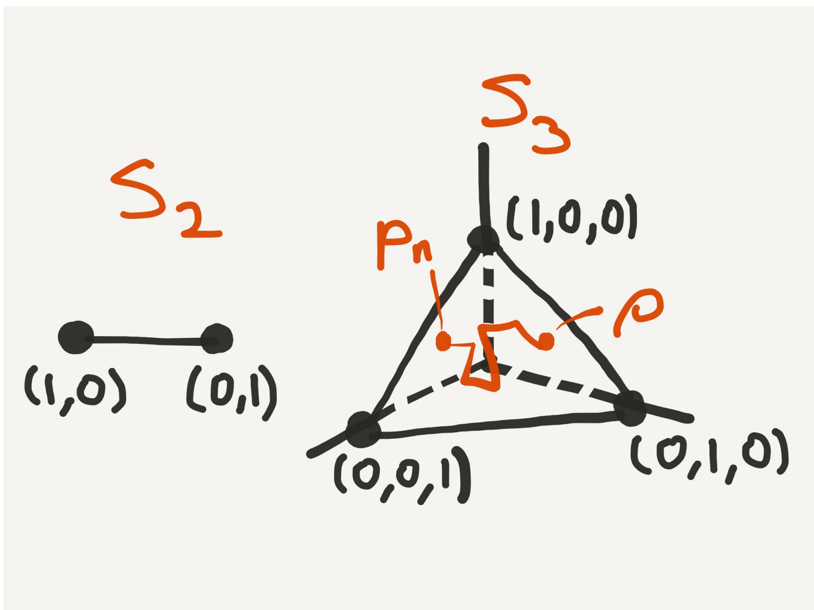

Figure 1: S_2 and S_3 are the simplices corresponding to the range of possible probability distributions on 2 and 3 categories, respectively. Each point in a simplex is a probability distribution, since each point is a vector whose entries sum to 1.

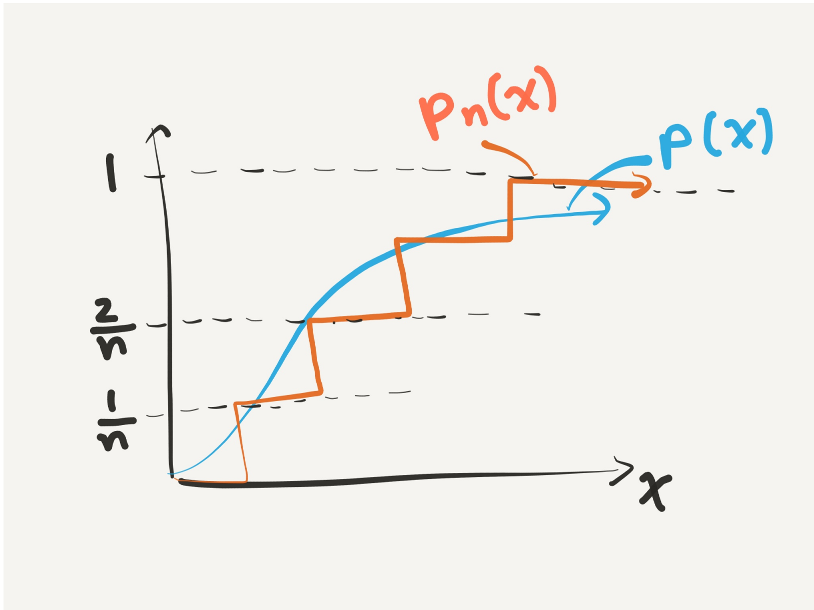

Figure 2: The empirical cdf (orange) is a step function which converges to the true cdf (blue) as the sample size increases.

So, at any finite n, $p_n(x)$ takes jumps of size $1/n$ at (at most) n distinct points. See Figure 2 for an example.

Theorem 1.1. *(Glivenko-Cantelli) If $X_i \overset{iid}{\sim} \rho =$ true cdf, then*

$$\max_x |p_n(x) - \rho(x)| \overset{a.s.}{\to} 0.$$

The Glivenko-Cantelli thoerem is also known as the *fundamental theorem of statistics*, since it essentially claims that we can use a sample from a distribution to learn that distribution.

Note that, similarly to the case of convergence of categorical distributions, ρ isn't necessarily a step function, even though p_n will be for any finite n. The possible cdf's that these step functions may converge to constitutes a broader class of functions than just step functions.

1.3 Alternative view of convergence in distribution

Definition 2. *Say we have a sequence of probability distributions $(\mu_1, \mu_2, \ldots, \mu_n, \ldots)$ with corresponding pdf's $(m_1, m_2, \ldots, m_n, \ldots)$. These converge on a limit μ with pdf m, if:*

$$\int_{\mathbb{R}} f(x) m_n(x)\, dx \overset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}} f(x) m(x)\, dx,$$

for all bounded and continuous f (such f are known as "test functions"). If this convergence holds, then we say that $\mu_n \xrightarrow{d} \mu$, i.e. that μ_n converges in distribution *to μ.*

How do we apply this notion to the convergence of empirical distributions? In particular, the empirical distribution is a jump function, meaning it does not have a pdf in the standard sense. However, we can consider the empirical pdf to be a mixture of Dirac delta functions, as follows: put μ_n to be the empirical distribution from n samples. Naturally, we let μ denote the distribution that all data were drawn from. Define the pdf of μ_n to be

$$\frac{1}{n}\sum_{i=1}^{n}\delta(x-x_i),$$

where $\delta(x)$ is the *Dirac delta function*[1],

$$\delta(x) = \begin{cases} 0, & x \neq 0, \\ +\infty, & x = 0, \end{cases}$$

so that

$$\int_{\mathbb{R}} f(x)\delta(x)\,dx = f(0)$$

for any $f : \mathbb{R} \to \mathbb{R}$.

Now, putting m_n to be this mixture of Dirac delta functions, we get

$$\int_{\mathbb{R}} f(x)m_n(x)\,dx = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} f(x)\delta(x-x_i)\,dx \tag{1}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} f(y+x_i)\delta(y)\,dy \tag{2}$$

$$= \frac{1}{n}\sum_{i=1}^{n} f(x_i). \tag{3}$$

So for $\mu_n \xrightarrow{d} \mu$, we need

$$\frac{1}{n}\sum_{i=1}^{n}\delta(x-x_i) \xrightarrow{d} \mu,$$

which we now know means that

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \xrightarrow{n\to\infty} \int_{\mathbb{R}} f(x)\,\mu(dx),$$

[1] Alternatively, $\delta(x)$ can be thought of as

$$\lim_{\sigma\to 0}\int_{\mathbb{R}} f(x)\phi(x;\sigma)\,dx,$$

where $\phi(x)$ is the pdf of a random variable with distribution $\mathcal{N}(0,\sigma)$.

for all bounded and continuous test functions f. Since

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \in \mathbb{R}, \qquad\qquad \int_{\mathbb{R}} f(x)\,\mu(dx) \in \mathbb{R},$$

this is now just convergence in the sense of a sequence of real numbers.

1.4 Lessons

1. Observed data sets get represented as objects with lots of discreteness.
2. They tend towards continuous limit objects.
3. The discrete ones are special cases of the continuous limits.
4. Convergence in distribution is equivalent to convergence of averages/integrals/expectations for bounded continuous test functions:

   $$\mu_n \xrightarrow{d} \mu \iff \int_{\mathbb{R}} f(x)\,\mu_n(dx) \to \int_{\mathbb{R}} f(x)\,\mu(dx)$$

   for all bounded and continuous f.

2 Convergence of sequences of graphs

How can we apply the idea of convergence in distribution in order to define convergence in graphs?

We have a sequence of graphs $g_1, g_2, \ldots, g_m, \ldots$. Denote by $V(g_m)$ and $E(g_m)$ the set of nodes and edges, respectively, for g_m. In other words, $E(g_m) \subseteq V(g_m) \times V(g_m)$.

Fix a favorite graph f, which we will call a *motif.* For example, f may be the triange graph K_3, or the four-cycle C_4. In general, a motif is any fixed, finite graph.

Definition 3. *An* isomorphism *between two graphs f and g is some bijective function*

$$\phi : V(f) \hookrightarrow\!\!\!\!\!\rightarrow V(g)$$

such that $(i, j) \in E(f)$ if and only if $(\phi(i), \phi(j)) \in E(g)$. We will sometimes say $f \simeq g$ if f and g are isomorphic in this sense. An example is given in Figure 3.

Definition 4. *The* density *of a k-node motif, f, in an n-node graph, g, is the fraction of k-node induced subgraphs in g that are isomorphic to f (by definition, the density is 0 if $k \geq n$).*

These motifs are going to be our test functions. Figure 4 demonstrates this concept when the motif is a 2-star.

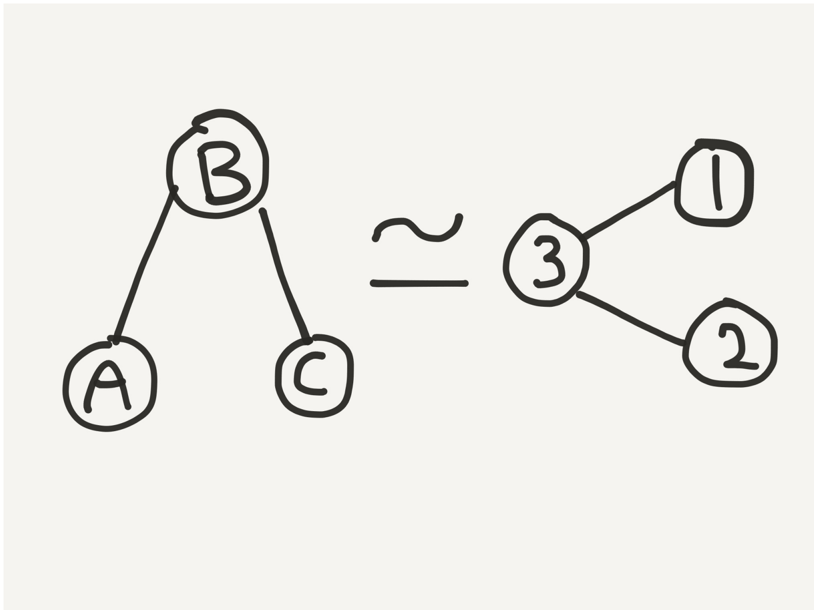

Figure 3: Graph isomorphism is determined by the graphs' structure, not by the label associated with each node.

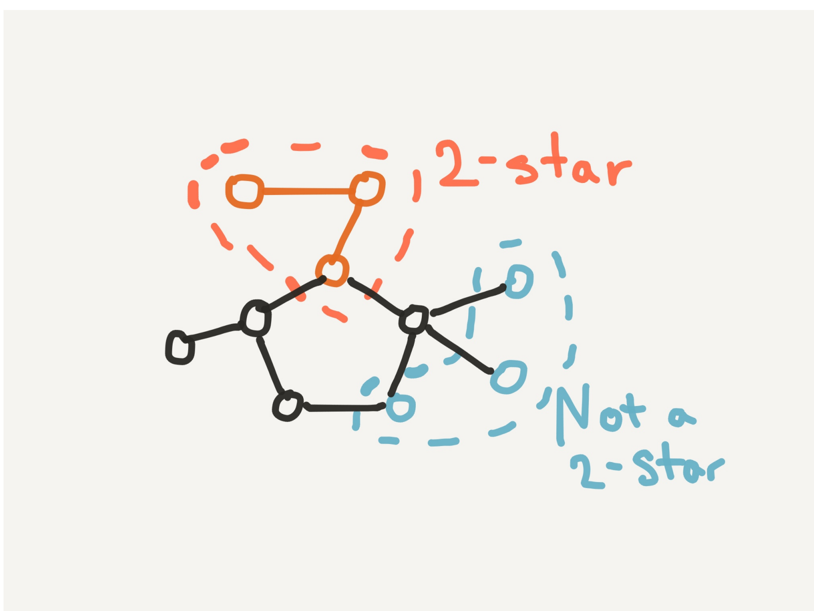

Figure 4: Take three arbitrary nodes in g: do they have the same structure as f? Here, we identify two successful matchings between a triplet in g and f.

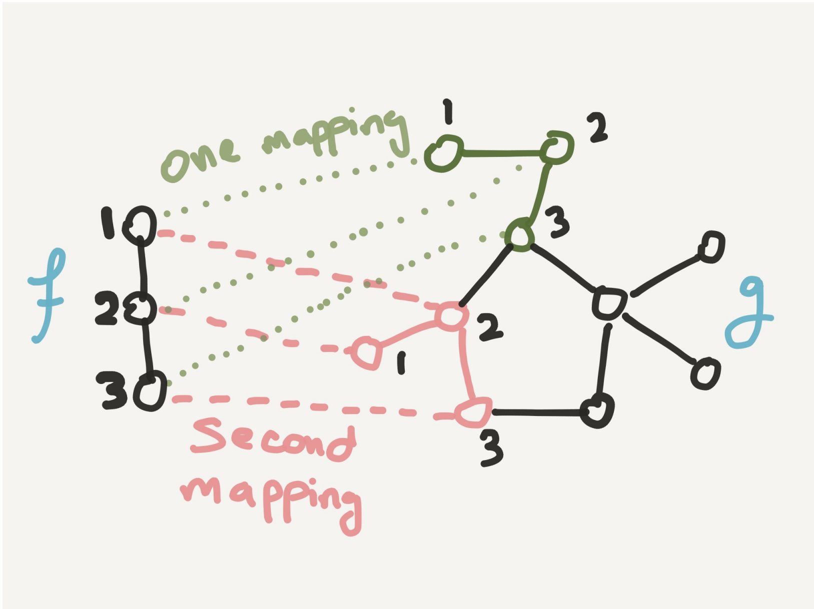

Figure 5: The subgraph induced by the three green nodes in g is isomorphic to f. The same applies to the three pink nodes, and to a several other sets of node triplets in g. The total number of such triplets that are isomophic to f is denoted by $\mathrm{Iso}(f,g)$.

2.1 Some symbols and notation

Denote by $\mathrm{Iso}(f,g)$ the number of mappings from $V(f)$ to $V(g)$ such that the induced subgraph of g is isomorphic to f. Figure 5 demonstrates 2 of the several such mappings from f to g.

Definition 5. *The* motif density *of f in g is*

$$t_{\mathrm{iso}}(f,g) = \frac{\mathrm{Iso}(f,g)}{\#\ \textit{potential mappings from } f \textit{ to } g} = \frac{\mathrm{Iso}(f,g)}{n(n-1)\dots(n-(k-1))} = \frac{\mathrm{Iso}(f,g)}{\binom{n}{k}k!}.$$

This motif density is sometimes referred to as $t_{\mathrm{ind}}(f,g)$.

Definition 6. *A sequence of graphs $(g_1, g_2, \dots, g_m, \dots)$* converges *(in the graph sense) when $t_{\mathrm{iso}}(f, g_m)$ converges as $m \to \infty$ for every fixed motif f.*[2]

One additional notation we will require is $G[k]$ = the subgraph induced by picking k distinct nodes from g at random. In this new notation, we can say

$$t_{\mathrm{iso}}(f,g) = \mathbb{P}(f \simeq G[k]).$$

[2]As a sanity check, if $g_m = g$ for every m, then the sequence still converges.

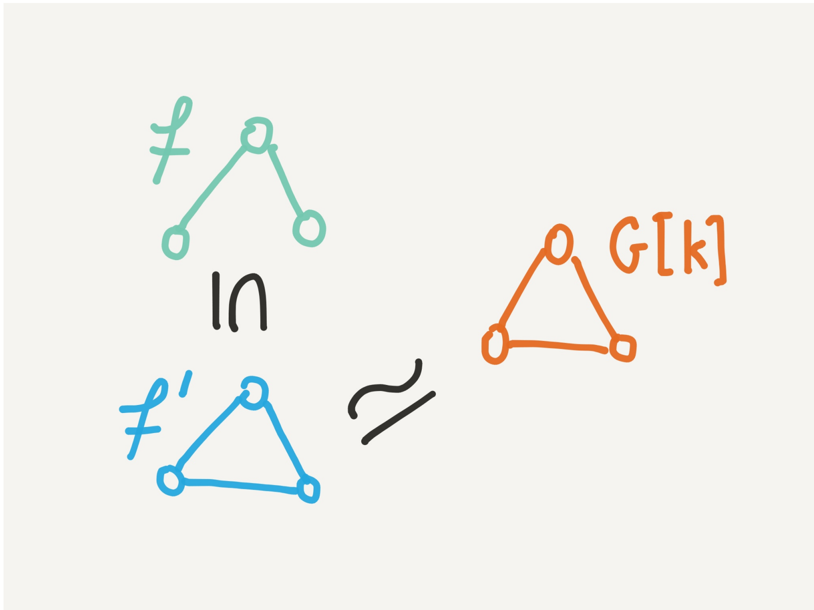

Figure 6: There exists some $f' \supseteq f$ such that $f' \simeq G[k]$.

2.2 Injective homomorphisms

Unfortunately, t_{iso} is very hard to calculate. So, define a weaker notion of matching than t_{iso}:

$$t_{\text{injective}(f,g)} = \mathbb{P}(f \subseteq G[k]),$$

where $f \subseteq G[k]$ means that $E(f) \subseteq E(G[k])$; Figure 6 provides a small example of this concept. Now, if we know all of the isomorphism densities $t_{\text{iso}}(f', g)$, for all f', we can calculate $t_{\text{injective}}(f, g)$. If $f \subseteq G[k]$, then there is some f' with the same edges as f (plus some more edges), with an isomorphism $f' \simeq G[k]$.

So,

$$t_{\text{injective}}(f, g) = \sum_{f' : f \subseteq f'} t_{\text{iso}}(f', g).$$

Since this is a linear system of equations, we can invert to get $t_{\text{iso}}(f, g)$ as a linear combination of $t_{\text{injective}}(f, g)$. This imples that if we have all of the $t_{\text{injective}}(f, g)$, we can always calculate all of the $t_{\text{iso}}(f, g)$, and vice versa. Therefore, the isomorphism densities converge if and only if the injective homomorphism densities converge.

2.3 Homomorphisms

We previously went from a strong notion of graph matching, graph isomorphism, to a weaker notion, injective homomorphism. We showed that convergence of

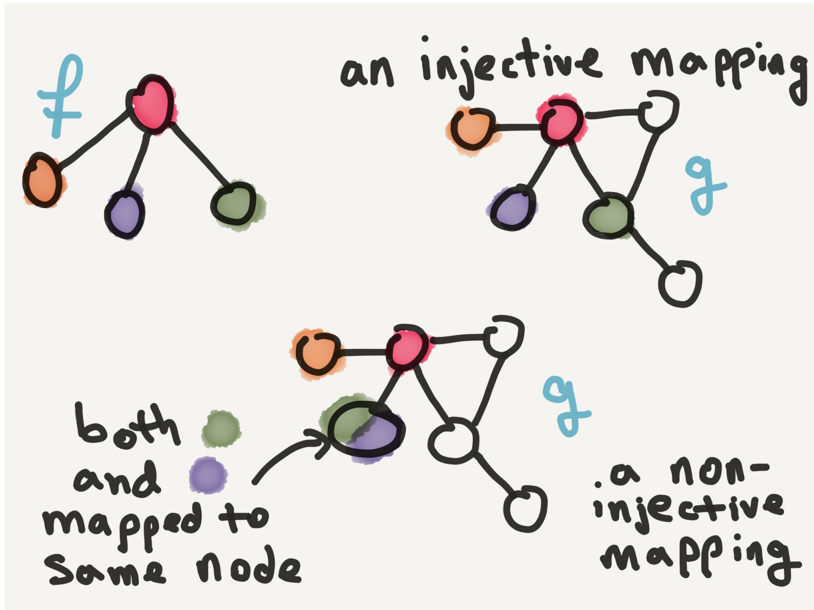

Figure 7: In a non-injective mapping, two different nodes in the domain graph f can be mapped to the same node in the image graph g.

one implies the other, so that we could get away with working with the simpler notion of injective homomorphism.

Naturally, the next step is to define an even weaker notion of graph matching, which will be easier to calculate as well. To that end, we remove the restriction of *injectivity* (in the sense of injective functions mapping to unique elements in the image), to obtain the graph *homomorphism*.

Definition 7. *A* homomorphism *from f to g is a mapping $\phi : V(f) \to V(g)$ such that if $(i, j) \in E(f)$, then $(\phi(i), \phi(j)) \in E(g)$. Figure 7 shows an example where f is injectively mapped to g, and another example where f is non-injectively mapped to g.*

The notion of graph homomorphism, in contrast to the injective homomorphism, is akin to sampling nodes of g *with replacement*, in the sense that we are allowing ϕ to assign the same node in g to more than one node from f.

Just as we defined $G[k]$ as the subgraph induced by picking k nodes *without* replacement, we define $G'[k]$ to be the subgraph induced by picking k (not necessarily distinct) nodes of g, i.e. with replacement. This leads to a new notion of motif density:

$$t_{\text{hom}} = \frac{\text{Hom}(f, g)}{n^k} = \mathbb{P}(f \subseteq G'[k]).$$