Inference in Regression Discontinuity Designs with Clustered Data

Claudia Noack Tomasz Olma Christoph Rothe

Abstract

Clustered sampling is prevalent in empirical regression discontinuity (RD) designs, but it has not received much attention in the theoretical literature. In this paper, we introduce a general model-based framework for such settings and derive high-level conditions under which the standard local linear RD estimator is asymptotically normal. We verify that our high-level assumptions hold across a wide range of empirical designs, including settings of growing cluster sizes. We further show that clustered standard errors that are currently used in practice can be either inconsistent or overly conservative in finite samples. To address these issues, we propose a novel nearest-neighbor-type variance estimator and illustrate its properties in a diverse set of empirical applications.

This version: March 19, 2026. We thank Debopam Bhattacharya, Morten Nielsen, Zhuan Pai and numerous seminar and conference participants for helpful comments and suggestions. The second author gratefully acknowledges financial support from the European Research Council ERC through grant SH-1852332. Author contact information: Claudia Noack, Department of Economics, University of Bonn. E-Mail: claudia.noack@uni-bonn.de. Website: https://claudianoack.github.io. Tomasz Olma, Department of Statistics, Ludwig Maximilian University of Munich. E-Mail: t.olma@lmu.de. Website: https://tomaszolma.github.io. Christoph Rothe, Department of Economics, University of Mannheim. E-Mail: rothe@vwl.uni-mannheim.de. Website: http://www.christophrothe.net.

1. INTRODUCTION

Regression discontinuity (RD) designs are widely applied in economics and other social sciences. In these settings, treatment is assigned whenever a unit's realization of the running variable crosses a known cutoff; for instance, a candidate is elected only if their vote share exceeds 50%. Under continuity conditions on the conditional expectations of the potential outcomes, the average treatment effect at the cutoff is identified by the jump in the conditional expectation of the observed outcome given the running variable at the cutoff. This jump is typically estimated as the difference of the local linear estimates on either side of the cutoff.

Most theoretical results on estimation and inference in RD designs are derived in settings where the researcher observes a sample of independent and identically distributed (i.i.d.) observations drawn from a large population (e.g., Hahn et al., 2001; Imbens and Kalyanaraman, 2012; Calonico et al., 2014; Armstrong and Kolesár, 2020). In practice, however, applied researchers often regard the i.i.d. assumption as unrealistic and routinely report clustered standard errors. For example, we revisited the survey of recent RD studies published in the journals of the American Economic Association conducted by Noack et al. (2025) and found that clustered standard errors were used in around 80% of the surveyed articles. Despite the prevalence of clustered sampling in applied RD studies, formal results for local linear RD estimators – and more generally local polynomial regression – under such dependence structures remain limited. In particular, the existing literature offers little guidance on the conditions under which these estimators are asymptotically normal across the range of clustering patterns encountered in practice, or on how to conduct valid inference in such settings.

This paper makes two main contributions. First, we provide an asymptotic theory for the local linear RD estimator for clustered data with a large number of independent groups. From a statistical perspective, RD estimators are weighted averages of the outcome variable where the weights depend on the running variable, the kernel, and the bandwidth. When units are clustered, the interaction between local weighting and within-cluster dependence alters the asymptotic behavior of the local linear RD estimator. We derive high-level conditions under which the local linear RD estimator is asymptotically normally distributed. These conditions are formulated in terms of the weights assigned to units from different clusters and translate into restrictions on cluster sizes within the estimation window. To relate our high-level conditions to empirically relevant designs, we introduce four stylized asymptotic frameworks motivated by empirical RD applications. These frameworks capture how the asymptotic behavior of local linear RD estimators depends on (i) the effective number of units per cluster within the estimation window, (ii) the dependence structure of the running variable within clusters, and (iii) assumptions on

the within-cluster covariance structure of the outcome. Within these frameworks, we find that distinct convergence rates and optimal bandwidth choices are qualitatively different from i.i.d. settings. Our results complement the analysis of Hansen and Lee (2019) by considering nonparametric models. In contrast to their analysis, where the estimators are based on the full sample, the localization of the RD estimator leads to non-standard convergence rates.

Second, we consider estimation of the conditional variance of the local linear RD estimator. We show that the conventional clustered regression residual-based standard error is consistent under the same cluster size conditions that ensure asymptotic normality of the RD estimator. However, in settings with independent data, the regression residual-based standard errors are known to exhibit less finite-sample bias than the so-called nearest-neighbors standard errors. A naive adaptation of the nearest-neighbors standard error for independent data to clustered settings is invalid, and to our knowledge, no valid nearest-neighbors-type standard error for clustered RD designs currently exists. As the second main contribution, we propose a novel clustered nearest-neighbors (CNN) standard error for RD estimators. Our proposed method chooses nearest neighbors taking into account the clustering structure and exploiting independence between clusters. We establish consistency of our proposed CNN standard error under our high-level assumptions on the cluster sizes.[1]

We complement our theoretical analysis with empirical applications that assess the finite-sample performance of the proposed standard error relative to existing alternatives.

Related Literature. Cluster-robust inference is routinely employed in empirical RD designs. Despite this fact, formal results remain limited, even for the standard nonparametric regression using local polynomial estimators. For RD designs with clustering, Bartalotti and Brummet (2017) show asymptotic normality of local polynomial RD estimators under the assumption that all clusters are of the same size and the realizations of the running variable are on the same side of the cutoff within each cluster. Clustered standard errors have been implemented in popular RD packages `rdrobust` and `RDHonest` without much theoretical foundation. Our paper contributes to this literature by providing a unified theory for all common RD variants with arbitrary clustering.

Lin and Carroll (2000), Wang (2003), and Bhattacharya (2005) study general local polynomial estimators under bounded cluster sizes. In these regimes, as the bandwidth converges to zero, the probability of having more than one unit within the estimation window converges to zero for any given cluster, and in consequence, the clustering does

[1]Although our CNN standard error is developed for RD designs in this paper, the general idea behind it extends naturally to other conditional inference problems under misspecification, such as those studied by Abadie et al. (2014).

not affect the asymptotic distribution. Shimizu (2025) studies nonparametric density and local polynomial estimation under clustered sampling with heterogeneous and potentially unbounded cluster sizes. However, his framework imposes restrictions that rule out several empirically relevant RD settings. In particular, cluster sizes within the estimation window are required to remain uniformly bounded in expectation, the covariates are not allowed to be perfectly dependent within a cluster, and the within-cluster covariance structure of the residuals takes on a very specific form. Furthermore, his proposed variance estimator relies on parametric assumptions. By contrast, our framework accommodates weaker conditions on cluster sizes, allows for more general dependence structures in the running variable within clusters, and imposes milder assumptions on the covariance functions, while still delivering asymptotic normality. Moreover, our proposed nearest-neighbor standard error is fully nonparametric and does not rely on parametric modeling assumptions.

In settings of regular parameters, there is a vast literature about cluster-robust inference. Liang and Zeger (1986); White (2014); Arellano (1987) provide the foundational large-G theory for cluster-robust covariances for bounded cluster sizes; see Cameron and Miller (2015); MacKinnon et al. (2023) for reviews. Djogbenou et al. (2019); Hansen and Lee (2019); Bugni et al. (2025); Hansen (2025) extend the results to unequal, possibly large clusters. Abadie et al. (2023) discuss clustering adjustments from a design-based perspective. Chiang et al. (2025) point out that in many empirical applications the size of the largest cluster is not negligible relative to the total sample size, and they provide an alternative bootstrap inference procedure.

This paper is the first to propose and formally study a cluster-robust nearest-neighbors-type standard error. In this regard, we extend the work of Abadie et al. (2014), who showed consistency of the nearest-neighbors standard errors under i.i.d. sampling (in a more general class of misspecified models).

Plan of the Paper. Section 2 introduces the model and gives a preview of our main results. Section 3 states the high-level assumptions and establishes asymptotic normality of the local linear RD estimator. Section 4 studies the asymptotic frameworks. Section 5 introduces our proposed standard error and we show that it is consistent. Section 6 contains the empirical applications. All proofs are collected in the Appendix.

2. SETTING AND PREVIEW OF THE RESULTS

2.1. **Clustered Sampling.** We consider a sharp RD design, in which a unit receives the treatment if and only if their running variable exceeds a known cutoff value, which we normalize to zero. The observed data is divided into G clusters, and in each cluster, we observe n_g units. Let X_{gi} and Y_{gi} denote the running variable and the outcome variable

of observation i in cluster g, respectively. The total sample size is given by $n = \sum_{g\in[G]} n_g$, where $[G] = (1, \ldots, G)$. In our asymptotic analysis, we will treat G and $(n_g)_{g\in[G]}$ as deterministic sequences indexed by the sample size n.

Observations in different clusters are independent, but they can be dependent within a cluster. The outcome is generated according to the model

$$Y_{gi} = \mu(X_{gi}) + \varepsilon_{gi}, \tag{2.1}$$

where $\mu(x) = \mathbb{E}[Y_{gi}|X_{gi} = x]$, $\mathbb{E}[\varepsilon_{gi}|\mathcal{X}_g] = 0$, $\mathcal{X}_g = (X_{gi})_{i\in I_g}$, and $I_g = (1, \ldots, n_g)$. The error terms ε_{gi} can be arbitrarily dependent within a cluster. We denote their variance and covariances, conditional on $\mathcal{X}_g$ by

$$\begin{aligned}\sigma^2_{g,i} &= \mathrm{Var}(Y_{gi}|\mathcal{X}_g),\\ \sigma_{g,ij} &= \mathrm{Cov}(Y_{gi}, Y_{gj}|\mathcal{X}_g),\end{aligned}$$

and we denote the covariance matrix of $\mathcal{Y}_g = (Y_{gi})_{i\in I_g}$ conditional on $\mathcal{X}_g$ by Σ_g.

Under continuity assumptions on the conditional expectation of the potential outcomes, the jump in the conditional expectation $\mu(x)$ at the cutoff identifies the average treatment effect of units at the cutoff (Hahn et al., 2001). Our parameter of interest is therefore given by

$$\tau = \mu(0^+) - \mu(0^-),$$

where for a generic function f, $f(0^+)$ and $f(0^-)$ denote the right and left limits of the function f at zero.

Remark 1. In the model of observed data in (2.1), we assume that the conditional expectation function μ is the same in each cluster in the sample. We note that this model allows for sampling from a population where clusters are heterogeneous in terms of μ if, in each repeated sample keeping $(n_g)_{g\in[G]}$ fixed, the observed clusters are drawn at random from the superpopulation of clusters of infinite size. Specifically, suppose that we draw a random set of clusters $\mathcal{G} = \{\tilde{g}_1, \ldots, \tilde{g}_G\}$ and observe a sample of random units from these clusters, $\{(X_{\tilde{g}i}, Y_{\tilde{g}i})_{i\in[I_{\tilde{g}}]}\}_{\tilde{g}\in\mathcal{G}}$. If we assume that

$$Y_{\tilde{g}i} = \mu_{\tilde{g}}(X_{\tilde{g}i}) + \eta_{\tilde{g}i}, \quad \mathbb{E}[\eta_{\tilde{g}i}|\mathcal{X}_{\tilde{g}}] = 0,$$

then this model fits into our framework if we define

$$\mu(x) = \mathbb{E}[\mu_{\tilde{g}}(x)] \quad \text{and} \quad \varepsilon_{\tilde{g}i} = \eta_{\tilde{g}i} + \mu_{\tilde{g}}(X_{\tilde{g}i}) - \mu(X_{\tilde{g}i}),$$

where the expectation is taken w.r.t. the distribution of clusters in the population.

2.2. **Local Linear RD Estimator.** In practice, it is common to estimate the RD parameter via local linear RD regressions. This estimator is defined as

$$\widehat{\tau}(h) = e_1^\top \underset{\beta \in \mathbb{R}^4}{\operatorname{argmin}} \sum_{i=1}^{n} k_h(X_i)(Y_i - V_i^\top \beta)^2 \equiv \sum_{g \in [G]} \sum_{i \in I_g} w_{gi}(h) Y_{gi}, \tag{2.2}$$

where $V_i = (T_i, X_i, T_i X_i, 1)^\top$, $k_h(v) = k(v/h)/h$ with $k(\cdot)$ a kernel function and $h > 0$ a bandwidth, and $e_1 = (1, 0, 0, 0)^\top$ is the first unit vector. The weights $w_{gi}(h)$ depend only on the running variable and the bandwidth; the exact expressions for the weights are given in Appendix C.1. In our setting, the variance of the RD estimator conditional on the running variable $\mathcal{X}_n = (\mathcal{X}_g)_{g \in [G]}$ equals

$$se^2(h) \equiv \operatorname{Var}(\widehat{\tau}(h) | \mathcal{X}_n) = \sum_{g \in [G]} \sum_{i,j \in I_g} w_{gi}(h) w_{gj}(h) \sigma_{g,ij}. \tag{2.3}$$

2.3. **Preview of Asymptotic Results.** Our first main result establishes the large-sample behavior of the RD estimator under clustered sampling. We derive it under high-level conditions on the weights $w_{gi}(h)$ that translate into restrictions on cluster sizes within the estimation window. Under these general high-level conditions, we show that the local linear RD estimator is asymptotically normal:

$$se(h)^{-1}\big(\widehat{\tau}(h) - \mathbb{E}[\widehat{\tau}(h) \mid \mathcal{X}_n]\big) \xrightarrow{d} \mathcal{N}(0, 1).$$

The convergence rate of the conditional variance $se^2(h)$ depends on the covariance of outcomes within each cluster, Σ_g, the joint distribution of the realizations of the running variable within each cluster, $\mathcal{X}_g$, and the cluster sizes, and it can be slower than the convergence rate of $(nh)^{-1}$ obtained in i.i.d. settings. For example, we show in Section 4 that if the joint distribution of $\mathcal{X}_g$ admits a bounded density and some mild regularity conditions hold, then

$$se^2(h) = O_P\left(\frac{1 + \lambda_n}{nh}\right),$$

where

$$\lambda_n = \frac{h}{n} \sum_{g \in [G]} n_g(n_g - 1).$$

If, in turn, the distribution of $\mathcal{X}_g$ is degenerate, meaning that the realizations of the running variable are equal within each cluster, then

$$se^2(h) = O_P\left(\frac{1 + \lambda_n/h}{nh}\right).$$

The above results provide bounds on the convergence rate of the conditional variance

$se^2(h)$. Whether these rates are binding depends on the exact form of the conditional covariance matrix Σ_g. In Section 4.4, we give an example where these bounds are achieved, but we note that the convergence rate of $se^2(h)$ can be faster. For example, in the case of bounded joint density, if the residuals are uncorrelated within each cluster, then $se^2(h) = O_P((nh)^{-1})$, as in the i.i.d. case, even if λ_n diverges to infinity.

2.4. **Standard Errors.** Reliable and efficient inference requires a variance estimator that is both consistent and well-behaved in finite samples. Natural estimates of $se^2(h)$ are of the form

$$\widehat{se}^2(h) = \sum_{g\in[G]} \sum_{i,j\in I_g} w_{gi}(h) w_{gj}(h) \widehat{\sigma}_{g,ij} \tag{2.4}$$

with $\widehat{\sigma}_{g,ij}$ being some estimate of $\sigma_{g,ij}$. Setting $\widehat{\sigma}_{g,ij}$ to the product of residuals from the local linear RD regression associated with observations i and j in cluster g yields a clustered regression residual-based standard error, analogous to the cluster-robust standard errors proposed for the linear regression (Liang and Zeger, 1986). We show that this standard error is valid under the same high-level conditions on the weights as those ensuring asymptotic normality.

Even though the regression residual-based standard error is consistent under mild assumptions, it has been long recognized in i.i.d. settings that this approach may be overly conservative in finite samples, and the nearest-neighbors standard error has been proposed to alleviate this issue (Abadie et al., 2014). At its core, the nearest-neighbors approach estimates the conditional variance of the outcome variable for any given observation using the outcome variability among its nearest neighbors.

A direct way to adapt the nearest-neighbors variance estimation approach to clustered data is to set $\widehat{\sigma}_{g,ij}$ in equation (2.4) to

$$\widehat{\sigma}^{\text{naive}}_{g,ij} = Y^{\Delta,\text{naive}}_{gi} Y^{\Delta,\text{naive}}_{gj}, \quad Y^{\Delta,\text{naive}}_{gi} = \sqrt{\frac{|\mathcal{N}^{\text{naive}}_{gi}|}{1+|\mathcal{N}^{\text{naive}}_{gi}|}} \left(Y_{gi} - \frac{1}{|\mathcal{N}^{\text{naive}}_{gi}|} \sum_{(g',i')\in\mathcal{N}^{\text{naive}}_{gi}} Y_{g'i'} \right),$$

where $\mathcal{N}^{\text{naive}}_{gi}$ is the set of J units that are closest to unit i in cluster g in terms of the running variable.[2] However, this procedure does not take into account the clustering structure, and there is no reason to expect that such a variance estimator is consistent in general. First, if the choice of neighbors is associated with cluster membership, then an additional bias term can be present. Second, the bias-correction factor is devised for variance estimation, but it turns out it is not suitable for covariance estimation. We give two illustrative examples showing these problems in Section 5.1.1.

[2]A standard error of that type was proposed by Calonico et al. (2019, Supplemental Appendix 7.10) and is implemented in the software package `rdrobust`. We note that this approach also assumes that residuals are uncorrelated across observations on opposite sides of the cutoff.

To remedy the deficiencies of the naive approach, we propose a different clustered nearest-neighbors standard error where $\widehat{\sigma}_{g,ij}$ is set to

$$\widehat{\sigma}^{\text{CNN}}_{g,ij} = Y^{\Delta_1}_{gi} Y^{\Delta_2}_{gj}, \quad Y^{\Delta_d}_{gi} \equiv Y_{gi} - \frac{1}{|\mathcal{N}^d_{gi}|} \sum_{(g',i') \in \mathcal{N}^d_{gi}} Y_{g'i'} \text{ for } d \in \{1,2\},$$

where $\mathcal{N}^1_{gi}$ and $\mathcal{N}^2_{gi}$ are carefully chosen sets of neighbors of observation i in cluster g. Crucially, we require that the observations in $\cup_{i \in I_g} \mathcal{N}^1_{gi}$ and $\cup_{i \in I_g} \mathcal{N}^2_{gi}$ belong to two disjoint sets of clusters not including g. Under this requirement and additional assumptions we show this standard error is consistent, and one can see in simulations that our procedure has favorable finite-sample properties, exhibiting a smaller bias relative to the regression residual-based approach is settings where the curvature of μ is substantial.

3. ASYMPTOTIC NORMALITY UNDER HIGH-LEVEL CONDITIONS

In this section, we show asymptotic normality of the local linear RD estimator under high-level assumptions on the weights.

3.1. **High-Level Conditions.**

The first assumption controls the cluster sizes in terms of the weights assigned to units in each cluster.

Assumption 1.

(i) $\max_{g \in [G]} \sum_{i,j \in I_g} |w_{gi}(h) w_{gj}(h)| / se^2(h) = o_P(1),$

(ii) $\sum_{g \in [G]} \sum_{i,j \in I_g} |w_{gi}(h) w_{gj}(h)| / se^2(h) = O_P(1).$

Assumption 1 puts restrictions on the size of the terms $\sum_{i,j \in I_g} |w_{gi}(h) w_{gj}(h)|$. In our asymptotic analysis, it is used to control the contributions of different clusters to the conditional variance $se^2(h)$. This assumption allows the number of non-zero weights within each cluster to grow with the sample size, as long as none of the clusters dominates all others in terms of the sums of the absolute value of cross-products of the weights. This assumption is very general and we provide a range of different asymptotic frameworks in which it is satisfied in Section 4. However, we note that it is a sufficient condition to obtain asymptotic normality, but it is not necessary. In particular, it does not cover settings where the number of clusters does not increase with the sample size. For example, if there is one cluster of weakly-dependent data (e.g., strongly mixing process, as extensively studied in time-series settings), one might still obtain asymptotic normality, but part (i) of the assumption will generally not hold.

Remark 2. Assumption 1 does not require the weights $w_{gi}(h)$ to be derived via a local linear regression. It is compatible with any estimator that is a weighted mean of outcomes where the weights do not depend on the outcome variable, e.g., higher-order local polynomial estimators or the optimized RD estimators (Imbens and Wager, 2019; Ghosh et al., 2025).

The second assumption controls the moments of the error terms and is standard for results invoking central limit theorems with non-i.i.d. data.

Assumption 2. $\mathbb{E}[|\varepsilon_{gi}|^{\delta}|\mathcal{X}_g]$ *is uniformly bounded for some* $\delta > 2$ *for all* $g \in [G]$ *and* $i \in I_g$.

3.2. **Asymptotic Normality.** Our first main result establishes asymptotic normality of the RD estimator. To control its bias, we introduce the Hölder-type class of real functions that are potentially discontinuous at zero, are twice differentiable on either side of the threshold, and whose second derivatives are uniformly bounded by some constant $M > 0$:

$$\mathcal{F}_H(M) = \{f_1(x)\mathbf{1}\{x \geq 0\} - f_0(x)\mathbf{1}\{x < 0\} : \|f_w''\|\infty \leq M, w = 0, 1\}.$$

Theorem 1. *(i) Suppose that Assumptions 1 and 2 hold. Then, conditional on* $\mathcal{X}_n$,

$$\frac{\widehat{\tau}(h) - \mathbb{E}[\widehat{\tau}(h)|\mathcal{X}_n]}{se(h)} \xrightarrow{d} \mathcal{N}(0, 1).$$

(ii) For any $M \geq 0$,

$$\sup_{\mu \in \mathcal{F}_H(M)} \left|\mathbb{E}[\widehat{\tau}(h)|\mathcal{X}_n] - \tau\right| \leq \bar{b}(h) \equiv -\frac{M}{2} \sum_{g \in [G]} \sum_{i \in I_g} w_{gi}(h) X_{gi}^2 sign(X_{gi}).$$

Part (i) of the theorem shows that $\widehat{\tau}(h)$, appropriately recentered and rescaled, is asymptotically normally distributed. Part (ii) shows that the conditional bias is bounded by a quantity that is the product of the bound on the second derivative of the conditional expectation function and an expression that depends only on the weights and the running variable. The bias bound is the same as in the i.i.d. setting (Armstrong and Kolesár, 2020; Noack and Rothe, 2024) since the local linear RD estimator is a linear operator and its conditional expectation is not affected by the dependence across outcomes.

Given the standard deviation of the local linear RD estimator and a bound on its bias, we will consider the conditional worst-case mean squared error over data-generating process with $\mu \in \mathcal{F}_H(M)$:

$$\overline{MSE}(h) = \bar{b}(h)^2 + se^2(h).$$

We will explicitly derive the limit of $\overline{MSE}(h)$ and characterize the corresponding asymptotically optimal bandwidth under low-level conditions in the next section.

4. ASYMPTOTIC FRAMEWORKS

Our high-level Assumption 1 introduced in Section 3 accommodates a wide range of empirical clustering structures. To illustrate this flexibility, we formalize four asymptotic frameworks and provide low-level conditions under which the high-level assumption holds. Each of the asymptotic frameworks is calibrated to a class of empirical RD applications. They differ in (i) how quickly the number of near-cutoff units can grow inside a cluster, (ii) the dependence structure of the running variable within clusters, and (iii) assumptions on the conditional covariance matrix of the outcome.

Asymptotic Frameworks I and *II* are motivated by empirical applications where there are many clusters that can be potentially large, but the number of units from any cluster within the estimation windows is relatively small. They cover two distinct dependence patterns in the running variable. While Asymptotic Framework I assumes that the joint distribution of the running variable within each cluster admits a bounded joint density, Asymptotic Framework II does not put any restrictions on this distribution at the cost of imposing stronger restrictions on the rates of the cluster sizes. *Asymptotic Frameworks III* and *IV* relax the restrictions on cluster sizes imposed in Asymptotic Frameworks I and II, respectively, while instead requiring additional assumptions on the covariance structure of the outcome residuals. We discuss representative empirical examples of each of the asymptotic frameworks in Section 6.2.

4.1. General Assumptions.

We will first present general assumptions that are maintained in all the four asymptotic frameworks.

Assumption 3. *(i) The marginal distribution of X_{gi} is the same for all $g \in [G]$ and $i \in I_g$, and it admits a continuous density f_X that is bounded away from zero around the cutoff. (ii) The kernel function k is a bounded and symmetric density function with bounded support, say $[-1, 1]$. (iii) $h \to 0$ and $nh \to \infty$.*

Part (i) of Assumption 3 matches standard conditions for local polynomial estimation with continuously distributed regressors (Fan and Gijbels, 1996). We impose continuity (and continuity at the cutoff) only to obtain simple closed-form variance limits; our high-level results can also accommodate discrete, mixed, or cutoff-discontinuous running variables under suitable modifications to the variance calculations. In the same spirit, we assume the density is continuous at the cutoff to simplify the formulas; formally, an RD analysis can proceed without this requirement, and our main results remain unaffected if the running variable is discontinuous at the threshold.

The kernel and bandwidth requirements in parts (ii) and (iii) are standard in the nonparametric regression literature. We note that the assumption that the bandwidth

shrinks to zero is not necessary for our high-level conditions to hold. We could accommodate a fixed bandwidth; we maintain the usual condition that the bandwidth shrinks to zero solely to obtain closed-form expressions for the leading bias and variance terms. In the asymptotic results, we will use the following kernel constants. For $j \in \mathbf{N}$, let $\bar{\mu}_j = \int_0^1 k(v)v^j dv$. Further, define $\bar{\mu} = (\bar{\mu}_2^2 - \bar{\mu}_1\bar{\mu}_3)/(\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2)$ and $\bar{\kappa} = \int_0^1 \bar{k}(v)^2 dv$, where $\bar{k}(v) = k(v)(\bar{\mu}_2 - \bar{\mu}_1 v)/(\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2)$.

Assumption 4. *(i)* $\mathbb{E}[|\varepsilon_{gi}|^2|\mathcal{X}_g]$ *is uniformly bounded for all* $g \in [G]$ *and* $i \in I_g$*. (ii) The minimal eigenvalue of the conditional covariance matrix of cluster* g*,* $\lambda_{min}(\Sigma_g)$*, is bounded away from zero uniformly in* n *and* $g \in [G]$*.*

Part (i) is slightly weaker than Assumption 2. Part (ii) is imposed to rule out degenerate conditional dependence structures in the outcome variable, such as situations where two residuals within the same cluster are perfectly negatively correlated. The variance of the RD estimator could collapse to zero in such cases, precluding asymptotic normality.

4.2. **Small Effective Cluster Sizes.** The first two asymptotic frameworks apply to settings where the number of units from any cluster within the estimation window is relatively small.

4.2.1. *Assumptions.* The number of units within the estimation window from any given cluster depends on the total number of units in the cluster as well as the joint distribution of the running variable within the cluster. In Asymptotic Framework I, we assume that the joint distribution of the realizations of the running variable admits a bounded joint density, while Asymptotic Framework II leaves the joint distribution unrestricted.

Assumption AF-I. (i) The joint density of $(X_{gi_1}, \ldots, X_{gi_k})$ is bounded uniformly in $n \in \mathbb{N}$, $g \in [G]$, and $1 \leq i_1 < \ldots < i_k \leq n_g$ for $k \in \{2, \ldots, n_g\}$. (ii) $\lambda_n = O(1)$. (iii) $\dfrac{\max_{g\in[G]} n_g^2 h^2 + \log^2 G}{nh} = o(1)$.

Assumption AF-I(i) excludes cases of perfect within-cluster correlation in the running variable where all units share the same realization of the running variable; such extreme dependence is allowed under Assumption AF-II below at the cost of more restrictive rate conditions. Assumption AF-I is similar to Assumption 2 of Hansen and Lee (2019), who study the convergence of the average of clustered observations and other regular, full-sample estimators. Since we consider estimation using the data close to the cutoff, our conditions are formulated in terms of the "local sample size" nh and the "local cluster sizes" $n_g h$, rather than the full sample size n and the cluster sizes n_g used by Hansen and Lee (2019). Another conceptual difference is that part (iii) includes an additional $\log G$ term. This extra factor is due to the randomness of cluster sizes within the estimation

window in our framework, whereas Hansen and Lee (2019) consider the setting of regular parameters and use all observations within each cluster. We note that this asymptotic framework shares some similarities with the framework of Shimizu (2025) specialized to one continuous covariate. He also assumes that the joint distribution of the realizations of the running variable admits a joint density (albeit only for subsets of four observations) and $\lambda_n = O(1)$. However, he imposes a restrictive assumption that $\max_{g\in[G]} n_g h = O(1)$, whereas our framework allows this quantity to diverge.

Assumption AF-II. (i) $\lambda_n = O(h)$. (ii) $\dfrac{\max_{g\in[G]} n_g^2}{nh} = o(1)$.

Assumption AF-II imposes no restrictions on the joint density of the running variable, but the rate conditions on the cluster sizes are more stringent than in Assumption AF-I. In particular, it allows for the realizations of the running variable to be equal within each cluster. It turns out that this extreme case determines the restrictions on the cluster sizes that are necessary to verify Assumption 1.

To illustrate the restrictions imposed in Assumptions AF-I and AF-II, we consider two examples that differ in the degree of allowed heterogeneity in cluster sizes. We revisit these examples in the next subsection to facilitate comparisons between asymptotic frameworks.

Example 1. Suppose that all clusters are of the same size, i.e., $n_g = n/G$ for all $g \in [G]$, then $\lambda_n = O(n_1 h)$. The rate conditions in Assumption AF-I reduce to $n_1 h = O(1)$ and $\log^2 G/(nh) = o(1)$ and in Assumption AF-II to $n_1 = O(1)$ and $Gh \to \infty$.

We note that in Example 1, the expected number of units within the estimation window remains bounded in any given cluster. The next example shows that both Asymptotic Frameworks I and II allow the maximal expected number of observations within the estimation window to diverge for some clusters.

Example 2. Let $a \geq b$ for $a, b \in (0, 1)$. Suppose that we observe $G_1 = n - \lfloor n^a \rfloor$ clusters of size 1 and $G_2 = \lfloor n^b \rfloor$ clusters of size $\lfloor n^{a-b} \rfloor$. Further, for illustration, assume that $h = n^{-b}$. Then $\lambda_n = O(n^{-b} + n^{2(a-b)-1})$. The rate conditions of Assumption AF-I hold if $a \leq b + \frac{1}{2}$. In addition, if $a > 2b$, then each large cluster contributes a diverging number of units local to the cutoff, since $\max_{g\in[G]} n_g h \asymp n^{a-2b} \to \infty$. The rate conditions of Assumption AF-II holds provided $a < \frac{1}{2}(1 + b)$.

4.2.2. *Theoretical results.* The following proposition presents our key theoretical results for Asymptotic Frameworks I and II.

Proposition 1. *Suppose that Assumptions 3, 4, and either Assumption AF-I or AF-II holds. Then Assumption 1 is satisfied,* $se^2(h) \asymp_p (nh)^{-1}$, *and* $\bar{b}(h) = -M\bar{\mu}h^2(1 + o_P(1))$.

Proposition 1 verifies our high-level condition on the weights, shows that the conditional variance of the RD estimator is of order $(nh)^{-1}$, and it characterizes the leading term of the conditional worst-case bias $\bar{b}(h)$. It follows that worst-case asymptotic mean squared error is of order $h^4 + (nh)^{-1}$, which is minimized for bandwidths of order $n^{-1/5}$. With such a bandwidth, the estimator converges at the rate $n^{-2/5}$, given that our assumptions hold for this bandwidth choice. We note that the order of the variance is the same as in the i.i.d. case, but its exact form is in general affected by clustering; we derive its limit under additional assumptions in Section 4.4.

4.3. **Large Effective Cluster Sizes.** While Asymptotic Frameworks I and II cover many relevant clustering patterns, the conditions on the growth rates of the cluster sizes might be restrictive in some applications. In this section, we show these rate conditions can be significantly relaxed under direct assumptions on the standard error. We then argue in Section 4.4 that these assumptions are reasonable in many settings.

4.3.1. *Assumptions.* Asymptotic Framework III considers cases where the joint distribution of the realizations of the running variable is continuous, as in Asymptotic Framework I, but it relaxes the rate conditions imposed on the cluster sizes. This asymptotic framework is motivated by settings where the clusters have a large number of units even in a shrinking neighborhood of the cutoff.

Assumption AF-III. (i) The joint density of $(X_{gi_1}, \ldots, X_{gi_k})$ is bounded uniformly in $n \in \mathbb{N}$, $g \in [G]$, and $1 \leq i_1 < \ldots < i_k \leq n_g$ for $k \in \{2, \ldots, n_g\}$.

(ii) $\dfrac{\lambda_n}{nh} + \dfrac{\max_{g\in[G]}(n_g h)^2 + \log^2 G}{nh(1+\lambda_n)} = o(1)$.

(iii) $se^2(h) \asymp_p (1+\lambda_n)/(nh)$.

Part (i) of Assumption AF-III coincides with part (i) of Assumption AF-I. However, the rate conditions on cluster sizes in part (ii) are considerably weaker than those imposed in parts (ii) and (iii) of Assumption AF-I. In particular, λ_n is allowed to diverge to infinity within this framework.

Asymptotic Framework IV applies to settings where each cluster may contain many units whose realizations of the running variable might be highly correlated. As a result, even within a shrinking neighborhood of the cutoff, some clusters can contribute a large number of units concentrated in a narrow region of the support of the running variable. Such settings occur naturally, for example, if the outcomes are measured at the individual level, whereas the running variable is assigned at the cluster level.

Assumption AF-IV. (i) $\dfrac{\lambda_n}{nh} + \dfrac{\max_{g\in[G]} n_g^2}{\sum_{g\in[G]} n_g^2} = o(h)$. (ii) $se^2(h) \asymp_p (1+\lambda_n/h)/(nh)$.

The rate conditions on cluster sizes are significantly weaker than those in Assumption AF-II as we combine these conditions with an additional assumption on the standard error. We note that the rate conditions in Assumption AF-IV are stronger than those in Assumption AF-III; this is needed because we do not impose any assumptions on the joint distribution of the running variables. We next illustrate the implications of these rate conditions in our examples studied above.

Example 1 (Equal Cluster Sizes, cont'd)**.** Assumption AF-III requires that $G \to \infty$ and $n^2h^2/G \to \infty$. In this setting, $se^2(h) = O\left(1/(nh) + 1/G\right)$. Part (i) of Assumption AF-IV reduces to the requirement that $Gh \to \infty$. We emphasize that there is no separate restriction on the cluster size. Part (ii) implies that $se^2(h) = O(1/(Gh))$.

Example 2 (Heterogeneous Cluster Sizes, cont'd)**.** The rate conditions of Assumption AF-III do not impose any additional restrictions. Moreover, if $a > 2b$, at least one large cluster is locally influential, in the sense that $\max_{g\in[G]} n_g h \asymp n^{a-2b} \to \infty$. The rate conditions of Assumption AF-IV impose the same restrictions as in Asymptotic Framework II. It has to hold that $a < \frac{1}{2}(1+b)$.

4.3.2. *Theoretical results.* The following proposition presents our key theoretical results for Asymptotic Frameworks III and IV.

Proposition 2. *Suppose that Assumptions 3 and 4 hold, and either Assumption AF-III or AF-IV holds. Then Assumption 1 is satisfied and* $\bar{b}(h) = -M\bar{\mu}h^2(1 + o_P(1))$.

Proposition 2 verifies our high-level Assumption 1 and characterizes the leading term of the conditional worst-case bias $\bar{b}(h)$. We note that the rate of the optimal bandwidth and the resulting convergence rate of the estimator are different than in the i.i.d. case or the case of small effective cluster sizes discussed in the previous section. Specifically, in Framework III, we have that

$$\widehat{\tau}(h) - \tau = O_P\left(h^2 + \frac{1}{\sqrt{nh}} + \sqrt{\frac{\sum_{g\in[G]} n_g^2}{n^2}}\right).$$

In contrast to Asymptotic Framework I, the third term in the remainder on the right-hands side can dominate the other terms. Since the bandwidth h appears only in the first two terms, the convergence rate is optimized if $h \sim n^{-1/5}$, in which case we obtain $\widehat{\tau}(h) - \tau = O_P(n^{-2/5} + \big(\sum_{g\in[G]} n_g^2/n^2\big)^{1/2})$. We note that if $n^{-2/5} = o\big(\big(\sum_{g\in[G]} n_g^2/n^2\big)^{1/2}\big)$ and $h \sim n^{-1/5}$, then the variance dominates the bias under the optimal bandwidth choice, such that

$$\frac{\widehat{\tau}(h) - \tau}{se(h)} \xrightarrow{d} \mathcal{N}(0,1),$$

assuming that Assumption AF-III holds for this bandwidth choice.

In Framework IV, in turn, we have that

$$\widehat{\tau}(h) - \tau = O_P\left(h^2 + \frac{1}{\sqrt{nh}}\left(1 + \sqrt{\frac{\sum_{g\in[G]} n_g^2}{n}}\right)\right).$$

The fastest possible convergence rate is achieved for $h \sim \left(\left(1+\sum_{g\in[G]} n_g^2/n\right)/n\right)^{1/5}$, yielding $\widehat{\tau}(h) - \tau = O_P\big(n^{-2/5} + \big(\sum_{g\in[G]} n_g^2/n^2\big)^{2/5}\big)$, given that this bandwidth choice satisfies Assumption AF-IV.

4.4. **Standard Error and Bandwidth Choice in a Special Case.** To illustrate the behavior of the local linear RD estimator with clustered data, we will further consider a simplified setup, where exact characterization of the limit of the variance is possible. The following assumption formalizes the setup.

Assumption 5. *For all $g \in [G]$ and $i \in I_g$, $\sigma_{i,g}^2 = \sigma^2(X_{gi})$ and $\sigma_{ij,g} = \sigma(X_{gi}, X_{gj})$ for some functions $\sigma^2(\cdot)$ and $\sigma(\cdot,\cdot)$ that are L-Lipschitz continuous away from the cutoff.*

Assumption 5 imposes a common covariance between any two units and across clusters, and it imposes continuity of the conditional variance and covariance functions.[3]

4.4.1. *Continuous Joint Distribution of the Running Variable.* We first study the standard error under Asymptotic Frameworks I and III.

Lemma 1. *Suppose that Assumptions 3, 4(i), and AF-III(i)-(ii) hold. Then*

(i) $se^2(h) = O_P\left(\dfrac{1+\lambda_n}{nh}\right)$.

(ii) If in addition Assumption 5 holds and (X_{gi}, X_{gj}), $i \neq j$, $i,j \in I_g$, are identically distributed with continuous joint density $f(x_1,x_2)$, then

$$se^2(h) = \frac{1}{nh}\left(\bar{\kappa}\sum_{\star\in\{+,-\}}\frac{\sigma^2(0^\star)}{f_X(0)} + \lambda_n \sum_{\star,\diamond\in\{+,-\}} \mathbf{1}^{\pm}_{\{\star=\diamond\}}\sigma(0^\star,0^\diamond)\frac{f(0,0)}{f_X(0)^2} + o_P\left(1+\lambda_n\right)\right),$$

where $\mathbf{1}^{\pm}_{\{\star=\diamond\}} = 1$ if $\star = \diamond$ and $\mathbf{1}^{\pm}_{\{\star=\diamond\}} = -1$ otherwise.

Lemma 1 imposes only the weak rate conditions on the cluster sizes of Asymptotic Framework III, which, in particular, cover Asymptotic Framework I. First, the lemma provides an upper bound on the convergence rate of the conditional variance $se^2(h)$, and then it derives its exact limit in a special case. Part (ii) demonstrates that the conditional

[3]This type of assumption can be rationalized by models studied in the functional data literature (see, e.g., Zhang and Chen, 2007). Shimizu (2025) also imposes such an assumption.

variance is asymptotically equal to the sum of two terms. The first one is driven by the variances of individual units and is the same as in the i.i.d. case. The second component is due to the covariance between outcomes within a cluster; we note this part depends neither on the bandwidth nor the choice of the kernel function. If λ_n converges to a positive constant, then the two terms are of the same order. If λ_n converges to zero, then the effect of clustering on variance becomes asymptotically negligible; a similar result was obtained by Shimizu (2025) under stronger rate restrictions on the cluster sizes. If λ_n diverges to infinity and $\sum_{\star,\diamond\in\{+,-\}} \mathbf{1}^{\pm}_{\{\star=\diamond\}}\sigma(0^\star,0^\diamond) \neq 0$, the covariance part dominates.

Under the assumptions of part (ii) of Lemma 1, the worst-case mean squared error of $\widehat{\tau}(h)$ satisfies

$$\overline{MSE}(h) = \left(M^2\bar{\mu}^2 h^4 + \frac{V_1}{nh} + \frac{\sum_{g\in[G]} n_g^2}{n^2} \sum_{\star,\diamond\in\{+,-\}} \mathbf{1}^{\pm}_{\{\star=\diamond\}}\sigma(0^\star,0^\diamond)\frac{f(0,0)}{f_X(0)^2} \right)(1+o_P(1)),$$

where $V_1 = \frac{\bar{\kappa}}{f_X(0)}\sum_{\star\in\{+,-\}}\sigma^2(0^\star)$. The bandwidth minimizing the leading term is given by

$$h_1^* = \left(\frac{V_1}{4M^2\bar{\mu}^2}\right)^{1/5} n^{-1/5},$$

given that this bandwidth choice satisfied the assumptions of part (ii) of Lemma 2. The optimal bandwidth h_1^* is the same as the AMSE-optimal bandwidth in the i.i.d. case. We emphasize that this result relies on Assumption 5; under more general covariance structures, it need not hold.

4.4.2. *Degenerate Joint Distribution of the Running Variable.* We now study the standard error under Asymptotic Frameworks II and IV. To derive a closed-form expression for the limit of the conditional variance in this setting, we need to impose an additional assumption on the joint distribution of the running variable. For illustration, we consider the setting where all the realizations of the running variable are the same within cluster.

Lemma 2. *Suppose that Assumptions 3, 4, and AF-IV(i) hold. Then*

(i) $se^2(h) = O_P\left(\dfrac{1+\lambda_n/h}{nh}\right)$

(ii) If in addition Assumption 5 holds and the realizations of the running variable are equal within each cluster. Then

$$se^2(h) = \frac{1}{nh}\frac{\bar{\kappa}}{f_X(0)}\sum_{\star\in\{+,-\}}\left(\sigma^2(0^\star) + \frac{\lambda_n}{h}\sigma(0^\star,0^\star) + o_P\left(1+\frac{\lambda_n}{h}\right)\right).$$

Lemma 2 imposes only the weak rate conditions on the cluster sizes of Asymptotic Framework IV, which, in particular, cover Asymptotic Framework II. First, the lemma

provides an upper bound on the convergence rate of the conditional variance $se^2(h)$, and then it derives its exact limit in a special case.

Under the assumptions of Lemma 2, the worst-case mean squared error of $\widehat{\tau}(h)$ satisfies

$$\overline{MSE}(h) = \left(M^2 \bar{\mu}^2 h^4 + \frac{V_2}{nh} \right) (1 + o_P(1)),$$

where $V_2 = \frac{\bar{\kappa}}{f_X(0)} \sum_{\star \in \{+,-\}} \left(\sigma^2(0^\star) + \sigma(0^\star, 0^\star) \sum_{g \in [G]} n_g(n_g - 1)/n \right)$. In this setting, the bandwidth minimizing the leading term of the AMSE-optimal bandwidth is given by

$$h_2^* = \left(\frac{V_2}{4M^2 \bar{\mu}^2} \right)^{1/5} n^{-1/5},$$

assuming the assumptions of Lemma 2 hold for this bandwidth choice.

5. CLUSTERED STANDARD ERROR

In this section, we study variance estimation based on the nearest-neighbors and regression residual-based approaches.

5.1. **Clustered Nearest-Neighbors Standard Error.** We first show why the naive adaptation of the nearest-neighbors approach devised for i.i.d. settings is in general not valid with clustered data. We then introduce our proposed clustered nearest-neighbors (CNN) standard error and show its consistency.

5.1.1. *Failure of the Naive Clustered Nearest-Neighbors Standard Error.* To highlight the main problems with the naive clustered nearest-neighbors standard error described in Section 2.4, we consider a simple setup with $\mu(X) = 0$, $\mathrm{Var}(Y_{gi}|\mathcal{X}_g) = \sigma^2$, and $J = 1$ nearest neighbor. We discuss two examples that differ in the assumptions on the joint distribution of the realizations of the running variable.

First, suppose that each cluster consists of only two observations and $X_{g1} = X_{g2}$ for all $g \in [G]$, such that $\mathcal{N}_{g1}^{\text{naive}} = \{(g, 2)\}$ and $\mathcal{N}_{g2}^{\text{naive}} = \{(g, 1)\}$. Then

$$\begin{aligned} \widehat{se}^2_{\text{naive}}(h) &= \frac{1}{2} \sum_{g \in [G]} w_{g1}^2(h) \sum_{i,j \in I_g} Y_{gi}^{\Delta,\text{naive}} Y_{gj}^{\Delta,\text{naive}} \\ &= \frac{1}{2} \sum_{g \in [G]} w_{g1}^2(h) \underbrace{\sum_{l=1}^{4} (-1)^l (\varepsilon_{g1} - \varepsilon_{g2})^2}_{=0} = 0. \end{aligned}$$

Clearly, this standard error cannot be consistent. While this is a very specific example, the same type of problem arises more generally whenever the realizations of the running variable are highly concentrated within clusters.

Second, suppose that the joint distribution of the running variable within each cluster admits a bounded joint density. Consider two observations $i, j \in I_g$, $i \neq j$, and let (g_1, i') and (g_2, j') be their respective nearest neighbors. Assume further that the clusters g, g_1, and g_2 are pairwise different, which occurs with high probability in this setup if all clusters are relatively small. Then, while it is easy to see that the conditional variance estimate is correctly centered, $\mathbb{E}\big[\big(Y_{gi}^{\Delta,\text{naive}}\big)^2|\mathcal{X}_n\big] = \sigma^2$, the conditional covariance estimates are biased:

$$\mathbb{E}\Big[Y_{gi}^{\Delta,\text{naive}}Y_{gj}^{\Delta,\text{naive}}|\mathcal{X}_n\Big] = \frac{1}{2}\mathbb{E}[\varepsilon_{gi}\varepsilon_{gj} - \varepsilon_{gi}\varepsilon_{g_2j'} - \varepsilon_{gj}\varepsilon_{g_1i'} + \varepsilon_{g_1i'}\varepsilon_{g_2j'}|\mathcal{X}_n] = \frac{1}{2}\mathbb{E}[\varepsilon_{gi}\varepsilon_{gj}|\mathcal{X}_n].$$

It follows that $\widehat{se}^2_{\text{naive}}(h)$ is not correctly centered in general.[4]

5.1.2. *Our Proposed Clustered Nearest-Neighbors Standard Error.* It is evident from the preceding discussion that selecting neighbors from distinct clusters induces certain independence restrictions that are instrumental for establishing conditional unbiasedness of the standard error. We leverage this insight to construct our proposed standard error, explicitly enforcing the desired independence structure.

For every cluster $g \in [G]$, we define two sets of its "companion clusters", $\mathcal{R}_g^1 \subset [G]$ and $\mathcal{R}_g^2 \subset [G]$. Next, for every $i \in I_g$ and $d \in \{1, 2\}$, we define $\mathcal{N}_{gi}^d$ as the set of at least J nearest neighbors of unit i in cluster g in terms of the running variable that are on the same side of the cutoff as X_{gi} and belong to a cluster in the set $\mathcal{R}_g^d$. Our proposed clustered nearest-neighbors (CNN) standard error is then defined as:

$$\widehat{se}^2_{CNN}(h) = \sum_{g\in[G]}\sum_{i,j\in I_g} w_{gi}(h)w_{gj}(h)Y_{gi}^{\Delta_1}Y_{gj}^{\Delta_2}, \quad Y_{gi}^{\Delta_d} \equiv Y_{gi} - \frac{1}{|\mathcal{N}_{gi}^d|}\sum_{(g',i')\in\mathcal{N}_{gi}^d} Y_{g'i'}. \tag{5.1}$$

To show consistency of this standard error, we require the sets of nearest neighbors to satisfy two properties. First we require that the neighbors in the two sets $\mathcal{N}_{gi}^1$ and $\mathcal{N}_{gi}^2$ are independent of the units in cluster g and of each other. This is achieved by selecting companion clusters that do not include cluster g, $g \notin \mathcal{R}_g^1 \cup \mathcal{R}_g^2$, and are disjoint, $\mathcal{R}_g^1 \cap \mathcal{R}_g^2 = \emptyset$. These properties ensure that our standard error is asymptotically conditionally unbiased.

Second, we need to control the dependence between $\sum_{i,j\in I_g} w_{gi}(h)w_{gj}(h)Y_{gi}^{\Delta_1}Y_{gj}^{\Delta_2}$ across different clusters. We achieve that by imposing that each cluster can be a companion cluster for at most R clusters for some fixed number R, i.e., for all $g \in [G]$, we

[4]We note that the problem arising in the naive conditional covariance estimation can be easily resolved by leaving out the correction factor when $i \neq j$. However, a complete proof of approximate unbiasedness of the standard error would still require controlling the probability of the respective clusters being distinct across all observations, limiting the applicability of this method to relatively small clusters with a bounded joint density of the realizations of the running variable.

require

$$\#\big\{\tilde{g} \in [G] : g \in \mathcal{R}^1_{\tilde{g}} \cup \mathcal{R}^2_{\tilde{g}}\big\} \leq R. \tag{5.2}$$

These general requirements on the choice of companion clusters can be satisfied by many different selection methods. We provide one concrete algorithm in Appendix A.

Remark 3. We note that $\widehat{se}^2_{CNN}(h)$ does not reduce to the nearest-neighbors standard error that is typically used in i.i.d. settings. The standard nearest-neighbors standard errors replaces the unknown conditional variances with the squares of distances to nearest neighbors and it involves a bias-correction factor. Since we effectively take the product of two different residuals, it turns out that this bias-correction factor is not needed.

5.1.3. *Consistency of the CNN Standard Error.* As is standard for nearest-neighbors-type estimators, our method relies on the assumption that the nearest neighbors selected in the construction of $\widehat{se}^2_{CNN}(h)$ are uniformly close to the respective units.

Assumption 6. $D(h) \equiv \max_{g\in[G]} \max_{\substack{i\in I_g \\ w_{gi}(h)\neq 0}} \max_{(\tilde{g},j)\in\mathcal{N}^1_{gi}\cup\mathcal{N}^2_{gi}} |X_{gi} - X_{\tilde{g}j}| = o_P(1).$

This assumption is automatically satisfied whenever the nearest neighbors are selected within the estimation window, as is the case for the algorithm given in Appendix A, and the bandwidth converges to zero. Since all matched units then lie within distance h of each other, we have $D(h) = O(h)$ by construction. In general, we expect $D(h)$ to converge to zero at a much faster rate. For illustration, consider a setting where the marginal density of X_{gi} is bounded and bounded away from zero in a neighborhood of the cutoff. If the realizations of the running variable are constant within each cluster and $Gh \to \infty$, then we expect that $D(h) = O_P(\log(Gh)/G)$. As a different example, consider a setting where the joint density of $\mathcal{X}_g$ is continuous for all $g \in [G]$ and $n_{\min}h \equiv \min_{g\in[G]} n_g h \to \infty$, then we expect that $D(h) = O_P(\log(n_{min}h)/n_{min})$.

Our second main result states that $\widehat{se}^2_{CNN}(h)$ is consistent for $se^2(h)$.

Theorem 2. *Suppose that Assumptions 1 and 6 hold and Assumption 2 holds with $\delta = 4$. Then*

$$\frac{\widehat{se}^2_{CNN}(h)}{se^2(h)} = 1 + o_{P,\mathcal{F}}(1),$$

where the term $o_{P,\mathcal{F}}(1)$ converges to zero uniformly over any class of DGPs where μ is L-Lipschitz continuous away from the cutoff for some constant L.

Remark 4. In contrast to the standard nearest neighbor variance estimator, that is typically applied in settings of independent samples, we do not need to impose continuity assumptions of the conditional variance function.

Remark 5. It is common practice in RD designs, to impose that μ has a bounded second derivative. If one wants to impose this assumption instead of assuming that μ is L-Lipschitz continuous away from the cutoff for some constant L, we can easily modify the standard error following the suggestions of Noack and Rothe (2024).

5.2. **Clustered Regression Residual-Based Standard Error.** In this subsection, we show consistency of the clustered regression residual-based (CRR) standard error for the local linear RD estimator. For $\star \in \{+,-\}$, define $b_0^\star = \mu(0^\star)$ and $b_1^\star = \mu'(0^\star)$, and let $\hat{b}_0^\star$ and $\hat{b}_1^\star$ denote the intercept and slope coefficient on the respective side of the cutoff in the local linear RD regression in equation (2.2). The CRR standard error is defined as

$$\widehat{se}^2_{CRR} = \sum_{g\in[G]} \sum_{i,j\in I_g} w_{gi}(h) w_{gj}(h) \hat{\varepsilon}_{gi} \hat{\varepsilon}_{gj}, \quad \hat{\varepsilon}_{gi} = Y_{gi} - \widehat{\mu}(X_{gi}),$$

where $\widehat{\mu}(x) = (\hat{b}_0^- + \hat{b}_1^- x)\mathbf{1}\{x < 0\} + (\hat{b}_0^+ + \hat{b}_1^+ x)\mathbf{1}\{0 \le x\}$.

Theorem 3. *Suppose that Assumption 1 holds and Assumption 2 holds with $\delta = 4$. Further, assume that $\mu \in \mathcal{F}_H(M)$, $\hat{b}_0^\star - b_0^\star = o_P(1)$, $h(\hat{b}_1^\star - b_1^\star) = o_P(1)$ for $\star \in \{+,-\}$, the weights $w_{gi}(h)$ are zero whenever $|X_{gi}| > h$, and $h \to 0$. Then*

$$\frac{\widehat{se}^2_{CRR}(h)}{se^2(h)} = 1 + o_P(1).$$

Theorem 3 imposes the same high-level assumption on the weights $w_{gi}(h)$ as we used to show consistency of the CNN standard error. The consistency requirements for $\hat{b}_0^\star$ and $\hat{b}_1^\star$ are very mild and they are satisfied in all the considered asymptotic frameworks given our smoothness assumption. The main difference in assumptions relative to the result for the CNN standard error is that Theorem 3 requires the bandwidth to converge to zero, while the assumptions of Theorem 2 may hold even for an asymptotically fixed bandwidth.

6. NUMERICAL ILLUSTRATIONS

In this section, we apply our clustered nearest-neighbors (CNN) standard error in four empirical applications, and we compare it to four alternative standard errors. The first two, the classical nearest-neighbors (NN) and Eicker-Huber-White (EHW) standard errors, do not account for clustering. The third approach is the naive clustered nearest-neighbors (Naive CNN) approach described in Section 5.1.1, and the fourth is the clustered regression residual-based (CRR) standard error described in Section 5.2.

In order to connect the asymptotic frameworks introduced in Section 4 to observable features of the data, we first propose a simple diagnostic rule of thumb for assessing

whether clusters are sufficiently small and reasonably balanced for Asymptotic Framework I or II to provide accurate approximations to the underlying data-generating process. We then revisit four recent RD applications, each exemplifying one of the asymptotic frameworks.[5]

6.1. **Rule of Thumb for the High-Level Conditions.** In this section, we provide a practical rule of thumb to assess whether our high-level Assumption 1 on the weights can be plausibly satisfied in a given empirical application. When our conditions hold, the data-generating process may be well approximated by Asymptotic Frameworks I or II under mild assumptions on the covariance of the residuals. If, in turn, these conditions are violated, one may need to justify stronger assumptions on the dependence structure of the residuals to ensure that the standard error converges at a suitable rate, as discussed in our Asymptotic Frameworks III and IV.

To describe our proposed criterion, for $g \in [G]$, define

$$w_{g,\text{ratio}}(h) = \frac{\sum_{i,j \in I_g} |w_{gi}(h) w_{gj}(h)|}{\sum_{g \in [G]} \sum_{i \in I_g} w_{gi}(h)^2},$$

and let

$$w_{\max}(h) \equiv \max_{g \in [G]} w_{g,\text{ratio}}(h), \qquad w_{\text{sum}}(h) \equiv \sum_{g \in [G]} w_{g,\text{ratio}}(h).$$

If the conditional covariance matrix of the residuals has eigenvalues bounded away from zero, the conditions $w_{\max}(h) = o_P(1)$ and $w_{\text{sum}}(h) = O_P(1)$ are sufficient to ensure that our high-level Assumption 1 is satisfied. To operationalize these asymptotic conditions, in finite samples, one needs to choose some threshold values $\eta_{\max}$ and η_{sum}, and check whether $w_{\max}(h) \leq \eta_{\max}$ and $w_{\text{sum}}(h) \leq \eta_{\text{sum}}$. We consider $\eta_{\max} = 0.1$ and $\eta_{\text{sum}} = 10$ to be reasonable benchmark values in practice.[6]

6.2. **Empirical Applications.** In this section, we revisit four empirical applications that motivated the asymptotic frameworks introduced in Section 4. Figure 1 illustrates the number of clusters, the distribution of cluster sizes, and the distribution of the running variable within a cluster. Each point represents an individual observation. For discrete

[5]For each application, we use the data provided in the respective replication package and consider one of the main RD specifications reported in the paper. For simplicity, we ignored any additional covariates that were included to improve estimation precision. We fix the bandwidth at the value used by the authors of the original study. To simplify the analysis, when the original paper used two different bandwidths on each side of the cutoff, we chose the bigger one.

[6]To give a heuristic interpretation of these threshold values, we note that, asymptotically, $w_{\max}(h) \approx \max_{g \in [G]} n_{g,h}^2 / n_h$ and $w_{\text{sum}}(h) \approx \sum_{g \in [G]} n_{g,h}^2 / n_h$, where $n_{g,h}$ and n_h denote the number of units from cluster g and the total number of units within the estimation window, respectively. Now suppose that within the estimation window, there are 100 clusters with 10 observations each, a setting in which asymptotic normality can plausibly be a good approximation. Then $w_{\max}(h) \approx 0.1$ and $w_{\text{sum}}(h) \approx 10$.

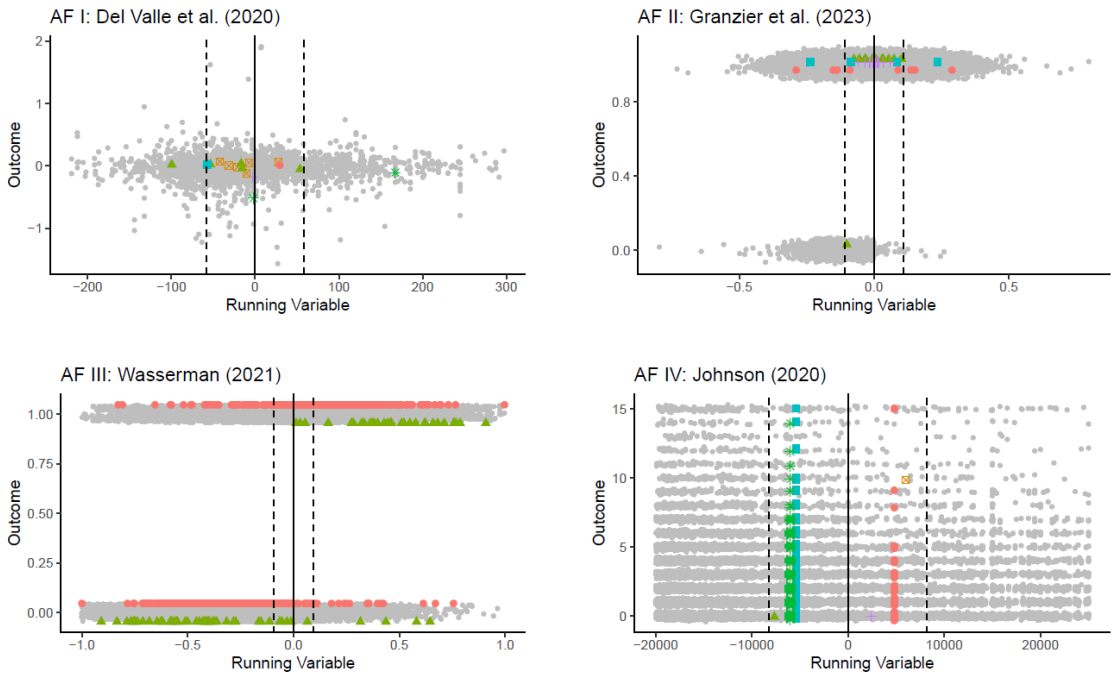

Figure 1: Visualization of cluster structures in the empirical applications.

Notes: Each point represents one observation. For discrete outcome variables, values are jittered to avoid overplotting. For selected clusters, all units are displayed in the same color and use the same marker symbol. The running variables, outcomes, and cutoffs are described in Section 6.2.

outcome variables, values are jittered to improve visual clarity and mitigate overplotting. In selected clusters, all observations are displayed in a common color and use the same marker symbol to emphasize cluster membership. The dashed lines mark the bandwidths used.

6.2.1. *Motivating Example for Asymptotic Framework I.* del Valle et al. (2020) study the impact of Mexico's indexed disaster fund (Fonden) on post-disaster economic recovery. The outcomes are constructed as changes in log night lights in the year following a disaster, measured using satellite-based night lights at the municipality level. They leverage a fuzzy regression discontinuity design where the eligibility for disaster transfers depended on whether the realized rainfall exceeded a pre-specified cutoff. The data contains information on municipal requests for Fonden funding in the period between 2004 and 2012. The authors cluster the standard errors at the municipality level. The estimation window contains around 1000 municipalities with an average of 1.5 requests per municipality.

6.2.2. *Motivating Example for Asymptotic Framework II.* Granzier et al. (2023) study French two-round elections to evaluate how candidates' first-round ranking affects their second-round results. Specifically, they measure the impact of barely achieving a higher rank on remaining in the race or winning, and the running variable is the first-round vote

margin between adjacent candidates. For concreteness, we focus on the effect of ranking 1st vs 2nd in the first round on the probability of running in the second round. The data contain information on electoral races from several decades of local and parliamentary elections. The standard errors are clustered at the district level. There are about 2300 clusters in the estimation window, with an average of 3 observations per cluster. By construction, the realizations of the running variable are symmetric around the cutoff: for every observation with running variable value X_{gi}, there is a corresponding observation with running variable value $-X_{gi}$. Such degenerate distributions of the running variable are allowed under our Asymptotic Framework II.

6.2.3. *Motivating Example for Asymptotic Framework III.* Wasserman (2021) studies the causal effect of an electoral defeat on subsequent political participation using a close-election RD design. The analysis focuses on first-time candidates for US state legislative offices. The running variable is the candidate's margin of victory, and the primary outcome of interest is whether the candidate runs again for any state legislative office within four years of the initial run. The data covers state legislative elections over several decades across the United States, and the standard errors are clustered at the state level. This yields clusters with a large number of observations spread across the support of the running variable, with approximately 250 observations per state on average within a local neighborhood of the cutoff.

6.2.4. *Motivating Example for Asymptotic Framework IV.* Johnson (2020) studies the deterrence effects of a policy under which the Occupational Safety and Health Administration (OSHA) issues press releases about violations of workplace safety and health regulations that exceed a penalty threshold. In this RD design, the running variable is the penalty amount assigned at inspection, with a discontinuity at the press-release threshold. The primary outcome is the count of violations recorded in subsequent inspections. In the dataset, the facilities are organized into "peer groups"–groups of facilities in the same sector located within a 5 km radius of a facility where a penalty was levied—such that all facilities within a peer group share the same penalty value, while the outcome is measured at the facility level. The standard errors are clustered at the peer group level. There are 707 peer groups of facilities, each containing roughly 16 facilities on average in a local neighborhood around the cutoff.

6.2.5. *Empirical Results.* The results for all four empirical applications are reported in Table 1. Before discussing the values of the different standard errors, we first note that, according to the rule-of-thumb diagnostics reported in the last two columns of the table, the studies of del Valle et al. (2020) and Granzier et al. (2023) represent settings with relatively small and fairly balanced cluster sizes. These settings appear to fit into our Asymptotic Frameworks I and II well, suggesting that asymptotic normality is likely a

Table 1: Empirical Results

	Est	IID SE		Clustered SE			Cluster Sizes			
		EHW	NN	Naive CNN	CRR	CNN	n_h	G_h	$w_{\max}$	w_{sum}
AF I: del Valle et al.	0.059	0.021	0.023	0.022	0.023	0.022	1563	1014	0.03	1.72
AF II: Granzier et al.	0.146	0.041	0.039	0.040	0.036	0.036	2338	903	0.06	2.89
AF III: Wasserman	0.507	0.015	0.015	0.022	0.025	0.024	12 140	50	68.46	239.22
AF IV: Johnson	−0.244	0.120	0.123	0.176	0.244	0.241	11 346	707	11.04	55.77

Notes: Column *Estimate* reports the RD estimate; it can differ slightly from the one reported in the original paper since, for example, we do not include additional covariates used to improve precision in their regressions. *EHW* and *NN* standard errors do not account for clustering. *Naive CNN* selects the nearest neighbors as in the i.i.d. case as implemented in the software `rdrobust`. *CRR* is the clustered regression-residual based approach, and *CNN* is our proposed standard error. n_h and G_h denote the number of observations and the number of clusters within the estimation window; $w_{\max}$ and w_{sum} are the rule-of-thumb measures described in Section 6.1. The data sets are described in Section 6.2.

reasonable approximation to the finite-sample distribution of the RD estimator. By contrast, in the applications studied by Wasserman (2021) and Johnson (2020), cluster sizes are either large or markedly unbalanced. In such cases, justifying asymptotic normality requires an additional assumption on the convergence rate of the conditional variance $se^2(h)$; Assumption 5 provides one illustrative sufficient condition for it to hold.

The standard errors computed under the assumption of i.i.d. sampling are substantially smaller than their clustered counterparts in applications with large clusters, which suggest that they fail to account for relevant within-cluster dependence. Our proposed CNN standard error is close in magnitude to the conventional CRR approach, while the Naive CNN standard error is markedly smaller in the fourth application, consistent with the discussion in Section 5.1.1.

7. CONCLUSION

This paper proposes a general framework for sharp RD designs with clustered data. Under general high-level conditions, we establish the asymptotic normality of the local linear RD estimator, and we illustrate the high-level conditions in empirically motivated asymptotic frameworks. Furthermore, we develop a novel nearest-neighbors-type standard error tailored to clustered samples. Our approach is easily extendable: with minor modifications, it readily accommodates fuzzy RD and kink designs as well as settings with covariate adjustments.

Appendix

A. COMPANION CLUSTERS SELECTION ALGORITHM

There are several approaches for selecting the sets $\mathcal{R}_g^1$ and $\mathcal{R}_g^2$ that satisfy our high-level conditions described in Section 5.1.2. The effectiveness of a given algorithm, however, depends on the specific empirical context. To provide a concrete illustration, we propose one specific algorithm which fulfills these high-level assumptions in a broad class of empirically relevant settings.

For $\star \in \{+,-\}$, let $\mathcal{X}_{g,h}^\star$ denote the set of distinct realizations of the running variable in cluster g within the estimation window on the respective side of the cutoff, and define $l_{g,h}^\star$ as the number of elements in $\mathcal{X}_{g,h}^\star$. Our proposed selection procedure is given in Algorithm 1.

Algorithm 1 Selection of Companion Clusters $\mathcal{R}_g^1$ and $\mathcal{R}_g^2$

Inputs: $(\mathcal{X}_g)_{g\in[G]}$, R, and J.

Support Dimension Reduction: Let $L \equiv \lfloor R/(4J) \rfloor$. For all $g \in [G]$ and $\star \in \{-,+\}$, define $\mathcal{S}_g^\star$ as follows: If $l_{g,h}^\star \leq L$ then $\mathcal{S}_g^\star = \mathcal{X}_{g,h}^\star$. Otherwise, let $\mathcal{S}_g^\star$ be the empirical quantiles of $\mathcal{X}_{g,h}^\star$ evaluated at the probabilities $\left\{0, \frac{1}{L-1}, \frac{2}{L-1}, \ldots, 1\right\}$.[7]

Choice of Companion Clusters: For $g \in [G]$:

Step 1: For $\star \in \{-,+\}$ and each $x \in \mathcal{S}_g^\star$, find the J closest values in $\bigcup_{\tilde{g}\in[G]\setminus\{g\}} \mathcal{S}_{\tilde{g}}^\star$ and let $r_g^1(x)$ denote the set of clusters these values belong to. Define:

$$\mathcal{R}_g^1 = \bigcup_{x\in\mathcal{S}_g^-\cup\mathcal{S}_g^+} r_g^1(x).$$

Step 2: For $\star \in \{-,+\}$ and each $x \in \mathcal{S}_g^\star$, find the J closest values in $\bigcup_{\tilde{g}\in[G]\setminus\{g\}\setminus\mathcal{R}_g^1} \mathcal{S}_{\tilde{g}}^\star$, and let $r_g^2(x)$ denote the set of clusters these values belong to. Define:

$$\mathcal{R}_g^2 = \bigcup_{x\in\mathcal{S}_g^-\cup\mathcal{S}_g^+} r_g^2(x).$$

The above algorithm can be applied if there are at least $2JL$ clusters within the bandwidth on each side of the cutoff. This condition is in line with all the asymptotic frameworks we consider in Section 4, where the number of clusters in the local neighborhood of the cutoff diverges to infinity.

[7] If the running variable has mass points, we jitter the elements of $\mathcal{S}_g^\star$ by adding small independent noise.

Lemma A.1. *Suppose that Algorithm 1 is used. Then for any $g \in [G]$,*

$$\#\left\{\tilde{g} : g \in \mathcal{R}^1_{\tilde{g}} \cup \mathcal{R}^2_{\tilde{g}}\right\} \leq R.$$

Proof. Each value in $\mathcal{S}^-_g \cup \mathcal{S}^+_g$ can be the J-th or closer nearest neighbor for at most $2J$ support points from other clusters. Since $|\mathcal{S}^-_g \cup \mathcal{S}^+_g| \leq 2L$, cluster g can be selected as a companion cluster at most $4LJ \leq R$ times. □

Lemma A.1 guarantees that if Algorithm 1 is used to define the companion clusters, then each cluster in the sample will be "used" at most R times for the prespecified value R. This shows that the general construction described in Subsection 5.1.2 is feasible.

B. PROOFS OF THEOREMS 1–3

B.1. **Proof of Theorem 1.** Let $W_g = \sum_{i \in I_g} w_{gi}(h)(Y_{gi} - \mathbb{E}[Y_{gi}|\mathcal{X}_n])$. It holds that $\mathbb{E}[W_g|\mathcal{X}_n] = 0$ and

$$se^2(h) = \sum_{g \in [G]} \mathrm{Var}(W_g|\mathcal{X}_n).$$

Let $\delta > 2$ be as in Assumption 2. We will verify Lyapunov's condition by showing that

$$A_n \equiv \frac{1}{se(h)^\delta} \sum_{g \in [G]} \mathbb{E}[|W_g|^\delta|\mathcal{X}_n] \xrightarrow{n\to\infty} 0.$$

First, note that

$$\begin{aligned} A_n &= \frac{\sum_{g\in[G]} \mathbb{E}[|\sum_{i\in I_g} w_{gi}(h)(Y_{gi} - \mathbb{E}[Y_{gi}|\mathcal{X}_n])|^\delta|\mathcal{X}_n]}{se(h)^\delta} \\ &\leq \frac{\sum_{g\in[G]} \mathbb{E}[|\sum_{i\in I_g} |w_{gi}(h)|^{1-1/\delta} \left(|w_{gi}(h)|^{1/\delta} |Y_{gi} - \mathbb{E}[Y_{gi}|\mathcal{X}_n]|\right)|^\delta|\mathcal{X}_n]}{se(h)^\delta} \end{aligned}$$

where the first equality uses the definition of W_g and the inequality follows by the triangle inequality. Next, by Hölder inequality with exponents δ and $\delta' = \delta/(\delta-1)$, such that $1/\delta + 1/\delta' = 1$, we obtain that

$$\begin{aligned} A_n &\leq \frac{\sum_{g\in[G]} \mathbb{E}\left[\left|\left(\sum_{i\in I_g} \left(|w_{gi}(h)|^{1-1/\delta}\right)^{\delta'}\right)^{1/\delta'} \left(\sum_{i\in I_g} |w_{gi}(h)| \, |Y_{gi} - \mathbb{E}[Y_{gi}|\mathcal{X}_n]|^\delta\right)^{1/\delta}\right|^\delta |\mathcal{X}_n\right]}{se(h)^\delta} \\ &= \frac{\sum_{g\in[G]} \mathbb{E}\left[\left|\left(\sum_{i\in I_g} |w_{gi}(h)|\right)^{1/\delta'} \left(\sum_{i\in I_g} |w_{gi}(h)| \, |Y_{gi} - \mathbb{E}[Y_{gi}|\mathcal{X}_n]|^\delta\right)^{1/\delta}\right|^\delta |\mathcal{X}_n\right]}{se(h)^\delta} \\ &= \frac{\sum_{g\in[G]} \left(\sum_{i\in I_g} |w_{gi}(h)|\right)^{\delta/\delta'} \mathbb{E}\left[\sum_{i\in I_g} |w_{gi}(h)||Y_{gi} - \mathbb{E}[Y_{gi}|\mathcal{X}_n]|^\delta|\mathcal{X}_n\right]}{se(h)^\delta}, \end{aligned}$$

where the first equality follows by the definition of δ', and the second equality uses the fact that the weights are deterministic given $\mathcal{X}_n$.

Next, using Assumption 2 that $\mathbb{E}[|Y_{gi} - \mathbb{E}[Y_{gi}|\mathcal{X}_n]|^\delta|\mathcal{X}_n]$ is uniformly bounded and the fact that $\delta/\delta' + 1 = \delta$, we obtain that

$$A_n \leq C\frac{\sum_{g\in[G]}\left(\sum_{i\in I_g}|w_{gi}(h)|\right)^\delta}{se(h)^\delta}.$$

Finally, by basic algebra,

$$\begin{aligned} A_n &\leq C\frac{\max_{g\in[G]}\left(\sum_{i\in I_g}|w_{gi}(h)|\right)^{\delta-2}}{se(h)^{\delta-2}}\frac{\sum_{g\in[G]}\left(\sum_{i\in I_g}|w_{gi}(h)|\right)^2}{se(h)^2} \\ &= C\left(\frac{\max_{g\in[G]}\left(\sum_{i\in I_g}|w_{gi}(h)|\right)^2}{se(h)^2}\right)^{\frac{\delta-2}{2}}\frac{\sum_{g\in[G]}\left(\sum_{i\in I_g}|w_{gi}(h)|\right)^2}{se(h)^2}. \end{aligned}$$

The conclusion follows by Assumption 1 and the fact that $\delta > 2$. □

B.2. **Proof of Theorem 2.** For a generic variable D_{gi}, let

$$D_{gi}^{\Delta_d} \equiv D_{gi} - \frac{1}{|\mathcal{N}_{gi}^d|}\sum_{(g',i')\in\mathcal{N}_{gi}^d} D_{g'i'} \quad \text{for } d \in \{1,2\}.$$

Let $q_{gi}(h) = w_{gi}(h)/se(h)$ and recall that $\widehat{\sigma}_{g,ij}^{CNN} = Y_{gi}^{\Delta_1}Y_{gj}^{\Delta_2}$. We prove Theorem 2 by showing that

$$\sum_{g\in[G]}\sum_{i,j\in I_g} q_{gi}(h)q_{gj}(h)(\widehat{\sigma}_{g,ij}^{CNN} - \sigma_{g,ij}) = o_P(1).$$

Throughout the proof, we rely on the following conditions on the error term and the weights. First, under cross-cluster independence and $\mathbb{E}[\varepsilon_{gi} \mid \mathcal{X}_n] = 0$, it holds that

$$\mathbb{E}[\varepsilon_{g_1i_1}\varepsilon_{g_2i_2}\varepsilon_{g_3i_3}\varepsilon_{g_4i_4} \mid \mathcal{X}_n] = 0 \tag{B.1}$$

unless each cluster index in $\{g_1, g_2, g_3, g_4\}$ appears at least twice.

Second, it directly follows from Assumption 1 that

$$\sum_{g\in[G]}\sum_{\tilde{g}\in[G]}\sum_{i,j\in I_g}\sum_{k,l\in I_{\tilde{g}}}|q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)| = O_P(1). \tag{B.2}$$

Third, we further note that by the triangular inequality and Assumption 1, it holds

that

$$|\sum_{g\in[G]}\sum_{i,j,k,l\in I_g} q_{gi}(h)q_{gj}(h)q_{gk}(h)q_{gl}(h)| \leq \sum_{g\in[G]}\sum_{i,j,k,l\in I_g} |q_{gi}(h)q_{gj}(h)q_{gk}(h)q_{gl}(h)|$$
$$\leq \max_g \Big\{ \sum_{i,j\in I_g} |q_{gi}(h)q_{gj}(h)| \Big\} \sum_{g\in[G]}\sum_{k,l\in I_g} |q_{gk}(h)q_{gl}(h)| = o_P(1). \tag{B.3}$$

In the following derivations, we write C for a generic positive constant whose value might differ between equations. For $d \in \{1,2\}$, we let

$$Y_{gi}^{\Delta_d} = \left(\mu(X_{gi}) - \frac{1}{|\mathcal{N}_{gi}^d|}\sum_{(g_d,i_d)\in\mathcal{N}_{gi}^d} \mu(X_{g_d i_d})\right) + \left(\varepsilon_{gi} - \frac{1}{|\mathcal{N}_{gi}^d|}\sum_{(g_d,i_d)\in\mathcal{N}_{gi}^d} \varepsilon_{g_d i_d}\right) \equiv \mu_{gi}^{\Delta_d} + \varepsilon_{gi}^{\Delta_d}.$$

First, we show that the standard error is asymptotically unbiased. It holds that

$$\begin{aligned}
\mathbb{E}\left[\frac{\widehat{se}^2_{CNN}(h)}{se^2(h)} - 1|\mathcal{X}_n\right] &= \mathbb{E}\Big[\sum_{g\in[G]}\sum_{i,j\in I_g} q_{gi}(h)q_{gj}(h)\left(Y_{gi}^{\Delta_1}Y_{gj}^{\Delta_2} - \sigma_{g,ij}\right)|\mathcal{X}_n\Big] \\
&= \sum_{g\in[G]}\sum_{i,j\in I_g} q_{gi}(h)q_{gj}(h)\mu_{gi}^{\Delta_1}\mu_{gj}^{\Delta_2} \\
&\quad + \sum_{g\in[G]}\sum_{i,j\in I_g} q_{gi}(h)q_{gj}(h)\mathbb{E}[\varepsilon_{gi}^{\Delta_1}|\mathcal{X}_n]\mu_{gj}^{\Delta_2} \\
&\quad + \sum_{g\in[G]}\sum_{i,j\in I_g} q_{gi}(h)q_{gj}(h)\mathbb{E}[\varepsilon_{gj}^{\Delta_2}|\mathcal{X}_n]\mu_{gi}^{\Delta_1} \\
&\quad + \sum_{g\in[G]}\sum_{i,j\in I_g} q_{gi}(h)q_{gj}(h)\mathbb{E}[\varepsilon_{gi}^{\Delta_1}\varepsilon_{gj}^{\Delta_2} - \sigma_{g,ij}|\mathcal{X}_n].
\end{aligned}$$

Since ε_{gi} is conditional mean zero for all $g \in [G]$ and $i \in I_g$ and by independence between clusters and construction of the sets $\mathcal{N}_{gi}^d$, it holds that for all pairs $i, j \in I_g$,

$$\begin{aligned}
\mathbb{E}[\varepsilon_{gi}^{\Delta_1}\varepsilon_{gj}^{\Delta_2}|\mathcal{X}_n] &= \mathbb{E}[\left(\varepsilon_{gi} - \frac{1}{|\mathcal{N}_{gi}^1|}\sum_{(g_1,i_1)\in\mathcal{N}_{gi}^1}\varepsilon_{g_1 i_1}\right)\left(\varepsilon_{gj} - \frac{1}{|\mathcal{N}_{gj}^2|}\sum_{(g_2,j_2)\in\mathcal{N}_{gj}^2}\varepsilon_{g_2 j_2}\right)|\mathcal{X}_n] \\
&= \mathbb{E}[\varepsilon_{gi}\varepsilon_{gj}|\mathcal{X}_n] = \sigma_{g,ij}.
\end{aligned}$$

We further note that

$$\left|\mathbb{E}\left[\frac{\widehat{se}^2_{CNN}(h)}{se^2(h)} - 1|\mathcal{X}_n\right]\right| = \left|\sum_{g\in[G]}\sum_{i,j\in I_g} q_{gi}(h)q_{gj}(h)\mu^{\Delta_1}_{gi}\mu^{\Delta_2}_{gj}\right|$$
$$\leq C \max_{g\in[G]} \max_{\substack{i\in I_g \\ w_{gi}(h)\neq 0}} \max_{(g',i')\in\mathcal{N}^1_{gi}\cup\mathcal{N}^2_{gi}} |X_{gi} - X_{g'i'}|^2 \sum_{g\in[G]}\sum_{i,j\in I_g} |q_{gi}(h)q_{gj}(h)|$$
$$= o_P(1).$$

The first inequality follows from the L-Lipschitz continuity of μ and the triangular inequality. The last line follows as $\sum_{g\in[G]}\sum_{i,j\in I_g}|q_{gi}(h)q_{gj}(h)| = O_P(1)$ and by Assumption 6. We have shown that the standard error is asymptotically unbiased.

Second, we study the conditional variance of $\widehat{se}^2_{CNN}(h)$. We consider the following decomposition

$$\frac{\widehat{se}^2_{CNN}(h) - \mathbb{E}[\widehat{se}^2_{CNN}(h)|\mathcal{X}_n]}{se^2(h)} = \sum_{g\in[G]}\sum_{i,j\in I_g} q_{gi}(h)q_{gj}(h)\left((\varepsilon^{\Delta_1}_{gi} + \mu^{\Delta_1}_{gi})(\varepsilon^{\Delta_2}_{gj} + \mu^{\Delta_2}_{gj}) - \mathbb{E}[\widehat{\sigma}^{CNN}_{g,ij}|\mathcal{X}_n]\right)$$

$$= \sum_{g\in[G]}\sum_{i,j\in I_g} q_{gi}(h)q_{gj}(h) \times \Bigg((\varepsilon_{gi}\varepsilon_{gj} - \mathbb{E}[\varepsilon_{gi}\varepsilon_{gj}|\mathcal{X}_n]) - \left(\frac{1}{|\mathcal{N}^1_{gi}|}\sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}}\varepsilon_{g_1i_1}\right)\varepsilon_{gj}$$
$$- \left(\frac{1}{|\mathcal{N}^2_{gj}|}\sum_{(g_2,j_2)\in\mathcal{N}^2_{gj}}\varepsilon_{g_2j_2}\right)\varepsilon_{gi} + \left(\frac{1}{|\mathcal{N}^2_{gj}|}\sum_{(g_2,j_2)\in\mathcal{N}^2_{gj}}\varepsilon_{g_2j_2}\right)\left(\frac{1}{|\mathcal{N}^1_{gi}|}\sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}}\varepsilon_{g_1i_1}\right)$$
$$+ \varepsilon_{gi}\mu^{\Delta_2}_{gj} - \left(\frac{1}{|\mathcal{N}^1_{gi}|}\sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}}\varepsilon_{g_1i_1}\right)\mu^{\Delta_2}_{gj} + \varepsilon_{gj}\mu^{\Delta_1}_{gi} - \left(\frac{1}{|\mathcal{N}^2_{gj}|}\sum_{(g_2,j_2)\in\mathcal{N}^2_{gj}}\varepsilon_{g_2j_2}\right)\mu^{\Delta_1}_{gi}\Bigg)$$
$$\equiv A_1 - A_2 - A_3 + A_4 + A_5 - A_6 + A_7 - A_8.$$

It is easy to see that these eight terms are all mean zero conditional on $\mathcal{X}_n$. It thus suffices to show that their second moments converge to zero. We will consider each of the terms $A_1, \ldots, A_8$ separately.

We start with A_1. It holds that

$$\mathrm{Var}(A_1|\mathcal{X}_n)$$
$$= \sum_{g\in[G]}\sum_{\tilde{g}\in[G]}\sum_{i,j\in I_g}\sum_{l,k\in I_{\tilde{g}}} q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)\mathbb{E}[(\varepsilon_{gi}\varepsilon_{gj} - \mathbb{E}[\varepsilon_{gi}\varepsilon_{gj}|\mathcal{X}_n])(\varepsilon_{\tilde{g}l}\varepsilon_{\tilde{g}k} - \mathbb{E}[\varepsilon_{\tilde{g}l}\varepsilon_{\tilde{g}k}|\mathcal{X}_n])|\mathcal{X}_n]$$
$$= \sum_{g\in[G]}\sum_{i,j,l,k\in I_g} q_{gi}(h)q_{gj}(h)q_{gk}(h)q_{gl}(h)\mathbb{E}[(\varepsilon_{gi}\varepsilon_{gj} - \mathbb{E}[\varepsilon_{gi}\varepsilon_{gj}|\mathcal{X}_n])(\varepsilon_{gl}\varepsilon_{gk} - \mathbb{E}[\varepsilon_{gl}\varepsilon_{gk}|\mathcal{X}_n])|\mathcal{X}_n]$$
$$\leq C\sum_{g\in[G]}\sum_{i,j,l,k\in I_g}|q_{gi}(h)q_{gj}(h)q_{gk}(h)q_{gl}(h)| = o_P(1)$$

The second equality follows as the units are independent across clusters and by Condition B.1. The first inequality follows by boundedness of fourth moments of the error term and the last equality follows from Condition B.3.

We now consider A_2

$$\begin{aligned}
&\mathrm{Var}(A_2|\mathcal{X}_n) \\
&= \sum_{g,\tilde{g}\in[G]} \sum_{i,j\in I_g} \frac{1}{|\mathcal{N}^1_{gi}|} \sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} \sum_{l,k\in I_{\tilde{g}}} \frac{1}{|\mathcal{N}^1_{\tilde{g}k}|} \sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{\tilde{g}k}} q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)\mathbb{E}[\varepsilon_{g_1i_1}\varepsilon_{gj}\varepsilon_{\tilde{g}_1k_1}\varepsilon_{\tilde{g}l}|\mathcal{X}_n]
\end{aligned}$$

Based on the logic of Condition B.1, each of those terms of the sums are nonzero if either $(g_1 = \tilde{g}_1 \,\&\, g = \tilde{g})$ or $(g_1 = \tilde{g} \,\&\, g = \tilde{g}_1)$. By the boundedness of the conditional expectations of the first four moments of the error term, it then follows that

$$\begin{aligned}
&\mathrm{Var}(A_2|\mathcal{X}_n) \leq C \sum_{g\in[G]} \sum_{i,j,l,k\in I_g} \sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} \sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{\tilde{g}k}} \frac{1}{|\mathcal{N}^1_{gi}|}\frac{1}{|\mathcal{N}^1_{\tilde{g}k}|}\mathbf{1}\{g_1=\tilde{g}_1\}|q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)| \\
&+ C \sum_{g,\tilde{g}\in[G]} \sum_{i,j\in I_g} \sum_{l,k\in I_{\tilde{g}}} \frac{1}{|\mathcal{N}^1_{gi}|} \sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} \frac{1}{|\mathcal{N}^1_{\tilde{g}k}|} \sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{\tilde{g}k}} \mathbf{1}\{g_1=\tilde{g}\,\&\,\tilde{g}_1=g\}|q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)| \\
&\equiv H_1 + H_2
\end{aligned}$$

Furthermore, by Condition B.3,

$$\begin{aligned}
H_1 &\leq C \sum_{g\in[G]} \sum_{i,j,k,l\in I_g} |q_{gi}(h)q_{gj}(h)| \frac{1}{|\mathcal{N}^1_{gi}|} \sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} \frac{1}{|\mathcal{N}^1_{gk}|} \sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{gk}} |q_{gk}(h)q_{gl}(h)| \\
&\leq C \sum_{g\in[G]} \sum_{i,j,k,l\in I_g} |q_{gi}(h)q_{gj}(h)q_{gk}(h)q_{gl}(h)| = o_P(1).
\end{aligned}$$

We further note that as by construction, Equation 5.2, each cluster is contributes neighbors only to units of R other clusters

$$\begin{aligned}
H_2 &\leq C \sum_{g\in[G]} \sum_{i,j\in I_g} \frac{|q_{gi}(h)q_{gj}(h)|}{|\mathcal{N}^1_{gi}|} \sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} \sum_{\substack{\tilde{g}\in[G] \\ g\in\mathcal{R}^1_{\tilde{g}} \\ \tilde{g}\in\mathcal{R}^1_g}} \left(\sum_{l,k\in I_{\tilde{g}}} \frac{|q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)|}{|\mathcal{N}^1_{\tilde{g}k}|} \sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{\tilde{g}k}} \mathbf{1}\{g_1=\tilde{g},\ \tilde{g}_1=g\} \right) \\
&\leq C \left(\max_{\tilde{g}\in[G]} \sum_{l,k\in I_{\tilde{g}}} |q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)| \right) \sum_{g\in[G]} \sum_{i,j\in I_g} |q_{gi}(h)q_{gj}(h)| \\
&= o_P(1).
\end{aligned}$$

where the last inequality follows by the restriction of Equation (5.2) as each cluster is used as a companion cluster at most R times. The last line follow from condition (B.3). It follows that $\mathrm{Var}(A_2|\mathcal{X}_n) = o_P(1)$. Using the same arguments, it also holds

that $\mathrm{Var}(A_3|\mathcal{X}_n) = o_P(1)$.

We now consider A_4.

$$\mathrm{Var}(A_4|\mathcal{X}_n) = \sum_{g\in[G]}\sum_{\tilde{g}\in[G]}\sum_{i,j\in I_g}\sum_{l,k\in I_{\tilde{g}}} \frac{1}{|\mathcal{N}^2_{gj}|}\sum_{(g_2,j_2)\in\mathcal{N}^2_{gj}} \frac{1}{|\mathcal{N}^1_{gi}|}\sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} \frac{1}{|\mathcal{N}^2_{\tilde{g}l}|}\sum_{(\tilde{g}_2,l_2)\in\mathcal{N}^2_{\tilde{g}l}} \frac{1}{|\mathcal{N}^1_{\tilde{g}k}|}\sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{\tilde{g}k}}$$
$$\cdot\, q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)\mathbb{E}[\varepsilon_{g_2j_2}\varepsilon_{g_1i_1}\varepsilon_{\tilde{g}_2l_2}\varepsilon_{\tilde{g}_1k_1}|\mathcal{X}_n]$$

Based on the logic of Condition B.1, each of those terms of the sums are nonzero if either $(g_1 = \tilde{g}_1 \,\&\, g_2 = \tilde{g}_2)$ or $(g_1 = \tilde{g}_2 \,\&\, g_2 = \tilde{g}_1)$. By the boundedness of the conditional expectations of the first four moments of the error term, it then follows that

$$\begin{aligned}
\mathrm{Var}(A_4|\mathcal{X}_n) &\le \sum_{g\in[G]}\sum_{\tilde{g}\in[G]}\sum_{i,j\in I_g}\sum_{l,k\in I_{\tilde{g}}} \frac{1}{|\mathcal{N}^2_{gj}|}\sum_{(g_2,j_2)\in\mathcal{N}^2_{gj}} \frac{1}{|\mathcal{N}^1_{gi}|}\sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} \frac{1}{|\mathcal{N}^2_{\tilde{g}l}|}\sum_{(\tilde{g}_2,l_2)\in\mathcal{N}^2_{\tilde{g}l}} \frac{1}{|\mathcal{N}^1_{\tilde{g}k}|}\sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{\tilde{g}k}} \\
&\quad \cdot \mathbf{1}\{g_1 = \tilde{g}_1 \,\&\, g_2 = \tilde{g}_2\}|q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)| \\
&+ \sum_{g\in[G]}\sum_{\tilde{g}\in[G]}\sum_{i,j\in I_g}\sum_{l,k\in I_{\tilde{g}}} \frac{1}{|\mathcal{N}^2_{gj}|}\sum_{(g_2,j_2)\in\mathcal{N}^2_{gj}} \frac{1}{|\mathcal{N}^1_{gi}|}\sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} \frac{1}{|\mathcal{N}^2_{\tilde{g}l}|}\sum_{(\tilde{g}_2,l_2)\in\mathcal{N}^2_{\tilde{g}l}} \frac{1}{|\mathcal{N}^1_{\tilde{g}k}|}\sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{\tilde{g}k}} \\
&\quad \cdot \mathbf{1}\{g_1 = \tilde{g}_2 \,\&\, g_2 = \tilde{g}_1\}|q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)| \\
&\equiv H_3 + H_4
\end{aligned}$$

The first inequality follows from Condition B.1 and the boundedness of the conditional expectations of the first four moments of the error term. We further note that

$$\begin{aligned}
H_3 &= \sum_{g\in[G]}\sum_{i,j\in I_g} \frac{1}{|\mathcal{N}^2_{gj}|}\sum_{(g_2,j_2)\in\mathcal{N}^2_{gj}} \frac{1}{|\mathcal{N}^1_{gi}|}\sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} |q_{gi}(h)q_{gj}(h)| \\
&\cdot \left(\sum_{\substack{\tilde{g}\in[G]: \\ \mathcal{R}^1_g\cap\mathcal{R}^1_{\tilde{g}}\neq\emptyset \\ \mathcal{R}^2_g\cap\mathcal{R}^2_{\tilde{g}}\neq\emptyset}} \sum_{l,k\in I_{\tilde{g}}} \frac{1}{|\mathcal{N}^2_{\tilde{g}l}|}\sum_{(\tilde{g}_2,l_2)\in\mathcal{N}^2_{\tilde{g}l}} \frac{1}{|\mathcal{N}^1_{\tilde{g}k}|}\sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{\tilde{g}k}} |q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)|\mathbf{1}\{g_1 = \tilde{g}_1 \,\&\, g_2 = \tilde{g}_2\} \right) \\
&\le C\left(\max_{\tilde{g}\in[G]} \sum_{l,k\in I_{\tilde{g}}} \frac{1}{|\mathcal{N}^2_{\tilde{g}l}|}\sum_{(\tilde{g}_2,l_2)\in\mathcal{N}^2_{\tilde{g}l}} \frac{1}{|\mathcal{N}^1_{\tilde{g}k}|}\sum_{(\tilde{g}_1,k_1)\in\mathcal{N}^1_{\tilde{g}k}} |q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)|\mathbf{1}\{g_1 = \tilde{g}_1 \,\&\, g_2 = \tilde{g}_2\} \right) \\
&\quad \cdot \sum_{g\in[G]}\sum_{i,j\in I_g} \frac{1}{|\mathcal{N}^2_{gj}|}\sum_{(g_2,j_2)\in\mathcal{N}^2_{gj}} \frac{1}{|\mathcal{N}^1_{gi}|}\sum_{(g_1,i_1)\in\mathcal{N}^1_{gi}} |q_{gi}(h)q_{gj}(h)| \\
&= o_P(1)
\end{aligned}$$

where the inequality follows from the condition (5.2). If each cluster is used as a compan-

ion cluster at most R times, cluster can share also only a bounded number of common companion clusters. The last equality follows from Condition B.3. By the same arguments, it holds that $H_4 = o_P(1)$. It follows that $\mathrm{Var}(A_4|\mathcal{X}_n) = o_P(1)$.

We now consider A_5. It holds that

$$\begin{aligned}
\mathrm{Var}(A_5|\mathcal{X}_n) &= \sum_{g\in[G]}\sum_{\tilde{g}\in[G]}\sum_{i,j\in I_g}\sum_{l,k\in I_{\tilde{g}}} q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)\mathbb{E}[\varepsilon_{gi}\varepsilon_{\tilde{g}k}|\mathcal{X}_n]\mu_{gj}^{\Delta_2}\mu_{\tilde{g}l}^{\Delta_2} \\
&= \sum_{g\in[G]}\sum_{i,j,l,k\in I_g} q_{gi}(h)q_{gj}(h)q_{gk}(h)q_{gl}(h)\mathbb{E}[\varepsilon_{gi}\varepsilon_{gk}|\mathcal{X}_n]\mu_{gj}^{\Delta_2}\mu_{gl}^{\Delta_2} \\
&\leq C\max_{g\in[G]}\max_{j\in I_g}(\mu_{gj}^{\Delta_2})^2\sum_{g\in[G]}\sum_{i,j,l,k\in I_g}|q_{gi}(h)q_{gj}(h)q_{gk}(h)q_{gl}(h)| \\
&= o_P(1).
\end{aligned}$$

The second equality follows from Condition B.1. The inequality follows as the first four moments of the error terms are bounded. The last step follows from Condition B.3 and as $\max_{g\in[G]}\max_{j\in I_g}(\mu_{gj}^{\Delta_2})^2 = o_P(1)$ by Assumption 6 and Lipschitz continuity of μ.

We now consider A_6. It holds that

$$\begin{aligned}
\mathrm{Var}(A_6|\mathcal{X}_n) &= \sum_{g\in[G]}\sum_{\tilde{g}\in[G]}\sum_{i,j\in I_g}\sum_{l,k\in I_{\tilde{g}}}\frac{1}{|\mathcal{N}_{gi}^1|}\sum_{(g_1,i_1)\in\mathcal{N}_{gi}^1}\frac{1}{|\mathcal{N}_{\tilde{g}k}^1|}\sum_{(\tilde{g}_1,k_1)\in\mathcal{N}_{\tilde{g}k}^1} \\
&\quad\cdot\left(q_{gi}(h)q_{gj}(h)q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)\right)\mathbb{E}[\varepsilon_{g_1i_1}\varepsilon_{\tilde{g}_1k_1}|\mathcal{X}_n]\mu_{gj}^{\Delta_2}\mu_{\tilde{g}l}^{\Delta_2} \\
&\leq C\max_{g\in[G]}\max_{j\in I_g}(\mu_{gj}^{\Delta_2})^2\sum_{g\in[G]}\sum_{i,j\in I_g}|q_{gi}(h)q_{gj}(h)\frac{1}{|\mathcal{N}_{gi}^1|}\sum_{(g_1,i_1)\in\mathcal{N}_{gi}^1}\sum_{\substack{\tilde{g}\in[G]\\ \mathcal{R}_g^1\cap\mathcal{R}_{\tilde{g}}^1\neq\emptyset}}\sum_{l,k\in I_{\tilde{g}}} \\
&\quad\cdot\left(\frac{1}{|\mathcal{N}_{\tilde{g}k}^1|}\sum_{(\tilde{g}_1,k_1)\in\mathcal{N}_{\tilde{g}k}^1}\mathbf{1}\{g_1=\tilde{g}_1\}|q_{\tilde{g}k}(h)q_{\tilde{g}l}(h)|\right) \\
&= o_P(1).
\end{aligned}$$

The first inequality follows as the first four moments of the error terms are bounded and by Condition B.1. The last equality follows as $\max_{g\in[G]}\max_{j\in I_g}(\mu_{gj}^{\Delta_2})^2 = o_P(1)$ by Assumption 6 and from Condition B.2.

One can show that $\mathrm{Var}(A_7|\mathcal{X}_n) = o_P(1)$ using the same arguments as in the discussion of A_5. Similarly, one can show that $\mathrm{Var}(A_8|\mathcal{X}_n) = o_P(1)$ using the same arguments as in the discussion of A_6. This reasoning concludes the proof. □

B.3. **Proof of Theorem 3.** Define the infeasible version of $\widehat{se}^2_{CRR}(h)$ using the true residuals $\varepsilon_{gi} = Y_{gi} - \mu(X_{gi})$,

$$\widetilde{se}^2_{CRR}(h) = \sum_{g\in[G]} \sum_{i,j\in I_g} w_{gi}(h) w_{gj}(h) \varepsilon_{gi}\varepsilon_{gj}.$$

First, we show that $\widehat{se}^2_{CRR}(h)$ and $\widetilde{se}^2_{CRR}(h)$ are first-order equivalent,

$$\frac{\widehat{se}^2_{CRR}(h) - \widetilde{se}^2_{CRR}(h)}{se^2(h)} = o_P(1). \tag{B.4}$$

To prove this claim, observe that

$$\begin{aligned}
\hat{\varepsilon}_{gi}\hat{\varepsilon}_{gj} - \varepsilon_{gi}\varepsilon_{gj} &= (Y_{gi}Y_{gj} - \widehat{\mu}(X_{gi})Y_{gj} - Y_{gi}\widehat{\mu}(X_{gj}) + \widehat{\mu}(X_{gi})\widehat{\mu}(X_{gj})) \\
&\quad - (Y_{gi}Y_{gj} - \mu(X_{gi})Y_{gj} - Y_{gi}\mu(X_{gj}) + \mu(X_{gi})\mu(X_{gj})) \\
&= (\mu(X_{gi}) - \widehat{\mu}(X_{gi}))(Y_{gj} - \mu(X_{gj})) + (\mu(X_{gj}) - \widehat{\mu}(X_{gj}))(Y_{gi} - \mu(X_{gi})) \\
&\quad + (\widehat{\mu}(X_{gi}) - \mu(X_{gi}))\,(\widehat{\mu}(X_{gj}) - \mu(X_{gj}))\,.
\end{aligned}$$

It follows that

$$\begin{aligned}
\widehat{se}^2_{CRR}(h) - \widetilde{se}^2_{CRR}(h) &= \sum_{g\in[G]} \sum_{i,j\in I_g} w_{gi}(h) w_{gj}(h)\,(\hat{\varepsilon}_{gi}\hat{\varepsilon}_{gj} - \varepsilon_{gi}\varepsilon_{gj}) \\
&= 2\sum_{g\in[G]} \sum_{i,j\in I_g} w_{gi}(h) w_{gj}(h)(\mu(X_{gi}) - \widehat{\mu}(X_{gi}))(Y_{gj} - \mu(X_{gj})) \\
&\quad + \sum_{g\in[G]} \sum_{i,j\in I_g} w_{gi}(h) w_{gj}(h)\,(\widehat{\mu}(X_{gi}) - \mu(X_{gi}))\,(\widehat{\mu}(X_{gj}) - \mu(X_{gj}))\,.
\end{aligned}$$

We can decompose $\widehat{\mu}(X_{gi}) - \mu(X_{gi})$ as follows:

$$\begin{aligned}
\widehat{\mu}(X_{gi}) - \mu(X_{gi}) &= \left((\hat{b}_0^+ - b_0^+) + (\hat{b}_1^+ - b_1^+)X_{gi}\right)\mathbf{1}\{0 \le X_{gi}\} \\
&\quad + \left((\hat{b}_0^- - b_0^-) + (\hat{b}_1^- - b_1^-)X_{gi}\right)\mathbf{1}\{X_{gi} < 0\} + B_{gi},
\end{aligned}$$

where B_{gi} is the remainder from a Taylor expansion of μ on the respective side of the cutoff that depends only on X_{gi} and $\max_{g\in[G]}\max_{i\in I_g:|X_{gi}|\leq h}|B_{gi}| = O(h^2)$. It follows that

$$\begin{aligned}
\widehat{se}^2_{CRR} - \widetilde{se}^2_{CRR} &= 2\sum_{\star\in\{+,-\}}(\hat{b}^\star_0 - b^\star_0)\sum_{g\in[G]}\sum_{i,j\in I_g}(-1)^{\mathbf{1}\{\star=-\}}w^\star_{gi}(h)w_{gj}(h)(Y_{gj}-\mu(X_{gj})) \\
&\quad + 2\sum_{\star\in\{+,-\}}h(\hat{b}^\star_1 - b^\star_1)\sum_{g\in[G]}\sum_{i,j\in I_g}(-1)^{\mathbf{1}\{\star=-\}}w^\star_{gi}(h)w_{gj}(h)(X_{gi}/h)(Y_{gj}-\mu(X_{gj})) \\
&\quad + 2\sum_{g\in[G]}\sum_{i,j\in I_g}w_{gi}(h)w_{gj}(h)B_{gi}(Y_{gj}-\mu(X_{gj})) \\
&\quad + \sum_{g\in[G]}\sum_{i,j\in I_g}w_{gi}(h)w_{gj}(h)\left(\widehat{\mu}(X_{gi})-\mu(X_{gi})\right)\left(\widehat{\mu}(X_{gj})-\mu(X_{gj})\right) \\
&\equiv 2A_1 + 2A_2 + 2A_3 + A_4.
\end{aligned}$$

We begin by studying the first two terms. Let

$$T^\star_l = \sum_{g\in[G]}\sum_{i,j\in I_g}w^\star_{gi}(h)w_{gj}(h)(X_{gi}/h)^l(Y_{gj}-\mu(X_{gj})).$$

Note that, $\mathbb{E}[T^\star_l|\mathcal{X}_n] = 0$ and, using the fact that $\mathrm{Var}(\varepsilon_{gi}|\mathcal{X}_g)$ is uniformly bounded,

$$\begin{aligned}
\mathrm{Var}(T^\star_l|\mathcal{X}_n) &= \sum_{g\in[G]}\mathbb{E}\left[\left(\sum_{i,j\in I_g}w^\star_{gi}(h)w_{gj}(h)(X_{gi}/h)^l(Y_{gj}-\mu(X_{gj}))\right)^2\Big|\mathcal{X}_g\right] \\
&\leq C\sum_{g\in[G]}\left(\sum_{i,j\in I_g}|w^\star_{gi}(h)w_{gj}(h)|\right)^2 \\
&\leq C\max_{g\in[G]}\sum_{i,j\in I_g}|w^\star_{gi}(h)w_{gj}(h)|\sum_{g\in[G]}\sum_{i,j\in I_g}|w^\star_{gi}(h)w_{gj}(h)|.
\end{aligned}$$

By Assumption 1, it then follows that

$$T^\star_l/se^2(h) = o_P(1),$$

and, in consequence, $A_1/se^2(h) = o_P(1)$ and $A_2/se^2(h) = o_P(1)$, using $\hat{b}^\star_0 - b^\star_0 = o_P(1)$ and $h(\hat{b}^\star_1 - b^\star_1) = o_P(1)$. By the same reasoning, $A_3/se^2(h) = O_P(h^2) = o_P(1)$. The last term is bounded as follows:

$$|A_4| \leq \max_{g\in[G]}\max_{i\in I_g:|X_{gi}|\leq h}\left(\widehat{\mu}(X_{gi})-\mu(X_{gi})\right)^2\sum_{g\in[G]}\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)|,$$

such that $A_4/se^2(h) = o_P(1)$.

Second, we show that

$$\frac{\widetilde{se}^2_{CRR}(h) - se^2(h)}{se^2(h)} = o_P(1). \tag{B.5}$$

To see that, note that

$$\mathbb{E}[\widetilde{se}^2_{CRR}(h)|\mathcal{X}_n] - se^2(h) = \sum_{g\in[G]}\sum_{i,j\in I_g} w_{gi}(h)w_{gj}(h)\mathbb{E}\left[\varepsilon_{gi}\varepsilon_{gj} - \sigma_{g,ij}|\mathcal{X}_g\right] = 0,$$

and

$$\begin{aligned}
\operatorname{Var}\left(\widetilde{se}^2_{CRR}(h)|\mathcal{X}_n\right) &\leq \sum_{g\in[G]}\left(\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)|\right)^2 \max_{i,j\in I_g}\operatorname{Var}(\varepsilon_{gi}\varepsilon_{gj}|\mathcal{X}_g) \\
&\leq C\max_{g\in[G]}\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)|\sum_{g\in[G]}\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)| \\
&= o_P(1).
\end{aligned}$$

The proof is concluded by combining (B.4) and (B.5). □

C. PROOFS FOR SECTION 4

C.1. **Additional Notation and Lemmas.** Let $k^+(v) = k(v)\mathbf{1}\{0 \leq v\}$, $k^-(v) = k(v)\mathbf{1}\{v < 0\}$, and $k^\star_h(v) = k^\star(v/h)/h$ for $\star \in \{+,-\}$. The local linear RD estimator is defined as

$$\begin{aligned}
\widehat{\tau}(h) &= \sum_{g\in[G]}\sum_{i\in I_g} w_{gi}(h)Y_i, \quad w_{gi}(h) = w^+_{gi}(h) - w^-_{gi}(h), \\
w^\star_{gi}(h) &= e_1^\top\left(\sum_{g\in[G]}\sum_{i\in I_g} k^\star_h(X_i)\widetilde{X}_{gi}\widetilde{X}_{gi}^\top\right)^{-1} k^\star_h(X_{gi})\widetilde{X}_{gi} \quad \text{for } \star\in\{+,-\},
\end{aligned}$$

where $\widetilde{X}_{gi} = (1, X_{gi})^\top$. The local linear weights can be further expressed as

$$w^\star_{gi}(h) = \frac{\frac{1}{n}k^\star_h(X_{gi})\left(S^\star_{n,2} - S^\star_{n,1}(X_{gi}/h)\right)}{S^\star_{n,2}S^\star_{n,0} - (S^\star_{n,1})^2}, \quad S^\star_{n,l} = \frac{1}{n}\sum_{g\in[G]}\sum_{i\in I_g} k^\star_h(X_{gi})(X_{gi}/h)^l.$$

To prove the results in Section 4, we first establish two technical lemmas. The first lemma provides basic convergence results for S_n and other kernel-weighted sums that are

used in all four asymptotic frameworks. For $l, m \in \mathbb{N}_0$ and $\star, \diamond \in \{+, -\}$,

$$T^{\star}_{n,l} = \frac{1}{n^2} \sum_{g \in [G]} \sum_{i \in I_g} (k^{\star}_h(X_{gi}))^2 (X_{gi}/h)^l,$$
$$U^{\star\diamond}_{n,l,m} = \frac{1}{n^2} \sum_{g \in [G]} \sum_{\substack{i \neq j \\ i,j \in I_g}} k^{\star}_h(X_{gi}) k^{\diamond}_h(X_{gj}) (X_{gi}/h)^l (X_{gj}/h)^m.$$

Furthermore, for $l \in \mathbb{N}_0$ and $\star, \diamond \in \{+, -\}$, let $\bar{\mu}^{\star}_l = \int k^{\star}(v) v^l dv$ and $\bar{\kappa}^{\star}_l = \int (k^{\star}(v))^2 v^l dv$.

Lemma C.1. *Suppose that Assumption 3 holds, and either*

(A) Assumption AF-I(i) holds and $\frac{1}{n^2} \sum_{g \in [G]} n_g^2 = o(1)$*; or*

(B) $\frac{1}{n^2} \sum_{g \in [G]} n_g^2 = o(h)$.

Then the following hold for $l, m \in \mathbb{N}_0$ *and* $\star, \diamond \in \{+, -\}$.

(i) $S^{\star}_{n,l} = \bar{\mu}^{\star}_l f(0) + o_P(1)$,

(ii) $T^{\star}_{n,l} = \frac{1}{nh} \left(\bar{\kappa}^{\star}_l f(0) + o_P(1) \right)$.

(iii) In case (A),

$$nh U^{\star\diamond}_{n,l,m} = O_P \left(\frac{1}{nh} \sum_{g \in [G]} (n_g h)^2 + \frac{1}{nh} \right).$$

If in addition all pairs (X_{gi}, X_{gj}), $i \neq j$, *are identically distributed with continuous joint density* $f(x_1, x_2)$ *and* $\frac{\max_{g \in [G]} (n_g h)^2}{nh + \sum_{g \in [G]} (n_g h)^2} = o(1)$, *then*

$$nh U^{\star\diamond}_{n,l,m} = \lambda_n \bar{\mu}^{\star}_l \bar{\mu}^{\diamond}_m f(0^{\star}, 0^{\diamond}) + o_P(1 + \lambda_n).$$

(iv) In case (B),

$$nh U^{\star\diamond}_{n,l,m} = O_P \left(\frac{1}{n} \sum_{g \in [G]} n_g^2 + \sqrt{\frac{\max_{g \in [G]} n_g^2}{nh} \frac{1}{n} \sum_{g \in [G]} n_g^2} \right).$$

If in addition the realizations of the running variable are equal within each cluster and $\frac{\max_{g \in [G]} n_g^2}{\sum_{g \in [G]} n_g^2} = o(h)$, *then*

$$nh U^{\star\diamond}_{n,l,m} = \bar{\kappa}^{\star}_{l+m} f(0) \lambda_n / h + o_P(\lambda_n / h).$$

The second lemma is relevant for the proofs under Asymptotic Frameworks I and III, where we assume that the density of the running variable within each cluster admits a bounded density. Let $n_{g,h}$ denote the number of observations within the estimation window from cluster $g \in [G]$, i.e., $n_{g,h} = \sum_{i \in I_g} \mathbf{1}\{|X_{gi}| \leq h\}$, assuming that the support of the kernel function used is contained in $[-1, 1]$.

Lemma C.2. *Suppose that Assumption AF-I(i) holds, the kernel k has bounded support, and $G \geq 2$. Then*

$$\max_{g \in [G]} n_{g,h} = O_P\left(\max_{g \in [G]} n_g h + \log G\right).$$

C.2. **Proof of Proposition 1.** In the following, let C denote a generic positive constant that might differ between equations. Recall from Appendix C.1 that for $\star \in \{+, -\}$, the local linear weights can be expressed as

$$w^{\star}_{gi}(h) = \frac{\frac{1}{n} k^{\star}_h(X_{gi}) \left(S^{\star}_{n,2} - S^{\star}_{n,1}(X_{gi}/h)\right)}{S^{\star}_{n,2} S^{\star}_{n,0} - (S^{\star}_{n,1})^2}.$$

By Lemma C.1, $S^{+}_{n,2} S^{+}_{n,0} - (S^{+}_{n,1})^2 = S^{-}_{n,2} S^{-}_{n,0} - (S^{-}_{n,1})^2 + o_P(1) = C + o_P(1)$. It follows that

$$\widetilde{w}^{\star}_{gi}(h) \equiv \frac{1}{n} k^{\star}_h(X_{gi}) \left(S^{\star}_{n,2} - S^{\star}_{n,1}(X_{gi}/h)\right) = w^{\star}_{gi}(h)/(C + o_P(1)).$$

Let $\widetilde{w}_{gi}(h) = \widetilde{w}^{+}_{gi}(h) - \widetilde{w}^{-}_{gi}(h)$.

Verification of Assumption 1: To begin with, note that the assumption that the eigenvalues of Σ_g are bounded away from zero implies that

$$se^2(h) = \frac{1}{C + o_P(1)} \sum_{g \in [G]} \sum_{i,j \in I_g} \widetilde{w}_{gi}(h) \widetilde{w}_{gj}(h) \sigma_{g,ij} \geq \frac{1}{C + o_P(1)} \sum_{g \in [G]} \sum_{i \in I_g} \widetilde{w}_{gi}(h)^2.$$

It follows that for any $g \in [G]$,

$$\frac{\sum_{i,j \in I_g} |w_{gi}(h) w_{gj}(h)|}{se^2(h)} \leq \frac{\sum_{i,j \in I_g} |\widetilde{w}_{gi}(h) \widetilde{w}_{gj}(h)|}{\sum_{g \in [G]} \sum_{i \in I_g} \widetilde{w}_{gi}(h)^2} (C + o_P(1)). \tag{C.1}$$

We use the bound in (C.1) to verify Assumption 1. First, we note that the denominator of the bound satisfies

$$\sum_{g \in [G]} \sum_{i \in I_g} \widetilde{w}_{gi}(h)^2 = \sum_{\star \in \{+,-\}} \sum_{g \in [G]} \sum_{i \in I_g} \widetilde{w}^{\star}_{gi}(h)^2 = \frac{C + o_P(1)}{nh}. \tag{C.2}$$

This holds because by Lemma C.1, for $\star \in \{+,-\}$, we have that

$$\begin{aligned}\sum_{g\in[G]}\sum_{i\in I_g}\widetilde{w}^{\star}_{gi}(h)^2 &= \frac{1}{n^2}\sum_{g\in[G]}\sum_{i\in I_g}(k^{\star}_h(X_{gi}))^2\left((S^{\star}_{n,2})^2 - 2S^{\star}_{n,2}S^{\star}_{n,1}(X_{gi}/h) + (S^{\star}_{n,1}(X_{gi}/h))^2\right)\\ &= (S^{\star}_{n,2})^2T^{\star}_{n,0} - 2S^{\star}_{n,2}S^{\star}_{n,1}T^{\star}_{n,1} + (S^{\star}_{n,1})^2T^{\star}_{n,2}\\ &= \frac{1}{nh}\left((\bar{\mu}^{\star}_2)^2\bar{\kappa}^{\star}_0 - 2\bar{\mu}^{\star}_2\bar{\mu}^{\star}_1\bar{\kappa}^{\star}_1 + (\bar{\mu}^{\star}_1)^2\bar{\kappa}^{\star}_2\right)f(0)^3(1+o_P(1)),\end{aligned}$$

where $(\bar{\mu}^{\star}_2)^2\bar{\kappa}^{\star}_0 - 2\bar{\mu}^{\star}_2\bar{\mu}^{\star}_1\bar{\kappa}^{\star}_1 + (\bar{\mu}^{\star}_1)^2\bar{\kappa}^{\star}_2 = \int(k^{\star}(v)(\bar{\mu}^{\star}_2 - \bar{\mu}^{\star}_1 v))^2dv > 0$.

Second, it holds that

$$\max_{g\in[G]}\sum_{i,j\in I_g}|\widetilde{w}_{gi}(h)\widetilde{w}_{gj}(h)| \le \max_{g\in[G]}\max_{i\in I_g} n^2_{g,h}\widetilde{w}_{gi}(h)^2 = O_P\left(\frac{1}{n^2h^2}\right)\max_{g\in[G]} n^2_{g,h}.$$

By Assumption AF-I, using Lemma C.2, or directly by Assumption AF-II it then follows that

$$\max_{g\in[G]}\sum_{i,j\in I_g}|\widetilde{w}_{gi}(h)\widetilde{w}_{gj}(h)| = o_P\left(\frac{1}{nh}\right). \tag{C.3}$$

Third, we show that

$$\sum_{g\in[G]}\sum_{i,j\in I_g}|\widetilde{w}_{gi}(h)\widetilde{w}_{gj}(h)| = \sum_{\star,\diamond\in\{+,-\}}\sum_{g\in[G]}\sum_{i,j\in I_g}|\widetilde{w}^{\star}_{gi}(h)\widetilde{w}^{\diamond}_{gj}(h)| = O_P\left(\frac{1}{nh}\right). \tag{C.4}$$

To prove this claim, note that

$$\begin{aligned}&\sum_{g\in[G]}\sum_{i,j\in I_g}|\widetilde{w}^{\star}_{gi}(h)\widetilde{w}^{\diamond}_{gj}(h)|\\ &\quad\le \frac{1}{n^2}\sum_{g\in[G]}\sum_{i,j\in I_g}k^{\star}_h(X_{gi})k^{\diamond}_h(X_{gj})\left|\left(S^{\star}_{n,2} - S^{\star}_{n,1}(X_{gi}/h)\right)\left(S^{\diamond}_{n,2} - S^{\diamond}_{n,1}(X_{gj}/h)\right)\right|\\ &\quad\le S^{\star}_{n,2}S^{\diamond}_{n,2}V^{\star\diamond}_{n,0,0} + |S^{\star}_{n,2}S^{\diamond}_{n,1}V^{\star\diamond}_{n,0,1}| + |S^{\star}_{n,1}S^{\diamond}_{n,2}V^{\star\diamond}_{n,1,0}| + |S^{\star}_{n,1}S^{\diamond}_{n,1}V^{\star\diamond}_{n,1,1}|,\end{aligned} \tag{C.5}$$

where

$$V^{\star\diamond}_{n,l,m} = \frac{1}{n^2}\sum_{g\in[G]}\sum_{i,j\in I_g}k^{\star}_h(X_{gi})k^{\diamond}_h(X_{gj})(X_{gi}/h)^l(X_{gj}/h)^m = U^{\star\diamond}_{n,l,m} + T^{\star}_{n,l}\mathbf{1}\{\star=\diamond\}.$$

Under the assumptions of the proposition, $V^{\star\diamond}_{n,l,m} = O_P((nh)^{-1})$ by Lemma C.1, and the conclusion in (C.4) follows.

Combining steps (C.1)–(C.4), both conditions of Assumption 1 follow:

$$\max_{g\in[G]}\sum_{i,j\in I_g}\frac{|w_{gi}(h)w_{gj}(h)|}{se^2(h)}\leq(C+o_P(1))\frac{\max_{g\in[G]}\sum_{i,j\in I_g}|\widetilde{w}_{gi}(h)\widetilde{w}_{gj}(h)|}{\sum_{g\in[G]}\sum_{i\in I_g}\widetilde{w}_{gi}(h)^2}=o_p(1),$$

$$\sum_{g\in[G]}\sum_{i,j\in I_g}\frac{|w_{gi}(h)w_{gj}(h)|}{se^2(h)}\leq(C+o_P(1))\frac{\sum_{g\in[G]}\sum_{i,j\in I_g}|\widetilde{w}_{gi}(h)\widetilde{w}_{gj}(h)|}{\sum_{g\in[G]}\sum_{i\in I_g}\widetilde{w}_{gi}(h)^2}=O_P(1).$$

Rate of the conditional variance: We have already shown that

$$se^2(h)\geq(C+o_P(1))/(nh).$$

Moreover,

$$se^2(h)\leq(C+o_P(1))\sum_{g\in[G]}\sum_{i,j\in I_g}|\widetilde{w}_{gi}(h)\widetilde{w}_{gj}(h)|=O_P\left(\frac{1}{nh}\right).$$

where the first step uses the assumption that the conditional variances (and hence also covariances) are bounded and the second step uses (C.4). Together, these results imply that $se^2(h)\asymp_p(nh)^{-1}$.

Limit of the conditional worst-case bias: Note that

$$\sum_{g\in[G]}\sum_{i\in I_g}w^\star_{gi}(h)X_{gi}^2=\frac{(S^\star_{n,2})^2-S^\star_{n,1}S^\star_{n,3}}{S^\star_{n,2}S^\star_{n,0}-(S^\star_{n,1})^2}h^2=(\bar{\mu}+o_P(1))h^2. \tag{C.6}$$

where the second step follows by Lemma C.1. □

C.3. **Proof of Proposition 2.** The proof has a similar structure as the proof of Proposition 1, except that we directly use the assumed order of the conditional variance, rather than derive it. To begin with, we note that under either Assumption AF-III or AF-IV, Lemma C.1 yields $S^\star_{n,l}=\bar{\mu}^\star_l f(0)+o_P(1)$ for $\star\in\{+,-\}$ and $l\in\mathbb{N}_0$. It follows that $S^+_{n,2}S^+_{n,0}-(S^+_{n,1})^2=S^-_{n,2}S^-_{n,0}-(S^-_{n,1})^2+o_P(1)=C+o_P(1)$ for a positive constant C, and the limit of the worst-case conditional bias follows as in (C.6).

We verify Assumption 1 in two steps; first under Assumption AF-III and then under Assumption AF-IV

Part 1: Suppose that Assumption AF-III holds. First, by Lemma C.2,

$$\max_{g\in[G]}\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)|=O_P\left(\frac{1}{n^2h^2}\right)\max_{g\in[G]}n_{g,h}^2=O_P\left(\frac{\max_{g\in[G]}n_g^2h^2+\log^2G}{n^2h^2}\right).$$

Second, using the inequality (C.5), Lemma C.1 yields

$$\sum_{g\in[G]}\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)| = O_P\left(\frac{1+\lambda_n}{nh}\right).$$

Since by assumption $se^2(h) \asymp_P (1+\lambda_n)/(nh)$, Assumption 1 follows under the assumptions made.

Part 2: Now suppose that Assumption AF-IV holds. First, note that

$$\max_{g\in[G]}\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)| = O_P\left(\frac{\max_{g\in[G]} n_g^2}{n^2h^2}\right).$$

Since by assumption $se^2(h) \asymp_P (1+\lambda_n/h)/(nh) = \sum_{g\in[G]} n_g^2/(n^2h)$, we obtain that

$$\frac{\max_{g\in[G]}\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)|}{se^2(h)} = O_P\left(\frac{\max_{g\in[G]} n_g^2}{h\sum_{g\in[G]} n_g^2}\right).$$

Part (i) of Assumption 1 follows.

Second, using the inequality (C.5), Lemma C.1 yields

$$\sum_{g\in[G]}\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)| = O_P\left(\frac{1+\lambda_n/h}{nh}\right).$$

This shows that part (ii) of Assumption 1 is satisfied. □

C.4. **Proofs of Lemmas 1 and 2. Part (i):** Under Assumption 4(i), by the triangle inequality,

$$se^2(h) \le C\sum_{g\in[G]}\sum_{i,j\in I_g}|w_{gi}(h)w_{gj}(h)|$$

for a positive constant C. The conclusions follow using the inequality (C.5) and Lemma C.1.

Part (ii): Under Assumption 5, we have that

$$\begin{aligned}
se^2(h) &= \sum_{g\in[G]}\sum_{i\in I_g} w_{gi}(h)^2\sigma_{g,i}^2 + \sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} w_{gi}(h)w_{gj}(h)\sigma_{g,ij} \\
&= \sum_{\star\in\{+,-\}}\sigma^2(0^\star)\sum_{g\in[G]}\sum_{i\in I_g} w_{gi}^\star(h)^2(1+o_P(1)) \\
&\quad + \sum_{\star,\diamond\in\{+,-\}}\mathbf{1}^{\pm}_{\{\star=\diamond\}}\sigma(0^\star,0^\diamond)\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} w_{gi}^\star(h)w_{gj}^\diamond(h)(1+o_P(1)).
\end{aligned}$$

Recall that $w_{gi}^\star(h) = \frac{1}{n}k_h^\star(X_{gi})\left(S_{n,2}^\star - S_{n,1}^\star(X_{gi}/h)\right)/(S_{n,2}^\star S_{n,0}^\star - (S_{n,1}^\star)^2)$. Using Lemma C.1,

we obtain that

$$\sum_{g\in[G]}\sum_{i\in I_g} w^{\star}_{gi}(h)^2 = \frac{1}{nh}\frac{\bar{\kappa}}{f(0)}(1+o_P(1)).$$

The second component satisfies:

$$\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} w^{\star}_{gi}(h)w^{\diamond}_{gj}(h) = \frac{\bar{\mu}^{\star}_2\bar{\mu}^{\diamond}_2 U^{\star\diamond}_{n,0,0} - \bar{\mu}^{\star}_2\bar{\mu}^{\diamond}_1 U^{\star\diamond}_{n,0,1} - \bar{\mu}^{\star}_1\bar{\mu}^{\diamond}_2 U^{\star\diamond}_{n,1,0} + \bar{\mu}^{\star}_1\bar{\mu}^{\diamond}_1 U^{\star\diamond}_{n,1,1}}{\left(\bar{\mu}^{\star}_2\bar{\mu}^{\star}_0 - (\bar{\mu}^{\star}_1)^2\right)\left(\bar{\mu}^{\diamond}_2\bar{\mu}^{\diamond}_0 - (\bar{\mu}^{\diamond}_1)^2\right) f_X^2(0)}(1+o_P(1)).$$

Under the assumptions of Lemma 1, plugging in $nhU^{\star\diamond}_{n,l,m} = \lambda_n\bar{\mu}^{\star}_l\bar{\mu}^{\diamond}_m f(0,0) + o_P(1+\lambda_n)$, we obtain that

$$\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} w^{\star}_{gi}(h)w^{\diamond}_{gj}(h) = \frac{1}{nh}\left(\frac{f(0,0)}{f_X^2(0)}\lambda_n + o_p(1+\lambda_n)\right).$$

Under the assumptions of Lemma 2, plugging in $nhU^{\star\diamond}_{n,l,m} = \lambda_n\bar{\kappa}^{\star}_{l+m}f_X(0)/h + o_P(1+\lambda_n/h)$, we obtain that

$$\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} w^{\star}_{gi}(h)w^{\diamond}_{gj}(h) = \frac{1}{nh}\left(\frac{\bar{\kappa}}{f_X(0)}\frac{\lambda_n}{h} + o_P\left(1+\frac{\lambda_n}{h}\right)\right)\mathbf{1}\{\star=\diamond\}.$$

Part (ii) of both lemmas follows. □

D. PROOFS OF AUXILIARY LEMMAS

D.1. **Proof of Lemma C.1.** In case (A), for any $g \in [G]$ and $i, j \in I_g$, $i \neq j$, let $f_{X_{gi},X_{gj}}$ denote the joint density of X_{gi} and X_{gj}. Note that it is uniformly bounded under Assumption AF-I(i).

Part (i). We study the expectation and the variance of $S^{\star}_{n,l}$. First, note that

$$\mathbb{E}[S^{\star}_{n,l}] = \mathbb{E}[k^{\star}_h(X_{gi})(X_{gi}/h)^l] = f(0)\bar{\mu}^{\star}_l + o(1).$$

Second, it holds that

$$\begin{aligned}
\operatorname{Var}(S^{\star}_{n,l}) &= \frac{1}{(nh)^2}\sum_{g\in[G]}\operatorname{Var}\left(\sum_{i\in I_g} k^{\star}(X_{gi}/h)(X_{gi}/h)^l\right)\\
&= \frac{1}{(nh)^2}\sum_{g\in[G]} n_g \operatorname{Var}(k^{\star}(X_{g1}/h)(X_{g1}/h)^l)\\
&\quad + \frac{1}{(nh)^2}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}}\operatorname{Cov}(k^{\star}(X_{gi}/h)(X_{gi}/h)^l,\, k^{\star}(X_{gj}/h)(X_{gj}/h)^l)
\end{aligned}$$

We analyze this variance in cases (A) and (B) separately, using standard kernel cal-

culations. In Case (A),

$$\begin{aligned}
\mathrm{Var}(S^{\star}_{n,j}) &\leq \frac{1}{(nh)^2}\sum_{g\in[G]} n_g \int (k^{\star}(x/h)(x/h)^l)^2 f_X(x)dx \\
&\quad + \frac{1}{(nh)^2}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} \left(\int\int k^{\star}(x/h)(x/h)^l k^{\star}(z/h)(z/h)^l f_{X_{gi},X_{gj}}(x,z)dxdz\right) \\
&= \frac{1}{n^2 h}\sum_{g\in[G]} n_g \int (k^{\star}(v)(v)^l)^2 f(vh)dv \\
&\quad + \frac{1}{n^2}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} \int\int k^{\star}(v)v^l k^{\star}(w)w^l f_{X_{gi},X_{gj}}(vh,wh)dvdw \\
&= O\left(\frac{1}{nh} + \frac{\sum_{g\in[G]} n_g(n_g-1)}{n^2}\right) = o(1).
\end{aligned}$$

In Case (B), we employ the inequality

$$\mathrm{Cov}(k^{\star}(X_{gi}/h)(X_{gi}/h)^l,\, k^{\star}(X_{gj}/h)(X_{gj}/h)^l) \leq \mathrm{Var}(k^{\star}(X_{g1}/h)(X_{g1}/h)^l).$$

With this bound, we obtain that

$$\begin{aligned}
\mathrm{Var}(S^{\star}_{n,j}) &\leq \frac{1}{(nh)^2}\sum_{g\in[G]} n_g^2 \,\mathrm{Var}(k^{\star}(X_{g1}/h)(X_{g1}/h)^l) \\
&\leq \frac{1}{(nh)^2}\sum_{g\in[G]} n_g^2 \int (k^{\star}(x/h)(x/h)^l)^2 f(x)dx \\
&= \frac{1}{n^2 h}\sum_{g\in[G]} n_g^2 \int (k^{\star}(v)(v)^l)^2 f(vh)dv \\
&= O\left(\frac{\sum_{g\in[G]} n_g^2}{n^2 h}\right) = o(1).
\end{aligned}$$

This concludes the proof of part (i).

Part (ii). We study the expectation and the variance of $T^{\star}_{n,l}$. First, note that

$$\begin{aligned}
\mathbb{E}[T^{\star}_{n,l}] &= \frac{1}{n}\mathbb{E}[(k^{\star}_h(X_{gi}))^2 (X_{gi}/h)^l] \\
&= \frac{1}{n}\int (k^{\star}_h(x))^2 (x/h)^l f(x)dx \\
&= \frac{1}{nh}(\bar{\kappa}^{\star}_l f(0) + o(1)).
\end{aligned}$$

Second, it holds that

$$\begin{aligned}
\mathrm{Var}(T^{\star}_{n,l}) &= \mathrm{Var}\left(\frac{1}{n^2h^2}\sum_{g\in[G]}\sum_{i\in I_g} k^{\star}(X_{gi}/h)^2(X_{gi}/h)^l\right)\\
&= \frac{1}{(nh)^4}\sum_{g\in[G]} n_g\,\mathrm{Var}((k^{\star}(X_{g1}/h))^2(X_{g1}/h)^l)\\
&\quad + \frac{1}{(nh)^4}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} \mathrm{Cov}((k^{\star}(X_{gi}/h))^2(X_{gi}/h)^l,\,(k^{\star}(X_{gj}/h))^2(X_{gj}/h)^l).
\end{aligned}$$

We study this variance separately in cases (A) and (B), using standard kernel derivations. In Case (A),

$$\begin{aligned}
\mathrm{Var}(nhT^{\star}_{n,j}) &\leq \frac{1}{(nh)^2}\sum_{g\in[G]} n_g\int((k^{\star}(x/h))^2(x/h)^l)^2 f(x)dx\\
&\quad + \frac{1}{(nh)^2}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}}\left(\int\int(k^{\star}(x/h))^2(x/h)^l(k^{\star}(z/h))^2(z/h)^l f(x,z)dxdz\right)\\
&= \frac{1}{n^2h}\sum_{g\in[G]} n_g\int((k^{\star}(v))^2(v)^l)^2 f(vh)dv\\
&\quad + \frac{1}{n^2}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}}\int\int(k^{\star}(v))^2v^l(k^{\star}(w))^2w^l f(vh,wh)dvdw\\
&= O\left(\frac{1}{nh} + \frac{\sum_{g\in[G]} n_g(n_g-1)}{n^2}\right) = o(1).
\end{aligned}$$

In Case (B),

$$\begin{aligned}
\mathrm{Var}(nhT^{\star}_{n,l}) &\leq \frac{1}{(nh)^2}\sum_{g\in[G]} n_g^2\,\mathrm{Var}(k^{\star}(X_{g1}/h)^2(X_{g1}/h)^l)\\
&\leq \frac{1}{(nh)^2}\sum_{g\in[G]} n_g^2\int(k^{\star}(x/h)^2(x/h)^l)^2 f(x)dx\\
&= \frac{1}{n^2h}\sum_{g\in[G]} n_g^2\int(k^{\star}(v)^2(v)^l)^2 f(vh)dv\\
&= O\left(\frac{\sum_{g\in[G]} n_g^2}{n^2h}\right) = o(1),
\end{aligned}$$

where the first inequality follows by the Cauchy-Schwarz inequality for the covariances. This concludes the proof of part (ii).

Part (iii). We study the expectation and the variance of $U^{\star\diamond}_{n,l,m}$ in Case (A). First, by

standard kernel derivations,

$$\begin{aligned}
\mathbb{E}[U^{\star\diamond}_{n,l,m}] &= \frac{1}{n^2}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} \mathbb{E}[k^{\star}_h(X_{gi})(X_{gi}/h)^l k^{\diamond}_h(X_{gj})(X_{gj}/h)^m] \\
&= \frac{1}{n^2}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} \int\int k^{\star}_h(x_1)(x_1/h)^l k^{\diamond}_h(x_2)(x_2/h)^m f_{X_{gi},X_{gj}}(x_1,x_2)dx_1dx_2 \\
&= \frac{1}{n^2}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} \int\int k^{\star}(v)v^l k^{\diamond}(w)(w)^m f_{X_{gi},X_{gj}}(vh,wh)dvdw \\
&= O\left(\frac{1}{n^2}\sum_{g\in[G]} n_g(n_g-1)\right).
\end{aligned}$$

Next, we consider the variance. Let

$$\begin{aligned}
C_g(i_1,j_1,i_2,j_2) = \mathrm{Cov}\Big(&k^{\star}_h(X_{gi_1})k^{\diamond}_h(X_{gj_1})(X_{gi_1}/h)^l(X_{gj_1}/h)^m, \\
&k^{\star}_h(X_{gi_2})k^{\diamond}_h(X_{gj_2})(X_{gi_2}/h)^l(X_{gj_2}/h)^m\Big).
\end{aligned}$$

Note that for all indices $i_1, j_1, i_2, j_2 \in I_g$ such that $i_1 \neq j_1$ and $i_2 \neq j_2$,

$$|C_g(i_1,j_1,i_2,j_2)| \leq \begin{cases} C/h^2 & \text{if } i_1 = i_2 \text{ and } j_1 = j_2, \\ C/h & \text{if there is exactly one pair of equal indices,} \\ C & \text{if } i_1, i_2, j_1, j_2 \text{ are pairwise different.} \end{cases}$$

for some constant C. We have that

$$\begin{aligned}
\mathrm{Var}(U^{\star\diamond}_{n,l,m}) &= \frac{1}{n^4}\sum_{g\in[G]}\mathrm{Var}\left(\sum_{\substack{i\neq j\\ i,j\in I_g}} k^{\star}_h(X_{gi})k^{\diamond}_h(X_{gj})(X_{gi}/h)^l(X_{gj}/h)^m\right)\\
&= \frac{1}{n^4}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}} C_g(i,j,i,j) + \frac{1}{n^4}\sum_{g\in[G]}\sum_{\substack{i_1,j_1,i_2,j_2\in I_g\\ \text{pairwise different}}} C_g(i_1,j_1,i_2,j_2)\\
&\quad + \frac{1}{n^4}\sum_{g\in[G]}\sum_{\substack{i_1\neq j_1, i_2\neq j_2\\ i_1=i_2, j_1\neq j_2\\ i_1,j_1,i_2,j_2\in I_g}} C_g(i_1,j_1,i_2,j_2) + \frac{1}{n^4}\sum_{g\in[G]}\sum_{\substack{i_1\neq j_1, i_2\neq j_2\\ i_1\neq i_2, j_1=j_2\\ i_1,j_1,i_2,j_2\in I_g}} C_g(i_1,j_1,i_2,j_2)\\
&\quad + \frac{1}{n^4}\sum_{g\in[G]}\sum_{\substack{i_1\neq j_1, i_2\neq j_2\\ i_1=j_2, i_2\neq j_1\\ i_1,j_1,i_2,j_2\in I_g}} C_g(i_1,j_1,i_2,j_2) + \frac{1}{n^4}\sum_{g\in[G]}\sum_{\substack{i_1\neq j_1, i_2\neq j_2\\ i_2=j_1, i_1\neq j_2\\ i_1,j_1,i_2,j_2\in I_g}} C_g(i_1,j_1,i_2,j_2)\\
&= O\left(\frac{1}{n^4h^2}\sum_{g\in[G]} n_g^2 + \frac{1}{n^4h}\sum_{g\in[G]} n_g^3 + \frac{1}{n^4}\sum_{g\in[G]} n_g^4\right)\\
&= O\left(\frac{1}{(nh)^4}\sum_{g\in[G]}(n_gh)^2 + \frac{1}{(nh)^4}\sum_{g\in[G]}(n_gh)^4\right),
\end{aligned}$$

where the last step uses the Cauchy-Schwarz inequality. The first statement of part (iii) follows by noting that

$$\begin{aligned}
nhU^{\star\diamond}_{n,l,m} &= O_P\left(\frac{1}{nh}\sum_{g\in[G]}(n_gh)^2 + \frac{1}{nh}\sqrt{\sum_{g\in[G]}(n_gh)^2} + \frac{1}{nh}\sqrt{\sum_{g\in[G]}(n_gh)^4}\right)\\
&= O_P\left(\frac{1}{nh}\sum_{g\in[G]}(n_gh)^2 + \frac{1}{nh}\left(1+\sum_{g\in[G]}(n_gh)^2\right) + \frac{1}{nh}\sqrt{\left(\sum_{g\in[G]}(n_gh)^2\right)^2}\right)\\
&= O_P\left(\frac{1}{nh}\sum_{g\in[G]}(n_gh)^2 + \frac{1}{nh}\right).
\end{aligned}$$

To prove the second claim, note that if all pairs (X_{gi}, X_{gj}), $i \neq j$, are identically

distributed with continuous joint density $f(x_1, x_2)$, then

$$\begin{aligned}
\mathbb{E}[nhU^{\star\diamond}_{n,l,m}] &= \frac{nh}{n^2}\sum_{g\in[G]} n_g(n_g-1)\int\int k^\star_h(x_1)(x_1/h)^l k^\diamond_h(x_2)(x_2/h)^m f(x_1,x_2)dx_1dx_2 \\
&= \lambda_n \int\int k^\star(v)v^l k^\diamond(w)(w)^m f(vh,wh)dvdw \\
&= \lambda_n \int\int k^\star(v)v^l k^\diamond(w)(w)^m dvdw f(0^\star,0^\diamond) + o(1+o(\lambda_n)) \\
&= \lambda_n \bar{\mu}^\star_l \bar{\mu}^\diamond_m f(0^\star,0^\diamond) + o(1+o(\lambda_n)).
\end{aligned}$$

If in addition,

$$\frac{\max_{g\in[G]}(n_gh)^2}{nh+\sum_{g\in[G]}(n_gh)^2} = o(1),$$

then we obtain that

$$\begin{aligned}
nhU^{\star\diamond}_{n,l,m} &= \lambda_n\bar{\mu}^\star_l\bar{\mu}^\diamond_m f(0^\star,0^\diamond)(1+o(1)) + O_P\left(\sqrt{\frac{1}{(nh)^2}\sum_{g\in[G]}(n_gh)^2\left(1+\max_{g\in[G]}(n_gh)^2\right)}\right) \\
&= \lambda_n\bar{\mu}^\star_l\bar{\mu}^\diamond_m f(0^\star,0^\diamond)(1+o(1)) \\
&\quad + O_P\left(\sqrt{\frac{(nh+\sum_{g\in[G]}(n_gh)^2)\sum_{g\in[G]}(n_gh)^2}{(nh)^2}\frac{1+\max_{g\in[G]}(n_gh)^2}{nh+\sum_{g\in[G]}(n_gh)^2}}\right) \\
&= \lambda_n\bar{\mu}^\star_l\bar{\mu}^\diamond_m f(0^\star,0^\diamond)(1+o_P(1)) + o_P(1).
\end{aligned}$$

This concludes the proof of part (iii).

Part (iv). We study the expectation and the variance of $U^{\star\diamond}_{n,l,m}$ in Case (B). First, note that

$$\begin{aligned}
|\mathbb{E}[U^{\star\diamond}_{n,l,m}]| &\leq \frac{1}{n^2}\sum_{g\in[G]}\sum_{\substack{i\neq j\\ i,j\in I_g}}\mathbb{E}[k^\star_h(X_{gi})k^\diamond_h(X_{gj})] \\
&\leq \frac{1}{n^2}\sum_{g\in[G]} n_g(n_g-1)\mathbb{E}[(k^\star_h(X_{g1}))^2] \\
&= O\left(\frac{1}{n^2h}\sum_{g\in[G]} n_g(n_g-1)\right),
\end{aligned}$$

where the second inequality follows by the Cauchy-Schwarz inequality. If all the realiza-

tions of the running variable are equal within each cluster and $f(0) > 0$, then

$$\begin{aligned}\mathbb{E}[U^{\star\diamond}_{n,l,m}] &= \frac{1}{n^2}\sum_{g\in[G]} n_g(n_g-1)\mathbb{E}\big[(k^\star_h(X_{g1}))^2(X_{g1}/h)^{l+m}\big] \\ &= \frac{1}{n^2}\sum_{g\in[G]} n_g(n_g-1)\bar{\kappa}^*_{l+m} f(0)(1+o(1)).\end{aligned}$$

The variance of $U^{\star\diamond}_{n,l,m}$ is bounded as follows:

$$\begin{aligned}\mathrm{Var}(U^{\star\diamond}_{n,l,m}) &= \frac{1}{n^4}\sum_{g\in[G]} \mathrm{Var}\left(\sum_{\substack{i\neq j\\ i,j\in I_g}} k^\star_h(X_{gi})k^\diamond_h(X_{gj})(X_{gi}/h)^l(X_{gj}/h)^m\right) \\ &= \frac{1}{n^4}\sum_{g\in[G]}\sum_{\substack{i_1\neq j_1\\ i_1,j_1\in I_g}}\sum_{\substack{i_2\neq j_2\\ i_2,j_2\in I_g}} C_g(i_1,j_1,i_2,j_2) \\ &\leq \frac{1}{n^4}\sum_{g\in[G]}(n_g(n_g-1))^2\mathbb{E}\big[k_h(X_{g1})^4\big] \\ &= O\left(\frac{1}{n^4h^3}\sum_{g\in[G]} n_g^4\right) \\ &= O\left(\frac{1}{n^2h^2}\frac{1}{n}\sum_{g\in[G]} n_g^2\frac{\max_{g\in[G]} n_g^2}{nh}\right),\end{aligned}$$

where $C_g(i_1,j_1,i_2,j_2)$ denotes respective variances, as defined in the proof of part (iii). Both statements in part (iv) follow from the above observations. □

D.2. **Proof of Lemma C.2.** By the union bound and Chernoff's inequality, for any $B > 0$ and $t > 0$,

$$\Pr\left(\max_{g\in[G]} n_{g,h} > B\right) \leq G\max_{g\in[G]}\Pr\left(n_{g,h} > B\right) \leq G\frac{\max_{g\in[G]}\mathbb{E}[e^{tn_{g,h}}]}{e^{tB}}.$$

Next, for any $g \in [G]$,

$$\begin{aligned}
\mathbb{E}[e^{tn_{g,h}}] &= \sum_{k=0}^{n_g} e^{tk} \Pr(n_{g,h} = k) \\
&\leq \sum_{k=0}^{n_g} e^{tk} \sum_{1 \leq i_1 < \ldots < i_k \leq n_g} \Pr\left(\bigcap_{l=1}^{k} \{|X_{gi_l}| \leq h\} \right) \\
&\leq \sum_{k=0}^{n_g} e^{tk} \binom{n_g}{k} C(2h)^k \\
&= C(2he^t + 1)^{n_g} \\
&\leq C e^{2n_g h e^t},
\end{aligned}$$

where the second step uses the union bound, the third step uses the assumption of bounded join densities, the fourth step uses the binomial formula, and the last step uses the bound $1 + x \leq e^x$.

It follows that

$$\Pr\left(\max_{g \in [G]} n_{g,h} > B \right) \leq CG \frac{e^{mhe^t}}{e^{tB}}.$$

where $m = 2 \max_{g \in [G]} n_g$.

Letting $B = \frac{e^t}{t}(mh + \log G)$, we obtain

$$\begin{aligned}
\Pr\left(\max_{g \in [G]} n_{g,h} > \frac{e^t}{t}(mh + \log G) \right) &\leq C \exp\left(\log G + mhe^t - t\frac{e^t}{t}(mh + \log G) \right) \\
&= C \exp\left(-(e^t - 1) \log G \right).
\end{aligned}$$

The last bound can be made arbitrarily small by choosing t large enough, which concludes the proof. □

REFERENCES

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023): "When should you adjust standard errors for clustering?" *The Quarterly Journal of Economics*, 138, 1–35.

Abadie, A., G. W. Imbens, and F. Zheng (2014): "Inference for misspecified models with fixed regressors," *Journal of the American Statistical Association*, 109, 1601–1614.

Arellano, M. (1987): "Computing robust standard errors for within-groups estimators," *Oxford Bulletin of Economics & Statistics*, 49.

Armstrong, T. B. and M. Kolesár (2020): "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*, 11, 1–39.

Bartalotti, O. and Q. Brummet (2017): "Regression discontinuity designs with clustered data," in *Regression discontinuity designs*, Emerald Publishing Limited, vol. 38, 383–420.

Bhattacharya, D. (2005): "Asymptotic inference from multi-stage samples," *Journal of Econometrics*, 126, 145–171.

Bugni, F., I. A. Canay, A. M. Shaikh, and M. Tabord-Meehan (2025): "Inference for cluster randomized experiments with nonignorable cluster sizes," *Journal of Political Economy Microeconomics*, 3, 255–288.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019): "Regression Discontinuity Designs Using Covariates," *The Review of Economics and Statistics*, 101, 442–451.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 82, 2295–2326.

Cameron, A. C. and D. L. Miller (2015): "A practitioners guide to cluster-robust inference," *Journal of Human Resources*, 50, 317–372.

Chiang, H. D., Y. Sasaki, and Y. Wang (2025): "Genuinely robust inference for clustered data," *arXiv preprint arXiv:2308.10138*.

del Valle, A., A. de Janvry, and E. Sadoulet (2020): "Rules for recovery: Impact of indexed disaster funds on shock coping in Mexico," *American Economic Journal: Applied Economics*, 12, 164–195.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019): "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 212, 393–412.

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications*, Chapman & Hall/CRC.

Ghosh, A., G. Imbens, and S. Wager (2025): "Plrd: partially linear regression discontinuity inference," *arXiv preprint arXiv:2503.09907*.

Granzier, R., V. Pons, and C. Tricaud (2023): "Coordination and bandwagon effects: How past rankings shape the behavior of voters and candidates," *American Economic Journal: Applied Economics*, 15, 177–217.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

Hansen, B. E. (2025): "Jackknife standard errors for clustered regression," *Working Paper*.

Hansen, B. E. and S. Lee (2019): "Asymptotic theory for clustered samples," *Journal of Econometrics*, 210, 268–290.

Imbens, G. and K. Kalyanaraman (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *Review of Economic Studies*, 79, 933–959.

Imbens, G. and S. Wager (2019): "Optimized regression discontinuity designs," *Review of Economics and Statistics*, 101.

Johnson, M. S. (2020): "Regulation by shaming: Deterrence effects of publicizing violations of workplace safety and health laws," *American economic review*, 110, 1866–1904.

Liang, K.-Y. and S. L. Zeger (1986): "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Lin, X. and R. J. Carroll (2000): "Nonparametric function estimation for clustered data when the predictor is measured without/with error," *Journal of the American statistical Association*, 95, 520–534.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023): "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, 232, 272–299.

Noack, C., T. Olma, and C. Rothe (2025): "Flexible covariate adjustments in regression discontinuity designs," *arXiv preprint arXiv:2107.07942*.

Noack, C. and C. Rothe (2024): "Bias-aware inference in fuzzy regression discontinuity designs," *Econometrica*.

Shimizu, Y. (2025): "Nonparametric regression under cluster sampling," *Journal of Econometrics*, 252, 106102.

Wang, N. (2003): "Marginal nonparametric kernel regression accounting for within-subject correlation," *Biometrika*, 90, 43–52.

Wasserman, M. (2021): "Up the political ladder: Gender parity in the effects of electoral defeats," *AEA Papers and Proceedings*, 111, 169–173.

White, H. (2014): *Asymptotic theory for econometricians*, Academic press.

Zhang, J.-T. and J. Chen (2007): "Statistical inferences for functional data," *The Annals of Statistics*.