Asymptotic order notation & more

Gustavo Pereira

November 12, 2019

These notes are a slight modification of Recitation 7 from Fall 2018 intro to metrics.

1 Asymptotic order notation

It is intuitive to argue that the sequence $x_n = n$ diverges to infinity faster than the sequence $y_n = \sqrt{n}$. But why is that the case? One way to frame this would be to look at the difference between y_n and x_n. As n grows, the values of $|x_n - y_n|$ become larger and larger.

However, this is not the only way of comparing "orders of magnitude". Take for example the sequences $a_n = n$ and $b_n = 5n$. Would you say that b_n diverges to infinity "an order of magnitude faster" than a_n? That would of course depend on what criterion you are using, but perhaps it would make sense to say that since they're both "diverging linearly" infinity, their asymptotic order is the same.

The standard way of comparing asymptotic orders of x_n and y_n depends on the ratio x_n/y_n instead of the difference. That way, for example, the asymptotic order of x_n is the same as Cx_n for any constant C. Define the ratio $r_n := x_n/y_n$. Modulo some technical qualifications, we say that

1. x_n is dominated by (or: is an order of magnitude below) y_n if $r_n \to 0$;
2. x_n is at most the same order as y_n if r_n is bounded after some index n_0; that is there exists M such that
$$|r_n| \leq M$$
for all $n \geq n_0$

The common way of designating 1 and 2 above is, respectively:

1. $x_n \in o(y_n)$
2. $x_n \in O(y_n)$

and in English would say that (1) x_n is "little oh" of y_n, or (2) x_n is "big oh" of y_n.

The technical qualification omitted above is one that deals with the signs of x_n and y_n. To avoid any technical complications, we typically require y_n to be strictly positive (at least after some index n_1). This is not a big deal because the right hand side of statements such as $x_n \in O(y_n)$ typically involve a class of sequences such as n, $log(n)$, $1/n$, etc, which are always positive anyway.

Also, if you are crazy about limits, one equivalent way of saying that $x_n \in O(y_n)$ is that

$$\limsup \left|\frac{x_n}{y_n}\right| < \infty$$

It will be useful to familiarize yourself with asymptotic order notation. First, convince yourself that $n \in o(n^2)$, $\log(n) \in O(n)$, etc. Then, solve the following exercises (which should follow very easily from the definition):

- $x_n \in O(x_n)$ for every sequence x_n; the same isn't true for $o(x_n)$.
- If $x_n \in An^2 + Bn + C$ with $A > 0$, then $x_n \in O(n^2)$. Also, $x_n \in o(x^3)$.
- The fact above extends to degree k polynomials.
- $x_n \in o(y_n)$ if and only if $x_n/y_n \in o(1)$; same holds for O.
- $x_n \in o(1)$ if and only if $x_n \to 0$
- If $x_n \to x$ then $x_n \in O(1)$

The properties below are stated as a Lemma just because they connect with later sections, but are also not hard to prove.

Lemma 1. *Regarding sequences x_n, y_n, z_n and w_n, where z_n and w_n are positive:*

1. *If $x_n \in o(z_n)$ and $y_n \in o(w_n)$, then $x_n y_n \in o(z_n w_n)$ and $x_n + y_n \in o(z_n + w_n)$; same holds for O*
2. *If $x_n \in o(z_n)$ and $y_n \in O(w_n)$, then $x_n y_n \in o(z_n w_n)$ and $x_n + y_n \in O(z_n + w_n)$*

One hint about proving the above lemma: since w_n and z_n are always positive, the ratio

$$\frac{w_n}{w_n + y_n}$$

is necessarily bounded (in fact, it is in $(0, 1)$). This may be helpful if one writes

$$\frac{x_n + y_n}{z_n + w_n} = \frac{x_n}{z_n}\frac{z_n}{z_n + w_n} + \frac{y_n}{w_n}\frac{w_n}{z_n + w_n}$$

because then the terms multiplying x_n/z_n and y_n/w_n are both bounded.

2 Convergence in Probability

Random variables are more complicated objects than real numbers. This is because random variables are *measurable maps between measurable spaces*; generally speaking, these are infinite dimensional spaces and defining notions such as "distance" and "convergence" there is sometimes tricky.

What we do here is to specialize. We will define what it means for a sequences of random variables X_n to **converge in probability** to a (constant) real number a. Before that, I'll introduce a weaker concept the stochastic analog of the $O(1)$ notation. Remember that a sequence of real numbers $x_n \sim O(1)$ if x_n is bounded after $n \geq n_0$.

Definition 1. The sequence of random variables (X_n) is said to be **stochastically bounded**, denoted in short by $X_n \in O_p(1)$, if for every $\delta > 0$, there exists M_δ and n_0 such that

$$\Pr\{|X_n| \leq M_\delta\} > 1 - \delta$$

holds for every $n \geq n_0$. ◁

Instead of requiring $|X_n| \leq M$ to hold strictly for large n, we require it to hold *with some probability* after for large n. Moreover, perhaps by choosing a large enough bound, this probability can be made arbitrarily close to 1.

The reason why this is analogous to the non-stochastic O notation is that if X_n is a deterministic sequence[1] then $X_n \in O_p(1)$ if, and only if, $X_n = O(1)$. In analogy with the deterministic case, we define X_n to be $O_p(Y_n)$ if $X_n/Y_n \in O_p(1)$.

The proposition below provides one type of sequence that is always stochastically bounded: the ones that converge in distribution.

Proposition 1. *Let (X_n) be a sequence of random variables and Y any random variable. If X_n converges in distribution to Y, then $X_n \in O_p(1)$*

Proof. Let F_Y be the cdf of Y. For fixed $\delta > 0$, take M_δ such that:

1. F_Y is continuous at M_δ and $-M_\delta$.

2. $F_Y(M_\delta) - F_Y(-M_\delta) > 1 - \delta/2$

   The existence of such M_δ comes from the "continuity property" of probability measures: since $(-n, n] \uparrow \mathbf{R}$, $F_Y(n) - F_Y(-n) \uparrow 1$; hence, $F_Y(n_0) - F_Y(-n_0) > 1 - \delta/2$ for some n_0.

[1]That is, for every n there exists some x_n such that the only possible outcome is $X_n = x_n$.

The two requirements above can be met simultaneously because F_Y is monotone, and a monotone function in the real line can have at most a countable number of discontinuities. Given that, and that the fact that there are uncountably many sets $[-x, x]$ where $x > n_0$; one of them has to be such that x and $-x$ are continuity points of F_Y. (Otherwise F_Y would have uncountably many discontinuities.)

The definition of convergence in distribution then implies that $F_{X_n}(M_\delta) \to F_Y(M_\delta)$ and $F_{X_n}(-M_\delta) \to F_Y(-M_\delta)$. Thus we can pick n_0 such that for $n \geq n_0$,

$$F_{X_n}(M_\delta) > F_Y(M_\delta) - \delta/4$$

and

$$F_{X_n}(-M_\delta) < F_Y(-M_\delta) + \delta/4$$

Thus

$$\begin{aligned}
\Pr\{|X_n| \leq M_\delta\} &\geq F_{X_n}(M_\delta) - F_{X_n}(-M_\delta) \\
&> F_Y(M_\delta) - F_Y(-M_\delta) - \delta/2 \\
&> 1 - \delta
\end{aligned}$$

□

We now move to convergence in probability.

Definition 2. Let (X_n) be a sequence of real valued random variables. The sequence **converges in probability** to a if, for every $\epsilon > 0$, and $\delta > 0$, there exists n_0 such that

$$\Pr\{|X_n - a| \leq \epsilon\} \geq 1 - \delta$$

if $n \geq n_0$. ◁

Again, we can't make $|X_n - a|$ arbitrarily small with certainty, but we can make it arbitrarily small with some probability, and we can make this probability close to 1. Convergence in probability is denoted by the $\xrightarrow{p}$ symbol; that is, convergence in probability of X_n in to a is denoted by

$$X_n \xrightarrow{p} a$$

Whenever X_n is a deterministic sequence, convergence in probability is equivalent to "regular" convergence. That is because the definition of convergence of real numbers states that for any $\epsilon > 0$, there exists n_0 such that

$$|X_n - a| \leq \epsilon$$

happens for all $n \geq n_0$. In particular, starting from n_0,

$$\Pr\{|X_n - a| \leq \epsilon\} = 1$$

which implies convergence in probability.

Convergence in probability is all we need to define the 'stochastic little o' notation. So without further ado:

Definition 3. Let (X_n) and (Y_n) be sequences of real valued random variables. We say that:

1. $X_n \in o_p(1)$ if $X_n \xrightarrow{p} 0$.
2. $X_n \in o_p(Y_n)$ if $X_n/Y_n \in o_p(1)$

◁

The facts of Lemma 1 go through substituting O with O_p and o with o_p, but proving them is slightly more difficult here. Because it may be instructive, I'll restate it in the stochastic form, and provide a proof.

Proposition 2. *Let (X_n), (Y_n), (Z_n) and (W_n) be sequences of real valued random variables where Z_n, W_n are positive. Then:*

1. *If $X_n \in o_p(Z_n)$ and $Y_n \in o_p(W_n)$, then $X_nY_n \in o_p(Z_nW_n)$ and $X_n + Y_n \in o_p(Z_n + W_n)$. Same holds for O_p.*
2. *If $X_n \in o_p(Z_n)$ and $Y_n \in O_p(W_n)$, then $X_nY_n \in o_p(Z_nW_n)$ and $X_n + Y_n \in O_p(Z_n + W_n)$.*

Proof. For part (1), I'll do the proof for o_p only; the O_p analog is very similar.

Start with $X_n + Y_n$. For any $n \in \mathbf{N}$, we can write

$$\begin{aligned}\frac{|X_n + Y_n|}{Z_n + W_n} &\leq \frac{|X_n|}{Z_n + W_n} + \frac{|Y_n|}{Z_n + W_n} \\ &= \frac{|X_n|}{Z_n}\frac{Z_n}{Z_n + W_n} + \frac{|Y_n|}{W_n}\frac{W_n}{Z_n + W_n}\end{aligned}$$

If $|X_n| \leq \epsilon Z_n$ and $|Y_n| \leq \epsilon W_n$, then the above inequality implies $|X_n + Y_n| \leq \epsilon(Z_n + W_n)$. In other words, the following relationship holds between events:

$$\{|X_n| \leq \epsilon Z_n\} \cap \{|Y_n| \leq \epsilon W_n\} \subseteq \{|X_n + Y_n| \leq \epsilon(Z_n + W_n)\} \tag{1}$$

Let $\delta > 0$ and $\epsilon > 0$. Fix n_0 such that, for $n \geq n_0$,

$$\Pr\{|X_n| \leq \epsilon Z_n\} > 1 - \frac{\delta}{2}$$

and similarly

$$\Pr\{|Y_n| \leq \epsilon W_n\} > 1 - \frac{\delta}{2}$$

Using the relationship $P(A_n \cap B_n) = 1 - P(A_n^c) - P(B_n^c)$, we get

$$\Pr\left(\{|X_n| \leq \epsilon Z_n\} \cap \{|Y_n| \leq \epsilon W_n\}\right) > 1 - \delta$$

If $n \geq n_0$, from 2 and the above inequality we conclude that

$$\Pr\{|X_n + Y_n| \leq \epsilon(Z_n + W_n)\} \geq 1 - \delta$$

which establishes that $X_n + Y_n \in o_p(Z_n + W_n)$.

Proving that $X_n Y_n \in o_p(Z_n W_n)$ is a quite similar exercise. Just note that

$$\left\{|X_n| \leq \sqrt{\epsilon} Z_n\right\} \cap \left\{|Y_n| \leq \sqrt{\epsilon} W_n\right\} \subseteq \{|X_n Y_n| \leq \epsilon Z_n W_n\}$$

so for $\delta, \epsilon > 0$ one can make $\Pr\{|X_n Y_n| \leq \epsilon Z_n W_n\} > 1 - \delta$ by taking $n \geq n_0$ where n_0 satisfies

$$\Pr\left\{|X_n| \leq \sqrt{\epsilon} Z_n\right\} > 1 - \frac{\delta}{2}$$

and

$$\Pr\left\{|X_n| \leq \sqrt{\epsilon} Z_n\right\} > 1 - \frac{\delta}{2}$$

for $n \geq n_0$.

I proceed to part (2). Let $X_n \in o_p(Z_n)$ and $Y_n \in O_p(W_n)$. For positive ϵ, δ, we can find:

1. $M_\delta > 0$ and n_0 such that $\Pr\{|Y_n| \leq M_\delta W_n\} > 1 - \frac{\delta}{2}$
2. n_1 such that $\Pr\{|X_n| \leq \epsilon Z_n / M_\eta\} > 1 - \frac{\delta}{2}$

Taking $n_0 = \max\{n_1, n_2\}$ we get, for $n \geq n_0$,

$$\begin{aligned}\Pr\{|X_n Y_n| \leq \epsilon Z_n W_n\} &\geq \Pr\left(\left\{|X_n| \leq \frac{\epsilon}{M_\eta} Z_n\right\} \cap \{|Y_n| \leq M_\eta W_n\}\right) \\ &> 1 - \delta\end{aligned}$$

Hence $X_n Y_n \in o_p(1)$. The fact that $X_n + Y_n \in O_p(Z_n + W_n)$ is left as an exercise. $\square$

2.1 Applications

The LLN states that if (X_i) is an iid sequence of variables with finite first moment, then

$$\bar{X}_n \xrightarrow{p} \mu$$

In terms of our stochastic order notation, this implies

$$\bar{X}_n - \mu = o_p(1)$$

it is common to rewrite this as

$$\bar{X}_n = \mu + o_p(1)$$

If in addition to the first moment, the second moments are also finite, we have

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

By Proposition 1, this implies $\sqrt{n}(\bar{X}_n - \mu) = O_p(1)$ hence using O_p properties from Proposition 2, we conclude that

$$\bar{X}_n - \mu = O_p\left(\frac{1}{\sqrt{n}}\right)$$

so the rate of convergence of $\bar{X}_n$ to μ is "no slower than" $n^{-1/2}$.

It has been proved in class that if $X_n \xrightarrow{d} X$ and $a_n \xrightarrow{p} a$ where a is constant,

$$X_n + a_n \xrightarrow{d} X + a$$

The fact above implies that if $X_n \xrightarrow{d} X$, then $X_n + o_p(1) \xrightarrow{d} X$.

One important use of the O_p machinery is when dealing with Taylor expansions. This is exemplified by the Lemma below:

Lemma 2. *Let $X_n = \mu + o_p(1)$, and g be a function that is differentiable at μ. Then*

$$g(X_n) - g(\mu) = (X_n - \mu)[g'(\mu) + o_p(1)]$$

Proof. Let

$$r(x) = \begin{cases} 0 & \text{if } x = \mu \\ \frac{g(x) - g(\mu) - g'(\mu)(x-\mu)}{x-\mu} & \text{if } |x - \mu| > 0 \end{cases}$$

Then $g(X_n) - g(\mu) = (X_n - \mu)[g'(\mu) + r(X_n)]$; we will show that $r(X_n) \in o_p(1)$.

Let $\epsilon > 0$ and $\delta > 0$ be arbitrary. By differentiability, there exists $\eta > 0$ such that $|X_n - \mu| < \eta$ implies $|r(X_n)| \leq \epsilon$. Take n_1 such that $n \geq n_1$ implies

$$\Pr\{|X_n - \mu| < \eta\} > 1 - \delta$$

It follows from the inclusion $\{|X_n - \mu| < \eta\} \subseteq \{|r(X_n)| \leq \epsilon\}$ that if $n \geq n_1$,

$$\Pr\{|r(X_n)| \leq \epsilon\} \geq \Pr\{|X_n - \mu| < \eta\} > 1 - \delta$$

□

We are now equipped to prove what is generally called the Delta method; this involves computing limiting distributions of sequences such as

$$Z_n = \sqrt{n}\left(\bar{x}_n^2 - \mu^2\right)$$

when we know that the CLT applies to $\bar{x}_n$. For $\mu \neq 0$, we'll show that

$$Z_n \xrightarrow{d} N(0, 4\mu^2\sigma^2)$$

That will be a consequence of the following theorem.

Theorem 1. *Suppose (X_n) is a sequence of random variables satisfying*

$$\sqrt{n}[X_n - \mu] \xrightarrow{d} N(0, \sigma^2)$$

for some μ, σ^2. Let $g : \mathbf{R} \to \mathbf{R}$ be differentiable at μ. If $\mu \neq 0$,

$$\sqrt{n}[g(X_n) - g(\mu)] \xrightarrow{d} N(0, \sigma^2 g'(\mu)^2)$$

Proof. From Lemma 2 we can write $g(X_n) - g(\mu) = (X_n - \mu)g'(\mu) + (X_n - \mu)o_p(1)$. Multiplying both sides by $\sqrt{n}$ yields

$$\sqrt{n}[g(X_n) - g(\mu)] = \sqrt{n}(X_n - \mu)g'(\mu) + \sqrt{n}(X_n - \mu)o_p(1)$$

We know that $\sqrt{n}(X_n - \mu) \in O_p(1)$ by 1, so that

$$\sqrt{n}[g(X_n) - g(\mu)] = \sqrt{n}(X_n - \mu)g'(\mu) + o_p(1) \xrightarrow{d} N(0, g'(\mu)^2\sigma^2)$$

□

3 The Law of Large Numbers

The statement of the weak law of large numbers is copied below.

Theorem 2. *Let $(X_n)_{n\in\mathbf{N}}$ be a sequence of iid random variables for which $E|X_n| < \infty$. Define $\mu := \mathbf{E}X_n$ and $\bar{X}_N = \sum_{n=1}^{N} X_n/N$.*

Then $\bar{X}_N \xrightarrow{p} \mu$.

It basically says that if one draws a random sample from an iid distribution, the probability that the sample mean deviates from the true mean (by any fixed threshold $\epsilon > 0$) can be made arbitrarily small.

The proof of this theorem is easier if one assumes bounded second moments. The reason is that imposing bounded second moments restricts how fat the tail of the distribution of X can plausibly be.

One inequality that generally relates the tail of random variables to their moments is the **Chebyshev** inequality. It provides the following tail bound, for every random variable Y with bounded first moment: for any $\lambda > 0$,

$$\Pr\{|Y| > \lambda\} \leq \frac{\mathbf{E}|Y|}{\lambda}$$

(of course the inequality still holds if Y has infinite first moment, but it does not provide a useful bound.)

If $|Y|$ has a bounded second moment, one can apply the bound to $\{|Y|^2 > \lambda^2\}$ and then use the fact that $\{|Y|^2 > \lambda^2\} = \{|Y| > \lambda\}$ to conclude that

$$\Pr\{|Y| > \lambda\} \leq \frac{\mathbf{E}(Y^2)}{\lambda^2}$$

The above bounds can be applied to the sample mean $\bar{X}_N$. It follows from the linearity of expectations that $\mathbf{E}(\bar{X}_N) = \mu$, so that

$$\Pr\left(\bar{X}_N \geq \lambda\right) = \frac{\mu}{\lambda}$$

The second moment version of the tail bound suffices to prove convergence of $\bar{X}_N$ to μ. We apply the bound to $Y_N = \bar{X}_N - \mu$; by the iid assumption and some algebra, it follows that

$$\mathbf{E}(Y_N^2) = \mathbf{V}(\bar{X}_N) = \frac{\mathbf{V}(X_1)}{N}$$

hence for each $\lambda > 0$,

$$\Pr\{|\bar{X}_N - \mu| > \lambda\} \leq \frac{V(X_1)}{N\lambda^2} \to 0$$

which establishes the WLLN when variance is assumed to be finite.

What if $\mathbf{E}X_i^2 = \infty$? While we can't exactly proceed as before, we can do a little trick. The idea is to *truncate* the variables X_i at K, to $\xi_i^{\leq}(K)$, where

$$\xi_i^{\leq}(K) = \begin{cases} X_i & \text{if } |X_i| \leq K \\ K & \text{otherwise} \end{cases}$$

We can succintly express $\xi_i^{\leq}(K)$ as $\xi_i^{\leq}(K) = X_i \mathbf{1}\{|X_i| \leq K\}$.

The LLN follows from the facts below, whose proofs will be provided later on.

1. For fixed K, $\xi_i^{\leq}(K)$ are iid and have finite second moments. By the LLN for variables with bounded variance, we have

   $$\Pr\{|\overline{\xi^{\leq}(K)}_N - \mu_K| \geq \lambda\} \to 0$$

   where of course $\overline{\xi^{\leq}(K)}_N = \frac{1}{N}\sum_{n=1}^{N} \xi_i^{\leq}(K)$ and $\mu_K = \mathbf{E}(\xi_i^{\leq}(K))$.

2. Define $\xi_i^{>} := X_i \mathbf{1}\{|X_i| > K\}$, so that $X_i = \xi_i^{\leq}(K) + \xi_i^{>}(K)$ for any threshold K. The following fact holds for any:

   $$\mathbf{E}|\xi_i^{>}(K)| \to 0$$

   as $K \to \infty$ (take K to be positive integers).

3. Taking $K \to \infty$,

   $$\mu_K = \mathbf{E}(\xi_i^{\leq}(K)) \to \mu$$

4. We can write

   $$\Pr\left\{|\bar{X}_N - \mu| > \lambda\right\} \leq \Pr\left\{|\overline{\xi^{\leq}(K)}_N - \mu_K| > \frac{\lambda}{3}\right\} + \Pr\left\{|\mu_K - \mu| > \frac{\lambda}{3}\right\} + \Pr\left\{|\overline{\xi^{>}(K)}_N| > \frac{\lambda}{3}\right\} \tag{2}$$

We can bound each of the terms in 4 conveniently. The third one involves the first moment tail bound:

$$\Pr\left\{|\overline{\xi^{>}(K)}_N| > \frac{\lambda}{3}\right\} \leq 3\frac{\mathbf{E}|\xi_i^{>}(K)|}{\lambda}$$

By Fact 2 above, this can be made arbitrarily small if K is chosen to be large enough.

The second term in 4 can be actually set to 0 by choosing K large enough, since $\mu_K \to \mu$ as $K \to \infty$ (Fact 3). The bound on the first term comes from Fact 1.