Measuring the Algorithmic Convergence of Random Forests via Bootstrap Extrapolation

Miles E. Lopes
melopes@ucdavis.edu
Department of Statistics
University of California, Davis
November 15, 2015

Abstract
When making predictions with a voting rule, a basic question arises: “What is the smallest number of votes needed to make a good prediction?” In the context of ensemble classifiers, such as Random Forests or Bagging, this question represents a tradeoff between computational cost and statistical performance. Namely, by paying a larger computational price for more classifiers, the prediction error of the ensemble tends to improve and become more stable. Conversely, by using fewer classifiers and tolerating some variability in accuracy, it is possible to speed up the tasks of training the ensemble and making new predictions. In this paper, we propose a bootstrap method to quantify this tradeoff for the methods of Bagging and Random Forests. To be specific, suppose the training dataset is fixed, and let the random variable $Err_t$ denote the prediction error of a randomly generated ensemble of $t = 1, 2, \ldots$ classifiers. (The randomness of $Err_t$ comes only from the algorithmic randomness of the ensemble.) Working under a “first order model” of Random Forests, we prove that the centered law of $Err_t$ can be consistently estimated via our proposed method as $t \to \infty$. As a consequence, this result offers practitioners a guideline for choosing the smallest number of base classifiers needed to ensure that the algorithmic fluctuations are negligible, e.g. $\text{var}(Err_t)$ less than a given threshold.

1 Introduction
Random Forests and Bagging are among the most widely used non-parametric approaches to classification [Bre96, Bre01]. These methods are commonly referred to as “ensemble methods”, as they make predictions by taking the plurality vote of a large ensemble of randomized classifiers. Due to this randomization, it follows that if $Err_t$ denotes the prediction error achieved by an ensemble of $t$ classifiers, then $Err_t$ is a random variable — even when the training dataset is viewed as fixed. Likewise, it is of basic interest to measure the fluctuations of $Err_t$, conditionally on the training data, in order to assess the algorithmic convergence of ensemble classifiers.*

In this paper, we propose a bootstrap method to approximate the centered distribution of $Err_t$, conditioned on the training dataset. The method is general insofar as it allows for an arbitrary number of label classes $k \geq 2$, and does not depend on the internal mechanisms of the classifiers. In particular, the method can be applied an ensemble of random classifiers that are i.i.d. “black boxes”, or even an ensemble of human decision makers who are drawn in an i.i.d. manner from a population.

*We remark that this problem is completely distinct from the more traditional problem of analyzing the error rate of Random Forests or Bagging as a function of the size of the training set, i.e. the “large-sample limit,” which has been studied in depth by a number of authors. We refer to [BDL08, Bia12] and references therein. In particular, our use of $t$ for the number of classifiers should not be confused with the size of the training set.
1.1 Motivations

Measuring the algorithmic fluctuations of $\text{Err}_t$ has several applications, from both statistical and computational viewpoints.

- **Statistical and computational tradeoffs.** Whenever Random Forests or Bagging are implemented in practice, a basic question arises: How many votes are needed to obtain a reliable prediction? In general, one expects that as $t \to \infty$, the error $\text{Err}_t$ will stabilize to some constant limiting value, say $\text{err}$. This value is typically regarded as a benchmark, in the sense that $t$ should be chosen large enough so that $\text{Err}_t$ is close to $\text{err}$ [FHT01, Section 15.3.1]. But of course, as $t$ increases, so do the computational costs of training the classifiers and storing them in memory — especially when the training set is large. Consequently, the choice of $t$ encodes a tradeoff between statistical performance and computational cost. (A numerical example is illustrated in Figure 1.) Our method quantifies this tradeoff in a precise way by measuring the fluctuations of $\text{Err}_t$ as a function of $t$. For example, a natural criterion is to ensure that $\sqrt{\text{var}(\text{Err}_t)}$ is small in comparison to $\text{Err}_t$. Hence, using estimates $\hat{\text{var}}(\text{Err}_t)$ and $\hat{\text{Err}}_t$ one may select $t$ when $\sqrt{\hat{\text{var}}(\text{Err}_t)} \leq \varepsilon \cdot \hat{\text{Err}}_t$ for some threshold $\varepsilon > 0$.

- **Fairly comparing algorithms.** The task of making a fair comparison between classification algorithms is an everyday part of data analysis. Using randomized classifiers adds a degree of complication. Even if we knew that the true error rate $\text{Err}_t$ of a realized ensemble is somewhat better than that of a competing method, how can we be sure that this advantage is significant, and not merely a fluke of randomization? In principle, this could be answered by repeatedly training many ensembles of size $t$, but this is clearly impractical in large-scale applications. Our approach provides an answer with much less computational effort by resampling the classifiers of a single ensemble. Moreover, our use of extrapolation allows this resampling to be done with a relatively small ensemble (see Section 2.2).

- **Other applications of voting algorithms.** The convergence of voting algorithms is relevant to many other areas where there is a tradeoff between the performance of the algorithm, and the cost of obtaining more votes. A few examples of such areas include crowdsourcing, recommender systems, prediction markets, and social choice theory [RV97, EK10]. Due to the fact that our proposed method is applicable to ensembles of black box classifiers or human decision makers, the method may be useful in some of these other contexts.

1.2 Problem setup

**Notation for classification.** In a generic classification problem involving $k \geq 2$ label classes, we have a set of labeled training samples $\mathcal{D} = \{(X_1,Y_1),\ldots,(X_N,Y_N)\}$ in a sample space $\mathcal{X} \times \{e_0,\ldots,e_{k-1}\}$. Here we use a somewhat non-standard class labeling scheme that will be useful later on. The label for class “0” is $e_0 := (0,0,\ldots,0) \in \mathbb{R}^{k-1}$, and the label for class “1” is $e_1 := (1,0,\ldots,0) \in \mathbb{R}^{k-1}$. Likewise, the label for class “2” is $e_2 := (0,1,0,\ldots,0) \in \mathbb{R}^{k-1}$, and so on up to class “$k-1$”. When a new test point $(X,Y)$ is randomly drawn from $\mathcal{X} \times \{e_0,\ldots,e_{k-1}\}$, a classification algorithm seeks to predict the value of $Y$ from $X$, such that the probability of error is small.

**Random Forests and Bagging for classification.** The essential idea of an ensemble method is to generate from $\mathcal{D}$ a collection of *base classifiers* $Q_i : \mathcal{X} \to \{e_0,\ldots,e_{k-1}\}$, $i = 1,\ldots,t$, and then predict the label of $X$ by aggregating the values $Q_i(X)$ with a voting rule. In the cases of Random
Figure 1: The black curve shows how $\text{Err}_t$ varies as a function of $t$ for a particular realization of $Q_1, \ldots, Q_t$. The red curve was obtained by generating many ensembles, and computing the average of the respective $\text{Err}_t$ values at each $t$. Similarly, the blue curves depict the 10th and 90th percentiles of $\text{Err}_t$ at each $t$.

Forests or Bagging, the functions $Q_i(\cdot)$ are typically decision trees trained on random (bootstrapped) subsets of $D$. As such, the $Q_i(\cdot)$ are random functions. In fact, additional randomization may occur in many other ways, such as the selection of various tuning parameters for each tree. Despite the complex variety of ways to construct the $Q_i(\cdot)$, our proposed method is derived only from the assumption that the classifiers are generated in an i.i.d. manner.† Specifically, we will use the following abstract definition of Random Forests, as introduced by Breiman [Bre01, Definition 1.1] (See also [BDL08]). Note that this definition includes Bagging as a special case, since sampling with replacement generates i.i.d. subsets of $X$.

**Definition 1 (random forests).** A Random Forest is a classifier whose predictions are obtained from the plurality vote of a sequence of i.i.d. functions $Q_i : X \to \{e_0, \ldots, e_{k-1}\}$, with $i = 1, 2, \ldots$.

The phrase *plurality vote* simply refers to the “most popular” class. That is, for each $j = 0, \ldots, k-1$, we define

$$N_j(X) := \text{card}\{i : Q_i(X) = e_j\}$$

as the number of votes assigning $X$ to class $j$, and we write

$$\text{plurality vote} = \arg\max_{0 \leq j \leq k-1} N_j(X),$$

where ties are broken arbitrarily. We distinguish between majority and plurality, since majority refers to the class that receives at least half the votes. When $k = 2$, the two notions coincide, but when $k \geq 3$, it may happen that a majority does not exist — and this is often a typical situation in applications. By contrast, a plurality always exists.

†Later on, our theoretical analysis will study a particular version of the Random Forests algorithm that satisfies additional assumptions.
**Error rates.** To define error rates precise terms, it is important to note that all of our analysis will be done conditionally on the training set $D$. As a result, the training set will play no explicit role in our analysis, and we will henceforth view $D$ as being non-random. There are only two other sources of randomness to consider: the test point $(X,Y)$, and the the classifiers $Q_i$. We assume throughout that $(X,Y)$ is independent of the sequence $Q_1, \ldots, Q_t$. If we let $P$ denote the joint distribution of $(X,Y,Q_1,Q_2,\ldots)$, then for any fixed $j \in \{0,\ldots,k-1\}$, we define the conditional error rate on the $j$th class to be

$\text{Err}_{t,j} := P\left(N_j(X) \leq \max_{l \neq j} (N_l(X)) \mid Q_1, \ldots, Q_t, Y = e_j\right).$  

(3)

In other words, if we are given a particular realization of the $Q_i$, and if $X$ is randomly drawn from the $j$th class, then $\text{Err}_{t,j}$ is the chance that the $j$th class fails to receive the most votes. (This definition eliminates the nuisance of ties by simply counting a tie as an error.) Summing these error rates across classes, and defining the class proportions $\pi_j$ as

$\pi_j := P(Y = e_j),$

(4)

we obtain the (total) conditional error rate,

$\text{Err}_t := \sum_{j=0}^{k-1} \pi_j \cdot \text{Err}_{t,j}.$

(5)

Again, this is a random variable, as it depends on a realization of the $Q_i$. Going forward, we will treat the $\pi_j$ as being known for simplicity, but of course, these can be estimated directly from the observed proportions of each type of label in the training set.

### 1.3 Contributions

The main contributions of the paper are twofold: a new bootstrap method for approximating the conditional law of $\text{Err}_t$ given $Q_1, \ldots, Q_t$, as well as a theoretical result showing that the method is consistent under a simplified model of Random Forests.

**Bootstrap method and extrapolation.** Conventionally, bootstrap methods work by viewing a statistic of interest as a function of random observations, and then re-evaluating the statistic on re-sampled observations. Our perspective in the current paper is different. Instead, we view the training data as being fixed, and we are interested in a quantity that is a function of a randomized algorithm, i.e. $\text{Err}_t$ as a function of $Q_1, \ldots, Q_t$. Accordingly, our proposed method works by resampling the classifiers rather than data. In this way, our method mimics the algorithmic fluctuations rather than the fluctuations arising from the training data.

Given that one of our motivations is optimizing the tradeoff between computational cost and prediction stability, the use of a bootstrap approach may sound foolhardy — since bootstrap methods are often regarded as “computationally intensive”. However, our proposed method has several attributes that keep computational costs in check.

First, with regard to the problem of estimating $\sqrt{\text{var}(\text{Err}_t)}$ as a function of $t$, we make use of the technique of extrapolation as a means of saving computation. This technique allows us to measure the fluctuations of $\text{Err}_t$ for a large ensemble while only paying resampling costs for a small ensemble. More specifically, we compute a bootstrap-based estimate of $\sqrt{\text{var}(\text{Err}_{t_0})}$ for a small ensemble size $t_0$, and then use a simple rescaling to derive an estimate of $\sqrt{\text{var}(\text{Err}_t)}$ for $t \gg t_0$.

Second, when held-out set of examples is available, our proposed bootstrap method only requires access to the predicted label on the held-out set. In particular, our method does not require any
re-evaluation of the classifiers, or any resampling of the original training data. Furthermore, our proposed resampling method enjoys the property of being "embarrassingly parallel" in the same way that conventional bootstrap methods are.

**Bootstrap consistency and a non-smooth first variation formula.** The paper’s main theoretical result, Theorem 1 asserts that the bootstrap works in the sense that $\text{Err}_t$ and its bootstrap analogue $\text{Err}^*_t$ have the same limiting distribution (after both are standardized). Conceptually, our proof proceeds through the ideas of Hadamard differentiability and the functional delta method vol VW96. Meanwhile, with regard to technique, the proof introduces to two ingredients of particular interest.

The first ingredient is a “lifting operator” that maps the usual empirical process $F_t : [0, 1] \to [0, 1]$ to a $(k - 1)$-dimensional version $V_t : \Delta \to \Delta$, where $\Delta \subset \mathbb{R}^{k-1}$ is a simplex, and $F_t(u) := \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}\{U_i \leq u\}$ with $U_1, \ldots, U_t \sim \text{Uniform}[0,1]$ being i.i.d. In addition to having several interesting properties in its own right, the lifting operator allows us to represent $\text{Err}_t$ as a functional the empirical process. In turn, this representation allows us to convert bootstrap consistency for the empirical process into bootstrap consistency for $\text{Err}_t$. Furthermore, the relevant functional of the empirical process turns out to be Hadamard differentiable.

The second ingredient is the calculation of the Hadamard derivative of the functional just mentioned. A suprising aspect of this calculation is that it turns out to require a non-smooth version of the so-called first variation formula of differential geometry Sim83 Section 9. Informally, the first variation formula may be stated as follows. Let $\mathcal{M}$ be a smooth manifold, and let $\{\varphi_\varepsilon\}_{\varepsilon \in [-1,1]}$ be a family of diffeomorphisms $\varphi_\varepsilon : \mathcal{M} \to \mathcal{M}$, satisfying $\varphi_\varepsilon \to \text{id}_\mathcal{M}$ as $\varepsilon \to 0$. Then,

$$\frac{d}{d\varepsilon} \text{vol}(\varphi_\varepsilon(\mathcal{M}))|_{\varepsilon=0} = - \int_{\partial \mathcal{M}} \langle Z(\theta), \mathbf{n}(\theta) \rangle d\sigma(\theta), \quad (6)$$

where $\text{vol}$ is a volume measure on $\mathcal{M}$, $\mathbf{n}$ is the outward normal to the boundary $\partial \mathcal{M}$, $d\sigma$ is the surface measure, and $Z$ is a vector field given by $Z(\theta) = \frac{\partial}{\partial \varepsilon} \varphi_\varepsilon(\theta)|_{\varepsilon=0}$. In our analysis, it is necessary to prove an analogous result where the $\varphi_\varepsilon$ are non-smooth maps, $\mathcal{M}$ is a non-smooth convex subset of $\mathbb{R}^{k-1}$, and the vector field $Z$ is a Gaussian process. Furthermore, the linear functional $Z(\cdot) \mapsto -\int_{\partial \mathcal{M}} \langle Z(\theta), \mathbf{n}(\theta) \rangle d\sigma(\theta)$ is the Hadamard derivative of the functional relevant to $\text{Err}_t$.

### 1.4 Related work

In the setting where the functions of an ensemble are independent, conditionally on the training data, the general problem of choosing the ensemble size has been studied from a variety of viewpoints. For instance, the papers NJ01, Lop13, CS15 analyze the convergence of the expected error rate $\mathbb{E}[\text{Err}_t]$ as a function of $t$. In the regression setting, the bias of an ensemble has been studied as a function of $t$ in the paper AGL4.

Another approach for classification is taken in the paper HLMMS13. That paper studies the smallest number of functions $t^*_\alpha$ such that the binary prediction of a finite ensemble agrees with that of an infinite ensemble, with probability at least $\alpha$. Some theoretical analysis is also given showing that the “agreement probability” scales in $t$ as $c/\sqrt{t}$ for some constant $c$, and an asymptotic formula is derived for $c$ is derived.

The approach in the current paper differs from previous works in several ways. First, our analysis here handles the general situation of $k \geq 2$ classes, whereas other works we are aware of focus on the binary case. Second, another advantage of our results is that they provide a distributional approximation of $\text{Err}_t$. Third, we believe that our resampling approach can provide a unified method
for measuring the convergence of ensembles in both the classification and regression settings, but we leave this for future work.

For other theoretical and empirical results on the convergence of ensemble methods that are less relevant to the approach we have taken here, we refer to the papers [RG02, LS97, LDD01, BHBK07, OPB12].

1.5 Outline

The remainder of the paper is organized as follows. Our proposed bootstrap method is described in Section 2, and our main consistency result is given in Section 3. The essential aspects of the proofs are presented in Sections 3.4 - 3.6 and the technical arguments are given in the appendices.

1.6 Notation and terminology

If \( S \subset \mathbb{R}^d \) is a subset of Euclidean space, we write \( S^\circ \) for the interior of \( S \) and \( \text{cl}(S) \) for the closure of \( S \), in the usual topology. Also, we define the boundary of \( S \) by \( \partial S := \text{cl}(S) \setminus S^\circ \). If \( U \) is an open subset of Euclidean space, we say a function \( f : U \to \mathbb{R} \) is \( C^1 \) if all partial derivatives of \( f \) exist and are continuous on \( U \). The identity map on a set \( S \) is denoted \( \text{id}_S \), or simply \( \text{id} \) when there is no ambiguity. We use the symbol \( \overset{w}{\longrightarrow} \) to denote weak convergence, and \( \overset{p}{\longrightarrow} \) to denote convergence in probability. The spaces \( F[0, 1], C[0, 1], D[0, 1], \) and \( \ell^\infty[0, 1] \) respectively denote the following sets of real-valued functions on the unit interval: cumulative distribution functions, continuous functions, càdlàg functions, and bounded functions (all equipped with the supremum norm). If \( W \) and \( V \) are random variables (or sets of random variables), then \( \mathcal{L}(W|V) \) denotes the conditional law of \( W \) given \( V \).

2 Method

A concrete algorithm for our bootstrap method will be given in Section 2.1.2. In order to explain the conceptual underpinnings of the method, we first introduce some notation. Let \( \Delta \) denote the “full-dimensional” simplex in \( \mathbb{R}^{k-1} \).

\[
\Delta := \left\{ \theta \in \mathbb{R}^{k-1} \left| \theta_1 + \cdots + \theta_{k-1} \leq 1, \text{ and } \min_{1 \leq i \leq k-1} \theta_i \geq 0 \right. \right\}. \tag{7}
\]

If \( \theta = (\theta_1, \ldots, \theta_{k-1}) \) is a vector in \( \Delta \), it will be convenient to use the symbol

\[
\theta_0 := 1 - \sum_{j=1}^{k-1} \theta_j. \tag{8}
\]

Next, define the average classifier

\[
\bar{Q}(\cdot) := \frac{1}{t} \sum_{i=1}^{t} Q_i(\cdot), \tag{9}
\]

which satisfies \( \bar{Q}(X) \in \Delta \) for all realizations of \( X \in \mathcal{X} \). For any \( j = 0, 1, \ldots, k-1 \), if the true label\(^1\) is \( Y = e_j \), then it is simple to check that the plurality vote makes an error iff

\[
\bar{Q}(X) \in S_j, \tag{10}
\]

where \( S_j \) is a convex subset of \( \Delta \) defined by

\[
\begin{align*}
S_j := \left\{ \theta \in \Delta \left| \theta_j &\leq \max_{l \neq j} \theta_l \right. \right\}. \tag{11}
\end{align*}
\]

More precisely, \( \max_{l \neq j} \) refers to the maximum over the indices \( l \in \{0, 1, \ldots, k-1\} \setminus \{j\} \), with \( \theta_0 \) being as defined above.

\(^1\)Recall the notation \( e_0 = (0, 0, \ldots, 0), e_1 = (1, 0, \ldots, 0), e_2 = (0, 1, 0, \ldots, 0) \) and so on.
2.1 Bootstrapping Err$_t$

In the notation just introduced, the class-wise error rate on class $j = 0, \ldots, k - 1$ is given by

$$\text{Err}_{t,j} = P(\bar{Q}(X) \in S_j \mid Q_1, \ldots, Q_t, Y = e_j).$$  \hspace{1cm} (12)

Our proposed method is based on interpreting these probabilities as functionals of $\bar{Q}(\cdot)$. To proceed, let $\nu_j$ denote the conditional distribution of $X$ given $Y = e_j$, which gives

$$\nu_j(\bar{Q}^{-1}(S_j)) = P(\bar{Q}(X) \in S_j \mid Q_1, \ldots, Q_t, Y = e_j).$$  \hspace{1cm} (13)

In other words, $\text{Err}_{t,j}$ is the $\nu_j$-measure of the preimage of $S_j$ under $\bar{Q}(\cdot)$, which is a well-defined functional of $\bar{Q}(\cdot)$. If we denote this functional by $\varphi_j$, then

$$\text{Err}_{t,j} = \varphi_j(\bar{Q}) := \nu_j(\bar{Q}^{-1}(S_j)).$$  \hspace{1cm} (14)

Furthermore, if we put $\varphi(\cdot) := \sum_{j=0}^{k-1} \pi_j \varphi_j(\cdot)$, then the total error rate is simply

$$\text{Err}_t = \varphi(\bar{Q}).$$  \hspace{1cm} (15)

The advantage of the representation (15) is that it reveals $\text{Err}_t$ to be a functional of a sample mean — which is a type of statistic that is commonly known as being well-suited to bootstrap methods. Nevertheless, there are considerable technical challenges arising from the fact that $\bar{Q}$ is a sample mean in the space of functions $X \to \Delta$, and these challenges will be addressed in our theoretical analysis.

To describe the bootstrap in greater detail, let $\{Q^*_1(\cdot), \ldots, Q^*_t(\cdot)\}$ denote a random sample with replacement from the (observed) set of classifiers $\{Q_1(\cdot), \ldots, Q_t(\cdot)\}$, and put

$$\bar{Q}^*(\cdot) := \frac{1}{t} \sum_{i=1}^{t} Q^*_i(\cdot).$$  \hspace{1cm} (16)

In turn, we define

$$\text{Err}^*_t := \varphi(\bar{Q}^*)$$  \hspace{1cm} (17)

as a bootstrap sample of $\text{Err}_t$. However, line (17) should be regarded as an idealized bootstrap sample, because the value $\varphi(\bar{Q}^*)$ is unknown — as it requires some knowledge of the unknown distributions $\nu_j$. Even so, it is straightforward to estimate the value $\varphi(\bar{Q}^*)$ via a set of held out training points, out-of-bag samples, or some variant thereof. (See Section 2.1.2 below for an explicit description.) More precisely, if $\hat{\varphi}$ denotes a rule for estimating $\varphi$, then a practitioner’s estimate of $\text{Err}_t$ would be written as

$$\hat{\text{Err}}_t := \hat{\varphi}(\bar{Q}),$$  \hspace{1cm} (18)

and the corresponding bootstrap sample is

$$\hat{\text{Err}}^*_t := \hat{\varphi}(\bar{Q}^*).$$  \hspace{1cm} (19)

Since there is some choice involved in the estimation of $\varphi$, we present two versions of our resampling method. First, we describe the method in an abstract form where $\varphi$ is estimated by an unspecified rule $\hat{\varphi}$. Second, we describe the algorithm in a more concrete form when $\varphi$ is estimated using a held-out set. The rationale for splitting the presentation is to illustrate the conceptual simplicity of the method in the first form, while showing that the method is easy to implement in the second form. In either case, $B \geq 1$ denotes a user-specified number of bootstrap replicates, and the samples drawn at each iteration are assumed to be independent.
2.1.1 General resampling algorithm

For \( l = 1, \ldots, B \):

- Draw a set of \( t \) classifiers \( \{Q_1^*, \ldots, Q_t^*\} \) by random sampling with replacement from \( \{Q_1, \ldots, Q_t\} \).
- Compute \( z_l := \hat{\varphi}(\bar{Q}^*) \).

Return: empirical distribution of \( z_1, \ldots, z_B \).

2.1.2 Resampling algorithm with held-out set

Suppose we have a held-out set of \( m \) labeled points, denoted \( \tilde{D} := \{(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m)\} \). Also consider an array \( M \) of size \( t \times m \), where the \( i \)th row of the array, denoted \( r_i \), is given by the predicted labels of the \( i \)th classifier on the held-out samples,

\[
r_i := [Q_i(\tilde{X}_1), \ldots, Q_i(\tilde{X}_m)],
\]

and

\[
M := \begin{bmatrix}
-r_1 & \\
\vdots & \\
-r_t
\end{bmatrix}.
\]

The estimated error rate is easily computed as a function of this matrix, say \( \hat{\text{Err}}_t = \hat{\text{Err}}_t(M) \). In detail, each column of \( M \) corresponds to a test point \( \tilde{X}_l \), and within the \( l \)th column, the plurality vote of the labels \( Q_1(\tilde{X}_l), \ldots, Q_t(\tilde{X}_l) \) is either correct or incorrect (i.e. matching \( \tilde{Y}_l \) or not). Hence, \( \hat{\text{Err}}_t(M) \) is simply the proportion of columns for which plurality vote is incorrect. (This step implicitly estimates \( \varphi(\bar{Q}) \).) Finally, since there is a one-to-one correspondence between the rows \( r_i \) and the classifiers \( Q_i \), the proposed method of bootstrapping \( \hat{\text{Err}}_t \) by resampling the \( Q_i(\cdot) \) corresponds to resampling the rows \( r_i \), as summarized below.

For \( l = 1, \ldots, B \):

- Draw a \( t \times m \) array \( M^* \) whose rows \( \{r_1^*, \ldots, r_t^*\} \) are sampled randomly with replacement from \( \{r_1, \ldots, r_t\} \).
- Compute \( z_l := \hat{\text{Err}}(M^*) \).

Return: empirical distribution of \( z_1, \ldots, z_B \).

Remarks. To comment on the algorithm, first observe that it only relies on the matrix of predicted labels, \( M \). Consequently, the algorithm does not require re-evaluating the classifier functions \( Q_i \), and is hence insensitive to the complexity of their internal mechanisms. Secondly, the algorithm does not require any access to the original training data — only the held-out set. Thirdly, the algorithm is well-suited to the situation where the classifiers \( Q_i \) are computed and stored in parallel. In this case, each of the rows \( r_i \) can be computed and stored on separate nodes (corresponding to the node where \( Q_i \) was trained).
2.2 Saving on computation with extrapolation

The computational cost of our bootstrap method can be reduced through the technique of extrapolation [Sid03, BZ13]. To explain the technique, suppose \( \psi(\mathcal{L}(\text{Err}_t)) \) is a quantity of interest associated with the law of \( \text{Err}_t \), such as the standard deviation or a quantile. In a nutshell, extrapolation involves computing an inexpensive estimate of \( \psi(\mathcal{L}(\text{Err}_t)) \) by bootstrapping a relatively small ensemble of size \( t_0 \), and then modifying the estimate so that it approximates \( \psi(\mathcal{L}(\text{Err}_t)) \) for a large ensemble with \( t \gg t_0 \). For simplicity, we will focus on the estimation of \( \psi(\mathcal{L}(\text{Err}_t)) = \sqrt{\text{var}(\text{Err}_t)} \), but the extension to other quantities, such as quantiles, proceeds along similar lines.

To explain the idea in greater detail, we will make use of the fact that under certain assumptions, \( \text{Err}_t \) satisfies a central limit theorem of the form
\[
\mathcal{L}(\sqrt{t}(\text{Err}_t - \text{err})) \xrightarrow{w} N(0, \sigma^2) \quad \text{as} \quad t \to \infty,
\]
for some constants \( \text{err} \) and \( \sigma \) (cf. Theorem 1 in Section 3). Likewise, we are led to consider the following approximation for large \( t \),
\[
\sqrt{\text{var}(\text{Err}_t)} \approx \frac{\sigma}{\sqrt{t}}.
\]
In order to make use of the \( 1/\sqrt{t} \) scaling property of the standard deviation, let \( \tilde{s}_{t_0} \) denote an estimate of \( \sqrt{\text{var}(\text{Err}_{t_0})} \) for a relatively small ensemble of size \( t_0 \). Then, by the following rescaling, we obtain an inexpensive estimate \( \tilde{s}_t \) for a larger ensemble size \( t \gg t_0 \) given by
\[
\tilde{s}_t := \frac{\sqrt{t_0}}{\sqrt{t}} \cdot \tilde{s}_{t_0} \approx \sqrt{\text{var}(\text{Err}_t)}.
\]

3 Main Result

Our main theoretical objective is to prove that the law of \( \sqrt{t}(\text{Err}^*_t - \text{Err}_t) \), conditionally on \((Q_1, \ldots, Q_t)\), yields a consistent distributional approximation of \( \sqrt{t}(\text{Err}_t - \text{err}) \), where \( \text{err} \) is the limiting value of \( \text{Err}_t \). Toward this goal, we will rely on two simplifications. Both are customary in analyses of bootstrap methods and random forests. First, we will ignore extraneous sources of “Monte-Carlo” error in our bootstrap resampling algorithm. These sources of error are the finite number of \( B \) bootstrap replicates, and the finite number of \( m \) held-out points used to estimate \( \hat{\psi} \) (as discussed in Section 2.1). Secondly, we will analyze a simplified version of the Random Forests algorithm, which we will refer to as a first order model. Such an approach has been useful in gaining theoretical insights into the behavior of Random Forests in a variety of previous works [BDL08, Bia12, AG14, LJ06, Gen12, Sco14]. In our context, the value of this simplification is that it abstracts away the complex mechanisms occurring within the base classifiers, and clarifies the relationship between the size of the ensemble and quality of the bootstrap approximation.

3.1 A first order model of Random Forests

Any random classifier function \( Q_1 : \mathcal{X} \to \{e_0, \ldots, e_{k-1}\} \) may be viewed as a discrete-valued stochastic process on \( \mathcal{X} \). From this point of view, a first order model, denoted \( T_1 : \mathcal{X} \to \{e_0, \ldots, e_{k-1}\} \), is another process with the same marginal distributions as \( Q_1 \), but possibly different correlation structure. Equality of marginals means that we have the equality in law
\[
Q_1(x) \overset{\mathcal{L}}{=} T_1(x) \quad \text{for all} \quad x \in \mathcal{X}.
\]
Note that since \( Q_1(x) \) takes values in the set of binary vectors \( \{e_0, \ldots, e_{k-1}\} \), it follows that equality of marginal laws is equivalent to
\[
\mathbb{E}[Q_1(x)] = \mathbb{E}[T_1(x)] \quad \text{for all} \quad x \in \mathcal{X}.
\]
The way that \( Q_1 \) and \( T_1 \) may differ is in terms of higher moments. For instance, if \( x \) and \( x' \) are distinct points in \( \mathcal{X} \), then it may occur that \( \mathbb{E}[Q_1(x)Q_1(x')] \neq \mathbb{E}[T_1(x)T_1(x')] \).

### 3.1.1 Constructing a first order model

Interestingly, it is possible to start with an arbitrary random classifier \( Q_1 : \mathcal{X} \to \{e_0, \ldots, e_{k-1}\} \), and explicitly construct a corresponding first order model \( T_1 \) in a relatively simple way. To do this, first consider a deterministic function \( \vartheta : \mathcal{X} \to \Delta \) defined by

\[
\vartheta(x) := \mathbb{E}[Q_1(x)].
\]

Next, we make special use of the fact that \( \vartheta \) takes values in the simplex \( \Delta \). For any fixed \( \theta \in \Delta \), there is an associated partition of the unit interval into sub-intervals

\[
I_1(\theta) \cup \cdots \cup I_{k-1}(\theta) \cup I_0(\theta) = [0, 1],
\]

such that the width of interval \( I_j(\theta) \) is equal to \( \theta_j \). Namely, we put \( I_1(\theta) := [0, \theta_1] \), and for \( j = 2, \ldots, k-1 \),

\[
I_j(\theta) := \left( (\theta_1 + \cdots + \theta_{j-1}), (\theta_1 + \cdots + \theta_j) \right).
\]

Lastly, for \( j = 0 \), we put

\[
I_0(\theta) := \left( \sum_{j=1}^{k-1} \theta_j, 1 \right],
\]

which can be empty when \( \theta_1 + \cdots + \theta_{k-1} = 1 \). Now, if we let \( U_1 \sim \text{Uniform}[0, 1] \), then we define a point \( T_1(x) \in \{e_0, \ldots, e_{k-1}\} \) to have its \( j \)th coordinate equal to the following indicator variable

\[
[T_1(x)]_j := 1\{U_1 \in I_j(\vartheta(x))\},
\]

where \( j = 1, \ldots, k-1 \). It is simple to check that the first order matching condition \( [26] \) holds, and so \( T_1 \) is indeed a first order model of \( Q_1 \).

Having defined \( T_1 \) in terms of a single random variable \( U_1 \sim \text{Uniform}[0, 1] \), we obtain a corresponding “first order ensemble” of i.i.d. functions \( T_1, \ldots, T_t \) via an i.i.d. sample of uniform variables \( U_1, \ldots, U_t \). In other words, the \( j \)th coordinate of \( T_1 \) is given by \( [T_1(x)]_j = 1\{U_i \in I_j(\vartheta(x))\} \). Likewise, we define the analogue of \( Q \) as

\[
\bar{T}() = \frac{1}{t} \sum_{i=1}^t T_i().
\]

As a point of clarification, it is important to note that the uniform variables \( U_1, \ldots, U_t \) are only used for the purpose of analyzing the first order ensemble. These variables play no role in the implementation of our bootstrap method (which can be applied to an arbitrary random forest).

### 3.2 Bootstrap consistency

We now state our main result, which asserts that the bootstrap “works” for \( \text{Err}_t \) under a first order model of Random Forests. In order to give precise meaning to our statement of bootstrap consistency, we first review the notion of conditional weak convergence.

**Conditional weak convergence.** Let \( \lambda_0 \) be a fixed probability law on \( \mathbb{R} \), and let \( \{\lambda(T_t)\}_{t \geq 1} \) be a sequence of laws on \( \mathbb{R} \) that depend on the random functions \( T_t := \{T_1, \ldots, T_t\} \). Also, let \( d_{BL} \) denote the bounded Lipschitz metric for weak convergence. Then, as \( t \to \infty \), we say that \( \lambda(T_t) - \overset{w}{\longrightarrow} \lambda_0 \) in \( \mathbb{P}_{T_t} \)-probability if the sequence of random variables \( d_{BL}(\lambda(T_t), \lambda_0) \) converges to 0 in \( \mathbb{P}_{T_t} \)-probability.

We refer to Section 23.2.1 of the book [vdV00] for further background on the notion of conditional weak convergence and the bounded Lipschitz metric.
Remark. For the remainder of the paper, we use $\mu_j$ to denote the conditional law of $\theta(X)$ given $Y = e_j$. That is, $\mu_j$ is a law on the simplex $\Delta \subset \mathbb{R}^{k-1}$ defined earlier. Note that the properties of $\mu_j$ are not affected by the assumption of a first order model, since $\theta(x) = \mathbb{E}[T_1(x)] = \mathbb{E}[Q_1(x)]$ for all $x \in \mathcal{X}$. We will assume that the measures $\mu_j$ satisfy the following regularity condition.

Assumption 1. For each $j = 0, \ldots, k - 1$, the distribution $\mu_j$ has a density $f_j$ with respect to Lebesgue measure on $\Delta$. Also, for each $j$, the density $f_j$ is $C^1$ on $\Delta$ and $\|\nabla f_j\|_2$ is bounded on $\Delta$.

Theorem 1 (Bootstrap consistency for $\text{Err}_t$). Suppose that the first order model $Q_i(\cdot) = T_i(\cdot)$ holds for all $i \geq 1$, and that Assumption 1 holds. Then, there are constants $\text{err}$ and $\sigma$ such that as $t \to \infty$,

$$\mathcal{L}(\sqrt{t}(\text{Err}_t - \text{err})) \xrightarrow{w} N(0, \sigma^2),$$

and furthermore,

$$\mathcal{L}(\sqrt{t}(\text{Err}_t^* - \text{Err}_t) \mid T_t) \xrightarrow{w} N(0, \sigma^2) \text{ in } \mathbb{P}_{T_t}$-probability.$$

Remarks. In the following three subsections, 3.3, 3.4, and 3.5, we will assemble the main conceptual ingredients used in our proof of Theorem 1. (The proof is concluded in Section 3.5.) Technical arguments will be given in Appendices A, B, and C.

3.3 Lifting the standard empirical process

Here we explain how the stochastic process $\bar{T} : \mathcal{X} \to \Delta$ can be obtained from a linear transformation that “lifts” the standard one-dimensional empirical process $\mathbb{F}_t : [0, 1] \to [0, 1]$ to a $(k-1)$-dimensional version. If the random variables $U_1, \ldots, U_k$ are i.i.d. Uniform$[0,1]$, then for any $u \in [0, 1]$, we put $\mathbb{F}_t(u) := \frac{1}{t} \sum_{i=1}^{t} 1\{U_i \leq u\}$. The relationship between $\bar{T}$ and $\mathbb{F}_t$ is unraveled by noting that for any $j = 1, \ldots, k - 1$, the definition of the interval $I_j$ gives

$$[\bar{T}(x)]_j = \frac{1}{t} \sum_{i=1}^{t} [T_i(x)]_j$$

$$= \frac{1}{t} \sum_{i=1}^{t} 1\{U_i \in I_j(\theta(x))\},$$

$$= \mathbb{F}_t(\theta_1(x) + \cdots + \theta_j(x)) - \mathbb{F}_t(\theta_1(x) + \cdots + \theta_{j-1}(x)).$$

Based on the previous line, it is natural to consider a transformation, denoted $L$, that lifts a generic function $h : \mathbb{R} \to \mathbb{R}$ to a new function $L(h) : \mathbb{R}^{k-1} \to \mathbb{R}^{k-1}$ whose $j$th coordinate is given by

$$[(Lh)(\theta)]_j := h \circ \xi_j(\theta) - h \circ \xi_{j-1}(\theta),$$

where the $\xi_j$ are the coordinate-sum functions for $j = 1, \ldots, k - 1$,

$$\xi_j(\theta) := \theta_1 + \cdots + \theta_j,$$

and $\xi_0(\theta)$ is identically 0. In particular, the definition of $L$ shows that $\bar{T}(x)$ can be expressed by lifting $\mathbb{F}_t$ and evaluating at $\theta(x)$,

$$\bar{T}(x) = L(\mathbb{F}_t)(\theta(x)).$$

Also, recall that under the first order model,

$$\text{Err}_{t,j} = \mathbb{P}(\bar{T}(X) \in S_j \mid Y = e_j).$$
By considering our representation for $\bar{T}$ in terms of $L$, and the fact that $\mu_j$ is the conditional law of $\vartheta(X)$ given $Y = e_j$, this implies that $\text{Err}_{t,j}$ can be expressed as a functional of $\mathbb{F}_t$ through the formula
\[
\text{Err}_{t,j} = \mu_j\left([L(\mathbb{F}_t)]^{-1}(S_j)\right). \tag{40}
\]
The importance of this formula is that it allows us to analyze bootstrap consistency via the functional delta method, which will be done in Section 3.5.

### 3.3.1 Properties of the lifting map

The transformation $L$ possesses several properties that will be useful in our analysis. Below, let $f, g,$ and $F$ be generic scalar-valued functions, and let $a \in \mathbb{R}$.

- **Linearity.**
  \[ L(ag + h) = aL(g) + L(h). \]

- **Composition.**
  \[ L(g \circ h) = L(g) \circ L(h). \]

- **Identities.**
  \[ L(id_{[0,1]}) = id_\Delta. \]

- **Simplices.**
  If $F$ is non-decreasing and maps $[0, 1]$ to $[0, 1]$, then $L(F)$ maps $\Delta$ to $\Delta$.

- **Inverses.**
  If $F$ is invertible, then $L(F)$ is invertible and $[L(F)]^{-1} = L(F^{-1})$.

The fact that $L$ respects composition of functions is the only property that takes some care to verify. From the “Composition” and “Identities” properties, it follows that if $F$ is an invertible function on $[0, 1]$, then $L(F)$ is invertible on $\Delta$ with inverse equal to $L(F^{-1})$, since
\[
id_\Delta = L(id_{[0,1]}) = L(F \circ F^{-1}) = L(F) \circ L(F^{-1}),
\]
and similarly for composition in the reverse order.

### 3.4 Hadamard differentiability

Let $\mathcal{F}[0, 1]$ denote the space of cumulative distribution functions on $[0, 1]$. In light of our earlier formula (40) for $\text{Err}_{t,j}$ in terms of $\mathbb{F}_t$, it is natural to define a functional $\phi_j : \mathcal{F}[0, 1] \to \mathbb{R}$ according to
\[
\phi_j(G) := \mu_j\left([L(G)]^{-1}(S_j)\right), \tag{41}
\]
where $j = 0, \ldots, k - 1$. In particular, the formula (40) gives
\[
\text{Err}_{t,j} = \phi_j(\mathbb{F}_t), \tag{42}
\]
and if we define $\phi : \mathcal{F}[0, 1] \to \mathbb{R}$ by
\[
\phi := \sum_{j=1}^{k-1} \pi_j \phi_j, \tag{43}
\]
then the total error rate $\operatorname{Err}_t$ satisfies

$$\operatorname{Err}_t = \phi(F_t).$$

(44)

Since the random distribution function $F_t$ approaches $\text{id}_{[0,1]}$ as $t \to \infty$, the asymptotic behavior of $\operatorname{Err}_t$ is closely linked to the smoothness $\phi$ in a neighborhood of the “point” $\text{id}_{[0,1]}$. Specifically, we will use a standard notion of local smoothness known as Hadamard differentiability, as defined below. (Further background on this notion may be found in the book [vdVW96].)

**Definition 2** (Hadamard differentiability). Let $\mathbb{D}$ be a normed space. A map $\psi : \mathcal{F}[0,1] \to \mathbb{D}$ is Hadamard differentiable at $G_0 \in \mathcal{F}[0,1]$ tangentially to $C[0,1]$, if there is a continuous linear map $\psi'_{G_0} : C[0,1] \to \mathbb{D}$ such that as $t \to \infty$,

$$\frac{\psi(G_0 + \varepsilon_t h_t) - \psi(G_0)}{\varepsilon_t} \to \psi'_{G_0}(h),$$

(45)

for all converging sequences of positive numbers $\varepsilon_t \to 0$ and functions $h_t \to h \in C[0,1]$ such that $G_0 + \varepsilon_t h_t \in \mathcal{F}[0,1]$ for every $t \geq 1$. In particular, the linear map $\psi'_{G_0}$ is referred to as the Hadamard derivative of $\psi$ at $G_0$.

The following theorem is the core technical result of the paper. It provides conditions under which the functionals $\phi_j$ are Hadamard differentiable at $G_0 = \text{id}_{[0,1]}$, tangentially to $C[0,1]$. But more importantly, the theorem implies that some differentiability property holds for the linear combination $\phi = \sum_{j=1}^{k-1} \pi_j \phi_j$ (which will be used in the next subsection). An interesting aspect of the theorem is that while each $\phi_j$ is defined in terms of the particular set $S_j$ and the particular measure $\mu_j$, the Hadamard differentiability of $\phi_j$ is only mildly dependent on the structure of $S_j$ and $\mu_j$. The only specific properties we will use are that the sets $S_j$ are convex, and that the measures $\mu_j$ have sufficiently regular densities with respect to Lebesgue measure. In fact, the assumption of convexity could be weakened further, but at the expense of a more intricate statement.

**Theorem 2** (Hadamard differentiability of $\phi_j$). Let $S \subset \Delta$ be a convex set, and let $\mu$ be a distribution on $\Delta$ that has a density $f : \Delta \to [0,\infty)$ with respect to Lebesgue measure. Assume that $f$ is $C^1$ on $\Delta^\circ$ and $\|\nabla f\|_2$ is bounded on $\Delta^\circ$. Also, let $\psi : \mathcal{F}[0,1] \to \mathbb{R}$ be a functional defined for any $G \in \mathcal{F}[0,1]$ according to

$$\psi(G) := \mu([L(G)]^{-1}(S)).$$

(46)

Then, $\psi$ is Hadamard differentiable at $\text{id} \in \mathcal{F}[0,1]$ tangentially to $C[0,1]$. Furthermore, the Hadamard derivative $\psi'_{\text{id}} : C[0,1] \to \mathbb{R}$ is given by

$$\psi'_{\text{id}}(h) = -\int_{\partial S} \langle L(h)(\theta), n(\theta) \rangle f(\theta) d\sigma(\theta),$$

(47)

where $n(\theta)$ is the outward normal to $\partial S$ at a point $\theta$, and $d\sigma$ is the surface measure on $\partial S$ induced by Lebesgue measure.

**Remark.** A proof is given in Section 3.6.

3.5 Functional delta method

Using the representation $\operatorname{Err}_t = \phi(F_t)$ from line (44) and the Hadamard differentiability of $\phi$ from Theorem 2, the functional delta method provides a way to convert “bootstrap consistency for $F_t$” into “bootstrap consistency for $\operatorname{Err}_t$”. To clarify the meaning of bootstrap consistency for $F_t$, let

$$F_t^*(u) := \frac{1}{t} \sum_{i=1}^{t} 1\{U_i^* \leq u\},$$

(48)
where $U^*_t, \ldots, U^*_t$ are drawn with replacement from the empirical measure on $U_t := \{U_1, \ldots, U_t\}$, and $U_1, \ldots, U_t$ are i.i.d. Uniform[0,1]. In this notation, the bootstrap “works” for $F_t$ in the sense that $L(\sqrt{t}(F_t - \text{id}) \mid U_t)$ converges to the same weak limit as $L(\sqrt{t}(F_t - \text{id}))$ (in $\mathbb{P}_{U_t}$-probability). The following well-known result shows that the bootstrap consistency remains valid when $F_t$ and $F^*_t$ are precomposed with a Hadamard differentiable functional.

**Lemma 1** (Functional delta method [vdV00], Sec. 23.2.1). Let $\psi : \mathcal{F}[0,1] \to \mathbb{R}$ be Hadamard differentiable at $\text{id} \in \mathcal{F}[0,1]$, tangentially to $C[0,1]$. Also, let $B$ be a standard Brownian bridge on $[0,1]$. Then, as $t \to \infty$,

$$L(\sqrt{t}(\psi(F_t) - \psi(\text{id}))) \xrightarrow{w} L(\psi'_\text{id}(B)),$$

and

$$L(\sqrt{t}(\psi(F^*_t) - \psi(F_t)) \mid U_t) \xrightarrow{w} L(\psi'_\text{id}(B)), \text{ in } \mathbb{P}_{U_t}-\text{probability.}$$

**Concluding the proof of Theorem 1.** With this lemma in hand, Theorem 1 follows quickly from Theorem 2. Specifically, if we put $\text{err} = \phi(\text{id})$, then the definitions of $\text{Err}^*_t$ and $T_1, \ldots, T_t$ give

$$L(\sqrt{t}(\text{Err}_t - \text{err})) = L(\sqrt{t}(\phi(F_t) - \phi(\text{id}))),$$

$$L(\sqrt{t}(\text{Err}^*_t - \text{Err}_t) \mid T_t) = L(\sqrt{t}(\phi(F^*_t) - \phi(F_t)) \mid U_t).$$

From these relations, we see that Theorem 2 and Lemma 1 imply that the left sides of lines (51) and (52) tend to the same weak limit, namely $\phi'_\text{id}(B)$. The only remaining detail to be proven in Theorem 1 is that $\phi'_\text{id}(B)$ has a centered Gaussian distribution. This follows from the fact that $B$ is a centered Gaussian process in $C[0,1]$, and the fact that $\phi'_\text{id}$ is a continuous linear functional on $C[0,1]$ [vdVV96, Lemma 3.9.8].

### 3.6 Proof of Theorem 2

Here we give a proof of Theorem 2 that focuses on the main ideas and delegates the technical pieces to the appendices. Consider a sequence of positive numbers $\varepsilon_t \to 0$ and functions $h_t \to h \in C[0,1]$ such that $\text{id} + \varepsilon_t h_t \in \mathcal{F}[0,1]$ for every $t \geq 1$. Define the distribution function $F_t : [0,1] \to [0,1]$ by

$$F_t := \text{id} + \varepsilon_t h_t,$$

and define its lifted version $V_t : \Delta \to \Delta$ by

$$V_t := L(F_t).$$

The fact that $V_t$ takes values in $\Delta$ follows from the properties of $L$ listed in Section 3.3.1. Our aim is to evaluate the limit of the following difference as $t \to \infty$,

$$\frac{1}{\varepsilon_t} (\psi(F_t) - \psi(\text{id})) = \frac{1}{\varepsilon_t} \left( \mu(V_t^{-1}(S)) - \mu(S) \right).$$

Here, we have used the fact that $\psi(\text{id}) = \mu(S)$, which follows from $L(\text{id}_{[0,1]}) = \text{id}_\Delta$. Since $V_t$ approaches $\text{id}_\Delta$ as $t \to \infty$, we may view the preimage $V_t^{-1}(S)$ as a perturbed version of the set $S$. From this perspective, it is natural to interpret the right side of line (55) through the lens of the first variation formula, introduced in Section 1.3.

In its classical form, the first variation formula deals with smooth maps on smooth manifolds. However, since the map $V_t$ need not be smooth, our proof proceeds by constructing a smoothed version of $V_t$. In order to smooth $V_t$, it is enough to smooth the univariate function $F_t$ and apply the
linear transformation $L$. Specifically, the smoothing will be done using the linear Bernstein smoothing operator, denoted $B_s$, where $s \geq 1$ is an integer-valued smoothing parameter \cite{Lor12, DL93}. For any function $G : [0, 1] \to \mathbb{R}$, the operator returns a new function $B_s(G) : [0, 1] \to \mathbb{R}$ defined by

$$B_s(G)(u) := \sum_{l=0}^{s} G(l/s) \cdot b_l(u; s),$$

(56)

where $b_l(u; s) := (\binom{s}{l}) u^l (1-u)^{s-l}$ is the $l$th Bernstein basis polynomial with $u \in [0, 1]$, and $l$ ranges over $\{0, 1, \ldots, s\}$. In particular, when $G$ is a cumulative distribution function, $B_s(G)$ is also a cumulative distribution function, and when $G$ is continuous, $B_s(G) \to G$ uniformly as $s \to \infty$. All of the specific properties of $B_s$ we will use are summarized in Lemma 2 of Appendix C.

The smoothed version of $F_t$ is denoted

$$F_{t,s} := B_s(F_t).$$

(57)

Likewise, the smoothed version $V_{t,s} : \Delta \to \Delta$ is given by

$$V_{t,s} := L(F_{t,s}) = [L \circ B_s](F_t).$$

(58)

The fact that $V_{t,s}$ takes values in $\Delta$ follows from $F_{t,s}$ being increasing (by Lemma 2 of Appendix C), as well as the “simplices” property of $L$, stated in Section 3.3.1.

The remainder of the proof involves two essential parts. First, we prove a special version of the first variation formula for the smoothed maps $V_{t,s}$. Second, we show that this smoothing leads to negligible approximation error. To proceed, define the remainder $R_{t,s}$ according to the following equation

$$\frac{\mu(V_{t}^{-1}(S)) - \mu(S)}{\varepsilon_t} = \frac{\mu(V_{t,s}^{-1}(S)) - \mu(S)}{\varepsilon_t} + R_{t,s}.$$  

(59)

Due to the smoothness of $V_{t,s}$, the difference on the right may be represented with a change of variable formula

$$\frac{\mu(V_{t,s}^{-1}(S)) - \mu(S)}{\varepsilon_t} = \int_{S^o} f(V_{t,s}^{-1}(\theta)) |\det J(V_{t,s}^{-1}(\theta))| - f(\theta) \, d\theta,$$

(60)

which is justified by Lemmas 5 and 7 in Appendix C. (These lemmas also prove invertibility of $V_{t,s}$.) From this integral formula, Proposition 1 in Appendix A provides the following limit

$$\lim_{s \to \infty} \lim_{t \to \infty} \int_{S^o} f(V_{t,s}^{-1}(\theta)) |\det J(V_{t,s}^{-1}(\theta))| - f(\theta) \, d\theta = - \int_{\partial S} \langle L(h)(\theta), \mathbf{n}(\theta) \rangle f(\theta) \, d\sigma(\theta).$$

Next, Proposition 2 in Appendix B shows that replacing $V_t^{-1}$ with its smoothed version $V_{t,s}^{-1}$ leads to negligible approximation error, i.e.

$$\lim_{s \to \infty} \limsup_{t \to \infty} |R_{t,s}| = 0.$$  

Consequently, by first letting $t \to \infty$, and then letting $s \to \infty$ in line (59), it follows that

$$\lim_{t \to \infty} \frac{\mu(V_{t}^{-1}(S)) - \mu(S)}{\varepsilon_t} = - \int_{\partial S} \langle L(h)(\theta), \mathbf{n}(\theta) \rangle f(\theta) \, d\sigma(\theta).$$

Lastly, it is simple to check that the right side is a continuous linear functional of $h$ (as required by the definition of Hadamard differentiability). \hfill \square
Acknowledgements. MEL is very grateful to Philip Kegelmeyer for many research discussions, and for stimulating interest in this problem during a practicum at Sandia National Laboratory. Peter Bickel is also thanked for many research discussions. Third, an anonymous reviewer who read an earlier version of this work is thanked for insightful suggestions.

4 Appendix

A A first variation formula

Proposition 1. Assume the conditions of Theorem 4. Then, in the notation of Section 3 the following limit holds,

$$\lim_{s \to \infty} \lim_{t \to \infty} \int_{S^o} \frac{f(V_{t,s}^{-1}(\theta)) - f(\theta)}{\varepsilon_t} d\theta = - \int_{\partial S} \left( L(h)(\theta), \nu(\theta) \right) f(\theta) d\sigma(\theta).$$

(61)

Proof. Using an expansion for $|\det J(V_{t,s}^{-1}(\theta))|$ given in Lemma 6 of Appendix C, as well as the boundedness of $f$, the integral on the left side of line (61) may be written as

$$\int_{S^o} \frac{f(V_{t,s}^{-1}(\theta)) - f(\theta)}{\varepsilon_t} d\theta - \int_{S^o} \text{div}(V_{t,s}(\theta)) f(V_{t,s}^{-1}(\theta)) d\theta + O(\varepsilon_t K_s),$$

(62)

where $K_s < \infty$ is a sequence of numbers not depending on $t$. We now evaluate the limits of the two integrals separately. For any $\theta \in \Delta$, let

$$W_{t,s}(\theta) := \frac{1}{\varepsilon_t} (V_{t,s}(\theta) - \theta),$$

(63)

and

$$\tilde{W}_{t,s}(\theta) := \frac{1}{\varepsilon_t} (V_{t,s}^{-1}(\theta) - \theta).$$

(64)

Due to the multivariate mean value theorem, for each fixed $\theta \in \Delta^o$, there are points $\zeta_{t,s}(\theta) \in \Delta^o$ on the line segment between $V_{t,s}^{-1}(\theta)$ and $\theta$ such that

$$\frac{1}{\varepsilon_t} (f(V_{t,s}^{-1}(\theta)) - f(\theta)) = \left\langle \nabla f(\zeta_{t,s}(\theta)), \tilde{W}_{t,s}(\theta) \right\rangle.$$

(65)

Also, these points may be taken to satisfy $\zeta_{t,s}(\theta) \to \theta$ as $t \to \infty$, since $\lim_{t \to \infty} V_{t,s}^{-1}(\theta) = \theta$ for every fixed $\theta \in \Delta^o$ and fixed $s \geq 1$, which follows from Lemma 10 of Appendix C. By the Cauchy-Schwarz inequality, the inner product above is dominated by a number depending only on $s$, since $\|\nabla f\|_2$ is bounded on $\Delta^o$ and $\sup_{t \geq 1} \sup_{\theta \in \Delta} \|\tilde{W}_{t,s}(\theta)\|_2$ is finite by Lemma 10. Furthermore, Lemma 10 gives the convergence $\lim_{t \to \infty} \tilde{W}_{t,s}(\theta) \to -W_s(\theta)$ for every $\theta \in \Delta$, where

$$W_s := L(B_s(h)).$$

(66)

Hence, the continuity of $\nabla f$ and the dominated convergence theorem lead to

$$\lim_{t \to \infty} \int_{S^o} \frac{f(V_{t,s}^{-1}(\theta)) - f(\theta)}{\varepsilon_t} d\theta = - \int_{S^o} \left\langle \nabla f(\theta), W_s(\theta) \right\rangle d\theta.$$

(67)

To handle the second integral in line (62), Lemma 4 in Appendix C ensures that the divergence

$$\text{div}(W_{t,s})(\theta) := \sum_{j=1}^{k} \frac{\partial}{\partial \theta_j} [W_{t,s}(\theta)]_j$$

(68)
converges uniformly to \( \text{div}(W_s)(\theta) \) on \( \Delta \) as \( t \to \infty \), and so the dominated convergence theorem gives
\[
\lim_{t \to \infty} \int_{S^c} \text{div}\left(W_{t,s}(\theta)\right) f(V_{t,s}^{-1}(\theta)) d\theta = \int_{S^c} \text{div}\left(W_s(\theta)\right) f(\theta) d\theta.
\]
The last two limits may be combined using the basic differential identity
\[
\text{div}\left(f(\theta)W_s(\theta)\right) = \langle \nabla f(\theta), W_s(\theta) \rangle + \text{div}(W_s(\theta)) f(\theta),
\]
which shows that line (62) tends to \(- \int_{S^c} \text{div}(f(\theta)W_s(\theta)) d\theta \) as \( t \to \infty \) with \( s \) held fixed. In turn, Stokes’ theorem may be applied to this divergence integral since the vector field \( f(\theta)W_s(\theta) = L(B_s(h))(\theta) \) is smooth, and the domain \( S \) is convex. (An applicable version of Stokes’ theorem that holds for convex domains may be found in Theorem 2.1 of the book [Chi00].) When referencing that result, note that convex domains have Lipschitz boundaries [KL11, p.21].) Hence, for every \( s \geq 1 \),
\[
- \int_{S^c} \text{div}(f(\theta)W_s(\theta)) d\theta = - \int_{\partial S} \langle W_s(\theta), n(\theta) \rangle f(\theta) d\sigma(\theta).
\]
(69)

Finally, since \( h \in C[0,1] \), the uniform approximation property of Bernstein polynomials for continuous functions from Lemma 2 in Appendix C implies that \( W_s(\theta) = L(B_s(h))(\theta) \to L(h)(\theta) \) uniformly on \( \Delta \) as \( s \to \infty \), and so the dominated convergence theorem implies that the right side of line (69) tends to \(- \int_{\partial S} \langle L(h)(\theta), n(\theta) \rangle f(\theta) d\sigma(\theta) \), which completes the proof.

\[\square\]

\section*{B Smoothing error is negligible}

\textbf{Proposition 2.} Let \( R_{t,s} \) be as defined in line (59), and suppose the conditions of Theorem 2 hold. Then,
\[
\lim_{s \to \infty} \limsup_{t \to \infty} |R_{t,s}| = 0.
\]
(70)

\textit{Proof.} It is simple to check that \( R_{t,s} \) may be written as
\[
R_{t,s} = \frac{1}{e_t} \mu\left(V_t^{-1}(S) \setminus V_{t,s}^{-1}(S)\right) - \frac{1}{e_t} \mu\left(V_{t,s}^{-1}(S) \setminus V_t^{-1}(S)\right).
\]
(71)

We will argue that both terms on the right side are small. Note that the set \( V_t^{-1}(S) \setminus V_{t,s}^{-1}(S) \) consists of the points \( \theta \in \mathbb{R}^{k-1} \) such that \( V_t(\theta) \in S \) and \( V_{t,s}(\theta) \notin S \). Hence, if the Euclidean distance between the points \( V_t(\theta) \) and \( V_{t,s}(\theta) \) is written as \( d_{t,s}(\theta) \), then in particular, both of the points \( V_{t,s}(\theta) \) and \( V_t(\theta) \) must be within a distance \( d_{t,s}(\theta) \) from the boundary \( \partial S \). In other words, both of the points \( V_t(\theta) \) and \( V_{t,s}(\theta) \) lie within the tubular neighborhood of \( \partial S \) of radius \( d_{t,s}(\theta) \). (The same reasoning applies to the other set \( V_{t,s}^{-1}(S) \setminus V_t^{-1}(S) \).

To express the reasoning above more formally, let \( T(\partial S; r) \subset \mathbb{R}^{k-1} \) denote the tubular neighborhood of \( \partial S \) of radius \( r \geq 0 \),
\[
T(\partial S; r) := \{ \theta \in \mathbb{R}^{k-1} : d_2(\theta, \partial S) \leq r \},
\]
where \( d_2(\theta, \partial S) = \inf\{||\theta - v||_2 : v \in \partial S\} \) is the Euclidean distance from a point \( \theta \) to the boundary \( \partial S \). Also, for any \( \theta \in \mathbb{R}^{k-1}, \) recall that
\[
d_{t,s}(\theta) = ||V_{t,s}(\theta) - V_t(\theta)||_2.
\]
(72)

Then,
\[
|R_{t,s}| \leq 2 \cdot \frac{1}{e_t} \mu\{ \theta \in \Delta \mid V_{t,s}(\theta) \in T(\partial S; d_{t,s}(\theta))\}.
\]
(73)
Furthermore, if $\Theta \in \Delta$ is a random vector distributed according to $\mu$, then this upper bound may be written as

$$|R_{t,s}| \leq 2 \cdot \frac{1}{\varepsilon_t} \mathbb{P}\left( V_{t,s}(\Theta) \in \mathcal{T}(\partial S; d_{t,s}(\Theta)) \right).$$  \tag{74}$$

To modify the last bound somewhat, let

$$\delta_{t,s} := \sup_{\theta \in \Delta} d_{t,s}(\theta),$$  \tag{75}$$

and put

$$r_{t,s} := \max\{\delta_{t,s}, \varepsilon_t / s\}.$$  \tag{76}$$

(This definition of $r_{t,s}$ will be used later on for handling the possibility that $\delta_{t,s}$ may be zero. Note that $\varepsilon_t / s$ is positive for all $s$ and $t$.) Clearly, the definition of $r_{t,s}$ gives

$$|R_{t,s}| \leq 2 \cdot \frac{1}{\varepsilon_t} \mathbb{P}(V_{t,s}(\Theta) \in \mathcal{T}(\partial S; r_{t,s})).$$  \tag{77}$$

Next, we obtain an upper bound on this probability in terms of volume. Lemma 7 in Appendix C shows that $V_{t,s}(\Theta)$ has a density, say $g_{t,s}$, with respect to $(k-1)$-dimensional Lebesgue measure. Also, Lemma 8 implies that the random vector $V_{t,s}(\Theta)$ lies in $\Delta^\circ$ with probability 1 for all $t \geq 1$ and $s \geq 1$. Therefore,

$$|R_{t,s}| \leq 2 \cdot \frac{1}{\varepsilon_t} \left( \sup_{\theta \in \Delta^\circ} g_{t,s}(\theta) \cdot \left( \text{vol}_{k-1} \mathcal{T}(\partial S; r_{t,s}) \right) \right),$$  \tag{78}$$

where $\text{vol}_{k-1}$ denotes Lebesgue measure.

To control the volume of the right hand side of line (78), it is convenient to consider the Hausdorff measure of $\partial S$. Specifically, it is a fact from geometric measure theory that the $(k-2)$-dimensional Hausdorff measure of $\partial S$, denoted $\mathcal{H}^{(k-2)}(\partial S)$, can be expressed as

$$\mathcal{H}^{(k-2)}(\partial S) = \lim_{r \downarrow 0} \frac{1}{r} \text{vol}_{k-1} \mathcal{T}(\partial S; r).$$  \tag{79}$$

(See the book [Fed96] Theorem 3.2.39 for additional details, and note that bounded convex sets have rectifiable boundaries when referencing that result.) In order to make use of this expression, we will rely on Lemma 9, which states that for every $s \geq 1$, there is a number $\kappa_s < \infty$ such that

$$\lim_{t \to \infty} \frac{1}{t} r_{t,s} = \kappa_s.$$

In particular, $r_{t,s}$ is a sequence of positive numbers with $r_{t,s} \to 0$ as $t \to \infty$, and so

$$\lim_{t \to \infty} \frac{1}{t} \text{vol}_{k-1} \mathcal{T}(\partial S; r_{t,s}) = \lim_{t \to \infty} \left( \frac{1}{t} r_{t,s} \right) \cdot \left( \frac{1}{r_{t,s}} \text{vol}_{k-1} \mathcal{T}(\partial S; r_{t,s}) \right)$$

$$= \kappa_s \cdot \mathcal{H}^{(k-2)}(\partial S).$$  \tag{80}$$

We now turn our attention to the factor $\sup_{\theta \in \Delta^\circ} g_{t,s}(\theta)$ in the bound (78). Lemma 7 ensures there is a constant $c_0 < \infty$, such that the following bound holds for every fixed $s \geq 1$,

$$\limsup_{t \to \infty} \sup_{\theta \in \Delta^s} g_{t,s}(\theta) \leq c_0.$$  \tag{81}$$

Combining lines (78), (80), and (81), we conclude that for every $s \geq 1$,

$$\limsup_{t \to \infty} |R_{t,s}| \leq 2c_0 \cdot \kappa_s \cdot \mathcal{H}^{(k-2)}(\partial S).$$  \tag{82}$$

Finally, the proof is completed using the fact that $\kappa_s \to 0$ as $s \to \infty$, which is shown in Lemma 9.
C Technical lemmas

Lemma 2 (Properties of Bernstein polynomials). The Bernstein smoothing operator $B_s$ satisfies the following properties.

(a) For any function $h \in C[0,1]$, the following limit holds

$$\sup_{u \in [0,1]} |B_s(h)(u) - h(u)| \to 0 \text{ as } s \to \infty. \quad (83)$$

(b) If $h : [0,1] \to \mathbb{R}$ is linear, then for every $s \geq 1$, $B_s(h) = h$, and in particular $B_s(id_{[0,1]}) = id_{[0,1]}$.

(c) If $h : [0,1] \to \mathbb{R}$ is non-decreasing and $h \neq 0$, then for every $s \geq 1$, $B_s(h)$ is strictly increasing in $(0,1)$.

Proof. We refer to the book [DL93] for further background on these properties. The first property is given in Theorem 2.3 of Chapter 1. The second property is proven after equation 1.7 of Chapter 1. Regarding the third property, it is shown in equation 2.2 of Chapter 10 that $B_s(h)$ satisfies

$$\frac{d}{ds}B_s(h)(u) = s \sum_{l=0}^{s-1} \delta h\left(\frac{l}{s}\right) \cdot b_l(u; s - 1) \quad (84)$$

where $\delta h(u) := h(u + 1/s) - h(u)$, and we recall that $b_l(u; s) := \binom{s}{l}u^l(1-u)^{s-l}$. If $h$ is non-decreasing and not identically 0, then at least one of the terms on the right side of line (84) is positive when $u \in (0,1)$, and all terms are non-negative. \hfill \square

Lemma 3. Let $D$ be the differentiation operator on univariate functions, and let $F_t = id + \varepsilon_t h_t \in \mathcal{F}[0,1]$ with $h_t \to h \in C[0,1]$ and $\varepsilon_t \to 0$ as $t \to \infty$, as in the proof of Theorem 2. Also let $F_{t,s} = B_s(F_t)$. Then, for any fixed $s \geq 1$, we have the identity,

$$\frac{1}{\varepsilon_t}([F_t]' - 1_{[0,1]}) = [D \circ B_s]\left(\frac{1}{\varepsilon_t}(F_t - id_{[0,1]}))\right), \quad (85)$$

and the uniform limit,

$$\lim_{t \to \infty} [D \circ B_s]\left(\frac{1}{\varepsilon_t}(F_t - id_{[0,1]}))\right) = [D \circ B_s](h) \text{ in } \ell^\infty_{[0,1]}. \quad (86)$$

Proof. The identity (85) is a direct consequence of $B_s(id_{[0,1]}) = id_{[0,1]}$ from Lemma 2. To prove the limit, first note that since the functions $b_l(\cdot; s)$ are polynomials on $[0,1]$, the supremum $C_s := \max_{0 \leq t \leq s} \sup_{x \in [0,1]} |b_l'(x; s)|$ is finite. Hence, for fixed $s$,

$$\sup_{x \in [0,1]} \left|D \circ B_s\left(\frac{1}{\varepsilon_t}(F_t - id_{[0,1]})) - h\right)(x)\right| = \sup_{x \in [0,1]} \left|\sum_{l=0}^{s} \frac{1}{\varepsilon_t}(F_t(l/s) - id_{[0,1]})l/s - h(l/s)\right| b_l'(x; s) \right| \leq C_s \sum_{l=0}^{s} \left|\frac{1}{\varepsilon_t}(F_t(l/s) - id_{[0,1]}l/s) - h(l/s)\right| \leq C_s \sum_{l=0}^{s} \left|h_t(l/s) - h(l/s)\right| = o(1) \text{ as } t \to \infty. \quad \square$$

Remark. In the statement of the following lemma, note that the derivatives $\frac{\partial}{\partial \theta_j}[W_{t,s}(\theta)]_j$ exist at the boundary points of $\Delta$, since the coordinates of $W_{t,s}(\theta)$ are polynomial functions of $\theta$.

Lemma 4. For any fixed $s \geq 1$, there is a number $K_s < \infty$ depending only on $s$ such that the inequality

$$\sup_{t \geq 1} \sup_{\theta \in \Delta} \left|\frac{\partial}{\partial \theta_j}[W_{t,s}(\theta)]_j\right| \leq K_s$$


holds for all \( j = 1, \ldots, k - 1 \). Furthermore, as \( t \to \infty \), we have the uniform limit

\[
\text{div}(W_{t,s}) \to \text{div}(W_s) \quad \text{in} \quad \ell^\infty(\Delta; \mathbb{R})
\]

(87)

Proof. From the definition of \( W_{t,s} \) in line (63), we have

\[
\frac{\partial [W_{t,s}(\theta)]_j}{\partial \theta_j} = \frac{1}{\varepsilon_t} \left( \frac{\partial [V_{t,s}(\theta)]_j}{\partial \theta_j} - 1 \right) = \frac{1}{\varepsilon_t} \left( F'_{t,s}(\theta_1 + \cdots + \theta_j) - 1 \right).
\]

The last expression is bounded in absolute value for every \( j = 1, \ldots, k - 1 \), and every \( t \geq 1 \), by

\[
K_s := \sup_t \sup_{x \in [0,1]} \frac{1}{\varepsilon_t} |F'_{t,s}(x) - 1|,
\]

which is finite by the uniform limit inLemma 3. The limit (87) also follows directly from Lemma 3.

Lemma 5. For any \( t \geq 1 \), \( s \geq 1 \), the function \( V_{t,s} : \Delta \to \Delta \) is bijective and \( C^1 \) on \( \Delta^o \). Also, the Jacobian matrix \( J(V_{t,s})(\theta) \) is non-singular for all \( \theta \in \Delta^o \).

Proof. The differentiability of \( V_{t,s} = L(F_{t,s}) \) is simple to verify. To show that \( V_{t,s} \) is bijective, it is enough to show that \( F_{t,s} \) is bijective (due to the “inverses” property of \( L \) stated in Section 3.3.1). The fact that \( F_{t,s} = B_s(F_t) \) is bijective can be verified by showing that \( \frac{d}{dx} B_s(G)(u) > 0 \) for all \( u \in (0,1) \) whenever \( G \) is a cumulative distribution function, which follows from Lemma 2.

Lastly, to see the Jacobian \( J(V_{t,s})(\theta) \) is non-singular, it is enough to check that \( V_{t,s}^{-1} \) is \( C^1 \), because we may then differentiate the identities \( V_{t,s} \circ V_{t,s}^{-1} = \text{id}_\Delta \) and \( V_{t,s}^{-1} \circ V_{t,s} = \text{id}_\Delta \) to obtain the inverse of \( J(V_{t,s}(\theta)) \) via the chain rule. Using \( V_{t,s}^{-1} = L(F_{t,s}^{-1}) \), it is simple to check that \( V_{t,s}^{-1} \) is \( C^1 \) on \( \Delta^o \) as long as \( F_{t,s}^{-1} \) is \( C^1 \) on the interval \((0,1)\), which follows from the strict monotonicity of \( F_{t,s} \) and the inverse function theorem.

Remark. The previous lemma shows that the inverse matrix \( J(V_{t,s}^{-1}(\theta)) \) exists for all \( \theta \in \Delta^o \). Next, we give a uniform expansion for the determinant of this matrix.

Lemma 6. For every \( s \geq 1 \) and \( t \geq 1 \), define the function \( r_{t,s} : \Delta^o \to \mathbb{R} \) according to

\[
|\det J(V_{t,s}^{-1}(\theta))| = 1 - \varepsilon_t \text{div} W_{t,s}(\theta) + r_{t,s}(\theta), \quad \theta \in \Delta^o.
\]

Then for any \( s \geq 1 \), there is a number \( K_s \in (0, \infty) \) not depending on \( t \), such that the following bound holds for all large \( t \),

\[
\sup_{\theta \in \Delta^o} |r_{t,s}(\theta)| \leq \varepsilon_t^2 \cdot \text{poly}(K_s, \varepsilon_t),
\]

where \( \text{poly}(\cdot, \cdot) \) is a polynomial function whose degree and coefficients do not depend on \( t \) or \( s \).

Proof. It is simple to check that the Jacobian matrix \( J(V_{t,s}(\theta)) \) is lower-triangular for all \( \theta \in \Delta^o \), and so the determinant of \( J(V_{t,s}(\theta)) \) is the product of the diagonal entries. Consequently, using the invertibility of \( J(V_{t,s}(\theta)) \) shown in Lemma 5, we obtain the following expression for all \( \theta \in \Delta^o \),

\[
\det (J(V_{t,s}^{-1}(\theta))) = 1/ \det (J(V_{t,s}(\theta))) = 1/ \prod_{j=1}^{k-1} \frac{\partial [V_{t,s}(\theta)]_j}{\partial \theta_j}.
\]

(88)

By the definition of \( W_{t,s}(\theta) \) in line (63), we have for each \( j = 1, \ldots, k - 1 \), and each \( \theta \in \Delta \),

\[
\frac{\partial [V_{t,s}(\theta)]_j}{\partial \theta_j} = 1 + \varepsilon_t \frac{\partial [W_{t,s}(\theta)]_j}{\partial \theta_j}.
\]

(89)
Due to Lemma 4, there is a number $K_s$ depending only on $s$, such that the bound

$$
\sup_{t \geq 1} \sup_{\theta \in \Delta} \sup_{j} \left| \frac{\partial[W_{t,s}(\theta)]}{\partial \theta_j} \right| \leq K_s
$$

(90)

holds for all $j$ simultaneously. Hence, expanding out the product in line (88), there is a quantity $\text{rem}_{t,s}(\theta)$ such that for all $\theta \in \Delta$,

$$
\prod_{j=1}^{k-1} \frac{\partial[W_{t,s}(\theta)]}{\partial \theta_j} = 1 + \varepsilon_t \text{div} W_{t,s}(\theta) + \text{rem}_{t,s}(\theta),
$$

(91)

and

$$
|\text{rem}_{t,s}(\theta)| \leq \varepsilon_t^2 \cdot \text{poly}(\varepsilon_t, K_s),
$$

(92)

where poly(⋅, ⋅) is a polynomial function whose degree and coefficients do not depend on $t$ or $s$. In particular, the last bound not depend on $\theta$. The proof is completed by combining lines (88) and (91) with the elementary bound

$$
\left| \frac{1}{1 + \eta} - (1 - \eta) \right| \leq 2\eta^2,
$$

which holds for all $|\eta| < \frac{1}{2}$. (Note that the function poly(⋅, ⋅) in line (92) need not be the same as in the statement of the lemma.)

**Lemma 7.** Assume the conditions of Theorem 2 hold, and let $\Theta$ be a random vector distributed according to $\mu$. Then, the random vector $V_{t,s}(\Theta)$ has a density $g_{t,s}(\theta)$ with respect to Lebesgue measure on $\Delta^\circ$ given by

$$
g_{t,s}(\theta) = f(V_{t,s}^{-1}(\theta)) |\det J(V_{t,s}^{-1}(\theta))|, \quad \theta \in \Delta^\circ.
$$

(93)

Furthermore, the density $g_{t,s}$ is asymptotically bounded in the sense that there is a number $c_0 \in (0, \infty)$, such that for all fixed $s \geq 1$,

$$
\lim \sup_{t \to \infty} \sup_{\theta \in \Delta^\circ} g_{t,s}(\theta) \leq c_0.
$$

(94)

**Proof.** Lemma 5 and the standard change of variables formula ([Fol99, Theorem 2.47]) give the stated expression for $g_{t,s}(\theta)$. The boundedness condition (94) follows from Lemmas 4 and 6, and the assumption that $f$ is continuous, and hence bounded, on $\Delta$.

**Lemma 8.** Suppose the conditions of Theorem 2 hold, and let $A \subset \Delta$ be a convex set. Then,

$$
\mu(\partial A) = 0,
$$

and for all $s, t \geq 1$,

$$
\mu(V_{t,s}^{-1}(\partial A)) = 0.
$$

**Proof.** The first equation follows from the fact that the boundary of any convex set in $\mathbb{R}^{k-1}$ has Lebesgue measure zero ([Dud99, Lemma 2.4.3]), and $\mu$ is assumed to have a density with respect to Lebesgue measure. To handle the second equation, we view $V_{t,s}^{-1}(\partial A)$ as the disjoint union $V_{t,s}^{-1}(\partial A \cap \Delta^\circ) \cup V_{t,s}^{-1}(\partial A \cap \partial \Delta)$ and show that each piece as measure 0 with respect to $\mu$. Regarding the second piece, since $V_{t,s}^{-1}$ is a homeomorphism from $\Delta$ to $\Delta$, the invariance of domain theorem from topology ([Mun00, Section 62]) implies that $V_{t,s}^{-1}$ must map $\partial \Delta$ to $\partial \Delta$. Hence, $V_{t,s}^{-1}(\partial A \cap \partial \Delta)$ is a subset of $\partial \Delta$, which has $\mu$-measure 0, as it is the boundary of a convex set. Regarding the first piece, we may use Lemma 7 to write $\mu(V_{t,s}^{-1}(\partial A \cap \Delta^\circ))$ as the integral of $g_{t,s}$ over $\partial A \cap \Delta^\circ$ with respect to Lebesgue measure. Therefore, $\mu(V_{t,s}^{-1}(\partial A \cap \Delta^\circ)) = 0$, as $\partial A \cap \Delta^\circ$ has Lebesgue measure 0.
Remark. In the remainder of the appendix, we let \( \ell^\infty(\Delta, \mathbb{R}^{k-1}) \) denote the space of bounded functions from \( \Delta \) to \( \mathbb{R}^{k-1} \), equipped with the uniform metric \( d(f,g) = \max_{1 \leq j \leq k-1} \sup_{\theta \in \Delta} |f_j(\theta) - g_j(\theta)| \).

Lemma 9. Let \( r_{t,s} \) be as defined in line (76). Then, there is a sequence of numbers \( \kappa_s < \infty \) such that
\[
\lim_{t \to \infty} \frac{1}{\epsilon_t} r_{t,s} = \kappa_s,
\]
and furthermore
\[
\lim_{s \to \infty} \kappa_s = 0.
\]

Proof. Suppose we can show there are numbers \( \tilde{\kappa}_s < \infty \) such that
\[
\lim_{t \to \infty} \frac{1}{\epsilon_t} \delta_{t,s} = \tilde{\kappa}_s
\]
as \( t \to \infty \), where \( \delta_{t,s} \) is defined in line (75). Then, the definition of \( r_{t,s} \) gives
\[
\lim_{t \to \infty} \frac{1}{\epsilon_t} r_{t,s} = \lim_{t \to \infty} \max \{ \frac{1}{\epsilon_t} \delta_{t,s}, 1/s \} = \max \{ \tilde{\kappa}_s, 1/s \}.
\] (95)

Hence, it is enough to show that such numbers \( \tilde{\kappa}_s \) exist and that \( \tilde{\kappa}_s \to 0 \) as \( s \to \infty \).

To proceed, recall the identities
\[
V_{t,s} - \text{id}_\Delta = L \circ B_s(F_t - \text{id}_{[0,1]})
\]
\[
V_t - \text{id}_\Delta = L(F_t - \text{id}_{[0,1]}).
\]

Letting \( I \) denote the identity map on \( \ell^\infty([0,1]) \), it is simple to check that the linear operator \( L(B_s - I) : \ell^\infty[0,1] \to \ell^\infty(\Delta; \mathbb{R}^{k-1}) \) is continuous. Consequently, we have the following uniform limit in \( \ell^\infty(\Delta; \mathbb{R}^{k-1}) \),
\[
\frac{1}{\epsilon_t}(V_{t,s} - V_t) = \frac{1}{\epsilon_t}(V_{t,s} - \text{id}_\Delta) - \frac{1}{\epsilon_t}(V_t - \text{id}_\Delta)
\]
\[
= L(B_s - I) \left[ \frac{1}{\epsilon_t}(F_t - \text{id}_{[0,1]}) \right]
\]
\[
\to L(B_s - I) h.
\] (96)

Then, by continuity, as \( t \to \infty \),
\[
\sup_{\theta \in \Delta} \frac{1}{\epsilon_t} \| V_{t,s}(\theta) - V_t(\theta) \|_2 \quad \longrightarrow \quad \sup_{\theta \in \Delta} \| [L(B_s - I)h](\theta) \|_2 =: \tilde{\kappa}_s,
\] (97)

which proves the desired numbers \( \tilde{\kappa}_s \) exist. To show that \( \tilde{\kappa}_s \) tends to 0 as \( s \to \infty \), note that since \( h \) lies in \( C[0,1] \), the uniform approximation property of Bernstein polynomials on \( C[0,1] \) in Lemma 2 ensures that as \( s \to \infty \),
\[
(B_s - I)h \to 0 \quad \text{in} \quad C[0,1].
\]

Consequently, by using the continuity of the operator \( L \), we see that \( \tilde{\kappa}_s \to 0 \).

Lemma 10. The following limits hold in the space \( \ell^\infty(\Delta, \mathbb{R}^{k-1}) \) as \( t \to \infty \),
\[
W_{t,s} = \frac{1}{\epsilon_t}(V_{t,s} - \text{id}_\Delta) \to L \circ B_s(h),
\] (98)

and
\[
\widetilde{W}_{t,s} = \frac{1}{\epsilon_t}(V_{t,s}^{-1} - \text{id}_\Delta) \to -L \circ B_s(h).
\] (99)
Proof. The limit (98) follows from the identity $\frac{1}{\varepsilon_t} (V_{t,s} - \text{id}_\Delta) = L \circ B_s(\frac{1}{\varepsilon_t} (F_t - \text{id}_{[0,1]}))$, and the continuity of $L \circ B_s$.

To prove the second limit (99), let $\Psi$ denote the inversion map from $\mathcal{F}[0,1]$ to $\ell^\infty[0,1]$, that sends a distribution function $G$ to its generalized inverse $G^{-1}$, defined by

$$G^{-1}(y) := \inf\{x \in [0,1] | G(x) \geq y\}. \tag{100}$$

Also note that $F_{t,s}$ is a proper cumulative distribution function by Lemma 2, and so it makes sense for $\Psi$ to act on $F_{t,s}$. Due to Lemma 3.9.23 in the book [vdVW96], the map $\Psi$ is Hadamard differentiable at $\text{id} \in \mathcal{F}[0,1]$ tangentially to $C[0,1]$, and the derivative acts on $C[0,1]$ at the negative of the identity map, i.e. $\Psi'_{\text{id}_{[0,1]}} = -\text{id}_{C[0,1]}$. Hence, the linearity of $B_s$ and the chain rule for Hadamard differentiation give

$$\frac{1}{\varepsilon_t} (F_{t,s}^{-1} - \text{id}_{[0,1]}) = \frac{1}{\varepsilon_t} \Psi(F_{t,s}) - \Psi(\text{id}_{[0,1]})$$

$$= \frac{1}{\varepsilon_t} \left( \Psi(B_s(F_t)) - \Psi(B_s(\text{id}_{[0,1]})) \right) \overset{t \to \infty}{\longrightarrow} \Psi'_{\text{id}_{[0,1]}}(B_s(h))$$

$$= -B_s(h), \tag{101}$$

where the limit is with respect to the sup-norm on $\ell^\infty[0,1]$. The linearity and continuity of the operator $L$ then give

$$\frac{1}{\varepsilon_t} (V_{t,s}^{-1} - \text{id}_\Delta) = \frac{1}{\varepsilon_t} (L(F_{t,s}^{-1} - \text{id}_{[0,1]})) \overset{t \to \infty}{\longrightarrow} -L(B_s(h)).$$

References


