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 illustrate the application of the procedure to several widely applied econometric methods.

Keywords: Sample-splitting, Cross-Fitting, Cross-Validation, Reproducibility
JEL: C01, C13, C52

Date: November 15, 2024
*Email: ritzwoll@stanford.edu, romano@stanford.edu. We thank Sourav Chatterjee, Jiafeng Chen, Han Hong, Guido Imbens, Lihua Lei, Evan Munro, Aaditya Ramdas, Brad Ross and audiences at the Econometric Society North American and European summer meetings for helpful comments. We thank Bhaskar Chakravorty for sharing his data and code. 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).

1. INTRODUCTION

Sample-splitting is ubiquitous in modern econometric 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 the validity of 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-fitting for adaptive estimation of semiparametric models (Bickel, 1982; Schick, 1986; Chernozhukov et al., 2018), among a growing set of additional 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. 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?

This paper makes two contributions. First, we show that many widely applied sample-split econometric methods exhibit significant residual randomness. We give examples from the applied economics literature where the randomness induced by sample-splitting determines the statistical significance of a treatment effect estimated with cross-fitting, the interpretation of a model selected with cross-validation, and the qualitative features of treatment targeting rules learned from cross-fit nuisance parameter estimates.

Second, and accordingly, we propose an efficient method for removing the residual randomness from sample-split statistics. The procedure takes as input a bound and an error rate. The statistic of interest is sequentially aggregated over randomly drawn splits of the sample. The procedure is stopped after an estimate of the residual variation of the aggregate statistic falls below a pre-determined threshold. If the procedure were run twice, we show that the chance that the outputs differ by more than the bound is well-approximated by the error rate. That is, by setting the bound and error rate to be sufficiently small, sample-split statistics aggregated with the procedure are reproducible.

We begin, in Section 2, by discussing several widely applied sample-split econometric methods and demonstrating that, for each, the randomness induced by sample-splitting can substantively effect results. In Section 3, we propose an efficient method for sequentially aggregating sample-split statistics that ensures that the residual randomness is small. We illustrate that, in each of our examples, the proposed method stabilizes results—ensuring reproducibility at a minimal computational expense. We establish that the procedure is valid, in a particular asymptotic sense, without imposing any restrictions on the data generating process or statistic of interest. Similarly, we show that, for a large class of applications, sample-split statistics aggregated with the procedure

maintain (or, improve upon) unconditional statistical guarantees. That is, reproducibly aggregated sample-split statistics are still consistent and associated inferences are still valid.

To implement the procedure, a user must make several choices that may affect performance, including the specification of a suitable bound and error rate. Additionally, in most applications, the procedure is applied to stabilize a statistic that is itself constructed with cross-splitting. In these cases, users must also specify how many folds to use for cross-splitting.

To shed light on these choices, in Section 4, we give an analysis of the performance of the procedure under a set of simplifying conditions. We give two main results. First, the computation needed to achieve a given bound on residual randomness is very sensitive to the desired error tolerance, but is insensitive to the number of folds that are used for cross-splitting. Second, and on the other hand, the accuracy of the nominal error rate of the procedure deteriorates as the number of folds used for cross-splitting increases. We conclude, in Section 5, by synthesizing these results into a set of concrete recommendations for practice. We emphasize simple rules-of-thumb for choosing suitable error tolerances and approaches to cross-splitting.

The main theoretical challenge posed by this analysis is the accommodation of the dependence between statistics computed on cross-splits of a sample. We address this through an application of the method of exchangeable pairs (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), that may be of independent interest.

1.1 Related Literature

The procedure studied in this paper is applicable to a large variety of sample-split statistical methods. In Section 2, we give a selective review of various sample-split methods that are frequently used in applied economics. There are many, additional, sample-split methods, proposed in the statistics literature, that have the potential to be useful in econometric applications. Generic sample-split methods for testing statistical hypotheses are studied in Guo and Romano (2017), DiCiccio et al. (2020), and Wasserman et al. (2020). Additional applications include sample-split procedures for selective inference (Rinaldo et al., 2019), inference on high-dimensional linear models (Meinshausen and Bühlmann, 2010), conformal and predictive inference (Lei et al., 2018), and knockoff tests of conditional independence (Barber and Candès, 2015).

Chernozhukov et al. (2018) and Chernozhukov et al. (2023) advocate for the aggregation of estimators and p-values computed with sample splitting, over a pre-determined number of splits, in the context of applications related to semiparametric estimation and characterization of treatment effect heterogeneity, respectively. We second these recommendations and contribute a general purpose method that provides a statistical guarantee that residual randomness has been controlled up to a specified level of error, at a minimal computational cost.

Our setting is related to a large literature that studies methods for constructing confidence intervals for cross-validated estimates of generalization error. Several examples include Dietterich (1998), Nadeau and Bengio (1999), Lei (2020), Bayle et al. (2020), Austern and Zhou (2020), and Bates et al. (2023). By contrast, we are interested in the randomness conditional on the data. Formally, our non-asymptotic 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 4. Our bounds apply under weaker conditions and are substantially simpler.

Some of our results 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). Some of the concentration inequalities that 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, and resultant bounds, are substantially simpler.

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 (Motwani and Raghavan, 1995). See Beran and Millar (1987) for a classical analysis of the asymptotics of randomized tests and estimators.

2. The Residual Randomness of Sample-Split Statistics

Sample-splitting has proliferated as a useful subroutine for simplifying various tasks associated with modern statistical inference. The high-level situation is as follows. Consider a researcher who observes the independent data $D = (D_i)_{i=1}^n$ and wishes to report the statistic

$$\Psi(D, \eta) \,, \tag{2.1}$$

where η is some unknown nuisance parameter.

For example, each observation D_i could contain a measurement of an outcome Y_i, a treatment W_i, and a vector collecting a large number of covariates X_i. The researcher could be interested in measuring the effect of W_i on Y_i. To ensure that their estimate is not unnecessarily imprecise, they might like to include controls for only the subset of covariates that are correlated with the outcome. Here, formally, the nuisance parameter η collects the indices of the subset of "relevant" controls and the statistic $\Psi(D, \eta)$ denotes a treatment effect estimate associated with, say, a regression of the outcome on the treatment with controls for covariates with indices in η.

In practice, the researcher cannot compute the statistic (2.1), as the nuisance parameter η is unknown. Often, however, an estimator $\hat{\eta}(D)$ is available and so it may be tempting to report the feasible statistic

$$\Psi(D, \hat{\eta}(D)) \,. \tag{2.2}$$

In the example, the estimator $\hat{\eta}(D)$ could collect the indices of the covariates whose absolute sample correlation with the outcome is greater than some pre-determined threshold.

Statistical inferences that, counterfactually, would be appropriate if the infeasible statistic (2.1) were available will not necessarily be valid if they were instead based on the feasible statistic (2.2). In particular, the process of computing the nuisance parameter estimate $\hat{\eta}(D)$ might produce a confounding effect. In the example, screening control variables according to their correlation with the outcome produces a bias in the resultant treatment effect estimate.

Sample-splitting solves this problem.[1] Let s denote a randomly drawn subset of the numbers $[n] = \{1, \ldots, n\}$ of size b and let $\tilde{\mathsf{s}}$ denote its complement. Sample-split statistics take the form

$$T(\mathsf{s}, D) = \Psi(D_{\mathsf{s}}, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \ , \tag{2.3}$$

where the quantities $D_{\mathsf{s}} = (D_i)_{i \in \mathsf{s}}$ and $D_{\tilde{\mathsf{s}}} = (D_i)_{i \in \tilde{\mathsf{s}}}$ collect the data with indices in s and $\tilde{\mathsf{s}}$, respectively. Splitting the data D into two independent subsamples $(D_{\mathsf{s}}, D_{\tilde{\mathsf{s}}})$ prevents the construction of nuisance parameter estimates from contaminating statistical inferences that the researcher may wish to make with the feasible statistic (2.3).

There are two, well-known, practical issues with this approach. First, by splitting the data, statistical precision may be meaningfully reduced. Second, the statistic (2.3) is random through both the data D and the choice of the random subset s. That is, if the same sample-split statistic were computed on the same data by two different researchers, and the statistic was sensitive to the choice of the random subset s, then the researchers could report meaningfully different results.

To address these concerns, researchers often aggregate several replications of sample-split statistics through cross-splitting (Stone, 1974; Schick, 1986). In particular, researchers typically report aggregate statistics of the form

$$a(\mathsf{r}, D) = \frac{1}{k} \sum_{j=1}^{k} T(\mathsf{s}_j, D) \ , \tag{2.4}$$

where the quantity $\mathsf{r} = (\mathsf{s}_j)_{j=1}^{k}$ denotes a random k-fold partition of $[n]$, i.e., a collection of k mutually exclusive sets whose union is equal to $[n]$. By reusing subsamples of the data for computation of both the statistic of interest and the estimation of nuisance parameters, cross-split statistics mitigate potential losses in statistical precision.

In this section, we demonstrate, in several real applications from applied economics, that the second concern—the residual randomness generated by sample-splitting—can substantively affect results. This residual variability is, often, not resolved by cross-splitting. Consequently, in Section 3,

[1] Of course, under certain conditions involving the sparsity of the covariance between the treatment, outcome, and controls, "double post selection" of control variables with a Lasso regression (Tibshirani, 1996) can produce consistent treatment effect estimates (Belloni et al., 2014). However, consistent estimates can be obtained under weaker conditions through a closely related approach based on sample-splitting (Chernozhukov et al., 2018; Belloni et al., 2012).

we propose an approach to aggregating sample-split statistics that controls the scope of residual randomness at a minimal computational cost. Revisiting the applications, we demonstrate that sample-split statistics aggregated through this procedure are reproducible.

2.1 Cross-Validation

The most prevalent instance of sample-splitting in applied economics is the use of cross-validation for model selection. Often, in this setting, the data D_i consist of a measurement of an outcome Y_i and a vector X_i collecting measurements of p covariates. Interest is in choosing a parsimonious subset of the covariates that, together, best predicts the outcome.

Lasso regression is a standard approach to this problem (Tibshirani, 1996; Hastie et al., 2015). The Lasso coefficient is the solution to the regularized least-squares regression

$$\hat{\eta}_\lambda(D) = \underset{\eta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \eta^\top X_i)^2 + \lambda \sum_{j=1}^{p} |\eta_j| \right\} , \tag{2.5}$$

where λ is some tuning parameter chosen by the user. Penalization of the ℓ_1-norm of the coefficient η encourages sparsity. That is, often, many elements of $\hat{\eta}_\lambda(D)$ are exactly equal to zero. The set of covariates associated with non-zero coefficients are referred to as the model "selected" by the Lasso. The selected model is sensitive to the choice of the tuning parameter λ. If λ is sufficiently large, no covariates are selected. If λ is sufficiently small, all covariates are selected.[2]

Cross-validation is a widely applied and, perhaps, uncontroversial approach for choosing tuning parameters. Here, cross-validation entails assigning λ the value that minimizes the cross-split estimate of the out-of-sample mean-squared error

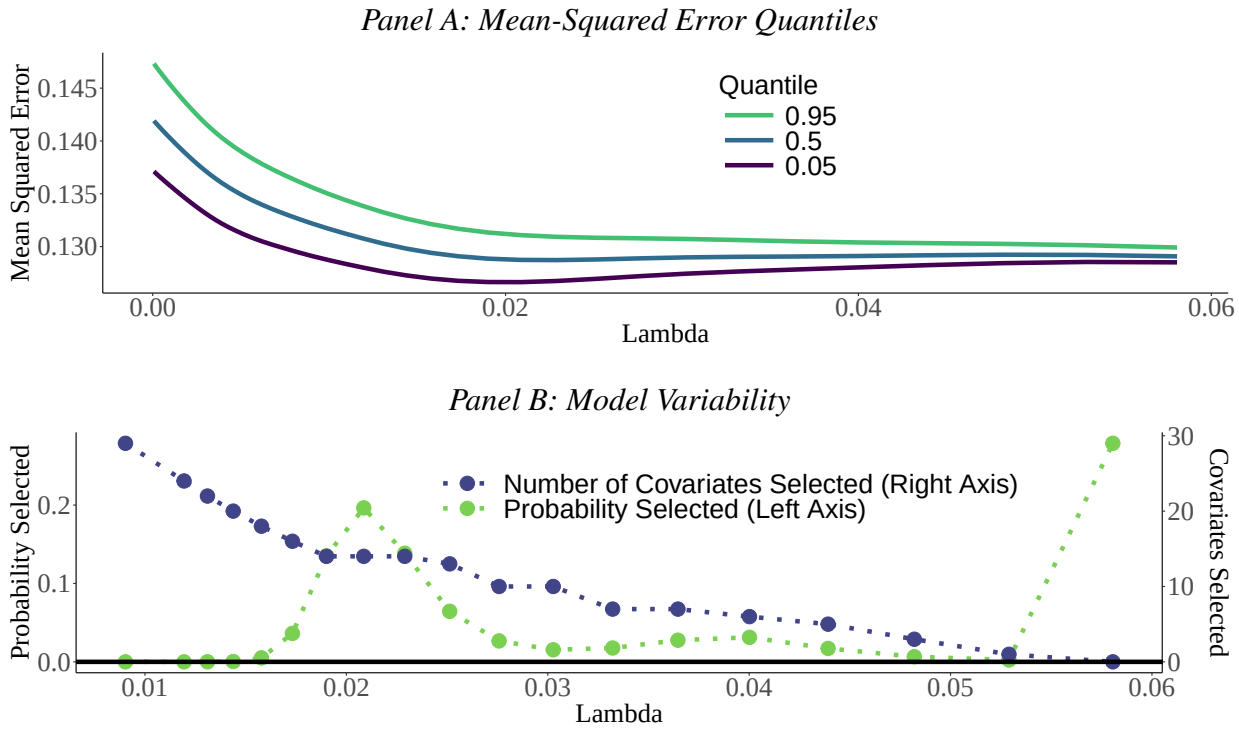
$$a_\lambda(\mathsf{r}, D) = \frac{1}{k} \sum_{j=1}^{k} T_\lambda(\mathsf{s}_j, D) , \quad \text{where} \quad T_\lambda(\mathsf{s}, D) = \frac{1}{|\mathsf{s}|} \sum_{i \in \mathsf{s}} (Y_i - \hat{\eta}_\lambda(D_{\tilde{\mathsf{s}}})^\top X_i)^2 \tag{2.6}$$

and, as before, $\mathsf{r} = (\mathsf{s}_j)_{j=1}^k$ is a random k-fold partition of $[n]$. By measuring error out-of-sample, i.e., in the independent subsample D_{s}, the cross-validated risk estimate (2.6) avoids any "over-fitting" bias induced by the estimation of the coefficient $\hat{\eta}_\lambda(D_{\tilde{\mathsf{s}}})$. This approach is widely applied throughout various subfields of applied economics.[3]

[2]For sufficiently small λ, a covariate is selected so long as it is not collinear with other covariates and its sample correlation with the outcome is not exactly equal to zero.

[3]See, for example, applications to Crime (Mastrobuoni, 2020; Arnold et al., 2020), Development (Casey et al., 2021; Blattman et al., 2024; Sadka et al., 2024), Economic Theory (Fudenberg and Liang, 2019), Education (Ellison and Pathak, 2021), Environment (Deryugina et al., 2019; Cicala, 2022), Finance (Koijen et al., 2024), Health (Abaluck et al., 2016; Cooper et al., 2020), Industrial Organization (Kelly et al., 2023; Dubé and Misra, 2023), Innovation (Chen et al., 2021; Myers and Lanahan, 2022), Labor (Muendler and Becker, 2010; Card et al., 2020; Adermon et al., 2021; Derenoncourt, 2022), Market Design (Agarwal et al., 2019), Macroeconomics (Hansen et al., 2018), and Political Economy (Gentzkow et al., 2019; Cantoni and Pons, 2022).

FIGURE 1. Cross-Validation



Notes: Figure 1 measures the residual randomness of the cross-validated Lasso, implemented in data from Casey et al. (2021). Panel A displays quantiles of the 10-fold cross-validated estimate of the mean-squared error (2.6) of the lasso regression (2.5) over a grid of values of λ. In Panel B, the probabilities that each value of λ minimize the cross-validated risk estimate are displayed with light green dots, relative to the left y-axis. The number of covariates selected at each value of λ are displayed with dark blue dots, relative to the right y-axis. See Appendix A.1 for further details.

The risk estimate (2.6) is random both through the data D and through the choice of the k-fold partition r. That is, the choice of λ, and thereby, the selected model, have the potential to change for different choices of the random collection r. To evaluate the scope of this sensitivity, we consider data from Casey et al. (2021). Casey et al. (2021) study a large-scale experiment, implemented in Sierra Leone, in which a randomly selected subset of parliamentary elections were preceded by direct vote, party-specific primaries. They select covariates to include in various, downstream, econometric analyses with the cross-validated Lasso. In their setting, the outcome Y_i is the vote share, in a poll of party officials, for each of 390 candidates. The vector X_i collects measurements of 48 characteristics for each candidate.

Panel A of Figure 1 displays quantiles, across random draws of the 10-fold partition r, of the cross-validated estimate of the mean-squared error (2.6) over a grid of values of λ. The curve associated with the 5th quantile has a minimum around $\lambda = 0.02$. By contrast, the curve associated with the 95th quantile is monotonically decreasing. This induces instability in the value of λ chosen by cross-validation. Panel B of Figure 1 displays, in light green, the probability—again, across random

10-fold partitions r—that each value of λ is selected, i.e., minimizes the cross-validated estimate of the mean-squared error. The number of covariates in the model associated with each value of λ is displayed in dark blue. The distribution of the selection probabilities has two maxima—both are associated with probabilities greater than 0.2. One entails selecting a model with 14 covariates. The other entails selecting a model with zero covariates. The random choice of the partition r substantively affects the size of the selected model.[4]

2.2 Cross-Fitting

Meaningful residual randomness is not particular to cross-validated risk estimation. A second class of sample-split methods, seeing increasing use in applied economics, are characterized by the application of "cross-fitting" to accommodate—or characterize—treatment effect heterogeneity. Often, in this case, the data D_i consist of an outcome Y_i, a binary treatment W_i, and a vector of covariates X_i. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes induced by the treatment W_i. As before, let s denote a random a subset of $[n]$, with complement $\tilde{\mathsf{s}}$.

The basic idea underlying this class of methods is to use the data from the units i in $\tilde{\mathsf{s}}$ to predict the treatment effects $Y_i(1) - Y_i(0)$ for the units i in s. For example, a machine learning algorithm can be applied to the data $D_{\tilde{\mathsf{s}}}$ to construct an estimate of the conditional expectation

$$\mu_w(x) = \mathbb{E}[Y_i \mid X_i = x, W_i = w] \tag{2.7}$$

for each w in $\{0, 1\}$. Collect these estimates into $\hat{\eta}(D_{\tilde{\mathsf{s}}}) = (\hat{\mu}_1, \hat{\mu}_0)$. Predictions of the treatment effects for the units i in s can be constructed through the sample-split statistic

$$\psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \ . \tag{2.8}$$

These predictions can then be used to estimate average treatment effects and characterize treatment effect heterogeneity, among other related problems. The point is, by splitting the data, these "second-stage" analyses are not confounded by the construction of the "first-stage" estimators.

"Double Machine Learning" (DML) estimates of average treatment effects are a leading example of a method that takes this structure (Chernozhukov et al., 2018). Here, the sample-split statistics (2.8) are aggregated with cross-splitting, through

$$a(\mathsf{r}, D) = \frac{1}{k} \sum_{j=1}^{k} T(\mathsf{s}_j, D) \ , \quad \text{where} \quad T(\mathsf{s}, D) = \frac{1}{|\mathsf{s}|} \sum_{i \in \mathsf{s}} \psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \tag{2.9}$$

[4] Casey et al. (2021) take measures to stabilize their results. In particular, they select covariates that are selected in more than 200 of 400 randomly drawn 10-fold partitions r. We use this setting as an example, in part, because a serious attempt was made to address residual randomness. This is not commonplace in the literature surveyed in Footnote 3 (a similar strategy is used in, e.g., Hansen et al. (2018), however).

and, again, r is a random k-fold partition of $[n]$. An appropriate standard error for (2.9) is itself given by the cross-split statistic

$$\mathsf{se}(\mathsf{r}, D) = \frac{1}{n}\sqrt{\sum_{j=1}^{k}\sum_{i\in\mathsf{s}_j}\left(\psi(D_i, \hat{\eta}(D_{\bar{\mathsf{s}}_j})) - a(\mathsf{r}, D)\right)^2}\,. \tag{2.10}$$

Chernozhukov et al. (2018) show that an asymptotically efficient test of the null hypothesis that an average effect treatment effect is less than zero can be constructed by comparing (2.9) to the critical value $\mathsf{cv}_\alpha(\mathsf{r}, D) = z_{1-\alpha}\cdot\mathsf{se}(\mathsf{r}, D)$, where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution.[5] DML estimators have been increasingly used in applied economics, as they place weaker restrictions on treatment effect heterogeneity than, say, methods based on linear regression.[6]

The DML estimator (2.9) and standard error (2.10) are, again, random through both the data D and the k-fold partition r. To evaluate the extent of resultant variability, we consider data from Chakravorty et al. (2024). Chakravorty et al. (2024) use DML to study the effect of a program, implemented in two Indian states, involving the provision of information concerning prospective jobs to vocational trainees, on employment outcomes. Here, the binary outcome Y_i indicates employment five months after completing training for each of 890 trainees placed into jobs, W_i denotes assignment to the program, and X_i collects measurements of 77 pre-treatment covariates.

Panel A of Figure 2 displays a heat map of the joint distribution of the 5-fold cross-fit estimator (2.9) and the associated critical value $\mathsf{cv}_\alpha(\mathsf{r}, D)$ over random draws of the partition r.[7] A black line is placed at the threshold where the estimate is equal to the critical value. The residual variability in the estimate is large relative to estimates of the sampling variability, and is sufficient to determine the purported statistical significance. Concretely, the difference between the 5th and 95th quantiles of the distribution of the estimator (0.094 and 0.131, respectively) is equal to 68% of the median of the distribution of the standard error (0.054).[8]

The same behavior is exhibited in related applications that use cross-fitting in more complicated ways. In these cases, measures to stabilize results are more common, although there is little guidance on how to best operationalize this stabilization. To illustrate this, we consider examples

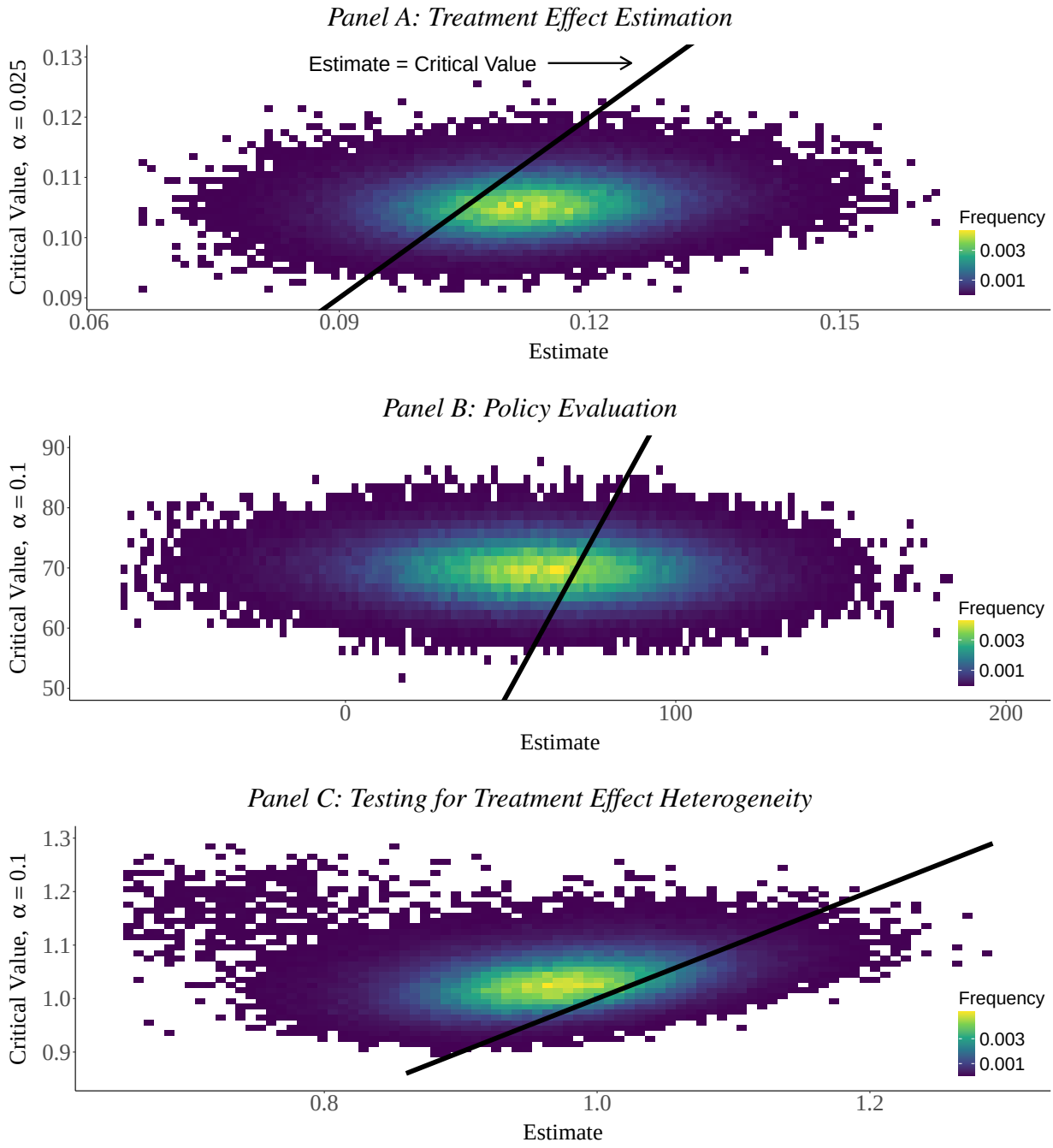
[5]This result does not apply to estimates of the form (2.8), but rather to estimators constructed with a "Neyman Orthogonal" moment, which, in this case, additionally require a non-parametric estimate of the propensity score.

[6]See, for example, applications in Okunogbe and Pouliquen (2022), Beraja et al. (2023), Covert and Sweeney (2023), Delfino (2024), Farronato et al. (2024).

[7]We follow the replication package associated with Chakravorty et al. (2024). Nuisance parameters are estimated with random forests using the "Ranger" R package (Wright and Ziegler, 2017). Estimates and standard errors are constructed using the "DoubleML" R package (Bach et al., 2024). See Appendix A.2 for further details.

[8]Some papers take measures to address analogous residual randomness (see e.g., Covert and Sweeney (2023)). In fact, Chernozhukov et al. (2018) suggest taking the median over several partitions r. Most implementations of DML in standard statistical software allow the user to aggregate estimates, although this aggregation does not occur by default (e.g., Bach et al., 2024 in R or Ahrens et al., 2024 in STATA).

FIGURE 2. Cross-Fitting



Notes: Figure 2 displays discretized heat maps quantifying the residual randomness of three estimators constructed with cross-fitting. All three panels give the joint distribution of an estimator and an associated critical value. In each case, a black line has been placed at the threshold where the estimator is equal to the critical value. Panel A displays the distribution of the DML estimate (2.9) cross 5-fold cross-splits using data from Chakravorty et al. (2024). Panel B displays the difference between treatment effects estimates for the "most impacted" and "most deprived" individuals, following Haushofer et al. (2022), using data from Egger et al. (2022), again across 5-fold cross-splits. Panel C displays the distribution of the estimate associated with a test of treatment effect heterogeneity using data from Beaman et al. (2023). Here, each estimate and critical value is computed by averaging over 250 independently drawn half-samples $\mathtt{s}$. See Appendix A for further details on the construction of each panel.

from Haushofer et al. (2022) and Beaman et al. (2023).[9] Haushofer et al. (2022) use data from a randomized cash transfer implemented in Kenya. These data were originally considered in Egger et al. (2022). Beaman et al. (2023) use data from an experiment concerning agricultural lending in Mali. Both papers use sample-splitting to construct estimates, of the form (2.8), of the treatment effects $Y_i(1) - Y_i(0)$ for each of the units i in s. Haushofer et al. (2022) additionally construct sample-split estimates of the untreated outcome $Y_i(0)$ for each of the units in s.

Haushofer et al. (2022) identify the 50% of units in s that have the largest predicted treatment effect as well as the 50% of units that have the smallest predicted untreated outcome. They refer to these groups as the "most impacted" and "most deprived," respectively. They estimate the difference in the average treatment effect for the two groups, and construct a standard error, and associated critical value, for this difference with the bootstrap. See Appendix A.3 for details. Panel B of Figure 2 displays a heat map of the joint distribution of the 5-fold cross-fit estimate of the difference between the treatment effect estimates for the two groups, and the associated critical value, over random draws of the partition r.[10] The residual randomness is considerable. Here, the difference between the 5th and 95th quantiles of the distribution of the estimator (37.82 and 109.69, respectively) is equal to 190% of the median of the distribution of the standard error (54.44). To stabilize these results, Haushofer et al. (2022) average over 400 draws of the 5-fold partition r. Due to this increased computation, they only report confidence intervals for their main results. The methods developed in this paper ensure that stabilization of sample-split statistics, in this way, occurs at a minimal computational expense.

Likewise, Beaman et al. (2023) implement a test of treatment effect heterogeneity proposed by Chernozhukov et al. (2023). In particular, in data for units in s, the outcome is regressed on the treatment, the treatment effect estimate, and an interaction between the treatment and the treatment effect estimate. The idea is that, if treatment effect estimates are well-calibrated, then the coefficient on the interaction should be statistically greater than zero. Chernozhukov et al. (2023) recommend aggregating p-values associated with the coefficients on the interaction over 250 replications of this sample-split test. Panel C of Figure 2 displays a heat map measuring the joint distribution of

[9] Additional, related, methods that uses cross-fitting to estimate and evaluate treatment targeting rules are proposed by Athey and Wager (2021) and Yadlowsky et al. (2024).

[10] We were not able to access a replication package associated with Haushofer et al. (2022), and so implement a simplified version of the exercise considered in that paper using data from the replication package associated with Egger et al. (2022). Treatment effects, and untreated outcomes, are estimated with random forests using the "GRF" R package (Athey et al., 2019). Analogous estimates reported in Haushofer et al. (2022) are statistically significant. We emphasize that, as we implement only a simplified version of the exercise conducted in Haushofer et al. (2022), our estimates should only be interpreted an an illustration of the scope of residual randomness in analyses of this type, rather than as a substantive characterization of underlying treatment effect heterogeneity. In particular, we make no attempt to reweigh observations according to their sampling probabilities and appear to be using a different measure of the time between the administration of the experiment and the measurement of post-treatment outcomes.

the average coefficient and critical value associated with this test.[11] That is, each coefficient and critical value is computed by taking the average over 250 sample-splits. Despite this aggregation, the residual randomness remains meaningful.[12] The difference between the 5th and 95th quantiles of the distribution of the estimator (0.86 and 1.08, respectively) is equal to 28% of the median of the distribution of the standard error (0.80). Perhaps motivated by this instability, Beaman et al. (2023) aggregate over 1000 sample-splits.

Two themes emerge from these examples. First, sample-splitting is a versatile tool for simplifying various tasks associated with modern statistical inference. Second, the residual randomness induced by sample splitting is often large and can affect the substantive interpretation of results. In the next section, we provide a general-purpose method for aggregating sample-split statistics that ensures that residual randomness is controlled at a minimal computational cost.

3. Reproducible Aggregation

We propose a sequential method for aggregating sample-split statistics. Our objective is to ensure that the auxiliary randomness induced by sample-splitting is small. To introduce the method, we require some additional notation. The set $\mathcal{S}_{n,b}$ consists of all subsets of $[n] = \{1, \ldots, n\}$ of size b. In turn, the set $\mathcal{R}_{n,k,b}$ contains all collections of k mutually exclusive elements of $\mathcal{S}_{n,b}$. That is, if $n = k \cdot b$, then the set $\mathcal{R}_{n,k,b}$ collects all partitions of $[n]$ into k mutually exclusive sets of size b. We refer to the elements of the set $\mathcal{R}_{n,k,b}$ as cross-splits. Throughout, the quantity $\mathsf{R}_{g,k} = (\mathsf{r}_i)_{i=1}^{g}$ collects g cross-splits, each given by $\mathsf{r}_i = (\mathsf{s}_{i,j})_{j=1}^{k}$. Unless otherwise specified, the elements of $\mathsf{R}_{g,k}$ are random, sampled independently and uniformly from the collection $\mathcal{R}_{n,k,b}$.

Suppose that we are interested in some sample-split statistic $T(\mathsf{s}, D)$. We study the construction of aggregate statistics of the form

$$a(\mathsf{R}_{g,k}, D) = \frac{1}{g}\sum_{i=1}^{g} a(\mathsf{r}_i, D)\ , \quad \text{where} \quad a(\mathsf{r}_i, D) = \mathcal{A}(\{T(\mathsf{s}_{i,j}, D)\}_{j=1}^{k}) \tag{3.1}$$

and the function $\mathcal{A}(\cdot)$ aggregates the statistics $T(\mathsf{s}_{i,j}, D)$ across the cross-split r_i. To simplify exposition, for the time being, we will restrict attention to the case that the statistic $a(\mathsf{R}_{g,k}, D)$ is real-valued. This is natural, if, for example, the statistic $a(\mathsf{r}, D)$ is a treatment effect estimate or p-value constructed with cross-fitting. Later, in our application to cross-validated risk estimation,

[11] We follow the details of the replication package associated with Beaman et al. (2023). Nuisance parameters are estimated with random forests using the "GRF" R package (Athey et al., 2019). Further details are given in Appendix A.4.

[12] Chernozhukov et al. (2023) recommend aggregating p-values the median to ensure robustness to outliers. In practice, if aggregation of a sample-split statistic with a median, rather than a mean, makes a material difference, we strongly encourage researcher to investigate why some sample-splits generate extreme (or highly skewed) estimates (e.g., outliers in the underlying data or poor overlap in an intervention).

we treat vector-valued sample-split statistics, e.g., cross-validated risk estimates queried at a vector of values of a penalization parameter.

Our task is to formulate a method for choosing the number of cross-splits g to ensure that the residual variability of the aggregate statistic (3.1) is small. We formalize this objective as follows.

Definition 3.1 (Reproducible Aggregation)**.** Let the sequences $\{\mathsf{r}_i\}_{i=1}^{\infty}$ and $\{\mathsf{r}'_i\}_{i=1}^{\infty}$ be drawn independently and uniformly, conditional on the data D, from the collection of cross-splits $\mathcal{R}_{n,k,b}$. Define $\mathsf{R}_{g,k} = \{\mathsf{r}_g\}_{i=1}^{g}$ and $\mathsf{R}'_{g,k} = \{\mathsf{r}'_g\}_{i=1}^{g}$ for each integer g. Suppose that the integers $\hat{g}$ and $\hat{g}'$ are independent and identically distributed, again conditional on the data D. We say that the aggregate statistic $a(\mathsf{R}_{\hat{g},k}, D)$ is (ξ, β)-reproducible if

$$P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \geq \xi \mid D\right\} \leq \beta \tag{3.2}$$

almost surely.

Definition 3.1 is motivated by the following thought experiment. Suppose that the data D are given to two researchers. Each researcher is tasked with producing an estimate of the form (3.1). They generate the collections of splits, $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$, independently using the same, potentially data-dependent, procedure. That is, the two collections of splits are independent and identically distributed, conditional on the data D. If the estimate $a(\mathsf{R}_{\hat{g},k}, D)$ is (ξ, β)-reproducible, then the probability that the two researchers' estimates differ by more than ξ is less than β.

Of course, ensuring that estimates are reproducible, in the sense of Definition 3.1, does not preclude deceptive behavior. Researchers might compute a reproducibility aggregated statistic many times, until a desirable estimate is obtained. However, reporting reproducible statistics greatly increases the cost of searches of this form, as the scope of residual randomness has been reduced.

We propose a sequential method for constructing reproducible sample-split statistics. The proposal is based on the fixed-length sequential confidence intervals of Anscombe (1952) and Chow and Robbins (1965). The procedure works by repeatedly drawing a cross-split r_g uniformly from $\mathcal{R}_{n,k,b}$, appending the cross-split to the collection $\mathsf{R}_{g,k} = (\mathsf{R}_{g-1,k}, \mathsf{r}_g)$, and estimating the conditional variance

$$v_{g,k}(D) = \mathrm{Var}(a(\mathsf{R}_{g,k}, D) \mid D) \tag{3.3}$$

with the plug-in estimator

$$\hat{v}(\mathsf{R}_{g,k}, D) = \frac{1}{g}\frac{1}{g-1}\sum_{i=1}^{g}(a(\mathsf{r}_i, D) - a(\mathsf{R}_{g,k}, D))^2 \tag{3.4}$$

until the condition

$$\hat{v}(\mathsf{R}_{g,k}, D) \leq \mathsf{cv}(\xi, \beta) = \frac{1}{2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \tag{3.5}$$

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 $\mathsf{R}_{g,k} = (\mathsf{R}_{g-1,k}, \mathsf{r}_g)$.

end

Set $\hat{g} \leftarrow g$

return $a(\mathsf{R}_{\hat{g},k}, D)$

Notes: Algorithm 1 gives a method for sequentially aggregating sample-split statistics. The critical value $\mathsf{cv}(\xi, \beta)$ is defined in display (3.5).

is satisfied, where z_α denotes the α quantile of the standard normal distribution. In particular, let $g_{\mathsf{init}} \geq 2$ denote some "burn-in" period chosen by the user. The number of cross-splits $\hat{g}$ chosen by the procedure is the smallest value of g, greater than g_{init}, such that (3.5) is satisfied. The procedure is summarized in Algorithm 1.

In Section 3.1, we show that Algorithm 1 is applicable, off-the-shelf, to a large class of problems. In particular, without imposing any conditions on the data generating process or the statistic of interest, we show that sample-split statistics aggregated with Algorithm 1 are reproducible, in a particular asymptotic sense. Moreover, we show that, in many applications, statistics aggregated with Algorithm 1 maintain, or improve on, various unconditional statistical guarantees. We then revisit, in Section 3.2, the examples considered in Section 2. We show that, in each case, an application of Algorithm 1 produces a reproducible, stabilized, estimate.

3.1 Asymptotic Validity

The following theorem establishes that statistics aggregated with Algorithm 1 are reproducible, in a particular asymptotic sense. The proof is closely related to the arguments of Anscombe (1952) and Chow and Robbins (1965) and is given in Appendix C. See e.g., Theorem 3.1 of Gut (2009) for a textbook treatment.

Theorem 3.1. *Suppose that the conditional variance* $\mathrm{Var}(a(\mathsf{r}, D) \mid D)$ *is strictly positive, almost surely, where* r *denotes a random collection drawn uniformly from* $\mathcal{R}_{n,k,b}$*. If the collections* $\mathsf{R}_{\hat{g},k}$ *and* $\mathsf{R}'_{\hat{g}',k}$ *are independently obtained using Algorithm 1, then*

$$P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \geq \xi \mid D\right\} \xrightarrow{a.s} \beta \quad \textit{as} \quad \xi \to 0\,. \tag{3.6}$$

Two aspects of Theorem 3.1 are worthy of emphasis. First, no restrictions on the data generating process or the statistic of interest are imposed. For instance, the data D do not necessarily need

to be i.i.d. and the statistic $a(\mathsf{r}, D)$ can be random though quantities other than just r and D. The result, thus, provides a strong assurance that Algorithm 1 can be applied widely.

Second, in the statement (3.6), asymptotics are taken as $\xi \to 0$, with all other quantities, including the sample size n, fixed. This asymptotic framework is somewhat opaque, or at least nonstandard, as the parameter ξ is a choice variable. The operational interpretation is that, if ξ is chosen to be sufficiently small—at a point negligible relative to, say, the conditional variance $v_{1,k}(D)$—then the nominal reproducibility error β is accurate. In Section 4, by imposing some simplifying restrictions, we give a set of non-asymptotic results that make the dependence of the performance of Algorithm 1 on the choices of ξ, β, and the statistic $a(\mathsf{r}, D)$ more transparent.

The intuition underlying Theorem 3.1 is straightforward. The result follows directly from the observation that the cross-splits r_i are independent and identically distributed conditional on the data D. Thus, for large values of g, the variance estimator $\hat{v}\left(\mathsf{R}_{g,k}, D\right)$ will be close to the conditional variance $v_{g,k}(D)$. As a consequence, if ξ is sufficiently small, then the number of cross-splits $\hat{g}$ chosen by the procedure will be close to the "oracle" stopping time

$$g^\star = \underset{g \geq g_{\mathsf{init}}}{\arg\min} \left\{ \mathsf{cv}(\xi, \beta) = \frac{1}{2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \geq v_{g,k}(D) \right\} \tag{3.7}$$

$$= \underset{g \geq g_{\mathsf{init}}}{\arg\min} \left\{ \xi \geq z_{1-\beta/2}\sqrt{\frac{2v_{1,k}(D)}{g}} \right\}, \tag{3.8}$$

where we have used the fact that $v_{g,k}(D) = g^{-1}v_{1,k}(D)$ in writing (3.8). This, then, ensures that the aggregate statistic is approximately (ξ, β)-reproducible, as

$$P\left\{\left|a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)\right| \geq \xi \mid D\right\}$$
$$= P\left\{\left|\sqrt{\frac{g^\star}{2v_{1,k}(D)}}\left(\frac{1}{g^\star}\sum_{i=1}^{g^\star} a(\mathsf{r}_i, D) - a(\mathsf{r}'_i, D)\right)\right| \geq z_{1-\beta/2} \mid D\right\} \xrightarrow{\text{a.s}} \beta$$

as $\xi \to 0$, where the limit follows from the central limit theorem (as $g^\star \to \infty$ as $\xi \to 0$).[13]

In most cases, sample-split statistics are used because they have some desirable property. For example, a sample-split estimator $a(\mathsf{r}, D)$ might be asymptotically normal or might perform well in terms of some loss function. The following result—roughly, a sequential version of Jensen's inequality—can be applied to show that, in many situations, statistics aggregated with Algorithm 1 inherit these properties. The proof is given in Appendix C and follows from an application of the martingale stopping theorem.

[13]The discrepancy between $a(\mathsf{R}_{\hat{g},k}, D)$ and $a(\mathsf{R}_{g^\star,k}, D)$ can be shown to be negligible by applying an appropriate maximal inequality—an idea due to Rényi (1957).

Theorem 3.2. *If the collection* $\mathsf{R}_{\hat{g},k}$ *is obtained with Algorithm 1, the cross-split* r *is a random element of the set* $\mathcal{R}_{n,k,b}$*, and* $f(\cdot)$ *is any convex function, then*

$$\mathbb{E}\left[f(a(\mathsf{R}_{\hat{g},k}, D))\right] \leq \mathbb{E}\left[f(a(\mathsf{r}, D))\right] \ . \tag{3.9}$$

Suppose that the estimator $a(\mathsf{r}, D)$ has good performance in terms of some convex loss. Theorem 3.2 demonstrates that the aggregate estimator $a(\mathsf{R}_{\hat{g},k}, D)$ will also perform well (and, in fact, may perform better). For example, Chetverikov et al. (2021) demonstrate that the cross-validated lasso has nearly optimal rates of convergence in sparse regression problems. Theorem 3.2 shows that all higher-order moments of risk estimates obtained with our procedure are smaller than those of risk estimates obtained from a single cross-split. Thus, we expect the same result to hold for tuning parameters chosen with risk estimates aggregated with Algorithm 1.

Alternatively, in many cases, sample-split statistics are shown to be asymptotically normal by demonstrating that the variance of the discrepancy

$$\sqrt{n}\left(a(\mathsf{r}, D) - \frac{1}{n}\sum_{i=1}^{b}\psi(D_i, \eta)\right) \tag{3.10}$$

is small, where $\psi(D_i, \eta)$ is some score, or influence, function and η is some unknown nuisance parameter.[14] This is how Chernozhukov et al. (2018) demonstrate that DML estimates of average treatment effects are asymptotically normal, for example. Theorem 3.2 demonstrates that the same argument will apply to statistics aggregated with Algorithm 1.

In some cases, substantial residual randomness in a sample-split statistic might cause concern for the validity of associated inferences. For example, for DML estimators, the variance of the term (3.10) is bounded from below by the expectation of its variance conditional on the data. So, if the conditional variance is large, then the unconditional variance of (3.10) is likely not small. On the other hand, the conditional variance can be reduced through the application of our procedure. In other cases, substantial residual randomness may not be a reason for concern, because inferences are based on estimating a nuisance parameter in one split of a data set, and then conducting inference in the other split by conditioning on this estimate. This is the case for the applications considered in Beaman et al. (2023) and Haushofer et al. (2022).

There is a large statistical literature on the use and interpretation of average p-values. In general, a level α test can be constructed by comparing an average p-value to $2 \cdot \alpha$ (see e.g. Rüger, 1978; Vovk and Wang, 2020; DiCiccio et al., 2020).[15] Using a result analogous to Theorem 3.2, for a

[14] That is, if second term in (3.10) is asymptotically normal and the variance of (3.10) converges to zero, then the statistic $a(\mathsf{r}, D)$ is also asymptotically normal by Chebychev's inequality and the continuous mapping theorem.

[15] A variety of methodological papers use this observation to construct sample-split hypothesis tests that control the Type I error rate under very general conditions. In Appendix B.1, we show that the sample-split hypothesis tests considered in e.g., DiCiccio et al. (2020), Meinshausen et al. (2009), and Wasserman et al. (2020) continue to be valid when they are constructed sequentially with Algorithm 1.

pre-determined number of sample-splits g, Chernozhukov et al. (2023) show that, under some high-level conditions, valid p-values aggregated by averaging over sample-splits are also valid p-values. Under the same conditions, Theorem 3.2 can be applied to give an analogous result for p-values aggregated with Algorithm 1.

3.2 Reproducible Aggregation in Practice

Equipped with Algorithm 1, we return to the examples considered in Section 2. We show that, in each case, an appropriate application of the procedure produces a stabilized, reproducible estimate.

3.2.1 Cross-Fitting. We begin by treating the three examples that use cross-fitting. The most important choice to make when implementing Algorithm 1 is the specification of the statistic $a(\mathsf{r}, D)$. There are many reasonable choices that one might make here. For example, one could choose to ensure that estimates or standard errors are stable up to a desired level of precision. For the sake of comparability across settings, we opt to stabilize the p-value

$$a(\mathsf{r}_i, D) = 1 - \Phi\left(\frac{\mathsf{est}(\mathsf{r}_i, D)}{\mathsf{se}(\mathsf{r}_i, D)}\right) , \tag{3.11}$$

where the function $\Phi(\cdot)$ is the standard normal c.d.f. and the quantities $\mathsf{est}(\mathsf{r}_i, D)$ and $\mathsf{se}(\mathsf{r}_i, D)$ denote the cross-split estimate and standard error that corresponding to the axes of Figure 2. In effect, controlling the residual randomness of the average p-value ensures that the residual randomness in the aggregate estimate is small relative to the sampling error.

Table 1 summarizes the results. We choose ξ equal to either 0.001 and 0.01, as these are the levels of precision relevant for the determination of statistical significance at levels $\alpha = 0.025$ and $\alpha = 0.10$, respectively. For now, we follow the approaches to cross-splitting taken by all three papers. In the applications to Chakravorty et al. (2024) and Haushofer et al. (2022), we aggregate over 5-fold cross splits. In the application to Beaman et al. (2023), the split r_i contains a single half-sample, i.e., subset of $[n]$ of size $n/2$.[16] The estimates displayed in Table 1 indicate that all three applications require aggregation over hundreds of cross-splits to ensure reproducibility, in the sense of Definition 3.1, at these levels of error tolerance. Moreover, we find that nominal error rate β associated with Algorithm 1, here set to 0.05, is very accurate.

3.2.2 Cross-Validation. We now turn to our application to cross-validated risk estimation. In this setting, there is more ambiguity in how to best apply Algorithm 1. We find that the following approach works well in the data from Casey et al. (2021). Recall that, in this application, the sample-split statistic $a_\lambda(\mathsf{r}, D)$ denotes the cross-validated estimate of the mean-squared error, queried at a specified value of the regularization parameter λ. We are interested in stabilizing these estimates

[16] If it is desirable to aggregate the median p-value, Algorithm 1 will continue to apply with a small modification. In particular, the variance estimator (3.4) can be replaced by an estimator constructed with the bootstrap. In this case, a result analogous to Theorem 3.1 will continue to hold.

TABLE 1. Reproducible Aggregation in Practice: Cross-Fitting

Application	ξ	p-value	Average $\hat{g}$	Reproducibility
Chakravorty et al. (2024)	0.001	0.979	860.7	0.945
Haushofer et al. (2022)	0.01	0.80	1360.6	0.947
Beaman et al. (2023)	0.01	0.83	2794.8	0.950

Notes: Table 1 summarizes the application of Algorithm 1 to three of the examples considered in Section 2. The first column indicates the application. The second column specifies choices for the error tolerance ξ. In each case, we set the reproducibility error rate to $\beta = 0.05$ and the burn-in sample size g_{init} to 10. The third column gives the p-value produced by one-application of the procedure. The fourth column gives the average number of cross-splits $\hat{g}$ drawn in each implementation, taken across 2,000 replications of the aggregation procedure. The fifth column displays an estimate of the true reproducibility probability, i.e., the probability that two independent implementations of Algorithm 1 produce estimates that differ by less than ξ, computed with these replicates.

for each value of λ in an increasing sequence $(\lambda_\ell)_{\ell=1}^p$, where λ_p is the smallest value such that no covariates have non-zero coefficients when estimated using the full data.[17]

To do this, we apply Algorithm 1 to each component of the vector

$$(b_{\lambda_1}(\mathsf{r}, D), \ldots, b_{\lambda_{p-1}}(\mathsf{r}, D)) \,, \quad \text{where} \quad b_{\lambda_i}(\mathsf{r}, D) = (a_{\lambda_i}(\mathsf{r}, D) - a_{\lambda_p}(\mathsf{r}, D)) \,. \tag{3.12}$$

The rationale for this is that the relative, rather than absolute, values of the risk estimates are what are relevant for determining where the estimates takes their minimum value. We find that, in practice, consideration of the differences (3.12) can substantially reduce the amount of aggregation needed for stabilization.

We implement Algorithm 1, independently, for each component of the vector (3.12). In other words, we ensure that each component of the risk estimate is marginally reproducible.[18] We let ξ_i denote the error tolerance used for the ith component of the vector (3.12). We use a simple approach for determining suitable values for these tolerances. Roughly speaking, we compute the statistic (3.12) for each element of a small, initial sample of cross-splits (e.g., $g = 20$). Using these estimates, we find that the level of precision needed to distinguish between the close-to-optimal

[17] Throughout, we use the grid $(\lambda_\ell)_{\ell=1}^p$ of values of the regularization parameter λ chosen by default by the "glmnet" R package (Friedman et al., 2021).

[18] In principle, it is straightforward to ensure that the components of the vector (3.12) are simultaneously reproducible. For example, simultaneous reproducibility at level β can be obtained by ensuring that each component is reproducible at level $\beta/(p-1)$, mimicking the standard Bonferonni adjustment. More sophisticated schemes, based on the bootstrap, say, are also applicable. In practice, ensuring simultaneous reproducibility can substantially increase the required computation and tends not to lead to materially different risk estimates.

values of λ is approximately $\xi_i = 10^{-4}$. The error tolerances specified at values of λ that can be determined to be sub-optimal are set to larger values. Further details are given in Appendix B.2.

Figure 3 gives measurements of the performance of Algorithm 1 for aggregation of the statistic (3.12), implemented in the data from Casey et al. (2021). Panel A displays quantiles of the reproducibly aggregated statistic (3.12) across replications of the procedure. The y-axis has been truncated to focus attention on the close-to-optimal values of the regularization parameter. There is little residual randomness. Panel B displays quantiles of the number of 10-fold cross-splits $\hat{g}$ chosen by the procedure at each value of the regularization parameter. Precise estimates at the close-to-optimal values require aggregation over thousands of cross-splits. Observe that less aggregation is used at small, sub-optimal values of λ, as we have used larger error tolerances for these values. We find that, if cross-validated model selection is implemented twice, and, in each case, aggregated at this level of error tolerance, then the same value of λ is selected with probability 0.84 and the same 14 covariates are selected with probability 0.999. In Appendix B.2, we show that the nominal reproducability error, again $\beta = 0.05$, is very accurate over the full range of the regularization parameter.

4. Theoretical Analysis

The asymptotic results given in Section 3 are quite general. In particular, Theorem 3.1 holds in the absence of any restrictions on the data generating process or statistic under consideration. It is worth asking, however, whether these results confer a clear statistical understanding. At least two issues arise. First, Theorem 3.1 relies entirely on the fact that, in Algorithm 1, successive cross-splits are sampled independently. That is, we have said nothing, yet, about the role of cross-splitting. Second, the interpretation of the asymptotic approximation with $\xi \to 0$ is somewhat opaque. What would be a reasonable value of ξ to choose to ensure that Algorithm 1 is accurate? And how do these choices impact the amount of computation required to implement the procedure?

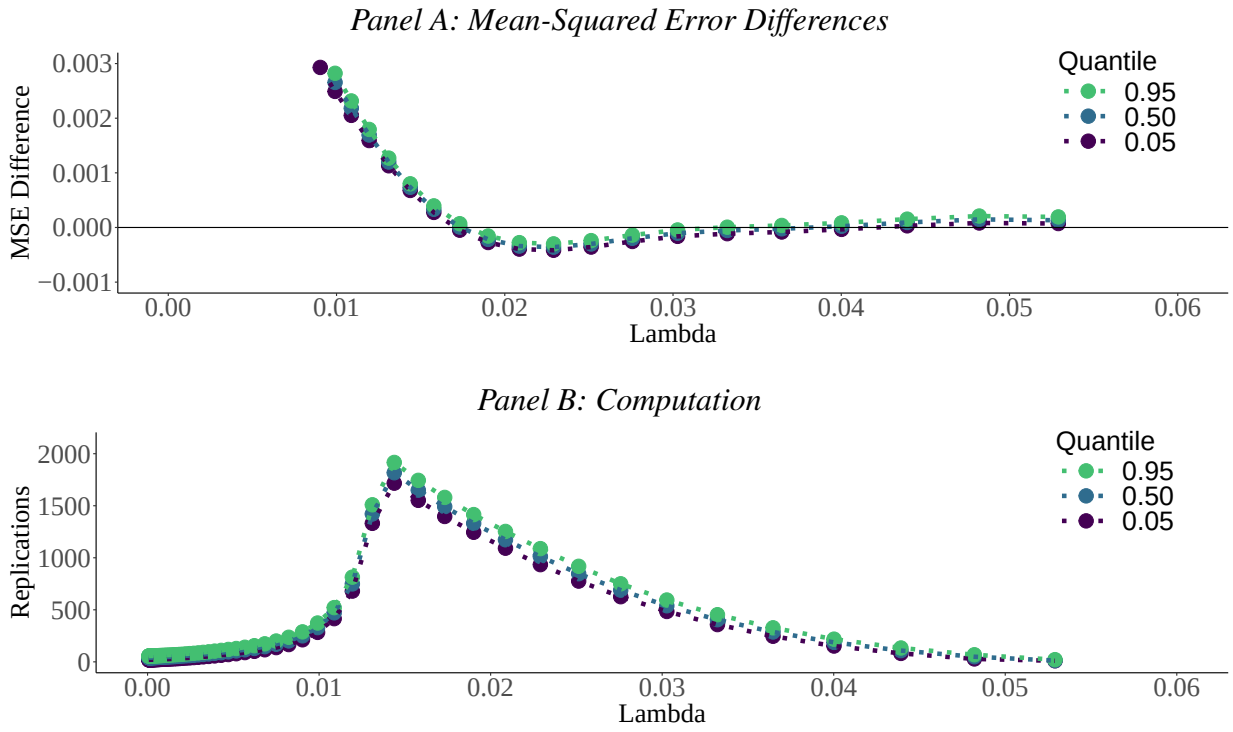
In this section, we give a non-asymptotic description of the performance of Algorithm 1. This is accomplished by placing some simplifying restrictions on the statistic of interest. The more specialized analysis that follows is aimed at providing qualitative and quantitative intuition for how the computational cost and reproducibility error of statistics aggregated with Algorithm 1 depend on the choices of k and ξ. Proofs for results stated in this section are given in Appendix D.

4.1 Symmetry, Linearity, and Stability

We impose a set of simplifying restrictions. Recall that, in general, we are considering the aggregation of cross-split statistics of the form

$$a(\mathsf{r}, D) = \mathcal{A}(\{T(\mathsf{s}_j, D)\}_{j=1}^{k}) \,, \quad \text{where} \quad T\left(\mathsf{s}, D\right) = \Psi\left(D_{\mathsf{s}}, \hat{\eta}\left(D_{\bar{\mathsf{s}}}\right)\right) \tag{4.1}$$

FIGURE 3. Reproducible Aggregation in Practice: Cross-Validation



Notes: Figure 3 displays the performance of Algorithm 1 in the application to cross-validated Lasso, implemented in data from Casey et al. (2021). Algorithm 1 is applied independently to each component of the vector (3.12). The error tolerances ξ_i are specified in Appendix B.2. We set the nominal reproducibility error to $\beta = 0.05$. Panel A displays quantiles of the statistic (3.12) at each value of a grid of values of the regularization parameter λ. The y-axes is truncated to focus attention on large values of λ. We give an un-truncated version in Appendix B.2. Panel B displays quantiles of the number of 10-fold cross-splits $\hat{g}$ chosen by the procedure at each value of the regularization parameter.

---

is a sample-split statistic and the function $\hat{\eta}(\cdot)$ is an estimator of an unknown nuisance parameter η. In the main text, we restrict attention to the case that $n = k \cdot b$, i.e., where each element r in $\mathcal{R}_{n,k,b}$ is a complete partition of $[n]$ into k sets of size b. Each of the results given here will follow directly from more general results stated in Appendix D, where this restriction is not imposed.

First, we assume that the sample-split statistic under consideration is symmetric and deterministic in each part of a split sample.

Assumption 4.1 (Symmetry and Determinism)**.** *For all sets* s *in* $\mathcal{S}_{n,b}$ *and data* D*, the statistic* $T(\mathsf{s}, D)$ *is deterministic and invariant to permutations of the data with indices in* s *and* $\tilde{\mathsf{s}}$*, respectively.*

The intention of Assumption 4.1 is to restrict the residual randomness under consideration to the randomness introduced by sample-splitting. This holds in cases where $\hat{\eta}(\cdot)$ is deterministic, e.g., when $\hat{\eta}(\cdot)$ 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). Assumption 4.1 rules out procedures where the estimator $\hat{\eta}(\cdot)$ is random conditional on the data. This excludes

settings where, e.g., $\hat{\eta}(\cdot)$ is estimated with stochastic gradient descent, bagging or subsampling, or is itself constructed with data splitting.

Second, we assume that the aggregation function $\mathcal{A}(\cdot)$ is an average and that the statistic $T(\mathsf{s}, D)$ is linearly separable in the first part of the split sample.

Assumption 4.2 (Linearity)**.** *For all cross-splits* $\mathsf{r} = (\mathsf{s}_j)_{j=1}^k$ *in* $\mathsf{R}_{n,k,b}$*, the cross-split statistic* $a(\mathsf{r}, D)$ *can be represented by*

$$a(\mathsf{r}, D) = \frac{1}{k} \sum_{i \in \mathsf{s}} T(\mathsf{s}_j, D) \ . \tag{4.2}$$

Moreover, for all sets s *in* $\mathcal{S}_{n,b}$*, the sample-split statistic* $T(\mathsf{s}, D)$ *can be represented by*

$$T(\mathsf{s}, D) = \frac{1}{b} \sum_{i \in \mathsf{s}} \psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \tag{4.3}$$

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

We make these restrictions to ease exposition. Assumption 4.2 is satisfied if, for example, the statistic under consideration is a cross-fit treatment effect or cross-validated mean-squared error estimate. In principle, Assumption 4.2 rules out some applications of interest. In practice, so long as the linear representations (4.2) and (4.3) hold up to a suitable degree of approximation, the qualitative and quantitative predictions of the results that follow will continue to hold.[19] In particular, we show below that the predictions of our results play out in the data from Chakravorty et al. (2024), where the statistic $T(\mathsf{s}, D)$ is a p-value of the form (3.11).

The remaining assumptions, and our ensuing results, are expressed in terms of two objects that measure the sensitivity of the statistic under consideration to perturbations of the data and of the splits, respectively. We refer to these objects as stabilities. They are defined as follows.

Definition 4.1 (Sample Stability)**.** Fix a set $\mathsf{s} \subseteq \mathcal{S}_{n,b}$ and let i be an arbitrary element of s. 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_j with D'_j in D for each j in q. Let q be a randomly selected subset of $\tilde{\mathsf{s}}$ of cardinality q. We refer to the quantity

$$\sigma^{(r,q)} = \mathbb{E}\left[\left|\psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) - \psi(D_i, \hat{\eta}(\tilde{D}^{(\mathsf{q})}_{\tilde{\mathsf{s}}}))\right|^r\right] \tag{4.4}$$

as the (r, q)-order sample stability.

Definition 4.2 (Split Stability)**.** We refer to the quantity

$$\zeta^{(r)} = \mathbb{E}\left[\max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T(\mathsf{s}, D) - T(\mathsf{s}', D)\right)^r\right] . \tag{4.5}$$

[19] Extensions of our results to cases where the function $\mathcal{A}(\cdot)$ or the statistic $T(\mathsf{s}, D)$ satisfy component-wise Lipschitz or bounded differences conditions, say, are feasible, and will exhibit the same qualitative behavior.

as the rth-order split stability.

We restrict attention to statistics whose sample stabilities decay in a suitable way with the sample size n. Throughout, we say $x \lesssim y$ if there exists a universal constant C such that $x \leq Cy$.

Assumption 4.3 (Sample Stability Decay)**.** *The (r, q)-order sample stability satisfies the bound*

$$\sigma^{(r,q)} \lesssim \left(\frac{\sqrt{q}}{n-b}\right)^r \tag{4.6}$$

uniformly for each q in $[b]$ and r in $\{2, 4\}$.

We call a statistic sample stable if it satisfies Assumption 4.3.[20] Many sample-split statistics of interest are sample stable. In Appendix B.3, we show that statistics satisfying Assumption 4.2 are sample stable if the nuisance parameter estimator $\hat{\eta}(\cdot)$ is an empirical risk minimizer of a, potentially regularized, strictly convex loss. There is a large literature that gives analogous bounds for other standard machine learning estimators, including bagged or subsampled estimators, like random forests (Chen et al., 2022; Ritzwoller and Syrgkanis, 2024), ensemble estimators (Elisseeff et al., 2005), and estimators computed with stochastic gradient descent (Hardt et al., 2016).[21]

Nevertheless, sample-stability should be viewed as a strong assumption, that is only applicable to highly regular estimators. Below, we show that the predictions that follow from the imposition of sample-stability, concerning the qualitative behavior of the residual randomness, play out in the applications to Casey et al. (2021) and Chakravorty et al. (2024). These predications may not have the same quality in settings that use less well-behaved nuisance parameter estimators.

The split stability $\zeta^{(r)}$ is a less frequently studied object. 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 the statistic $T(\mathsf{s}, D)$ is bounded.

4.2 Computation and Concentration

Algorithm 1 entails sequentially computing the statistic $a(\mathsf{R}_{g,k}, D)$, as g increases, until a stopping criteria is satisfied. How much computation should we expect to do? And how does this quantity depend on the parameters k and ξ?

[20] The $(2, 1)$-order sample stability $\sigma^{(2,1)}$ is a widely studied object in the statistical learning literature, where it is referred to as mean-square stability (see e.g., Bousquet and Elisseeff, 2002; Kale et al., 2011; Kumar et al., 2013).

[21] Chen et al. (2022) show that, under some regularity conditions, sample-splitting is unnecessary for the consistency and asymptotic normality of DML estimators, if a condition related to, but partially stronger than, Assumption 4.3 is satisfied. We comment on the relationship between Assumption 4.3 and the conditions considered in Chen et al. (2022) in Appendix B.3.

Recall from Section 3.1 that we should expect the total number of splits $\hat{m} = \hat{g} \cdot k$ used by Algorithm 1 to be close to the "oracle" quantity

$$m^\star = g^\star \cdot k \approx 2k \cdot v_{1,k}(D) \left(\frac{z_{1-\beta/2}}{\xi} \right)^2 , \tag{4.7}$$

where $v_{1,k}(D) = \mathrm{Var}(a(\mathsf{r}, D) \mid D)$ denotes the conditional variance of the statistic computed using a single cross-split. Two aspects of the expression (4.7) are worth highlighting. First, the total number of splits depends on the error tolerance ξ through the factor ξ^{-2}. In other words, in order to reduce the reproducibility error by a factor of 10, e.g., to move from $\xi = 0.1$ to $\xi = 0.01$, the total number of splits must be increased by a factor of 100. Second, the total number of splits depends on the parameter k through the factor $k \cdot v_{1,k}(D)$. How should we expect this quantity to scale with k?

We answer this question with the following result, which characterizes the rate of convergence of the statistic $a(\mathsf{R}_{g,k}, D)$ around its conditional mean, given by

$$\overline{a}\,(D) = \mathbb{E}\left[a(\mathsf{R}_{g,k}, D) \mid D\right] = \mathbb{E}\left[T(\mathsf{s}_{i,j}, D) \mid D\right] \tag{4.8}$$

under Assumption 4.2. The nonstandard aspect of this result 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 applying a coupling argument due to Chatterjee (2005, 2007).[22]

Theorem 4.1. *Suppose that Assumptions 4.1 to 4.3 hold, the data D are independently and identically distributed, and $n = k \cdot b$. If the 4th-order split stability $\zeta^{(4)}$ is finite, then for each $\delta > 0$, the inequality*

$$v_{g,k}(D) = \mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\,(D)\right)^2 \mid D\right] \lesssim \frac{1}{\delta} \frac{b-1}{n^2} \frac{1}{g} . \tag{4.9}$$

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

The left-hand side of the inequality (4.9) is random through the data D. Theorem 4.1 says that, if $\mathcal{F}$ is the event that the inequality (4.9) holds, then $P\{\mathcal{F}\} > 1 - \delta$ unconditionally. This bound results from an application of Markov's inequality, at one point in the proof, to bound a complicated, data-dependent term with a term that depends on the sample stability. This strategy—bounding the conditional quantities of interest with unconditional quantities—is helpful because the resultant

[22] Theorem 4.1 is closely related to the unconditional variance bounds given in Kale et al. (2011) and Kumar et al. (2013), who give bounds with the same dependence on k for cross-validated risk estimation. Our result follows from a different method of argument. In particular, Theorem 4.1 is a corollary of a more general result, presented in Appendix D, that gives an analogous large deviations bound. In particular, we show that, for each $\varepsilon > 1$, the bound

$$P\left\{ |a(\mathsf{R}_{g,k}, D) - \overline{a}(D)| \leq \sqrt{\frac{b-1}{n^2} \frac{1}{g} \frac{\log(\varepsilon^{-1})}{\delta}} \mid D \right\} \geq 1 - \varepsilon$$

holds with probability greater than $1 - \delta$ as D varies. That is, the rate of convergence suggested by Theorem 4.1 holds for all higher-order moments as well. This large deviations bound is applied repeatedly to establish the result given in the following subsection.

unconditional object, the sample stability, is tractable and has been characterized in many settings of interest.

Theorem 4.1 demonstrates that the variance $v_{g,k}(D)$ converges to zero at the rate

$$\frac{1}{n}\frac{b-1}{n}\frac{1}{g} \leq \frac{1}{n}\frac{1}{k}\frac{1}{g}\,. \tag{4.10}$$

Consequently, for a fixed total number of splits $m = g \cdot k$, the rate of convergence is proportional to $(mn)^{-1}$. That is, the conditional randomness of aggregate statistics constructed with m sample-splits concentrates like averages of m i.i.d. random variables, despite the dependence across cross-splits. Observe, also, that if $b = 1$, corresponding to "jackknife" or "leave-one-out" sample-splitting, then there is no residual randomness and the left-hand-side of (4.10) collapses to zero.

Plugging the bound (4.9) into the expression (4.7), we find that

$$m^\star \approx \frac{1}{n}\left(\frac{z_{1-\beta/2}}{\xi}\right)^2\,, \tag{4.11}$$

In other words, Theorem 4.1 implies that—in contrast to the error tolerance ξ—the number of cross-folds k does not affect the required computation. On the other hand, all else equal, less aggregation is needed to remove residual randomness from settings with larger sample sizes n.

Theorem 4.1 plays out empirically. Figure 4 displays measurements of the conditional variance $v_{1,k}(D)$ in dark blue, as k varies, for our applications to Casey et al. (2021) and Chakravorty et al. (2024). For the application to Casey et al. (2021), we use the cross-validated estimate of the mean-squared error at the value of λ that minimizes the curves displayed in Panel A of Figure 3. For the application to Chakravorty et al. (2024), we use the p-value (3.11), associated with the cross-fit estimate of the average treatment effect. If the variance bound (4.9) is accurate, then the approximation

$$v_{1,k}(D) \approx \left(\frac{k'}{k}\right) v_{1,k'}(D) \tag{4.12}$$

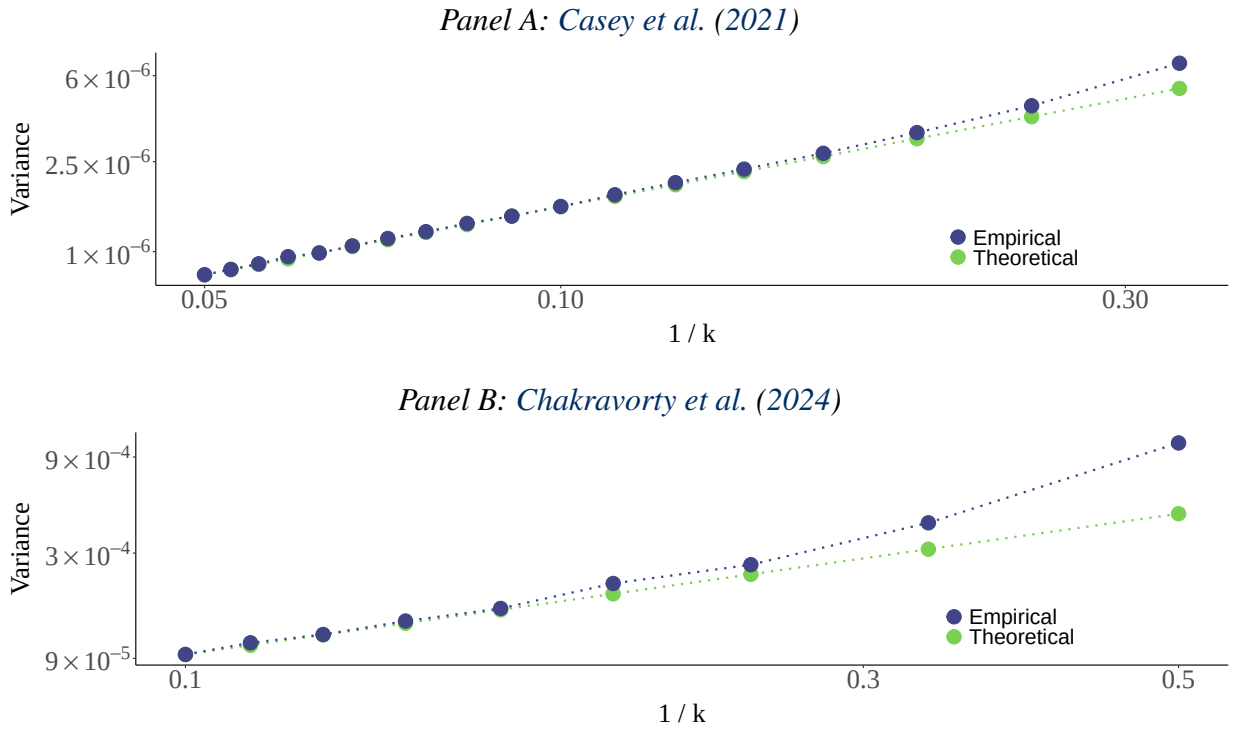
should hold for each pair k, k'. To test this, we display estimates of right-hand-side of (4.12) in light green, where k' is the largest value of k considered in each sub-figure. In each case, the approximation is remarkably accurate.[23]

4.3 Reproducibility

Algorithm 1 confers a guarantee. In particular, statistics aggregated with Algorithm 1 should be interpreted as being reproducible, up to an error tolerance ξ, with probability greater than β.

[23]It is worth emphasizing that neither Assumption 4.1 nor Assumption 4.2 holds for the application to Chakravorty et al. (2024). That is, in this case, nuisance parameters are estimated with random forests, which are random conditional on the data, and the p-value (3.11) does not admit an exact representation of the form (4.3). Rather, in this setting, both of these assumptions are good approximations (so long as the number of trees used to construct the random forests is sufficiently large), and so the predictions of Theorem 4.1 hold.

FIGURE 4. Concentration



Notes: Figure 4 illustrates the concentration of the residual randomness of various cross-split statistics with the number of cross-splits k, using data from Casey et al. (2021) and Chakravorty et al. (2024). For the application to Casey et al. (2021), we use the cross-validated estimate of the mean-squared error at the value of λ that minimizes the curves displayed in Panel A of Figure 3. For the application to Chakravorty et al. (2024), we use the p-value of the form (3.11), associated with the cross-fit estimate of the average treatment effect. The x-axes give $1/k$. The y-axes give measurements of the conditional variance of each statistic. Both the axes are displayed on a logarithmic scale, base 10. The theoretical prediction (4.12), based on Theorem 4.1, is given in light green.

How accurate is this guarantee? In particular, how does the accuracy of the nominal reproducibility error β depend on the choice parameters ξ and k? These questions are answered by the following Berry-Esseen type bound on the accuracy of the nominal reproducibility of Algorithm 1.

Theorem 4.2. *Suppose that the collections* $\mathsf{R}_{\hat{g},k}$ *and* $\mathsf{R}'_{\hat{g}',k}$ *are independently obtained using Algorithm 1. If Assumptions 4.1 to 4.3 hold, the conditional variance* $v_{1,k}(D) = \mathrm{Var}(a(\mathsf{r}, D) \mid D)$ *is strictly positive, almost surely, the data* D *are independent and identically distributed, and the eighth-order split stability* $\zeta^{(8)}$ *is finite, then for all sufficiently small* ξ*, the inequality*

$$\left| P\left\{ \left| a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D) \right| \geq \xi \mid D \right\} - \beta \right| \lesssim \frac{1}{\delta^{3/4}} \frac{1}{kn} \left(\frac{1}{v_{1,k}(D)} \right)^{5/4} \left(\frac{\xi}{z_{1-\beta/2}} \right)^{1/2}, \tag{4.13}$$

holds with probability greater than $1-\delta$ *as* D *varies, where in writing* (4.13)*, we have omitted a multiplicative term that converges to zero logarithmically as* ξ *decreases to zero.*

Theorem 4.2 describes the settings under which Algorithm 1 is accurate. To unpack this result, observe that the variance bound (4.9) gives

$$\frac{1}{kn}\left(\frac{1}{v_{1,k}(D)}\right)^{5/4}\left(\frac{\xi}{z_{1-\beta/2}}\right)^{1/2} \gtrsim \frac{k^{1/4}}{n}\left(\frac{\xi}{z_{1-\beta/2}}\right)^{1/2}. \tag{4.14}$$

This suggests that the performance of Algorithm 1 improves as k and ξ decrease and as n increases.[24] These predictions hold empirically. Panel A of Figure 5 displays estimates of the reproducibility error in the application to Casey et al. (2021) as k and ξ vary. An analogous figure for the application to Chakravorty et al. (2024) is displayed in Appendix A.5. As predicted, over most of the range of ξ, the reproducibility error is increasing as k increases.[25]

Likewise, Panel B displays estimates of the average number of cross-splits $\hat{g}$ used in Algorithm 1, at each value of ξ and k. Again, as predicted, the total number of splits used by the procedure stays constant as k varies. To see this, observe that the average value of $\hat{g}$ is roughly 10 times smaller for $k=20$ than for $k=2$, as predicted by the approximation (4.11). Comparing Panels A and B, observe that, once the average value of $\hat{g}$ is larger than approximately 500, the reproducibility error is close to the nominal error rate β.

If the total number of splits required to ensure reproducibility at a given error tolerance ξ is constant as k varies, then why does the performance of Algorithm 1 decrease with k? As part of the proof of Theorem 4.2, we show that, for each $\varepsilon>0$, the bound

$$P\left\{\left|\frac{\hat{g}}{g^\star}-1\right| \lesssim \frac{1}{n}\frac{1}{k}\left(\frac{1}{v_{1,k}(D)}\right)^{3/2}\frac{\xi}{z_{1-\beta/2}}\sqrt{\frac{\log(\varepsilon^{-1})}{\delta}} \mid D\right\} \geq 1-\varepsilon \tag{4.15}$$

holds with probability greater than $1-\delta$, as D varies. Again the variance bound (4.9) gives

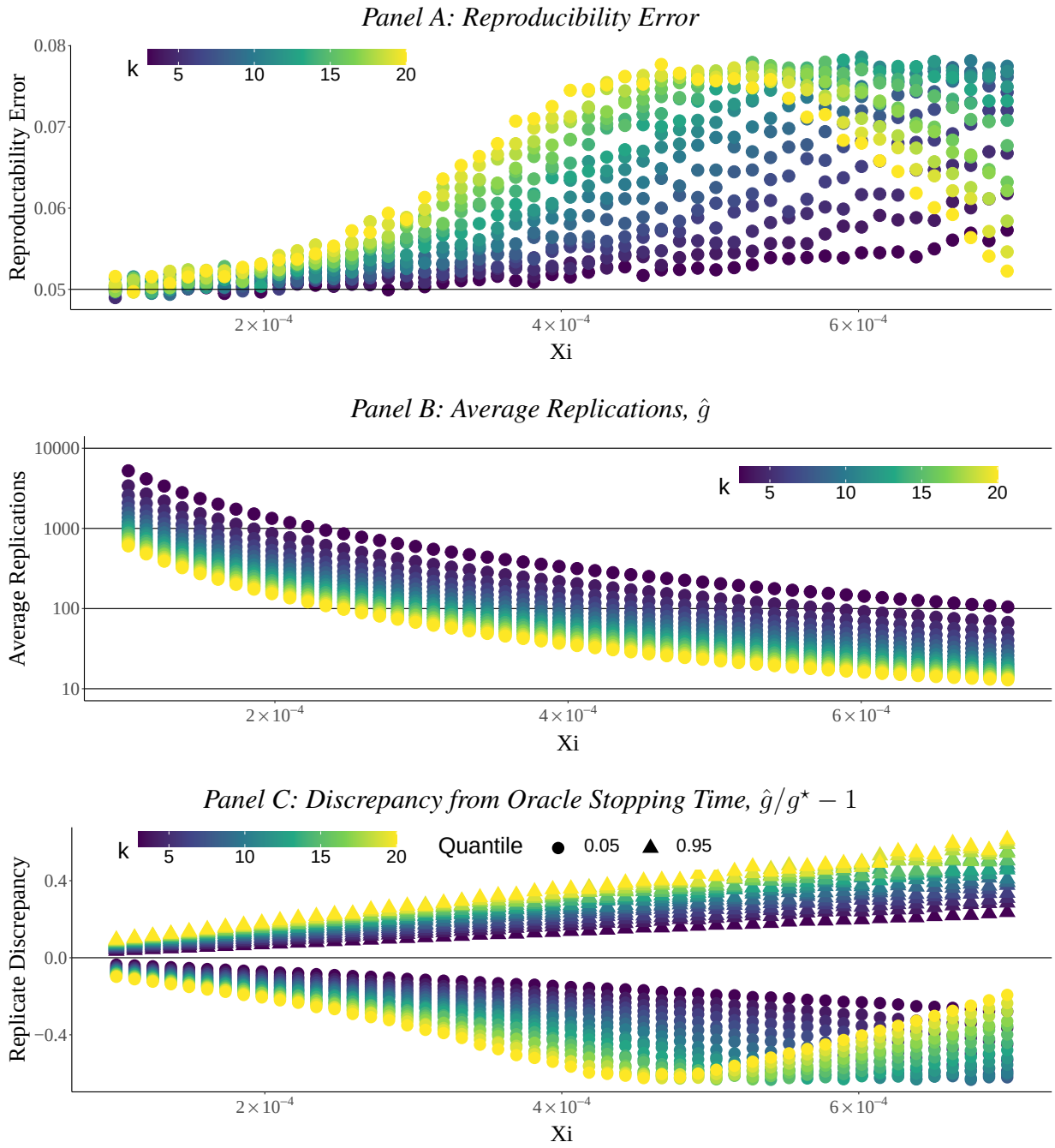
$$\frac{1}{n}\frac{1}{k}\left(\frac{1}{v_{1,k}(D)}\right)^{3/2}\frac{\xi}{z_{1-\beta/2}} \gtrsim \frac{k^{1/2}}{n}\frac{\xi}{z_{1-\beta/2}}. \tag{4.16}$$

Written differently, as k increases the discrepancy between the realized and oracle number of cross-splits, $|\hat{g}/g^\star-1|$, increases, reducing the accuracy of the nominal error rate. Roughly speaking, this happens because the quality of the estimator $\hat{v}_{g,k}(D)$ for the conditional variance $v_{g,k}(D)$ depends

[24] The dependence of the bound (4.13) on ξ is sharp, at least up to the logarithmic factor. This follows from general results concerning randomly stopped sums given in Landers and Rogge (1976, 1988).

[25] Note, however, that for large values of ξ, the reproducibility error decreases as k increases. This is due to early stopping. That is, if the variance $v_{1,k}(D)$ is large relative to ξ, then there is an increased chance that $\hat{g}$ stops immediately after the burn-in period, i.e., $\hat{g}=g_{\mathsf{init}}$.

FIGURE 5. Performance in Application to Casey et al. (2021)



Notes: Figure 5 displays measurements of the performance of Algorithm 1 on the data from Casey et al. (2021). Panel A displays measurements of the reproducibility error, $P\{|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)| \geq \xi \mid D\}$, as ξ and k vary. A solid horizontal line is displayed at the nominal error rate $\beta = 0.05$. Panel B displays measurements of the average number of replications $\hat{g}$ s ξ and k vary. The y-axis is displayed with a log scale, base 10. Solid horizontal lines are placed at each exponential factor of 10. Panel C displays measurements of the 5th and 95th quantiles of the discrepancy $\hat{g}/g^\star - 1$ as k and ξ vary. Further details on the construction of this figure are given in Appendix A.5.

only on the number of cross-splits g.[26] That is, for a given number of sample-splits $\hat{m} = k \cdot \hat{g}$, Algorithm 1 is most accurate when $\hat{g}$ is large, as the variance estimate $\hat{v}_{\hat{g},k}(D)$ is more precise.

As before, these predictions play out in practice. Panel C of Figure 5 displays estimates of the 5th and 95th quantiles of the distribution of the discrepancy $\hat{g}/g^\star - 1$ as k and ξ vary. Over most of the range of ξ the discrepancy between $\hat{g}$ and $g^\star$ is increasing in k. At small values of k and large values of ξ, there is an increased chance that $\hat{g}$ stops immediately after the burn-in period, i.e., $\hat{g} = g_{\mathsf{init}}$.

It is worth pausing to note that these results do not support eschewing cross-splitting altogether, i.e., setting k equal to one and aggregating over independent splits, as we have restricted attention to the case that $n = k \cdot b$. In Appendix D, we give analogous results that relax this assumption and show that cross-splitting, i.e., setting $n = k \cdot b$, reduces residual randomness at a faster rate than independent splitting, i.e., setting k equal to one. In other words, all else equal, Algorithm 1 performs best, in the sense that the nominal reproducibility error is most accurate, when k is equal to 2 and b is equal to $b/2$.

To summarize, the computation needed to achieve a desired bound on residual randomness is highly sensitive to the error tolerance ξ, but is not affected by the number of cross-folds k. On the other hand, the accuracy of the nominal error rate of Algorithm 1 decreases as the number of cross-folds k increases. If the error bound ξ is chosen to be suitably small, such that the realized number of cross-splits $\hat{g}$ is greater than roughly 500, then the nominal reproducibility probability tends to be quite accurate, and insensitive to changes in the number of cross-splits.

5. Recommendations for Practice

Sample-splitting is a helpful tool for simplifying many widely encountered problems in applied econometrics. This simplification comes at the cost of the introduction of residual randomness. We have shown, in several applications, that this residual randomness is large enough to substantively affect results. To address this, we have proposed a simple procedure, summarized in Algorithm 1, for removing the auxiliary randomness from sample-split statistics. The procedure takes as input a bound and an error rate. We have shown that, if the procedure were run twice, the chance that the results differ by more than the bound is well-approximated by the error rate.

We conclude, in this section, by detailing several recommendations for how to best implement Algorithm 1 in practice. The most important choice to make is the specification of the sample-split statistic of interest. For example, suppose that we are interested in estimating an average treatment effect, or other causal contrast, using an estimator based on sample-splitting. Denote this quantity

[26] Similarly, in Appendix D, we show that the quality of a normal approximation to $a(\mathsf{R}_{g,k}, D)$ depends only on g, although this discrepancy is not the leading term in the reproducibility error. The close approximation exhibited in Figure 4 suggests that is may be reasonable to estimate $v_{1,k}(D)$ by taking the sample variance both across and within cross-splits, i.e., computing the sample variance across all m sample-splits. Although this worth further consideration, asymptotic validity, i.e., Theorem 3.1, would not hold at the same level of generality.

by $\mathsf{est}(\mathsf{r}, D)$ and let $\mathsf{se}(\mathsf{r}, D)$ denote an associated, potentially sample-split, standard error estimate. There are several reasonable choices that one might make. In this case, we recommend applying Algorithm 1 to sequentially aggregate the p-value

$$a(\mathsf{r}, D) = 1 - \Phi\left(\frac{\mathsf{est}(\mathsf{r}, D)}{\mathsf{se}(\mathsf{r}, D)}\right) \tag{5.1}$$

associated with a test that the contrast of interest is greater than zero, where $\Phi(\cdot)$ denotes the standard normal c.d.f. This approach has the benefit of benchmarking residual randomness with an estimate of the sampling error.

In some settings, it may be of interest to report the scope of residual randomness for alternative quantities associated with statistics that have aggregated with Algorithm 1. For example, suppose that we have applied Algorithm 1 to stabilize the p-value (5.1), and obtain $a(\mathsf{R}_{\hat{g},k}, D)$. We may also wish to report the associated estimate $\mathsf{est}(\mathsf{R}_{\hat{g},k}, D)$. In this case, we recommend reporting and interpreting the standard error

$$\sqrt{\frac{1}{\hat{g}}\frac{1}{\hat{g}-1}\sum_{i=1}^{\hat{g}}(\mathsf{est}(\mathsf{r}_i, D) - \mathsf{est}(\mathsf{R}_{\hat{g},k}, D))^2} \tag{5.2}$$

in the usual way. That is, the standard error (5.2) estimates the residual randomness of the estimate $\mathsf{est}(\mathsf{R}_{\hat{g},k}, D)$, conditional on the data. If this is too large, the error tolerance ξ should be decreased.

The optimal choice of statistic in settings where cross-validation is used for model selection and risk estimation is less immediate. In this case, we recommend applying Algorithm 1 to each element of the statistic (3.12), i.e., the relative values of the risk estimates.

The second most important choice to make is the choice of the error tolerance ξ. In some cases, an appropriate choice is clear. For example, it may be desirable to ensure that a p-value is reproducible at the level relevant for the determination of statistical significance at level 0.10 or 0.025, e.g., setting $\xi = 0.01$ or $\xi = 0.001$. In other cases, like in applications to cross-validated model selection, this choice is less clear. We propose a procedure for making this choice in that setting in Appendix B.2.

The error tolerance ξ should be set to a value that is sufficiently small so that the nominal reproducibility error is accurate. On the other hand, ξ should not be set so small that the computation associated with implementing Algorithm 1 becomes infeasible. To ensure that the chosen value of ξ satisfies these constraints, we recommend computing the statistic of interest for each of a small, initial sample of cross-splits (e.g., g equal to 20 or 30). Let $\hat{v}_{1,k}(D)$ denote an estimate of the conditional variance $v_{1,k}(D)$ computed using this sample. An estimate of the total number of cross-splits needed to implement Algorithm 1 at a specified level of ξ can then be obtained by

$$g(\xi) = 2\hat{v}_{1,k}(D)\left(\frac{z_{1-\beta/2}}{\xi}\right)^2 . \tag{5.3}$$

Motivated by the numerical results reported in Section 4, we recommend choosing a value of ξ such that this estimate is greater than 500.

In some cases, the estimate (5.3) may be too large at practically relevant values of ξ to be computationally feasible. Unfortunately, as we have shown in Section 4, changing the number of cross-folds k will not address this issue, and it may be best to consider alternative choices of nuisance parameter estimators. In this situation, one might consider using "leave-one-out" or "jackknife" sample-splitting, i.e., setting $k = n$, which exhibits no residual randomness. This is not an omnibus fix, however. For example, proofs of the asymptotic normality of DML estimates of average treatment effects require that k is small relative to n (Chernozhukov et al., 2018). Similarly, leave-one-out cross-validation is not necessarily optimal for model selection (see e.g., Shao (1993) for early discussion of this point). Chetverikov (2024) gives a review of various, alternative, approaches to tuning parameter selection.

If the number of cross-folds k is set too high, the realized number of cross-splits $\hat{g}$ may be too small and Algorithm 1 may perform poorly. But, so long as the user ensures that the error tolerance ξ is sufficiently small such that the number of cross-splits $\hat{g}$ tends to be large (e.g., greater than 500, say), small changes in the number of folds k should not have adverse effects. Thus, this choice should be made on substantive grounds, that will depend on the application. In practice, the conventional choices of k equal to 2, 5, or 10 should work well in most settings.

We have had relatively little to say about the choice of the nominal reproducibility error β. We have found that setting $\beta = 0.05$ is suitable for most applications. Finally, although it has not played a central role in our theoretical analysis, it is important to choose the burn-in period g_{init} to be suitably large to ensure that the variance estimate $\hat{v}_{g_{\mathsf{init}},k}(D)$ is reasonably accurate, otherwise Algorithm 1 will tend to stop too early. We recommend setting g_{init} to be equal to at least 10.

REFERENCES

Abaluck, J., Agha, L., Kabrhel, C., Raja, A., and Venkatesh, A. (2016). The determinants of productivity in medical testing: Intensity and allocation of care. *American Economic Review*, 106(12):3730–3764.

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.

Adermon, A., Lindahl, M., and Palme, M. (2021). Dynastic human capital, inequality, and intergenerational mobility. *American Economic Review*, 111(5):1523–1548.

Agarwal, N., Ashlagi, I., Azevedo, E., Featherstone, C. R., and Karaduman, Ö. (2019). Market failure in kidney exchange. *American Economic Review*, 109(11):4026–4070.

Ahrens, A., Hansen, C. B., Schaffer, M. E., and Wiemann, T. (2024). ddml: Double/debiased machine learning in stata. *The Stata Journal*, 24(1):3–45.

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.

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

Arnold, D., Dobbie, W., and Hull, P. (2020). Measuring racial discrimination in bail decisions. Technical report, National Bureau of Economic Research.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

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

Bach, P., Kurz, M. S., Chernozhukov, V., Spindler, M., and Klaassen, S. (2024). DoubleML: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108(3):1–56.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of statistics*, pages 2055–2085.

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.

Beaman, L., Karlan, D., Thuysbaert, B., and Udry, C. (2023). Selection into credit markets: Evidence from agriculture in mali. *Econometrica*, 91(5):1595–1627.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Beraja, M., Kao, A., Yang, D. Y., and Yuchtman, N. (2023). Ai-tocracy. *The Quarterly Journal of Economics*, 138(3):1349–1402.

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.

Blattman, C., Duncan, G., Lessing, B., and Tobón, S. (2024). Gang rule: Understanding and countering criminal governance. *Review of Economic Studies*, page rdae079.

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.

Cantoni, E. and Pons, V. (2022). Does context outweigh individual characteristics in driving voting behavior? evidence from relocations within the united states. *American Economic Review*, 112(4):1226–1272.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2020). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics*, 135(1):269–327.

Casey, K., Kamara, A. B., and Meriggi, N. F. (2021). An experiment in candidate selection. *American Economic Review*, 111(5):1575–1612.

Chakravorty, B., Arulampalam, W., Bhatiya, A. Y., Imbert, C., and Rathelot, R. (2024). Can information about jobs improve the effectiveness of vocational training? experimental evidence from india. *Journal of Development Economics*, 169:103273.

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, 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.

Chen, Z., Liu, Z., Suárez Serrato, J. C., and Xu, D. Y. (2021). Notching r&d investment with corporate income tax cuts in china. *American Economic Review*, 111(7):2065–2100.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). 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 Fernández-Val, I. (2023). Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in india. *arXiv preprint arxiv:1712.04802*.

Chetverikov, D. (2024). Tuning parameter selection in econometrics. *arXiv preprint arXiv:2405.03021*.

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

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.

Cicala, S. (2022). Imperfect markets versus imperfect regulation in us electricity generation. *American Economic Review*, 112(2):409–441.

Cooper, Z., Scott Morton, F., and Shekita, N. (2020). Surprise! out-of-network billing for emergency care in the united states. *Journal of Political Economy*, 128(9):3626–3677.

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

Covert, T. R. and Sweeney, R. L. (2023). Relinquishing riches: Auctions versus informal negotiations in texas oil and gas leasing. *American Economic Review*, 113(3):628–663.

Delfino, A. (2024). Breaking gender barriers: Experimental evidence on men in pink-collar jobs. *American Economic Review*, 114(6):1816–1853.

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

Derenoncourt, E. (2022). Can you move to opportunity? evidence from the great migration. *American Economic Review*, 112(2):369–408.

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D., and Reif, J. (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review*, 109(12):4178–4219.

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.

Dubé, J.-P. and Misra, S. (2023). Personalized pricing and consumer welfare. *Journal of Political Economy*, 131(1):131–189.

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

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. (2022). General equilibrium effects of cash transfers: experimental evidence from kenya. *Econometrica*, 90(6):2603–2643.

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

Ellison, G. and Pathak, P. A. (2021). The efficiency of race-neutral alternatives to race-based affirmative action: Evidence from chicago's exam schools. *American Economic Review*, 111(3):943–975.

Farronato, C., Fradkin, A., Larsen, B. J., and Brynjolfsson, E. (2024). Consumer protection in an online world: An analysis of occupational licensing. *American Economic Journal: Applied Economics*, 16(3):549–579.

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

Fudenberg, D. and Liang, A. (2019). Predicting and understanding initial play. *American Economic Review*, 109(12):4112–4141.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

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.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

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.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Haushofer, J., Niehaus, P., Paramo, C., Miguel, E., and Walker, M. W. (2022). Targeting impact versus deprivation. Technical report, National Bureau of Economic Research.

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

Kelly, M., Mokyr, J., and Ó Gráda, C. (2023). The mechanics of the industrial revolution. *Journal of Political Economy*, 131(1):59–94.

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

Koijen, R. S., Richmond, R. J., and Yogo, M. (2024). Which investors matter for equity valuations and expected returns? *Review of Economic Studies*, 91(4):2387–2424.

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.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

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

Mastrobuoni, G. (2020). Crime is terribly revealing: Information technology and police productivity. *The Review of Economic Studies*, 87(6):2727–2753.

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.

Motwani, R. and Raghavan, P. (1995). *Randomized algorithms*. Cambridge university press.

Muendler, M.-A. and Becker, S. O. (2010). Margins of multinational labor substitution. *American Economic Review*, 100(5):1999–2030.

Myers, K. R. and Lanahan, L. (2022). Estimating spillovers from publicly funded r&d: Evidence from the us department of energy. *American Economic Review*, 112(7):2393–2423.

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

Okunogbe, O. and Pouliquen, V. (2022). Technology, taxation, and corruption: evidence from the introduction of electronic tax filing. *American Economic Journal: Economic Policy*, 14(1):341–372.

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.

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.

Ritzwoller, D. M. and Syrgkanis, V. (2024). Uniform inference for subsampled moment regression. *arXiv preprint arXiv:2405.07860*.

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.

Sadka, J., Seira, E., and Woodruff, C. (2024). Information and bargaining through agents: Experimental evidence from mexico's labour courts. *Review of Economic Studies*, page rdae003.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494.

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.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence*. Springer.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.

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.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2024). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, pages 1–14.

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: November 15, 2024

APPENDIX A. DATA AND SIMULATIONS

A.1 Casey et al. (2021)

We obtain the data associated with Casey et al. (2021) from the replication package posted on OpenICPSR.[27] The units of observation are the aspirant political candidates in Sierra Leone's 2018 parliamentary elections. The data that we consider consist of an outcome and 48 covariates. The outcome is the vote share for each aspirant candidate in a poll of party officials. The covariates collect various measurements relating to the candidate's personal, political, and financial characteristics. All data cleaning steps match Casey et al. (2021).

We consider the exercise summarized in Appendix Table A.4 of Casey et al. (2021). These authors repeatedly compute 10-fold cross-validated estimates of the risk associated with a regularized regression of the outcome on the collection of covariates, at each value of a grid of penalization parameters. They include both Ridge and Lasso type penalties. For each replication, they compute the covariates that are selected in the regression associated with the penalty parameter having the minimum risk estimate. They display the covariates that are selected in more than 200 of the 400 replications. These covariates are included as controls in various downstream analyses. In our replication of this exercise, to ease exposition, we only include a Lasso type penalty.

A.2 Chakravorty et al. (2024)

We obtain the data and replication package associated with Chakravorty et al. (2024) directly from the authors.[28] The units of observation are trainees in a job training program implemented in the Indian states of Bihar and Jharkhand. There are 2,488 total trainees, who were divided into batches of approximately 30 trainees. Batches of trainees were randomly assigned to treatment with a program involving the provision of information concerning potential placement jobs. All data cleaning steps match those taken in Chakravorty et al. (2024) .

We replicate the analysis reported in Column 4 of Table 1 of Chakravorty et al. (2024). This estimate restricts attention to the 890 trainees that completed the training and were subsequently placed into a job. The outcome variable is an indicator for whether the trainee is still in the job five months after training completion. The covariates collect measurements of 77 attributes related to the demographics, human capital, and expectations of the trainees. These covariates are listed in Table 3.1 of Chakravorty et al. (2024).

Following Chakravorty et al. (2024), we estimate the effect of the treatment using the implementation of Double Machine Learning available through the "DoubleML" R package (Bach et al., 2024). We use 5-fold cross-fitting. Nuisance parameters (i.e., outcome regressions and propensity scores) are

[27] The replication package associated with Casey et al. (2021) is posted at the URL `https://www.openicpsr.org/openicpsr/project/124501/version/V1/view`.

[28] We thank Bhaskar Chakravorty for sharing this material.

estimated with Random Forests using the default tuning parameters associated with the "Ranger" R package (Wright and Ziegler, 2017). Standard errors are clustered at the batch level.

A.3 Haushofer et al. (2022)

Haushofer et al. (2022) consider data originally studied in Egger et al. (2022). At the time of the preparation of this paper, no replication materials for Haushofer et al. (2022) were publicly available. We replicate a simplified version of the exercise considered in Haushofer et al. (2022), using data obtained from the replication package associated with Egger et al. (2022).[29] The units of observation are households in three sub-counties of Siaya County, western Kenya. Households were randomly allocated large cash transfers. See Egger et al. (2022) and Haushofer et al. (2022) for further details on the design of this experiment.

Using the replication data from Egger et al. (2022), we begin with a sample of 5,423 treatment eligible households. We then restrict attention to 4,758 households who were surveyed in both the first and second round of the experiment. We drop any households that are missing measurements of post-treatment assets, consumption, income, or food security, leaving a sample of 4,755 households. These data are merged with pre-treatment measurements of the demographics, assets, food security, and labor market participation for each household, leaving a sample of 4,754 households.

We obtain a cleaned dataset that records 24 measurements associated with each household. The dataset identifies the village that contains each household, whether each household was assigned to treatment, and post-treatment measurements of indices of assets, consumption, income, and food security. The covariates are measurements of the household size, the number of meals eaten the day before the survey, the number of high-protein meals eaten the day before the survey, the time between the administration of the program and the measurement of post-treatment outcomes,[30] and indicators for whether the household contains a widow, contains a female, has children, has children under 3, has children under 6, contains an elderly resident, has livestock, has land, has more than a quarter acre of land, has a radio and tv, has a self-employed resident, and has an employed resident.

We implement a simplified version of the exercise studied in Haushofer et al. (2022). Consider a subset s. Let the complement of s in $[n]$ be denoted by $\tilde{\mathsf{s}}$. Let Y_i denote a measurement of an index quantifying post-treatment consumption, X_i denote a vector of measurements of pre-treatment covariates, and W_i denote assignment to treatment. Let H_i denote the number of occupants of household i. Let D_i collect these observations. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes

[29] The replication package for Egger et al. (2022) is available at https://www.econometricsociety.org/publications/econometrica/2022/11/01/General-Equilibrium-Effects-of-Cash-Transfers-Experimental-Evidence-From-Kenya.

[30] We use the measurement of the time between the administration of the program and the measurement of post-treatment outcomes available as "exptoend" in the dataset "GE_Survey_and_Transfer_Dates.dta" in the replication package for Egger et al. (2022). This variable appears to differ from the analogous quantity used in Haushofer et al. (2022), whose density is displayed in their Figure A.2.

induced by the treatment W_i. For each household i in s, let

$$\mathsf{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \quad \text{and} \quad \mathsf{uto}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \tag{A.1}$$

denote sample-split estimates of the treatment effect $Y_i(1) - Y_i(0)$ and *per-capita* untreated outcome $Y_i(0)/H_i$, respectively. In this case, the nuisance parameters $\hat{\eta}(D_{\tilde{\mathsf{s}}})$ consist of nonparametric estimates of outcome regressions, i.e., the conditional expectations (2.7), computed using the data for units i in $\tilde{\mathsf{s}}$.

Following Haushofer et al. (2022), we compute these estimates with Random Forests using the "GRF" R package (Athey et al., 2019).[31] As best as we can tell, there are two differences between our implementation and the results presented in Haushofer et al. (2022). First, we make no attempt to include sample weights that reflect the sampling probabilities. Second, we include the time between treatment and the measurement of post-treatment outcomes as a covariate, rather than implement some form of de-meaning and re-weighing of estimates across time.

Let $\mathsf{Impacted}(\mathsf{s})$ collect the 50% of units in s associated with the largest values of $\mathsf{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}}))$. Similarly, let $\mathsf{Deprived}(\mathsf{s})$ collect the 50% of units i in s associated with the smallest values of $\mathsf{uto}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}}))$. Let $\mathsf{Impacted}_w(\mathsf{s})$ collect the subset of $\mathsf{Impacted}(\mathsf{s})$ with $W_i = w$. Define $\mathsf{Deprived}_w(\mathsf{s})$ analogously. Consider the sample-split statistic

$$\begin{aligned} T(\mathsf{s}, D) = &\left(\frac{1}{|\mathsf{Impacted}_1(\mathsf{s})|} \sum_{i \in \mathsf{Impacted}_1(\mathsf{s})} Y_i - \frac{1}{|\mathsf{Impacted}_0(\mathsf{s})|} \sum_{i \in \mathsf{Impacted}_0(\mathsf{s})} Y_i \right) \\ &- \left(\frac{1}{|\mathsf{Deprived}_1(\mathsf{s})|} \sum_{i \in \mathsf{Deprived}_1(\mathsf{s})} Y_i - \frac{1}{|\mathsf{Deprived}_0(\mathsf{s})|} \sum_{i \in \mathsf{Deprived}_0(\mathsf{s})} Y_i \right) . \end{aligned} \tag{A.2}$$

That is, the sample-split statistic (A.2) is an estimator of the difference in the average treatment effect of the most impacted and most deprived units in s. The statistic (A.2) can be aggregated with cross-splitting, through

$$a(\mathsf{r}, D) = \frac{1}{k} \sum_{i=1}^{k} T(\mathsf{s}_i, D) \ , \tag{A.3}$$

where $\mathsf{r} = (\mathsf{s}_i)_{i=1}^k$ is a k-fold partition of $[n]$. Following Haushofer et al. (2022), in Section 2, we use 5-fold cross-fitting.

Haushofer et al. (2022) say that they construct a standard error for (A.3) with the bootstrap. There are a variety of ways that one might do this. In our implementation, we keep the values of the treatment effect and untreated outcome estimates (A.1) fixed for each household, and compute a

[31] Following Haushofer et al. (2022), we use the default tuning parameters, but set "sample.fraction" equal to 0.1 and "min.node.size" equal to 10.

bootstrap standard error for the statistic (A.3), re-computing the groups $\mathsf{Impacted}(\mathsf{s})$ and $\mathsf{Deprived}(\mathsf{s})$ in each bootstrap replicate.

A.4 Beaman et al. (2023)

We obtain the data associated with Beaman et al. (2023) from the associated replication package posted on the Econometric Society's webpage.[32] The units of observations are low-income farmers in Mali. The cleaned data consist of observations of 5210 farmers across 198 villages. All data cleaning steps match Beaman et al. (2023).

Beaman et al. (2023) consider a two-stage experiment. In the first stage, a microcredit organization randomly offered group-liability loans to women in 88 of the 198 villages. In the second stage, households were randomly offered cash grants. We replicate the analysis reported in Column (4) of Appendix Table VI of Beaman et al. (2023). In this setting, attention is restricted to the 2,142 households who received a loan. They are interested in testing whether the effects of cash-grants on farm profits are heterogeneous. For this problem, the data D_i consist of a measurement Y_i of the post-transfer profit for farm i, the variable W_i denotes assignment to the cash transfer, and the vector X_i collects a vector of pre-treatment measurements of the physical and financial attributes of each farm. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes induced by the treatment W_i.

Beaman et al. (2023) implement a sample-split test for treatment effect heterogeneity proposed by Chernozhukov et al. (2023). Let s denote a random half-sample of $[n]$, i.e., a random subset of $[n]$ of size $n/2$. As before $\tilde{\mathsf{s}}$ denotes the complement of s in $[n]$. For each farm i in s, let

$$\mathsf{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \tag{A.4}$$

denote sample-split estimates of the treatment effect $Y_i(1) - Y_i(0)$. The nuisance parameters $\hat{\eta}(D_{\tilde{\mathsf{s}}})$ consist of nonparametric estimates of outcome regressions, i.e., the conditional expectations (2.7), computed using the data for units i in $\tilde{\mathsf{s}}$. We compute these estimates with Random Forests using the "GRF" R package (Athey et al., 2019). Tuning parameters are chosen with cross-validation in the same manner implemented in Beaman et al. (2023).

To test for treatment effect heterogeneity, we estimate the coefficients of the linear regression

$$Y_i = \alpha + \tau_1 \cdot W_i + \tau_2 \cdot \mathsf{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) + \tau_3 \cdot W_i \cdot \mathsf{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) + \varepsilon_i \,. \tag{A.5}$$

The idea is that, if the treatment effect estimates are heterogeneous, then the "true" value of the interaction coefficient τ_3 should be positive. Let $\hat{\tau}_3(\mathsf{s}, D)$ denote the estimate of τ_3 computed with linear regression in the sample s. Let $\mathsf{se}(\mathsf{s}, D)$ denote the associated standard error, clustered at the

[32] The replication package for Beaman et al. (2023) is available at https://www.econometricsociety.org/publications/econometrica/2023/09/01/Selection-into-Credit-Markets-Evidence-from-Agriculture-in-Mali

village level. Chernozhukov et al. (2023) advocate for computing the median estimate and standard error over 250 half-splits s. Beaman et al. (2023) use 1000 half-splits.

A.5 Simulations

In this section, we give further details concerning the simulations whose results are reported in Section 4. For both the applications to Casey et al. (2021) and Chakravorty et al. (2024), we sample 100,000 replications of the cross-split statistic of over a range of values of k. In the application to Chakravorty et al. (2024), we have increased the number of trees used to compute the random forest nuisance parameter estimates, from 100 to 3000, in order to ensure that the residual randomness induced by cross-splitting is greater than the residual randomness induced by the nuisance parameter estimate. We use these replicates to estimate the variance $v_{1,k}(D)$ for each value of k, in addition to the oracle stopping time

$$g^\star = 2v_{1,k}(D)\left(\frac{z_{1-\beta/2}}{\xi}\right)^2, \tag{A.6}$$

at each value of k and ξ

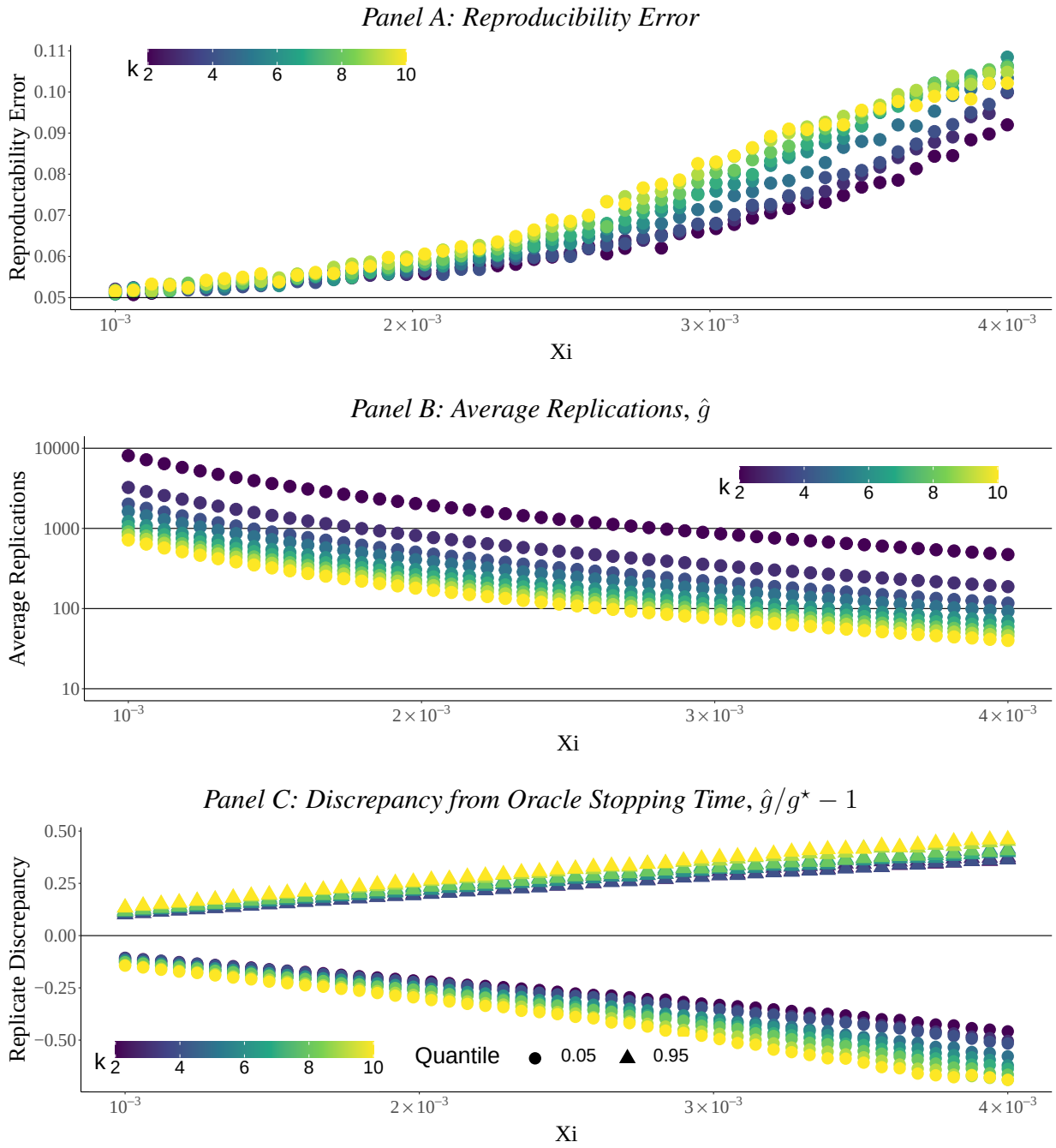
We implement Algorithm 1 100,000 times for each value of k and ξ, for each application, by sampling with replacement from the replicates of 100,000 draws. We use these estimates to estimate the reproducibility error

$$P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \geq \xi \mid D\right\} \tag{A.7}$$

as well as the distribution of $\hat{g}$ for each k and ξ. These estimates are used to construct Figure 5.

Figure A.1 displays measurements analogous to those displayed in Figure 5 for the application to Chakravorty et al. (2024). The accuracy of the reproducibility error, relative to the application to Casey et al. (2021), is somewhat worse. This is likely a consequence of the fact that the quality of a normal approximation to the conditional distribution of the p-value (3.11) is worse than the quality of the normal approximation to the conditional distribution of the mean-squared error estimate. As before, the reproducibility error tends to be most accurate at values of k and ξ where the average number of cross-splits $\hat{g}$ is greater than 500.

FIGURE A.1. Performance in Application to Chakravorty et al. (2024)



Notes: Figure A.1 displays measurements of the performance of Algorithm 1 on the data from Chakravorty et al. (2024). Panel A displays measurements of the reproducibility error, $P\{|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)| \geq \xi \mid D\}$, as ξ and k vary. A solid horizontal line is displayed at the nominal error rate $\beta = 0.05$. Panel B displays measurements of the average number of replications $\hat{g}$ s ξ and k vary. The y-axis is displayed with a log scale, base 10. Solid horizontal lines are placed at each exponential factor of 10. Panel C displays measurements of the 5th and 95th quantiles of the discrepancy $\hat{g}/g^\star - 1$ as k and ξ vary. Further details on the construction of this figure are given in Appendix A.5.

APPENDIX B. AUXILIARY RESULTS AND DISCUSSION

B.1 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 $\mathbf{P}$.

B.1.1 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

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

for all u in $(0, 1)$ and P in $\mathbf{P}$. For example, a test-statistic could be chosen using the data in $D_{\tilde{\mathsf{s}}}$ 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

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

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

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

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

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

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

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

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

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

for any $a > 0$.

Consequently, we have that

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

by Theorem B.2, as required. ■

B.1.2 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

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

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

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

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

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

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

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

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

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

B.2 Determining Suitable Levels of Precision for Cross-Validated Risk Estimation

In this section, we give the details supporting our application of Algorithm 1 to cross-validated risk estimation, using data from Casey et al. (2021). Recall from Section 3.2.2, that we apply the procedure, independently, to each component of the vector

$$(b_{\lambda_1}(\mathsf{r}, D), \ldots, b_{\lambda_{p-1}}(\mathsf{r}, D)) \,, \quad \text{where} \quad b_{\lambda_i}(\mathsf{r}, D) = (a_{\lambda_i}(\mathsf{r}, D) - a_{\lambda_p}(\mathsf{r}, D)) \,. \tag{B.9}$$

Let ξ_i denote the error tolerance used for the ith component of the vector (B.9). We use a simple, data-driven procedure for choosing these tolerances.

In particular, we draw a small number of cross-splits, i.e., $g = 20$ and compute the aggregated difference

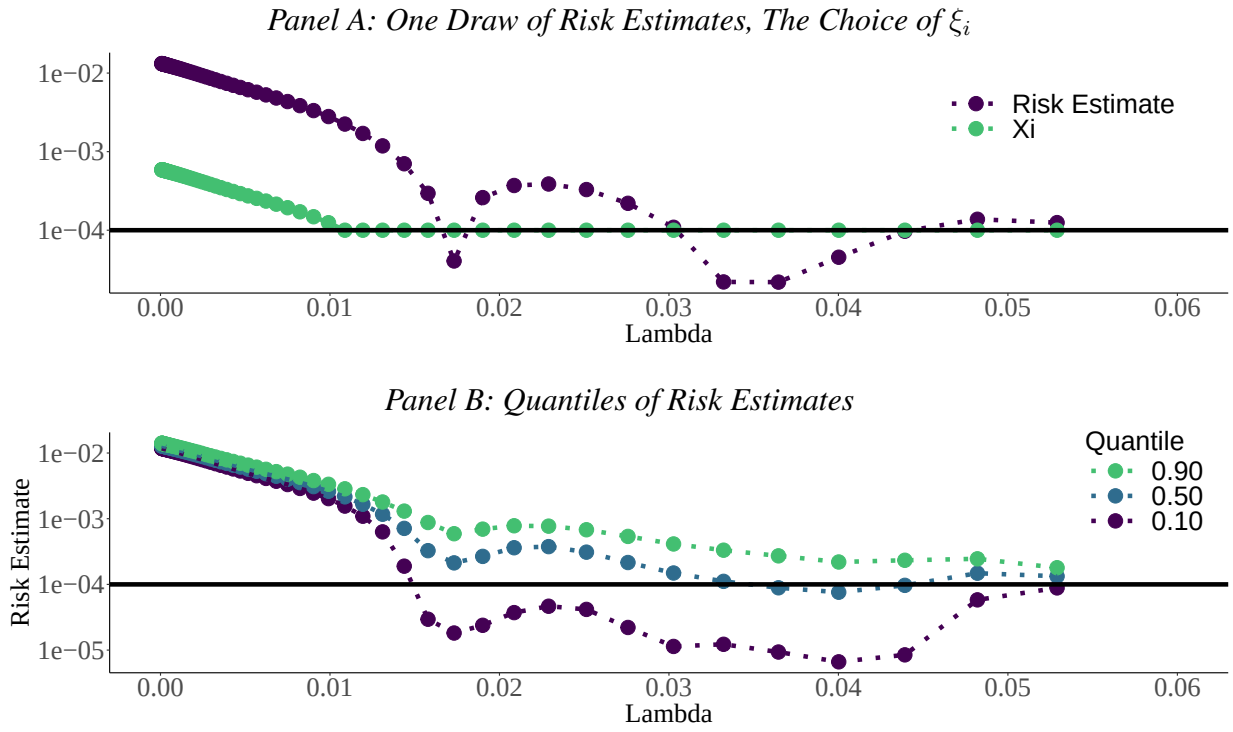
$$b_{\lambda_i}(\mathsf{R}_{g,k}, D) = \frac{1}{g}\sum_{i=1}^{g} b_{\lambda_i}(\mathsf{r}_i, D) \tag{B.10}$$

for each i in $1, \ldots, p-1$. Panel A of Figure B.2 displays the absolute value of these estimates in purple, where the y-axis has been transformed to a logarithmic scale. The absolute value of the differences (B.10) are on the order 10^{-4} at the close-to-optimal values of λ, i.e., values of λ above about 0.01. The differences for smaller values of λ are much larger, and, as indicated by Figure 1, have larger conditional variances. These qualitative features are not particularly sensitive to residual randomness. Panel B of Figure B.2 displays the quantiles of the difference (B.10) across draws of $g = 20$ cross-splits.

These observations suggest the following approach for choosing the tolerances ξ_i. We would like to set $\xi_i = 10^{-4}$ for the large, close-to-optimal values of the regularization parameter, but would be willing to tolerate a larger error tolerance at smaller, sub-optimal values. This is practically important, as achieving the same level of precision for the small values of λ requires aggregation over many more cross-splits (as the conditional variances are much larger). We operationalize this approach in the following way. First, using the sample of $g = 20$ cross-splits that we used to estimate (B.10), we compute the p-values

$$q_{\lambda_i}(\mathsf{R}_{g,k}, D) = 1 - \Phi\left(\frac{b_{\lambda_i}(\mathsf{R}_{g,k}, D)}{\sqrt{\hat{v}_{\lambda_i}(\mathsf{R}_{g,k}, D)}}\right) \tag{B.11}$$

FIGURE B.2. Determining Error Tolerances for Cross-Validated Risk Estimation



Notes: Figure B.2 displays auxiliary measurements relating to the the performance of Algorithm 1 in the application to cross-validated Lasso, implemented in data from Casey et al. (2021). Panel A displays the absolute value of the differences (B.10) computed on a random sample of $g = 20$ cross-splits, in purple, and the realized values of the error tolerances (B.13), in green. Panel B displays the quantiles of the absolute value of the differences (B.10), over replications of random samples of $g = 20$ cross-splits. In both cases, the y-axes is displayed in a logarithmic scale.

---

for each i in $1, \ldots, p-1$, where $\hat{v}_{\lambda_i}(\mathsf{R}_{g,k}, D)$ denotes the sample-variance of the statistics $b_{\lambda_i}(\mathsf{r}_i, D)$ across the sample-splits. Let $\overline{\lambda}$ denote the largest value of the regularization parameter such that $q_{\overline{\lambda}}(\mathsf{R}_{g,k}, D) < 0.2$ and let the constant "scale" solve the equality

$$10^{-4} = \frac{b_{\overline{\lambda}}(\mathsf{R}_{g,k}, D)}{\mathsf{scale}} \,. \tag{B.12}$$

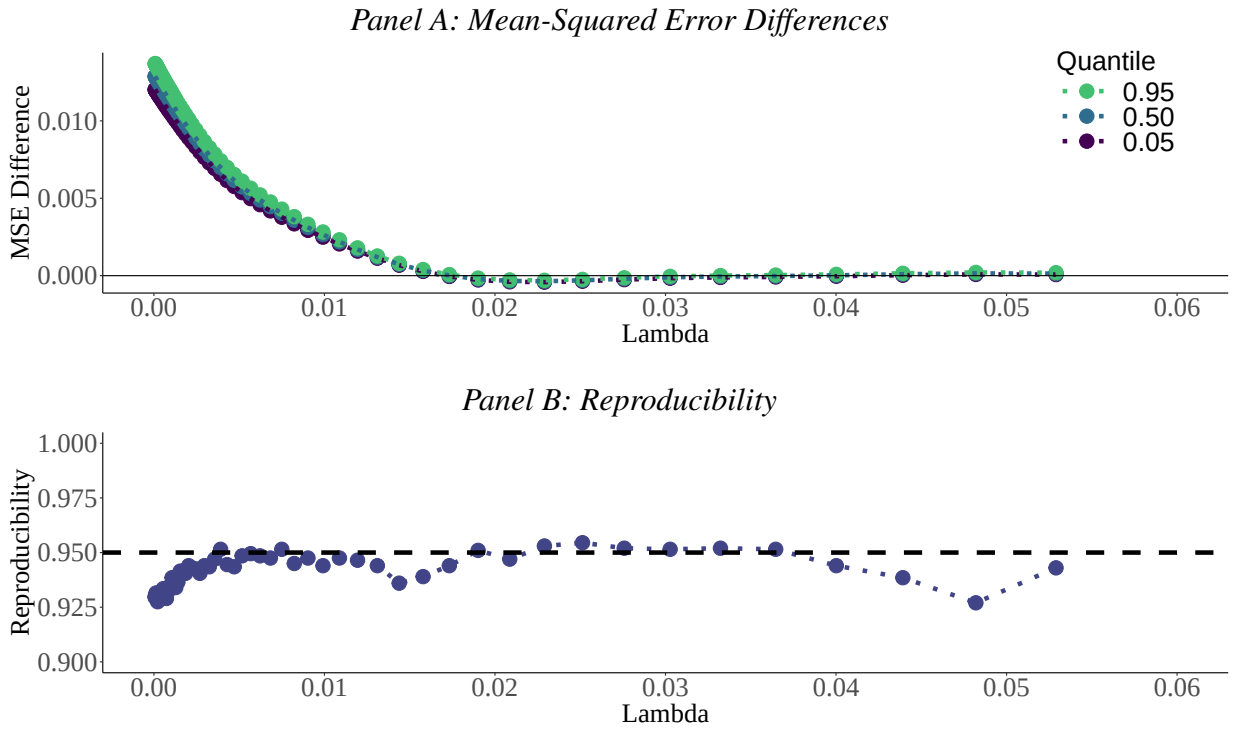
We set the error tolerances to

$$\xi_i = \max\{10^{-4}, b_{\lambda_i}(\mathsf{R}_{g,k}, D) \times \mathsf{scale}\} \,. \tag{B.13}$$

In other words, we set $\xi_i = 10^{-4}$ for all i with $q_{\lambda_i}(\mathsf{R}_{g,k}, D) \geq 0.2$ and increase the error tolerance in proportion to $b_{\lambda_i}(\mathsf{R}_{g,k}, D)$ for all other values of i. The realized values of ξ_i are displayed in Panel A of Figure B.2. These error tolerances are used throughout the main text.

Panel A displays quantiles of the reproducibly aggregated statistic (3.12) across replications of the procedure, without truncation of the y-axis. Panel B of Figure B.2 displays estimates of the marginal reproducibility probability at each value of the regularization parameter, computed

FIGURE B.3. Auxiliary Figures for Application to Cross-Validation



Notes: Figure B.2 displays auxiliary measurements relating to the the performance of Algorithm 1 in the application to cross-validated Lasso, implemented in data from Casey et al. (2021). Panel A is analogous to Panel A of Figure 3, but the y-axis is not truncated. Panel B displays estimates of the marginal reproducibility probability for each value of the regularization parameter, where the nominal reproducibility error is set to $\beta = 0.05$.

---

using $2,000$ replications of the procedure. Over the full range of the regularization parameter, the reproducibility error is very close to the nominal value $\beta = 0.05$.

B.3 Sample Stability

This appendix collects discussion and analysis concerning the (r, q)-sample stability $\sigma^{(r,q)}$ defined in Definition 4.1. First, in Appendix B.3.1, we specialized our consideration to the case that the estimator $\hat{\eta}$ is a regularized empirical risk minimizer. Some closely related arguments are used in proof of Proposition 4 of Austern and Zhou (2020). Second, in Appendix B.3.2 we compare we compare the stable-stability decay condition Assumption 4.3 to the conditions considered by Chen et al. (2022). The proof of the Theorem stated in Appendix B.3.1 is given in Appendix B.3.3

B.3.1 Specialization to Regularized Empirical Risk Minimizers. Assume that the parameter η is an element of some closed convex space $H \subseteq \mathbb{R}^p$. Consider the estimator

$$\Psi\left(D_{\mathsf{s}}, \eta\right) = \frac{1}{b} \sum_{i \in \mathsf{s}} \psi\left(D_i, \eta\right) \tag{B.14}$$

$$\hat{\eta} = \underset{\eta \in H}{\arg\min} \left\{ \frac{1}{n-b} \sum_{i \in \tilde{\mathsf{s}}} \ell\left(D_i, \eta\right) + \lambda_{1,n} \|\eta\|_1 + \lambda_{2,n} \|\eta\|_2 \right\}, \tag{B.15}$$

where $\psi\left(\cdot,\cdot\right)$ are $\ell\left(\cdot,\cdot\right)$ functions and $\lambda_{1,n}, \lambda_{2,n} \geq 0$ are penalty parameters.

Let $\nabla \ell\left(d,\eta\right)$ and $\nabla^2 \ell\left(d,\eta\right)$ denote the gradient and Hessian of the function $\eta \mapsto \ell\left(d,\eta\right)$. Let $\nabla_j \ell\left(\cdot,\cdot\right)$ denote the jth component of $\nabla \ell\left(\cdot,\cdot\right)$. Similarly, we write

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

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

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

We impose the following set of restrictions. The first two restrictions concern the curvature of the loss function.

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

Assumption B.2. *The loss function $\ell\left(\cdot,\cdot\right)$ is strictly convex and twice continuously differentiable in its second argument.*

Remark B.1. Under Assumption B.1, if $\rho > 0$, then 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 the sample stability in the high-dimensional regime is an interesting direction for further research. ■

The second two restrictions are statistical. The first concerns the expected curvature of the moment function $\psi\left(\cdot,\cdot\right)$. The second is a sub-exponential tail bound on the gradient $\nabla_j \ell\left(\cdot,\cdot\right)$.

Assumption B.3. *The function $\psi\left(\cdot,\cdot\right)$ satisfies the inequality*

$$\mathbb{E}\left[\left(\psi(D_i,\eta) - \psi(D_i,\eta')\right)^r\right] \lesssim \mathbb{E}\left[\|\hat{\eta} - \hat{\eta}'\|_2^r\right] \tag{B.16}$$

for each pair $\eta, \eta' \in H$.

Assumption B.4. *The ψ_1-Orlicz norm bound*

$$\|\nabla_j \ell_j(D_i',\eta) - \nabla_j \ell(D_i,\eta)\| \lesssim \kappa^2 \tag{B.17}$$

holds for each pair $\eta \in H$.

Remark B.2. Assumption B.3 is satisfied in many problems of interest. For example, Chen et al. (2022) show that similar conditions hold when $\psi(\cdot,\cdot)$ is given by various moment functions

used for semiparametric causal inference. Assumption B.4 is equivalent to the assumption that $\nabla_j \ell_j(D_i', \eta) - \nabla \ell_j(D_i, \eta)$ is sub-exponential with parameter κ^2. ■

The following theorem gives a bound on the (r, q)-sample stability.

Theorem B.4. *If Assumptions B.1 to B.4, if $\rho + \lambda_{2,n} > 0$, then*

$$\sigma^{(r,q)} \lesssim C_r p \left(\frac{\sqrt{q}}{n-b} \frac{\kappa}{\rho + \lambda_{2,n}} \right)^r + C_r \left(\lambda_{1,n} \frac{p}{\rho + \lambda_{2,n}} \right)^r ,$$

uniformly over all even integers r and positive integers $q \leq n - b$, where C_r is some positive constant that depends only on r.

Remark B.3. A necessary and sufficient condition for $\hat{\eta}$ to be consistent for the population risk minimizer associated with (B.15) is that $\lambda_{1,n} = o\left((n-b)^{-1}\right)$. See e.g., Knight and Fu (2000). Thus, so long as the penalty $\lambda_{1,n}$ is chosen in this regime, the sample stability $\sigma^{(r,q)}$ will be

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

up to terms that depend on the dimension of η, as required. ■

B.3.2 Comparison to Chen et al. (2022). In the main text, we restrict attention to sample stable estimators. In a recent article, Chen et al. (2022) show that, under some regularity conditions, if a condition related to, but partially stronger than, sample stability is satisfied, then sample-splitting is unnecessary for the consistency and asymptotic normality of DML estimators. In this section, we comment on the difference between the notion of sample stability used in this article and the related condition used in Chen et al. (2022).

First, for ease of reference, we recall our definition of sample stability. We restrict attention to the case that $q = 1$ and $r = 2$, as this is what is relevant to make the comparison. Fix a set $\mathsf{s} \subseteq \mathcal{S}_{n,b}$ and let i be an arbitrary element of s. Let D' denote an independent and identical copy of the data D. For each i in $[n]$, let $\tilde{D}^{(i)}$ be constructed by replacing D_i with D_i' in D. Let I be a randomly selected element of $\tilde{\mathsf{s}}$. In this paper, we refer to the quantity

$$\sigma^{(2,1)} = \mathbb{E}\left[\left(\psi(D_i, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) - \psi(D_i, \hat{\eta}(\tilde{D}^{(I)}_{\tilde{\mathsf{s}}})) \right)^2 \right] \tag{B.18}$$

as the $(2,1)$-sample stability. In particular, in the main text, we restrict attention to settings where the bound

$$\sigma^{(2,1)} \lesssim \left(\frac{1}{n-b} \right)^2 \tag{B.19}$$

holds.

Chen et al. (2022) consider settings where the nuisance parameter estimator $\hat{\eta}$ is not constructed with sample-splitting. Likewise, they consider moments of the form

$$\tilde{\sigma}^{(2,1)} = \mathbb{E}\left[\left(\psi(D_i, \hat{\eta}\,(D)) - \psi(D_i, \hat{\eta}(\tilde{D}^{(i)}))\right)^2\right] , \tag{B.20}$$

where i is an arbitrary element of $[n]$. Observe that, in contrast to (B.18), the moment (B.20) is evaluated using the same observation D_i used to perturb the nuisance parameter estimate—and that the nuisance parameter estimate is evaluated using the complete data. Roughly speaking, Chen et al. (2022) show that, under some regularity conditions, DML estimators are consistent and asymptotically normal if the bound

$$\tilde{\sigma}^{(2,1)} \lesssim n^{-1} \tag{B.21}$$

holds for a suitable choice of the function $\psi(\cdot, \cdot)$. Moreover, they show that, under certain conditions on the choice of tuning parameters, nuisance parameter estimators constructed with a bagged 1-nearest-neighbor estimator satisfies this bound.

Although the right-hand-side of the bound (B.21) is larger than the bound (B.19), control of the moment (B.20) can be more difficult to achieve than control of the moment (B.18). To see the substantive differences between the conditions (B.19) and (B.20), we consider an artificial, but illustrative, example adapted from an example given in the introduction to Chernozhukov et al. (2018). Suppose that the observation D_i consists of the real-valued quantities Y_i, W_i, and X_i. Furthermore, suppose that the nuisance parameter $\eta(x)$ denotes the conditional expectation $\mathbb{E}[W_i \mid X_i = x]$ and that

$$\psi(D_i, \eta) = Y_i(W_i - \mathbb{E}[W_i \mid X_i]) = Y_i(W_i - \eta(x)) . \tag{B.22}$$

Suppose that the nuisance parameter estimator takes the form

$$\hat{\eta}(D_{\tilde{\mathsf{s}}})(x) = \sum_{j \in \tilde{\mathsf{s}}} w_j(x)\hat{\eta}_j(x) \tag{B.23}$$

where $\tilde{\mathsf{s}}$ is an arbitrary subset of $[n]$, $\hat{\eta}_j(x)$ is an estimator of the nuisance parameter $\eta(x)$ based on the observation D_j, and the weight $w_j(x)$ again depends on the observation D_j and the set $\tilde{\mathsf{s}}$. We refer the reader to discussion in Section 1 of Chernozhukov et al. (2018) and Section 2 of Chen et al. (2022) for further context on why understanding quantities of the form (B.22) is essential for establishing the consistency of DML estimators.[33]

To study the sample stability (B.18) in this example, i.e., in the sense considered in this article, we are interested in the difference

$$\psi(D_i, \hat{\eta}\,(D_{\tilde{\mathsf{s}}})) - \psi(D_i, \hat{\eta}(\tilde{D}^{(I)}_{\tilde{\mathsf{s}}})) = Y_i(w_I(X_i)\hat{\eta}_I(X_i) - w'_I(X_i)\hat{\eta}'_I(X_i)) \tag{B.24}$$

[33] In particular, establishing the Stochastic Equicontinuity of a given Neyman Orthogonal moment amounts to bounding scaled averages of quantities of the form (B.22).

where $w'_I(X_i)$ and $\hat{\eta}'_I(X_i))$ denote the updated weight and estimator associated with the perturbed data $\tilde{D}^{(I)}_{\tilde{\mathsf{s}}}$. On the other hand, in order to consider the stability (B.20), i.e., in the sense of Chen et al. (2022), we are interested in the difference

$$\psi(D_i, \hat{\eta}\,(D)) - \psi(D_i, \hat{\eta}(\tilde{D}^{(i)})) = Y_i(w_i(X_i)\hat{\eta}_i(X_i) - w'_i(X_i)\hat{\eta}'_i(X_i)) \tag{B.25}$$

where, as before, $w'_i(X_i)$ and $\hat{\eta}'_i(X_i)$ denote the updated weight and estimator associated with the perturbed data $\tilde{D}^{(i)}$.

The quantities (B.24) and (B.25) will behave differently if the nuisance parameter estimator exhibits overfitting. In particular, to take a highly stylized example, suppose that the weights satisfy

$$w_i(X_j) = \begin{cases} |\tilde{\mathsf{s}}|^{-1/3}, & i = j, \\ |\tilde{\mathsf{s}}|^{-1}, & \text{otherwise.} \end{cases} \tag{B.26}$$

That is, the nuisance parameter estimator (B.23) meaningfully up-weights the estimate obtained from the evaluation point X_j, if X_j is part of the training data $\tilde{\mathsf{s}}$, and is otherwise equally balanced. Overfitting to the training data, in a perhaps less stylized form, is a well-known property of some machine learning estimators.

It is easy to see that, in this case, if all other quantities are bounded, we should expect that

$$\begin{aligned} \sigma^{(2,1)} &= \mathbb{E}\left[\left(\psi(D_i, \hat{\eta}\,(D_{\tilde{\mathsf{s}}})) - \psi(D_i, \hat{\eta}(\tilde{D}^{(I)}_{\tilde{\mathsf{s}}}))\right)^2\right] \\ &= \mathbb{E}\left[\left(Y_i(w_I(X_i)\hat{\eta}_I(X_i) - w'_I(X_i)\hat{\eta}'_I(X_i))\right)^2\right] \lesssim \left(\frac{1}{|\tilde{\mathsf{s}}|}\right)^2 \end{aligned} \tag{B.27}$$

whereas

$$\begin{aligned} \tilde{\sigma}^{(2,1)} &= \mathbb{E}\left[\left(\psi(D_i, \hat{\eta}\,(D)) - \psi(D_i, \hat{\eta}(\tilde{D}^{(i)}))\right)^2\right] \\ &= \mathbb{E}\left[\left(Y_i(w_i(X_i)\hat{\eta}_i(X_i) - w'_i(X_i)\hat{\eta}'_i(X_i))\right)^2\right] \lesssim \left(\frac{1}{|\tilde{\mathsf{s}}|}\right)^{2/3}. \end{aligned} \tag{B.28}$$

In particular, we can see that all that is needed for us to expect the moment (B.27) to be approximately of order $O((n-b)^{-2})$, is that the *randomly selected* weights $w_I(X_i)$ *tend* to be close to, or smaller than, $(n-b)^{-1}$. The moment (B.28), by contrast, is highly sensitive to the weight $w_i(X_i)$ placed on the evaluation point X_i.

The main point of this example is that establishing that a nuisance parameter estimator is stable, in the sense of Chen et al. (2022), requires showing that it does not overfit its training data, i.e., that evaluations of points *in-sample* are not overly-determined by a few observations. By contrast, all that is needed to establish that a nuisance parameter estimator is sample stable, in the sense used in this article, is that evaluations of points *out-of-sample* are not overly concentrated on a small

number of training samples. In general, we should expect that out-of-sample balance is more likely to occur in practice than in-sample balance.

B.3.3 Proof of Theorem B.4. By Assumptions B.1 and B.2, the objective function for the estimator (B.15) is strongly convex. Thus, there is a unique solution to (B.15) for any data D. Let $\hat{\eta}$ and $\hat{\eta}'$ denote the solutions to (B.15) 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 (B.15) is given by

$$\overline{\nabla}_{\tilde{\mathsf{s}}} \ell(D, \eta) + \lambda_{2,n} \hat{\eta} + \lambda_{1,n} \hat{z} = 0, \tag{B.29}$$

where $\hat{z} \in \partial \|\hat{\eta}\|_1$ is the subgradient associated with the Lasso penalty. Observe that, in this case, $\hat{z} \in \text{sign}(\hat{\eta})$, where we set $\text{sign}(0) = [-1, 1]$. 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 B.2 (i), we have that

$$\begin{aligned}
&\overline{\nabla}_{\tilde{\mathsf{s}}} \ell(D, \hat{\eta}') + \lambda_{2,n} \hat{\eta}' + \lambda_{1,n} \hat{z} \\
&= \overline{\nabla}_{\tilde{\mathsf{s}}} \ell(D, \hat{\eta}) + \lambda_{2,n} \hat{\eta} + \lambda_{1,n} \hat{z} + \left(\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}} \ell(D, \tilde{\eta}) + \lambda_{2,n} I_d \right) (\hat{\eta} - \hat{\eta}') \\
&= \left(\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}} \ell(D, \tilde{\eta}) + \lambda_{2,n} I_d \right) (\hat{\eta} - \hat{\eta}')
\end{aligned} \tag{B.30}$$

for some vector $\tilde{\eta}$, by a Taylor expansion and the optimality condition (B.29). On the other hand, we have that

$$\begin{aligned}
\overline{\nabla}_{\tilde{\mathsf{s}}} \ell(D, \hat{\eta}') + \lambda_{2,n} \hat{\eta}' + \lambda_{1,n} \hat{z} &= \overline{\nabla}_{\tilde{\mathsf{s}}} \ell(\tilde{D}^{(\mathsf{q})}, \hat{\eta}') + \lambda_{2,n} \hat{\eta}' + \lambda_{1,n} \hat{z}' \\
&\quad + \frac{1}{n-b} \left(\sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}') \right) + \lambda_{1,n} (\hat{z} - \hat{z}') \\
&= \frac{1}{n-b} \left(\sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}') \right) + \lambda_{1,n} (\hat{z} - \hat{z}')
\end{aligned} \tag{B.31}$$

again by the optimality condition (B.29). Thus, we have that

$$\begin{aligned}
&\left(\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}} \ell(D, \tilde{\eta}) + \lambda_{2,n} I_p \right) (\hat{\eta} - \hat{\eta}') \\
&= \frac{1}{n-b} \left(\sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}') \right) + \lambda_{1,n} (\hat{z} - \hat{z}')
\end{aligned} \tag{B.32}$$

by (B.30) and (B.31). Observe that the the matrix $\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}} \ell(D, \tilde{\eta}) + \lambda_{2,n} I_p$ is invertible by Assumption B.1. Thus, we find that

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

$$+ \lambda_{1,n} \left(\overline{\nabla}^{(2)} \ell \left(D, \tilde{\eta} \right) + \lambda_2 I_d \right)^{-1} (\hat{z} - \hat{z}')$$

and that consequently

$$\begin{aligned} \|\hat{\eta} - \hat{\eta}'\|_2 &\lesssim \frac{1}{n-b} \frac{\| \sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}') \|_2}{\rho + \lambda_{2,n}} + \lambda_{1,n} \frac{\|\hat{z} - \hat{z}'\|_2}{\rho + \lambda_{2,n}} \\ &\lesssim \frac{1}{n-b} \frac{\| \sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}') \|_2}{\rho + \lambda_{2,n}} + \lambda_{1,n} \frac{p}{\rho + \lambda_{2,n}} \end{aligned} \tag{B.33}$$

by Assumption B.1.

Now, observe that

$$\begin{aligned} \sigma^{(r,q)} &= \mathbb{E}\left[\left(\psi(D_i, \hat{\eta}) - \psi(D_i, \hat{\eta}') \right)^r \right] \\ &\lesssim \mathbb{E}\left[\|\hat{\eta} - \hat{\eta}'\|_2^r \right] \\ &\lesssim \mathbb{E}\left[\left(\frac{1}{n-b} \frac{\| \sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}') \|_2}{\rho + \lambda_{2,n}} + \lambda_{1,n} \frac{p}{\rho + \lambda_{2,n}} \right)^r \right] \qquad \text{(B.34)} \\ &\lesssim 2^r \left(\frac{1}{n-b} \frac{1}{\rho + \lambda_{2,n}} \right)^r \mathbb{E}\left[\| \sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}') \|_2^r \right] \\ &\quad + 2^r \left(\lambda_{1,n} \frac{p}{\rho + \lambda_{2,n}} \right)^r , \qquad \text{(B.35)} \end{aligned}$$

where the first inequality follows from Assumption B.3, the second inequality follows from (B.33), and the third inequality follows from the Binomial Theorem and Cauchy-Schwarz. By Assumption B.4, we have that

$$\| \sum_{i \in \mathsf{q}} \nabla_j \ell_j(D_i', \hat{\eta}') - \nabla \ell_j(D_i, \hat{\eta}') \|_{\psi_1} \lesssim q \| \nabla_j \ell_j(D_i', \hat{\eta}') - \nabla \ell_j(D_i, \hat{\eta}') \|_{\psi_1} \lesssim q\kappa \, . \tag{B.36}$$

Consequently, we have that

$$\begin{aligned} &\mathbb{E}\left[\| \sum_{i \in \mathsf{q}} \nabla \ell(D_i', \hat{\eta}') - \nabla \ell(D_i, \hat{\eta}') \|_2^r \right] \\ &\lesssim \sum_{j \in [p]} \mathbb{E}\left[\left| \sum_{i \in \mathsf{q}} \nabla_j \ell_j(D_i', \hat{\eta}') - \nabla \ell_j(D_i, \hat{\eta}') \right|^{r/2} \right] \qquad \text{(Jensen)} \\ &\lesssim p((r/2)!)^{r/2} (\kappa q)^{r/2} \qquad \text{(B.37)} \end{aligned}$$

where, for the final inequality, we have used the fact that $\|X\|_{r/2} \lesssim ((r/2)!) \|X\|_{\psi_1}$ (see e.g., Section 2.2 of Van Der Vaart and Wellner (1996)). Putting the pieces together, we find that

$$\sigma^{(r,q)} \lesssim 2^r ((r/2)!)^{r/2} p \left(\frac{q^{1/2}}{n-b} \frac{\kappa^{1/2}}{\rho + \lambda_{2,n}} \right)^r + 2^r \left(\lambda_{1,n} \frac{p}{\rho + \lambda_{2,n}} \right)^r , \tag{B.38}$$

as required. ■

APPENDIX C. PROOFS FOR RESULTS STATED IN SECTION 3

C.1 Proof of Theorem 3.1

For each collection $\mathsf{r} = (\mathsf{s}_j)_{j=1}^k$ in $\mathcal{R}_{n,k,b}$, we write

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

Define the 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. Moreover, we have that

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

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

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

by definition. Consequently, 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

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

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

Define the objects

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

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

Observe that we can decompose

$$\left(a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right) / \sqrt{2v_{g^\star,k}(D)}$$

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

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

First, we have that

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

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

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

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

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

almost surely, where the first inequality follows from Kolmogorov's maximal inequality (see e.g., Proposition 2.3.16 of Dembo (2021)) and the second inequality follows from the fact that the cross-splits r_i are independent. An analogous bound holds for the remaining term. Thus, we have

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

almost surely 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 $\xi \to 0$ almost surely, by the central limit theorem. ■

C.2 Proof of Theorem 3.2

Recall that the cross-split r is drawn independently and uniformly from the collection $\mathcal{R}_{n,k,b}$. The result is based on the bound

$$\begin{aligned}
\mathbb{E}\left[f(a(\mathsf{R}_{g,k}, D))\right] &= \mathbb{E}\left[\mathbb{E}\left[f\left(\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}} a(\mathsf{r}_i, D)\right) \mid \hat{g}\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}} f\left(a(\mathsf{r}_i, D)\right) \mid \hat{g}\right]\right] \qquad \text{(Jensen)} \\
&= \mathbb{E}\left[\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}} f\left(a(\mathsf{r}_i, D)\right)\right] \\
&= \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right)\right] + \mathbb{E}\left[\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}}\left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right)\right]. \qquad \text{(C.8)}
\end{aligned}$$

In particular, it will suffice to show that the second term in (C.8) is weakly negative. To do this, we apply the following optimal stopping theorem.

Theorem C.1 (Theorem 5.7.5, Durrett (2019))**.** *Suppose that X_g is a real-valued supermartingale, that $\mathcal{F}_g$ is the σ-algebra generated by $X_1, \ldots, X_g$, and that there exists some constant B such that*

$$\mathbb{E}\left[|X_{g+1} - X_g| \mid \mathcal{F}_g\right] \leq B \qquad \text{(C.9)}$$

almost surely. If $\hat{g}$ is a stopping time with respect to the filtration $\{\mathcal{F}_g\}_{g=1}^{\infty}$, then $\mathbb{E}\left[X_{\hat{g}}\right] \leq \mathbb{E}\left[X_1\right]$.

Consider the sequence of random variables

$$X_g = \frac{1}{g}\sum_{i=1}^{g}\left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right) \qquad \text{(C.10)}$$

Observe that

$$\mathbb{E}\left[X_{g+1} \mid \mathcal{F}_g, D\right] = \mathbb{E}\left[\frac{1}{g+1}\sum_{i=1}^{g+1}\left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right) \mid \mathcal{F}_g, D\right]$$

$$= \frac{1}{g+1}\mathbb{E}\left[f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right] \mid \mathcal{F}_g, D\right]$$
$$+ \frac{g}{g+1}\frac{1}{g}\sum_{i=1}^{g}\left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right)$$
$$= \frac{g}{g+1}X_g\,. \tag{C.11}$$

Consequently, X_g is a supermartingale. To verify the condition (C.9), observe that, as the collection $\mathcal{R}_{n,k,b}$ is finite, there exists some positive, finite function $B(D)$ such that

$$f(a(\mathsf{r}, D)) \leq B(D) \tag{C.12}$$

almost surely. Thus, we find that

$$\mathbb{E}\left[|X_{g+1} - X_g| \mid \mathcal{F}_g, D\right] = \mathbb{E}\left[\left|\frac{1}{g+1}\sum_{i=1}^{g+1} f\left(a(\mathsf{r}_i, D)\right) - \frac{1}{g}\sum_{i=1}^{g} f\left(a(\mathsf{r}_i, D)\right)\right| \mid \mathcal{F}_g, D\right]$$
$$\leq \frac{1}{g+1}\mathbb{E}\left[|f\left(a(\mathsf{r}, D)\right)|\right] + \frac{1}{g(g+1)}\sum_{i=1}^{g}\mathbb{E}\left[|f\left(a(\mathsf{r}_i, D)\right)|\right]$$
$$\leq \frac{2}{g+1}B(D) \tag{C.13}$$

almost surely. Hence, by Theorem C.1, we have

$$\mathbb{E}\left[\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}}\left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right) \mid D\right]$$
$$\leq \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right] \mid D\right] = 0 \tag{C.14}$$

almost surely, completing the proof. ■

APPENDIX D. GENERAL NON-ASYMPTOTIC RESULTS

In this appendix, we give a general, non-asymptotic analysis of the performance of Algorithm 1. We begin, in Appendix D.1, by studying the concentration and normal approximation of the aggregate statistic $a(\mathsf{R}_{k,g}, D)$. Equipped with these results, in Appendix D.2, we characterize the accuracy of the nominal error rate of Algorithm 1. In Appendix D.3, we show that these general results imply the results stated in Section 4, by restricting attention to sample-stable, cross-split statistics. Proofs are given in Appendices D.4 and D.5.

D.1 Concentration and Normal Approximation

We begin by giving a non-asymptotic characterization of the concentration and normal approximation of the statistic $a(\mathsf{R}_{g,k}, D)$. To state these results, we require a slightly more general definition of sample stability.

Definition D.1 (General Sample Stability)**.** 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. Fix a set $\mathsf{s} \subseteq \mathcal{S}_{n,b}$. Let i be an arbitrary element of s. Let q be a randomly selected subset of $\tilde{\mathsf{s}}$ of cardinality q. We refer to the quantities

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

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

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

as the rth-order full sample stability.

In other words, we define separate stabilities associated with resampling data in the subsets s and $\tilde{\mathsf{s}}$. We refer to these subsets as the "validation" and "training" samples, respectively. The training sample stability $\sigma^{(r,q)}_{\mathsf{train}}$ is identical to the sample stability defined in Section 4.

With this in place, we give a large deviations bound for the statistic $a(\mathsf{R}_{g,k}, D)$ about its conditional mean. 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). In particular, we construct a suitable exchangeable pair with a coupling argument due to Chatterjee (2005) and apply a method for deriving concentration inequalities with exchangeable pairs, also due to Chatterjee (2005, 2007). Some preliminary results that facilitate the application of this approach are given in Appendix D.4. Proofs for results stated in this appendix are then given in Appendix D.5.

Theorem D.1. *Suppose that Assumptions 4.1 and 4.2 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, for each $\varepsilon > 0$, the inequality*

$$P\left\{\left|a(\mathsf{R}_{g,k}, D) - \overline{a}\left(D\right)\right| \leq \sqrt{\frac{2^4(2-\varphi k-\varphi)^2}{\delta}\frac{\Gamma_{k,\varphi,b}\log(\varepsilon^{-1})}{g}} \mid D\right\} \geq 1 - \frac{\varepsilon}{2} \tag{D.2}$$

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

$$\Gamma_{k,\varphi,b} = \left(\frac{1-k\varphi}{1-\varphi}\right)4\sigma^{(2)}_{\mathsf{max}} + \left(\frac{k\varphi-\varphi}{1-\varphi}\right)\sigma^{(2,b-1)}_{\mathsf{train}} \tag{D.3}$$

and $\varphi = b/n$.

Remark D.1. 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 $1/\varphi$. In the case of independent splitting, i.e., $k = 1$, 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 $\sigma^{(2,b-1)}_{\mathsf{train}}$. In a previous draft of this paper, we give estimates of $\sigma^{(2)}_{\mathsf{valid}}$ and $\sigma^{(2,b-1)}_{\mathsf{train}}$ in applications to treatment effect estimation and cross-validated risk estimation. We show that the training stability is dramatically smaller than the validation stability and decreases rapidly as b decreases.[34] To see this, observe that we should next expect the validation sample-stability $\sigma^{(r)}_{\mathsf{valid}}$ to change with k. That is, residual randomness is smaller for full cross-splitting, i.e., $k = n/b$, than for independent splitting. ■

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 D.1. Bounds of this form are known as Burkholder-Davis-Gundy inequalities (Burkholder, 1973).

Theorem D.2. *Suppose that Assumptions 4.1 and 4.2 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, for each $\delta >$, the inequality*

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

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

$$\Gamma^{(r)}_{k,\varphi,b} = \left(\frac{1-k\varphi}{1-\varphi}\right) 2^{2r}\sigma^{(2r,1)}_{\mathsf{max}} + \left(\frac{k\varphi-\varphi}{1-\varphi}\right)\sigma^{(2r,b-1)}_{\mathsf{valid}} \tag{D.5}$$

and $\varphi = b/n$.

Remark D.2. By setting $r = 1$, the inequality (D.4) gives a variance bound. In this case, the right-hand side of the inequality (D.4) is equal to negative one times the inverse of the right-hand side of the inequality (D.2). In this sense, the concentration inequality Theorem D.1 can be thought of as an exponential Efron-Stein inequality adapted to the dependence inherent in our problem (Boucheron et al., 2003). For values of r greater than 1, the unconditional quantity $\Gamma^{(r)}_{k,\varphi,b}$ again interpolates between the $2r$th order full sample stability $\sigma^{(2r,1)}_{\mathsf{max}}$ and the $(2r, b-1)$th order training sample stability $\sigma^{(2r,b-1)}_{\mathsf{valid}}$ as k increases from 1 to $1/\varphi$. ■

Remark D.3. An analogous, unconditional, result will follow from the same argument. In particular, under the same conditions, we have that

$$\mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)\right)^{2r}\right] \leq (2r-1)^r \left(\frac{2^4(2-\varphi k-\varphi)^2}{g}\right)^r \Gamma^{(r)}_{k,\varphi,b}\,. \tag{D.6}$$

We opt to state the conditional version of the result thought, as, arguably, the conditional behavior of the statistic is of primary interest. ■

[34] See Figure 3 of the preprint posted at https://arxiv.org/abs/2311.14204v2.

In turn, we bound the distance between the conditional distribution of $a\left(\mathsf{R}_{g,k}, D\right)$ and the standard normal distribution. The result follows by combining the bounds derived in Theorem D.2 with a standard Berry-Esseen inequality (Shevtsova, 2011).

Theorem D.3. *Suppose that Assumptions 4.1 and 4.2 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 inequality*

$$\sup_{z\in\mathbb{R}}\left|P\left\{\frac{a\left(\mathsf{R}_{g,k}, D\right)-\overline{a}\left(D\right)}{\sqrt{v_{g,k}\left(D\right)}}\leq z \mid D\right\}-\Phi(z)\right| \lesssim \frac{2^6 3^2}{\delta}\frac{(2-\varphi k-\varphi)^3}{g^{1/2}}\left(\frac{(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{v_{1,k}(D)}\right)^{3/2} \tag{D.7}$$

is satisfied with probability greater than $1-\delta$ as D varies, where $\Phi(\cdot)$ is the standard normal c.d.f.

Remark D.4. 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.6, we show that this central limit theorem implies a bound that does not reduce as either g or k increase. ■

D.2 Reproducibility

We now characterize accuracy of the nominal reproducibility error of Algorithm 1.

Theorem D.4. *Suppose that the conditional variance $\mathrm{Var}(a(\mathsf{r}, D) \mid D)$ is strictly positive, almost surely, where r denotes a random collection drawn uniformly from $\mathcal{R}_{n,k,b}$. Suppose that the collections $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$ are independently obtained using Algorithm 1. If Assumptions 4.1 and 4.2 hold, the data D are independent and identically distributed, and the eighth-order split stability $\zeta^{(8)}$ is finite, then, for all sufficiently small ξ, the inequality*

$$\begin{aligned} &P\left\{|a(\mathsf{R}_{\hat{g},k}, D)-a(\mathsf{R}'_{\hat{g}',k}, D)|\leq \xi \mid D\right\}-(1-\beta)\,| \\ &\quad\lesssim \left(\frac{\xi}{z_{1-\beta/2}}\frac{(2-\varphi k-\varphi)^4\Gamma^{(1)}_{k,\varphi,b}(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\delta^{3/2}(v_{1,k}\left(D\right))^{5/2}}\right)^{1/2} \\ &\qquad\cdot\log^{3/4}\left(\frac{z_{1-\beta/2}}{\xi}\frac{\delta(v_{1,k}(D))^2}{(2-\varphi k-\varphi)^3(\Gamma^{(2)}_{k,\varphi,b})^{3/4}}\right), \end{aligned} \tag{D.8}$$

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

Remark D.5. Recall the definition of the oracle stopping time $g^\star$ given in (3.7). The bound (D.17) is obtained by decomposing the reproducibility error into two terms. The first term involves the error in a normal approximation to the difference $a(\mathsf{R}_{g^\star,k}, D)-a(\mathsf{R}'_{g^\star,k}, D)$. As $g^\star$ is deterministic

conditional on D, a bound on this term follows from an argument very similar to the proof of Theorem D.3. The second term involves the differences $a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}_{g^\star,k}, D)$ and $a(\mathsf{R}'_{\hat{g}',k}, D) - a(\mathsf{R}'_{g^\star,k}, D)$. 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 D.1 for the estimator $\hat{v}_{g,k}(D)$ with a maximal inequality due to Steiger (1970). ■

D.3 Specialization to Cross-Split, Sample-Stable Statistics

In this section, we specialize the results stated in Appendices D.1 and D.2 to cross-fit, sample-stable statistics. That is, we focus attention to the setting where $n = k \cdot b$ and impose Assumption 4.3. By doing this, we prove each of the results stated in Section 4.

We begin by stating the following corollary of Theorem D.1, which is referenced in Footnote 22.

Corollary D.1. *Suppose that Assumptions 4.1 to 4.3 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, for each $\varepsilon > 0$, the inequality*

$$P\left\{ \left| a(\mathsf{R}_{g,k}, D) - \overline{a}\,(D) \right| \leq \sqrt{\frac{b-1}{n^2}\frac{1}{g}\frac{\log(\varepsilon^{-1})}{\delta}} \mid D \right\} \geq 1 - \varepsilon \tag{D.9}$$

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

Proof. Corollary D.1 follows immediately from Theorem D.1 by restricting attention to the case that $n = k \cdot b$ and applying Assumption 4.3. In particular, observe that if $n = k \cdot b$, then $\Gamma_{k,\varphi,b} = \sigma^{(2,b-1)}_{\mathsf{train}}$ and

$$(2 - \varphi k - \varphi)^2 = (1 - \varphi)^2 = \left(\frac{n-b}{b}\right)^2 . \tag{D.10}$$

Applying Assumption 4.3 then gives

$$\Gamma_{k,\varphi,b} = \sigma^{(2,b-1)}_{\mathsf{train}} \lesssim \left(\frac{\sqrt{b-1}}{n-b}\right)^2 . \tag{D.11}$$

Plugging (D.10) and (D.11) into Theorem D.1 gives the desired bound. ■

Next, we show that Theorem D.2 implies the following generalization of Theorem 4.1 to higher-order moments. Theorem 4.1 is a strictly weaker result, as it only holds for the second moment.

Corollary D.2. *Suppose that Assumptions 4.1 to 4.3 hold, the data D are independently and identically distributed, and $n = k \cdot b$. If the $4r$th-order split stability $\zeta^{(4)}$ is finite, then, for each $\delta > 0$, the inequality*

$$\mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right) \right)^{2r} \mid D \right] \lesssim \left(\frac{b-1}{n^2}\frac{1}{g} \right)^{r/2} \tag{D.12}$$

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

Proof. If $n = k \cdot b$, then $\Gamma^{(r)}_{k,\varphi,b} = \sigma^{(2r,b-1)}_{\mathsf{valid}}$. Applying Assumption 4.3 then gives

$$\Gamma^{(r)}_{k,\varphi,b} \lesssim \left(\frac{\sqrt{b-1}}{n-b}\right)^{2r} . \tag{D.13}$$

Plugging (D.10) and (D.13) into (D.4) gives

$$\mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)\right)^{2r}\right] \lesssim \left(\frac{1}{nkg}\right)^{r} , \tag{D.14}$$

where we omit constants the depend only on r. ■

In turn, we specialize the non-asymptotic central limit theorem Theorem D.3 cross-fit, sample-stable statistics. This result suggests that, although the aggregate statistic $a(\mathsf{R}_{g,k}, D)$ is more concentrated at larger values of k, the quality of a normal approximation to its conditional distribution does not necessarily improve. This result is referenced in Footnote 26.

Corollary D.3. *Suppose that Assumptions 4.1 to 4.3 hold, the data D are independently and identically distributed, $n = k \cdot b$, and the conditional variance $v_{1,k}(D) = \mathrm{Var}(a(\mathsf{r}, D) \mid D)$ is strictly positive almost surely, where r denotes a random collection drawn uniformly from $\mathcal{R}_{n,k,b}$. If the 8th-order split stability $\zeta^{(8)}$ is finite, then, for all g, the inequality*

$$\sup_{z\in\mathbb{R}} \left| P\left\{ \frac{a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)}{\sqrt{v_{g,k}\left(D\right)}} \leq z \mid D \right\} - \Phi(z) \right| \lesssim \frac{1}{\delta}\left(\frac{1}{n}\frac{1}{k}\frac{b-1}{v_{1,k}(D)}\right)^{3/2} \sqrt{\frac{1}{g}} \tag{D.15}$$

is satisfied with probability greater than $1-\delta$ as D varies, where $\Phi(\cdot)$ is the standard normal c.d.f.

Proof. As before, Corollary D.3 follows by restricting attention to the case that $n = k \cdot b$ and imposing Assumption 4.3. In particular, Corollary D.3 is obtained by plugging the bounds (D.10) and (D.13) into (D.7). ■

Remark D.6. In this case, the quantities g and k enter differently. In particular, observe that, by applying the variance bound (4.9) for the case $g = 1$, to lower bound the rate (D.15), we get

$$\left(\frac{1}{n}\frac{1}{k}\frac{1}{v_{1,k}(D)}\right)^{3/2} \sqrt{\frac{1}{g}} \gtrsim \sqrt{\frac{1}{g}} . \tag{D.16}$$

Written differently, Corollary D.3 suggests that the quality of a normal approximation to the conditional distribution of the aggregate statistic $a\left(\mathsf{R}_{g,k}, D\right)$ does not increase with k—it only depends on g. ■

Finally, we show that Theorem 4.2 is a corollary of Theorem D.4 by restricting attention to the case that $n = k \cdot b$ and applying Assumption 4.3. We restate Theorem 4.2 below as a corollary for ease of reference.

Corollary D.4. *Suppose that the collections $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$ are independently obtained using Algorithm 1. If Assumptions 4.1 to 4.3 hold, the conditional variance $v_{1,k}(D) = \mathrm{Var}(a(\mathsf{r}, D) \mid D)$ is strictly positive, almost surely, the data D are independent and identically distributed, and the eighth-order split stability $\zeta^{(8)}$ is finite, then for all sufficiently small ξ, the inequality*

$$\left| P\left\{ \left| a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D) \right| \geq \xi \mid D \right\} - \beta \right| \lesssim \frac{1}{\delta^{3/4}} \frac{1}{kn} \left(\frac{1}{v_{1,k}(D)} \right)^{5/4} \left(\frac{\xi}{z_{1-\beta/2}} \right)^{1/2}, \tag{D.17}$$

holds with probability greater than $1 - \delta$ as D varies, where in writing (D.17), we have omitted a multiplicative term that converges to zero logarithmically as ξ decreases to zero.

Proof. Observe that if $n = k \cdot b$, then the right-hand-side of the inequality (D.17) can be written

$$\left(\frac{\xi}{z_{1-\beta/2}} \frac{\sigma_{\mathsf{train}}^{(2,b-1)} (\sigma_{\mathsf{train}}^{(4,b-1)})^{1/2}}{\delta^{3/2} (v_{1,k}(D))^{5/2}} \left(\frac{n-b}{n} \right)^4 \right)^{1/2} \cdot \log^{3/4} \left(\frac{z_{1-\beta/2}}{\xi} \left(\frac{n}{n-b} \right)^3 \frac{\delta (v_{1,k}(D))^2}{(\sigma_{\mathsf{train}}^{(4,b-1)})^{3/4}} \right). \tag{D.18}$$

By applying Assumption 4.3, the non-logarithmic factor reduces to

$$\left(\frac{\xi}{z_{1-\beta/2}} \frac{1}{\delta^{3/2} (v_{1,k}(D))^{5/2}} \left(\frac{\sqrt{b-1}}{n} \right)^4 \right)^{1/2} \leq \left(\frac{1}{\delta^{3/2}} \frac{\xi}{z_{1-\beta/2}} \left(\frac{1}{kn} \right)^2 \left(\frac{1}{v_{1,k}(D)} \right)^{5/2} \right)^{1/2} \tag{D.19}$$

as required. ∎

D.4 Constructing a Stein Representer

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

$$f(X) = \mathbb{E}[F(X, X') \mid X] \tag{D.20}$$

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

In many applications, the Stein representer (D.20) 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.

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 such a 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

$$\sum_{m=0}^{\infty} \left|\mathbb{E}\left[f(X_m) - f(X'_m) \mid X_0 = x, X'_0 = y\right]\right| \leq C , \tag{D.21}$$

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 $f\left(X\right)$. 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 (D.21). 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 $\mathsf{r}_k\left(\pi\right) = \left(\mathsf{s}_1\left(\pi\right), \ldots, \mathsf{s}_k\left(\pi\right)\right)$, through

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

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 $\mathsf{R}_{g,k}\left(\boldsymbol{\pi}\right) = \left(\mathsf{r}_k\left(\pi_i\right)\right)_{i=1}^g$ 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 (D.21). For any permutation $\pi \in \mathcal{P}_n$ and indices $i, j \in [n]$, define the updated permutation

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

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

$$\boldsymbol{\pi}_m = (\pi_{m-1,-L_m}, \hat{\pi}_{m-1,L_m}(I_m, J_m)) \quad \text{and} \quad \boldsymbol{\pi}'_m = (\pi'_{m-1,-L_m}, \hat{\pi}'_{m-1,L_m}(I_m, J_m)) \tag{D.22}$$

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 (D.21). This argument is formalized in the proof of the following Lemma.

Lemma D.1. *The function*

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

is finite, antisymmetric, and satisfies the equality

$$\mathbb{E}\left[A(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D) \mid \boldsymbol{\pi}, D\right] = a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}(D) \tag{D.23}$$

almost surely.

D.5 Proofs

D.5.1 Proofs for Non-Sequential Results. To obtain the concentration inequality Theorem D.1 and Burkholder-Davis-Gundy inequality Theorem D.2, 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 D.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

$$U_f^{(r)}(X) = \frac{1}{2}\mathbb{E}\left[(f(X) - f(X'))^{2r} \mid X\right] \quad \textit{and} \quad U_F^{(r)}(X) = \frac{1}{2}\mathbb{E}\left[F(X, X')^{2r} \mid X\right]. \tag{D.24}$$

If there exist nonnegative constants u and s such that

$$U_f^{(1)}(X) \leq s^{-1}u \quad \textit{and} \quad U_F^{(1)}(X) \leq su, \tag{D.25}$$

then the concentration inequality

$$P\{|f(X)| \geq \delta\} \leq 2\exp\left(-t^2/2u\right) \tag{D.26}$$

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

$$\mathbb{E}\left[f(X)^{2r}\right] \leq (2r-1)^r \left(s\mathbb{E}\left[U_F^{(r)}(X)\right] + s^{-1}\mathbb{E}\left[U_f^{(r)}(X)\right]\right) \tag{D.27}$$

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 D.2, we are required to develop bounds for the objects

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

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 D.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*

$$\left(\sum_{m=0}^{\infty} \mathbb{E}\left[f(Z_m) - f(Z'_m) \mid Z_0 = Z, Z'_0 = Z'\right]\right)^r \leq W \tag{D.28}$$

almost surely. If the inequality

$$\mathbb{E}\left[\mathbb{E}\left[f(Z_m) - f(Z'_m) \mid Z_0 = Z, Z'_0 = Z'\right]^r \mid \boldsymbol{\pi}\right] \leq h_m^r \tag{D.29}$$

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

$$\frac{1}{2}\mathbb{E}\left[(f(Z) - f(Z'))^r \mid Z\right] \leq \frac{h_0^r}{\delta} \quad \textit{and} \tag{D.30}$$

$$\frac{1}{2}\mathbb{E}\left[\left(\sum_{m=0}^{\infty} \mathbb{E}\left[f(Z_m) - f(Z'_m) \mid Z_0 = Z, Z'_0 = Z'\right]\right)^r \mid Z\right] \leq \frac{1}{\delta}\left(\sum_{m=0}^{\infty} h_m\right)^r \tag{D.31}$$

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

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

$$\begin{aligned}&\left(\sum_{m=0}^{\infty}\mathbb{E}\left[a\left(Z_m\right)-a\left(Z_m'\right)\mid Z_0=Z,Z_0'=Z'\right]\right)^2\\&\lesssim g^2n^4\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T\left(\mathsf{s},U,D\right)-T\left(\mathsf{s}',U,D\right)\right)^2\end{aligned}\tag{D.32}$$

follows from Lemma E.1, stated in Appendix E.1. Moreover, the right hand side of (D.32) is square integrable, as the fourth order split-stability $\zeta^{(4)}$ is finite by assumption. Deterministic bounds of the form (D.29) are obtained through the following Lemma.

Lemma D.4. *Under Assumptions 4.1 and 4.2, 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 $\Gamma_{k,\varphi,b}^{(r)}$ *is defined in the statement of Theorem D.2.*

Combining Lemmas D.3 and D.4, we have that

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

and

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

with probability $1-\delta$. Corollary D.1 is obtained by setting $r=1$ and applying (D.26) of Lemma D.2 with $s=gn^2/2$ and

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

Similarly, Theorem D.2 is obtained by applying (D.27) of Lemma D.2 with $s=(gn^2/2)^r$.

The normal approximation error bound Theorem D.3 follows from Theorem D.2. To see this, consider the centered statistic

$$a\left(Z\right)-\overline{a}\left(D\right)=\frac{1}{g}\sum_{\ell=1}^{g}\overline{a}\left(\mathsf{r}_\ell,D\right),\quad\text{where}\quad\overline{a}\left(\mathsf{r}_\ell,D\right)=\frac{1}{k}\sum_{i=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}\left(D\right)\right)\tag{D.33}$$

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

$$\mathbb{E}\left[\left|\overline{a}\left(\mathsf{r}_\ell,D\right)\right|^3\right]\leq 3^2 2^6\left(2-\varphi k-\varphi\right)^3\left(\Gamma_{k,\varphi,b}^{(2)}\right)^{3/4}\tag{D.34}$$

by Hölder's inequality and Theorem D.2. Hence, we find that

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

holds with probability greater than $1-\delta$, by combining (D.34) with the Markov inequality. The proof then follows by the Berry-Esseen inequality. See e.g., Corollary 1 of Shevtsova (2011). ■

D.5.2 Proofs for Sequential Results. The first step for verifying Theorem D.4 is deriving a large deviation bound for the variance estimator $\hat{v}\left(\mathsf{R}_{g,k}, D\right)$ defined in (3.4). This is obtained in the following Lemma, which follows from an argument very similar to the proof of Theorem D.1.

Lemma D.5. *Suppose that Assumptions 4.1 and 4.2 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*

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

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

We focus our analysis on the error

$$|P\left\{a(\mathsf{R}_{\hat{g},k}, D)-a(\mathsf{R}_{\hat{g}',k}', D) \leq \xi \mid D\right\}-(1-\beta/2)|. \tag{D.35}$$

An analogous argument will yield the same bound for the lower tail. We begin by bounding (D.35) with quantities that will be easier to handle in isolation. Define the objects

$$U\left(\mathsf{R}_{g,k}, D\right)=\frac{1}{g^{\star}}\left(\sum_{i=1}^{g} \overline{a}\left(\mathsf{r}_{i,k}, D\right)-\sum_{i=1}^{g^{\star}} \overline{a}\left(\mathsf{r}_{i,k}, D\right)\right) \quad \text{and} \tag{D.36}$$

$$Q\left(\mathsf{R}_{g,k}, D\right)=\left(1-\frac{g}{g^{\star}}\right)\left(a(\mathsf{R}_{g,k}, D)-\overline{a}\left(D\right)\right) . \tag{D.37}$$

The following Lemma bounds the error (D.35) 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 (D.36) and (D.37).

Lemma D.6. *Define the events*

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

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

The quantity (D.35) *is bounded above by*

$$\sup_{z\in\mathbb{R}}\left|P\left\{\frac{a(\mathsf{R}_{g^\star,k},D)-a(\mathsf{R}'_{g^\star,k},D)}{\sqrt{2v_{g^\star,k}(D)}}\le z\mid D\right\}-\Phi(z)\right| + 2\lambda+\left(1-P\left\{\mathcal{U}_{k,\lambda}(D)\cap\mathcal{Q}_{k,\lambda}(D)\mid D\right\}\right), \tag{D.38}$$

where $\Phi(\cdot)$ *is the standard normal c.d.f.*

Thus, it remains to give suitable bounds for the objects in (D.38). We state these bounds in terms of the quantities

$$\rho_{k,\varphi,b}(\xi,\beta\mid D)=\frac{\xi}{z_{1-\beta/2}}\frac{(2-\varphi k-\varphi)^3(\Gamma^{(2)}_{k,\varphi,b})^{3/4}}{\delta\left(v_{1,k}(D)\right)^2} \quad \text{and} \tag{D.39}$$

$$\lambda_{k,\varphi,b}(\xi,\beta\mid D)=\left(\frac{\xi}{z_{1-\beta/2}}\frac{(2-\varphi k-\varphi)^4\Gamma^{(1)}_{k,\varphi,b}(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\delta^{3/2}(v_{1,k}(D))^{5/2}}\right)^{1/2}, \tag{D.40}$$

respectively. Each part of following Lemma follows from an application of Lemma D.6.

Lemma D.7. *Suppose that Assumptions 4.1 and 4.2 hold, that the data* D *are independent and identically distributed, that the conditional variance* $v_{1,k}(D)=\mathrm{Var}(a(\mathsf{r},D)\mid D)$ *is strictly positive almost surely, and that the eighth-order split stability* $\zeta^{(8)}$ *is finite.*

(i) *The inequality*

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

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

(ii) *The conditional concentration inequality*

$$P\left\{\frac{|U(\mathsf{R}_{\hat{g},k},D)-U(\mathsf{R}'_{\hat{g}',k},D)|}{\sqrt{2v_{g^\star,k}(D)}}\ge\lambda_{k,\varphi,b}(\xi,\beta\mid D)\log^{3/4}\left(\rho_{k,\varphi,b}(\xi,\beta\mid D)^{-1}\right)\mid D\right\} \lesssim\rho_{k,\varphi,b}(\xi,\beta\mid D)$$

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

(iii) *For all sufficiently small* ξ*, the conditional concentration inequality*

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

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

Now, observe that $\rho_{k,\varphi,b}(\xi, \beta \mid D)$ is smaller than $\lambda_{k,\varphi,b}(\xi, \beta \mid D)$ for all sufficiently small ξ, almost surely. Consequently, by combining Lemma D.6 and Lemma D.7, we find that

$$\begin{aligned} &|P\left\{|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)| \leq \xi \mid D\right\} - (1-\beta)| \\ &\qquad \lesssim \left(\frac{\xi}{z_{1-\beta/2}} \frac{(2-\varphi k-\varphi)^4 \Gamma^{(1)}_{k,\varphi,b} (\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\delta^{3/2} (v_{1,k}(D))^{5/2}}\right)^{1/2} \\ &\qquad\qquad \cdot \log^{3/4}\left(\frac{z_{1-\beta/2}}{\xi} \frac{\delta (v_{1,k}(D))^2}{(2-\varphi k-\varphi)^3 (\Gamma^{(2)}_{k,\varphi,b})^{3/4}}\right) \end{aligned} \tag{D.41}$$

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

D.6 Comparison with Zhang (2022)

We state a Berry-Esseen bound for $a(\mathsf{R}_{g,k}, D)$ through an application of a result due to Zhang (2022). In contrast to the bound stated in Theorem D.3, 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 4.1 and 4.2 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*

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

is satisfied.

D.6.1 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*

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

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

$$\overline{f}(X) = f(X) / \sqrt{\mathrm{Var}(f(X))}. \tag{D.42}$$

$$G(X) = \frac{1}{2}\mathbb{E}[(f(X) - f(X'))F(X, X') \mid X] \quad \text{and} \tag{D.43}$$

$$\overline{G}(X) = \frac{1}{2}\mathbb{E}\left[\left(f(X) - f(X')\right)\left|F(X,X')\right| \mid X\right]. \tag{D.44}$$

Let W denote a standard normal random variable. The bound

$$d_K\left(\overline{f}(X), W\right) \leq \frac{\mathbb{E}\left[\left|G(X) - \mathbb{E}\left[G(X)\right]\right|\right] + \mathbb{E}\left[\left|\overline{G}(X)\right|\right]}{\mathrm{Var}\left(f(X)\right)} \tag{D.45}$$

is satisfied.

To this end, define the objects

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

It will suffice to bound the quantity

$$\frac{\mathbb{E}\left[\left|B(Z) - \mathbb{E}\left[B(Z)\right]\right|\right] + \mathbb{E}\left[\left|\overline{B}(Z)\right|\right]}{\mathrm{Var}\left(a(Z) - \overline{a}(Z)\right)}. \tag{D.46}$$

Observe that

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

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

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

it will suffice to bound

$$2\mathbb{E}\left[\left(a(Z) - a(Z')\right)^2 A(Z,Z')^2\right]. \tag{D.47}$$

By Young's inequality, we have that

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

where $U_a^{(2)}(Z)$ and $U_A^{(2)}(Z)$ are defined in Appendix D.5.1.

Observe that the bound

$$\left(\sum_{m=0}^{\infty}\mathbb{E}\left[a(Z_m) - a(Z_m') \mid Z_0 = Z, Z_0' = Z'\right]\right)^4$$
$$\leq \left(2gn^2 \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}}\left(T(\mathsf{s},D) - T(\mathsf{s}',D)\right)\right)^4 \tag{D.48}$$

holds by Lemma E.1 and that the right-hand side of (D.48) is square-integrable as the eighth-order split stability $\zeta^{(8)}$ is finite. Thus, by combining Lemma E.3 and Lemma D.4, we have that

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

and

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

Hence, by taking $s = \left(gn^2/2\right)^2$, we find that (D.47) is bounded above by

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

Now, we have that

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

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

$$\mathbb{E}\left[v_{g,k}\left(D\right)\right] = \frac{\mathbb{E}[\phi_{n,b}(D)] + (k-1)\,\mathbb{E}[\gamma_{n,b}(D)]}{kg}. \tag{D.49}$$

Hence, we have that (D.46) 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. ■

Appendix E. Proofs for Lemmas Stated in Appendix D

E.1 Proof of Lemma D.1

We use the following Lemma in several places.

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

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

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 $A(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D)$ follows from Lemma E.1. Define the operator

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

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

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

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

$$\left(K^{m+1}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right)\right)\right)_{m \geq 0} \tag{E.2}$$

also converges almost everywhere. Lemma E.1 also implies that the limit of (E.2) 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

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

that depends only on D. Observe that

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

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

$$= 0, \qquad \text{(Exchangeability)}$$

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

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

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

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

completing the proof. ■

E.2 Proof of Lemma D.2

First, observe that

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

where the inequality follows from Jensen's inequality and the definition of the Stein representer F. By (D.30), we have that

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

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

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

$$\mathbb{E}\left[h\left(X\right) f\left(X\right)\right] = \frac{1}{2}\mathbb{E}\left[\left(h\left(X\right) - h\left(X'\right)\right) F\left(X, X'\right)\right]. \tag{E.3}$$

Let

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

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

$$\begin{aligned} m'(\theta) &= \mathbb{E}\left[\exp\left(\theta f(X)\right) f(X)\right]. \\ &= \frac{1}{2}\mathbb{E}\left[\left(\exp\left(\theta f(X)\right) - \exp\left(\theta f(X')\right)\right) F(X, X')\right], \end{aligned} \tag{E.4}$$

where the second inequality follows from (E.3). To bound $m'(\theta)$ we apply the following exponential mean-value inequality, stated in a more general form in Paulin et al. (2016).

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

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

In particular, by (E.4) and Lemma E.2, we obtain the bound

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

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

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

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

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

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 (D.26). To prove (D.27) we apply the following result stated in Chatterjee (2007).

Theorem E.1 (Theorem 1.5, (iii), Chatterjee, 2007)**.** *Reintroduce the notation and assumptions from the statement of Theorem D.2. Define*

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

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

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

then completes the proof. ■

E.3 Proof of Lemma D.3

We apply the following Lemma.

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

$$\mathbb{E}\left[X_m^{2^c}\right] \leq h_m^{2^c} \tag{E.5}$$

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

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

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

By the (D.28), 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, (D.29) guarantees that the conditions of Lemma E.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 E.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 $1 - \delta$. ■

E.4 Proof of Lemma D.4

Recall the definition of the collections $(\boldsymbol{\pi}_m, \boldsymbol{\pi}'_m)_{m\geq 0}$ given in (D.22). 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

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

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

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

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

By Assumption 4.1, the event $\mathcal{H}_m = \mathcal{E}_m \cap \mathcal{F}_m$ is a necessary condition for $a(Z_m) - a(Z'_m) \neq 0$. Thus, we can compute

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

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

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

almost surely. Consequently, we find that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[a\left(Z_m\right)-a\left(Z_m'\right) \mid Z, Z'\right]^{2r} \mid \boldsymbol{\pi}\right] \\
&=\mathbb{E}\left[\left(P\left\{\mathcal{H}_m \mid Z, Z'\right\} \mathbb{E}\left[a\left(Z_m\right)-a\left(Z_m'\right) \mid Z, Z', \mathcal{H}_m\right]\right)^{2r} \mid \boldsymbol{\pi}\right] \\
&\leq\left(P\left\{\mathcal{H}_m \mid \mathcal{E}_0\right\}\right)^{2r} \mathbb{E}\left[\mathbb{E}\left[\left(a\left(Z_m\right)-a\left(Z_m'\right)\right)^{2r} \mid Z, Z', \mathcal{H}_m\right] \mid \boldsymbol{\pi}\right] \qquad \text{(Jensen)} \\
&=\left(1-\frac{2}{gn^2}\right)^{2mr}\left(\frac{kb\left(2n-kb-b\right)}{n(n-1)}\right)^{2r} \mathbb{E}\left[\mathbb{E}\left[\left(a\left(Z_m\right)-a\left(Z_m'\right)\right)^{2r} \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \mid \boldsymbol{\pi}\right] \\
&\leq\left(1-\frac{2}{gn^2}\right)^{2mr}\left(\frac{2kb\left(2n-kb-b\right)}{n^2}\right)^{2r} \mathbb{E}\left[\left(a\left(Z_m\right)-a\left(Z_m'\right)\right)^{2r} \mid \mathcal{H}_m\right], \qquad \text{(E.6)}
\end{aligned}$$

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 (E.6).

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

$$\begin{aligned}
\mathbb{E}\left[\left(a\left(Z\right)-a\left(Z'\right)\right)^{2r} \mid \mathcal{H}, \boldsymbol{\pi}\right] &= \left(\frac{kb-b}{n-b}\right) \mathbb{E}\left[\left(a\left(Z\right)-a\left(Z'\right)\right)^{2r} \mid \mathcal{G}\right] \\
&\quad+\left(\frac{n-kb}{n-b}\right) \mathbb{E}\left[\left(a\left(Z\right)-a\left(Z'\right)\right)^{2r} \mid \mathcal{H} \setminus \mathcal{G}\right]. \qquad \text{(E.7)}
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\left(a\left(Z\right)-a\left(Z'\right)\right)^{2r} \mid \mathcal{H} \setminus \mathcal{G}\right] \\
&=\mathbb{E}\left[\left(a\left(Z\right)-a\left(Z'\right)\right)^{2r} \mid \mathcal{H} \setminus \mathcal{G}\right]=\left(\frac{1}{gkb}\right)^{2r} \mathbb{E}\left[h(D_B, D_C, D_{\tilde{s}_C})^{2r}\right]. \qquad \text{(E.8)}
\end{aligned}$$

On the other hand, we can decompose

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

$$= \left(\frac{1}{gkb}\right)^{2r} \mathbb{E}\left[\left(h(D_B, D_C, D_{\hat{\mathsf{s}}_C}) - h(D_B, D_B, D_{\hat{\mathsf{s}}_B})\right)^{2r}\right] . \tag{E.9}$$

Consequently, by (E.7), (E.8), and (E.9), it suffices to express suitable bounds for the expectations

$$\begin{aligned} &\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \quad \text{and} \\ &\mathbb{E}\Bigg[\Bigg(\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right) \\ &\qquad - \left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_B}))\right) \Bigg)^{2r}\Bigg] \end{aligned} \tag{E.10}$$

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

$$\begin{aligned} &\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \\ &= \mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) + \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \\ &= \sum_{q=0}^{2r} \binom{2r}{q} \mathbb{E}\Bigg[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r-q} \\ &\qquad \left(\psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{q} \Bigg] \qquad \text{(Binomial Theorem)} \\ &\leq 2^{2r} \mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \end{aligned} \tag{E.11}$$

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

$$\begin{aligned} &\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r-q}\right] \\ &= \mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) + \psi(D_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \\ &= \sum_{q=0}^{2r} \binom{2r}{q} \mathbb{E}\Bigg[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{2r-q} \\ &\qquad \left(\psi(D_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C})) - \psi(\tilde{D}_B, \hat{\eta}(\tilde{D}_C, D_{\hat{\mathsf{s}}_C}))\right)^{q} \Bigg] \qquad \text{(Binomial Theorem)} \\ &\leq \sum_{q=0}^{2r} \binom{2r}{q} \left(\sigma_{\mathsf{valid}}^{(2r)}\right)^{\frac{2r-q}{2r}} \left(\sigma_{\mathsf{train}}^{(2r,1)}\right)^{\frac{q}{2r}} \qquad \text{(Hölder)} \\ &\leq 2^{2r} \sigma_{\mathsf{max}}^{(2r)}, \end{aligned} \tag{E.12}$$

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

$$\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right] \leq 2^{4r}\sigma^{(2r)}_{\mathsf{max}} \tag{E.13}$$

by (E.11) and (E.12).

Next, we consider the double difference term (E.10). In this case, we have that

$$\begin{aligned}
&\mathbb{E}\Bigg[\Bigg((\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C}))) \\
&\qquad - (\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_B}))) \Bigg)^{2r}\Bigg] \\
&= \mathbb{E}\Bigg[\Bigg((\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B}))) \\
&\qquad - (\psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_B}))) \Bigg)^{2r}\Bigg] \\
&= \sum_{q=0}^{2r} \binom{2r}{q} \mathbb{E}\Bigg[\Bigg((\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B}))) \\
&\qquad \cdot (\psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_C})) - \psi(D_C, \hat{\eta}(D_B, D_{\hat{\mathsf{s}}_B}))) \Bigg)^{2r}\Bigg] \qquad \text{(Binomial Theorem)} \\
&\leq 2^{2r}\mathbb{E}\left[\left(\psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_C})) - \psi(D_B, \hat{\eta}(D_C, D_{\hat{\mathsf{s}}_B}))\right)^{2r}\right] = 2^{2r}\sigma^{(2r,b-1)}_{\mathsf{train}}
\end{aligned} \tag{E.14}$$

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

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

as required. ■

E.5 Proof of Lemma D.5

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

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

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

$$\begin{aligned}
\hat{v}_{g,k}(Z) &= \frac{1}{g(g-1)k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-a(Z)\right)\left(T(\mathsf{s}_{\ell,i'},D)-a(Z)\right) \\
&= \frac{1}{g(g-1)k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}(D)\right)\left(T(\mathsf{s}_{\ell,i'},D)-a(Z)\right) \\
&\quad + \frac{1}{g(g-1)k}\sum_{\ell=1}^{g}\sum_{i=1}^{g}\left(\overline{a}(D)-a(Z)\right)\left(T(\mathsf{s}_{\ell,i},D)-a(Z)\right) \\
&= \frac{1}{g(g-1)k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}(D)\right)\left(T(\mathsf{s}_{\ell,i'},D)-a(Z)\right) \\
&= \frac{g}{g-1}\tilde{v}(Z)+\frac{1}{g(g-1)k}\sum_{\ell=1}^{g}\sum_{i=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}(D)\right)\left(\overline{a}(D)-a(Z)\right) \\
&= \frac{g}{g-1}\tilde{v}_{g,k}(Z)-\frac{1}{g-1}\check{v}_{g,k}(Z). \qquad \text{(E.15)}
\end{aligned}$$

Hence, it will suffice to characterize the exponential rates of conditional concentration for the statistics $\tilde{v}_{g,k}(Z)$ and $\check{v}_{g,k}(Z)$. These are established through the application of the following Lemma, which follows from an argument very similar to the proof of Theorem D.1.

Lemma E.4. *Suppose that Assumptions 4.1 and 4.2 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*

$$\log\frac{1}{2}P\left\{|\tilde{v}(Z)-v(D)|\geq t\mid D\right\}\lesssim -\frac{\delta}{(2-\varphi k-\varphi)^4}\frac{g^3t^2}{\Gamma^{(2)}_{k,\varphi,b}}, \qquad \text{(E.16)}$$

holds for all $t>0$ with probability greater than $1-\delta$.
(ii) *The conditional concentration inequality*

$$\log\frac{1}{2}P\left\{|\check{v}(Z)-v(D)|\geq t\mid D\right\}\lesssim -\frac{\delta}{(2-\varphi k-\varphi)^4}\frac{g^2t^2}{\Gamma^{(2)}_{k,\varphi,b}}, \qquad \text{(E.17)}$$

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

Putting the pieces together, by (E.15), we have that

$$\begin{aligned}
&P\left\{|\hat{v}_{g,k}(Z)-v_{g,k}(D)|\leq t\mid D\right\} \\
&= P\left\{|\frac{g}{g-1}\left(\tilde{v}_{g,k}(Z)-v_{g,k}(D)\right)-\frac{1}{g-1}\left(\check{v}_{g,k}(Z)-v_{g,k}(D)\right)|\leq t\mid D\right\} \\
&\geq P\left\{\frac{g}{g-1}|\tilde{v}_{g,k}(Z)-v_{g,k}(D)|+\frac{1}{g-1}|\check{v}_{g,k}(Z)-v_{g,k}(D)|\leq t\mid D\right\}
\end{aligned}$$

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

with probability greater than $1-\delta$ for some universal constant C, where the last inequality follows by Lemma E.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 $g\geq 2$. ■

E.6 Proof of Lemma D.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

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

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

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

as required. ■

E.7 Proof of Lemma D.7

E.7.1 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 (D.33). We have that

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

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 D.2. Consequently, we find that

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

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

E.7.2 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 E.5. *Suppose that Assumptions 4.1 and 4.2 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*

$$\log\frac{1}{2}P\left\{|U(\mathsf{R}_{\hat{g},k}, D)| \geq t\sqrt{v_{g^\star,k}(D)}, |\hat{g}/g^\star - 1| \leq c \mid D\right\} \tag{E.18}$$

$$\leq -\frac{\delta v_{1,k}(D)}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}}\frac{t^2}{c} \tag{E.19}$$

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*

$$\log\frac{1}{8}P\left\{|\hat{g}-g^\star| > cg^\star \mid D\right\} \lesssim -\frac{\delta(v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4}\frac{z^2_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}}\frac{c^2}{\xi^2} \tag{E.20}$$

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

Observe that

$$\begin{aligned}&P\left\{|U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D)| \geq \lambda\sqrt{2v_{g^\star,k}(D)}\right\}\\ &\lesssim P\left\{|U(\mathsf{R}_{\hat{g},k}, D)| \geq \lambda\sqrt{2v_{g^\star,k}(D)}\right\}\end{aligned}$$

$$\leq P\left\{|U(\mathsf{R}_{\hat{g},k},D)|\geq\lambda\sqrt{2v_{g^\star,k}\left(D\right)},|\hat{g}/g^\star-1|\leq c\mid D\right\}+P\left\{|\hat{g}/g^\star-1|\geq c\mid D\right\}$$

$$\lesssim\exp\left(-\frac{\delta v_{1,k(D)}}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}}\frac{\lambda^2}{c}\right)+\exp\left(-C\frac{\delta(v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4}\frac{z^2_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}}\frac{c^2}{\xi^2}\right)\tag{E.21}$$

for some universal constant C. Hence, it remains to choose c and λ such that the quantity (E.21) is less than $\rho_{k,\varphi,b}\left(\xi,\beta\mid D\right)$. First, we choose c such that

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

Choosing c by

$$c=\xi\left(\frac{(2-\varphi k-\varphi)^2\left(\Gamma^{(2)}_{k,\varphi,b}\right)^{1/2}}{z_{1-\beta/2}\delta^{1/2}(v_{1,k}(D))^{3/2}}\right)\log^{1/2}\left(\rho_{k,\varphi,b}\left(\xi,\beta\mid D\right)^{-1}\right)\tag{E.22}$$

suffices. Next, we choose λ such that

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

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

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

Choosing λ by

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

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

will then suffice, as required. ■

E.7.3 Part(iii). Observe that

$$P\left\{|\left(1-\frac{\hat{g}}{g^\star}\right)\left(a(\mathsf{R}_{\hat{g},k},D)-\overline{a}\left(D\right)\right)|\geq\lambda\sqrt{2v_{g^\star,k}\left(D\right)},|g^\star-\hat{g}|\leq cg^\star\mid D\right\}$$

$$=P\left\{|\left(a(\mathsf{R}_{\hat{g},k},D)-\overline{a}\left(D\right)\right)|\geq\frac{\lambda}{c}\sqrt{2v_{g^\star,k}\left(D\right)},|g^\star-\hat{g}|\leq cg^\star\mid D\right\}$$

$$\leq P\left\{\max_{|g^\star-g|\leq cg^\star}|\left(a(\mathsf{R}_{g,k},D)-\overline{a}\left(D\right)\right)|\geq\frac{\lambda}{c}\sqrt{2v_{g^\star,k}\left(D\right)}\mid D\right\}$$

$$\leq P\left\{\max_{|g^\star-g|\leq cg^\star}|\left(a(\mathsf{R}_{g,k},D)-\overline{a}\left(D\right)\right)|\geq\frac{\lambda}{c}\left(\frac{\xi}{z_{1-\beta/2}}\right)\mid D\right\}$$

$$\leq \sum_{|g^\star - g| \leq cg^\star} P\left\{ \left| \left(a(\mathsf{R}_{g,k}, D) - \overline{a}\left(D\right) \right) \right| \geq \frac{\lambda}{c} \left(\frac{\xi}{z_{1-\beta/2}} \right) \right\} . \tag{E.23}$$

By Theorem D.1, we have that

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

and so is (E.23) bounded from above by

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

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

$$\begin{aligned} & P\left\{ |Q(\mathsf{R}_{\hat{g},k}, D) - Q(\mathsf{R}'_{\hat{g}',k}, D)| \geq \lambda \sqrt{2 v_{g^\star,k}\left(D\right)} \mid D \right\} \\ & \lesssim P\left\{ \left| \left(1 - \frac{\hat{g}}{g^\star} \right) \left(a(\mathsf{R}_{\hat{g},k}, D) - \overline{a}\left(D\right) \right) \right| \geq \lambda \sqrt{2 v_{g^\star,k}\left(D\right)}, |g^\star - \hat{g}| \leq cg^\star \mid D \right\} \\ & + P\left\{ |g^\star - \hat{g}| \geq cg^\star \mid D \right\} \\ & \lesssim v_{1,k}\left(D\right) c \left(\frac{z_{1-\beta/2}}{\xi} \right)^2 \exp\left(- \frac{\delta v_{1,k}(D)}{2^5 (2 - \varphi k - \varphi)^2 \Gamma^{(1)}_{k,\varphi,b}} \frac{\lambda^2}{c^2} \right) \qquad \text{(E.24)} \\ & \quad + \exp\left(- \frac{1}{C} \frac{\delta (v_{1,k}(D))^3}{(2 - \varphi k - \varphi)^4} \frac{z_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}} \frac{c^2}{\xi^2} \right) \qquad \text{(E.25)} \end{aligned}$$

for some universal constant C by Lemma E.5, Part (ii). If c is chosen to satisfy (E.22), the term (E.25) will be bounded above by $\rho_{\varphi,k}\left(\xi, \beta \mid D\right)$ for all sufficiently small ξ. By plugging this value of c and $\lambda = \lambda_{\varphi,k}\left(\xi, \beta \mid D\right)$ into (E.24), we obtain the term

$$\begin{aligned} & \left(\frac{z^2_{1-\beta/2}}{\xi} \right) \left(\frac{(2 - \varphi k - \varphi)^2 \left(\Gamma^{(2)}_{k,\varphi,b}\right)^{1/2}}{z_{1-\beta/2} \delta^{1/2} (v_{1,k}(D))^{1/2}} \right) \log^{1/2}\left(\rho_{k,\varphi,b}\left(\xi, \beta \mid D\right)^{-1} \right) \\ & \exp\left(- \frac{1}{\xi} \frac{\delta^{1/2} (v_{1,k}(D))^{3/2}}{(2 - \varphi k - \varphi)^2 (\Gamma^{(2)}_{k,\varphi,b})^{1/2}} \tilde{\lambda}^U_{k,\varphi,b}\left(D\right) \log\left(\rho_{k,\varphi,b}\left(\xi, \beta \mid D\right)^{-1} \right) \right) \qquad \text{(E.26)} \end{aligned}$$

as required. Observe that the term (E.26) is bounded above by $\rho_{\varphi,k}\left(\xi, \beta \mid D\right)$, for all sufficiently small ξ, completing the proof. ■

E.8 Proofs for Supporting Lemmas

E.8.1 Proof of Lemma E.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

$$\frac{1}{k}\sum_{i=1}^{k}\left(T_i(\pi_{m,\ell}) - T_i\left(\pi'_{m,\ell}\right)\right) = 0 \tag{E.27}$$

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

$$\begin{aligned}
&\sum_{m=0}^{\infty}\left|\mathbb{E}\left[a(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D) - a(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D) \mid \boldsymbol{\pi}_0 = \boldsymbol{\psi}, \boldsymbol{\pi}'_0 = \boldsymbol{\psi}', D\right]\right| \qquad \text{(E.28)}\\
&\leq \sum_{m=0}^{\infty}\frac{1}{g}\frac{1}{k}\sum_{\ell=1}^{g}\sum_{i=1}^{k}\mathbb{E}\left[\left|\left(T_i(\pi_{m,\ell}) - T_i\left(\pi'_{m,\ell}\right)\right)\right| \mid \boldsymbol{\pi}_0 = \boldsymbol{\psi}, \boldsymbol{\pi}'_0 = \boldsymbol{\psi}', D\right]\\
&\leq \frac{1}{g}\frac{1}{k}\sum_{\ell=1}^{g}\sum_{i=1}^{k}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left|T(\mathsf{s}, D) - T(\mathsf{s}', D)\right|\mathbb{E}\left[N(\ell)\right]
\end{aligned}$$

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

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

and

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

and thereby obtain the bound

$$\mathbb{E}[N(\ell)] = \sum_{r=1}^{n}\mathbb{E}\left[N_r(\ell) - N_{r-1}(\ell)\right] \leq \sum_{r=1}^{n}\frac{gn^2}{(n-r+1)^2} = gn^2\sum_{r=1}^{n}\frac{1}{r^2} \leq 2gn^2. \tag{E.29}$$

Hence, we find that (E.28) is upper bounded by

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

as required. ∎

E.8.2 Proof of Lemma E.2. Observe that

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

Thus, we find that

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

as required. ■

E.8.3 Proof of Lemma E.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

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

and so

$$(x_1 \cdots x_{2^c})^{2s} \leq \frac{1}{2} \left(x_{i_1} \cdots x_{i_{2^{c-1}}} \right)^{2(s+1)} + \frac{1}{2} \left(x_{i_{2^{c-1}+1}} \cdots x_{2^c} \right)^{2(s+1)}. \tag{E.30}$$

Consequently, we have that

$$x_1 \cdots x_{2^c} \leq \frac{1}{2^c} \sum_{i=1}^{2^c} x_i^{2c} \tag{E.31}$$

through 2^r applications of (E.30).

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

$$\mathbb{E}\left[\left(\sum_{m=0}^{\infty} X_m \right)^{2^c} \right] = \sum_{i_1=0}^{\infty} \cdots \sum_{i_{2^c}=0}^{\infty} \mathbb{E}\left[X_{i_1} \cdots X_{i_{2^c}} \right] \tag{E.32}$$

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 (E.31). Hence, by (E.32), 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. ■

E.8.4 Proof of Lemma E.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 D.4. The subsequent result follows from an argument similar to Lemma D.1, again based on an idea expressed by Lemma 4.1 of Chatterjee (2005).

Lemma E.6.

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

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

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

hold almost surely.

(ii) *The functions*

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

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

are finite and satisfy the equalities

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

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

almost surely.

Next, we apply the concentration inequality stated in Theorem D.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

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

with probability $1-\delta$, where

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

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

We obtain such a characterization through the application of Lemma D.3. To this end, observe that, by Lemma E.6, part (i), the bound

$$\left(\sum_{m=0}^{\infty} \mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m) - \tilde{v}_{g,k}(Z'_m)\right) \mid Z_0 = Z, Z'_0 = Z'\right]\right)^2 \tag{E.33}$$

$$\leq n^4 \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T(\mathsf{s},D) - T(\mathsf{s}',D)\right)^4 \tag{E.34}$$

holds almost surely. The right hand side of (E.34) is square integrable, as the eighth-order split-stability $\zeta^{(8)}$ is finite almost surely. An analogous statement holds for the statistic $\check{v}_{g,k}(Z)$. Thus, deterministic bounds of the form (D.29) can be obtained through the application of the following Lemma.

Lemma E.7. *Suppose that Assumptions 4.1 and 4.2 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*

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

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

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

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

Thus, by Lemma D.3, part (i), the inequalities

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

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

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 (E.16) by applying Theorem D.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 D.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 (E.17) by Theorem D.2. ■

E.8.5 Proof of Lemma E.6. Reinstate the notation of the proof of Lemma E.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

$$\sum_{m=0}^{\infty}\left|\mathbb{E}\left[\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\mid Z_0=\left(\mathsf{r}(\boldsymbol{\psi}),D\right),Z'_0=\left(\mathsf{r}(\boldsymbol{\psi}'),D\right)\right]\right|$$
$$\leq\sum_{m=0}^{\infty}\frac{1}{g^2k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\mathbb{E}\Bigg[\Bigg|\left(T_i(\pi_{m,\ell})-\overline{a}(D)\right)\left(T_{i'}(\pi_{m,\ell})-\overline{a}(D)\right)$$
$$-\left(T_i\left(\pi'_{m,\ell}\right)-\overline{a}(D)\right)\left(T_{i'}\left(\pi'_{m,\ell}\right)-\overline{a}(D)\right)\Bigg|\mid Z_0=\left(\mathsf{r}(\boldsymbol{\psi}),D\right),Z'_0=\left(\mathsf{r}(\boldsymbol{\psi}'),D\right)\Bigg]$$

$$\leq \frac{2}{g^2k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left|\left(T\left(\mathsf{s},D\right)-\overline{a}\left(D\right)\right)\left(T\left(\mathsf{s}',D\right)-\overline{a}\left(D\right)\right)\right|\mathbb{E}\left[N\left(\ell\right)\right].$$

$$\leq \frac{2}{g}\max_{\mathsf{s}\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-\overline{a}\left(D\right)\right)^2\mathbb{E}\left[N\left(\ell\right)\right].$$

$$\leq \frac{1}{2g}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^2\mathbb{E}\left[N\left(\ell\right)\right].$$

Similarly, we have that

$$\sum_{m=0}^{\infty}\left|\mathbb{E}\left[\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z_m')\mid Z_0=\left(\mathsf{r}\left(\boldsymbol{\psi}\right),D\right),Z_0'=\left(\mathsf{r}\left(\boldsymbol{\psi}'\right),D\right)\right]\right|$$

$$\leq \sum_{m=0}^{\infty}\frac{1}{g^2k^2}\sum_{\ell,\ell'=1}^{g}\sum_{i,i'=1}^{k}\mathbb{E}\Bigg[\Big|\left(T_i\left(\pi_{m,\ell}\right)-\overline{a}\left(D\right)\right)\left(T_{i'}\left(\pi_{m,\ell'}\right)-\overline{a}\left(D\right)\right)$$

$$-\left(T_i\left(\pi_{m,\ell}'\right)-\overline{a}\left(D\right)\right)\left(T_{i'}\left(\pi_{m,\ell'}'\right)-\overline{a}\left(D\right)\right)\Big|\mid Z_0=\left(\mathsf{r}\left(\boldsymbol{\psi}\right),D\right),Z_0'=\left(\mathsf{r}\left(\boldsymbol{\psi}'\right),D\right)\Bigg]$$

$$\leq \frac{2}{g^2k^2}\sum_{\ell,\ell'=1}^{g}\sum_{i,i'=1}^{k}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left|\left(T\left(\mathsf{s},D\right)-\overline{a}\left(D\right)\right)\left(T\left(\mathsf{s}',D\right)-\overline{a}\left(D\right)\right)\right|\mathbb{E}\left[\max\left\{N\left(\ell\right),N\left(\ell'\right)\right\}\right]$$

$$\leq \frac{2}{g^2k^2}\sum_{\ell,\ell'=1}^{g}\sum_{i,i'=1}^{k}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left|\left(T\left(\mathsf{s},D\right)-\overline{a}\left(D\right)\right)\left(T\left(\mathsf{s}',D\right)-\overline{a}\left(D\right)\right)\right|\mathbb{E}\left[N\left(\ell\right)+N\left(\ell'\right)\right]$$

$$\leq \frac{1}{2}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^2\left|\mathbb{E}\left[N\left(\ell\right)+N\left(\ell'\right)\right]\right.$$

Thus, part (i) of the Lemma follows by the bound (E.29). Part (ii) of the Lemma then follows by an argument analogous to the proof of Lemma D.1. ■

E.8.6 Proof of Lemma E.7. We reinstate the notation introduced in the proof of Lemma D.4 from Section E.4. First, observe that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z_m')\mid Z,Z'\right]^2\mid\boldsymbol{\pi}\right]\\
&=\mathbb{E}\left[\left(P\left\{\mathcal{H}_m\right\}\mathbb{E}\left[\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z_m')\mid\mathcal{H}_m,Z,Z'\right]\right)^2\mid\boldsymbol{\pi}\right]\\
&\leq\left(1-\frac{2}{gn^2}\right)^{2m}\left(\frac{2kb\left(2n-bk-b\right)}{n^2}\right)^2\mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z_m')\right)^2\mid\mathcal{H}_m,\boldsymbol{\pi}\right]
\end{aligned}\tag{E.35}$$

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

$$
\begin{aligned}
&\mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \\
&= \frac{1}{g^4}\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)^2-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)^2\right)^2 \mid \mathcal{H}_m\right] \\
&= \frac{1}{g^4}\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)+\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^2\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^2 \mid \mathcal{H}_m\right] \\
&\leq \frac{1}{g^4}\left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)+\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} && \text{(Cauchy-Schwarz)} \\
&\quad\cdot\left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} \\
&\leq \frac{2^4}{g^4}\left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(D)\right)^4\right]\right)^{1/2} && \text{(Hölder)} \\
&\quad\cdot\left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} . && \text{(E.36)}
\end{aligned}
$$

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 D.2, as Assumption 4.1 is maintained and the eighth-order sample-split stability $\zeta^{(8)}$ is finite. In turn, we have that

$$
\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}),D)-\overline{a}(\mathsf{r}(\pi'_{m,L}),D)\right)^4 \mid \mathcal{H}_m\right] \leq \frac{4}{k^4b^4}\Gamma^{(2)}_{k,\varphi,b} \tag{E.37}
$$

by (E.7), (E.13), and (E.14). Hence, we have that

$$
\mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \leq \frac{3\cdot 2^9}{k^2b^2g^4}(2-\varphi k-\varphi)^2\Gamma^{(2)}_{k,\varphi,b} . \tag{E.38}
$$

Combining (E.35), (E.37), (E.38), 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

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z'_m) \mid Z, Z'\right]^2 \mid \boldsymbol{\pi}\right] \\
&\leq \left(1-\frac{2}{gn^2}\right)^{2m}\left(\frac{2kb\,(2n-bk-b)}{n^2}\right)^2\mathbb{E}\left[\left(\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] && \text{(E.39)}
\end{aligned}
$$

In this case, we can compute

$$
\mathbb{E}\left[\left(\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right]
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\left(a(Z_m)-\overline{a}(D)\right)^2-\left(a(Z'_m)-\overline{a}(D)\right)^2\right)^2 \mid \mathcal{H}_m\right] \\
&= \mathbb{E}\left[\left(a(Z_m)-\overline{a}(D)+a(Z'_m)-\overline{a}(D)\right)^2\left(a(Z_m)-a(Z'_m)\right)^2 \mid \mathcal{H}_m\right] \\
&\leq \left(\mathbb{E}\left[\left(a(Z_m)-\overline{a}(D)+a(Z'_m)-\overline{a}(D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} \qquad \text{(Cauchy-Schwarz)} \\
&\quad \cdot \left(\mathbb{E}\left[\left(a(Z_m)-a(Z'_m)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} \\
&\leq 2^4\left(\mathbb{E}\left[\left(a(Z_m)-\overline{a}(D)\right)^4\right]\right)^{1/2}\left(\mathbb{E}\left[\left(a(Z_m)-a(Z'_m)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2}, \qquad \text{(E.40)}
\end{aligned}
$$

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

$$
\mathbb{E}\left[\left(a(Z_m)-a(Z'_m)\right)^4\right] \leq 3^2\left(\frac{2^4\varphi^2(2-\varphi k-\varphi)^2}{g}\right)^2\Gamma^{(2)}_{k,\varphi,b} \qquad \text{(E.41)}
$$

by Theorem D.2, as Assumption 4.1 is maintained and the eighth-order sample-split stability $\zeta^{(8)}$ is finite. Similarly. we have that

$$
\mathbb{E}\left[\left(a(Z_m)-\overline{a}(D)\right)^4 \mid \mathcal{H}_m\right] \leq \frac{4}{g^4k^4b^4}\Gamma^{(2)}_{k,\varphi,b} \qquad \text{(E.42)}
$$

by (E.7), (E.13), and (E.14). Combining (E.39), (E.41), (E.42), 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. ■

E.8.7 Proof of Lemma E.5, Part (i). Observe that

$$
\begin{aligned}
&P\left\{|U(\mathsf{R}_{\hat{g},k},D)| \geq t\sqrt{v_{g^\star,k}(D)}, |\hat{g}-g^\star| \leq cg^\star \mid D\right\} \\
&= P\left\{\sum_{i=1}^{g}\overline{a}\left(\mathsf{r}_{i,k},D\right)-\sum_{i=1}^{g^\star}\overline{a}\left(\mathsf{r}_{i,k},D\right) \geq tg^\star\sqrt{v_{g^\star,k}(D)}, |\hat{g}-g^\star| \leq cg^\star \mid D\right\} \\
&\leq P\left\{\max_{g^\star(1-c)\leq g\leq g^\star}\left|\sum_{i=1}^{g}\overline{a}\left(\mathsf{r}_{i,k},D\right)-\sum_{i=1}^{g^\star}\overline{a}\left(\mathsf{r}_{i,k},D\right)\right| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\} \\
&+ P\left\{\max_{g^\star\leq g\leq g^\star(1+c)}\left|\sum_{i=1}^{g}\overline{a}\left(\mathsf{r}_{i,k},D\right)-\sum_{i=1}^{g^\star}\overline{a}\left(\mathsf{r}_{i,k},D\right)\right| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\} \\
&= P\left\{\max_{g^\star(1-c)\leq g\leq g^\star}\left|\sum_{i=1}^{g^\star-g}\overline{a}\left(\mathsf{r}_{i,k},D\right)\right| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\} \\
&+ P\left\{\max_{g^\star\leq g\leq g^\star(1+c)}\left|\sum_{i=1}^{g-g^\star}\overline{a}\left(\mathsf{r}_{i,k},D\right)\right| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\}. \qquad \text{(E.43)}
\end{aligned}
$$

To bound the two probabilities (E.43), we apply the following Chernoff-type variation to the Kolmogorov maximal inequality, due to Steiger (1970).

Theorem E.2 (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 D.2, we find that Theorem Corollary D.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 4.1 holds and the fourth-order split stability $\zeta^{(4)}$ is finite. Thus, by setting

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

we find that

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

with probability greater than $1-\delta$, by Theorem E.2 and the inequality (E.43).

E.8.8 Proof of Lemma E.5, Part (ii). Observe that

$$P\left\{|\hat{g}-g^\star| > cg^\star \mid D\right\} = P\left\{g^\star(1+c) < \hat{g} \mid D\right\} + P\left\{\hat{g} < g^\star(1-c) \mid D\right\}. \tag{E.44}$$

We begin by handling the first term. We have that

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

where the first inequality follows the definition of $g^\star$. Thus, as Assumption 4.1 holds and the eighth-order split stability $\zeta^{(8)}$ is finite, we have that

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

with probability greater than $1-\delta$, by Lemma D.5, the definition of $g^\star$ and the assumption that $0 < c < 1/2$.

Next, we bound the second term in (E.44). Define the partial sums

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

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

Observe that the equality

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

from the proof of Lemma D.5, implies that

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

for all positive g. Thus, we can write

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

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

$$\min_{2 \le g' \le g^\star(1-c)} \frac{(g' A_{g'} - B_{g'})}{(g'-1)(g')^2} + \frac{1}{g'} v_{1,k}(D) \le \left(\frac{\xi}{z_{1-\beta/2}}\right)^2, \tag{E.45}$$

it must also be the case that

$$\min_{2 \le g' \le g^\star(1-c)} (g'-1)(g')^2 \left(\frac{(g' A_{g'} - B_{g'})}{(g'-1)(g')^2} + \frac{1}{g'} v_{1,k}(D) \right) \le (\hat{g}-1)(\hat{g})^2 \left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \tag{E.46}$$

as $\hat{g} \le g^\star(1-c)$ necessarily. But then, similarly, (E.46) implies that

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

and in turn

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

are then also true. Finally, again as (E.45) is equivalent to $\hat{g} < g^\star(1-c)$, we have that

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

Now, observe that we can write

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

We bound this term by combining the argument used to establish Lemma D.5 with an application of Theorem E.2. To this end, observe that

$$\begin{aligned}
&P\left\{\max_{2\leq g'\leq g}\frac{B_{g'}-g'A_{g'}}{\left(g-1\right)g^2}\leq t\mid D\right\}\\
&\geq P\left\{\max_{2\leq g'\leq g}\frac{-1}{g-1}\frac{g'}{g^2}A_{g'}\leq\frac{1}{1+\sqrt{g}}t\mid D\right\}+P\left\{\max_{2\leq g'\leq g}\frac{1}{g-1}\frac{1}{g^2}B_{g'}\leq\frac{\sqrt{g}}{1+\sqrt{g}}t\mid D\right\}-1\\
&\geq P\left\{\max_{2\leq g'\leq g}\frac{g}{g-1}\frac{-1}{g^2}A_{g'}\leq\frac{\sqrt{g}}{1+\sqrt{g}}t\mid D\right\}+P\left\{\max_{2\leq g'\leq g}\frac{1}{g-1}\frac{1}{g^2}B_{g'}\leq\frac{1}{1+\sqrt{g}}t\mid D\right\}-1\\
&\geq P\left\{\max_{2\leq g'\leq g}-A_{g'}\leq g^2\frac{1}{\sqrt{g}}\frac{g-1}{1+\sqrt{g}}t\mid D\right\}+P\left\{\max_{2\leq g'\leq g}|B_{g'}|\leq g^2\frac{g-1}{1+\sqrt{g}}t\mid D\right\}-1 \qquad \text{(E.47)}
\end{aligned}$$

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 4.1 holds and the eighth-order split stability $\zeta^{(8)}$ is finite, though inspection of the proof of Lemma D.2, Lemma E.4 implies that implies that

$$\begin{aligned}
&\inf_{\theta>0}\left(\log\mathbb{E}\left[\exp\left(\theta A_g\right)\mid D\right]-\theta g^2\frac{1}{\sqrt{g}}\frac{g-1}{1+\sqrt{g}}t\right)\\
&=\inf_{\theta>0}\left(\log\mathbb{E}\left[\exp\left(\frac{\theta}{g^2}A_g\right)\mid D\right]-\theta\frac{1}{\sqrt{g}}\frac{g-1}{1+\sqrt{g}}t\right)\\
&\lesssim-\frac{\delta}{(2-\varphi k-\varphi)^4}\frac{g^3t^2}{\Gamma^{(2)}_{k,\varphi,b}} \qquad \text{(E.48)}
\end{aligned}$$

and

$$\begin{aligned}
&\inf_{\theta>0}\left(\log\mathbb{E}\left[\exp\left(\theta B_g\right)\mid D\right]-\theta g^2\frac{g-1}{1+\sqrt{g}}t\right)\\
&=\inf_{\theta>0}\left(\log\mathbb{E}\left[\exp\left(\frac{\theta}{g^2}B_g\right)\mid D\right]-\theta\frac{g-1}{1+\sqrt{g}}t\right)\\
&\lesssim-\frac{\delta}{(2-\varphi k-\varphi)^4}\frac{g^3t^2}{\Gamma^{(2)}_{k,\varphi,b}} \qquad \text{(E.49)}
\end{aligned}$$

each with probability greater than $1-\delta$, where we have used the facts that $4g\leq\left(1+\sqrt{g}\right)^2$ for all $g\geq1$ and $(1/4)\,g^2\geq(g-1)^2$ for all $g\geq2$. Hence, by Theorem E.2, and plugging (E.48) and

(E.49) into (E.47), we find that

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

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

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

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

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

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