Rates of Convergence by Moduli of Continuity

John Duchi: Notes for Statistics 300b

March 2, 2017

1 Introduction

In this note, we give a presentation showing the importance, and relationship between, the modulis of continuity of a stochastic process and certain growth-like properties of the (population) quantity being modeled or optimized. Our treatment roughly follows van der Vaart and Wellner [1, Chapter 3.2], though we make a few simplifications in attempt to make the approach somewhat cleaner.

To set notation, let Θ be some parameter space with distance d, and let $R_n : \Theta \to \mathbb{R}$ be a sequence of (random) risk functionals with expectation $R(\theta) := \mathbb{E}[R_n(\theta)]$. A typical example of such a process is when we have data $X_i \in \mathcal{X}$ and a loss function $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$, for example, the loss may be the negative log likelihood $-\log p_\theta(x)$ for some model p_θ. We then draw $X_i \stackrel{\text{iid}}{\sim} P$, and we define

$$R_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta, X_i) \quad \text{and} \quad R(\theta) := \mathbb{E}[\ell(\theta, X)].$$

We would like to understand the convergence *rate* properties of $\widehat{\theta}_n = \mathrm{argmin}_{\theta \in \Theta} R_n(\theta)$ to $\theta_0 := \mathrm{argmin}_{\theta \in \Theta} R(\theta)$, the population minimizer.

It is natural, based on a Taylor expansion, to assume that in a neighborhood of θ_0, the population risk grows at least quadratically (because $\nabla R(\theta_0) = 0$). Thus, throughout this note, we assume that there is a constant $\eta > 0$ and a growth constant $\nu > 0$ such that

$$R(\theta) \ge R(\theta_0) + \nu d(\theta, \theta_0)^2 \quad \text{for } \theta \in \Theta \text{ s.t. } d(\theta, \theta_0) \le \eta. \tag{1}$$

With such a condition, it is possible to give rates of convergence of $\widehat{\theta}_n$ to θ_0, at least so long as the random functions R_n do not have so much variability in a neighborhood of θ_0 that they swamp the quadratic growth away from θ_0.

2 Rates of convergence and comparison of functions

Because we would like to understand minimizing the population risk R and finding θ_0, we do not particularly care if $R(\theta)$ and $R_n(\theta)$ are close. While having $R_n(\theta) \approx R(\theta)$ uniformly is sufficient to guarantee that $\widehat{\theta}_n \to \theta_0$, it is not necessary. Indeed, all we really care about is that $R_n(\theta) > R_n(\theta_0)$ for θ sufficiently far from θ_0. That is, as we expect to have roughly $R_n(\theta) - R_n(\theta_0) \approx R(\theta) - R(\theta_0)$, where $R(\theta) \ge R(\theta_0) + \nu d(\theta, \theta_0)^2$, so that we hope that $R_n(\theta) > R_n(\theta_0)$ whenever $d(\theta, \theta_0)^2$ is large enough that it swamps the stochastic error in $R_n(\theta) - R_n(\theta_0)$. Moreover, as long as $\widehat{\theta}_n$ is close enough to θ_0, we can give stronger convergence guarantees, because we expect $\mathrm{Var}(R_n(\theta) - R_n(\theta_0))$ to be smaller than $\mathrm{Var}(R_n(\theta))$ by itself. A bit more precisely, we must have deviations roughly

$\frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}(\ell(\theta, X))}$ in any uniform estimate of $R(\theta)$, by the central limit theorem. However, if θ is near θ_0, then

$$R_n(\theta) - R_n(\theta_0) = R(\theta) - R(\theta_0) + O_P\left(n^{-\frac{1}{2}}\sqrt{\mathrm{Var}(\ell(\theta; X) - \ell(\theta_0; X))}\right),$$

and the latter variance may be substantially smaller than $\mathrm{Var}(\ell(\theta, X))$ when $d(\theta, \theta_0)$ is small.

With the above motivation in mind, as we wish to compare $R_n(\theta) - R_n(\theta_0)$ to $R(\theta) - R(\theta_0)$, our first step in providing rates of convergence is to understand the *modulus of continuity* of the process $\theta \mapsto R_n(\theta)$ in a neighborhood of θ_0. We make the following definition.

Definition 2.1. *Let* $\Theta_\delta := \{\theta \in \Theta : d(\theta, \theta_0) \le \delta\}$. *The* expected modulus of continuity *of the process* R_n *in a radius* δ *around* θ_0 *is*

$$\mathbb{E}\left[\sup_{\theta \in \Theta_\delta} |(R_n(\theta) - R(\theta)) - (R_n(\theta_0) - R(\theta_0))|\right].$$

For notational convenience, we also define the error processes

$$\begin{aligned} \Delta(\theta, x) &:= (\ell(\theta, x) - R(\theta)) - (\ell(\theta_0, x) - R(\theta_0)) \ \text{ and} \\ \Delta_n(\theta) &:= (R_n(\theta) - R(\theta)) - (R_n(\theta_0) - R(\theta_0)). \end{aligned} \tag{2}$$

Both of these processes are evidently mean zero.

We are most often concerned with upper bounds on the modulus of continuity relative to $1/\sqrt{n}$—the typical central limit theorem rate. That is, we consider functions ϕ of the form that

$$\mathbb{E}\left[\sup_{\theta \in \Theta_\delta} |(R_n(\theta) - R(\theta)) - (R_n(\theta_0) - R(\theta_0))|\right] \le \frac{\phi(\delta)}{\sqrt{n}}$$

Often, these functions satisfy $\phi(\delta) \le \sigma\delta$, where σ is a type of standard deviation/variance measure (though for fuller generality, we will consider functions $\phi(\delta) = \sigma\delta^\alpha$ for parameters $\alpha \in (0, 2)$). An example makes this more apparent.

Example 1: Let ℓ be L-Lipschitz in $\Theta \subset \mathbb{R}^d$ and take the norm $\|\cdot\|$ as the distance function, that is, $|\ell(\theta; x) - \ell(\theta'; x)| \le L\|\theta - \theta'\|$. Recalling the comparison process (2), we then have

$$\mathbb{E}\left[\exp\left(\lambda\Delta(\theta, X)\right)\right] \le \exp\left(\frac{\lambda^2 L^2 \|\theta - \theta'\|^2}{2}\right)$$

by the standard sub-Gaussian inequality for bounded random variables. Thus, letting $N(\Theta_\delta, \|\cdot\|, \epsilon)$ be the covering number of Θ_δ for the norm $\|\cdot\|$, we have

$$\log N(\Theta_\delta, \|\cdot\|, \epsilon) \le d\log\left(1 + \frac{2\delta}{\epsilon}\right)$$

and $\log N(\Theta_\delta, \|\cdot\|, \epsilon) = 0$ for $\epsilon \ge \delta$. Thus, a standard entropy integral calculation, using that $\frac{\sqrt{n}}{L}\Delta_n(\theta)$ is a $\|\cdot\|$-sub-Gaussian process, yields

$$\mathbb{E}\left[\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|\right] \le c\frac{L}{\sqrt{n}}\int_0^\delta \sqrt{\log N(\Theta_\delta, \|\cdot\|, \epsilon)}d\epsilon \le c\frac{L\sqrt{d}}{\sqrt{n}}\int_0^\delta \sqrt{\log\left(1 + \frac{2\delta}{\epsilon}\right)}d\epsilon \le c\frac{L\sqrt{d}}{\sqrt{n}}\delta,$$

where c is a numerical constant. That is, we have modulus of continuity bound with $\phi(\delta) = L\sqrt{d}\delta$, or $\sigma = L\sqrt{d}$. ♣

2.1 For intuition: non-stochastic bounds on differences in empirical risk

Because we would like to understand the relative differences between R_n and R, we begin for intuition by assuming that we have the unconditional bound that

$$|\Delta_n(\theta)| \le \frac{\phi(\delta)}{\sqrt{n}} \ \text{ whenever } d(\theta, \theta_0) \le \delta.$$

Then intuitively, we must have $d(\widehat{\theta}_n, \theta_0)$ small whenever the quadratic growth $\nu d(\theta, \theta_0)^2$ in $R(\theta)$ away from θ_0 dominates (or overcomes) the "stochastic" error $\phi(\delta)/\sqrt{n}$ in our estimation.

Let us make this rigorous, and begin by assuming that $d(\theta, \theta_0) \le \eta$, that is, θ is in the region of quadratic growth (1) of R away from $R(\theta_0)$, and let $\nu = 1$ for simplicity and w.l.o.g. Now, let $\delta = d(\theta, \theta_0)$, and assume that $R_n(\theta) \le R_n(\theta_0)$, that is, θ has smaller empirical risk than θ_0. Then we have

$$\begin{aligned} R_n(\theta) \le R_n(\theta_0) &= R_n(\theta_0) - R(\theta_0) + R(\theta) + \underbrace{R(\theta_0) - R(\theta)}_{\le -d(\theta,\theta_0)^2} \\ &\le R_n(\theta_0) - R(\theta_0) + R(\theta) - d(\theta, \theta_0)^2, \end{aligned}$$

where we have used the condition (1). Rearranging, we find that

$$d(\theta, \theta_0)^2 \le R_n(\theta_0) - R(\theta_0) + R(\theta) - R_n(\theta) \le |\Delta_n(\theta)| \le \frac{\phi(\delta)}{\sqrt{n}}.$$

That is, we have the key inequality

$$\delta^2 \le \frac{\phi(\delta)}{\sqrt{n}}. \tag{3}$$

This inequality is the key insight to all of our considerations of moduli of continuity: if $\phi(\delta)$ does not grow as fast as δ^2 and δ were large, this would contradict inequality (3), so $\delta = d(\theta, \theta_0)$ must be small. Said differently, for suitably large δ ("suitably large" will decrease as n grows), the quadratic growth δ^2 will eventually swamp the stochastic error $\phi(\delta)/\sqrt{n}$ based on inequality (3).

More carefully, suppose that

$$\phi(\delta) \le \sigma \delta^\alpha$$

for some $\alpha \in (0, 2)$. Then inequality (3) implies

$$\delta^2 \le \frac{\sigma \delta^\alpha}{\sqrt{n}}, \quad \text{or} \quad \delta \le \left(\frac{\sigma^2}{n}\right)^{\frac{1}{2(2-\alpha)}}.$$

2.2 Moduli of continuity and convergence guarantees

We now show how to make the (non-stochastic) heuristic argument of the preceding section rigorous. Assume that we have the modulus of continuity bound

$$\mathbb{E}\left[\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|\right] = \mathbb{E}\left[\sup_{\theta \in \Theta_\delta} |(R_n(\theta) - R(\theta)) - (R_n(\theta_0) - R(\theta_0))|\right] \le \frac{\phi(\delta)}{\sqrt{n}} \tag{4}$$

for all $\delta \le \eta$, where $\eta > 0$ is the region of strong convexity of R (inequality (1)). Assume additionally that $\phi(\delta) \le \sigma \delta^\alpha$ for some variance parameter σ and a power $\alpha \in (0, 2)$. Then we choose the rate

δ_n to be the point at which the quadratic growth "dominates" the stochastic error in the modulus of continuity (4), that is,

$$\delta_n^* := \inf\left\{\delta \geq 0 : \delta^2 \geq \frac{\phi(\delta)}{\sqrt{n}}\right\}. \tag{5}$$

Noting that $\phi(\delta) \leq \sigma\delta^\alpha$, then we certainly have that

$$\delta_n^\star = \left(\frac{\sigma^2}{n}\right)^{\frac{1}{2(2-\alpha)}}$$

is sufficient to satisfy this domination condition, that is, we have $\delta_n^\star \geq \delta_n^*$. Moreover, we have $\phi(\delta_n^*)/((\delta_n^*)^2\sqrt{n}) \leq 1$, and similarly for $\delta_n^\star$.

Thus, at least intuitively, we expect that the rate of convergence of $\widehat{\theta}_n$ to θ_0 should be roughly of order $\delta_n^* \leq \delta_n^\star$, because this is the point at which the curvature of the risk dominates the stochastic error in its estimation. We may formalize this in the following theorem.

Theorem 1 (Rates of convergence)**.** *Let δ_n^* be the smallest dominating radius (5), where the empirical risk R_n satisfies the modulus condition (4) and $\phi(\delta) \leq \sigma\delta^\alpha$. Assume also that $\widehat{\theta}_n = \operatorname{argmin}_\theta R_n(\theta)$ is consistent, that is, $\widehat{\theta}_n \xrightarrow{p} \theta_0$. Then*

$$d(\widehat{\theta}_n, \theta_0) = O_P(\delta_n^*) = O_P(\delta_n^\star) = O_P\left(\left(\frac{\sigma^2}{n}\right)^{\frac{1}{2(2-\alpha)}}\right).$$

Proof Our proof builds off of a so-called *peeling* argument, where we argue that the behavior of the local relative errors $\Delta_n(\theta)$ is nice on shells around θ_0. Indeed, for each n and all $j \in \mathbb{N}$, define the shells

$$S_{j,n} := \left\{\theta \in \Theta : \delta_n^* 2^{j-1} \leq d(\theta, \theta_0) \leq \delta_n^* 2^j\right\}.$$

Recall the definition $\eta > 0$ of the radius in the quadratic growth condition (1). Now, fix any $t \in \mathbb{R}_+$, and consider the event that $d(\widehat{\theta}_n, \theta_0) \geq 2^t\delta_n^*$. Then either $d(\widehat{\theta}_n, \theta_0) \geq \eta$ or we have $\widehat{\theta}_n \in S_{j,n}$ for some j with $j \geq t$ but $2^j\delta_n^* \leq \eta$. In particular,

$$\mathbb{P}\left(d(\widehat{\theta}_n, \theta_0) \geq 2^t\delta_n^*\right) \leq \sum_{j : j \geq t, 2^j\delta_n^* \leq \eta} \mathbb{P}(\widehat{\theta}_n \in S_{j,n}) + \mathbb{P}(d(\widehat{\theta}_n, \theta_0) \geq \eta). \tag{6}$$

The final term is $o(1)$, so we may ignore it in what follows.

Now, consider the event that $\widehat{\theta}_n \in S_{j,n}$. This implies that there exists some $\theta \in S_{j,n}$ such that $R_n(\theta) \leq R_n(\theta_0)$, in turn implying

$$R_n(\theta) \leq R_n(\theta_0) - R(\theta_0) + R(\theta) + R(\theta_0) - R(\theta) \leq R_n(\theta_0) - R(\theta_0) + R(\theta) - \nu d(\theta, \theta_0)^2,$$

where we have used the growth condition (1) that $R(\theta) \geq R(\theta_0) + \nu d(\theta, \theta_0)^2$, which holds for $\theta \in S_{j,n}$ as $d(\theta, \theta_0) \leq \eta$. Noting that $d(\theta, \theta_0)^2 \geq (\delta_n^*)^2 2^{2j-2}$, we rearrange the preceding inequality to obtain that $\widehat{\theta}_n \in S_{j,n}$ implies

$$\nu(\delta_n^*)^2 2^{2j-2} \leq R_n(\theta_0) - R(\theta_0) - (R_n(\theta) - R(\theta)) \leq \sup_{\theta \in S_{j,n}} |\Delta_n(\theta)|.$$

Returning to the probability sum (6), we thus have

$$\mathbb{P}(\widehat{\theta}_n \in S_{j,n}) \leq \mathbb{P}\left(\sup_{\theta \in S_{j,n}} |\Delta_n(\theta)| \geq \nu(\delta_n^*)^2 2^{2j-2}\right) \leq \frac{\mathbb{E}[\sup_{\theta \in S_{j,n}} |\Delta_n(\theta)|]}{\nu(\delta_n^*)^2 2^{2j-2}}.$$

But of course, by assumption (4), this in turn has the bound

$$\mathbb{P}(\widehat{\theta}_n \in S_{j,n}) \le \frac{2^{2-2j}}{\nu} \frac{\phi(2^j \delta_n^*)}{(\delta_n^*)^2 \sqrt{n}} \le \frac{2^{2-2j} \cdot 2^{j\alpha}}{\nu} \cdot \frac{\phi(\delta_n^*)}{(\delta_n^*)^2 \sqrt{n}} \le \frac{2^{2-2j} \cdot 2^{j\alpha}}{\nu}$$

by the definition (5) of the critical radius for δ_n^*.

Summing inequality (6), we thus obtain

$$\mathbb{P}\left(d(\widehat{\theta}_n, \theta_0) \ge 2^t \delta_n^*\right) \le \frac{4}{\nu} \sum_{j \ge t} 2^{-j(2-\alpha)} + o(1).$$

For any $\epsilon > 0$, we may take t sufficiently large that $\sum_{j \ge t} 2^{-j(2-\alpha)} \le \epsilon$, which is the definition of $d(\widehat{\theta}_n, \theta_0) = O_P(\delta_n^*)$. □

References

[1] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York, 1996.