Reproducible Aggregation of Sample-Split Statistics*

David M. Ritzwoller
Stanford University

Joseph P. Romano
Stanford University

ABSTRACT. Statistical inference is often simplified by sample-splitting. This simplification comes at the cost of the introduction of randomness not native to the data. We propose a simple procedure for sequentially aggregating statistics constructed with multiple splits of the same sample. The user specifies a bound and a nominal error rate. If the procedure is implemented twice on the same data, the nominal error rate approximates the chance that the results differ by more than the bound. We analyze the accuracy of the nominal error rate and illustrate the application of the procedure to several widely applied statistical methods.

Keywords: Sample-splitting, Cross-Validation, Replicability, Exchangeable Pairs, Stability
JEL: C01, C13, C52

1. INTRODUCTION

Sample-splitting is ubiquitous in modern statistical theory. Routine statistical tasks—model selection, dimension reduction, nuisance parameter estimation—can be implemented on a randomly selected subsample of a data set, without contaminating a statistical inference produced on its complement. This principle underlies the widely applied practices of cross-validation for predictive risk estimation (Stone, 1974; Arlot and Celisse, 2010) and cross-splitting for adaptive estimation of semiparametric models (Bickel, 1982; Klaassen, 1987; Chernozhukov et al., 2018a), among many other applications.

For a fixed data set, statistics constructed with sample-splitting are not deterministic. Two researchers can compute the same statistic on the same data and obtain different values. If this residual variability is large, substantive results based on sample-split statistics may not be credible. Researchers are incentivized to report significant results. If there is scope to materially alter the statistics that they report through the choice of the split of their sample, should this choice be left to chance?

In this paper, we propose a method for sequentially aggregating sample-split statistics over multiple splits of the same sample. Our method is based on the fixed-length sequential confidence intervals of Anscombe (1952) and Chow and Robbins (1965). The procedure takes as input a bound and an error rate. The statistic of interest is iteratively computed on different splits of the sample and aggregated in a running average. The procedure is stopped after an estimate of the residual variability of the running average falls below a fixed threshold. If the procedure were run twice, we show that the chance that the outputs differ by more the bound

Date: December 8, 2023
*Email: ritzwoll@stanford.edu, romano@stanford.edu. We thank Jiafeng Chen, Han Hong, Guido Imbens, Lihua Lei, Evan Munro, Aaditya Ramdas, and Brad Ross for helpful comments and conversations. Ritzwoller gratefully acknowledges support from the National Science Foundation under the Graduate Research Fellowship. Computational support was provided by the Data, Analytics, and Research Computing (DARC) group at the Stanford Graduate School of Business (RRI:SCR_022938).

is well-approximated by the error rate. That is, by setting the bound and the error rate to be sufficiently small, sample-split statistics aggregated with the procedure are reproducible.

To fix ideas, consider the setting of Banerjee et al. (2015), who study data collected from a randomized evaluation of a poverty alleviation program implemented in several developing countries.[1] We focus on their evaluation of the effect of the program on the level of household consumption in Bangladesh. For each household in their sample, they observe a vector $D_i = (Y_i, W_i, X_i)$, where Y_i is a measurement of consumption three years after the implementation of the program, W_i is an indicator denoting assignment to the program, and X_i is a vector of pretreatment covariates. Let $D = (D_i)_{i=1}^n$ collect the data in their sample.

The augmented inverse propensity score weighted (AIPW) estimator of Robins et al. (1994) is a standard approach to estimating average treatment effects in randomized experiments. This estimator is formed by randomly splitting the sample into two parts. Let s and $\tilde{\mathsf{s}}$ denote sets containing the indices of the households in each part. Using data from the households i in $\tilde{\mathsf{s}}$, nonparametric estimates of the propensity score and conditional outcome regression

$$\pi(x) = P\{W_i = 1 \mid X_i = x\} \quad \text{and} \quad \mu_w(x) = \mathbb{E}[Y_i \mid X_i = x, W_i = w]$$

are formed. For example, conditional outcome regressions are often estimated with the Lasso estimator of Tibshirani (1996), through

$$\hat{\mu}_w(X_i) = \hat{\beta}_w(\lambda)^\top X_i, \quad \text{where} \quad \hat{\beta}_w(\lambda) = \arg\min_\beta \left\{ \sum_{i \in \tilde{\mathsf{s}}} \mathbb{I}\{W_i = w\}(Y_i - \beta^\top X_i)^2 + \lambda \|\beta\|_1 \right\} \tag{1.1}$$

for some penalization parameter λ, with $\mathbb{I}\{\cdot\}$ denoting the indicator function.[2] Collect these estimates into $\hat{\eta}(D_{\tilde{\mathsf{s}}}) = (\hat{\pi}, \hat{\mu}_w)$, where $D_{\tilde{\mathsf{s}}}$ collects the data with indices in $\tilde{\mathsf{s}}$. The AIPW estimator is given by

$$T(\mathsf{s}, D) = \frac{1}{|\mathsf{s}|} \sum_{i \in \mathsf{s}} \psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})), \quad \text{where} \tag{1.2}$$

$$\psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{W_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{(1 - W_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\pi}(X_i)}.$$

Let s and s' denote independently drawn subsets of $[n] = \{1, \ldots, n\}$ of size $n/2$. Panel A of Figure 1 compares the densities of the single-split, double-split, and two-fold cross-split estimators

$$T(\mathsf{s}, D), \quad \frac{1}{2}(T(\mathsf{s}, D) + T(\mathsf{s}', D)), \quad \text{and} \quad \frac{1}{2}(T(\mathsf{s}, D) + T(\tilde{\mathsf{s}}, D)), \tag{1.3}$$

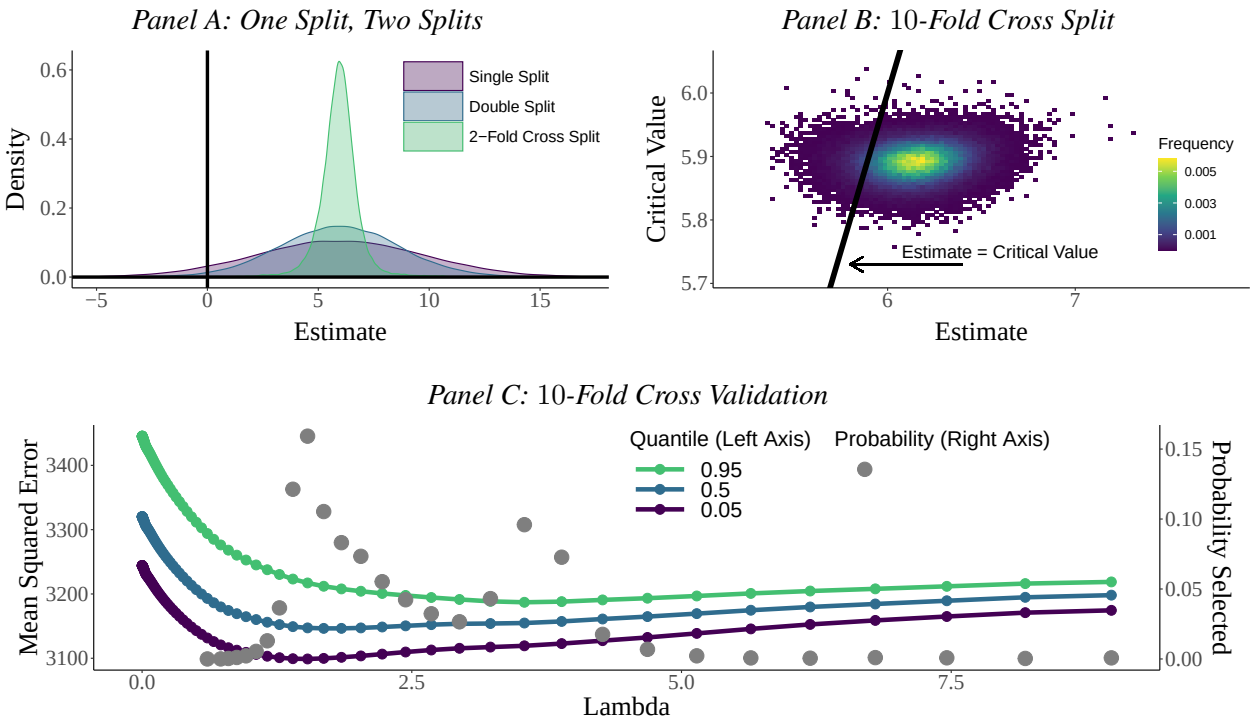
over repeated samples of $(\mathsf{s}, \mathsf{s}')$ for the Banerjee et al. (2015) data.

The variability in the estimator across sample-splits is substantial and sufficient to switch the sign. A moderate increase in the number of splits does not address the problem. The k-fold cross-split estimator is

[1] Further details on our treatment of the Banerjee et al. (2015) data are given in Appendix A.

[2] For the purposes of this section, we choose the penalization parameter with a single implementation of 10-fold cross-validation of the mean-squared error. Similarly, we estimate the propensity score with a linear probability model constructed with 10-fold cross-validated lasso. Further details on design of our simulation are given in Appendix A.

FIGURE 1. Residual Randomness



Notes: Panel A of Figure 1 displays kernel estimates of the density of the single split, double split, and two-fold cross split estimators defined in (1.3) over repeated samples of $(\mathsf{s}, \mathsf{s}')$ for the Banerjee et al. (2015) data. Panel B displays a discretized heat map of the joint distribution of the 10-fold cross split estimator $a(\mathsf{r}_{10}, D)$ defined in (1.4) and the critical value estimator $\mathsf{CV}_\alpha(\mathsf{r}_{10}, D)$ defined in (1.5). The black line denotes the threshold where the estimator is equal to the estimated critical value. Panel C displays quantiles of the 10-fold cross-validated estimate of the mean-squared error (1.6) of the Lasso regression (1.1) over a grid of values of λ. The probabilities that each value of λ minimizes the cross-validated risk estimate are displayed with large grey dots.

given by

$$a(\mathsf{r}_k, D) = \frac{1}{k} \sum_{j=1}^{k} T(\mathsf{s}_j, D) \,, \tag{1.4}$$

where $\mathsf{r}_k = (\mathsf{s}_i)_{i=1}^k$ is a collection of k equally sized sets that form a partition of $[n]$.[3] An asymptotically exact test of the null hypothesis that the average effect of the program is less than or equal to zero can be constructed by comparing (1.4) with the critical value

$$\mathsf{CV}_\alpha(\mathsf{r}_k, D) = \frac{z_{1-\alpha}}{n} \left(\sum_{j=1}^{k} \sum_{i \in \mathsf{s}_j} \left(\psi(Y_i, \hat{\eta}(D_{\tilde{\mathsf{s}}_j})) - a(\mathsf{r}_k, D) \right)^2 \right)^{1/2} , \tag{1.5}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution (Chernozhukov et al., 2018a). Panel B of Figure 1 displays a heat map of the joint distribution of the 10-fold cross-split statistic $a(\mathsf{r}_{10}, D)$ and the critical value $\mathsf{CV}_{0.05}(\mathsf{r}_{10}, D)$ over random draws of the partition r_{10}. The residual variability in the estimator

[3] If π and μ_w are estimated sufficiently well, this estimator is semiparametrically efficient for the average effect of the program on household consumption (Chernozhukov et al., 2018a; Hahn, 1998).

has been greatly reduced, but is still large relative to estimates of the sampling variability. The choice of the partition r_{10} determines whether a test that the treatment effect of the program is negative is rejected.

The problem of residual variability is even more severe when choosing the penalization parameter λ in the Lasso regression (1.1) with cross-validation. Here, the sample-split statistic of interest is the out-of-sample mean-squared error estimate

$$T(\mathsf{s}, D) = \frac{1}{\sum_{i \in \mathsf{s}} \mathbb{I}\{W_i = w\}} \sum_{i \in \mathsf{s}} \mathbb{I}\{W_i = w\}(Y_i - \hat{\beta}_w(\lambda)^\top X_i)^2 \,. \tag{1.6}$$

Often, the penalization parameter λ is chosen as the value that minimizes the cross-fit statistic (1.4). Panel C of Figure 1 displays the distribution of 10-fold cross-validated estimates of the mean-squared error of the regression (1.1), with w equal to one, over a grid of values of λ.[4] The probabilities that each value of λ minimizes the cross-validated risk estimate are displayed with large grey dots. There is substantial variability in the penalization parameter chosen by 10-fold cross-validation.

The primary methodological contribution of this paper is the introduction of a computationally efficient procedure for aggregating sample-split statistics across multiple splits to ensure that residual variability is small. The proposal is based on the sequential fixed-width confidence intervals developed by Anscombe (1952) and Chow and Robbins (1965). In Section 2, we detail the proposed algorithm and give very general conditions for its validity, in a particular asymptotic sense.

A user must make several choices when implementing the procedure. These include specifying a suitable bound and error rate as well as choosing the size and joint distribution of the random splits. The primary theoretical contribution of this paper is a non-asymptotic analysis that motivates approaches for making these choices. In particular, in Section 3, we restrict attention to linearly separable statistics that are symmetric in each part of each split. We give two main results. First, we derive concentration inequalities for averages of independently drawn cross-split statistics of the form (1.4) around their conditional mean. Our bounds characterize the dramatic difference between the residual variability of the double split and two-fold cross split statistics displayed in Panel A of Figure 1. Second, we provide a Berry-Esseen type bound on the accuracy of the nominal error rate for the procedure. As the procedure is sequential, this calculation is nonstandard. This result illustrates a trade-off between the computational efficiency of the procedure and the accuracy of the nominal error rate.

The main theoretical challenge posed by this analysis is the accommodation of the dependence between statistics computed on cross-splits of a sample. That is, conditioned on the data, the summands, over j from 1 to k, of the aggregate statistic (1.4) will be dependent. We address this through an application of the method of exchangeable pairs (see e.g., Stein, 1986; Ross, 2011; Chen et al., 2011). To construct an appropriate exchangeable pair for our problem, we develop a novel application of a coupling argument due to Chatterjee (2005). See also Paulin et al. (2013, 2016). The various concentration and moment inequalities we develop are then obtained by applying results due to Chatterjee (2005, 2007). These arguments are outlined in Section 5.

[4] We consider the values of λ reported by the "glmnet" R software package (Friedman et al., 2021).

In Section 4, we measure the performance of our procedure in the applications considered in Figure 1. Section 6 concludes. Details concerning the data and simulations considered in this paper are given in Appendix A. Unless otherwise specified, proofs for all results stated in the main text are given in Appendix B. Proofs supporting Lemmas used in Appendix B are given in Appendix C. Appendix D gives additional results that will be introduced at appropriate points throughout the paper.

1.1 Related Literature. The procedure studied in this paper is based on the sequential fixed-width confidence intervals of Anscombe (1952) and Chow and Robbins (1965). Our non-asymptotic analysis is related to the Berry-Esseen bounds for randomly indexed sums given in Landers and Rogge (1976, 1988). Applications of these results to fixed-width confidence intervals have been given in Csenki (1980), Mukhopadhyay (1981), and Callaert and Janssen (1981). A more recent non-asymptotic consideration of the application of fixed width confidence intervals to Monte Carlo techniques is given in Hickernell et al. (2013). Our more specialized analysis differs from these results by accommodating and quantifying the dependence induced by cross-splitting.

Our setting is related to a large literature that studies methods for constructing confidence intervals for cross-validated estimates of generalization error. Several recent examples include Dietterich (1998), Nadeau and Bengio (1999), Lei (2020), Bayle et al. (2020), Austern and Zhou (2020), and Bates et al. (2023). In contrast, we are interested in the randomness conditional on the data. Formally, our results are most similar to the Berry-Esseen bounds given in Austern and Zhou (2020), who study the unconditional normal approximation of statistics similar to the those considered in Section 3. Our bounds apply under strictly weaker conditions and are substantially simpler. See also Chetverikov et al. (2021) and Chetverikov and Sørensen (2021) for further analyses of cross-validated regression.

Some of our arguments build on a literature studying the role of algorithmic stability in the accuracy of cross-validation (Kale et al., 2011; Kumar et al., 2013). These papers are related to a broader literature that derives generalization bounds for stable algorithms, originating with Bousquet and Elisseeff (2002). The stability of several specific estimators is studied in Elisseeff et al. (2005), Hardt et al. (2016), Celisse and Guedj (2016), Chen et al. (2022), and Du et al. (2023). Some of the concentration inequalities we derive can be compared to the results of Cornec (2010) and Abou-Moustafa and Szepesvári (2019). Again, the setting we study is different and our conditions are substantially simpler.

Chen et al. (2022) show that sample-splitting is unnecessary for the asymptotic normality of estimators of the form (1.2) if the nuisance parameter estimator $\hat{\eta}$ satisfies a stability condition related to some of the conditions studied in this paper. The intention of our paper is not to argue that sample-splitting should, necessarily, be used for any particular application. Rather, our premise is that sample-splitting is widely applied in practice, irrespective of its optimality.

The procedure studied in this paper is applicable to a large variety of statistical methods constructed with sample-splitting. General methods for testing statistical hypothesis with sample-splitting are studied in Guo and Romano (2017), DiCiccio and Romano (2019), DiCiccio et al. (2020), Wasserman et al. (2020), and Tse and Davison (2022). Rinaldo et al. (2019) study a general procedure for selective inference based on sample-splitting. Meinshausen et al. (2009) and Meinshausen and Bühlmann (2010) study methods for inference

in high-dimensional linear models that use multiple data splitting. Chernozhukov et al. (2018b) study a procedure for estimating best linear predictors of heterogeneous treatment effects based on sample-splitting.

Although our emphasis is on statistics constructed with sample-splitting, the algorithmic and formal methods studied in this paper are potentially applicable to randomized algorithms more generally. See Beran and Millar (1987) for an analysis of the asymptotics of randomized tests and estimators. Guo and Shah (2023) and Zhang et al. (2023) propose methods for conducting inference with randomized algorithms through subsampling (Politis et al., 1999).

1.2 Notation. We let $\tilde{\mathsf{s}}$ denote the complement of the set s in $[n] = \{1, \ldots, n\}$. The set $\mathcal{S}_{n,b}$ consists of all subsets of $[n]$ of size b. The set $\mathcal{R}_{n,k,b}$ contains all collections of k mutually exclusive elements of $\mathcal{S}_{n,b}$. We refer to elements of $\mathcal{R}_{n,k,b}$ as cross-splits. That is, if $n = k \cdot b$, then $\mathcal{R}_{n,k,b}$ is the set of all partitions of $[n]$ into k mutually exclusive subsets of size b. Likewise, $\mathcal{R}_{n,1,b}$ is the collection of sets containing one element of $\mathcal{S}_{n,b}$. We say $a \lesssim b$ if there exists a universal constant C such that $a \leq Cb$. We let z_α denote the α quantile of the standard normal distribution and $\lfloor x \rfloor$ denote the largest integer smaller than or equal to x.

2. Reproducible Aggregation

Consider a sample $D = (D_i)_{i=1}^n$. Let s be some subset of $[n]$ with complement $\tilde{\mathsf{s}}$. We are interested in reporting a real-valued sample-split statistic of the form

$$T(\mathsf{s}, D) = \Psi\left(D_{\mathsf{s}}, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)\right), \tag{2.1}$$

where the collection $D_{\mathsf{s}} = (D_i)_{i \in \mathsf{s}}$ contains data with indices in s, $\hat{\eta}$ is an estimator of an unknown nuisance parameter η, and $\Psi(\cdot)$ is some function. The structure encoded in statistics of the form (2.1) is quite general and encompasses most applications of sample-splitting encountered in practice, including out-of-sample evaluation of prediction error, model selection, and nuisance parameter estimation.

We study methods for aggregating statistics of the form (2.1) across multiple splits of the same sample. Our objective is to ensure that the auxiliary randomness introduced by sample-splitting is small. In particular, for each collection $\mathsf{R}_{g,k} = (\mathsf{r}_i)_{i=1}^g$ of g elements of $\mathcal{R}_{n,k,b}$, where we write $\mathsf{r}_i = (\mathsf{s}_{i,j})_{j=1}^k$, we consider aggregate statistics of the form

$$a(\mathsf{R}_{g,k}, D) = \frac{1}{g}\frac{1}{k}\sum_{i=1}^{g}\sum_{j=1}^{k} T(\mathsf{s}_{i,j}, D)\,. \tag{2.2}$$

The statistic (2.2) generalizes several standard methods for aggregating sample-split statistics. If $k = n/b$ and $g = 1$, then (2.2) is a k-fold cross-split statistic (also known as, e.g., k-fold cross-validation or k-fold cross-fitting). If $k = 1$, then (2.2) averages over g independently drawn sample-splits. In this case, if $\hat{\eta}$ is constant, then (2.2) is an incomplete U-statistic of order b.[5] We entertain the intermediate cases, i.e., where k is between 1 and n/b, so that the theoretical results to follow interpolate between independent splitting and cross-splitting. Throughout, we denote the proportion b/n by φ.

[5] A special case of the class of statistics with $k = 1$ are subsampled statistics (Politis et al., 1999).

Algorithm 1: Anscombe-Chow-Robbins Aggregation

Input: Data D, tolerance ξ, error rate β, collection size k, split size b, initialization g_{init}

Set $g \leftarrow g_{\mathsf{init}}$

Draw r_1, ..., $\mathsf{r}_{g_{\mathsf{init}}}$ independently and uniformly from $\mathcal{R}_{n,k,b}$. Collect $\mathsf{R}_{g_{\mathsf{init}},k} = (\mathsf{r}_j)_{j=1}^{g_{\mathsf{init}}}$.

while $\hat{v}(\mathsf{R}_{g,k}, D) > \mathsf{cv}_{\xi,\beta}$ **do**

  Set $g \leftarrow g + 1$

  Draw r_g uniformly from $\mathcal{R}_{n,k,b}$. Collect $\mathsf{R}_{g,k} = (\mathsf{R}_{g-1}, \mathsf{r}_g)$.

end

Set $\hat{g} \leftarrow g$

return $\mathsf{R}_{\hat{g},k}$

Our task is to formulate a method for choosing the number of collections of mutually exclusive splits g to ensure that the residual variability of the aggregate statistic (2.2) is small. This objective is formalized in the following criterion.

Definition 2.1 (Reproducibility)**.** Suppose that the integers $\hat{g}$ and $\hat{g}'$ and the collections $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$, are independent and identically distributed, conditional on the data D. We say that $\mathsf{R}_{\hat{g},k}$ is (ξ, β)-reproducible if

$$P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \leq \xi \mid D\right\} \geq 1 - \beta \tag{2.3}$$

almost surely.

Remark 2.1. Suppose that the data D are given to two researchers. Each researcher is tasked with producing an estimate of the form (2.2). They generate the collections of splits $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$ independently using the same procedure. That is, the two collections of splits used by the two researchers are independent and identically distributed conditional on the data D. If this procedure is (ξ, β)-reproducible, then the probability that the two researchers' estimates differ by more than ξ is less than β. ■

We propose a sequential method for constructing a reproducible collection $\mathsf{R}_{\hat{g},k}$. Our proposal is based on the fixed-length sequential confidence intervals of Anscombe (1952) and Chow and Robbins (1965). Define the variance estimator

$$\hat{v}\left(\mathsf{R}_{g,k}, D\right) = \frac{1}{g\left(g-1\right)k^2} \sum_{i=1}^{g} \sum_{j,j'=1}^{k} \left(T(\mathsf{s}_{i,j}, D) - a\left(\mathsf{R}_{g,k}, D\right)\right)\left(T(\mathsf{s}_{i,j'}, D) - a\left(\mathsf{R}_{g,k}, D\right)\right). \tag{2.4}$$

Observe that (2.4) is simply the sample variance of the summands in (2.2) across the g collections of splits. Our procedure iteratively draws a collection uniformly at random and computes the variance (2.4) until the condition

$$\hat{v}\left(\mathsf{R}_{g,k}, D\right) \leq \mathsf{cv}_{\xi,\beta} = \frac{1}{2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \tag{2.5}$$

is satisfied. We let $\hat{g}$ denote the smallest value of g greater than or equal to g_{init} such that (2.5) is satisfied, where $g_{\mathsf{init}} \geq 2$ is an integer chosen by the user. This procedure is summarized in Algorithm 1.

The following theorem gives an asymptotic sense in which the collection of sample-splits chosen with Algorithm 1 are (ξ, β)-reproducible. The proof is closely related to the arguments of Anscombe (1952) and Chow and Robbins (1965). See e.g., Theorem 3.1 of Gut (2009) for a textbook treatment.

Theorem 2.1. *Suppose that the real-valued statistic $a(\mathsf{R}_{1,k}, D)$ has a non-zero variance conditional on D almost surely. If the collections $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$ are independently computed with Algorithm 1, then*

$$P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \leq \xi \mid D\right\} \to 1 - \beta \quad \text{as} \quad \xi \to 0\,. \tag{2.6}$$

Remark 2.2. Theorem 2.1 applies to a fixed data set D. Asymptotics are taken as $\xi \to 0$ with all other quantities, including the sample size n, fixed. Roughly, Theorem 2.1 is established by showing that

$$\hat{g}/g^\star \xrightarrow{\text{a.s}} 1 \quad \text{as} \quad \xi \to 0, \quad \text{where} \quad g^\star = \min\left\{g : \mathrm{Var}\left(a(\mathsf{R}_{g,k}, D) \mid D\right) \leq \mathsf{cv}_{\xi,\beta}\right\}\,, \tag{2.7}$$

and using the fact that a central limit theorem will hold for the sequence $a(\mathsf{R}_{g^\star,k}, D)$ as $\xi \to 0$, conditional on the data. The discrepancy between $a(\mathsf{R}_{\hat{g},k}, D)$ and $a(\mathsf{R}_{g^\star,k}, D)$ can be handled with an appropriate maximal inequality, an idea due to Rényi (1957). ■

Remark 2.3. The proof of Theorem 2.1 uses only the fact that the g collections r_i in $\mathsf{R}_{g,k}$ are drawn independently and identically. However, Figure 1 suggests that we should expect the statistic $a(\mathsf{R}_{g,k}, D)$ to concentrate as k increases. Some insight into this phenomenon can be gained without any assumptions. In particular, let s and s' each be uniformly distributed on $\mathcal{S}_{n,b}$ such that their intersection is empty with probability one. Define the conditional variance and covariance

$$\phi_{n,b}(D) = \mathrm{Var}(T(\mathsf{s}, D) \mid D) \quad \text{and} \quad \gamma_{n,b}(D) = \mathrm{Cov}(T(\mathsf{s}, D), T(\mathsf{s}', D) \mid D) \tag{2.8}$$

respectively. We can evaluate

$$v_{g,k}(D) = \mathrm{Var}(a(\mathsf{R}_{g,k}, D) \mid D) = \frac{1}{g}\left(\frac{1}{k}\phi_{n,b}(D) + \frac{k-1}{k}\gamma_{n,b}(D)\right). \tag{2.9}$$

Observe that $\gamma_{n,b}(D) \leq \phi_{n,b}(D)$ by the Cauchy-Schwarz inequality. Thus, for a fixed total number of cross-splits g and split size b, the conditional variance $v_{g,k}(D)$ decreases as k increases. That is, cross-splitting (i.e., setting $k > 1$) always has a smaller conditional variance than independent splitting (i.e., setting $k = 1$) for a fixed value of g. On the other hand, for a fixed number of total splits $m = gk$ and split size b, the conditional variance $v_{g,k}(D)$ is smaller for cross-splitting than for independent splitting if and only if the conditional covariance $\gamma_{n,b}(D)$ is negative.

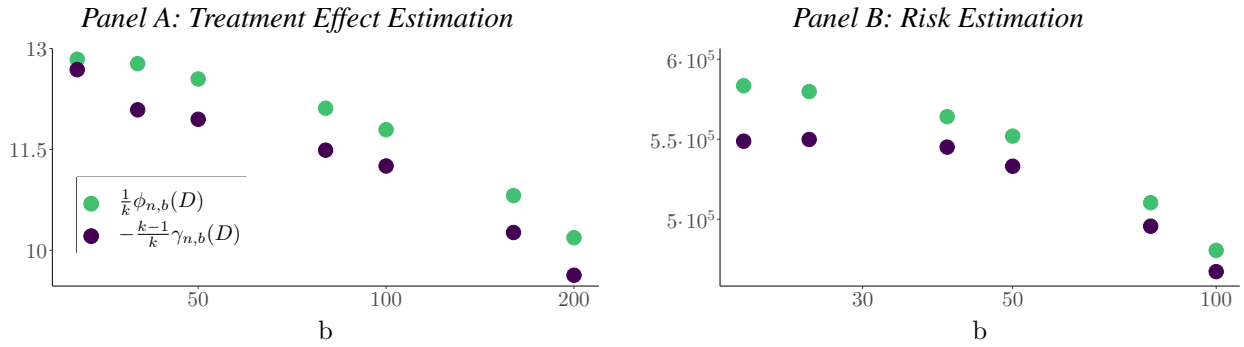
In many standard applications, the conditional covariance $\gamma_{n,b}(D)$ is negative.[6] Consider again the Banerjee et al. (2015) data. Figure 2 compares the variance components

$$\frac{1}{k}\phi_{n,b}(D) \quad \text{and} \quad -\frac{k-1}{k}\gamma_{n,b}(D) \tag{2.10}$$

under the restriction that $n = kb$ for both the treatment effect and risk estimation applications introduced in Section 1. In both cases, the two terms (2.10) are nearly equal, i.e., the two terms in the variance (2.9) nearly

[6] This will not always be the case. In Appendix D.1, we give examples where $\gamma_{n,b}(D)$ obtains both endpoints of the Cauchy-Schwarz bound $-\phi_{n,b}(D) \leq \gamma_{n,b}(D) \leq \phi_{n,b}(D)$.

FIGURE 2. Covariance Across Cross-Splits



Notes: Figure 2 displays measurements of the quantities (2.10) for the treatment effect and risk estimation applications to the Banerjee et al. (2015) data introduced in Section 1. For the risk estimation application, we study the regression (1.1) for the treated outcome, i.e., w equal to one, for a fixed choice of λ. In each case, the parameters k and b are chosen to satisfy $n = kb$. The x-axes are displayed in a log-scale base 10. Further details on this figure are given in Appendix A.

cancel each other out. In other words, cross-splitting is able to account for almost all of the residual variability introduced by independent splitting. By placing restrictions on the statistic (2.1), the subsequent section provides a non-asymptotic account of the error in the approximation (2.6) that quantifies and compares the concentration in $a(\mathsf{R}_{g,k}, D)$ with both g and k. ■

3. NON-ASYMPTOTIC THEORY

Theorem 2.1 is very general. However, interpretation of the asymptotic approximation with $\xi \to 0$ is somewhat nebulous, or at least unfamiliar. What would be a reasonable value of ξ to choose such that the approximation (2.6) is accurate? Or, how quickly should we expect the error in (2.6) to diminish as ξ shrinks?

We give a non-asymptotic analysis of the quality of the nominal error rate β in Algorithm 1. This is accomplished by placing restrictions on the statistics of interest (2.1). We emphasize that Theorem 2.1 should provide assurance that Algorithm 1 can be applied widely. The more specialized analysis that follows is aimed at (i) articulating why cross-split statistics often exhibit significantly lower residual variability than independently split statistics and (ii) providing intuition for how the error in the approximation (2.6) depends on the choices of k and ξ. Proofs for results stated in this section are deferred to Section 5.

3.1 Symmetry, Linearity, and Stability. We impose two simplifying restrictions on the statistic of interest. First, we assume that the statistic is symmetric and deterministic in each part of the split sample.

Assumption 3.1 (Symmetry and Determinism)**.** *For all sets* $\mathsf{s} \subseteq [n]$ *and data* D*, the statistic* $T(\mathsf{s}, D)$ *is deterministic and invariant to permutations of the data with indices in* s *and of the data with indices in* $\tilde{\mathsf{s}}$*.*

Remark 3.1. The intention of Assumption 3.1 is to restrict the residual randomness under consideration to the randomness introduced by sample-splitting. Assumption 3.1 rules out procedures where the estimator $\hat{\eta}$ is random conditional on the data. This includes settings where, e.g., $\hat{\eta}$ is estimated with stochastic gradient descent, bagging or subsampling, or is itself constructed with data splitting. It does not, however, exclude

cases where $\hat{\eta}$ is deterministic, e.g., when $\hat{\eta}$ is a coefficient vector determined by a regularized regression or in the applications to hypothesis testing considered by DiCiccio et al. (2020) or Wasserman et al. (2020).[7] ■

Second, we assume that the statistic (2.1) is linearly separable in the first part of the split sample.

Assumption 3.2 (Linearity)**.** *For all sets* $\mathsf{s} \subseteq [n]$ *and data* D*, the statistic* $T(\mathsf{s}, D)$ *can be written*

$$T(\mathsf{s}, D) = \Psi\left(D_{\mathsf{s}}, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)\right) = \frac{1}{|\mathsf{s}|} \sum_{i \in \mathsf{s}} \psi(D_i, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) \tag{3.1}$$

for some function $\psi(\cdot, \cdot)$*.*

Remark 3.2. Assumption 3.2 is satisfied in the applications to treatment effect and risk estimation studied in Section 1 in addition to the applications to hypothesis testing considered in DiCiccio et al. (2020) and Wasserman et al. (2020). Extending our analysis to cases where (2.1) satisfies a component-wise Lipschitz or bounded differences condition is worth further consideration. ■

Both the remaining assumptions, and the ensuing bounds, specified in this section are expressed in terms of two objects that measure the sensitivity of the statistics of interest (2.1) to perturbations of the data and of the splits, respectively. We refer to these objects as stabilities. Let D' denote an independent and identical copy of the data D. For each $\mathsf{q} \subseteq [n]$, let $\tilde{D}^{(\mathsf{q})}$ be constructed by replacing D_i with D'_i in D for each i in q.

Definition 3.1 (Sample Stability)**.** Fix a set $\mathsf{s} \subseteq \mathcal{S}_{n,b}$. Let I be an index drawn uniformly from s. Let q be a randomly selected subset of $\tilde{\mathsf{s}}$ of cardinality q. We refer to the quantities

$$\sigma^{(r)}_{\mathsf{valid}} = \mathbb{E}\left[\left|\psi(D_I, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) - \psi(D'_I, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right))\right|^r\right] \quad \text{and}$$
$$\sigma^{(r,q)}_{\mathsf{train}} = \mathbb{E}\left[\left|\psi(D_I, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) - \psi(D_I, \hat{\eta}(\tilde{D}^{(\mathsf{q})}_{\tilde{\mathsf{s}}}))\right|^r\right]$$

as the rth-order validation and (r, q)th-order training sample stabilities, respectively. Similarly, we refer to the quantity

$$\sigma^{(r)}_{\mathsf{max}} = \max\left\{\sigma^{(r)}_{\mathsf{valid}}, \sigma^{(r,1)}_{\mathsf{train}}\right\} \tag{3.2}$$

as the rth-order full sample stability.

Definition 3.2 (Split Stability)**.** We refer to the quantity

$$\zeta^{(r)} = \mathbb{E}\left[\max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T(\mathsf{s}, D) - T\left(\mathsf{s}', D\right)\right)^r\right].$$

as the rth-order split stability.

At several points in the analysis that follows, we focus attention on statistics whose training sample stabilities shrink at an appropriate rate as n increases.

[7] In Appendix D.2, we show that the hypothesis tests constructed with multiple splits of the same sample considered in e.g., DiCiccio et al. (2020), Meinshausen et al. (2009), Chernozhukov et al. (2018b), and Wasserman et al. (2020) continue to be valid when they are constructed sequentially with Algorithm 1. These results follow from an inequality due to Ramdas and Manole (2023).

Definition 3.3. A statistic $T(\mathsf{s}, D)$ that satisfies Assumption 3.2 is sample stable if

$$\sigma_{\mathsf{train}}^{(r,q)} \lesssim \left(\frac{q}{n-b}\right)^r \tag{3.3}$$

uniformly for all even values of the parameter r.

Remark 3.3. The $(2,1)$th-order training sample stability $\sigma_{\mathsf{train}}^{(2,1)}$ is a widely studied object and is referred to as mean-square stability in the statistical learning literature (see e.g., Kale et al., 2011; Kumar et al., 2013).[8] In Appendix D.3, we give simple sufficient conditions for the sample stability of statistics that satisfy Assumption 3.2 when $\hat{\eta}$ is an empirical risk minimizer of a potentially regularized strictly convex loss. The argument is closely related to a similar result given in Austern and Zhou (2020). Analogous bounds have been obtained for many estimators that do not necessarily satisfy Assumption 3.1, such as bagged or subsampled estimators (Chen et al., 2022), ensemble methods (Elisseeff et al., 2005), and estimators computed with stochastic gradient descent (Hardt et al., 2016). It most settings, it is reasonable to expect that the validation sample stability $\sigma_{\mathsf{valid}}^{(r)}$ is bounded for small values of r. The split stability $\zeta^{(r)}$ is a less frequently studied object, and we will only require that it is finite for r equal to 4 or 8, depending on the setting. This is a weak restriction that will hold, for example, if $T(\mathsf{s}, D)$ is bounded. ■

3.2 Concentration and Normal Approximation. We begin by studying the concentration of the aggregate statistic $a(\mathsf{R}_{g,k}, D)$ about its conditional mean

$$\overline{a}(D) = \mathbb{E}\left[a(\mathsf{R}_{g,k}, D) \mid D\right] = \mathbb{E}\left[T(\mathsf{s}_{i,j}, D) \mid D\right] . \tag{3.4}$$

The nonstandard aspect of our analysis is that we account for the dependence in the summands in $a(\mathsf{R}_{g,k}, D)$ across cross-splits, i.e., the dependence induced by cross-fitting. This is accomplished by constructing a suitable exchangeable pair with a coupling argument due to Chatterjee (2005) and applying a method for deriving concentration inequalities with exchangeable pairs, also due to Chatterjee (2005, 2007). Aspects of our argument are closely related to the analysis of Paulin et al. (2013, 2016).

Theorem 3.1. *Suppose that Assumptions 3.1 and 3.2 hold, that the data D are independently and identically distributed, and that the statistic $a(\mathsf{R}_{1,k}, D)$ has a non-zero variance conditional on D almost surely. If the fourth-order split stability $\zeta^{(4)}$ is finite, then the concentration inequality*

$$\log P\left\{a(\mathsf{R}_{g,k}, D) - \overline{a}(D) \geq t \mid D\right\} \leq -\frac{g}{2^4(2-\varphi k-\varphi)^2}\frac{\delta t^2}{\Gamma_{k,\varphi,b}} \tag{3.5}$$

holds with probability greater than $1-\delta$ as D varies, where

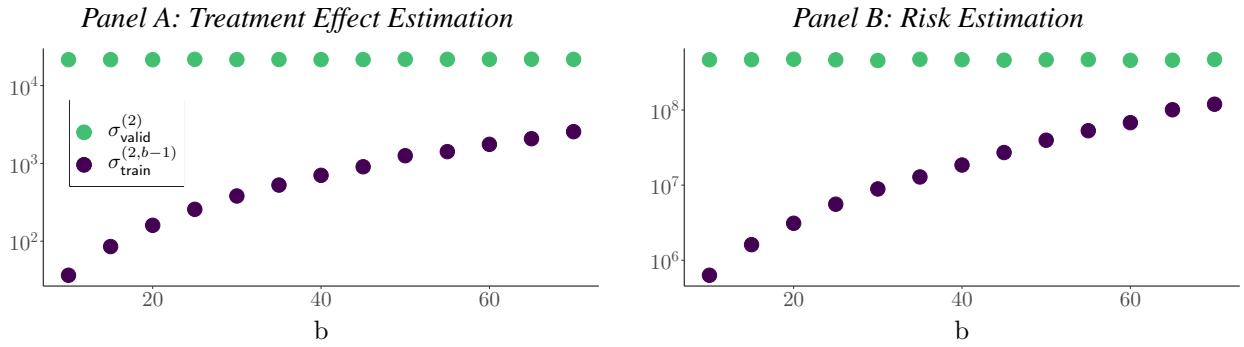
$$\Gamma_{k,\varphi,b} = \left(\frac{1-k\varphi}{1-\varphi}\right) 4\sigma_{\mathsf{max}}^{(2)} + \left(\frac{k\varphi-\varphi}{1-\varphi}\right)\sigma_{\mathsf{train}}^{(2,b-1)} \tag{3.6}$$

and $\varphi = b/n$.

Remark 3.4. In the inequality (3.5), a conditional probability that depends on the data D is bounded by a term that depends on the fixed parameters g, k, φ, δ, and t in addition to the unconditional quantity

[8] Connections between sample stability and generalization are studied in Bousquet and Elisseeff (2002) and Celisse and Guedj (2016).

FIGURE 3. Sample Stability



Notes: Figure 3 displays estimates of the validation and training sample stabilities, defined in Definition 3.1, for the treatment effect and risk estimation applications to the Banerjee et al. (2015) data introduced in Section 1. For the risk estimation application, we study the regression (1.1) for the treated outcome, i.e., w equal to one, for a fixed choice of λ. We approximate the unknown true data generating distribution with the empirical distribution of the data. The y-axes are displayed in a log-scale base 10. Further details on this figure are given in Appendix A.

$\Gamma_{k,\varphi,b}$. In other words, if $\mathcal{F}$ is the event that the inequality (3.5) holds, then $P\{\mathcal{F}\} > 1 - \delta$ unconditionally. This transference from a conditional quantity to an unconditional quantity is obtained by a Markov type bound at one point in the proof. In the sequel, we give many conditional bounds of a similar form. This strategy—bounding the conditional quantities of interest with unconditional quantities—is helpful because the resultant unconditional objects are more tractable and are widely studied. ■

Remark 3.5. The quantity $\Gamma_{k,\varphi,b}$ interpolates between $\sigma^{(2)}_{\mathsf{max}}$ and $\sigma^{(2,b-1)}_{\mathsf{train}}$ as k varies between its bounds 1 and $1/\varphi$.[9] In the case of independent splitting, the rate of concentration for the statistic $a(\mathsf{R}_{g,k}, D)$ is driven by the full sample stability $\sigma^{(2)}_{\mathsf{max}}$. On the other hand, in the case of cross-splitting, concentration depends only on the training sample stability $\sigma^{(2,b-1)}_{\mathsf{train}}$. At the extreme, if b is set equal to one and k is set equal to n, known as leave-one-out or jackknife cross-fitting, $\Gamma_{k,\varphi,b}$ will equal zero and there will be no residual randomness (that is, the assumption that $a(\mathsf{R}_{1,k}, D)$ has a non-zero conditional variance is violated).

To provide a quantitative sense of this tradeoff, Figure 3 displays estimates of $\sigma^{(2)}_{\mathsf{valid}}$ and $\sigma^{(2,b-1)}_{\mathsf{train}}$ for the risk and treatment effect estimation applications to the Banerjee et al. (2015) data, where we have approximated the unknown data generating distribution with the empirical distribution of the data. The training stability is dramatically smaller than the validation stability and decreases rapidly as b decreases.

An interesting structure emerges under the assumption that $T(\mathsf{s}, D)$ is sample stable. In particular, restricting attention to the case of cross-fitting $b = n/k$, and plugging (3.3) into the bound (3.5) yields

$$\frac{\delta g t^2}{2^4(1-\varphi)^2}\left(\frac{n-b}{b-1}\right)^2 \geq \frac{\delta}{2^4} g k^2 t^2$$

In other words, in the case of cross-splitting, large deviation probabilities converge to zero with an exponential rate that is quadratic in k. By contrast, the exponential rate of concentration is always linear in g. That is,

[9] The pre-factor $(2 - \varphi k - \varphi)$ also reduces by a factor of 2 as k increases from 1 to $1/\varphi$.

for a fixed total number of splits $m = gk$, the exponential rates of concentration for sample stable statistics aggregated with independent splitting and with cross-splitting are m and $m \cdot k$, respectively. ■

Next, we derive bounds for the centered conditional moments of $a\left(\mathsf{R}_{g,k}, D\right)$ through an argument closely related to the proof of Theorem 3.1. Bounds of this form are known as Burkholder-Davis-Gundy inequalities (Burkholder, 1973). By combining these bounds with a standard Berry-Esseen inequality (Shevtsova, 2011), we obtain a non-asymptotic central limit theorem as a corollary.

Theorem 3.2. *Suppose that Assumptions 3.1 and 3.2 hold and that the data D are independent and identically distributed. If $r = 2^{c-1}$ for some positive integer c and the $4r$th-order split stability $\zeta^{(4r)}$ is finite, then the inequality*

$$\mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)\right)^{2r}\right] \leq (2r-1)^r \left(\frac{2^4(2-\varphi k-\varphi)^2}{g}\right)^r \Gamma_{k,\varphi,b}^{(r)}\,, \tag{3.7}$$

holds, where

$$\Gamma_{k,\varphi,b}^{(r)} = \left(\frac{1-k\varphi}{1-\varphi}\right) 2^{2r} \sigma_{\mathsf{max}}^{(2r,1)} + \left(\frac{k\varphi-\varphi}{1-\varphi}\right) \sigma_{\mathsf{valid}}^{(2r,b-1)} \tag{3.8}$$

and $\varphi = b/n$.

Remark 3.6. By setting $r = 1$, the inequality (3.7) gives a variance bound. In this case, the right-hand side of the inequality (3.7) is equal to negative one times the inverse of the right-hand side of the inequality (3.5). In this sense, the concentration inequality (3.5) can be thought of as an exponential Efron-Stein inequality adapted to the dependence inherent in our problem (Boucheron et al., 2003).[10] For values of r greater than 1, the unconditional quantity $\Gamma_{k,\varphi,b}^{(r)}$ again interpolates between the $2r$th order full sample stability $\sigma_{\mathsf{max}}^{(2r,1)}$ and the $(2r, b-1)$th order training sample stability $\sigma_{\mathsf{valid}}^{(2r,b-1)}$ as k increases from 1 to $1/\varphi$. For sample stable statistics under cross-fitting, the bound (3.7) can be re-written

$$\mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)\right)^{2r}\right] \lesssim \left(\frac{1}{gk^2}\right)^r \tag{3.9}$$

by omitting constants that depend only on r. The rate of convergence of central moments of cross-split sample stable statistics is much faster with k than it is with g. ■

The Kolmogorov distance between two measures P and Q on the real line is given by

$$d_{\mathrm{K}}\left(P, Q\right) = \sup_{w \in \mathbb{R}} \left\{\left|P\left\{(-\infty, w]\right\} - Q\left\{(-\infty, w]\right\}\right|\right\}. \tag{3.10}$$

With a mild abuse of notation, we let $d_{\mathrm{K}}\left(X, Y\right)$ denote the Kolmogorov distance between the probability measures of the real-valued random variables X and Y.

Corollary 3.1. *Let W denote a standard normal random variable. Suppose that Assumptions 3.1 and 3.2 hold, that the data D are independently and identically distributed, and that the statistic $a(\mathsf{R}_{1,k}, D)$ has a non-zero variance conditional on D almost surely. If the eighth-order split stability $\zeta^{(8)}$ is finite, then for all g, the*

[10] Abou-Moustafa and Szepesvári (2019) give related exponential Efron-Stein inequality in a different setting.

conditional Berry-Esseen inequality

$$d_K\left(\frac{a\left(\mathsf{R}_{g,k}, D\right)-\overline{a}\left(D\right)}{\sqrt{v_{g,k}\left(D\right)}}, W \middle| D\right) \leq \frac{2^6 3^2}{\delta}\frac{(2-\varphi k-\varphi)^3}{g^{1/2}}\left(\frac{(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{v_{1,k}(D)}\right)^{3/2} \tag{3.11}$$

is satisfied with probability greater than $1-\delta$ *as* D *varies.*[11]

Remark 3.7. Consider again the case of cross-splitting, where $k = n/b = 1/\varphi$. By Theorem 3.2, the Berry-Esseen bound (3.11) is at least

$$\frac{(1-\varphi)^3}{g^{1/2}}\left(\frac{(\sigma^{(4,b-1)}_{\mathsf{train}})^{1/2}}{v_{1,k}(D)}\right)^{3/2} \gtrsim \frac{1}{g^{1/2}}\left(\frac{(\sigma^{(4,b-1)}_{\mathsf{train}})^{1/2}}{\sigma^{(2,b-1)}_{\mathsf{train}}}\right)^{3/2}. \tag{3.12}$$

Suppose that the statistic is sample stable. As long as $\sigma^{(4,b-1)}_{\mathsf{train}}$ and $\sigma^{(2,b-1)}_{\mathsf{train}}$ are well-approximated by their upper bounds (3.3), the quantity (3.12) is of order $g^{-1/2}$ and does not decrease as k increases.

This argument suggests that, despite the fast decay in the residual randomness demonstrated in Theorem 3.2, the quality of a normal approximation to $a\left(\mathsf{R}_{g,k}, D\right)$ does not increase as k increases. This is again born out in our running examples. Figure 4 displays the Kolmogorov distance between a standard normal random variable and the studentized distribution of the cross-fit statistic $a\left(\mathsf{R}_{1,k}, D\right)$ conditioned on D over a range of values of k for our two applications to the Banerjee et al. (2015) data. That is, Figure 4 plots the left-hand side of the inequality (3.11). In both cases, despite a substantial initial improvement, the quality of a normal approximation does not increase for large k. In fact, in the risk estimation application, the quality of a normal approximation decays for values of k larger than 20. On the other hand, it is worth noting that a normal approximation in the treatment effect application is remarkably accurate for values of k larger than 20. ■

3.3 Reproducibility. We now return our attention to the sequential procedure for aggregating sample-split statistics formulated in Section 2. Our main result is a Berry-Esseen type bound on the accuracy of the nominal error rate β for the Anscombe-Chow-Robbins method specified in Algorithm 1.

Theorem 3.3. *Suppose that the statistic* $a(\mathsf{R}_{1,k}, D)$ *has a non-zero and variance, conditional on* D*, almost surely and that the collections* $\mathsf{R}_{\hat{g},k}$ *and* $\mathsf{R}'_{\hat{g}',k}$ *are independently computed with Algorithm 1. If Assumptions 3.1 and 3.2 hold, the data* D *are independent and identically distributed, and the eighth-order split stability* $\zeta^{(8)}$ *is finite, then the Berry-Esseen inequality*

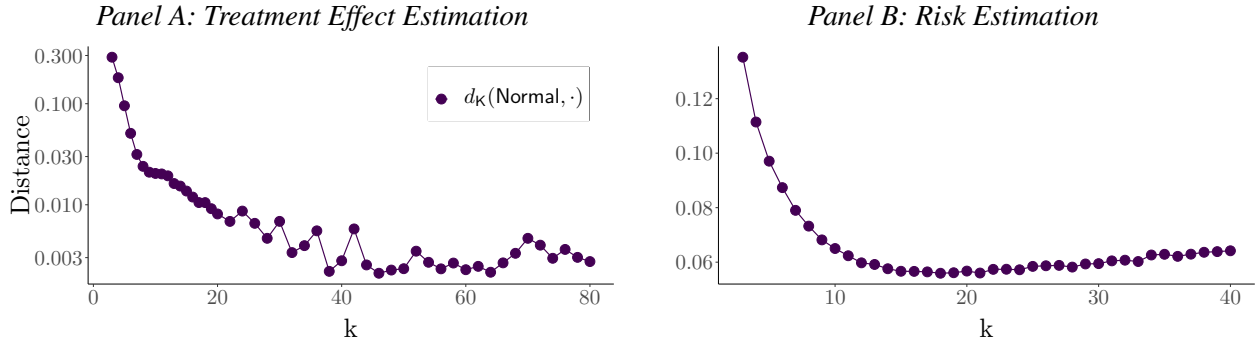
$$\left|P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \leq \xi \mid D\right\} - (1-\beta)\right| \lesssim \rho_{k,\varphi,b}(\xi, \beta \mid D) + \lambda_{k,\varphi,b}(\xi, \beta \mid D)\,, \tag{3.13}$$

holds with probability greater than $1-\delta$ *as* D *varies, for*

$$\rho_{k,\varphi,b}(\xi, \beta \mid D) = \frac{\xi}{z_{1-\beta/2}}\frac{(2-\varphi k-\varphi)^3(\Gamma^{(2)}_{k,\varphi,b})^{3/4}}{\delta\left(v_{1,k}\left(D\right)\right)^2} \quad \textit{and} \tag{3.14}$$

[11] To the best of our knowledge, the only Stein's method central limit theorem in the Kolmogorov distance, applicable to our setting, is given in Zhang (2022), which generalizes the argument of Shao and Zhang (2019). In Appendix D.4, we show that this central limit theorem, in conjunction with the bounds used to derive Theorem 3.2, implies a bound that does not reduce as either g or k increase.

FIGURE 4. Quality of Normal Approximation



Notes: Figure 4 displays measurements of the Kolmogorov distance (3.10) between a standard normal random variable and the studentized distribution of the cross-fit statistic $a\left(\mathsf{R}_{1,k}, D\right)$ conditioned on D over a range of values of k for our two applications to the Banerjee et al. (2015) data. This quantity is defined in the left hand side of the inequality (3.11). The y-axis of Panel A is displayed in a log-scale base 10. For the risk estimation application, we study the regression (1.1) for the treated outcome, i.e., w equal to one, for a fixed choice of λ. Further details on this figure are given in Appendix A.

$$\lambda_{k,\varphi,b}(\xi, \beta \mid D) = \left(\frac{\xi}{z_{1-\beta/2}} \frac{(2-\varphi k-\varphi)^4 \Gamma^{(1)}_{k,\varphi,b} (\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\delta^{3/2} (v_{1,k}\left(D\right))^{5/2}} \right)^{1/2}, \tag{3.15}$$

where a multiplicative factor that grows logarithmically as ξ shrinks has been omitted in writing (3.15) and an additional additive term that shrinks exponentially as ξ shrinks has been omitted in writing (3.13).

Remark 3.8. Recall the definition of the oracle stopping time $g^\star$ given in (2.7). The bound (3.13) is obtained by decomposing the approximation error into two terms. The first term involves the error in a normal approximation to the difference $a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)$ and contributes the term $\rho_{\varphi,k}(\xi, \beta \mid D)$. As $g^\star$ is deterministic conditional on D, a bound on this term follows from an argument very similar to the proof of Corollary 3.1. The second term involves the differences $a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}_{g^\star,k}, D)$ and $a(\mathsf{R}'_{\hat{g}',k}, D) - a(\mathsf{R}'_{g^\star,k}, D)$ and contributes the quantity $\lambda_{\varphi,k}(\xi, \beta \mid D)$. The key step in bounding these quantities involves deriving a high probability bound for the difference $\hat{g} - g^\star$. This follows by combining a concentration inequality analogous to Theorem 3.1 for the conditional variance estimator $\hat{v}_{g,k}(D)$ with a Bernstein-type maximal inequality, due to Steiger (1970). ■

Remark 3.9. The dependence of the bound (3.13) on ξ is sharp, at least up to the logarithmic factor. This follows from the examples given in Landers and Rogge (1976, 1988). Berry-Esseen type bounds on the coverage error of Anscombe-Chow-Robbins fixed-width sequential confidence intervals have been given in Csenki (1980), Mukhopadhyay (1981), and Callaert and Janssen (1981). Our result differs in several respects. Most importantly for our application, the dependence of our bound on the quantities k, φ, and $v_{1,k}(D)$ is explicit. We explore the structure of the bound through these terms in the following remark. Secondarily, we give an explicit characterization of the logarithmic factors in the bound by using Bernstein-type concentration inequalities, e.g., the bound given in Theorem 3.1, rather than moment bounds. ■

Remark 3.10. Restrict attention again to the case of cross-fitting and assume that the statistic of interest is sample stable. We will refer to the procedure that uses the oracle stopping time $g^\star$ in the place of $\hat{g}$ as the oracle procedure. Observe that the oracle stopping time $g^\star$ is the unique integer that satisfies

$$g^\star - 1 \leq 2v_{1,k}(D)\left(\frac{z_{1-\beta/2}}{\xi}\right)^2 \leq g^\star\,. \tag{3.16}$$

Replacing the variance $v_{1,k}(D)$ by the bound derived in Theorem 3.2, we get that $g^\star$ is proportional to

$$\frac{2}{k^2}\left(\frac{z_{1-\beta/2}}{\xi}\right)^2\,. \tag{3.17}$$

Thus, if $m^\star = kg^\star$ is the total number of splits used by the oracle procedure, then $m^\star$ is proportional to k^{-1} as k increases with all other quantities held fixed. In other words, the total number of splits used by the oracle procedure decreases rapidly as k increases. In the proof of Theorem 3.3, we give high probability bounds for $|g^\star - \hat{g}|$. Consequently, similar intuition will hold for Algorithm 1. That is, with high probability, the computational cost scales like k^{-1} as k increases.

On the other hand, by Theorem 3.2, the leading term $\lambda_{k,\varphi,b}(\xi,\beta \mid D)$ in the bound (3.13) is at least

$$\left(\frac{\xi}{z_{1-\beta/2}}\frac{1}{\delta^{3/2}(1-\varphi)}\frac{(\sigma^{(4,b-1)}_{\mathsf{train}})^{1/2}}{(\sigma^{(2,b-1)}_{\mathsf{train}})^{3/2}}\right)^{1/2}\,. \tag{3.18}$$

If $\sigma^{(4,b-1)}_{\mathsf{train}}$ and $\sigma^{(2,b-1)}_{\mathsf{train}}$ are proportional to their upper bounds (3.3), then the quantity (3.18) is of order

$$\left(\frac{\xi}{z_{1-\beta/2}}\frac{k}{\delta^{3/2}}\right)^{1/2}\,, \tag{3.19}$$

which is in turn of order $(g^\star)^{-1/4}$ by (3.17).

Comparison of the quantities (3.16) and (3.19) illustrates a fundamental trade-off in the performance of Algorithm 1 over the choice of k. Increased values of k reduce the conditional variance $v_{g,k}(D)$, and thereby, reduce the oracle sample size $g^\star$, easing computational cost. But as a consequence, the quality of the accuracy of the nominal error rate may deteriorate. In particular, at smaller values of g, the accuracy of the estimator $\hat{v}_{g,k}(D)$ for the conditional variance $v_{g,k}(D)$ will be lower, increasing the risk that $\hat{g}$ is smaller than $g^\star$. The increased risk of early stopping decreases the probability that a result is reproducible. It is desirable, then, to choose values for k and ξ such that the variance $v_{1,k}(D)$ is small enough to make Algorithm 1 computationally efficient, but not so small that the effective sample size $g^\star$ is too small for the nominal error rate β to be accurate. ■

4. PERFORMANCE

We now study the performance of Algorithm 1 in the applications to the Banerjee et al. (2015) data introduced in Section 1. Figure 5 displays measurements of reproducibility error

$$P\left\{\left|a(\mathsf{R}_{\hat{g},k},D) - a(\mathsf{R}'_{\hat{g}',k},D)\right| \geq \xi \mid D\right\} \tag{4.1}$$

when $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$ are obtained independently with Algorithm 1 for the risk estimation application. Figure A.2, given in Appendix A.3, displays analogous measurements for the treatment effect estimation application. We set the nominal error rate to $\beta = 0.1$ and vary the values of k and ξ.

Panel A displays measurements of the reproducibility error with the initialization g_{init} set equal to 2, the minimal feasible value. A transition is evident. For large values of ξ, the reproducibility error is monotonically decreasing in k. As ξ decreases, this pattern reverses and the reproducibility probability approaches the nominal error rate. This phenomenon is caused by severe early stopping at large values of ξ. In particular, at large values of ξ, there is a substantial probability that $\hat{g}$ is chosen to be a very small value, e.g., 2 or 3, often several orders of magnitude less than $g^\star$.[12] We document this behavior in Figure A.3, given in Appendix A.3.

This issue can be addressed through a very small increase in g_{init}. Panel B displays measurements of the reproducibility error with the initialization g_{init} increased to 10. The nominal error rate is now substantially more accurate over the full support of values for ξ considered. The moderate inaccuracy at larger values of ξ has two sources. First, there is still some risk of early stopping at values of g not much greater than 10. Second, if ξ is large relative to the conditional variance $v_{1,k}(D)$, ensuring that $\hat{g}$ is at least 10 can cause the reproducibility error to be close to 0. At moderate values of ξ, the trade off documented in Section 3 is apparent. As k increases, the effective sample size $g^\star$ decreases, reducing the accuracy of the nominal error rate. Figure A.3, given in Appendix A.3, demonstrates that this inaccuracy can be largely explained by the error in the approximation of $g^\star$ with $\hat{g}$. For small values of ξ, at the level of precision likely to be of practical interest, the nominal error rate is very accurate.

5. A Stein Representer, Concentration, and Reproducibility

In this section, we outline the proofs for results stated in Section 3. The primary difficulty is accommodating the dependence in the summands of $a(\mathsf{R}_{g,k}, D)$ across elements of the same cross-split. We tackle this problem through the method of exchangeable pairs (Stein, 1986).

Suppose that we are interested in studying the statistic $f(X)$, where X is random variable valued on the separable metric space $\mathcal{X}$. The method of exchangeable pairs has two ingredients. First, we need to construct a random variable X' such that (X, X') is an exchangeable pair. Second, we need to construct an antisymmetric function $F(X, X')$ such that

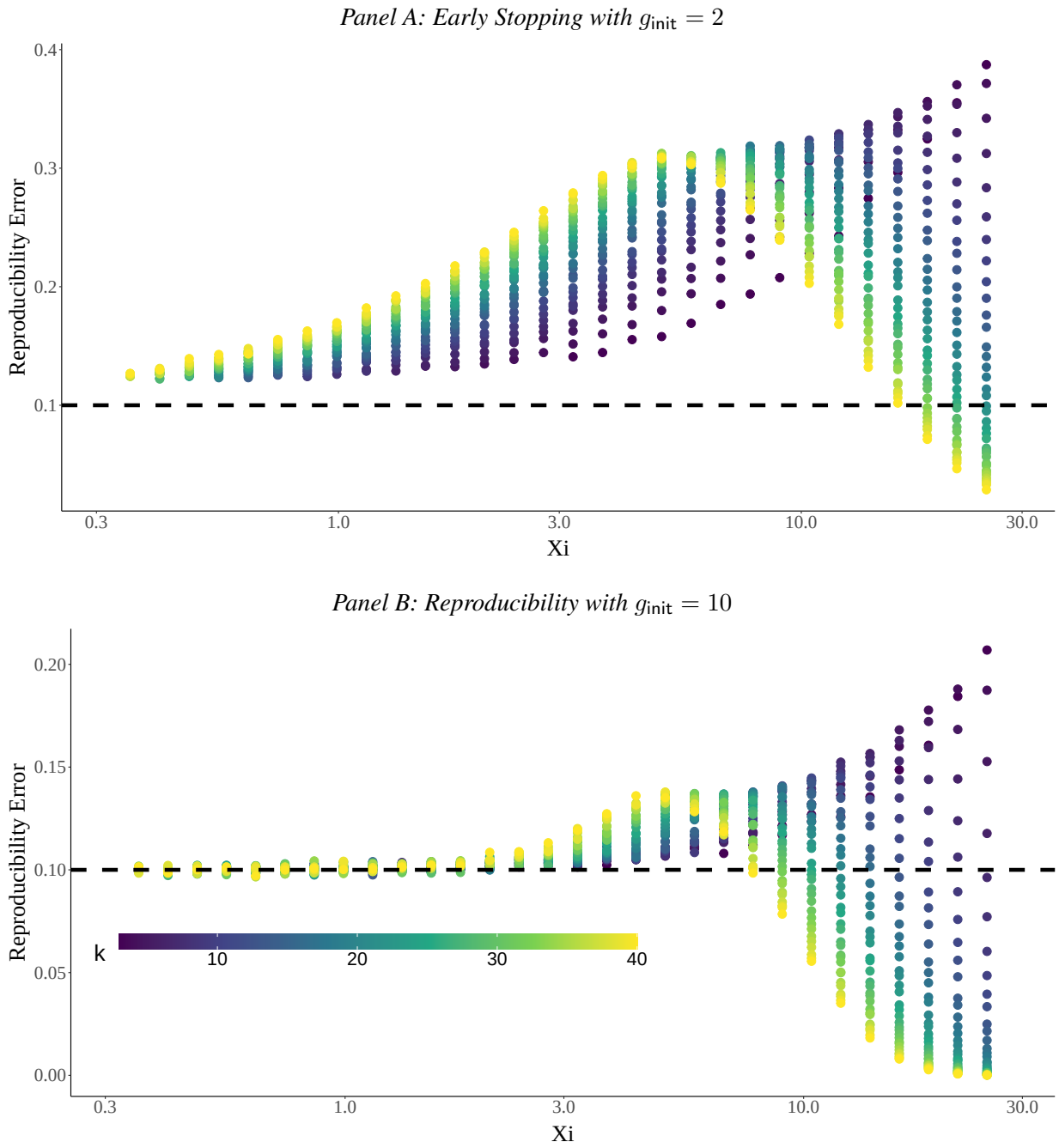
$$f(X) = \mathbb{E}\left[F\left(X, X'\right) \mid X\right] \tag{5.1}$$

almost surely. We will refer to the function $F(X, X')$ as a "Stein representer" for $f(X)$.

5.1 Constructing a Stein Representer. In many applications, the Stein representer (5.1) induced by a suitable exchangeable pair (X, X') can be derived in closed form. See e.g Chen et al. (2011) and Ross (2011). A closed form derivation is more challenging in our setting, where the collection $\mathsf{R}_{g,k}$ takes the place of the random variable X. To address this issue, we apply a method due to Chatterjee (2005) for constructing Stein representers through a pair of coupled Markov chains induced by an appropriately chosen exchangeable pair.

[12] In the risk estimation application, with $k = 10$, $g^\star$ is 54.4 when ξ is 10.4 and is 570.5 when ξ is 3.2.

FIGURE 5. Reproducibility Error in Risk Estimation



Notes: Figure 5 displays measurements of reproducibility error (4.1) for Algorithm 1 in the risk estimation application to the Banerjee et al. (2015) data. A dashed horizontal line is displayed at the nominal error rate $\beta = 0.1$. The x-axes vary the bound ξ and are displayed in a log-scale base 10. We study the regression (1.1) for the treated outcome, i.e., w equal to one, for a fixed choice of λ. To ease computation, we only display reproducibility error estimates for a (ξ, k) pair if $g^\star$ is less than 10^4. Further details on this figure are given in Appendix A.

Chatterjee's construction is founded on a observation, due to Stein (1986), that an exchangeable pair (X, X') induces a reversible Markov kernel K through

$$Kg(X) = \mathbb{E}\left[g(X') \mid X = x\right] \ ,$$

where g is any function satisfying $\mathbb{E}\left[|g(X)|\right] < \infty$. Suppose that $\{X_m\}_{m\geq 0}$ and $\{X'_m\}_{m\geq 0}$ are two Markov chains constructed with the kernel induced by (X, X') and coupled in a such way that the marginal distributions of X_m and X'_m depend only on the initial conditions X_0 and X'_0, respectively. Chatterjee makes the following observation. If there exists a constant C such that

$$\sum_{m=0}^{\infty} |\mathbb{E}\left[f(X_m) - f(X'_m) \mid X_0 = x, X'_0 = y\right]| \leq C \ , \tag{5.2}$$

for each x and y in $\mathcal{X}$, then the function

$$F(x, y) = \sum_{m=0}^{\infty} \mathbb{E}\left[f(X_m) - f(X'_m) \mid X_0 = x, X'_0 = y\right]$$

is a Stein representer for $f(X)$. See Paulin et al. (2013, 2016) for applications of this idea to the derivation of matrix concentration inequalities. To apply this construction to our setting, we have two tasks. First, we need to specify a suitable exchangeable pair. Second, we need to construct a pair of coupled Markov chains induced by this pair that satisfy the finiteness condition (5.2). Throughout, for a vector $x = (x_i)_{i=1}^n$ we let $(x_{-\ell}, y)$ denote the vector formed by replacing the ℓth component of x with y.

Our construction is premised on the observation that the random collection of splits $\mathsf{R}_{g,k}$ can be generated by a random collection of permutations. To see this, let $\mathcal{P}_n$ denote the set of permutations of the set $[n]$, treating each $\pi \in \mathcal{P}_n$ as a bijection from $[n]$ to $[n]$. Observe that each permutation $\pi \in \mathcal{P}_n$ can be associated with an element of $\mathcal{R}_{n,k,b}$, denoted by $\mathsf{r}_k(\pi) = (\mathsf{s}_1(\pi), \ldots, \mathsf{s}_k(\pi))$, through

$$\mathsf{s}_i(\pi) = \{\pi(k \cdot (i-1) + 1), \ldots, \pi(k \cdot (i-1) + b)\} .$$

If $\boldsymbol{\pi} = (\pi_i)_{i=1}^g$denotes a collection of permutations drawn independently and uniformly at random from $\mathcal{P}_n$, then the collection $\mathsf{R}_{g,k}(\boldsymbol{\pi}) = (\mathsf{r}_k(\pi_i))_{i=1}^g$ is equidistributed with the collection $\mathsf{R}_{g,k}$ defined in Section 2.

Now, we construct an exchangeable pair $(\boldsymbol{\pi}, \boldsymbol{\pi}')$, keeping in mind that our aim is to verify a condition of the form (5.2). For any permutation $\pi \in \mathcal{P}_n$ and indices $i, j \in [n]$, define the updated permutation

$$\hat{\pi}(i, j)(x) = \begin{cases} j, & x = i, \\ \pi(i), & x = \pi^{-1}(j), \\ \pi(x), & \text{otherwise.} \end{cases}$$

In other words, $\hat{\pi}(i, j)$ is identical to π, except that i maps to j and $\pi^{-1}(j)$ maps to $\pi(i)$. Let L be distributed uniformly on $[g]$ and let I and J be independently and uniformly distributed on $[n]$. Define the modified collection

$$\boldsymbol{\pi}' = (\pi_{-L}, \hat{\pi}_L(I, J)).$$

and observe that $(\boldsymbol{\pi}, \boldsymbol{\pi}')$ is an exchangeable pair.

With the exchangeable pair $(\boldsymbol{\pi}, \boldsymbol{\pi}')$ in place, we choose a coupled pair of Markov chains that it induces. For each $m \geq 1$, let L_m be distributed uniformly on $[g]$ and let I_m and J_m be distributed uniformly on $[n]$. Construct $(\boldsymbol{\pi}_m, \boldsymbol{\pi}'_m)$ from the pair $(\boldsymbol{\pi}_0, \boldsymbol{\pi}'_0) = (\boldsymbol{\pi}, \boldsymbol{\pi}')$ by setting

$$\boldsymbol{\pi}_m = (\pi_{m-1,-L_m}, \hat{\pi}_{m-1,L_m}(I_m, J_m)) \quad \text{and} \quad \boldsymbol{\pi}'_m = (\pi'_{m-1,-L_m}, \hat{\pi}'_{m-1,L_m}(I_m, J_m)) \tag{5.3}$$

recursively. In other words, the Markov chain $\boldsymbol{\pi}'_0$ is initialized by choosing one permutation in $\boldsymbol{\pi}_0$ and swapping one pair of indices. In the mth iteration of the Markov chain $(\boldsymbol{\pi}_m, \boldsymbol{\pi}'_m)_{m\geq 0}$, the permutations π_{m-1,L_m} and π'_{m-1,L_m} are selected and updated so that I_m now maps to J_m. As the iterations proceed, I_m will continue to map to the same index in the L_mth permutation in both collections. That is, in each iteration that a new permutation L and index I are selected, the two collections become more similar. Eventually, every L and I will have been selected, the two collections $\boldsymbol{\pi}_m$ and $\boldsymbol{\pi}'_m$ will be the same, and so $a(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D) - a(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D) = 0$. This will imply the finiteness (5.2). This argument is formalized in the proof of the following Lemma.

Lemma 5.1. *The function*

$$A\left(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D\right) = \sum_{m=0}^{\infty} \mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right)\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D\right],$$

is finite, antisymmetric, and satisfies the equality

$$\mathbb{E}\left[A\left(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D\right) \mid \boldsymbol{\pi}, D\right] = a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) \tag{5.4}$$

almost surely.

5.2 Proofs for Theorems 3.1 and 3.2 and Corollary 3.1. To obtain the concentration inequality (3.5) and Burkholder-Davis-Gundy inequality (3.7), we apply the following result due to Chatterjee (2005, 2007), which we have augmented to be applicable under a set of assumptions considered in Paulin et al. (2016).

Lemma 5.2. *Let $\mathcal{X}$ be a separable metric space and suppose that (X, X') is an exchangeable pair of $\mathcal{X}$-valued random variables. Let $f : \mathcal{X} \to \mathbb{R}$ be a square integrable function and let $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Stein representer for f. For each positive integer (r), define the quantities*

$$U_f^{(r)}(X) = \frac{1}{2}\mathbb{E}\left[\left(f(X) - f\left(X'\right)\right)^{2r} \mid X\right] \quad \text{and} \quad U_F^{(r)}(X) = \frac{1}{2}\mathbb{E}\left[F\left(X, X'\right)^{2r} \mid X\right]. \tag{5.5}$$

If there exist nonnegative constants u and s such that

$$U_f^{(1)}(X) \leq s^{-1}u \quad \text{and} \quad U_F^{(1)}(X) \leq su, \tag{5.6}$$

then the concentration inequality

$$P\left\{|f(X)| \geq \delta\right\} \leq 2\exp\left(-t^2/2u\right) \tag{5.7}$$

holds for all $t \geq 0$. Moreover, the moment inequality

$$\mathbb{E}\left[f(X)^{2r}\right] \leq (2r-1)^r \left(s\mathbb{E}\left[U_F^{(r)}(X)\right] + s^{-1}\mathbb{E}\left[U_f^{(r)}(X)\right]\right) \tag{5.8}$$

holds for all positive integers r for any positive constant s.

Throughout, to ease notation, we use the short hand

$$Z = (\mathsf{R}_{g,k}(\boldsymbol{\pi}), D) \quad \text{and} \quad Z' = (\mathsf{R}_{g,k}(\boldsymbol{\pi}'), D)\,.$$

The Markov chains $(Z_m)_{m\geq 0}$ and $(Z'_m)_{m\geq 0}$ are defined accordingly and we simplify $A(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D)$ to $A(Z, Z')$. To apply Lemma 5.2, we are required to develop bounds for the objects

$$U_a^{(r)}(Z) = \frac{1}{2}\mathbb{E}\left[\left(a(Z) - a\left(Z'\right)\right)^{2r} \mid Z\right] \quad \text{and} \quad U_A^{(r)}(Z) = \frac{1}{2}\mathbb{E}\left[A\left(Z, Z'\right)^{2r} \mid Z\right].$$

Our approach is based on the following Lemma, which combines a generalization of an idea due to Lemma 10.4 of Paulin et al. (2016) with a Markov type bound.

Lemma 5.3. *Let $r = 2^c$ for some positive integer c. Let $f : \mathcal{Z} \to \mathbb{R}$ be a function such that there exists a square integrable random variable W with*

$$\left(\sum_{m=0}^{\infty} \mathbb{E}\left[f(Z_m) - f\left(Z'_m\right) \mid Z_0 = Z, Z'_0 = Z'\right]\right)^r \leq W \tag{5.9}$$

almost surely. If the inequality

$$\mathbb{E}\left[\mathbb{E}\left[f(Z_m) - f(Z'_m) \mid Z_0 = Z, Z'_0 = Z'\right]^r \mid \boldsymbol{\pi}\right] \leq h_m^r \tag{5.10}$$

holds for each $m \geq 0$ and each collection $\boldsymbol{\pi}$, where $(h_m)_{m\geq 0}$ is a deterministic sequence of nonnegative numbers, then the inequalities

$$\frac{1}{2}\mathbb{E}\left[\left(f(Z) - f\left(Z'\right)\right)^r \mid Z\right] \leq \frac{h_0^r}{\delta} \quad \textit{and} \tag{5.11}$$

$$\frac{1}{2}\mathbb{E}\left[\left(\sum_{m=0}^{\infty} \mathbb{E}\left[f(Z_m) - f\left(Z'_m\right) \mid Z_0 = Z, Z'_0 = Z'\right]\right)^r \mid Z\right] \leq \frac{1}{\delta}\left(\sum_{m=0}^{\infty} h_m\right)^r \tag{5.12}$$

both hold with probability greater than $1 - \delta$ as D varies

To apply this Lemma, we begin by noting that the inequality

$$\left(\sum_{m=0}^{\infty} \mathbb{E}\left[a(Z_m) - a\left(Z'_m\right) \mid Z_0 = Z, Z'_0 = Z'\right]\right)^2 \lesssim g^2 n^4 \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T(\mathsf{s}, U, D) - T\left(\mathsf{s}', U, D\right)\right)^2 \tag{5.13}$$

follows from Lemma B.1, stated in Appendix B.2. Moreover, the right hand side of (5.13) is square integrable, as the fourth order split-stability $\zeta^{(4)}$ is finite by assumption. Deterministic bounds of the form (5.10) are obtained through the following Lemma.

Lemma 5.4. *Under Assumptions 3.1 and 3.2, for all integers $m \geq 0$ and $r \geq 1$, the inequality*

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[a(Z_m) - a(Z'_m) \mid Z_0 = Z, Z'_0 = Z'\right]^{2r} \mid \boldsymbol{\pi}\right] \\ &\leq 2^{4r}\left(1 - \frac{2}{gn^2}\right)^{2mr}\left(\frac{2n - bk - b}{gn^2}\right)^{2r} \Gamma_{k,\varphi,b}^{(r)}\end{aligned}$$

holds almost surely, where $\Gamma^{(r)}_{k,\varphi,b}$ *is defined in the statement of Theorem 3.2.*

Combining Lemmas 5.3 and 5.4, we have that

$$U^{(r)}_A(Z) \leq \frac{1}{\delta}\left(\frac{gn^2}{2}\right)^r \left(\frac{2^4\left(2n-bk-b\right)^2}{gn^2}\right)^r \Gamma^{(r)}_{k,\varphi,b}$$

and

$$U^{(r)}_a(Z) \leq \frac{1}{\delta}\left(\frac{2}{gn^2}\right)^r \left(\frac{2^4\left(2n-bk-b\right)^2}{gn^2}\right)^r \Gamma^{(r)}_{k,\varphi,b}$$

with probability $1-\delta$. Theorem 3.1 is obtained by setting $r=1$ and applying (5.7) of Lemma 5.2 with $s = gn^2/2$ and

$$u = \frac{1}{\delta}\frac{2^4\left(2n-bk-b\right)^2}{gn^2}\Gamma^{(1)}_{n,k,b}.$$

Similarly, Theorem 3.2 is obtained by applying (5.8) of Lemma 5.2 with $s = (gn^2/2)^r$.

The normal approximation error bound (3.11) is a corollary of Theorem 3.2. To see this, consider the centered statistic

$$a\left(Z\right) - \overline{a}\left(D\right) = \frac{1}{g}\sum_{\ell=1}^{g}\overline{a}\left(\mathsf{r}_\ell, D\right), \quad \text{where} \quad \overline{a}\left(\mathsf{r}_\ell, D\right) = \frac{1}{k}\sum_{i=1}^{k}\left(T(\mathsf{s}_{\ell,i}, D) - \overline{a}\left(D\right)\right) \tag{5.14}$$

for each ℓ in $[g]$. Conditional on the data D, the statistics $\overline{a}\left(\mathsf{r}_\ell, D\right)$ are independent, identically distributed, and mean zero. Moreover, we have that

$$\mathbb{E}\left[\left|\overline{a}\left(\mathsf{r}_\ell, D\right)\right|^3\right] \leq 3^2 2^6 \left(2-\varphi k-\varphi\right)^3 \left(\Gamma^{(2)}_{k,\varphi,b}\right)^{3/4} \tag{5.15}$$

by Hölder's inequality and Theorem 3.2. Hence, we find that

$$\frac{\mathbb{E}\left[\left|\overline{a}\left(\mathsf{r}_\ell, D\right)\right|^3 \mid D\right]}{g^{1/2}\left(v_{1,k}(D)\right)^{3/2}} \leq \frac{3^2 2^6}{\delta}\frac{\left(2-\varphi k-\varphi\right)^3}{\left(v_{1,k}(D)\right)^{3/2}}\frac{\left(\Gamma^{(2)}_{k,\varphi,b}\right)^{3/4}}{g^{1/2}}$$

holds with probability greater than $1-\delta$, by combining (5.15) with the Markov inequality. The proof then follows by the standard Berry-Esseen inequality. See e.g., Corollary 1 of Shevtsova (2011). ■

5.3 Proof of Theorem 3.3. The first step for verifying Theorem 3.3 is deriving the exponential rate of concentration for the variance estimator $\hat{v}\left(\mathsf{R}_{g,k}, D\right)$ defined in (2.4). This is obtained in the following Lemma, which follows from an argument very similar to the proof of Theorem 3.1.

Lemma 5.5. *Suppose that Assumptions 3.1 and 3.2 hold and that the data* D *are independent and identically distributed. If the eighth-order split-stability* $\zeta^{(8)}$ *is finite, then the conditional concentration inequality*

$$\log\frac{1}{4}P\left\{\left|\frac{\hat{v}(\mathsf{R}_{g,k}, D)}{v_{g,k}\left(D\right)} - 1\right| \geq t \mid D\right\} \lesssim -\frac{\delta\left(v_{1,k}(D)\right)^2}{\left(2-\varphi k-\varphi\right)^4}\frac{gt^2}{\Gamma^{(2)}_{k,\varphi,b}}$$

holds for all $t>0$ *with probability greater than* $1-\delta$ *as* D *varies.*

We focus our analysis on the error

$$|P\left\{a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D) \leq \xi \mid D\right\} - (1-\beta/2)\,|. \tag{5.16}$$

An analogous argument will yield the same bound for the lower tail. We begin by bounding (5.16) with quantities that will be easier to handle in isolation. Define the objects

$$U\left(\mathsf{R}_{g,k}, D\right) = \frac{1}{g^\star}\left(\sum_{i=1}^{g} \overline{a}\left(\mathsf{r}_{i,k}, D\right) - \sum_{i=1}^{g^\star} \overline{a}\left(\mathsf{r}_{i,k}, D\right)\right) \quad \text{and} \tag{5.17}$$

$$Q\left(\mathsf{R}_{g,k}, D\right) = \left(1 - \frac{g}{g^\star}\right)\left(a(\mathsf{R}_{g,k}, D) - \overline{a}\left(D\right)\right) \ . \tag{5.18}$$

The following Lemma bounds the error (5.16) in terms of the error in the normal approximation to the quantity $a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)$ and generic high probability bounds on (5.17) and (5.18).

Lemma 5.6. *Let W denote a standard normal random variable. Define the events*

$$\mathcal{U}_{k,\lambda}\left(D\right) = \left\{|U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D)| \leq \lambda\sqrt{2v_{g^\star,k}\left(D\right)}\right\} \quad \textit{and}$$

$$\mathcal{Q}_{k,\lambda}(D) = \left\{\left|Q(\mathsf{R}_{g,k}, D) - Q(\mathsf{R}'_{\hat{g}',k}, D)\right| \leq \lambda\sqrt{2v_{g^\star,k}\left(D\right)}\right\}.$$

The quantity (5.16) is bounded above by

$$d_K\left(\frac{a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)}{\sqrt{2v_{g^\star,k}\left(D\right)}}, W \mid D\right) + 2\lambda + \left(1 - P\left\{\mathcal{U}_{k,\lambda}\left(D\right) \cap \mathcal{Q}_{k,\lambda}(D) \mid D\right\}\right) \ , \tag{5.19}$$

where the Kolmogorov distance $d_K\left(\cdot,\cdot\right)$ is defined in (3.10).

Thus, it remains to give suitable bounds for the objects in (5.19). These are obtained in the following Lemma.

Lemma 5.7. *Let W denote a standard normal random variable. Recall the objects $\rho_{\varphi,k}\left(\xi, \beta \mid D\right)$ and $\lambda_{\varphi,k}\left(\xi, \beta \mid D\right)$ defined in the statement of Theorem 3.3. Suppose that Assumptions 3.1 and 3.2 hold, that the data D are independent and identically distributed, and that the eighth-order split stability $\zeta^{(8)}$ is finite.*
***(i)** The Berry-Esseen inequality*

$$d_K\left(\frac{a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)}{\sqrt{2v_{g^\star,k}\left(D\right)}}, W \mid D\right) \lesssim \rho_{k,\varphi,b}\left(\xi, \beta \mid D\right)$$

is satisfied with probability greater than $1-\delta$ as D varies.
***(ii)** The conditional concentration inequality*

$$P\left\{\frac{|U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D)|}{\sqrt{2v_{g^\star,k}\left(D\right)}} \geq \lambda_{k,\varphi,b}\left(\xi, \beta \mid D\right)\tilde{\lambda}_{k,\varphi,b}\left(D\right) \mid D\right\} \lesssim \rho_{k,\varphi,b}\left(\xi, \beta \mid D\right)$$

holds with probability greater than $1-\delta$ as D varies, where $\tilde{\lambda}_{\varphi,k}\left(D\right)$ is a term that grows logarithmically as ξ shrinks and is characterized explicitly in the proof.

(iii) *The conditional concentration inequality*

$$P\left\{\frac{|Q(\mathsf{R}_{\hat{g},k},D)-Q(\mathsf{R}'_{\hat{g}',k},D)|}{\sqrt{2v_{g^\star,k}(D)}}\geq\lambda_{k,\varphi,b}(\xi,\beta\mid D)\right\}\lesssim\rho_{k,\varphi,b}(\xi,\beta\mid D)+\tilde{\rho}_{k,\varphi,b}(\xi,\beta\mid D)$$

holds with probability greater than $1-\delta$ *as* D *varies, where* $\tilde{\rho}_{\varphi,k}(\xi,\beta\mid D)$ *is a term that shrinks exponentially as* ξ *shrinks and is characterized explicitly in the proof.*

The proof is complete by substituting the bounds specified in Lemma 5.7 into Lemma 5.6. ■

6. Conclusion

We propose a method for sequentially aggregating randomized statistics to ensure that residual randomness is small. The method is applicable under very general conditions. By restricting the set of statistics under consideration, we give a non-asymptotic analysis of the performance of the method. We have two main findings. First, aggregation of sample-split statistics with cross-splitting reduces residual randomness at a much faster rate than for independent splitting. Second, cross-splitting does not necessarily improve the quality of the nominal error rate of the procedure. Thus, users navigate a tradeoff. Finer cross-splitting reduces the conditional variance of the aggregate statistic, reducing the computation needed to achieve a given bound on the residual randomness. But as a consequence, the quality of the nominal error rate may have also been reduced, limiting the user's ability to give a high quality estimate of the probability that their results are reproducible.

Our computational analysis has highlighted the substantial practical importance of choosing the initialization parameter g_{init} to be sufficiently large. By contrast, our theoretical analysis is agnostic to this issue. In the context of fixed-length sequential confidence intervals, Mukhopadhyay and Datta (1996) study the choice of g_{init}. They show that if g_{init} is chosen appropriately, a non-uniform asymptotic approximation to the coverage error, based on Edgeworth expansions due to Aras and Woodroofe (1993), has a leading term of order $O(\xi^2)$. This is a substantial improvement on the $O(\xi^{1/2})$ rate given in Theorem 3.3. Further research should give a non-asymptotic treatment of the choice of the parameter g_{init}, aimed at assessing whether the $O(\xi^{1/2})$ rate can be improved through further assumptions and a principled choice of g_{init}.

REFERENCES

Abou-Moustafa, K. and Szepesvári, C. (2019). An exponential efron-stein inequality for ℓ_q stable learning rules. In *Algorithmic Learning Theory*, pages 31–63. PMLR.

Anscombe, F. J. (1952). Large-sample theory of sequential estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 600–607. Cambridge University Press.

Aras, G. and Woodroofe, M. (1993). Asymptotic expansions for the moments of a randomly stopped average. *The annals of Statistics*, pages 503–519.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40 – 79.

Austern, M. and Zhou, W. (2020). Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., and Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236):1260799.

Bates, S., Hastie, T., and Tibshirani, R. (2023). Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, pages 1–12.

Bayle, P., Bayle, A., Janson, L., and Mackey, L. (2020). Cross-validation confidence intervals for test error. *Advances in Neural Information Processing Systems*, 33:16339–16350.

Beran, R. and Millar, P. W. (1987). Stochastic estimation and testing. *The Annals of Statistics*, 15(3):1131–1154.

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics*, pages 647–671.

Boucheron, S., Lugosi, G., and Massart, P. (2003). Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526.

Burkholder, D. L. (1973). Distribution function inequalities for martingales. *The Annals of Probability*, 1(1):19–42.

Callaert, H. and Janssen, P. (1981). The convergence rate of fixed-width sequential confidence intervals for the mean. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 211–219.

Celisse, A. and Guedj, B. (2016). Stability revisited: new generalisation bounds for the leave-one-out. *arXiv preprint arXiv:1608.06412*.

Chatterjee, S. (2005). Concentration inequalities with exchangeable pairs (Ph. D. thesis). *arXiv preprint math/0507526*.

Chatterjee, S. (2007). Stein's method for concentration inequalities. *Probability Theory and Related Fields*, 1(138):305–321.

Chen, J. and Ritzwoller, D. M. (2023). Semiparametric estimation of long-term treatment effects. *Journal of Econometrics*, 237(2):105545.

Chen, L. H., Goldstein, L., and Shao, Q.-M. (2011). *Normal approximation by Stein's method*, volume 2. Springer.

Chen, Q., Syrgkanis, V., and Austern, M. (2022). Debiased machine learning without sample-splitting for stable estimators. *Advances in Neural Information Processing Systems*, 35:3096–3109.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1).

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018b). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research.

Chetverikov, D., Liao, Z., and Chernozhukov, V. (2021). On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317.

Chetverikov, D. and Sørensen, J. R.-V. (2021). Analytic and bootstrap-after-cross-validation methods for selecting penalty parameters of high-dimensional m-estimators. *arXiv preprint arXiv:2104.04716*.

Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics*, 36(2):457–462.

Cornec, M. (2010). Concentration inequalities of the cross-validation estimate for stable predictors. *arXiv preprint arXiv:1011.5133*.

Csenki, A. (1980). On the convergence rate of fixed-width sequential confidence intervals. *Scandinavian Actuarial Journal*, 1980(2):107–111.

Dembo, A. (2021). Probability theory: Stat310/math230.

DiCiccio, C. and Romano, J. P. (2019). *Multiple data splitting for testing*. Department of Statistics, Stanford University Stanford, CA.

DiCiccio, C. J., DiCiccio, T. J., and Romano, J. P. (2020). Exact tests via multiple data splitting. *Statistics & Probability Letters*, 166:108865.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Du, J.-H., Patil, P., Roeder, K., and Kuchibhotla, A. K. (2023). Extrapolated cross-validation for randomized ensembles. *arXiv preprint arXiv:2302.13511*.

Dunn, R., Ramdas, A., Balakrishnan, S., and Wasserman, L. (2023). Gaussian universal likelihood ratio testing. *Biometrika*, 110(2):319–337.

Elisseeff, A., Evgeniou, T., Pontil, M., and Kaelbing, L. P. (2005). Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1).

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., and Qian, J. (2021). Package 'glmnet'. *CRAN R Repositary*, 595.

Guo, F. R. and Shah, R. D. (2023). Rank-transformed subsampling: inference for multiple data splitting and exchangeable p-values. *arXiv preprint arXiv:2301.02739*.

Guo, W. and Romano, J. P. (2017). Analysis of error control in large scale two-stage multiple hypothesis testing. *arXiv preprint arXiv:1703.06336*.

Gut, A. (2009). *Stopped random walks*. Springer.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR.

Hickernell, F. J., Jiang, L., Liu, Y., and Owen, A. B. (2013). Guaranteed conservative fixed width confidence intervals via monte carlo sampling. In *Monte Carlo and Quasi-Monte Carlo Methods 2012*, pages 105–128.

Springer.

Kale, S., Kumar, R., and Vassilvitskii, S. (2011). Cross-validation and mean-square stability. In *ICS*, pages 487–495.

Klaassen, C. A. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, 15(4):1548–1562.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Kumar, R., Lokshtanov, D., Vassilvitskii, S., and Vattani, A. (2013). Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pages 27–35. PMLR.

Landers, D. and Rogge, L. (1976). The exact approximation order in the central-limit-theorem for random summation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 36(4):269–283.

Landers, D. and Rogge, L. (1988). Sharp orders of convergence in the random central limit theorem. *Journal of Approximation Theory*, 53(1):86–111.

Lei, J. (2020). Cross-validation with confidence. *Journal of the American Statistical Association*, 115(532):1978–1997.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Mukhopadhyay, N. (1981). Convergence rates of sequential confidence intervals and tests for the mean of a u-statistic. *Communications in Statistics-Theory and Methods*, 10(21):2231–2244.

Mukhopadhyay, N. and Datta, S. (1996). On sequential fixed-width confidence intervals for the mean and second-order expansions of the associated coverage probabilities. *Annals of the Institute of Statistical Mathematics*, 48:497–507.

Nadeau, C. and Bengio, Y. (1999). Inference for the generalization error. *Advances in Neural Information Processing Systems*, 12.

Paulin, D., Mackey, L., and Tropp, J. A. (2013). Deriving matrix concentration inequalities from kernel couplings. *arXiv preprint arXiv:1305.0612*.

Paulin, D., Mackey, L., and Tropp, J. A. (2016). Efron-Stein inequalities for random matrices. *The Annals of Probability*, 44(5):3431 – 3473.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Science & Business Media.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601.

Ramdas, A. and Manole, T. (2023). Randomized and exchangeable improvements of markov's, chebyshev's and chernoff's inequalities. *arXiv preprint arXiv:2304.02611*.

Rényi, A. (1957). On the asymptotic distribution of the sum of a random number of independent random variables. *Acta Math*, 8:193–199.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Ross, N. (2011). Fundamentals of Stein's method. *Probability Surveys*, 8:210–293.

Rüger, B. (1978). Das maximale signifikanzniveau des tests:"lehne h_o ab, wenn k unter n gegebenen tests zur ablehnung führen". *Metrika*, 25:171–178.

Shao, Q.-M. and Zhang, Z.-S. (2019). Berry–Esseen bounds of normal and nonnormal approximation for unbounded exchangeable pair. *The Annals of Probability*, 47(1):61–108.

Shevtsova, I. (2011). On the absolute constants in the berry-esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*.

Steiger, W. (1970). Bernstein's inequality for martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 16(2):104–106.

Stein, C. (1986). Approximate computation of expectations. *IMS Lecture Notes—Monograph Series*, 7.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tse, T. and Davison, A. C. (2022). A note on universal inference. *Stat*, 11(1):e501.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Zhang, Z., Lee, S., and Dobriban, E. (2023). A framework for statistical inference via randomized algorithms. *arXiv preprint arXiv:2307.11255*.

Zhang, Z.-S. (2022). Berry–esseen bounds for generalized u-statistics. *Electronic Journal of Probability*, 27:1–36.

Supplemental Appendix to:

Reproducible Aggregation of Sample-Split Statistics*

David M. Ritzwoller
Stanford University

Joseph P. Romano
Stanford University

Contents

Date: December 8, 2023

APPENDIX A. DETAILS CONCERNING DATA AND SIMULATIONS

A.1 Data. Banerjee et al. (2015) study randomized evaluations of several similar poverty-alleviation programs implemented by BRAC, a large non-governmental organization. The data from Banerjee et al. (2015) were acquired from `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NHIXNT` on September 10, 2021. See Appendix D of Chen and Ritzwoller (2023) for further details concerning the cleaning of these data. Our discussion in this subsection uses language similar to the language used to discuss the data there.

The programs studied by Banerjee et al. (2015) randomly allocated productive assets (typically livestock) to participating households and measured economic outcomes two and three years later. We restrict attention to data from the program evaluation in Pakistan. In the sample considered in this paper, there are 854 households. Of these households, 446 were randomly assigned to participate in the program. In both the treatment effect and risk estimation applications considered throughout the paper, the outcome of interest Y_i is the total monthly consumption for each household. The covariate vector X_i collects measurements of 20 pretreatment variables. These variables can be grouped into five categories: consumption, food security, assets, finance, and income and revenue. The consumption variables are pretreatment total monthly consumption and total monthly consumptions on food, non-food, and durable commodities. The food-security variables are an index for overall food-security and five binary variables indicating different aspects of food security (e.g., did a child in the household skip a meal). The assets variables are value of assets, value of productive assets, and value of the household's assets, each aggregated in two ways. The financial variables are total amount of formal and informal loans outstanding and total value of the household's savings. The income and revenue variables are income from agriculture, income from business, income from paid labor, revenue from animals, and self-assessed perception of economic status. We refer the reader to the appendix of Banerjee et al. (2015) for further information on the construction of these variables.

A.2 Simulations. Throughout the paper, we consider three Lasso regressions

$$\hat{\beta}_w(\lambda) = \arg\min_{\beta} \left\{ \sum_{i \in \tilde{\mathbf{s}}} \mathbb{I}\{W_i = w\}(Y_i - \beta^\top X_i)^2 + \lambda \|\beta\|_1 \right\}, \quad w \in \{0,1\}, \quad \text{and} \tag{A.1}$$

$$\tilde{\beta}(\lambda) = \arg\min_{\beta} \left\{ \sum_{i \in \tilde{\mathbf{s}}} (W_i - \beta^\top X_i)^2 + \lambda \|\beta\|_1 \right\} \tag{A.2}$$

which produce the nuisance parameter estimates

$$\hat{\mu}_w(X_i) = \hat{\beta}_w(\lambda)^\top X_i \quad \text{and} \quad \hat{\pi}(X_i) = \tilde{\beta}(\lambda)^\top X_i \,.$$

Estimates of the average treatment effect are formed with the estimator (1.2). Figure A.1 displays average cross-validated risk estimates, over one million cross-splits, for the regressions (A.1) and (A.2) as k is varied. For all figures, except for Figure 1, we choose λ based off of the estimates reported in Figure A.1.

Specifically, for the two outcome regressions (A.1). we choose the value of λ that minimizes the average 10-fold cross-validated mean-squared error.[13] For the propensity score regression, the average k-fold mean-squared error is monotonically decreasing in λ. To encourage some selection of covariates, we set the value of λ sub-optimally, at approximately 0.02.

A.2.1 Figure 1. To emulate the way that cross-splitting is often applied in practice, for the purpose of Panels A and B of Figure 1, we choose λ by implementing 10-fold cross validation once. We use one hundred thousand cross-splits to construct each Panel.

A.2.2 Figure 2. For the purpose of Figure 2, we reduce the sample under consideration so that the total sample size has many perfect divisors. In particular, for the treatment effect application, we randomly reduce the sample of 854 households to a sample of 800 households. For the risk estimation application, we randomly reduce the sample of 446 treated households to 400 treated households. We consider the set of values of b that perfectly divide 800 and 400 respectively. For each value of b, we generate one million sample sample-splitsplits. In both applications, to address some numerical instability in the estimation of $\phi_{n,b}(D)$ and $\gamma_{n,b}(D)$ we drop the 0.25% of simulation draws with the largest and smallest values of $T(\mathsf{s}, D)$.

A.2.3 Figure 3. One million sample-splits are used to estimate $\sigma^{(2)}_{\mathsf{valid}}$ and $\sigma^{(2,b-1)}_{\mathsf{train}}$ for each value of b considered in Figure 3. In particular, for a given value of b, we draw one million sample-splits s and compute $T(\mathsf{s}, D)$. For each sample-split s, we draw a random index I in s and compute

$$\left(\psi(D_I, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) - \psi(D'_I, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)\right)^2 \ , \tag{A.3}$$

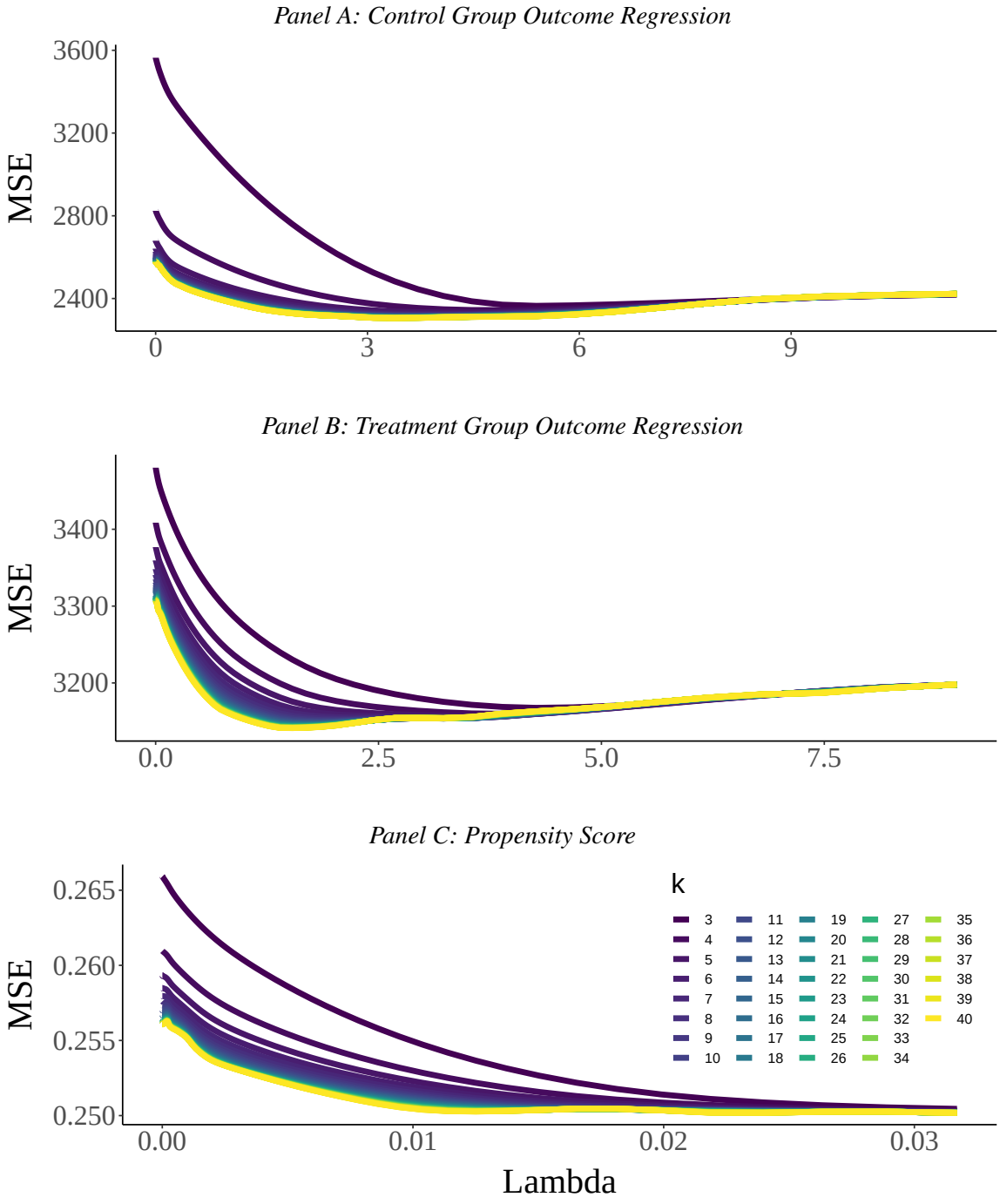
where D'_I is a random data point drawn from the uniformly from the data D.[14] The quantity $\sigma^{(2)}_{\mathsf{valid}}$ is estimated by averaging the quantity (A.3) over the one million sample-splits s. The estimate of $\sigma^{(2,b-1)}_{\mathsf{train}}$ is constructed similarly, where now we replace $b-1$ random data points in $\tilde{\mathsf{s}}$ with $b-1$ observations drawn uniformly from the data D.

A.2.4 Figure 4. The total number of simulation draws used in the measurements displayed in Figure 4 varies by the value of k. In the treatment effect estimation application, one million cross-split estimates are drawn for each value of k less than or equal to 20. For each value of k greater than 20, five hundred thousand cross-splits are used. Similarly, in the risk estimation application, one million cross-splits are used for k less than or equal to 20 and five hundred thousand cross-splits are used for k greater than 20. To address issues with numerical instability, when computing the variance $v_{1,k}(D)$ in the left-hand side of the Berry-Esseen inequality (3.11), we drop the 0.1% of simulation draws with the largest and smallest values of $a(\mathsf{R}_{1,k}, D)$.

[13] The chosen values of λ are equal to approximately 3.38 and 1.85 for the untreated and treated outcome regression, respectively.

[14] In the case of the risk estimation application to the treated outcome regression, the data set D is composed of only the data for the treated households.

FIGURE A.1. Mean Squared Error for Nuisance Parameter Estimates



Notes: Figure A.1 displays the average of cross-validated risk estimate, taken over one million cross-splits, for the regressions (A.1) and (A.2) implemented with the Banerjee et al. (2015) data. Each line displays this estimate for a different value of k. We consider the values of λ reported by the "glmnet" R software package (Friedman et al., 2021).

A.2.5 Figure 5. The measurements displayed in this figure are obtained by first computing one million replicates of $a(\mathsf{R}_{1,k}, D)$ for each value of k. For each value of k and and ξ, we implement Algorithm 1 twice fifty thousand times by sampling with replacement from the replicates of $a(\mathsf{R}_{1,k}, D)$. The measurements of the reproducibility error give the proportions of time that these two measurements differ by more than ξ.

A.3 Additional Figures and Discussion. Figure A.2 displays measurements of reproducibility error (4.1) when $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$ are obtained independently with Algorithm 1 for the treatment effect application to the Banerjee et al. (2015) data. Again, we set the nominal error rate to $\beta = 0.1$ and vary the values of k and ξ. These measurements are qualitatively similar to the results for the risk estimation application displayed in (5).

The details of the simulation are also similar. We first compute one million replicates of $a(\mathsf{R}_{1,k}, D)$ for each value of k. However, for numerical stability, we remove the 0.001% of replicates with the largest and smallest values of $a(\mathsf{R}_{1,k}, D)$. These extreme values are caused by the division by estimates of the propensity score in the formula (1.2). For each value of k and and ξ, we implement Algorithm 1 twice fifty thousand times by sampling with replacement from the replicates of $a(\mathsf{R}_{1,k}, D)$. To ease computation, we do not display reproducibility error estimates for a (ξ, k) pair if k is less than 10 and ξ is less than 0.03. We reduce the number of replications from fifty thousand to ten thousand for all estimates with k less than or equal to 7.

Figure A.3 displays the quantiles of the distribution of the sample size discrepancy

$$\hat{g}/g^\star - 1 \tag{A.4}$$

for the risk estimation application to the Banerjee et al. (2015) data. The 95th and 5th quantiles of the distribution of this difference are displayed with triangles and dots, respectively. Panel A gives results for g_{init} set equal to 2. The high probability of early stopping at large values of ξ is evident. Interestingly, there appears to be a phase transition as ξ decreases, wherein early stopping becomes much less likely. Panel B gives results for g_{init} set equal to 10. Here, the features that characterize the measurements in Panel A only occur at very large values of ξ. That is, the problem of early stopping has largely been resolved. At moderate and small values of ξ, the variance of $\hat{g}$ around $g^\star$ increases as k increases.

Appendix B. Proofs for Results Stated in the Main Text

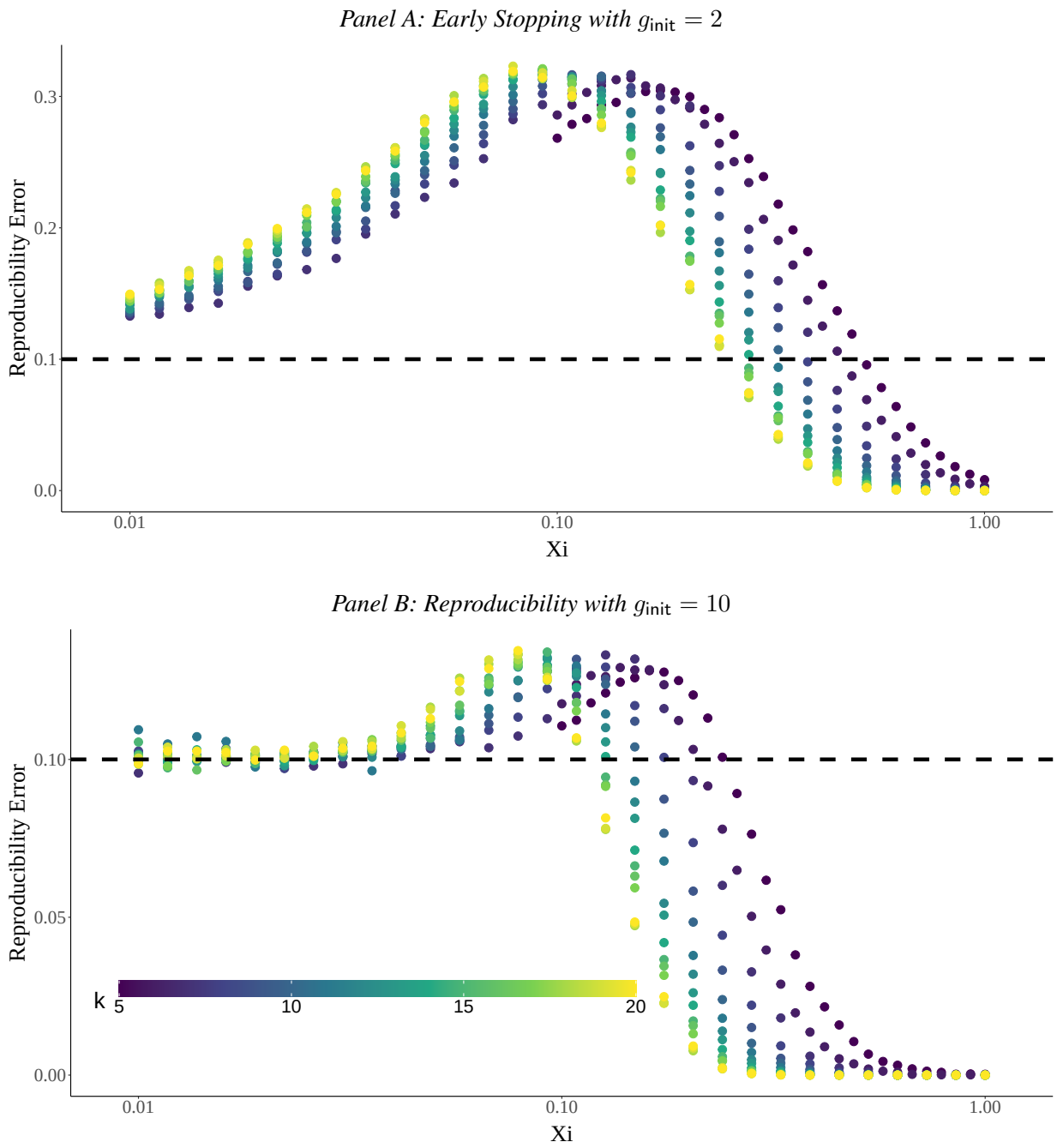
B.1 Proof of Theorem 2.1. For each collection $\mathsf{r}_k = (\mathsf{s}_j)_{j=1}^k$ in $\mathcal{R}_{n,k,b}$, we write

$$\overline{a}(\mathsf{r}_k, D) = \frac{1}{k}\sum_{j=1}^{k} T(\mathsf{s}_j, D) - \mathbb{E}\left[T(\mathsf{s}_j, D) \mid D\right].$$

Define the conditional variance

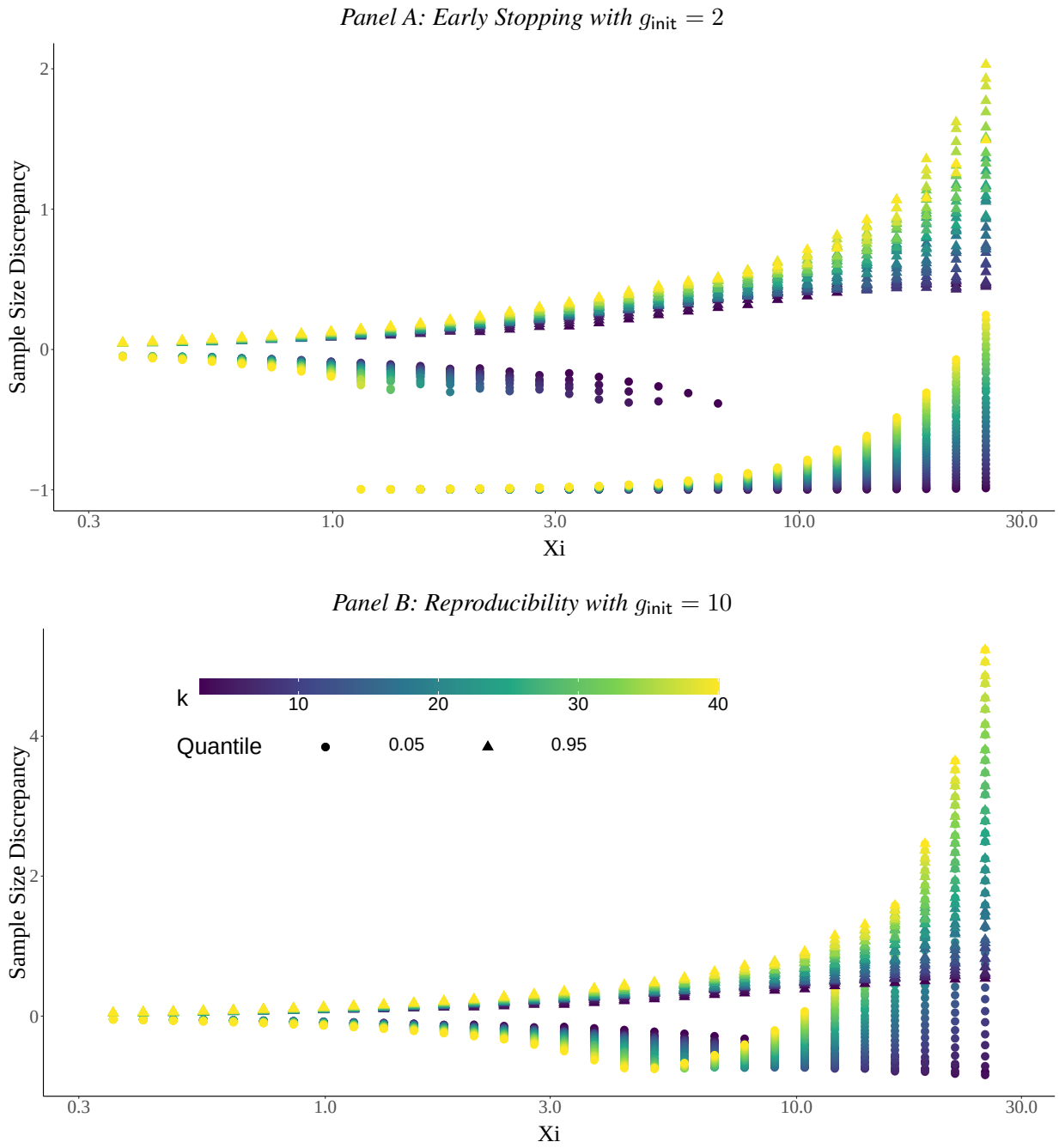
$$v_{g,k}(D) = \mathrm{Var}(a(\mathsf{R}_{g,k}, D) \mid D)$$

FIGURE A.2. Reproducibility Error in Treatment Effect Estimation



Notes: Figure A.2 displays measurements of reproducibility error (4.1) for Algorithm 1 in the treatment effect estimation application to the Banerjee et al. (2015) data. A dashed horizontal line is displayed at the nominal error rate $\beta = 0.1$. The x-axes vary the bound ξ and are displayed in a log-scale base 10. To ease computation, we do not display reproducibility error estimates for a (ξ, k) pair if k is less than 10 and ξ is less than 0.03. We reduce the number of simulation replications from 50,000 to 10,000 for all estimates with k less than or equal to 7.

FIGURE A.3. Sample Size Discrepancy in Risk Estimation



Notes: Figure A.3 displays quantiles of measurements of sample size discrepancy (A.4) for Algorithm 1 in the risk estimation application to the Banerjee et al. (2015) data. The 95th and 5th quantiles of the sample size error are displayed with triangles and circles, respectively. We set the error rate at $\beta = 0.1$. The x-axes vary the bound ξ and are displayed in a log-scale base 10. We study the regression (1.1) for the treated outcome, i.e., w equal to one, for a fixed choice of λ. To ease computation, we only display reproducibility error estimates for a (ξ, k) pair if $g^\star$ is less than 10^4.

and the oracle stopping time

$$g^\star = \arg\min_{g \geq 2} \left\{ v_{g,k}(D) \leq \frac{1}{2} \left(\frac{\xi}{z_{1-\beta/2}} \right)^2 \right\}.$$

Observe that $v_{g,k}(D) = (1/g) \cdot v_{1,k}(D)$ as the collections of cross-splits are drawn independently and identically. Similarly, we have that

$$\frac{\hat{v}(\mathsf{R}_{g,k}, D)}{v_{g,k}(D)} = \frac{g\hat{v}(\mathsf{R}_{g,k}, D)}{v_{1,k}(D)} \xrightarrow{\text{a.s}} 1 \tag{B.1}$$

as $g \to \infty$, by the strong law of large numbers. Observe that

$$\frac{\hat{g} \cdot \hat{v}(\mathsf{R}_{\hat{g},k}, D)}{v_{1,k}(D)} \leq \frac{\hat{g}}{2v_{1,k}(D)} \left(\frac{\xi}{z_{1-\beta/2}} \right)^2 \leq \frac{\hat{g} \cdot \hat{v}(\mathsf{R}_{\hat{g}-1,k}, D)}{v_{1,k}(D)}, \tag{B.2}$$

by definition. Thus, as $\hat{g} \to \infty$ and

$$g^\star \left(2v_{1,k}(D) \left(\frac{z_{1-\beta/2}}{\xi} \right)^2 \right) \to 1$$

as $\xi \to 0$, we have that

$$\hat{g}/g^\star \xrightarrow{\text{a.s}} 1 \tag{B.3}$$

as $\xi \to 0$ by (B.1) and (B.2).

Define the objects

$$U\left(\mathsf{R}_{g,k}, D\right) = \frac{1}{g^\star} \left(\sum_{i=1}^{g} \overline{a}\left(\mathsf{r}_{i,k}, D\right) - \sum_{i=1}^{g^\star} \overline{a}\left(\mathsf{r}_{i,k}, D\right) \right) \quad \text{and} \tag{B.4}$$

$$Q\left(\mathsf{R}_{g,k}, D\right) = \left(1 - \frac{g}{g^\star}\right) \left(a(\mathsf{R}_{g,k}, D) - \mathbb{E}\left[T(\mathsf{s}_j, D) \mid D\right]\right) \tag{B.5}$$

Observe that we can decompose

$$\begin{aligned}
&\left(a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right) / \sqrt{2v_{g^\star,k}(D)} \\
&= \sqrt{\frac{g^\star}{2v_{1,k}(D)}} \left(\frac{1}{g^\star} \sum_{i=1}^{g^\star} \left(\overline{a}\left(\mathsf{r}_{i,k}, D\right) - \overline{a}\left(\mathsf{r}'_{i,k}, D\right) \right) \right) \\
&+ \sqrt{\frac{g^\star}{2v_{1,k}(D)}} \left(\frac{1}{\hat{g}} \sum_{i=1}^{\hat{g}} \overline{a}\left(\mathsf{r}_{i,k}, D\right) - \frac{1}{g^\star} \sum_{i=1}^{g^\star} \overline{a}\left(\mathsf{r}_{i,k}, D\right) \right) \\
&- \sqrt{\frac{g^\star}{2v_{1,k}(D)}} \left(\frac{1}{\hat{g}'} \sum_{i=1}^{\hat{g}'} \overline{a}\left(\mathsf{r}'_{i,k}, D\right) - \frac{1}{g^\star} \sum_{i=1}^{g^\star} \overline{a}\left(\mathsf{r}'_{i,k}, D\right) \right) \\
&= \sqrt{\frac{g^\star}{2v_{1,k}(D)}} \left(a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)\right) + \sqrt{\frac{g^\star}{2v_{1,k}(D)}} \left(U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D)\right)
\end{aligned}$$

$$+\sqrt{\frac{g^\star}{2v_{1,k}(D)}}\left(Q(\mathsf{R}_{\hat{g},k},D)-Q(\mathsf{R}'_{\hat{g}',k},D)\right). \tag{B.6}$$

First, we have that

$$\sqrt{\frac{g^\star}{2v_{1,k}(D)}}\left(Q(\mathsf{R}_{\hat{g},k},D)-Q(\mathsf{R}'_{\hat{g}',k},D)\right)=o_p(1)$$

as $\xi\to 0$ by (B.3). To handle the terms involving (B.4), fix some $\varepsilon>0$ and observe that

$$\begin{aligned}
&P\left\{\left|\sum_{i=1}^{\hat{g}}a(\mathsf{r}_{i,k},D)-\sum_{i=1}^{g^\star}a(\mathsf{r}_{i,k},D)\right|>\varepsilon\sqrt{g^\star}\mid D\right\}\\
&\leq P\left\{\left|\sum_{i=1}^{\hat{g}}a(\mathsf{r}_{i,k},D)-\sum_{i=1}^{g^\star}a(\mathsf{r}_{i,k},D)\right|>\varepsilon\sqrt{g^\star},\hat{g}\in\left[g^\star\left(1-\varepsilon^3\right),g^\star\left(1+\varepsilon^3\right)\right]\mid D\right\}\\
&+P\left\{\left|\sum_{i=1}^{\hat{g}}a(\mathsf{r}_{i,k},D)-\sum_{i=1}^{g^\star}a(\mathsf{r}_{i,k},D)\right|>\varepsilon\sqrt{g^\star},\hat{g}'\notin\left[g^\star\left(1-\varepsilon^3\right),g^\star\left(1+\varepsilon^3\right)\right]\mid D\right\}\\
&\leq P\left\{\max_{g^\star(1-\varepsilon^3)\leq g\leq g^\star}\left|\sum_{i=1}^{g}a(\mathsf{r}_{i,k},D)\right|>\varepsilon\sqrt{g^\star}\mid D\right\}\\
&+P\left\{\max_{g^\star\leq g\leq(1+\varepsilon^3)g^\star}\left|\sum_{i=1}^{g}a(\mathsf{r}_{i,k},D)\right|>\varepsilon\sqrt{g^\star}\mid D\right\}+P\left\{\hat{g}\notin\left[g^\star\left(1-\varepsilon^3\right),g^\star\left(1+\varepsilon^3\right)\right]\mid D\right\}.
\end{aligned} \tag{B.7}$$

The third term in (B.7) is smaller than ε for all sufficiently small ξ by (B.3). To handle the first two terms, observe that

$$P\left\{\max_{g^\star\leq g\leq(1+\varepsilon^3)g^\star}\left|\sum_{i=1}^{g}a(\mathsf{r}_{i,k},D)\right|>\varepsilon\sqrt{\hat{g}}\mid D\right\}\leq\frac{1}{\varepsilon^2}\frac{1}{g^\star}\operatorname{Var}\left(\sum_{i=1}^{\lfloor\varepsilon^3\hat{g}\rfloor+1}a(\mathsf{r}_{i,k},D)\mid D\right)\leq\varepsilon,$$

where the first inequality follows from Kolmogorov's maximal inequality (see e.g., Proposition 2.3.16 of Dembo (2021)). An analogous bound holds for the remaining term. Thus, we have that

$$\sqrt{\frac{g^\star}{2v_{1,k}(D)}}\left(U(\mathsf{R}_{\hat{g},k},D)-U(\mathsf{R}'_{\hat{g}',k},D)\right)=o_p\left(1\right)$$

as $\xi\to 0$. Hence, we have that

$$\begin{aligned}
&P\left\{\left|a(\mathsf{R}_{\hat{g},k},D)-a(\mathsf{R}'_{\hat{g}',k},D)\right|>\xi\mid D\right\}\\
&=P\left\{\left|\sqrt{\frac{g^\star}{2v_{1,k}(D)}}\left(\frac{1}{g^\star}\sum_{i=1}^{g^\star}a(\mathsf{r}_{i,k},D)-a(\mathsf{r}'_{i,k},D)\right)\right|>z_{1-\beta/2}\mid D\right\}+o\left(1\right)\\
&=1-\beta+o\left(1\right)
\end{aligned}$$

as $\xi\to 0$ by the central limit theorem. ■

B.2 Proof of Lemma 5.1. We use the following Lemma in several places.

Lemma B.1. *Let ψ and ψ' be two sets, each containing g elements of $\mathcal{P}_n$. The inequality*

$$\sum_{m=0}^{\infty} \left| \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m)\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m)\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\psi}, \boldsymbol{\pi}'_0 = \boldsymbol{\psi}', D \right] \right| \\ \leq 2gn^2 \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T\left(\mathsf{s}, D\right) - T\left(\mathsf{s}', D\right) \right)$$

holds almost surely.

Observe that the quantity

$$\max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T\left(\mathsf{s}, D\right) - T\left(\mathsf{s}', D\right) \right)$$

is finite almost surely. Thus, the convergence of the series defining $A\left(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D\right)$ follows from Lemma B.1. Define the operator

$$K : \mathcal{F} \to \mathcal{F} \\ f\left(\cdot\right) \mapsto \mathbb{E}\left[f\left(\boldsymbol{\pi}'\right) \mid \boldsymbol{\pi} = \cdot \right],$$

where $\mathcal{F}$ is the set of all measurable functions supported on the domain of $\boldsymbol{\pi}$. Observe that

$$\begin{aligned} &\mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D \right] \\ &= \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - \overline{a}(D) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, D \right] - \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) - \overline{a}(D) \mid \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D \right] \\ &= K^m\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi})\right) - \overline{a}(D) \right) - K^{m+1}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}')\right) - \overline{a}(D) \right). \end{aligned}$$

Thus, for any m', we have that

$$\begin{aligned} &\sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D \right] \\ &= a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) - K^{m'+1}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) \right). \end{aligned} \tag{B.8}$$

By Lemma B.1, the partial sums (B.8) converge almost everywhere and so the sequence

$$\left(K^{m+1}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) \right) \right)_{m \geq 0} \tag{B.9}$$

also converges almost everywhere. Lemma B.1 also implies that the limit of (B.9) depends only on D, as

$$K^m\left(a\left(\left(\mathsf{R}_{g,k}\left(\boldsymbol{\psi}\right), D\right)\right) - \overline{a}\left(D\right) \right) - K^m a\left(\left(\mathsf{R}_{g,k}\left(\boldsymbol{\psi}'\right), D\right) - \overline{a}\left(D\right) \right) \to 0$$

for any $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'$ each containing g elements of $\mathcal{P}_n$. Therefore, we have that

$$\begin{aligned} &\mathbb{E}\left[A\left(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D\right) \mid D \right] \\ &= \mathbb{E}\left[\lim_{n \to \infty} \left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) - K^{m+1} a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) \right) \mid D \right] \\ &= \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) \mid D \right] - b\left(D\right), \end{aligned}$$

for some quantity

$$b(D) = \lim_{m\to\infty} K^m \left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}(D)\right)$$

that depends only on D. Observe that

$$\begin{aligned}
&\mathbb{E}\left[A\left(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D\right) \mid D\right] \\
&= \mathbb{E}\left[\lim_{m'\to\infty} \sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D\right] \mid D\right] \\
&= \lim_{m'\to\infty} \sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid D\right] && \text{(Dominated Conv.)} \\
&= 0, && \text{(Exchangeability)}
\end{aligned}$$

where the applicability of the Dominated Convergence Theorem follows from Lemma B.1. Thus, as

$$\mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}(D) \mid D\right] = 0,$$

we can conclude that $b(D) = 0$ almost surely. Hence, we find that

$$\begin{aligned}
&\mathbb{E}\left[A\left(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D\right) \mid \boldsymbol{\pi}, D\right] \\
&= \mathbb{E}\left[\lim_{m'\to\infty} \sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D\right] \mid \boldsymbol{\pi}, D\right] \\
&= \lim_{m'\to\infty} \sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D\right] && \text{(Dominated Conv.)} \\
&- \lim_{m'\to\infty} K^{m'+1}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}(D)\right) \\
&= a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}(D),
\end{aligned}$$

completing the proof. ■

B.3 Proof of Lemma 5.2. First, observe that

$$U_F(X) \geq \frac{1}{2}\left(\mathbb{E}\left[F\left(X, X'\right) \mid X\right]\right)^2 = \frac{1}{2} f(X)^2,$$

where the inequality follows from Jensen's inequality and the definition of the Stein representer F. By (5.11), we have that

$$su \geq \frac{1}{2} f(X)^2.$$

Hence, the random variable $f(X)$ is bounded almost surely.

Now, suppose $h : \mathcal{X} \to \mathbb{R}$ is any measurable function such that $\mathbb{E}\left[h(X) F(X, X')\right] < \infty$. Then, $\mathbb{E}\left[h(X) f(X)\right] = \mathbb{E}\left[h(X) F(X, X')\right]$. Using the exchangeability of X and X' and the fact that F is

antisymmetric, we have that

$$\mathbb{E}\left[h\left(X\right)F\left(X,X'\right)\right]=\mathbb{E}\left[h\left(X'\right)F\left(X',X\right)\right]=-\mathbb{E}\left[h\left(X'\right)F\left(X,X'\right)\right]$$

and that therefore

$$\mathbb{E}\left[h\left(X\right)f\left(X\right)\right]=\frac{1}{2}\mathbb{E}\left[\left(h\left(X\right)-h\left(X'\right)\right)F\left(X,X'\right)\right]. \tag{B.10}$$

Let

$$m\left(\theta\right)=\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\right]$$

denote the moment generating function of $f\left(X\right)$. As $f\left(X\right)$ is bounded almost surely, we can exchange differentiation and expectation in the differentiation of $m\left(\theta\right)$. Thus, we obtain

$$\begin{aligned}m'\left(\theta\right)&=\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)f\left(X\right)\right].\\&=\frac{1}{2}\mathbb{E}\left[\left(\exp\left(\theta f\left(X\right)\right)-\exp\left(\theta f\left(X'\right)\right)\right)F\left(X,X'\right)\right],\end{aligned} \tag{B.11}$$

where the second inequality follows from (B.10). To bound $m'\left(\theta\right)$ we apply the following exponential mean-value inequality, stated in a more general form in Paulin et al. (2016).

Lemma B.2. *For all constants x, y, and c in $\mathbb{R}$ and $s>0$, it holds that*

$$\left|\left(e^{x}-e^{y}\right)c\right|\leq\frac{1}{4}\left(s\left(x-y\right)^{2}+s^{-1}c^{2}\right)\left(e^{x}+e^{y}\right).$$

In particular, by (B.11) and Lemma B.2, we obtain the bound

$$\begin{aligned}\left|m'\left(\theta\right)\right|&\leq\frac{1}{2}\mathbb{E}\left[\left|\left(\exp\left(\theta f\left(X\right)\right)-\exp\left(\theta f\left(X'\right)\right)\right)F\left(X,X'\right)\right|\right]\\&\leq\frac{1}{8}\inf_{t>0}\mathbb{E}\left[\left(t\left(\theta f\left(X\right)-\theta f\left(X'\right)\right)^{2}+t^{-1}F\left(X,X'\right)^{2}\right)\left(\exp\left(\theta f\left(X\right)\right)+\exp\left(\theta f\left(X'\right)\right)\right)\right]\\&=\frac{\left|\theta\right|}{4}\inf_{t>0}\mathbb{E}\left[\left(t\left(f\left(X\right)-f\left(X'\right)\right)^{2}+t^{-1}F\left(X,X'\right)^{2}\right)\exp\left(\theta f\left(X\right)\right)\right]\\&=\frac{\left|\theta\right|}{2}\inf_{t>0}\mathbb{E}\left[\left(\frac{t}{2}\mathbb{E}\left[\left(f\left(X\right)-f\left(X'\right)\right)^{2}\mid X\right]+\frac{1}{2t}\mathbb{E}\left[F\left(X,X'\right)^{2}\mid X\right]\right)\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\mid X\right]\right].\\&=\frac{\left|\theta\right|}{2}\inf_{t>0}\mathbb{E}\left[\left(tU_{f}\left(X\right)+t^{-1}U_{F}\left(X\right)\right)\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\mid X\right]\right]\\&\leq\frac{\left|\theta\right|}{2}\mathbb{E}\left[\left(sU_{f}\left(X\right)+s^{-1}U_{F}\left(X\right)\right)\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\mid X\right]\right]\\&\leq\left|\theta\right|v\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\right]\end{aligned}$$

for all $\theta\in\mathbb{R}$. Thus, we have that

$$m'\left(\theta\right)\leq u\theta m\left(\theta\right)$$

for all $\theta>0$. As $m\left(\cdot\right)$ is a convex function and $m'\left(0\right)=0$, $m'\left(\theta\right)$ always has the same sign as θ, we find that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log m\left(\theta\right)\leq u\theta.$$

As a consequence, and by $m(0) = 1$, we have that

$$\log m(\theta) \leq \int_0^\theta ut\mathrm{d}t \leq \frac{u\theta^2}{2}.$$

By the Chernoff bound (see e.g., Equation 2.5, Wainwright, 2019), we have

$$\log P\{f(X) \geq \delta\} \leq \inf_{\theta \geq 0} (\log m(\theta) - \theta\delta) \leq \inf_{\theta \geq 0} \left(\frac{u\theta^2}{2} - \theta\delta\right)$$

Solving this minimization with $\theta = \delta/u$, we find

$$P\{f(X) \geq t\} \leq \exp\left(\frac{-\delta^2}{2u}\right),$$

as required. The analogous lower tail bound follows from an identical argument, which completes the proof of (5.7). To prove (5.8) we apply the following result stated in Chatterjee (2007).

Theorem B.1 (Theorem 1.5, (iii), Chatterjee, 2007)**.** *Reintroduce the notation and assumptions from the statement of Theorem 5.2. Define*

$$\Delta(X) = \frac{1}{2}\mathbb{E}\left[|F\left(X, X'\right)\left(f(X) - f\left(X'\right)\right)| \mid X\right].$$

The Burkholder-Davis-Gundy inequality

$$\mathbb{E}\left[f(X)^{2r}\right] \leq (2r-1)^r \mathbb{E}\left[\Delta(X)^r\right]$$

holds for any positive integer r.

The inequality

$$\begin{aligned}
\mathbb{E}\left[\Delta(X)^r\right] &= \mathbb{E}\left[\mathbb{E}\left[|\left(f(X) - f\left(X'\right)\right)F\left(X, X'\right)| \mid X\right]^r\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[|\left(f(X) - f\left(X'\right)\right)F\left(X, X'\right)|^r \mid X\right]\right] && \text{(Jensen)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\left(s^{-1}\left(f(X) - f\left(X'\right)\right)^{2r}\right)\left(sF\left(X, X'\right)^{2r}\right)\right)^{1/2} \mid X\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[s^{-1}\left(f(X) - f\left(X'\right)\right)^{2r} + sF\left(X, X'\right)^{2r} \mid X\right]\right] && \text{(Young)} \\
&= s^{-1}\mathbb{E}\left[U_f^{(r)}(X)\right] + s\mathbb{E}\left[U_F^{(r)}(X)\right]
\end{aligned}$$

then completes the proof. ■

B.4 Proof of Lemma 5.3. We apply the following Lemma.

Lemma B.3. *Let $(X_m)_{m\geq 0}$ be a sequence of real-valued random variables. Suppose that the inequality*

$$\mathbb{E}\left[X_m^{2^c}\right] \leq h_m^{2^c} \tag{B.12}$$

holds for each $m \geq 0$ and positive integer c, where $(h_m)_{m\geq 0}$ is a deterministic sequence of nonnegative real numbers. If there exists a square integrable random variable W such that

$$\left(\sum_{m=0}^{\infty} X_m\right)^{2^c} \leq W$$

almost surely, then the inequality

$$\mathbb{E}\left[\left(\sum_{m=0}^{\infty} X_m\right)^{2^c}\right] \leq \left(\sum_{m=0}^{\infty} h_m\right)^{2^c}$$

holds almost surely.

Define the objects

$$U_f^{(r)}(Z) = \frac{1}{2}\mathbb{E}\left[\left(f(Z) - f\left(Z'\right)\right)^{2r} \mid Z\right] \quad \text{and}$$

$$U_F^{(r)}(Z) = \frac{1}{2}\mathbb{E}\left[\left(\sum_{m=0}^{\infty} \mathbb{E}\left[f(Z_m) - f\left(Z'_m\right) \mid Z_0 = Z, Z'_0 = Z'\right]\right)^{2r} \mid Z\right].$$

By the (5.9), we have that

$$\mathbb{E}\left[\sum_{m=0}^{\infty} f(Z_m) - f(Z'_m) \mid Z_0 = Z, Z'_0 = Z'\right]^{2^c} \leq W.$$

Thus, (5.10) guarantees that the conditions of Lemma B.3 are satisfied, and we have that

$$\mathbb{E}\left[U_F^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \leq \frac{1}{2}\left(\sum_{i=0}^{\infty} h_i\right)^{2^c} \quad \text{and} \quad \mathbb{E}\left[U_f^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \leq \frac{1}{2}h_0^{2^c}.$$

By Markov's inequality, we obtain

$$P\left\{U_F^{(2^{c-1})}(Z) \geq \frac{2}{\delta}\mathbb{E}\left[U_F^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \quad \text{or} \quad U_f^{(2^{c-1})}(Z) \geq \frac{2}{\delta}\mathbb{E}\left[U_f^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \mid \boldsymbol{\pi}\right\} \leq \delta.$$

Hence, by Lemma B.3 and DeMorgan's law, the probability that both

$$U_F^{(2^{c-1})}(Z) \leq \frac{2}{\delta}\mathbb{E}\left[U_F^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \leq \frac{1}{\delta}\left(\sum_{m=0}^{\infty} h_m\right)^{2^c} \quad \text{and}$$

$$U_f^{(2^{c-1})}(Z) \leq \frac{2}{\delta}\mathbb{E}\left[U_f^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \leq \frac{h_0^{2^c}}{\delta}$$

hold is greater than $1 - \delta$. ■

B.5 Proof of Lemma 5.4. Recall the definition of the collections $(\boldsymbol{\pi}_m, \boldsymbol{\pi}'_m)_{m\geq 0}$ given in (5.3). Fix $Z_0 = Z$ and $Z'_0 = Z'$ throughout. For any i in $[n]$, let $\mathsf{s}_{m,\ell}(i)$ denote the element of the collection

$$\mathsf{r}(\pi_{m,\ell}) = (\mathsf{s}_1(\pi_{m,\ell}), \ldots, \mathsf{s}_g(\pi_{m,\ell}))$$

that contains the index i and set $\mathsf{s}_{m,\ell}(i)$ equal to $\varnothing$ if no element of $\mathsf{r}(\pi_{m,\ell})$ contains i.

We begin by defining three events that will determine the structure of our argument. By construction, the collections $\boldsymbol{\pi}_m$ and $\boldsymbol{\pi}'_m$ are either identical or differ in exactly two indices in their Lth element. Let $\mathcal{E}_m$ denote the event that $\boldsymbol{\pi}_m$ and $\boldsymbol{\pi}'_m$ differ. On the event $\mathcal{E}_m$, let $i_{1,m}$ and $i_{2,m}$ denote the two indices in which the Lth elements of $\boldsymbol{\pi}_m$ and $\boldsymbol{\pi}'_m$ differ. Define the random variables B_m and C_m such that, conditional on $\mathcal{E}_m$, each is uniformly distributed on $\{i_{1,m}, i_{2,m}\}$ such that $B_m \neq C_m$. On the complement of $\mathcal{E}_m$, set these indices uniformly at random. Let $\mathcal{F}_m$ denote the event that

$$\mathsf{s}_{m,L}\left(B_m\right) \neq \mathsf{s}_{m,L}\left(C_m\right),$$

i.e., the event that the indices that differ are not in the same element of the collection $\mathsf{r}(\pi_{m,\ell})$. Finally, let $\mathcal{G}_m$ denote the event that

$$\mathsf{s}_{m,L}\left(B_m\right) \neq \varnothing \quad \text{and} \quad \mathsf{s}_{m,L}\left(C_m\right) \neq \varnothing,$$

i.e., the event that the indices that differ are both in the collection $\mathsf{r}(\pi_{m,\ell})$.

By Assumption 3.1, the event $\mathcal{H}_m = \mathcal{E}_m \cap \mathcal{F}_m$ is a necessary condition for $a\left(Z_m\right) - a\left(Z'_m\right) \neq 0$. Thus, we can compute

$$\begin{aligned}
P\left\{\mathcal{H}_m \mid \mathcal{E}_0\right\} &= P\left\{\mathcal{F}_m \mid \mathcal{E}_m, \mathcal{E}_0\right\} P\left\{\mathcal{E}_m \mid \mathcal{E}_0\right\} \\
&= \Big(P\left\{\mathsf{s}_{m,L}\left(B_m\right) \neq \mathsf{s}_{m,L}\left(C_m\right) \mid \mathsf{s}_{m,L}\left(B_m\right) \neq \varnothing, \mathcal{E}_m\right\} P\left\{\mathsf{s}_{m,L}\left(B_m\right) \neq \varnothing \mid \mathcal{E}_m\right\} \\
&\quad + P\left\{\mathsf{s}_{m,L}\left(B_m\right) \neq \mathsf{s}_{m,L}\left(C_m\right) \mid \mathsf{s}_{m,L}\left(B_m\right) = \varnothing, \mathcal{E}_m\right\} P\left\{\mathsf{s}_{m,L}\left(B_m\right) = \varnothing \mid \mathcal{E}_m\right\} \Big) \\
&\quad \cdot P\left\{\mathcal{E}_m \mid \mathcal{E}_0\right\} \\
&= \left(\frac{n-b}{n-1} \frac{kb}{n} + \frac{kb}{n-1} \frac{n-kb}{b} \right) \left(1 - \frac{2}{gn^2}\right)^m \\
&= \left(\frac{kb\left(2n - kb - b\right)}{n(n-1)} \right) \left(1 - \frac{2}{gn^2}\right)^m
\end{aligned}$$

for all $m \geq 0$ by the law of total probability. Moreover, we have that

$$P\left\{\mathcal{H}_m \mid Z, Z'\right\} \leq P\left\{\mathcal{H}_m \mid \mathcal{E}_0\right\}$$

almost surely. Consequently, we find that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[a\left(Z_m\right) - a\left(Z'_m\right) \mid Z, Z'\right]^{2r} \mid \boldsymbol{\pi}\right] \\
&= \mathbb{E}\left[\left(P\left\{\mathcal{H}_m \mid Z, Z'\right\} \mathbb{E}\left[a\left(Z_m\right) - a\left(Z'_m\right) \mid Z, Z', \mathcal{H}_m\right]\right)^{2r} \mid \boldsymbol{\pi}\right] \\
&\leq \left(P\left\{\mathcal{H}_m \mid \mathcal{E}_0\right\}\right)^{2r} \mathbb{E}\left[\mathbb{E}\left[\left(a\left(Z_m\right) - a\left(Z'_m\right)\right)^{2r} \mid Z, Z', \mathcal{H}_m\right] \mid \boldsymbol{\pi}\right] \qquad \text{(Jensen)} \\
&= \left(1 - \frac{2}{gn^2}\right)^{2mr} \left(\frac{kb\left(2n - kb - b\right)}{n(n-1)} \right)^{2r} \mathbb{E}\left[\mathbb{E}\left[\left(a\left(Z_m\right) - a\left(Z'_m\right)\right)^{2r} \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \mid \boldsymbol{\pi}\right]
\end{aligned}$$

$$\leq \left(1 - \frac{2}{gn^2}\right)^{2mr} \left(\frac{2kb\,(2n - kb - b)}{n^2}\right)^{2r} \mathbb{E}\left[\left(a\left(Z_m\right) - a\left(Z'_m\right)\right)^{2r} \mid \mathcal{H}_m\right] , \tag{B.13}$$

where the final inequality follows from the fact that $\boldsymbol{\pi}$ is uniformly distributed independently of D and the elementary inequality $n/(n-1) \leq 2$. Thus, it remains to bound the expectation in (B.13).

To ease notation, we now drop the dependence on m and L. Observe that

$$P\left\{\mathcal{G} \mid \mathcal{H}\right\} = P\left\{\mathsf{s}\left(B\right) \neq \varnothing \mid \mathsf{s}\left(C\right) \neq \varnothing\right\} = \frac{kb - b}{n - b}$$

and that therefore

$$\begin{aligned}\mathbb{E}\left[\left(a\left(Z\right) - a\left(Z'\right)\right)^{2r} \mid \mathcal{H}, \boldsymbol{\pi}\right] &= \left(\frac{kb - b}{n - b}\right) \mathbb{E}\left[\left(a\left(Z\right) - a\left(Z'\right)\right)^{2r} \mid \mathcal{G}\right] \\ &\quad + \left(\frac{n - kb}{n - b}\right) \mathbb{E}\left[\left(a\left(Z\right) - a\left(Z'\right)\right)^{2r} \mid \mathcal{H} \setminus \mathcal{G}\right] . \end{aligned} \tag{B.14}$$

Define the sets

$$\hat{\mathsf{s}}_i = \mathsf{s}(i) \setminus i \quad \text{and} \quad \bar{\mathsf{s}} = \tilde{\mathsf{s}}(B) \cap \tilde{\mathsf{s}}(C).$$

With an abuse of notation, we let

$$\begin{aligned}\psi(D_i, \hat{\eta}(D_j, D_{\hat{\mathsf{s}}_k})) &= \psi(D_i, \hat{\eta}(D_j \cap D_{\hat{\mathsf{s}}_k} \cap D_{\bar{\mathsf{s}}})) \quad \text{and} \\ h(D_i, D_j, D_{\hat{\mathsf{s}}_k}) &= \psi(D_i, \hat{\eta}(D_j, D_{\hat{\mathsf{s}}_k})) - \psi(D_j, \hat{\eta}(D_i, D_{\hat{\mathsf{s}}_k})).\end{aligned}$$

Observe that

$$\begin{aligned}&\mathbb{E}\left[\left(a\left(Z\right) - a\left(Z'\right)\right)^{2r} \mid \mathcal{H} \setminus \mathcal{G}\right] \\ &= \mathbb{E}\left[\left(a\left(Z\right) - a\left(Z'\right)\right)^{2r} \mid \mathcal{H} \setminus \mathcal{G}\right] = \left(\frac{1}{gkb}\right)^{2r} \mathbb{E}\left[h(D_B, D_C, D_{\tilde{s}_C})^{2r}\right] .\end{aligned} \tag{B.15}$$

On the other hand, we can decompose

$$\begin{aligned}&\mathbb{E}\left[\left(a\left(Z\right) - a\left(Z'\right)\right)^{2} \mid \mathcal{G}\right] \\ &= \left(\frac{1}{gkb}\right)^{2r} \mathbb{E}\left[\left(h(D_B, D_C, D_{\hat{\mathsf{s}}_C}) + h(D_C, D_B, D_{\hat{\mathsf{s}}_B})\right)^{2r}\right] \\ &= \left(\frac{1}{gkb}\right)^{2r} \mathbb{E}\left[\left(h(D_B, D_C, D_{\hat{\mathsf{s}}_C}) - h(D_B, D_B, D_{\hat{\mathsf{s}}_B})\right)^{2r}\right] .\end{aligned} \tag{B.16}$$

Consequently, by (B.14), (B.15), and (B.16), it suffices to express suitable bounds for the expectations

$$\begin{aligned}&\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \quad \text{and} \\ &\mathbb{E}\Bigg[\Bigg(\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)\end{aligned} \tag{B.17}$$

$$- \left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_B}))\right)^{2r}\Bigg]$$

respectively. To this end, recall that $\tilde{D}_i$ are independent copies of D_i for each i. Observe that

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \\
&= \mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) + \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \\
&= \sum_{q=0}^{2r} \binom{2r}{q} \mathbb{E}\Bigg[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r-q} \\
&\qquad \left(\psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{q}\Bigg] && \text{(Binomial Theorem)} \\
&\leq 2^{2r}\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] && \text{(B.18)}
\end{aligned}$$

where the inequality follows from the fact that B and C are exchangeable and the Hölder inequality. Similarly, we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r-q}\right] \\
&= \mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) + \psi(D_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \\
&= \sum_{q=0}^{2r} \binom{2r}{q} \mathbb{E}\Bigg[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r-q} \\
&\qquad \left(\psi(D_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{q}\Bigg] && \text{(Binomial Theorem)} \\
&\leq \sum_{q=0}^{2r} \binom{2r}{q} \left(\sigma_{\mathsf{valid}}^{(2r)}\right)^{\frac{2r-q}{2r}} \left(\sigma_{\mathsf{train}}^{(2r,1)}\right)^{\frac{q}{2r}} && \text{(Hölder)} \\
&\leq 2^{2r}\sigma_{\mathsf{max}}^{(2r)}, && \text{(B.19)}
\end{aligned}$$

where the last inequality follows by the definitions of $\sigma_{\mathsf{valid}}^{(2r)}$ and $\sigma_{\mathsf{train}}^{(2r,1)}$. Thus, we have that

$$\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \leq 2^{4r}\sigma_{\mathsf{max}}^{(2r)} \tag{B.20}$$

by (B.18) and (B.19).

Next, we consider the double difference term (B.17). In this case, we have that

$$\mathbb{E}\Bigg[\Bigg(\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right) \\
- \left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_B}))\right)\Bigg)^{2r}\Bigg]$$

$$
\begin{aligned}
&= \mathbb{E}\Bigg[\Big(\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B})\right)\Big) \\
&\qquad - \left(\psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_B}))\right)\Big)^{2r}\Bigg] \\
&= \sum_{q=0}^{2r} \binom{2r}{q} \mathbb{E}\Bigg[\Big(\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B})\right) \\
&\qquad \cdot \left(\psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_B}))\right)\Big)^{2r}\Bigg] \qquad \text{(Binomial Theorem)} \\
&\leq 2^{2r}\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B}))\right)^{2r}\right] = 2^{2r}\sigma_{\mathsf{train}}^{(2r,b-1)} \qquad \text{(B.21)}
\end{aligned}
$$

where the final inequality follows from the fact that B and C are exchangeable and the Hölder inequality. Putting the pieces together, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[a\left(Z_m\right) - a\left(Z'_m\right) \mid Z, Z'\right]^{2r} \mid \boldsymbol{\pi}\right] \\
&\leq 2^{4r}\left(1 - \frac{2}{gn^2}\right)^{2mr}\left(\frac{2n - kb - b}{gn^2}\right)^{2r}\left(\left(\frac{n - kb}{n - b}\right)2^{2r}\sigma_{\mathsf{max}}^{(2r)} + \left(\frac{kb - b}{n - b}\right)\sigma_{\mathsf{train}}^{(2r,b-1)}\right)
\end{aligned}
$$

as required. ■

B.6 Proof of Lemma 5.5. Throughout, we us the short hand $\hat{v}_{g,k}(Z)$ to denote $\hat{v}(\mathsf{R}_{g,k}, D)$. We begin by decomposing the estimator $\hat{v}_{g,k}(Z)$ into two parts that will each be easier to handle when considered in isolation. To this end, define the statistics

$$
\begin{aligned}
\tilde{v}_{g,k}(Z) &= \frac{1}{g^2k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i}, Y) - \overline{a}(D)\right)\left(T(\mathsf{s}_{\ell,i'}, D) - \overline{a}(D)\right) \quad \text{and} \\
\check{v}_{g,k}(Z) &= \left(a(Z) - \overline{a}(D)\right)^2.
\end{aligned}
$$

Observe that both $\tilde{v}_{g,k}(Z)$ and $\check{v}_{g,k}(Z)$ are unbiased for $v(Z)$. Moreover, we can write

$$
\begin{aligned}
\hat{v}_{g,k}(Z) &= \frac{1}{g(g-1)k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i}, D) - a(Z)\right)\left(T(\mathsf{s}_{\ell,i'}, D) - a(Z)\right) \\
&= \frac{1}{g(g-1)k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i}, D) - \overline{a}(D)\right)\left(T(\mathsf{s}_{\ell,i'}, D) - a(Z)\right) \\
&\quad + \frac{1}{g(g-1)k}\sum_{\ell=1}^{g}\sum_{i=1}^{g}\left(\overline{a}(D) - a(Z)\right)\left(T(\mathsf{s}_{\ell,i}, D) - a(Z)\right) \\
&= \frac{1}{g(g-1)k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i}, D) - \overline{a}(D)\right)\left(T(\mathsf{s}_{\ell,i'}, D) - a(Z)\right)
\end{aligned}
$$

$$= \frac{g}{g-1}\tilde{v}\left(Z\right) + \frac{1}{g\left(g-1\right)k}\sum_{\ell=1}^{g}\sum_{i=1}^{k}\left(T(\mathsf{s}_{\ell,i}, D) - \overline{a}\left(D\right)\right)\left(\overline{a}\left(D\right) - a\left(Z\right)\right)$$

$$= \frac{g}{g-1}\tilde{v}_{g,k}\left(Z\right) - \frac{1}{g-1}\check{v}_{g,k}\left(Z\right). \tag{B.22}$$

Hence, it will suffice to characterize the exponential rates of conditional concentration for the statistics $\tilde{v}_{g,k}\left(Z\right)$ and $\check{v}_{g,k}\left(Z\right)$. These are established through the application of the following Lemma, which follows from an argument very similar to the proof of Theorem 3.1.

Lemma B.4. *Suppose that Assumptions 3.1 and 3.2 hold and that the data D are independently and identically distributed. If the eighth-order split-stability $\zeta^{(8)}$ is finite, then:*

(i) *The conditional concentration inequality*

$$\log\frac{1}{2}P\left\{\left|\tilde{v}\left(Z\right) - v\left(D\right)\right| \geq t \mid D\right\} \lesssim -\frac{\delta}{(2-\varphi k-\varphi)^4}\frac{g^3t^2}{\Gamma^{(2)}_{k,\varphi,b}}, \tag{B.23}$$

holds for all $t > 0$ with probability greater than $1-\delta$.

(ii) *The conditional concentration inequality*

$$\log\frac{1}{2}P\left\{\left|\check{v}\left(Z\right) - v\left(D\right)\right| \geq t \mid D\right\} \lesssim -\frac{\delta}{(2-\varphi k-\varphi)^4}\frac{g^2t^2}{\Gamma^{(2)}_{k,\varphi,b}}, \tag{B.24}$$

holds for all $t > 0$ with probability greater than $1-\delta$.

Putting the pieces together, by (B.22), we have that

$$\begin{aligned}
&P\left\{\left|\hat{v}_{g,k}\left(Z\right) - v_{g,k}\left(D\right)\right| \leq t \mid D\right\} \\
&= P\left\{\left|\frac{g}{g-1}\left(\tilde{v}_{g,k}\left(Z\right) - v_{g,k}\left(D\right)\right) - \frac{1}{g-1}\left(\check{v}_{g,k}\left(Z\right) - v_{g,k}\left(D\right)\right)\right| \leq t \mid D\right\} \\
&\geq P\left\{\frac{g}{g-1}\left|\tilde{v}_{g,k}\left(Z\right) - v_{g,k}\left(D\right)\right| + \frac{1}{g-1}\left|\check{v}_{g,k}\left(Z\right) - v_{g,k}\left(D\right)\right| \leq t \mid D\right\} \\
&\geq P\left\{\frac{g}{g-1}\left|\tilde{v}_{g,k}\left(Z\right) - v\left(D\right)\right| \leq \frac{\sqrt{g}}{1+\sqrt{g}}t, \frac{1}{g-1}\left|\check{v}\left(Z\right) - v\left(D\right)\right| \leq \frac{1}{1+\sqrt{g}}t \mid D\right\} \\
&\geq P\left\{\left|\tilde{v}_{g,k}\left(Z\right) - v\left(D\right)\right| \leq \frac{1}{\sqrt{g}}\frac{g-1}{1+\sqrt{g}}t \mid D\right\} + P\left\{\left|\check{v}_{g,k}\left(Z\right) - v\left(D\right)\right| \leq \frac{g-1}{1+\sqrt{g}}t \mid D\right\} - 1 \\
&\geq 1 - 4\exp\left(-\frac{\delta}{C(2-\varphi k-\varphi)^4}\frac{g^3t^2}{\Gamma^{(2)}_{k,\varphi,b}}\right)
\end{aligned}$$

with probability greater than $1-\delta$ for some universal constant C, where the last inequality follows by Lemma B.4 and the facts that $4g \geq \left(1+\sqrt{g}\right)^2$ for all $g \geq 1$ and $\left(1/4\right)g^2 \leq \left(g-1\right)^2$ for all $g \geq 2$. ■

B.7 Proof of Lemma 5.6. Define the event

$$\mathcal{W}_\lambda\left(D\right) = \mathcal{U}_{k,\lambda}\left(D\right) \cap \mathcal{Q}_{k,\lambda}(D),$$

the quantity

$$W(D) = \left(U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D\right) + \left(Q(\mathsf{R}_{\hat{g},k}, D) - Q(\mathsf{R}'_{\hat{g}',k}, D)\right).$$

By the decomposition (B.6), we have that

$$\begin{aligned}
&\left| P\left\{a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D) \leq \xi \mid D\right\} - (1-\beta/2) \right| \\
&= \left| P\left\{ \frac{a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)}{\sqrt{2v_{g^\star,k}(D)}} \leq \frac{\xi}{\sqrt{2v_{g^\star,k}(D)}} \mid D \right\} - (1-\beta/2) \right| \\
&= \left| P\left\{ \frac{a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)}{\sqrt{2v_{g^\star,k}(D)}} \leq \frac{\xi}{\sqrt{2v_{g^\star,k}(D)}} - \frac{W(D)}{\sqrt{2v_{g^\star,k}(D)}} \mid D \right\} - (1-\beta/2) \right| \\
&\leq \sup_{q\in[-2\lambda,2\lambda]} \left| P\left\{ \frac{a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)}{\sqrt{2v_{g^\star,k}(D)}} \leq z_{1-\beta/2} + q \right\} - (1-\beta/2) \right| + \left(1 - P\left\{\mathcal{W}_\lambda(D) \mid D\right\}\right) \\
&\leq d_K\left(\frac{a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)}{\sqrt{2v_{g^\star,k}(D)}}, W \mid D \right) + 2\lambda + \left(1 - P\left\{\mathcal{W}_\lambda(D) \mid D\right\}\right),
\end{aligned}$$

as required. ■

B.8 Proof of Lemma 5.7, Part (i). Observe that we can write

$$a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D) = \frac{1}{g^\star} \sum_{\ell=1}^{g} \left(\overline{a}(\mathsf{r}_\ell, D) - \overline{a}(\mathsf{r}'_\ell, D)\right),$$

where $\overline{a}(\cdot, D)$ is defined in (5.14). We have that

$$\begin{aligned}
\mathbb{E}\left[|\overline{a}(\mathsf{r}_\ell, D) - \overline{a}(\mathsf{r}'_\ell, D)|^3\right] &\leq \left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}_\ell, D) - \overline{a}(\mathsf{r}'_\ell, D)\right)^4\right]\right)^{3/4} && \text{(Hölder)} \\
&\leq 2^6 \left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}_\ell, D)\right)^4\right]\right)^{3/4} \\
&\leq 2^6 3^2 (2 - \varphi k - \varphi)^3 (\Gamma^{(2)}_{k,\varphi,b})^{3/4},
\end{aligned}$$

where the second inequality follows from Hölder's inequality, the binomial theorem, and the fact that r_ℓ and r'_ℓ are exchangeable. The final inequality follows from Theorem 3.2. Consequently, we find that

$$\begin{aligned}
\frac{\mathbb{E}\left[|\overline{a}(\mathsf{r}_\ell, D) - \overline{a}(\mathsf{r}'_\ell, D)|^3 \mid D\right]}{(g^\star)^{1/2}\left(2v_{1,k}(D)\right)^{3/2}} &\lesssim \frac{(2 - \varphi k - \varphi)^3 (\Gamma^{(2)}_{k,\varphi,b})^{3/4}}{\delta (g^\star)^{1/2} \left(v_{1,k}(D)\right)^{3/2}} && \text{(Markov)} \\
&\lesssim \frac{\xi}{z_{1-\beta/2}} \frac{(2 - \varphi k - \varphi)^3 (\Gamma^{(2)}_{k,\varphi,b})^{3/4}}{\delta \left(v_{1,k}(D)\right)^2},
\end{aligned}$$

holds with probability $1 - \delta$. The Lemma then follows by the standard Berry-Esseen inequality. ■

B.9 Proof of Lemma 5.7, Part (ii). Our bound is based on the following two conditional concentration inequalities. Both arguments are based on a Chernoff-type maximal inequality, due to Steiger (1970).

Lemma B.5. *Suppose that Assumptions 3.1 and 3.2 hold, that the data D are independent and identically distributed, and that the eighth-order split ζ^8 stability is finite.*

(i) *For all $t > 0$ and $c > 0$, the condition concentration inequality*

$$\log \frac{1}{2} P\left\{ |U(\mathsf{R}_{\hat{g},k}, D)| \geq t\sqrt{v_{g^\star,k}(D)}, |\hat{g}/g^\star - 1| \leq c \mid D \right\} \leq -\frac{\delta v_{1,k(D)}}{2^4(2-\varphi k-\varphi)^2 \Gamma^{(1)}_{k,\varphi,b}} \frac{t^2}{c} \tag{B.25}$$

holds with probability greater than $1-\delta$ as D varies.

(ii) *If $g^\star c \geq 2$ and $1/2 > c > 0$, then the conditional concentration inequality*

$$\log \frac{1}{8} P\left\{ |\hat{g} - g^\star| > c g^\star \mid D \right\} \lesssim -\frac{\delta (v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4} \frac{z_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}} \frac{c^2}{\xi^2} \tag{B.26}$$

holds with probability greater than $1-\delta$ as D varies.

Observe that

$$\begin{aligned}
&P\left\{ |U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D)| \geq \lambda\sqrt{2v_{g^\star,k}(D)} \right\} \\
&\lesssim P\left\{ |U(\mathsf{R}_{\hat{g},k}, D)| \geq \lambda\sqrt{2v_{g^\star,k}(D)} \right\} \\
&\leq P\left\{ |U(\mathsf{R}_{\hat{g},k}, D)| \geq \lambda\sqrt{2v_{g^\star,k}(D)}, |\hat{g}/g^\star - 1| \leq c \mid D \right\} + P\left\{ |\hat{g}/g^\star - 1| \geq c \mid D \right\} \\
&\lesssim \exp\left(-\frac{\delta v_{1,k(D)}}{2^4(2-\varphi k-\varphi)^2 \Gamma^{(1)}_{k,\varphi,b}} \frac{\lambda^2}{c} \right) + \exp\left(-C\frac{\delta (v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4} \frac{z_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}} \frac{c^2}{\xi^2} \right)
\end{aligned} \tag{B.27}$$

for some universal constant C. Hence, it remains to choose c and λ such that the quantity (B.27) is less than $\rho_{k,\varphi,b}(\xi, \beta \mid D)$. First, we choose c such that

$$\frac{\delta (v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4} \frac{z_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}} \frac{c^2}{\xi^2} \lesssim \log\left(\rho_{k,\varphi,b}(\xi, \beta \mid D)^{-1} \right)$$

Choosing c by

$$c = \xi \left(\frac{(2-\varphi k-\varphi)^2 (\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{z_{1-\beta/2}\delta^{1/2}(v_{1,k}(D))^{3/2}} \right) \log^{1/2}\left(\rho_{k,\varphi,b}(\xi, \beta \mid D)^{-1} \right) \tag{B.28}$$

suffices. Next, we choose λ such that

$$\frac{\delta v_{1,k(D)}}{2^4(2-\varphi k-\varphi)^2 \Gamma^{(1)}_{k,\varphi,b}} \frac{\lambda^2}{c} \lesssim \log\left(\rho_{k,\varphi,b}(\xi, \beta \mid D)^{-1} \right).$$

We can rewrite this condition by plugging in our choice of c through

$$\frac{\lambda^2}{\xi} \frac{z_{1-\beta/2}\delta^{3/2}(v_{1,k}(D))^{5/2}}{(2-\varphi k-\varphi)^4 \Gamma^{(1)}_{k,\varphi,b}(\Gamma^{(2)}_{k,\varphi,b})^{1/2}} \lesssim \log^{3/2}\left(\rho_{k,\varphi,b}(\xi, \beta \mid D)^{-1} \right)$$

Choosing λ by

$$\lambda = \left(\frac{\xi}{z_{1-\beta/2}} \frac{(2-\varphi k-\varphi)^4 \Gamma^{(1)}_{k,\varphi,b} (\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\delta^{3/2} (v_{1,k}(D))^{5/2}} \right)^{1/2} \tilde{\lambda}^{U}_{k,\varphi,b}(D), \quad \text{where}$$

$$\tilde{\lambda}^{U}_{k,\varphi,b}(D) = \log^{3/4}\left(\rho_{k,\varphi,b}(\xi, \beta \mid D)^{-1} \right)$$

will then suffice, as required. ■

B.10 Proof of Lemma 5.7, Part(iii). Observe that

$$\begin{aligned}
&P\left\{ \left| \left(1 - \frac{\hat{g}}{g^\star}\right) (a(\mathsf{R}_{\hat{g},k}, D) - \overline{a}(D)) \right| \geq \lambda \sqrt{2 v_{g^\star,k}(D)}, |g^\star - \hat{g}| \leq c g^\star \mid D \right\} \\
&= P\left\{ \left| (a(\mathsf{R}_{\hat{g},k}, D) - \overline{a}(D)) \right| \geq \frac{\lambda}{c} \sqrt{2 v_{g^\star,k}(D)}, |g^\star - \hat{g}| \leq c g^\star \mid D \right\} \\
&\leq P\left\{ \max_{|g^\star - g| \leq c g^\star} \left| (a(\mathsf{R}_{g,k}, D) - \overline{a}(D)) \right| \geq \frac{\lambda}{c} \sqrt{2 v_{g^\star,k}(D)} \mid D \right\} \\
&\leq P\left\{ \max_{|g^\star - g| \leq c g^\star} \left| (a(\mathsf{R}_{g,k}, D) - \overline{a}(D)) \right| \geq \frac{\lambda}{c} \left(\frac{\xi}{z_{1-\beta/2}} \right) \mid D \right\} \\
&\leq \sum_{|g^\star - g| \leq c g^\star} P\left\{ \left| (a(\mathsf{R}_{g,k}, D) - \overline{a}(D)) \right| \geq \frac{\lambda}{c} \left(\frac{\xi}{z_{1-\beta/2}} \right) \right\}. \qquad \text{(B.29)}
\end{aligned}$$

By Theorem 3.1, we have that

$$P\left\{ \left| (a(\mathsf{R}_{g,k}, D) - \overline{a}(D)) \right| \geq \frac{\lambda}{c} \left(\frac{\xi}{z_{1-\beta/2}} \right) \right\} \leq 2 \exp\left(- \frac{g\delta}{2^4 (2-\varphi k-\varphi)^2 \Gamma_{k,\varphi,b}(1)} \frac{\lambda^2}{c^2} \left(\frac{\xi}{z_{1-\beta/2}} \right)^2 \right)$$

and so is (B.29) bounded from above by

$$\begin{aligned}
&4 c g^\star \exp\left(- \frac{g^\star (1-c)\delta}{4(2-\varphi k-\varphi)^2 \Gamma_{k,\varphi,b}(1)} \frac{\lambda^2}{c^2} \left(\frac{\xi}{z_{1-\beta/2}} \right)^2 \right) \\
&\lesssim v_{1,k}(D)\, c \left(\frac{z_{1-\beta/2}}{\xi} \right)^2 \exp\left(- \frac{\delta v_{1,k}(D)}{2^5 (2-\varphi k-\varphi)^2 \Gamma_{k,\varphi,b}(1)} \frac{\lambda^2}{c^2} \right)
\end{aligned}$$

if $1 - c \geq 1/2$ by the definition of $g^\star$. Thus, we have that

$$\begin{aligned}
&P\left\{ |Q(\mathsf{R}_{\hat{g},k}, D) - Q(\mathsf{R}'_{\hat{g}',k}, D)| \geq \lambda \sqrt{2 v_{g^\star,k}(D)} \mid D \right\} \\
&\lesssim P\left\{ \left| \left(1 - \frac{\hat{g}}{g^\star}\right) (a(\mathsf{R}_{\hat{g},k}, D) - \overline{a}(D)) \right| \geq \lambda \sqrt{2 v_{g^\star,k}(D)}, |g^\star - \hat{g}| \leq c g^\star \mid D \right\} + P\left\{ |g^\star - \hat{g}| \geq c g^\star \mid D \right\} \\
&\lesssim v_{1,k}(D)\, c \left(\frac{z_{1-\beta/2}}{\xi} \right)^2 \exp\left(- \frac{\delta v_{1,k}(D)}{2^5 (2-\varphi k-\varphi)^2 \Gamma^{(1)}_{k,\varphi,b}} \frac{\lambda^2}{c^2} \right) \qquad \text{(B.30)} \\
&\quad + \exp\left(- \frac{1}{C} \frac{\delta (v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4} \frac{z_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}} \frac{c^2}{\xi^2} \right) \qquad \text{(B.31)}
\end{aligned}$$

for some universal constant C by Lemma B.5, Part (ii). If c is chosen to satisfy (B.28), the term (B.31) will be bounded above by $\rho_{\varphi,k}(\xi,\beta \mid D)$. By plugging this value of c and $\lambda = \lambda_{\varphi,k}(\xi,\beta \mid D)$ into (B.30), we obtain the sub-polynomial term

$$\tilde{\rho}_{k,\varphi,b}(\xi,\beta \mid D) = \left(\frac{z_{1-\beta/2}^2}{\xi}\right)\left(\frac{(2-\varphi k-\varphi)^2(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{z_{1-\beta/2}\delta^{1/2}(v_{1,k}(D))^{1/2}}\right)\log^{1/2}\left(\rho_{k,\varphi,b}(\xi,\beta \mid D)^{-1}\right)$$
$$\exp\left(-\frac{1}{\xi}\frac{\delta^{1/2}(v_{1,k}(D))^{3/2}}{(2-\varphi k-\varphi)^2(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}\tilde{\lambda}^{U}_{k,\varphi,b}(D)\log\left(\rho_{k,\varphi,b}(\xi,\beta \mid D)^{-1}\right)\right) \tag{B.32}$$

as required. ■

APPENDIX C. PROOFS FOR AUXILIARY RESULTS

C.1 Proof of Lemma B.1. Let $T_i(\pi_{m,\ell}) = T(\mathsf{s}_i(\pi_{\ell,m}), D)$ and define $T_i(\pi'_{m,\ell})$ analogously. Define

$$R_m(\ell) = \left\{i \in [n] : \pi_{m,\ell}(i) \neq \pi'_{m,\ell}(i)\right\}$$

and let $\overline{R}_m(\ell) = |R_m(\ell)|$ denote the number of shared indices in the permutations in $\pi_{m,\ell}$ and $\pi'_{m,\ell}$. Observe that

$$\frac{1}{k}\sum_{i=1}^{k}\left(T_i(\pi_{m,\ell}) - T_i\left(\pi'_{m,\ell}\right)\right) = 0 \tag{C.1}$$

almost surely for all $m' \geq m$ if and only if $\overline{R}_m(\ell) = n$. Let $N(\ell)$ denote the values of the smallest index m with $\overline{R}_m(\ell) = n$. Observe that

$$\sum_{m=0}^{\infty}\left|\mathbb{E}\left[a(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D) - a(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D) \mid \boldsymbol{\pi}_0 = \boldsymbol{\psi}, \boldsymbol{\pi}'_0 = \boldsymbol{\psi}', D\right]\right| \tag{C.2}$$
$$\leq \sum_{m=0}^{\infty}\frac{1}{g}\frac{1}{k}\sum_{\ell=1}^{g}\sum_{i=1}^{k}\mathbb{E}\left[\left|\left(T_i(\pi_{m,\ell}) - T_i\left(\pi'_{m,\ell}\right)\right)\right| \mid \boldsymbol{\pi}_0 = \boldsymbol{\psi}, \boldsymbol{\pi}'_0 = \boldsymbol{\psi}', D\right]$$
$$\leq \frac{1}{g}\frac{1}{k}\sum_{\ell=1}^{g}\sum_{i=1}^{k}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left|T(\mathsf{s}, D) - T\left(\mathsf{s}', D\right)\right|\mathbb{E}[N(\ell)]$$

Hence, it suffices to bound the quantity $\mathbb{E}[N(\ell)]$.

To that end, let $N_r(\ell)$ be the value of the smallest index m with $\overline{R}_m(\ell) \geq r$. We proceed analogously to standard analysis of the coupon collector's problem (see e.g., Section 2.2 of Levin and Peres, 2017). We can evaluate

$$P\left\{\overline{R}_m(\ell) \geq r+1 \mid \overline{R}_{m-1}(\ell) = r\right\} = \frac{1}{g}P\left\{\pi_\ell(I_m) \neq \pi'_\ell(I_m), \pi_\ell^{-1}(J_m) \neq \pi_\ell'^{-1}(J_m)\right\} = \frac{(n-r)^2}{gn^2}$$

and

$$P\left\{\overline{R}_m(\ell) < r \mid \overline{R}_{m-1}(\ell) = r\right\} = 0,$$

and thereby obtain the bound

$$\mathbb{E}\left[N\left(\ell\right)\right] = \sum_{r=1}^{n} \mathbb{E}\left[N_r\left(\ell\right) - N_{r-1}\left(\ell\right)\right] \leq \sum_{r=1}^{n} \frac{gn^2}{\left(n-r+1\right)^2} = gn^2 \sum_{r=1}^{n} \frac{1}{r^2} \leq 2gn^2. \tag{C.3}$$

Hence, we find that (C.2) is upper bounded by

$$2gn^2 \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} T\left(\mathsf{s}, D\right) - T\left(\mathsf{s}', D\right),$$

as required. ■

C.2 Proof of Lemma B.2. Observe that

$$\frac{\mathrm{d}}{\mathrm{d}t} e^{tx} e^{(1-t)y} = \left(x-y\right) e^{tx} e^{(1-t)y}.$$

Thus, we find that

$$\begin{aligned}
c\left(e^x - e^y\right) &= c \int_0^1 \left(\frac{\mathrm{d}}{\mathrm{d}t} e^{tx} e^{(1-t)y}\right) \mathrm{dt} && \text{(Fundamental Theorem of Calculus)} \\
&= c\left(x-y\right) \int_0^1 e^{tx} e^{(1-t)y} \mathrm{dt} \\
&\leq c\left(x-y\right) \int_0^1 \left(te^x + \left(1-t\right) e^y\right) && \text{(Convexity)} \\
&= \frac{c}{2}\left(x-y\right)\left(e^x + e^y\right) \\
&= \frac{1}{2}\left(\left(s^{-1}c^2\left(e^x + e^y\right)\right)\left(s\left(x-y\right)^2\left(e^x + e^y\right)\right)\right)^{1/2} \\
&\leq \frac{1}{4}\left(\left(s^{-1}c^2\right) + s\left(x-y\right)^2\right)\left(e^x + e^y\right), && \text{(AM-GM)}
\end{aligned}$$

as required. ■

C.3 Proof of Lemma B.3. To begin, consider any collection of real numbers $x_1, \ldots, x_{2^c}$, for some positive integer c. For any integer $s > 0$, we have

$$\begin{aligned}
\left(x_1 \cdots x_{2^c}\right)^{2s} &\leq \frac{1}{4}\left(\left(x_{i_1} \cdots x_{i_{2^{c-1}}}\right)^{2s} + \left(x_{i_{2^{c-1}+1}} \cdots x_{2^c}\right)^{2s}\right)^2 && \text{(Young's Inequality)} \\
&\leq \frac{1}{4}\left(\left(x_{i_1} \cdots x_{i_{2^{c-1}}}\right)^{2(s+1)} + 2\left(x_1 \cdots x_{2^c}\right)^{2s} + \left(x_{i_{2^{c-1}+1}} \cdots x_{2^c}\right)^{2(s+1)}\right)
\end{aligned}$$

and so

$$\left(x_1 \cdots x_{2^c}\right)^{2s} \leq \frac{1}{2}\left(x_{i_1} \cdots x_{i_{2^{c-1}}}\right)^{2(s+1)} + \frac{1}{2}\left(x_{i_{2^{c-1}+1}} \cdots x_{2^c}\right)^{2(s+1)}. \tag{C.4}$$

Consequently, we have that

$$x_1 \cdots x_{2^c} \leq \frac{1}{2^c} \sum_{i=1}^{2^c} x_i^{2c} \tag{C.5}$$

through 2^r applications of (C.4).

Now, to prove the Lemma, we may assume without loss that $h_i > 0$ for all i by continuity. Observe that

$$\mathbb{E}\left[\left(\sum_{m=0}^{\infty} X_m\right)^{2^c}\right] = \sum_{i_1=0}^{\infty} \cdots \sum_{i_{2^c}=0}^{\infty} \mathbb{E}\left[X_{i_1} \cdots X_{i_{2^c}}\right] \tag{C.6}$$

by dominated convergence. By writing

$$X_1 \cdots X_{2^c} = \prod_{j=1}^{2^c} \left(\frac{\prod_{k \neq j} h_k}{h_j^{2^{c-1}}}\right)^{1/2^c} X_j$$

we find that

$$\mathbb{E}\left[X_1 \cdots X_{2^c}\right] \leq \frac{1}{2^c} \sum_{j=1}^{2^c} \frac{\prod_{k \neq j} h_k}{h_j^{2^{c-1}}} h_j^{2^c} = \prod_{j=1}^{2^c} h_j$$

by (C.5). Hence, by (C.6), we have that

$$\mathbb{E}\left[\left(\sum_{m=0}^{\infty} X_m\right)^{2^c}\right] \leq \sum_{i_1=0}^{\infty} \cdots \sum_{i_{2^c}=0}^{\infty} \prod_{j=1}^{2^c} h_{i_j} = \left(\sum_{m=0}^{\infty} h_j\right)^{2^c},$$

as required. ■

C.4 Proof of Lemma B.4. We begin by constructing Stein representers $\tilde{V}(Z, Z')$ and $\check{V}(Z, Z')$ for the statistic $\tilde{v}_{g,k}(Z)$ and $\check{v}_{g,k}(Z)$, respectively. We will use the same exchangeable pair $(Z, Z')_{m \geq 0}$ and Markov chain $(Z_m, Z'_m)_{m \geq 1}$ defined in Section 5.1. The subsequent result follows from an argument similar to Lemma 5.1, again based on an idea expressed by Lemma 4.1 of Chatterjee (2005).

Lemma C.1.

(i) *Let ψ and ψ' be two sets, each containing g elements of $\mathcal{P}_n$. The inequalities*

$$\sum_{m=0}^{\infty} \left| \mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m) - \tilde{v}_{g,k}(Z'_m)\right) \mid Z_0 = Z, Z'_0 = Z'\right] \right| \leq n^2 \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T(\mathsf{s}, D) - T(\mathsf{s}', D)\right)^2 \quad \textit{and}$$

$$\sum_{m=0}^{\infty} \left| \mathbb{E}\left[\left(\check{v}_{g,k}(Z_m) - \check{v}_{g,k}(Z'_m)\right) \mid Z_0 = Z, Z'_0 = Z'\right] \right| \leq 2gn^2 \max_{\mathsf{s} \in \mathsf{S}_{n,b}} \left(T(\mathsf{s}, D) - T(\mathsf{s}', D)\right)^2$$

hold almost surely.

(ii) *The functions*

$$\tilde{V}(Z, Z') = \sum_{m=0}^{\infty} \mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m) - \tilde{v}_{g,k}(Z'_m)\right) \mid Z_0 = Z, Z'_0 = Z'\right] \quad \textit{and}$$

$$\check{V}(Z, Z') = \sum_{m=0}^{\infty} \mathbb{E}\left[\left(\check{v}_{g,k}(Z_m) - \check{v}_{g,k}(Z'_m)\right) \mid Z_0 = Z, Z'_0 = Z'\right]$$

are finite and satisfy the equalities

$$\mathbb{E}\left[\tilde{V}(Z, Z') \mid Z\right] = \tilde{v}_{g,k}(Z) - v_{g,k}(D) \quad \textit{and}$$

$$\mathbb{E}\left[\check{V}\left(Z,Z'\right)\mid Z\right]=\check{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)$$

almost surely.

Next, we apply the concentration inequality stated in Theorem 5.2, due to Chatterjee (2005, 2007). That is, to establish part (i) of the Lemma, it will suffice to characterize constants s and u that satisfy

$$U_{\tilde{v}}\left(Z\right)\leq s^{-1}u\quad\text{and}\quad U_{\tilde{V}}\left(Z\right)\leq su,$$

with probability $1-\delta$, where

$$U_{\tilde{v}}\left(Z\right)=\frac{1}{2}\mathbb{E}\left[\left(\tilde{v}\left(Z\right)-\tilde{v}\left(Z'\right)\right)\mid Z\right]\quad\text{and}\quad U_{\tilde{V}}\left(Z\right)=\frac{1}{2}\mathbb{E}\left[\tilde{V}\left(Z,Z'\right)^{2}\mid Z\right].$$

The same statement holds for part (ii) of the Lemma, for the objects $U_{\check{v}}\left(Z\right)$ and $U_{\check{V}}\left(Z\right)$ defined analogously.

We obtain such a characterization through the application of Lemma 5.3. To this end, observe that, by Lemma C.1, part (i), the bound

$$\left(\sum_{m=0}^{\infty}\mathbb{E}\left[\left(\tilde{v}_{g,k}\left(Z_{m}\right)-\tilde{v}_{g,k}\left(Z'_{m}\right)\right)\mid Z_{0}=Z,Z'_{0}=Z'\right]\right)^{2}\leq n^{4}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^{4}\tag{C.7}$$

holds almost surely. The right hand side of (C.7) is square integrable, as the eighth-order split-stability $\zeta^{(8)}$ is finite almost surely. An analogous statement holds for the statistic $\check{v}_{g,k}\left(Z\right)$. Thus, deterministic bounds of the form (5.10) can be obtained through the application of the following Lemma.

Lemma C.2. *Suppose that Assumptions 3.1 and 3.2 hold and that the data D are independent and identically distributed. If the eighth-order split stability $\zeta^{(8)}$ is finite, then*
***(i)** The inequality*

$$\mathbb{E}\left[\mathbb{E}\left[\tilde{v}_{g,k}(Z_{m})-\tilde{v}_{g,k}(Z'_{m})\mid Z,Z'\right]^{2}\mid\boldsymbol{\pi}\right]\lesssim\left(1-\frac{2}{gn^{2}}\right)^{2m}\left(\frac{\left(2-\varphi k-\varphi\right)^{4}}{n^{2}g^{4}}\right)\Gamma_{k,\varphi,b}^{(2)},$$

holds almost surely for all integers $m\geq 0$, and
***(ii)** The inequality*

$$\mathbb{E}\left[\mathbb{E}\left[\check{v}_{g,k}(Z_{m})-\check{v}_{g,k}(Z'_{m})\mid Z,Z'\right]^{2}\mid\boldsymbol{\pi}\right]\lesssim\left(1-\frac{2}{gn^{2}}\right)^{2m}\left(\frac{\left(2-\varphi k-\varphi\right)^{4}}{n^{2}g^{3}}\right)\Gamma_{k,\varphi,b}^{(2)},$$

holds almost surely for all integers $m\geq 0$.

Thus, by Lemma 5.3, part (i), the inequalities

$$\begin{aligned}U_{\tilde{V}}\left(Z\right)&\lesssim\frac{1}{\delta}\left(\frac{gn^{2}}{2}\right)^{2}\left(\frac{\left(2-\varphi k-\varphi\right)^{4}}{n^{2}g^{4}}\right)\Gamma_{k,\varphi,b}^{(2)}\\&\lesssim\left(\frac{gn^{2}}{2}\right)\left(\frac{1}{\delta}\frac{\left(2-\varphi k-\varphi\right)^{4}}{g^{3}}\right)\Gamma_{k,\varphi,b}^{(2)}\end{aligned}$$

and

$$U_{\tilde{v}}(Z) \lesssim \frac{1}{\delta}\left(\frac{(2-\varphi k-\varphi)^4}{n^2 g^4}\right)\Gamma^{(2)}_{k,\varphi,b}$$
$$\lesssim \left(\frac{2}{gn^2}\right)\left(\frac{1}{\delta}\frac{(2-\varphi k-\varphi)^4}{g^3}\right)\Gamma^{(2)}_{k,\varphi,b}$$

both hold with probability greater than $1-\delta$. Hence, we obtain the bound (B.23) by applying Theorem 5.2 and choosing $s = gn^2/2$ and

$$u = \frac{1}{\delta}\frac{(2-\varphi k-\varphi)^4}{g^3}\Gamma^{(2)}_{k,\varphi,b}\,.$$

Analogous inequalities for the objects $U_{\check{v}}(Z)$ and $U_{\check{V}}(Z)$ hold by Lemma 5.3, part (ii), where in that case

$$u = \frac{1}{\delta}\frac{(2-\varphi k-\varphi)^4}{g^2}n^4\Gamma^{(2)}_{k,\varphi,b}\,.$$

which similarly implies the bound (B.24) by Theorem 5.2. ■

C.5 Proof of Lemma C.1. Reinstate the notation of the proof of Lemma B.1. We begin by noting that

$$\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^2 = \max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-\overline{a}(D)\right)^2+\left(T(\mathsf{s}',D)-\overline{a}(D)\right)^2$$
$$+2\left(T(\mathsf{s},D)-\overline{a}(D)\right)\left(T(\mathsf{s}',D)-\overline{a}(D)\right)$$
$$\geq 4\max_{\mathsf{s}\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-\overline{a}(D)\right)^2.$$

Observe that

$$\sum_{i,i'=1}^{k}\left(T_i(\pi_{m,\ell})-\overline{a}(D)\right)\left(T_i(\pi_{m,\ell})-\overline{a}(D)\right)-\sum_{i,i'=1}^{k}\left(T_i\left(\pi'_{m,\ell}\right)-\overline{a}(D)\right)\left(T_{i'}\left(\pi'_{m,\ell}\right)-\overline{a}(D)\right)=0$$

for all for all $m'\geq m$ if and only if $\overline{R}_m(\ell)=n$. Thus, we have that

$$\sum_{m=0}^{\infty}\left|\mathbb{E}\left[\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\mid Z_0=\left(\mathsf{r}(\boldsymbol{\psi}),D\right),Z'_0=\left(\mathsf{r}\left(\boldsymbol{\psi}'\right),D\right)\right]\right|$$
$$\leq\sum_{m=0}^{\infty}\frac{1}{g^2k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\mathbb{E}\Bigg[\Bigg|\left(T_i(\pi_{m,\ell})-\overline{a}(D)\right)\left(T_{i'}(\pi_{m,\ell})-\overline{a}(D)\right)$$
$$-\left(T_i\left(\pi'_{m,\ell}\right)-\overline{a}(D)\right)\left(T_{i'}\left(\pi'_{m,\ell}\right)-\overline{a}(D)\right)\Bigg|\mid Z_0=\left(\mathsf{r}(\boldsymbol{\psi}),D\right),Z'_0=\left(\mathsf{r}\left(\boldsymbol{\psi}'\right),D\right)\Bigg]$$
$$\leq\frac{2}{g^2k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left|\left(T(\mathsf{s},D)-\overline{a}(D)\right)\left(T\left(\mathsf{s}',D\right)-\overline{a}(D)\right)\right|\mathbb{E}\left[N(\ell)\right].$$
$$\leq\frac{2}{g}\max_{\mathsf{s}\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-\overline{a}(D)\right)^2\mathbb{E}\left[N(\ell)\right].$$
$$\leq\frac{1}{2g}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^2\mathbb{E}\left[N(\ell)\right].$$

Similarly, we have that

$$\begin{aligned}
&\sum_{m=0}^{\infty}\left|\mathbb{E}\left[\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z'_m)\mid Z_0=\left(\mathsf{r}\left(\boldsymbol{\psi}\right),D\right),Z'_0=\left(\mathsf{r}\left(\boldsymbol{\psi}'\right),D\right)\right]\right| \\
&\leq \sum_{m=0}^{\infty}\frac{1}{g^2k^2}\sum_{\ell,\ell'=1}^{g}\sum_{i,i'=1}^{k}\mathbb{E}\Big[\Big|\left(T_i\left(\pi_{m,\ell}\right)-\overline{a}\left(D\right)\right)\left(T_{i'}\left(\pi_{m,\ell'}\right)-\overline{a}\left(D\right)\right) \\
&\qquad\qquad -\left(T_i\left(\pi'_{m,\ell}\right)-\overline{a}\left(D\right)\right)\left(T_{i'}\left(\pi'_{m,\ell'}\right)-\overline{a}\left(D\right)\right)\Big|\mid Z_0=\left(\mathsf{r}\left(\boldsymbol{\psi}\right),D\right),Z'_0=\left(\mathsf{r}\left(\boldsymbol{\psi}'\right),D\right)\Big] \\
&\leq \frac{2}{g^2k^2}\sum_{\ell,\ell'=1}^{g}\sum_{i,i'=1}^{k}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left|\left(T\left(\mathsf{s},D\right)-\overline{a}\left(D\right)\right)\left(T\left(\mathsf{s}',D\right)-\overline{a}\left(D\right)\right)\right|\mathbb{E}\left[\max\left\{N\left(\ell\right),N\left(\ell'\right)\right\}\right] \\
&\leq \frac{2}{g^2k^2}\sum_{\ell,\ell'=1}^{g}\sum_{i,i'=1}^{k}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left|\left(T\left(\mathsf{s},D\right)-\overline{a}\left(D\right)\right)\left(T\left(\mathsf{s}',D\right)-\overline{a}\left(D\right)\right)\right|\mathbb{E}\left[N\left(\ell\right)+N\left(\ell'\right)\right] \\
&\leq \frac{1}{2}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^2\left|\mathbb{E}\left[N\left(\ell\right)+N\left(\ell'\right)\right]\right.
\end{aligned}$$

Thus, part (i) of the Lemma follows by the bound (C.3). Part (ii) of the Lemma then follows by an argument analogous to the proof of Lemma 5.1. ■

C.6 Proof of Lemma C.2. We reinstate the notation introduced in the proof of Lemma 5.4 from Section B.5. First, observe that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\mid Z,Z'\right]^2\mid\boldsymbol{\pi}\right] \\
&=\mathbb{E}\left[\left(P\left\{\mathcal{H}_m\right\}\mathbb{E}\left[\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\mid\mathcal{H}_m,Z,Z'\right]\right)^2\mid\boldsymbol{\pi}\right] \\
&\leq\left(1-\frac{2}{gn^2}\right)^{2m}\left(\frac{2kb\left(2n-bk-b\right)}{n^2}\right)^2\mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\right)^2\mid\mathcal{H}_m,\boldsymbol{\pi}\right] \qquad \text{(C.8)}
\end{aligned}$$

as before. Recall the notation

$$\overline{a}(\mathsf{r}_\ell,D)=\frac{1}{k}\sum_{i=1}^{k}(T(\mathsf{s}_{\ell,i},D)-\overline{a}(D))$$

and observe that

$$\begin{aligned}
&\mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\right)^2\mid\mathcal{H}_m,\boldsymbol{\pi}\right] \\
&=\frac{1}{g^4}\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)^2-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)^2\right)^2\mid\mathcal{H}_m\right] \\
&=\frac{1}{g^4}\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)+\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^2\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^2\mid\mathcal{H}_m\right] \\
&\leq\frac{1}{g^4}\left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)+\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^4\mid\mathcal{H}_m\right]\right)^{1/2} \qquad \text{(Cauchy-Schwarz)} \\
&\quad\cdot\left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^4\mid\mathcal{H}_m\right]\right)^{1/2}
\end{aligned}$$

$$\leq \frac{2^4}{g^4}\left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(D)\right)^4\right]\right)^{1/2} \qquad \text{(Hölder)}$$

$$\cdot\left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2}. \qquad \text{(C.9)}$$

Observe that

$$\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(D)\right)^4\right] \leq 3^2\left(2^4(2-\varphi k-\varphi)^2\right)^2\Gamma^{(2)}_{k,\varphi,b}$$

by Theorem 3.2, as Assumption 3.1 is maintained and the eighth-order sample-split stability $\zeta^{(8)}$ is finite. In turn, we have that

$$\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^4 \mid \mathcal{H}_m\right] \leq \frac{4}{k^4b^4}\Gamma^{(2)}_{k,\varphi,b} \qquad \text{(C.10)}$$

by (B.14), (B.20), and (B.21). Hence, we have that

$$\mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m,\boldsymbol{\pi}\right] \leq \frac{3\cdot 2^9}{k^2b^2g^4}(2-\varphi k-\varphi)^2\Gamma^{(2)}_{k,\varphi,b}\,. \qquad \text{(C.11)}$$

Combining (C.8), (C.10), (C.11), we find that

$$\mathbb{E}\left[\mathbb{E}\left[\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\mid Z,Z'\right]^2 \mid \boldsymbol{\pi}\right] \leq \left(1-\frac{2}{gn^2}\right)^{2m}\left(\frac{\left(3\cdot 2^{11}\right)(2-\varphi k-\varphi)^4}{n^2g^4}\right)\Gamma^{(2)}_{k,\varphi,b}\,,$$

which completes the proof of the first part of the Lemma.

Second, following the same argument, we again have that

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z'_m)\mid Z,Z'\right]^2 \mid \boldsymbol{\pi}\right]\\ &\leq \left(1-\frac{2}{gn^2}\right)^{2m}\left(\frac{2kb\,(2n-bk-b)}{n^2}\right)^2\mathbb{E}\left[\left(\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m,\boldsymbol{\pi}\right] \qquad \text{(C.12)}\end{aligned}$$

In this case, we can compute

$$\begin{aligned}&\mathbb{E}\left[\left(\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m,\boldsymbol{\pi}\right]\\ &=\mathbb{E}\left[\left(\left(a(Z_m)-\overline{a}(D)\right)^2-\left(a(Z'_m)-\overline{a}(D)\right)^2\right)^2 \mid \mathcal{H}_m\right]\\ &=\mathbb{E}\left[\left(a(Z_m)-\overline{a}(D)+a(Z'_m)-\overline{a}(D)\right)^2\left(a(Z_m)-a(Z'_m)\right)^2 \mid \mathcal{H}_m\right]\\ &\leq\left(\mathbb{E}\left[\left(a(Z_m)-\overline{a}(D)+a(Z'_m)-\overline{a}(D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} \qquad \text{(Cauchy-Schwarz)}\\ &\quad\cdot\left(\mathbb{E}\left[\left(a(Z_m)-a(Z'_m)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2}\\ &\leq 2^4\left(\mathbb{E}\left[\left(a(Z_m)-\overline{a}(D)\right)^4\right]\right)^{1/2}\left(\mathbb{E}\left[\left(a(Z_m)-a(Z'_m)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2}, \qquad \text{(C.13)}\end{aligned}$$

where the last inequality follows from Hölder's inequality. Again we have that

$$\mathbb{E}\left[\left(a(Z_m)-a(Z'_m)\right)^4\right] \leq 3^2\left(\frac{2^4\varphi^2(2-\varphi k-\varphi)^2}{g}\right)^2\Gamma^{(2)}_{k,\varphi,b} \qquad \text{(C.14)}$$

by Theorem 3.2, as Assumption 3.1 is maintained and the eighth-order sample-split stability $\zeta^{(8)}$ is finite. Similarly. we have that

$$\mathbb{E}\left[\left(a(Z_m)-\overline{a}(D)\right)^4 \mid \mathcal{H}_m\right] \leq \frac{4}{g^4k^4b^4}\Gamma^{(2)}_{k,\varphi,b} \tag{C.15}$$

by (B.14), (B.20), and (B.21). Combining (C.12), (C.14), (C.15), we find that

$$\mathbb{E}\left[\left(\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \leq \left(1-\frac{2}{gn^2}\right)^{2m}\left(\frac{\left(3\cdot 2^9\right)\varphi^4\left(2-\varphi k-\varphi\right)^4}{n^2g^3}\right)\Gamma^{(2)}_{k,\varphi,b}\,,$$

which completes the proof of the second part of the Lemma. ■

C.7 Proof of Lemma B.5, Part (i). Observe that

$$\begin{aligned}
&P\left\{|U(\mathsf{R}_{\hat{g},k},D)| \geq t\sqrt{v_{g^\star,k}(D)}, |\hat{g}-g^\star| \leq cg^\star \mid D\right\}\\
&= P\left\{\sum_{i=1}^{g}\overline{a}\left(\mathsf{r}_{i,k},D\right)-\sum_{i=1}^{g^\star}\overline{a}\left(\mathsf{r}_{i,k},D\right) \geq tg^\star\sqrt{v_{g^\star,k}(D)}, |\hat{g}-g^\star| \leq cg^\star \mid D\right\}\\
&\leq P\left\{\max_{g^\star(1-c)\leq g\leq g^\star}|\sum_{i=1}^{g}\overline{a}\left(\mathsf{r}_{i,k},D\right)-\sum_{i=1}^{g^\star}\overline{a}\left(\mathsf{r}_{i,k},D\right)| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\}\\
&+ P\left\{\max_{g^\star\leq g\leq g^\star(1+c)}|\sum_{i=1}^{g}\overline{a}\left(\mathsf{r}_{i,k},D\right)-\sum_{i=1}^{g^\star}\overline{a}\left(\mathsf{r}_{i,k},D\right)| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\}\\
&= P\left\{\max_{g^\star(1-c)\leq g\leq g^\star}|\sum_{i=1}^{g^\star-g}\overline{a}\left(\mathsf{r}_{i,k},D\right)| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\}\\
&+ P\left\{\max_{g^\star\leq g\leq g^\star(1+c)}|\sum_{i=1}^{g-g^\star}\overline{a}\left(\mathsf{r}_{i,k},D\right)| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\}.
\end{aligned} \tag{C.16}$$

To bound the two probabilities (C.16), we apply the following Chernoff-type variation to the Kolmogorov maximal inequality, due to Steiger (1970).

Theorem C.1 (Steiger, 1970)**.** *Let S_i, $i=1,2,\ldots$, be a real-valued martingale sequence. If the moment generating function*

$$m_n(\theta) = \mathbb{E}\left[\exp\left(\theta S_n\right)\right]$$

is finite for all positive θ, then the inequality

$$\log\ P\left\{\max_{1\leq n'\leq n} S_{n'} > t\right\} \leq \inf_{\theta>0}\left(\log m_n(\theta)-\theta t\right),$$

holds.

Conditional on D, the random variables

$$\overline{a}\left(\mathsf{r}_{i,k},D\right), \quad i=1,2,\ldots,$$

are mean-zero, independent, and identically distributed. Thus, the partial sums

$$S_m = \sum_{i=1}^{m} \overline{a}\left(\mathsf{r}_{i,k}, D\right)$$

are a martingale sequence. Moreover, though inspection of the proof of Lemma 5.2, we find that Theorem Theorem 3.1 implies that

$$\begin{aligned}
&\inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\theta S_{\lfloor cg^\star \rfloor}\right) \mid D\right] - \theta\tau\right) \\
&= \inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\frac{\theta}{\lfloor cg^\star \rfloor} S_{\lfloor cg^\star \rfloor}\right) \mid D\right] - \theta\frac{\tau}{\lfloor cg^\star \rfloor}\right) \\
&\leq -\frac{\lfloor cg^\star \rfloor}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}} \frac{\delta\tau^2}{\left(\lfloor cg^\star \rfloor\right)^2} \\
&\leq -\frac{\delta}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}} \frac{\tau^2}{cg^\star}
\end{aligned}$$

with probability greater than $1-\delta$, as Assumption 3.1 holds and the fourth-order split stability $\zeta^{(4)}$ is finite. Thus, by setting

$$\tau = tg^\star\sqrt{v_{g^\star,k}\left(D\right)} = \sqrt{g^\star}t\left(v_{1,k}(D)\right)^{1/2}$$

we find that

$$\begin{aligned}
&\log\frac{1}{2}P\left\{|U(\mathsf{R}_{\hat{g},k}, D)| \geq t\sqrt{v_{g^\star,k}\left(D\right)}, |\hat{g}-g^\star| \leq cg^\star \mid D\right\} \\
&\leq -\frac{\delta v_{1,k}(D)}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}} \frac{t^2}{c}
\end{aligned}$$

with probability greater than $1-\delta$, by Theorem C.1 and the inequality (C.16).

C.8 Proof of Lemma B.5, Part (ii). Observe that

$$P\left\{|\hat{g}-g^\star| > cg^\star \mid D\right\} = P\left\{g^\star\left(1+c\right) < \hat{g} \mid D\right\} + P\left\{\hat{g} < g^\star\left(1-c\right) \mid D\right\}. \tag{C.17}$$

We begin by handling the first term. We have that

$$\begin{aligned}
P\left\{g^\star\left(1+c\right) < \hat{g} \mid D\right\} &\leq P\left\{\hat{v}_{g^\star(1+c),k}\left(Z\right) - v_{g^\star(1+c),k}\left(D\right) + v_{g^\star(1+c),k}\left(D\right) > \frac{1}{2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \mid D\right\} \\
&\leq P\left\{\hat{v}_{g^\star(1+c),k}\left(Z\right) - v_{g^\star(1+c),k}\left(D\right) > \frac{v_{1,k}(D)}{g^\star} - \frac{v_{1,k}(D)}{g^\star\left(1+c\right)} \mid D\right\} \\
&= P\left\{\hat{v}_{g^\star(1+c),k}\left(Z\right) - v_{g^\star(1+c),k}\left(D\right) > \frac{v_{1,k}(D)}{g^\star}\frac{c}{1+c} \mid D\right\},
\end{aligned}$$

where the first inequality follows the definition of $g^\star$. Thus, as Assumption 3.1 holds and the eighth-order split stability $\zeta^{(8)}$ is finite, we have that

$$\begin{aligned}\log \frac{1}{4} P\left\{g^\star\left(1+c\right) < \hat{g} \mid D\right\} &\lesssim -\frac{\delta(v_{1,k}(D))^2}{(2-\varphi k-\varphi)^4}\frac{g^\star(1+c)c^2}{\Gamma^{(2)}_{k,\varphi,b}} \\ &\lesssim -\frac{\delta(v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4}\frac{z_{1-\beta/2}c^2}{\xi^2\Gamma^{(2)}_{k,\varphi,b}}\end{aligned}$$

with probability greater than $1-\delta$, by Lemma 5.5, the definition of $g^\star$ and the assumption that $0 < c < 1/2$.

Next, we bound the second term in (C.17). Define the partial sums

$$A_g = \sum_{\ell=1}^{g}\left(\frac{1}{k^2}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}\left(D\right)\right)\left(T(\mathsf{s}_{\ell,i'},D)-\overline{a}\left(D\right)\right)-v_{1,k}(D)\right) \quad \text{and}$$

$$\begin{aligned}B_g = \sum_{\ell=1}^{g}\Bigg(&\frac{1}{k^2}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}\left(D\right)\right)\left(T(\mathsf{s}_{\ell,i'},D)-\overline{a}\left(D\right)\right) \\ &+2\sum_{\ell'=1}^{\ell-1}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}\left(D\right)\right)\left(T(\mathsf{s}_{\ell',i'},D)-\overline{a}\left(D\right)\right)-v_{1,k}(D)\Bigg).\end{aligned}$$

Observe that the equality

$$\hat{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)=\frac{g}{g-1}\left(\tilde{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\right)-\frac{1}{g-1}\left(\check{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\right),$$

from the proof of Lemma 5.5, implies that

$$\begin{aligned}\hat{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right) &= \frac{1}{g\left(g-1\right)}\sum_{\ell=1}^{g}\frac{1}{k^2}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-a\left(Z\right)\right)\left(T(\mathsf{s}_{\ell,i'},D)-a\left(Z\right)\right)-\frac{v_{1,k}(D)}{g} \\ &= \frac{1}{g-1}\frac{1}{g}A_g-\frac{1}{g-1}\frac{1}{g^2}B_g\end{aligned}$$

for all positive g. Thus, we can write

$$\begin{aligned}&P\left\{\hat{g} < g^\star\left(1-c\right) \mid D\right\} \\ &= P\left\{\min_{2\le g'\le g^\star(1-c)}\hat{v}_{g',k}\left(Z\right)\le\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \mid D\right\} \\ &= P\left\{\min_{2\le g'\le g^\star(1-c)}\frac{\left(g'A_{g'}-B_{g'}\right)}{\left(g'-1\right)\left(g'\right)^2}+\frac{1}{g'}v_{1,k}(D)\le\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \mid D\right\},\end{aligned}$$

where the first equality follows from the definition of $\hat{g}$. Now, in the event that

$$\min_{2\le g'\le g^\star(1-c)}\frac{\left(g'A_{g'}-B_{g'}\right)}{\left(g'-1\right)\left(g'\right)^2}+\frac{1}{g'}v_{1,k}(D)\le\left(\frac{\xi}{z_{1-\beta/2}}\right)^2, \tag{C.18}$$

it must also be the case that

$$\min_{2\leq g'\leq g^\star(1-c)} \left(g'-1\right)\left(g'\right)^2\left(\frac{\left(g'A_{g'}-B_{g'}\right)}{\left(g'-1\right)\left(g'\right)^2}+\frac{1}{g'}v_{1,k}(D)\right) \leq (\hat{g}-1)(\hat{g})^2\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \tag{C.19}$$

as $\hat{g}\leq g^\star\left(1-c\right)$ necessarily. But then, similarly, (C.19) implies that

$$\min_{2\leq g'\leq g^\star(1-c)} \left(g'A_{g'}-B_{g'}\right)+\frac{(\hat{g}-1)(\hat{g})^2}{g'}v_{1,k}(D) \leq (\hat{g}-1)(\hat{g})^2\left(\frac{\xi}{z_{1-\beta/2}}\right)^2,$$

and in turn

$$\begin{aligned}
&\min_{2\leq g'\leq g^\star(1-c)} \frac{\left(g'A_{g'}-B_{g'}\right)}{\left(g^\star\left(1-c\right)-1\right)\left(g^\star\left(1-c\right)\right)^2} \\
&\leq \frac{(\hat{g}-1)(\hat{g})^2}{\left(g^\star\left(1-c\right)-1\right)\left(g^\star\left(1-c\right)\right)^2}\left(\frac{1}{g^\star-1}-\frac{1}{g^\star\left(1-c\right)}\right)v_{1,k}(D) \\
&= \frac{(\hat{g}-1)(\hat{g})^2}{\left(g^\star\left(1-c\right)-1\right)\left(g^\star\left(1-c\right)\right)^2}\left(\frac{1-g^\star c}{g^\star\left(g^\star-1\right)\left(1-c\right)}\right)v_{1,k}(D),
\end{aligned}$$

are then also true. Finally, again as (C.18) is equivalent to $\hat{g}<g^\star\left(1-c\right)$, we have that

$$\begin{aligned}
&P\left\{\min_{2\leq g'\leq g^\star(1-c)} \frac{\left(g'A_{g'}-B_{g'}\right)}{\left(g'-1\right)\left(g'\right)^2}+\frac{1}{g'}v_{1,k}(D)\leq\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \mid D\right\} \\
&\leq P\left\{\min_{2\leq g'\leq g^\star(1-c)} \frac{g'A_{g'}-B_{g'}}{\left(g^\star\left(1-c\right)-1\right)\left(g^\star\left(1-c\right)\right)^2}\leq\left(\frac{1-g^\star c}{g^\star\left(g^\star-1\right)\left(1-c\right)}\right)v_{1,k}(D)\mid D\right\}.
\end{aligned}$$

Now, observe that we can write

$$\begin{aligned}
&P\left\{\min_{2\leq g'\leq g^\star(1-c)} \frac{g'A_{g'}-B_{g'}}{\left(g^\star\left(1-c\right)-1\right)\left(g^\star\left(1-c\right)\right)^2}\leq\left(\frac{1-g^\star c}{g^\star\left(g^\star-1\right)\left(1-c\right)}\right)v_{1,k}(D)\mid D\right\} \\
&= P\left\{\max_{2\leq g'\leq g^\star(1-c)} \frac{B_{g'}-g'A_{g'}}{\left(g^\star\left(1-c\right)-1\right)\left(g^\star\left(1-c\right)\right)^2}\geq\left(\frac{g^\star c-1}{g^\star\left(g^\star-1\right)\left(1-c\right)}\right)v_{1,k}(D)\mid D\right\}.
\end{aligned} \tag{C.20}$$

We bound (C.20) by combining the argument used to establish Lemma 5.5 with an application of Theorem C.1. To this end, observe that

$$\begin{aligned}
&P\left\{\max_{2\leq g'\leq g}\frac{B_{g'}-g'A_{g'}}{\left(g-1\right)g^2}\leq t\mid D\right\} \\
&\geq P\left\{\max_{2\leq g'\leq g}\frac{-1}{g-1}\frac{g'}{g^2}A_{g'}\leq\frac{1}{1+\sqrt{g}}t\mid D\right\}+P\left\{\max_{2\leq g'\leq g}\frac{1}{g-1}\frac{1}{g^2}B_{g'}\leq\frac{\sqrt{g}}{1+\sqrt{g}}t\mid D\right\}-1 \\
&\geq P\left\{\max_{2\leq g'\leq g}\frac{g}{g-1}\frac{-1}{g^2}A_{g'}\leq\frac{\sqrt{g}}{1+\sqrt{g}}t\mid D\right\}+P\left\{\max_{2\leq g'\leq g}\frac{1}{g-1}\frac{1}{g^2}B_{g'}\leq\frac{1}{1+\sqrt{g}}t\mid D\right\}-1 \\
&\geq P\left\{\max_{2\leq g'\leq g}-A_{g'}\leq g^2\frac{1}{\sqrt{g}}\frac{g-1}{1+\sqrt{g}}t\mid D\right\}+P\left\{\max_{2\leq g'\leq g}|B_{g'}|\leq g^2\frac{g-1}{1+\sqrt{g}}t\mid D\right\}-1
\end{aligned} \tag{C.21}$$

for any $t>0$ and any positive integer D. Observe that the partial sums A_g and B_g are both martingale sequences. As Assumption 3.1 holds and the eighth-order split stability $\zeta^{(8)}$ is finite, though inspection of the

proof of Lemma 5.2, Lemma B.4 implies that implies that

$$\begin{aligned}
&\inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\theta A_g\right) \mid D\right]-\theta g^2 \frac{1}{\sqrt{g}} \frac{g-1}{1+\sqrt{g}} t\right) \\
&=\inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\frac{\theta}{g^2} A_g\right) \mid D\right]-\theta \frac{1}{\sqrt{g}} \frac{g-1}{1+\sqrt{g}} t\right) \\
&\lesssim-\frac{\delta}{(2-\varphi k-\varphi)^4} \frac{g^3 t^2}{\Gamma_{k, \varphi, b}^{(2)}}
\end{aligned} \tag{C.22}$$

and

$$\begin{aligned}
&\inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\theta B_g\right) \mid D\right]-\theta g^2 \frac{g-1}{1+\sqrt{g}} t\right) \\
&=\inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\frac{\theta}{g^2} B_g\right) \mid D\right]-\theta \frac{g-1}{1+\sqrt{g}} t\right) \\
&\lesssim-\frac{\delta}{(2-\varphi k-\varphi)^4} \frac{g^3 t^2}{\Gamma_{k, \varphi, b}^{(2)}}
\end{aligned} \tag{C.23}$$

each with probability greater than $1-\delta$, where we have used the facts that $4g \leq \left(1+\sqrt{g}\right)^2$ for all $g \geq 1$ and $(1/4)\, g^2 \geq (g-1)^2$ for all $g \geq 2$. Hence, by Theorem C.1, and plugging (C.22) and (C.23) into (C.21), we find that

$$\log \frac{1}{4} P\left\{\max_{2 \leq g' \leq g} \frac{B_{g'}-g' A_{g'}}{\left(g'-1\right)\left(g'\right)^2} \geq t \mid D\right\} \lesssim-\frac{\delta}{(2-\varphi k-\varphi)^4} \frac{g^3 t^2}{\Gamma_{k, \varphi, b}^{(2)}}$$

with probability greater than $1-\delta$. Consequently, by (C.20) and the definition of $g^\star$, we find that

$$\begin{aligned}
&\log \frac{1}{4} P\left\{g^\star-\hat{g} \geq c \mid D\right\} \\
&=\log \frac{1}{4} P\left\{\max_{2 \leq g' \leq g^\star(1-c)} \frac{B_{g'}-g' A_{g'}}{\left(g^\star(1-c)-1\right)\left(g^\star(1-c)\right)^2} \geq\left(\frac{g^\star c-1}{g^\star\left(g^\star-1\right)(1-c)}\right) v_{1, k}(D) \mid D\right\} \\
&\lesssim-\frac{\delta\left(v_{1, k}(D)\right)^2}{(2-\varphi k-\varphi)^4} \frac{g^\star(1-c)\left(g^\star c-1\right)^2}{\left(g^\star-1\right)^2 \Gamma_{k, \varphi, b}^{(2)}} \\
&\lesssim-\frac{\delta\left(v_{1, k}(D)\right)^2}{(2-\varphi k-\varphi)^4} \frac{g^\star c^2}{\Gamma_{k, \varphi, b}^{(2)}} \\
&\lesssim-\frac{\delta\left(v_{1, k}(D)\right)^3}{(2-\varphi k-\varphi)^4} \frac{z_{1-\beta / 2} c^2}{\xi^2 \Gamma_{k, \varphi, b}^{(2)}}
\end{aligned}$$

where in the second to last inequality we have used the facts that $\frac{1}{2} x \leq x-1$ for $x \geq 2$ and $g^\star c \geq 2$. Thus, putting the pieces together, we find that

$$\log \frac{1}{8} P\left\{\left|\hat{g}-g^\star\right|>c g^\star \mid D\right\} \lesssim-\frac{\delta\left(v_{1, k}(D)\right)^3}{(2-\varphi k-\varphi)^4} \frac{z_{1-\beta / 2} c^2}{\xi^2 \Gamma_{k, \varphi, b}^{(2)}}$$

with probability greater than $1-\delta$, as required. ■

APPENDIX D. ADDITIONAL RESULTS

D.1 Examples of Positive and Negative Conditional Covariance. In this appendix, we give simple examples where the conditional covariance $\gamma_{n,b}(D)$ obtains either endpoint of the Cauchy-Schwarz bound

$$-\phi_{n,b}(D) \leq \gamma_{n,b}(D) \leq \phi_{n,b}(D) \, .$$

Suppose that the data are

$$D_1 = -4, \quad D_2 = -2, \quad D_3 = 2, \quad \text{and} \quad D_4 = 4 \, .$$

Set $b = 2$ and consider the statistic

$$T(\mathsf{s}, D) = \big| \sum_{i \in \mathsf{s}} D_i \big|$$

Observe that

$$T(\mathsf{s}, D) = T(\tilde{\mathsf{s}}, D)$$

for each set s in $\{1, \ldots, 4\}$ of size 2. Thus, we have that

$$\gamma_{n,b}(D) = \mathrm{Cov}(T(\mathsf{s}, D), T(\tilde{\mathsf{s}}, D) \mid D) = \mathrm{Var}(T(\mathsf{s}, D) \mid D) = \phi_{n,b}(D)$$

giving one end of the inequality.[15] On the other hand, suppose that

$$T(\mathsf{s}, D) = \sum_{i \in \mathsf{s}} D_i \, .$$

In this case, we have that

$$T(\mathsf{s}, D) = -T(\tilde{\mathsf{s}}, D)$$

and thereby

$$\gamma_{n,b}(D) = \mathrm{Cov}(T(\mathsf{s}, D), T(\tilde{\mathsf{s}}, D) \mid D) = -\mathrm{Var}(T(\mathsf{s}, D) \mid D) = -\phi_{n,b}(D) \, ,$$

giving the other end of the inequality.

D.2 Validity of Testing Procedures Based on Multiple Sample-Splitting. In this appendix, we discuss the application of Algorithm 1 to testing procedures based on averaging over multiple splits of the same sample. We show that methods based on both p-values and e-values constructed with sample splitting continue to control the Type I error rate if they are aggregated sequentially with Algorithm 1. Both results follow from the "Exchangeable Markov Inequality" of Ramdas and Manole (2023). As before, let $D = (D_i)_{i=1}^n$ be independent and identically distributed according to a probability distribution P. Interest is in testing the null hypothesis $H_0 : P \in \mathbf{P}$ for some collection of probability distributions $\mathbf{P}$.

[15] Observe that this will occur for any mean zero data set.

D.2.1 Methods Based on p-Values. Suppose that we have access to a valid p-value $\hat{p}(\mathsf{s}, D)$. That is, the statistic $\hat{p}(\mathsf{s}, D)$ satisfies

$$P\left\{\hat{p}(\mathsf{s}, D) \leq u \mid D_{\bar{\mathsf{s}}}\right\} \leq u$$

for all u in $(0, 1)$ and P in $\mathbf{P}$. For example, a test-statistic could be chosen using the data in $D_{\bar{\mathsf{s}}}$ and a p-value can be constructed based on this test statistic using the data in D_{s}. For any collection $\mathsf{R}_{g,k}$ in $\mathcal{R}_{n,k,b}$, let

$$a_\delta(\mathsf{R}_{g,k}, D) = \frac{1}{g}\frac{1}{k}\sum_{i=1}^{g}\sum_{j=1}^{k}\mathbb{I}\left\{\hat{p}(\mathsf{s}_{i,j}, D) \leq \delta\right\} \tag{D.1}$$

denote the proportion of p-values that are less than or equal than δ. Rüger (1978), Meinshausen et al. (2009), and DiCiccio et al. (2020) observe that if $\mathsf{R}_{g,k}$ is constructed independently of the data D, then

$$P\left\{a_\delta(\mathsf{R}_{g,k}, D) \geq c\right\} \leq \frac{\mathbb{E}\left[a_\delta(\mathsf{R}_{g,k}, D)\right]}{c} = \frac{\delta}{c}$$

by Markov's inequality, for all P in $\mathbf{P}$. Thus, if δ and c are chosen such that $\delta/c = \alpha$, then the test that rejects the null hypothesis H_0 if $a_\delta(\mathsf{R}_{g,k}, D)$ is larger than c has level α.

The following theorem establishes that this test continues to be valid if the collection of sample-splits $\mathsf{R}_{g,k}$ is constructed sequentially with Algorithm 1.

Theorem D.1. *If the statistic $a_\delta(\mathsf{R}_{\hat{g},k}, D)$ defined in (D.1) is constructed sequentially with Algorithm 1, then*

$$P\left\{a_\delta(\mathsf{R}_{\hat{g},k}, D) \geq c\right\} \leq \frac{\delta}{c} \tag{D.2}$$

for all P in $\mathbf{P}$.

Proof. We apply the following inequality, due to Ramdas and Manole (2023).

Theorem D.2 (Theorem 1.1, Ramdas and Manole (2023))**.** *If $X_1, X_2, \ldots$ form an exchangeable sequence of integrable random variables, then*

$$P\left\{\exists t \geq 1 : \frac{1}{t}\sum_{i=1}^{t}|X_i| \geq 1/a\right\} \leq a\mathbb{E}\left[|X_i|\right] \tag{D.3}$$

for any $a > 0$.

Consequently, we have that

$$\begin{aligned} P\left\{a_\delta(\mathsf{R}_{\hat{g},k}, D) \geq c\right\} &\leq P\left\{\exists g \geq 1 : a_\delta(\mathsf{R}_{g,k}, D) \geq c\right\} \\ &\leq \frac{\mathbb{E}\left[\mathbb{I}\left\{\hat{p}(\mathsf{s}_{i,j}, D) \leq \delta\right\}\right]}{c} \leq \frac{\delta}{c} \end{aligned} \tag{D.4}$$

by Theorem D.2, as required. ∎

D.2.2 Methods Based on e-Values. Next, we consider settings where we have access to a valid e-value $\hat{e}(\mathsf{s}, D)$ (see Ramdas et al. (2023) for a recent review). That is, the nonnegative statistic $\hat{e}(\mathsf{s}, D)$ satisfies

$$\mathbb{E}_P\left[\hat{e}(\mathsf{s}, D)\right] \leq 1$$

for all P in $\mathbf{P}$. For example, this setting applies to the "Universal Inference" procedure of Wasserman et al. (2020). Here, an estimator $\hat{P}(\tilde{\mathsf{s}})$ of P is formed using the data $D_{\tilde{\mathsf{s}}}$ and is used in the split-likelihood ratio test statistic

$$\hat{e}(\mathsf{s}, D) = \inf_{P \in \mathbf{P}} \prod_{i \in \mathsf{s}} \frac{\mathrm{d}\hat{P}(\tilde{\mathsf{s}})}{\mathrm{d}P}(D_i) \,. \tag{D.5}$$

Wasserman et al. (2020) prove that (D.5) is an e-value. For any collection $\mathsf{R}_{g,k}$ in $\mathcal{R}_{n,k,b}$, let

$$a(\mathsf{R}_{g,k}, D) = \frac{1}{g}\frac{1}{k}\sum_{i=1}^{g}\sum_{j=1}^{k} \hat{e}(\mathsf{s}_{i,j}, D) \tag{D.6}$$

denote an aggregate e-value. See Dunn et al. (2023) and Tse and Davison (2022) for further discussion of aggregate e-values. Observe that

$$P\left\{a(\mathsf{R}_{g,k}, D) \geq 1/\alpha\right\} \leq \alpha \mathbb{E}\left[\hat{e}(\mathsf{s}, D)\right] = \alpha \,, \tag{D.7}$$

by Markov's inequality, for all P in $\mathbf{P}$. Thus, the test that rejects the null hypothesis H_0 if $a(\mathsf{R}_{g,k}, D)$ is larger than $1/\alpha$ has level α.

We again establish that this test continues to be valid if the collection of sample splits $\mathsf{R}_{g,k}$ is constructed sequentially with Algorithm 1.

Theorem D.3. *If the aggregate e-value $a_\delta(\mathsf{R}_{\hat{g},k}, D)$ defined in (D.6) is constructed sequentially with Algorithm 1, then*

$$P\left\{a(\mathsf{R}_{\hat{g},k}, D) \geq \frac{1}{\alpha}\right\} \leq \alpha \tag{D.8}$$

for all P in $\mathbf{P}$.

Proof. The claim is established with an argument very similar to the proof of Theorem D.1. Namely, the Markov inequality used to establish (D.7) can then be replaced by Theorem D.2, as before. ■

D.3 Stability of Regularized M-Estimation. In this appendix, we study the (r, q)th-order training sample $\sigma^{(r,q)}_{\mathsf{train}}$ defined in Definition 3.1. Our analysis is specialized to the case that the estimator $\hat{\eta}$ is a regularized empirical risk minimizer and is closely related to the proof of Proposition 4 of Austern and Zhou (2020).

Assume that the parameter η is an element of some closed convex space $H \subseteq \mathbb{R}^p$. Consider the estimator

$$\Psi\left(D_{\mathsf{s}}, \eta\right) = \frac{1}{b}\sum_{i \in \mathsf{s}} \psi\left(D_i, \eta\right) \tag{D.9}$$

$$\hat{\eta} = \operatorname*{arg\,min}_{\eta \in H}\left\{\frac{1}{n-b}\sum_{i \in \tilde{\mathsf{s}}} \ell\left(D_i, \eta\right) + \lambda_{1,n}\|\eta\|_1 + \lambda_{2,n}\|\eta\|_2\right\}, \tag{D.10}$$

where $\psi(\cdot,\cdot)$ are $\ell(\cdot,\cdot)$ functions and $\lambda_{1,n}, \lambda_{2,n} \geq 0$ are penalty parameters. Let $\nabla\ell(d,\eta)$ and $\nabla^2\ell(d,\eta)$ denote the gradient and Hessian of the function $\eta \mapsto \ell(d,\eta)$. Similarly, we write

$$\overline{\nabla}_{\tilde{\mathsf{s}}}\ell(D,\eta) = \frac{1}{n-b}\sum_{i\in\tilde{\mathsf{s}}}\nabla\ell(D_i,\eta) \quad \text{and}$$

$$\overline{\nabla}^2_{\tilde{\mathsf{s}}}\ell(D,\eta) = \frac{1}{n-b}\sum_{i\in\tilde{\mathsf{s}}}\nabla^2\ell(D_i,\eta)$$

for the empirical averages of the gradients and the Hessian evaluated on the data in the set $\tilde{\mathsf{s}}$. Define the moment

$$\kappa_r = \mathbb{E}\left[\left(\ell(D_i',\hat{\eta}') - \nabla\ell(D_i,\hat{\eta}')\right)^r\right]$$

for every even integer r.

Assumption D.1. *The minimum eigenvalue of $\overline{\nabla}^2_{\tilde{\mathsf{s}}}\ell(D,\eta)$ is bounded below by ρ almost surely.*

Assumption D.2. *For some strictly positive constant C_ψ, the function $\psi(\cdot,\cdot)$ satisfies the inequality*

$$|\psi(d,\eta) - \psi\left(d,\eta'\right)| \leq C_\psi\|\eta-\eta'\|_2$$

for each $\eta \in H$ and d.

Assumption D.3. *The loss function $\ell(\cdot,\cdot)$ is strictly convex and twice continuously differentiable in its second argument.*

Theorem D.4. *Under Assumptions D.1, D.2, and D.3, if $\rho + \lambda_{2,n} > 0$, then*

$$\sigma^{(r,q)}_{\mathsf{train}} \lesssim \left(\frac{q}{n-b}\frac{\kappa_r}{\rho+\lambda_{2,n}}\right)^r + \left(\lambda_{1,n}\frac{p}{\rho+\lambda_{2,n}}\right)^r$$

as $n \to \infty$, for all even integers r and positive integers $q \leq n-b$.

Remark D.1. Assumption D.1 implies that the dimension p of the parameter vector η is less than the number of observations n. To the best of our knowledge, analysis of the stability of ℓ_1 regularized empirical risk minimization in the regime with p larger than n is an open problem. Further analysis of this quantity in the high-dimensional regime is a promising direction for further research, given the widespread use of cross-validation for estimation of the prediction error for the lasso and the connections between stability and the concentration of cross-splitting established in this paper. ■

Remark D.2. A necessary and sufficient condition for $\hat{\eta}$ to be consistent for the population risk minimizer associated with (D.10) is that $\lambda_{1,n} = o\left((n-b)^{-1}\right)$. See e.g., Knight and Fu (2000). Thus, so long as the penalty $\lambda_{1,b}$ is chosen in this regime, the training sample stability $\sigma^{(r,q)}_{\mathsf{train}}$ will be

$$O\left(\left(\frac{q\kappa_r}{n-b}\right)^{-r}\right)$$

as required. ■

D.3.1 Proof of Theorem D.4. By Assumptions D.1 and D.3, the objective function for the estimator (D.10) is strongly convex. Thus, there is a unique solution to (D.10) for any data D. Let $\hat{\eta}$ and $\hat{\eta}'$ denote the solutions to (D.10) for the data D and $\tilde{D}^{(\mathsf{q})}$ respectively. Let $\partial f(x)$ denote the subgradient set of the function $x \mapsto f(x)$. The Karush-Kuhn-Tucker condition for the program (D.10) is given by

$$\overline{\nabla}_{\tilde{\mathsf{s}}} \ell(D, \eta) + \lambda_{2,n} \hat{\eta} + \lambda_{1,n} \hat{z} = 0, \tag{D.11}$$

where $\hat{z} \in \partial \|\hat{\eta}\|_1$ is the subgradient associated with the Lasso penalty. Observe that, in this case, $\hat{z} \in \operatorname{sign}(\hat{\eta})$, where we set $\operatorname{sign}(0) = [-1, 1]$. Let $\hat{z}$ and $\hat{z}'$ denote the subgradients obtained from D and $\tilde{D}^{(\mathsf{q})}$, respectively. As $\ell(d, \cdot)$ is twice continuously differentiable under Assumption D.3 (i), we have that

$$\begin{aligned}
&\overline{\nabla}_{\tilde{\mathsf{s}}} \ell\left(D, \hat{\eta}'\right) + \lambda_{2,n} \hat{\eta}' + \lambda_{1,n} \hat{z} \\
&= \overline{\nabla}_{\tilde{\mathsf{s}}} \ell(D, \hat{\eta}) + \lambda_{2,n} \hat{\eta} + \lambda_{1,n} \hat{z} + \left(\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}} \ell(D, \tilde{\eta}) + \lambda_{2,n} I_d\right)\left(\hat{\eta} - \hat{\eta}'\right) \\
&= \left(\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}} \ell(D, \tilde{\eta}) + \lambda_{2,n} I_d\right)\left(\hat{\eta} - \hat{\eta}'\right)
\end{aligned} \tag{D.12}$$

for some vector $\tilde{\eta}$, by a Taylor expansion and the optimality condition (D.11). On the other hand, we have that

$$\begin{aligned}
\overline{\nabla}_{\tilde{\mathsf{s}}} \ell\left(D, \hat{\eta}'\right) + \lambda_{2,n} \hat{\eta}' + \lambda_{1,n} \hat{z} &= \overline{\nabla}_{\tilde{\mathsf{s}}} \ell(\tilde{D}^{(\mathsf{q})}, \hat{\eta}') + \lambda_{2,n} \hat{\eta}' + \lambda_{1,n} \hat{z}' \\
&\quad + \frac{1}{n-b}\left(\sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}')\right) + \lambda_{1,n}\left(\hat{z} - \hat{z}'\right) \\
&= \frac{1}{n-b}\left(\sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}')\right) + \lambda_{1,n}\left(\hat{z} - \hat{z}'\right)
\end{aligned} \tag{D.13}$$

again by the optimality condition (D.11). Thus, we have that

$$\left(\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}} \ell(D, \tilde{\eta}) + \lambda_{2,n} I_p\right)\left(\hat{\eta} - \hat{\eta}'\right) = \frac{1}{n-b}\left(\sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}')\right) + \lambda_{1,n}\left(\hat{z} - \hat{z}'\right) \tag{D.14}$$

by (D.12) and (D.13). Observe that the the matrix $\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}} \ell(D, \tilde{\eta}) + \lambda_{2,n} I_p$ is invertible by Assumption D.1. Thus, we find that

$$\begin{aligned}
\hat{\eta} - \hat{\eta}' = &\frac{1}{n-b}\left(\overline{\nabla}^{(2)} \ell(D, \tilde{\eta}) + \lambda_2 I_p\right)^{-1}\left(\sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}')\right) \\
&+ \lambda_{1,n}\left(\overline{\nabla}^{(2)} \ell(D, \tilde{\eta}) + \lambda_2 I_d\right)^{-1}\left(\hat{z} - \hat{z}'\right)
\end{aligned}$$

and that consequently

$$\|\hat{\eta} - \hat{\eta}'\|_2 \lesssim \frac{1}{n-b} \frac{\sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}')}{(\rho + \lambda_{2,n})} + \lambda_{1,n} \frac{p}{\rho + \lambda_{2,n}} \tag{D.15}$$

by Assumption D.1. Hence, we have that

$$\begin{aligned}\sigma_{\mathsf{train}}^{(r,q)} &= \mathbb{E}\left[\left(T\left(\mathsf{s}, D\right) - T(\mathsf{s}, \tilde{D}^{(\mathsf{q})})\right)^r\right] \\ &\lesssim \mathbb{E}\left[\|\hat{\eta} - \hat{\eta}'\|_2^r\right] \\ &\lesssim \left(\frac{1}{(n-b)}\frac{q\kappa_r}{\rho + \lambda_{2,n}}\right)^r + \left(\lambda_{1,n}\frac{p}{\rho + \lambda_{2,n}}\right)^r,\end{aligned}$$

by Assumption D.2. ■

D.4 Comparison with Zhang (2022). We state a Berry-Esseen bound for $a\left(\mathsf{R}_{g,k}, D\right)$ through an application of a result due to Zhang (2022). In contrast to the bound stated in Corollary 3.1, the bound obtained here does not shrink as g increases. On the other hand, the bound obtained below is unconditional. It is straightforward to modify our argument to give an analogous high-probability conditional bound.

Theorem D.5. *Let W denote a standard normal random variable. Suppose that Assumptions 3.1 and 3.2 hold and that the data D are independent and identically distributed. If the eighth-order split stability $\zeta^{(8)}$ is finite, then the Berry-Esseen inequality*

$$d_{\mathrm{K}}\left(\frac{a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)}{\sqrt{\mathbb{E}\left[v_{g,k}\left(D\right)\right]}}, W\right) \leq \frac{4\left(2 - \varphi k - \varphi\right)^2\left(\Gamma_{k,\varphi,b}^{(2)}\right)^{1/2}}{\mathbb{E}\left[v_{1,k}(D)\right]}$$

is satisfied.

D.4.1 Proof of Theorem D.5. We apply the following central limit theorem, due to Zhang (2022). This result generalizes Theorem 2.1 of Shao and Zhang (2019) to accommodate general Stein representers.

Theorem D.6 (Theorem 4.1, Zhang, 2022)**.** *Let $\mathcal{X}$ be a separable metric space and suppose that (X, X') is an exchangeable pair of $\mathcal{X}$-valued random variables. Suppose that $f : \mathcal{X} \to \mathbb{R}$ and $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are square-integrable functions such that F is antisymmetric and*

$$\mathbb{E}\left[F\left(X, X'\right) \mid X\right] = f\left(X\right)$$

almost surely. Assume that $\mathrm{Var}\left(f\left(X\right)\right)$ is finite and non-zero and that $\mathbb{E}\left[f\left(X\right)\right] = 0$. Define the objects

$$\overline{f}\left(X\right) = f\left(X\right)/\sqrt{\mathrm{Var}\left(f\left(X\right)\right)}. \tag{D.16}$$

$$G\left(X\right) = \frac{1}{2}\mathbb{E}\left[\left(f\left(X\right) - f\left(X'\right)\right)F\left(X, X'\right) \mid X\right] \quad \textit{and} \tag{D.17}$$

$$\overline{G}\left(X\right) = \frac{1}{2}\mathbb{E}\left[\left(f\left(X\right) - f\left(X'\right)\right)\left|F\left(X, X'\right)\right| \mid X\right]. \tag{D.18}$$

Let W denote a standard normal random variable. The bound

$$d_K\left(\overline{f}\left(X\right), W\right) \leq \frac{\mathbb{E}\left[\left|G\left(X\right) - \mathbb{E}\left[G\left(X\right)\right]\right|\right] + \mathbb{E}\left[\left|\overline{G}\left(X\right)\right|\right]}{\mathrm{Var}\left(f\left(X\right)\right)} \tag{D.19}$$

is satisfied.

To this end, define the objects

$$B(Z) = \frac{1}{2}\mathbb{E}\left[\left(a(Z) - a\left(Z'\right)\right) A\left(Z, Z'\right) \mid Z\right] \quad \text{and}$$
$$\overline{B}(Z) = \frac{1}{2}\mathbb{E}\left[\left(a(Z) - a\left(Z'\right)\right) |A\left(Z, Z'\right)| \mid Z\right].$$

It will suffice to bound the quantity

$$\frac{\mathbb{E}\left[|B(Z) - \mathbb{E}\left[B(Z)\right]|\right] + \mathbb{E}\left[|\overline{B}(Z)|\right]}{\mathrm{Var}\left(a(Z) - \overline{a}(Z)\right)}. \tag{D.20}$$

Observe that

$$\mathbb{E}\left[|B(X) - \mathbb{E}\left[B(X)\right]|\right] \leq \sqrt{\mathrm{Var}\left(B(Z)\right)} \quad \text{and}$$
$$\mathbb{E}\left[|\overline{B}(Z)|\right] \leq \sqrt{\mathrm{Var}\left(\overline{B}(Z)\right)}$$

by the Cauchy-Schwarz inequality and the fact that $\mathbb{E}\left[\overline{B}(X)\right] = 0$ by exchangeability. Consequently, as

$$\mathrm{Var}\left(B(Z)\right) \leq \mathbb{E}\left[B(Z)^2\right] = \mathbb{E}\left[\left(a(Z) - a\left(Z'\right)\right)^2 A\left(Z, Z'\right)^2\right] = \mathrm{Var}\left(\overline{B}(Z)\right)$$

it will suffice to bound

$$2\mathbb{E}\left[\left(a(Z) - a\left(Z'\right)\right)^2 A\left(Z, Z'\right)^2\right]. \tag{D.21}$$

By Young's inequality, we have that

$$\mathbb{E}\left[\left(a(Z) - a\left(Z'\right)\right)^2 A\left(Z, Z'\right)^2\right] \leq \frac{1}{2}\left(s^{-1}\mathbb{E}\left[U_a^{(2)}(Z)\right] + s\mathbb{E}\left[U_A^{(2)}(Z)\right]\right),$$

where $U_a^{(2)}(Z)$ and $U_A^{(2)}(Z)$ are defined in Section 5.2.

Observe that the bound

$$\left(\sum_{m=0}^{\infty} \mathbb{E}\left[a(Z_m) - a\left(Z'_m\right) \mid Z_0 = Z, Z'_0 = Z'\right]\right)^4 \leq \left(2gn^2 \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T(\mathsf{s}, D) - T\left(\mathsf{s}', D\right)\right)\right)^4 \tag{D.22}$$

holds by Lemma B.1 and that the right-hand side of (D.22) is square-integrable as the eighth-order split stability $\zeta^{(8)}$ is finite. Thus, by combining Lemma B.3 and Lemma 5.4, we have that

$$\overline{U}_A^{(2)} \leq \left(\frac{gn^2}{2}\right)^2 \left(\frac{4(2n - bk - b)^2}{gn^2}\right)^2 \Gamma_{k,\varphi,b}^{(2)}$$

and

$$\overline{U}_a^{(2)} \leq \left(\frac{2}{gn^2}\right)^2 \left(\frac{4(2n - bk - b)^2}{gn^2}\right)^2 \Gamma_{k,\varphi,b}^{(2)}.$$

Hence, by taking $s = \left(gn^2/2\right)^2$, we find that (D.21) is bounded above by

$$\frac{4}{gn^2}\left((2n - bk - b)^2 \left(\Gamma_{k,\varphi,b}^{(2)}\right)^{1/2}\right)$$

Now, we have that

$$\text{Var}\left(a\left(Z\right)-\overline{a}\left(D\right)\right)=\mathbb{E}\left[v_{g,k}\left(D\right)\right]$$

by the law of total variance and the fact that $\mathbb{E}\left[a\left(Z\right)-\overline{a}\left(D\right)\mid D\right]=0$. Consequently we can decompose

$$\mathbb{E}\left[v_{g,k}\left(D\right)\right]=\frac{\mathbb{E}[\phi_{n,b}(D)]+\left(k-1\right)\mathbb{E}[\gamma_{n,b}(D)]}{kg}. \tag{D.23}$$

Hence, we have that (D.20) is bounded above by

$$\frac{4\left(2-\varphi k-\varphi\right)^{2}\left(\Gamma_{k,\varphi,b}^{(2)}\right)^{1/2}}{\mathbb{E}\left[v_{1,k}(D)\right]},$$

as required. ■