REU in Random Walk

Xiaoyu Liu

November 7, 2022

Contents

Note 4: Concepts of Continuity

Wed 12 Oct 2022 18:59

1 Continuity and Hölder continuity test

1.1 Definitions and Examples

The main reference in this note is Durrett, 2010.

Definition 1 (Lipschitz continuous). f is said to be *Lipschitz continuous* if there is a constant C so that $|f(x) - f(y)| \leq C\rho(x, y)$.

Lipschitz continuity is a stronger notion of continuity than classical continuity. Lipschitz continuity implies continuity.

Geometrically, Lipchitz condition puts a finite bound on the slope of any secant line one can get from the graph of the function.

Definition 2. A real or complex valued function f on d-dimensional Euclidean space satisfies a Hölder condition with exponent α, or is α-Hölder continuous, when there are nonnegative real constants $C, a > 0$, such that

$$|f(x) - f(y)| \leq C\|x - y\|^{\alpha}$$

for all x and y in the domain of f.

- When $\alpha > 1$, an α-Hölder continuous function is *constant*.
- When $\alpha = 1$, an α-Hölder continuous function is *Lipchitz continuous*.
- When $\alpha > 0$, an α-Hölder continuous function is *uniformly continuous*.
- Whenver $0 < \alpha \le \alpha'$, α'-Hölder continuity implies α-Hölder continuity.

Remark. (from wikipedia) We have the following chain of strict inclusions for functions over a closed and bounded non-trivial interval of the real line:

Continuously differentiable $\subset$ Lipschitz continuous $\subset$ α-Hölder continuous $\subset$ uniformly continuous $\subset$ continuous

where $\alpha \in (0, 1]$.

Theorem 1. 8.1.5. Brownian paths are Hölder continuous for any exponent $\gamma < 1/2$.

Theorem 2. 8.1.6. With probability one, Brownian paths are not Lipschitz continuous (and hence not differentiable) at any point.

1.2 Test for hölder continuity exponent

From the definition of hölder continuity:

$$|f(x) - f(y)| \le C\|x - y\|^{\alpha}$$

Our goal is to find smallest possible α such that the equation above holds true empirically.

We take logarithm on both sides (using the fact that log is monotone increasing):

$$\log|f(x) - f(y)| \le \log C + \alpha \log\|x - y\|$$

rearrange and assuming $\log\|x - y\| < 0$ for small $\|x - y\|$:

$$\alpha \le \frac{\log|f(x) - f(y)| - \log C}{\log\|x - y\|} \sim \frac{\log|f(x) - f(y)|}{\log\|x - y\|}. \tag{1}$$

The smallest such α is then

$$\alpha = \inf_{x,y} \frac{\log|f(x) - f(y)| - \log C}{\log\|x - y\|}. \tag{2}$$

The term involving $\log C$ should fade away when $\|x - y\|$ is sufficiently small.

1.2.1 Compute α by finding a limit

Let $x_1, \ldots, x_i, \ldots, x_n$ be random points in the domain of f. Let $\varepsilon_1, \ldots, \varepsilon_j, \ldots, \varepsilon_m$ be small changes.

Denote

$$\hat{a}_i^{(j)} := \frac{\log|f(x_i) - f(x_i + \varepsilon_j)| - \log C}{\log \varepsilon_j}.$$

We know $\hat{a}_i^{(j)}$ is always an **upper estimate** of α. Thus a reasonable guess of α would be:

$$\hat{\alpha} = \inf_{i,j} \hat{a}_i^{(j)}.$$

Since $\hat{a}_i^{(j)} \sim a_i^{(j)}$ as $j \to \infty$, we can replace our estimate of α by

$$\hat{\alpha} = \lim_{j\to\infty} \inf_i \hat{a}_i^{(j)}.$$

The limit, however, converges very slowly. As ε_j decreases exponentially, $|\log \varepsilon_j|$ only increases linearly, so the error terms decreases very slowly.

1.2.2 Find α and C by Curve Fitting

Rewrite equation (2) as

$$\alpha = \inf_{x,\varepsilon} \frac{\log|f(x) - f(x+\varepsilon)| - \log C}{\log \varepsilon} \quad \text{where } \varepsilon > 0.$$

Hence

$$\alpha = \inf_{\varepsilon} \left(\inf_x \frac{\log|f(x) - f(x+\varepsilon)|}{\log \varepsilon} - \frac{\log C}{\log \varepsilon} \right).$$

The next claim justifies us to take smaller and smaller ε, and implies that α can be bounded from below arbitrarily closely in finitely many steps.

Proposition 1.

$$\alpha = \lim_{\varepsilon'\to 0} \inf_{\varepsilon<\varepsilon'} \left(\inf_x \frac{\log|f(x) - f(x+\varepsilon)|}{\log \varepsilon} - \frac{\log C}{\log \varepsilon} \right).$$

Proof. Fix some ε' such that for some $\delta > 0$ and all $\varepsilon < \varepsilon'$,

$$\inf_x \frac{\log|f(x) - f(x+\varepsilon)|}{\log \varepsilon} - \frac{\log C}{\log \varepsilon} > \alpha + \delta.$$

We claim that f is actually $(\alpha+\delta)$-Hölder continuous, which results in a contradiction. It suffices to show that there exists some constant C such that for all ε where $\varepsilon \geq \varepsilon'$,

$$|f(x) - f(x+\varepsilon)| \leq C\varepsilon^{\alpha+\delta}.$$

But since f is α-Hölder continuous, there exists some $C' > 0$ such that

$$|f(x) - f(x+\varepsilon)| \leq C'\varepsilon^{\alpha}.$$

Take $C = \frac{C'}{(\varepsilon')^{-\delta}}$, and we have

$$|f(x) - f(x+\varepsilon)| \leq C(\varepsilon')^{\delta}\varepsilon^{\alpha} \leq C\varepsilon^{\delta}\varepsilon^{\alpha} = C\varepsilon^{\alpha+\delta}.$$

□

To get a even faster approximation, we want to fix some particular ε and take infimum only over x. We want to be assured that dropping the infimum over ε when calculating α gives a good enough approximation. This motivates the next proposition:

Proposition 2. If f is a random process with independent increment, and is scale invariant in the sense that for any increments ε and ε', the scaled increments are equal in distribution, i.e. $\frac{f(x+\varepsilon)-f(x)}{\varepsilon} \overset{d}{=} \frac{f(x'+\varepsilon')-f(x')}{\varepsilon'}$ (need to be fixed), then for all $\delta > 0$, we have the bound

$$\Pr\left[\left|\inf_x \frac{\log|f(x)-f(x+\varepsilon)|}{\log\varepsilon} - \frac{\log C}{\log\varepsilon} - \alpha\right| > \delta\right] < O(\varepsilon)$$

Proof. We assume $\alpha + \delta < 1$, and $\varepsilon < 1$.

$$\begin{aligned}
&\Pr\left[\inf_x \frac{\log|f(x)-f(x+\varepsilon)|}{\log\varepsilon} - \frac{\log C}{\log\varepsilon} - \alpha < \delta\right] \\
&\geq \Pr\left[\frac{\log|f(x)-f(x+\varepsilon)|}{\log\varepsilon} - \frac{\log C}{\log\varepsilon} - \alpha < \delta\right] \\
&= \Pr\left[|f(x)-f(x+\varepsilon)| > C\varepsilon^{\alpha+\delta}\right]
\end{aligned}$$

Now we choose some $\varepsilon' < \varepsilon$, and note that this implies $C\varepsilon'^{\alpha+\delta}\frac{\varepsilon}{\varepsilon'} > C\varepsilon^{\alpha+\delta}$, so

$$\begin{aligned}
&\geq \Pr\left[|f(x)-f(x+\varepsilon)| > C\varepsilon'^{\alpha+\delta}\frac{\varepsilon}{\varepsilon'}\right] \\
&= \Pr\left[|f(x')-f(x'+\varepsilon')| > C\varepsilon'^{\alpha+\delta}\right] && \text{using scale invariance} \\
&\geq \Pr\left[\frac{\log|f(x')-f(x'+\varepsilon')|}{\log\varepsilon'} - \frac{\log C}{\log\varepsilon'} - \alpha < \delta\right].
\end{aligned}$$

Using the previous proposition, we can choose small enough ε' to make this probability arbitrarily close to 1. Thus for all $\delta > 0$,

$$\Pr\left[\left|\inf_x \frac{\log|f(x)-f(x+\varepsilon)|}{\log\varepsilon} - \frac{\log C}{\log\varepsilon} - \alpha\right| > \delta\right] = 0$$

□

Using the approximation above, for some specific ε, we can approximate α by

$$\alpha = \inf_x \frac{\log|f(x)-f(x+\varepsilon)|}{\log\varepsilon} - \frac{\log C}{\log\varepsilon} + O(\varepsilon).$$

We then use $\inf_i a_i^{(j)}$ as an upper estimation:

Theorem 3. Assume (some condition). Then for any specific ε_j,

$$\inf_{i\in[n]} a_i^{(j)} < \inf_x \frac{\log|f(x)-f(x+\varepsilon_j)|}{\log\varepsilon_j} + O\left(\frac{1}{\log n}\right).$$

Proof. To be done. Need to work out what assumptions I need. □

Put all these together, we have a curve to fit, with error terms bounded and going to zero:

$$\alpha + \frac{\log C}{\log \varepsilon_j} = \inf_i a_i^{(j)} + O(\varepsilon) + O\left(\frac{1}{\log n}\right). \tag{3}$$

More work needed to turn this into an error bound for α, by considering the curve regression error.

Thus our testing strategy is:

1. Pick some small ε_1 and randomly pick $x_1^{(1)}, \ldots, x_n^{(1)}$ in the domain of interested function.
2. Calculate
$$a_i^{(j)} := \frac{\log \left| f(x_i^{(j)}) - f(x_i^{(j)} + \varepsilon_j) \right|}{\log \varepsilon_j}.$$
3. Calculate $\inf_i a_i^{(j)}$
4. Do the same for successively smaller $\varepsilon_2, \ldots, \varepsilon_m, \ldots$ that approach 0.
5. Fit the curve (3). Then read the α and C.

Note 5: Statistical Tests and their Development

Mon 07 Nov 2022 20:50

2 Statistical Tests

2.1 Fundamental Results

Theorem 4 (Glivenko–Cantelli Theorem)**.** Assume that $X_1, X_2, \ldots$ are independent and identically-distributed random variables in $\mathbb{R}$ with common cumulative distribution function $F(x)$. The empirical distribution function for $X_1, \ldots, X_n$ is defined by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i, \infty)}(x) = \frac{1}{n} \left| \{1 \le i \le n \mid X_i \le x\} \right|$$

where I_C is the indicator function of the set C. Then the empirical cdf uniformly converges to the common cdf almost surely:

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0 \text{ almost surely.}$$

Sometimes called [[The Fundamental Theorem of Statistics]] . Source: wikipedia.

Remark. For every (fixed) $x, F_n(x)$ is a sequence of random variables which converge to $F(x)$ almost surely by the strong law of large numbers. Glivenko and Cantelli strengthened this result by proving uniform convergence of F_n to F.

2.2 Kolomogorov-Smirnov (K-S) One-Sample Statistic

Source: Gibbons and Chakraborti, 2014

A single random sample of size n is drawn from a population with unknown cdf F_X. We wish to test the null hypothesis

$$H_0 : F_X(x) = F_0(x) \text{ for all } x$$

where $F_0(x)$ is completely specified, against the general alternative

$$H_1 : F_X(x) \neq F_0(x) \quad \text{for some } x$$

Let S_n be the empirical distribution function with sample size n. Let F_0 be the expected cdf. The test statistic is

$$D_n = \sup_x |S_n(x) - F_0(x)| .$$

By Glivenko–Cantelli Theorem, D_n should be a reasonable measure of the accuracy of our estimate.

Theorem 5. (K-S statistic is distribution free) The statistics D_n, D_n^+, and D_n^- are completely distribution-free for any specified continuous cdf F_0.

The following theorem gives a good approximation (practically $n > 35$) to sampling distribution of D_n.

Theorem 6 (Kolmogorov Theorem)**.** If F_X is any continuous distribution function, then for every $d > 0$,

$$\lim_{n\to\infty} P\left(D_n \leq d/\sqrt{n}\right) = L(d)$$

where

$$L(d) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\mathrm{e}^{-2i^2d^2}$$

2.3 K-S Two-Sample Statistic

The order statistics corresponding to two random samples of size m and n from continuous populations F_X and F_Y are

$$X_{(1)}, X_{(2)}, \ldots, X_{(m)} \quad \text{and} \quad Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$$

Their respective empirical (sample) distribution functions, denoted by $S_m(x)$ and $S_n(x)$, are defined as before:

$$S_m(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ k/m & \text{if } X_{(k)} \leq x < X_{(k+1)} \quad \text{for } k = 1, 2, \ldots, m-1 \\ 1 & \text{if } x \geq X_{(m)} \end{cases}$$

and

$$S_n(x) = \begin{cases} 0 & \text{if } x < Y_{(1)} \\ k/n & \text{if } Y_{(k)} \leq x < Y_{(k+1)} \\ 1 & \text{if } x \geq Y_{(n)} \end{cases} \qquad \text{for } k = 1, 2, \ldots, n-1$$

The null and alternative hypothesis are

$$H_0 : F_Y(x) = F_X(x) \text{ for all } x$$

$$H_A : F_Y(x) \neq F_X(x) \text{ for some } x$$

The test statistic is based on the maximum absolute difference between the two empirical distributions

$$D_{m,n} = \max_x |S_m(x) - S_n(x)| .$$

The rejection region is defined by

$$D_{m,n} \geq c_\alpha$$

where α is the significance level, and

$$P\left(D_{m,n} \geq c_\alpha \mid H_0\right) \leq \alpha$$

Because of the Glivenko-Cantelli theorem (Theorem 2.3.2), the test is consistent for this alternative. The P value is

$$p = P\left(D_{m,n} \geq D_0 \mid H_0\right)$$

where D_0 is the observed value of the two-sample $K-S$ test statistic.

For the asymptotic null distribution, that is, $m, n \to \infty$ in such a way that m/n remains constant, Smirnov (1939) proved the result

$$\lim_{m,n \to \infty} P\left(\sqrt{\frac{mn}{m+n}} D_{m,n} \leq d\right) = L(d)$$

where

$$L(d) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}$$

Note that the asymptotic distribution of $\sqrt{mn/(m+n)}D_{m,n}$ is exactly the same as the asymptotic distribution of $\sqrt{N}D_N$ in the Kolmogorov Theorem. The only difference is in the normalizing factor.

References

Durrett, Richard (2010). *Probability: Theory and Examples*. 4th ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge ; New York: Cambridge University Press. 428 pp. ISBN: 978-0-521-76539-8.

Gibbons, Jean Dickinson and Subhabrata Chakraborti (2014). *Nonparametric Statistical Inference*. CRC press.