1 Introduction

In this note, we give a presentation showing the importance, and relationship between, the moduli 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 $\Theta$ 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 $p_\theta$. We then draw $X_i \overset{iid}{\sim} P$, 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 $\hat{\theta}_n = \arg\min_{\theta \in \Theta} R_n(\theta)$ to $\theta_0 := \arg\min_{\theta \in \Theta} R(\theta)$, the population minimizer.

It is natural, based on a Taylor expansion, to assume that in a neighborhood of $\theta_0$, 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

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

With such a condition, it is possible to give rates of convergence of $\hat{\theta}_n$ to $\theta_0$, at least so long as the random functions $R_n$ do not have so much variability in a neighborhood of $\theta_0$ that they swamp the quadratic growth away from $\theta_0$.

2 Rates of convergence and comparison of functions

Because we would like to understand minimizing the population risk $R$ and finding $\theta_0$, 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 $\hat{\theta}_n \to \theta_0$, it is not necessary. Indeed, all we really care about is that $R_n(\theta) > R_n(\theta_0)$ for $\theta$ sufficiently far from $\theta_0$. That is, as we expect to have roughly $R_n(\theta) - R_n(\theta_0) \approx R(\theta) - R(\theta_0)$, where $R(\theta) \geq R(\theta_0) + \nu d(\theta, \theta_0)^2$, 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 $\hat{\theta}_n$ is close enough to $\theta_0$, we can give stronger convergence guarantees, because we expect $\text{Var}(R_n(\theta) - R_n(\theta_0))$ to be smaller than $\text{Var}(R_n(\theta))$ by itself. A bit more precisely, we must have deviations roughly
\[ \frac{1}{\sqrt{n}} \sqrt{\text{Var}(\ell(\theta, X))} \] in any uniform estimate of \( R(\theta) \), by the central limit theorem. However, if \( \theta \) is near \( \theta_0 \), then
\[ R_n(\theta) - R_n(\theta_0) = R(\theta) - R(\theta_0) + O_P \left( n^{-\frac{1}{2}} \sqrt{\text{Var}(\ell(\theta; X) - \ell(\theta_0; X))} \right), \]
and the latter variance may be substantially smaller than \( \text{Var}(\ell(\theta, X)) \) 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 \( \theta_0 \). We make the following definition.

**Definition 2.1.** Let \( \Theta_{\delta} := \{ \theta \in \Theta : d(\theta, \theta_0) \leq \delta \} \). The expected modulus of continuity of the process \( R_n \) in a radius \( \delta \) around \( \theta_0 \) is
\[ \mathbb{E} \left[ \sup_{\theta \in \Theta_{\delta}} \left| (R_n(\theta) - R(\theta)) - (R_n(\theta_0) - R(\theta_0)) \right| \right]. \]
For notational convenience, we also define the error processes
\[ \Delta(\theta, x) := (\ell(\theta, x) - R(\theta)) - (\ell(\theta_0, x) - R(\theta_0)) \quad \text{and} \quad \Delta_n(\theta) := (R_n(\theta) - R(\theta)) - (R_n(\theta_0) - R(\theta_0)). \]
(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 \( \phi \) of the form that
\[ \mathbb{E} \left[ \sup_{\theta \in \Theta_{\delta}} \left| (R_n(\theta) - R(\theta)) - (R_n(\theta_0) - R(\theta_0)) \right| \right] \leq \frac{\phi(\delta)}{\sqrt{n}}. \]
Often, these functions satisfy \( \phi(\delta) \leq \sigma \delta \), where \( \sigma \) 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 \( \ell \) be \( L \)-Lipschitz in \( \Theta \subset \mathbb{R}^d \) and take the norm \( \| \cdot \| \) as the distance function, that is, \( |\ell(\theta; x) - \ell(\theta'; x)| \leq L \| \theta - \theta' \|. \) Recalling the comparison process (2), we then have
\[ \mathbb{E} \left[ \exp (\lambda \Delta(\theta, X)) \right] \leq \exp \left( \frac{\lambda^2 L^2 \| \theta - \theta' \|^2}{2} \right) \]
by the standard sub-Gaussian inequality for bounded random variables. Thus, let \( N(\Theta_{\delta}, \| \cdot \|, \epsilon) \) be the covering number of \( \Theta_{\delta} \) for the norm \( \| \cdot \| \), we have
\[ \log N(\Theta_{\delta}, \| \cdot \|, \epsilon) \leq d \log \left( 1 + \frac{2\delta}{\epsilon} \right) \]
and \( \log N(\Theta_{\delta}, \| \cdot \|, \epsilon) = 0 \) for \( \epsilon \geq \delta \). Thus, a standard entropy integral calculation, using that \( \frac{1}{\sqrt{n}} \Delta_n(\theta) \) is a \( \| \cdot \| \)-sub-Gaussian process, yields
\[ \mathbb{E} \left[ \sup_{\theta \in \Theta_{\delta}} |\Delta_n(\theta)| \right] \leq c \frac{L}{\sqrt{n}} \int_0^\delta \sqrt{\log N(\Theta_{\delta}, \| \cdot \|, \epsilon)} \, d\epsilon \leq c \frac{L \sqrt{d}}{\sqrt{n}} \int_0^\delta \sqrt{\log \left( 1 + \frac{2\delta}{\epsilon} \right)} \, d\epsilon \leq c \frac{L \sqrt{d} \delta}{\sqrt{n}}, \]
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

\[
|\Delta_n(\theta)| \leq \frac{\phi(\delta)}{\sqrt{n}} \quad \text{whenever} \quad d(\theta, \theta_0) \leq \delta.
\]

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

Let us make this rigorous, and begin by assuming that \( d(\theta, \theta_0) \leq \eta \), that is, \( \theta \) 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 \( R_n(\theta) \leq R_n(\theta_0) \), that is, \( \theta \) has smaller empirical risk than \( \theta_0 \). Then we have

\[
R_n(\theta) \leq R_n(\theta_0) = R_n(\theta_0) - R(\theta) + R(\theta_0) - R(\theta) \leq -d(\theta, \theta_0)^2,
\]

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

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

That is, we have the key inequality

\[
\delta^2 \leq \frac{\phi(\delta)}{\sqrt{n}}. \tag{3}
\]

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

More carefully, suppose that

\[
\phi(\delta) \leq \sigma \delta^\alpha
\]

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

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

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

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

for all \( \delta \leq \eta \), where \( \eta > 0 \) is the region of strong convexity of \( R \) (inequality (1)). Assume additionally that \( \phi(\delta) \leq \sigma \delta^\alpha \) for some variance parameter \( \sigma \) and a power \( \alpha \in (0, 2) \). Then we choose the rate
Let \( \delta_n \) to be the point at which the quadratic growth “dominates” the stochastic error in the modulus of continuity (4), that is,

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

(5)

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

\[
\delta_n^* = \left( \frac{\sigma^2}{n} \right)^{\frac{1}{2(2-\alpha)}}
\]

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

Thus, at least intuitively, we expect that the rate of convergence of \( \hat{\theta}_n \) to \( \theta_0 \) should be roughly of order \( \delta_n^* \sim \delta_n^{*\ast} \), 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 \( \delta_n^* \) 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 \( \hat{\theta}_n = \arg \min_{\theta} R_n(\theta) \) is consistent, that is, \( \hat{\theta}_n \xrightarrow{P} \theta_0 \). Then

\[
d(\hat{\theta}_n, \theta_0) = O_P(\delta_n^{*\ast}) = O_P(\delta_n^*) = O_P \left( \left( \frac{\sigma^2}{n} \right)^{\frac{1}{2(2-\alpha)}} \right)
\]

**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 \( \theta_0 \). Indeed, for each \( n \) and all \( j \in \mathbb{N} \), define the shells

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

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 \( d(\hat{\theta}_n, \theta_0) \geq 2^t \delta_n^{*\ast} \). Then either \( d(\hat{\theta}_n, \theta_0) \geq \eta \) or we have \( \hat{\theta}_n \in S_{j,n} \) for some \( j \) with \( j \geq t \) but \( 2^j \delta_n^{*\ast} \leq \eta \). In particular,

\[
\mathbb{P} \left( d(\hat{\theta}_n, \theta_0) \geq 2^t \delta_n^{*\ast} \right) \leq \sum_{j : j \geq t, \delta_n^{*\ast} \leq 2^j} \mathbb{P}(\hat{\theta}_n \in S_{j,n}) + \mathbb{P}(d(\hat{\theta}_n, \theta_0) \geq \eta).
\]

(6)

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

Now, consider the event that \( \hat{\theta}_n \in S_{j,n} \). 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 \( d(\theta, \theta_0)^2 \geq (\delta_n^{*\ast})^2 2^{2j-2} \), we rearrange the preceding inequality to obtain that \( \hat{\theta}_n \in S_{j,n} \) implies

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

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

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

4
But of course, by assumption (4), this in turn has the bound
\[
P(\hat{\theta}_n \in S_{j,n}) \leq \frac{2^{2-2j}}{\nu} \frac{\phi(2^j \delta^*_n)}{(\delta^*_n)^2 \sqrt{n}} \leq \frac{2^{2-2j} \cdot 2^{j\alpha}}{\nu} \cdot \frac{\phi(\delta^*_n)}{(\delta^*_n)^2 \sqrt{n}} \leq \frac{2^{2-2j} \cdot 2^{j\alpha}}{\nu}
\]
by the definition (5) of the critical radius for \(\delta^*_n\).

Summing inequality (6), we thus obtain
\[
P \left( d(\hat{\theta}_n, \theta_0) \geq 2^t \delta^*_n \right) \leq \frac{4}{\nu} \sum_{j \geq t} 2^{-j(2-\alpha)} + o(1).
\]
For any \(\epsilon > 0\), we may take \(t\) sufficiently large that \(\sum_{j \geq t} 2^{-j(2-\alpha)} \leq \epsilon\), which is the definition of \(d(\hat{\theta}_n, \theta_0) = O_P(\delta^*_n)\).

References