A note on the first-projection method for proving central limit theorems

Vasco Portilheiro

June 9, 2020

1 Introduction

The purpose of this note is to serve as an accessible exposition of a method for proving central limit theorems — the so-called "first projection" method. The idea is motivated through the study of subgraph-counts in Erdös-Rényi random graphs, as developed in the classic book of Janson, Łuczak, and Rucinski [1]. The language used therein is that of *graph functionals* — real-valued functions of graphs depending only on their isomorphism class. This is a nice bit of generality, which will allow us to use the idea of projection in the space of functionals.

Here, we will be a bit more general. The projection idea in fact goes through for any random variables, allowing us to prove the following result (a graph-less generalization of Theorem 6.39 from [1]).

Theorem 1. *Consider random variables L_n with a central limit theorem, that is, with*

$$\frac{L_n - \mathbb{E}[L_n]}{\mathrm{Var}(L_n)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1).$$

Any random variables X_n such that

$$\mathrm{Var}(X_n) \sim \frac{\mathrm{Cov}(X_n, L_n)^2}{\mathrm{Var}(L_n)}$$

also obey the central limit theorem,

$$\frac{X_n - \mathbb{E}[X_n]}{\mathrm{Var}(X_n)^{1/2}} \to \mathcal{N}(0,1).$$

We prove this result in Section 2. Intuitively, we may view the result above as a tool for "bootstrapping" central limit theorems; given some X_n of interest, we may prove a central limit theorem by judiciously choosing L_n.

We will try to give a flavor of how this idea applies (quite nicely) to random graphs. In Section 3 we sketch how one can apply Theorem 1 when X_n is the number of triangles in $G_n \sim \mathbb{G}(n,p)$.

2 The first projection method

The main idea behind the proof of Theorem 1 is to approximate the X_n by some Y_n that are easier to study. If we can show that X_n and Y_n are not too far apart, and we pick Y_n to have a central limit theorem, we might then expect to be able to prove one for X_n.

The tool that formalizes this intuition the following result, attributed in [1] to Cramér. (Others may refer to this as a version of Slutsky's theorem.)

Theorem 2. *Consider two sequences of random variables X_n and Y_n. If $|X_n - Y_n| \xrightarrow{p} 0$ and $Y_n \xrightarrow{d} Z$, then $X_n \xrightarrow{d} Z$.*

Proof. We will take as the definition of convergence in distribution $Z_n \xrightarrow{d} Z$ that for any continuous and bounded function f, $\mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]$. Consider any such f, with bounding constant B, and any $\epsilon > 0$. Let $\delta > 0$ be such that if $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$. Now, note that we can bound

$$\begin{aligned}
|\mathbb{E}[f(X_n) - \mathbb{E}[f(Y_n)]| &\leq \mathbb{E}[|f(X_n) - f(Y_n)|] \\
&= \mathbb{E}[|f(X_n) - f(Y_n)|\mathbf{1}_{|X_n - Y_n| < \delta}] + \mathbb{E}[|f(X_n) - f(Y_n)|\mathbf{1}_{|X_n - Y_n| \geq \delta}] \\
&\leq \epsilon \mathbb{P}\left(|X_n - Y_n| < \delta\right) + 2B\mathbb{P}\left(|X_n - Y_n| \geq \delta\right) \\
&\leq \epsilon + 2B\mathbb{P}\left(|X_n - Y_n| \geq \delta\right).
\end{aligned}$$

By the triangle inequality, we then have that

$$|\mathbb{E}[f(X_n)] - \mathbb{E}[f(Z)]| \leq \epsilon + 2B\mathbb{P}\left(|X_n - Y_n| \geq \delta\right) + |\mathbb{E}[f(Y_n)] - \mathbb{E}[f(Z)]|.$$

Sending $n \to \infty$, the last two terms vanish by assumption. We thus have that for any continuous and bounded f and any $\epsilon > 0$,

$$\lim_{n \to \infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(Z)]| < \epsilon,$$

and letting $\epsilon \to 0$, we have $X_n \xrightarrow{d} Z$. □

With the result above in hand, the question becomes: what approximating Y_n to choose? The first projection method is one answer.

Consider the linear space $\mathcal{L}(L) = \{aL + b : a, b \in \mathbb{R}\}$, where L is some fixed random variable. The *first projection* $Y = \text{Proj}_{\mathcal{L}(L)}(X) \in \mathcal{L}(L)$ of X (in $L^2(\mathbb{P})$) minimizes the L^2 distance $\mathbb{E}[(X - Y)^2]$, and is characterized as follows. (Our proof below is only a sketch, as the theory here is not the focus of this note.)

Proposition 3. *Consider random variables L and X, and let $Y = \text{Proj}_{\mathcal{L}(L)}(X)$. Then $Y = aL + b$ with*

$$\begin{aligned}
a &= \frac{\text{Cov}(X, L)}{\text{Var}(L)} \\
b &= \mathbb{E}[X] - a\mathbb{E}[L].
\end{aligned}$$

Furthermore, $\mathbb{E}[Y] = \mathbb{E}[X]$*, and*

$$\mathbb{E}[(X-Y)^2] = \mathrm{Var}(X) - a^2\,\mathrm{Var}(L) = \mathrm{Var}(X) - \frac{\mathrm{Cov}(X,L)^2}{\mathrm{Var}(L)}.$$

Proof. Recall that Y minimizes $\mathbb{E}[(X-Y)^2]$, and is thus in fact the solution to a linear least-squares problem in $L^2(\mathbb{P})$. In particular, we may write

$$Y = \frac{\langle X, L\rangle}{\langle L, L\rangle}(L - \mathbb{E}[L]) + \mathbb{E}[X]$$

where $\langle \cdot, \cdot \rangle = \mathrm{Cov}(\cdot, \cdot)$. The result follows immediately. □

The result above suggests the following proof of Theorem 1.

Proof of Theorem 1. Suppose the random variables L_n obey a central limit theorem, and X_n have

$$\mathrm{Var}(X_n) \sim \frac{\mathrm{Cov}(X_n, L_n)^2}{\mathrm{Var}(L_n)}.$$

Let $Y_n = \mathrm{Proj}_{\mathcal{L}(L)}(X_n)$. By Proposition 3 and our assumption of $\mathrm{Var}(X_n)$, we have that $\mathbb{E}[(X_n - Y_n)^2] = o(\mathrm{Var}(X_n))$, and thus

$$\frac{X_n - Y_n}{\mathrm{Var}(X_n)^{1/2}} \xrightarrow{p} 0.$$

Noting that $\mathbb{E}[Y_n] = \mathbb{E}[X_n]$ and $\mathrm{Var}(Y_n) = \frac{\mathrm{Cov}(X_n, L_n)^2}{\mathrm{Var}(L_n)} \sim \mathrm{Var}(X_n)$,

$$\frac{X_n - \mathbb{E}[X_n]}{\mathrm{Var}(X_n)^{1/2}} \sim \frac{Y_n - \mathbb{E}[Y_n]}{\mathrm{Var}(Y_n)^{1/2}} + \frac{X_n - Y_n}{\mathrm{Var}(X_n)^{1/2}}.$$

Since L_n has a central limit theorem, so does Y_n (being an affine transformation of L_n). Thus by Theorem 2

$$\frac{X_n - \mathbb{E}[X_n]}{\mathrm{Var}(X_n)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1).$$

□

3 A central limit for triangle counts

We now apply Theorem 1 to show that the number of triangles X_n in $G_n \sim \mathbb{G}(n,p)$ obeys a central limit theorem. As mentioned previously we will use as our functional L_n the number of edges in G_n. As $L_n \sim \mathrm{Bin}(\binom{n}{2}, p)$, it is clear that

$$\frac{L_n - \mathbb{E}[L_n]}{\mathrm{Var}(L_n)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1)$$

when $n^2p \to \infty$ and $n^2(1-p) \to \infty$. We thus show that in this regime,

$$\mathrm{Var}(X_n) \sim \frac{\mathrm{Cov}(X_n, L_n)^2}{\mathrm{Var}(L_n)},$$

which by Theorem 1 will imply that

$$\frac{X_n - \mathbb{E}[X_n]}{\mathrm{Var}(X_n)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1)$$

when $n^2p \to \infty$ and $n^2(1-p) \to \infty$.

To see this, we first note that $\mathrm{Var}(L_n) = \binom{n}{2}p(1-p)$. Observe next that fixing any $u, v \in V(G_n)$ and writing $\mathbf{1}_{uv}$ for the indicator of (u,v) being an edge, we can write

$$\mathrm{Cov}(X_n, L_n) = \binom{n}{2}\mathrm{Cov}(X_n, \mathbf{1}_{uv}) = \binom{n}{2}p(\mathbb{E}[X_n|\mathbf{1}_{uv}] - \mathbb{E}[X]).$$

Therefore,

$$\frac{\mathrm{Cov}(X_n, L_n)^2}{\mathrm{Var}(L_n)} = \frac{\binom{n}{2}p(\mathbb{E}[X_n|\mathbf{1}_{uv}] - \mathbb{E}[X])^2}{1-p}.$$

It can be shown (by counting of triangles intersecting at a single edge), that the above is asymptotically equal to the contribution to $\mathrm{Var}(X_n)$ of triangles overlapping in a single edge, $\binom{n}{2}(n-2)(n-3)p^5$. As this is the leading term in $\mathrm{Var}(X_n)$, the result holds.

References

[1] "Asymptotic Distributions". In: *Random Graphs*. John Wiley & Sons, Ltd, 2011. Chap. 6, pp. 139–177. ISBN: 9781118032718. DOI: 10.1002/9781118032718.ch6.