Proceedings of Machine Learning Research vol 195:1–92, 2023 36th Annual Conference on Learning Theory

# A High-dimensional Convergence Theorem for U-statistics with Applications to Kernel-based Testing

**Kevin Han Huang** HAN.HUANG.20@UCL.AC.UK
*Gatsby Unit, University College London*

**Xing Liu** XING.LIU16@IMPERIAL.AC.UK
*Department of Mathematics, Imperial College London*

**Andrew B. Duncan** A.DUNCAN@IMPERIAL.AC.UK
*Department of Mathematics, Imperial College London and Alan Turing Institute*

**Axel Gandy** A.GANDY@IMPERIAL.AC.UK
*Department of Mathematics, Imperial College London*

**Editors:** Gergely Neu and Lorenzo Rosasco

## Abstract

We prove a convergence theorem for U-statistics of degree two, where the data dimension $d$ is allowed to scale with sample size $n$. We find that the limiting distribution of a U-statistic undergoes a phase transition from the non-degenerate Gaussian limit to the degenerate limit, regardless of its degeneracy and depending only on a moment ratio. A surprising consequence is that a non-degenerate U-statistic in high dimensions can have a non-Gaussian limit with a larger variance and asymmetric distribution. Our bounds are valid for any finite $n$ and $d$, independent of individual eigenvalues of the underlying function, and dimension-independent under a mild assumption. As an application, we apply our theory to two popular kernel-based distribution tests, MMD and KSD, whose high-dimensional performance has been challenging to study. In a simple empirical setting, our results correctly predict how the test power at a fixed threshold scales with $d$ and the bandwidth.

**Keywords:** High-dimensional statistics, U-statistics, distribution testing, kernel method

## 1. Introduction

We consider a one-dimensional U-statistic of degree two built on $n$ i.i.d. data points in $\mathbb{R}^d$. Numerous estimators can be formulated as a U-statistic: Modern applications include high-dimensional change-point detection (Wang et al., 2022), sensitivity analysis of algorithms (Gamboa et al., 2022) and convergence guarantees for random forests (Peng et al., 2022).

The asymptotic theory of U-statistics is well-established in the classical setting, where $d$ is fixed and small relative to $n$ (e.g. Chapter 5 of Serfling (1980)). Classical theory shows that the large-sample asymptotic of a U-statistic depends on its martingale structure and moments: For U-statistics of degree two, this reduces to the notion of *degeneracy*, i.e. whether the variance of a certain conditional mean is zero. Non-degenerate U-statistics are shown to have a Gaussian limit, whereas degenerate ones converge to an infinite sum of weighted chi-squares.

However, these results fail to apply to the modern context of high-dimensional data, where $d$ is of a comparable size to $n$. The key issue is that the moment terms, which determine degeneracy, may scale with $d$. Existing efforts on high-dimensional results either focus on U-statistics of a growing degree (Song et al., 2019; Chen and Kato, 2019) and of growing output dimension (Chen, 2018) or rely on very specific data structures (Chen and Qin, 2010; Yan and Zhang, 2022). In particular,

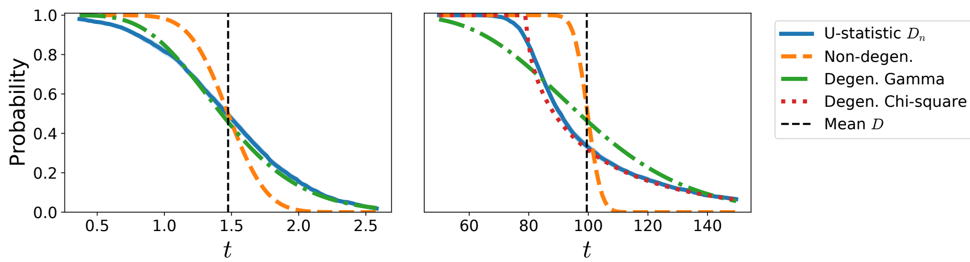


Figure 1: Behaviour of $\mathbb{P}(X > t)$ for $X = D_n$, a non-degenerate U-statistic, versus $X$ being different theoretical limits. *Left.* KSD with RBF kernel, $n = 50$ and $d = 2000$. *Right.* MMD with linear kernel, $n = 50$ and $d = 1000$. The left plot shows that $\mathbb{P}(D_n > t)$ *disagrees* with the non-degenerate limit from known classical results but aligns with the degenerate limit from ours (moment-matched by a Gamma variable – discussed in Section 3.2). The right plot is when the limit predicted by our result can be computed exactly as a shifted-and-rescaled chi-square and shows asymmetry, which confirms a departure from Gaussianity. See the last paragraph of Section 4.3 and Appendix A for simulation details.

these articles focus on a comparison to some Gaussian limit in high dimensions, and the effect of moments on a departure from Gaussianity has largely been ignored.

The practical motivation for our work stems from distribution tests, which typically employ U-statistics as a test statistic. In the machine learning community, it has been empirically observed that the power of kernel-based distribution tests can deteriorate in high dimensions, depending on hyperparameter choices and the class of alternatives (Reddi et al., 2015; Ramdas et al., 2015). A theoretical analysis in the most general case has not been possible, due to the lack of a general convergence result for high-dimensional U-statistics. In the statistics community, there are similar interests in analysing U-statistics used in mean testing of high-dimensional data (e.g. Chen and Qin (2010); Wang et al. (2015)). All existing results, to our knowledge, are limited by very specific data assumptions and a focus on obtaining Gaussian limits.

In this paper, we prove a general convergence theorem for U-statistics of degree two, which holds in the high-dimensional setting and under very mild assumptions on the data. We observe a high-dimensional analogue of the classical behaviour: Depending on a moment ratio, the limiting distribution of U-statistics can take either the non-degenerate Gaussian limit, the degenerate limit or an intermediate distribution. Crucially, this happens *regardless of* the statistic's degeneracy, as defined in the classical sense. We provide error bounds that are finite-sample valid and *dimension-independent* under a mild assumption.

In the context of kernel-based distribution tests, we show that our results hold for *Maximum Mean Discrepancy* (MMD) and for *(Langevin) Kernelized Stein Discrepancy* (KSD) under some natural conditions. We investigate several examples under Gaussian mean-shift – a setting purposely chosen to be as simple as possible to obtain good intuitions, while already capturing a rich amount of complex behaviours. Our theory correctly predicts the high-dimensional behaviour of the test power with a wider variance than classical results and, perhaps surprisingly, potential asymmetry (see Fig. 1 for one such example). Our results enable us to characterise such behaviours based on the size of $d$ and hyperparameter choices.

### 1.1. Overview of results

Given some i.i.d. data $\{\mathbf{X}_i\}_{i=1}^n$ drawn from a distribution $R$ on $\mathbb{R}^d$ and a symmetric measurable function $u : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the goal is to estimate the quantity $D := \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)]$. The U-statistic provides an unbiased estimator, defined as

$$D_n \;\coloneqq\; \frac{1}{n(n-1)} \sum\nolimits_{1 \leq i \neq j \leq n} u(\mathbf{X}_i, \mathbf{X}_j) \;. \tag{1}$$

Our main result is Theorem 2. Loosely speaking, it says that as $n, d \to \infty$, the statistic $D_n$ converges in distribution to a quadratic form of Gaussians:

$$D_n \;\overset{d}{\to}\; W + Z + D \;, \tag{2}$$

where $W$ is some infinite sum of weighted and centred chi-squares and $Z$ is some Gaussian. Define

$$\rho_d \coloneqq \sigma_{\text{full}} \,/\, \sigma_{\text{cond}} \;, \text{ where } \sigma_{\text{full}} \;=\; \sqrt{\text{Var}[u(\mathbf{X}_1, \mathbf{X}_2)]} \quad \text{and} \quad \sigma_{\text{cond}} \;=\; \sqrt{\text{Var}\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1]} \;,$$

and recall that the classical notion of degeneracy is defined by $\sigma_{\text{cond}} = 0$. We next observe that in (2), $W + D$ is closely related to the classical degenerate limit, whereas $Z + D$ gives exactly the classical non-degenerate limit. It turns out that, up to a mild assumption, the type of asymptotic distribution of $D_n$ is *completely determined* by the ratio $\rho_d$. This is reminiscent of the classical result, where the notion of degeneracy, i.e. whether $\sigma_{\text{cond}} = 0$, determines the limit of $D_n$. The difference in high dimensions is that $\sigma_{\text{full}}$ and $\sigma_{\text{cond}}$ may scale *differently* with $d$. Even if $\sigma_{\text{cond}} \neq 0$, $\rho_d$ can grow to infinity as $d$ grows, causing a non-degenerate $D_n$ to behave like a degenerate U-statistic. We show that, depending on $\rho_d$, (2) becomes

$$D_n \;\overset{d}{\to}\; W + D \quad \text{for } \rho_d = \omega(n^{1/2}) \qquad \text{and} \qquad D_n \;\overset{d}{\to}\; Z + D \quad \text{for } \rho_d = o(n^{1/2}) \;.$$

The second result is the classical Berry-Esséen bound for U-statistics, while the first result is new. It recovers the classical degenerate limit as a special case but also applies to very general U-statistics in high dimensions regardless of degeneracy.

The paper is organised as follows. Section 2 provides definitions and a sketch-of-intuition on the role of moment terms in the limiting behaviour of $D_n$. Section 3 presents the main results along with a proof overview in Section 3.3. Section 4.2 shows that these results apply to MMD and KSD under some natural conditions and Section 4.3 studies the Gaussian mean-shift case in detail.

### 1.2. Related literature

*Convergence results for U-statistics.* Existing high-dimensional results focus either on a different setting or on showing asymptotic normality under very specific assumptions on data; some references are provided at the start of this section. The results that resemble our work more closely are finite-sample bounds for classical degenerate U-statistics. Those works focus on providing bounds under conditions on specific eigenvalues of a spectral decomposition of $D_n$, and we defer a list of references to Remark 1. Among them, Yanushkevichiene (2012) provides a rate $O(n^{-1/12})$ under perhaps the least stringent assumption on eigenvalues, but the error is still pre-multiplied by the inverse square-root of the largest eigenvalue. These eigenvalues are intractable and yet depend on $d$ through the data distribution, which make them hard to apply to high-dimensional settings. In the *classical* setting where $d$ is fixed, a recent work by Bhattacharya et al. (2022) proves a Gaussian-quadratic-form limit similar to ours for a random quadratic polynomial, which includes a simple U-statistic as a special case. However, their results are asymptotic and in particular do not identify a parameter that leads to the phase transition. Our finite-sample results require a very different proof technique and show how a moment ratio governs the transition.

*High-dimensional power analysis for MMD and KSD.* Some recent work has investigated the asymptotic behaviour of $D_n$ for MMD. Yan and Zhang (2022) prove a convergence result under a specific data model and kernel choice, so that $u(\mathbf{x}, \mathbf{y}) = g(\|\mathbf{x} - \mathbf{y}\|_2)$ for some function $g : \mathbb{R} \to \mathbb{R}$ and $\| \bullet \|_2$ being the vector norm. The dimension-independence of $g$ enables a Taylor expansion argument reminiscent of delta method and therefore gives a Gaussian limit. Such structures are not assumed in our work. A related work of Gao and Shao (2021) provides a finite-sample bound under more general conditions. The results show asymptotic normality of a studentised version of $D_n$ rather than $D_n$ itself, and the error bound is only valid if a moment ratio, analogous to excess kurtosis, vanishes with $d$ (see their Theorem 13). Interestingly, this effect is obtained as a special case of our results for much more general settings: In Section 3.2, we point out that the degenerate limit is Gaussian if and only if the excess kurtosis vanishes. Another recent line of work (Kim and Ramdas, 2020; Shekhar et al., 2022) focuses on a studentised $D_n$ that is modified to exclude half of the terms. They show dimension-agnostic normality results at the cost of not using the full U-statistic $D_n$.

## 2. Setup and motivation

We use the asymptotic notations $o, O, \Theta, \omega, \Omega$ defined in the usual way (see e.g. Chapter 3 of Cormen et al. (2009)) for the limit $n \to \infty$, where the dimension is allowed to depend on $n$; we make the $n$-dependence explicit in the dimension $d_n$ whenever such asymptotics are considered.

### 2.1. Moment terms in high dimensions

Consider a U-statistic $D_n$ as defined in (1) with respect to $(R, u)$ with mean $D = \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)]$. For $\nu \geq 1$, denote the $L_\nu$ norms by $\| \bullet \|_{L_\nu} := \mathbb{E}[| \bullet |^\nu]^{1/\nu}$. The $\nu$-th central moment of $D_n$ are bounded from above and below in terms of two types of moment terms (see Lemma 30 in the appendix):

$$M_{\text{cond};\nu} := \big\|\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_2] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)]\big\|_{L_\nu}, M_{\text{full};\nu} := \big\|u(\mathbf{X}_1, \mathbf{X}_2) - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)]\big\|_{L_\nu}.$$

In the special case $\nu = 2$, the definitions from Section 1.1 implies $\sigma_{\text{cond}} = M_{\text{cond};2}$, $\sigma_{\text{full}} = M_{\text{full};2}$ and $\rho_d = \sigma_{\text{full}} \,/\, \sigma_{\text{cond}}$. The fact that these moments may scale with $d$ has a significant effect on convergence results: For example, bounds of the form $\frac{\text{moment}}{f(n)}$ for some increasing function $f$ of $n$ are no longer guaranteed to be small. This is yet another effect of the "curse of dimensionality". For U-statistics, the classical Berry-Esséen result (see e.g. Theorem 10.3 of Chen et al. (2011)) says that, if $\sigma_{\text{cond}} > 0$, then for a normal random variable $Z \sim \mathcal{N}(D, 4n^{-1}\sigma^2_{\text{cond}})$ and $\nu \in (2, 3]$, we have

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\Big(\frac{\sqrt{n}}{\sigma_{\text{cond}}} D_n < t\Big) - \mathbb{P}\Big(\frac{\sqrt{n}}{\sigma_{\text{cond}}} Z < t\Big)\right| \;\leq\; \frac{6.1 M^\nu_{\text{cond};\nu}}{n^{(\nu-2)/2}\sigma^\nu_{\text{cond}}} + \frac{(1+\sqrt{2})\rho_d}{2(n-1)^{1/2}} \;. \tag{3}$$

Indeed, the error bound in the classical Berry-Esséen result is an increasing function of $n^{-1/2}\rho_d = \sigma_{\text{full}}/(n^{1/2}\sigma_{\text{cond}})$, which is not guaranteed to be small as $d$ grows.

The ratio $M_{\text{cond};\nu}/\sigma_{\text{cond}}$ also appears in classical error bounds. However, we do *not* focus on how this ratio scales, since it appears in Berry-Esséen bounds even for sample averages. Error bounds in our main theorem will depend on similar ratios, and for our theorem to imply a convergence theorem, the following assumption is required:

**Assumption 1** *There exists some $\nu \in (2, 3]$ and some constant $C < \infty$ such that for all $n$ and $d$, we have the uniform bounds $\frac{M_{\text{full};\nu}}{\sigma_{\text{full}}} \leq C$ and $\frac{M_{\text{cond};\nu}}{\sigma_{\text{cond}}} \leq C$ .*

### 2.2. Sketch of intuition

We motivate our results by noting that the variance of $D_n$ defined in (1) satisfies

$$\begin{aligned}\mathrm{Var}[D_n] &= O\Big(\frac{\mathbb{E}[(u(\mathbf{X}_1,\mathbf{X}_2)-D)(u(\mathbf{X}_1,\mathbf{X}_3)-D)]}{n} + \frac{\mathbb{E}[(u(\mathbf{X}_1,\mathbf{X}_2)-D)(u(\mathbf{X}_1,\mathbf{X}_2)-D)]}{n(n-1)}\Big)\\ &= O\big(\frac{\sigma^2_{\mathrm{cond}}}{n} + \frac{\sigma^2_{\mathrm{full}}}{n(n-1)}\big)\;.\end{aligned}$$

To study the asymptotic distribution of $D_n$, we need to understand how its asymptotic variance behaves as $n$ and $d$ grow. Suppose we are in the classical *non-degenerate* setting, where $d$ is fixed and $\sigma_{\mathrm{cond}} > 0$. The dominating term in $\mathrm{Var}[D_n]$ is $O(n^{-1}\sigma^2_{\mathrm{cond}})$. The contribution of the $\sigma^2_{\mathrm{full}}$ term is small, i.e. the effect of the variance of each individual summand $u(\mathbf{X}_1,\mathbf{X}_2)$ is negligible. In fact, we can approximate $D_n$ by replacing each argument in the summand by an independent copy $\mathbf{X}'_i$ of $\mathbf{X}_i$ and applying CLT for an empirical average:

$$\begin{aligned}D_n &= D + \frac{1}{n(n-1)}\sum\nolimits_{1\le i\neq j\le n}(u(\mathbf{X}_i,\mathbf{X}_j)-D)\\ &\approx D + \frac{1}{n}\sum\nolimits_{i=1}^n\Big(\frac{1}{n-1}\sum\nolimits_{j\neq i}(u(\mathbf{X}_i,\mathbf{X}'_j)-D)\Big) + \frac{1}{n}\sum\nolimits_{j=1}^n\Big(\frac{1}{n-1}\sum\nolimits_{i\neq j}(u(\mathbf{X}'_i,\mathbf{X}_j)-D)\Big)\\ &= D + \frac{2}{n}\sum\nolimits_{i=1}^n\big(\mathbb{E}[u(\mathbf{X}_i,\mathbf{X}'_j)|\mathbf{X}_i]-D\big) \;\approx\; \mathcal{N}\big(D\,,\,\frac{4\sigma^2_{\mathrm{cond}}}{n}\big)\;.\end{aligned}$$

This argument underpins results on CLT for non-degenerate U-statistics. In the classical degenerate setting, however, $d$ is still fixed but $\sigma_{\mathrm{cond}} = 0$, and the above argument fails to apply. Instead, one considers a spectral decomposition $u(\mathbf{x},\mathbf{y}) = \sum_{k=1}^\infty \lambda_k\phi_k(\mathbf{x})\phi_k(\mathbf{y})$ for some eigenvalues $\{\lambda_k\}_{k=1}^\infty$ and eigenfunctions $\{\phi_k\}_{k=1}^\infty$, and compares the distribution of $D_n$ to a weighted sum of chi-squares:

$$\begin{aligned}D_n &= \frac{1}{n(n-1)}\sum\nolimits_{1\le i\neq j\le n}\sum\nolimits_{k=1}^\infty \lambda_k\phi_k(\mathbf{X}_i)\phi_k(\mathbf{X}_j)\\ &= \sum\nolimits_{k=1}^\infty \lambda_k\Big(\Big(\frac{1}{n}\sum\nolimits_{i=1}^n\phi_k(\mathbf{X}_i)\Big)\Big(\frac{1}{n}\sum\nolimits_{j=1}^n\phi_k(\mathbf{X}_j)\Big) - \frac{1}{n^2}\sum\nolimits_{i=1}^n\phi_k(\mathbf{X}_i)^2\Big)\\ &\approx \frac{1}{n}\sum\nolimits_{k=1}^\infty\lambda_k\left(\big(\sqrt{\mathrm{Var}[\phi_k(\mathbf{X}_1)]}\,\xi_k + \sqrt{n}\,\mathbb{E}[\phi_k(\mathbf{X}_1)]\big)^2 - \mathbb{E}[\phi_k(\mathbf{X}_1)^2]\right)\,,\end{aligned}$$

where $\xi_k$'s are i.i.d. standard normals. The limiting distributions in both settings enable one to construct consistent confidence intervals for $D_n$ and study $\mathbb{P}(D_n > t)$.

The key takeaway is that the asymptotic distribution of $D_n$ depends on the relative sizes of $\sigma^2_{\mathrm{cond}}$ and $(n-1)^{-1}\sigma^2_{\mathrm{full}}$. This comparison reduces to degeneracy when $d$ is fixed, but is no longer so when $d$ grows. In the high-dimensional setting, $\sigma_{\mathrm{cond}}$ and $\sigma_{\mathrm{full}}$ can scale with $d$ at *different orders*, making it possible for the ratio $\rho_d$ to vary with $d$. In particular, a non-degenerate U-statistic with $\sigma_{\mathrm{cond}} > 0$ may still satisfy $\rho_d = \omega(n^{1/2})$, i.e. $(n-1)^{-1}\sigma^2_{\mathrm{full}}/\sigma^2_{\mathrm{cond}} \to \infty$ as $n$ and $d$ grow. In this case, the classical argument for a non-degenerate Gaussian limit would fail and a degenerate limit would dominate. This is exactly what we observe in the practical applications in Section 4.3, and motivates the need for results that explicitly addresses the high-dimensional setting.

## 3. Main results

The main result presented in this section is a finite-sample bound that compares $D_n$ to a quadratic form of infinitely many Gaussians. The limiting distribution is a sum of the non-degenerate limit and a variant of the degenerate limit, and subject to Assumption 1, the error bound is *independent* of $\rho_d$. In the case $\rho_d = o(n^{1/2})$, the non-degenerate limit dominates and our result agrees with the Gaussian limit given by a Berry-Esséen theorem for U-statistics. However when dimension is high such that $\rho_d = \omega(n^{1/2})$, the degenerate limit dominates and implies a *larger asymptotic variance*. We also discuss how to obtain consistent distribution bounds that reflect the effect of a large dimension $d$ on the original statistic $D_n$.

Our results rest on a functional decomposition assumption. For a sequence of $\mathbb{R}^d \to \mathbb{R}$ functions $\{\phi_k\}_{k=1}^\infty$ and a sequence of real values $\{\lambda_k\}_{k=1}^\infty$, we define the $L_\nu$ approximation error for $\nu \geq 1$ and a given $K \in \mathbb{N}$ as

$$\varepsilon_{K;\nu} \;\coloneqq\; \big\| \textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u(\mathbf{X}_1, \mathbf{X}_2) \big\|_{L_\nu} \;.$$

**Assumption 2** *There exists some $\nu \in (2, 3]$ such that, for any fixed $n$ and $d$, as $K \to \infty$, the $L_\nu$ approximation error $\varepsilon_{K;\nu} \to 0$ for some choice of $\{\phi_k\}_{k=1}^\infty$ and $\{\lambda_k\}_{k=1}^\infty$.*

**Remark 1** *(i) If Assumption 2 holds for some $\nu > 3$, it certainly holds for $\nu = 3$. We restrict our focus to $\nu \in (2, 3]$ for simplicity. (ii) Assumption 2 always holds for $\nu = 2$ by the spectral decomposition of an operator on $L_2(\mathbb{R}^d, R)$. For degenerate U-statistics with $d$ fixed, the corresopnding orthonormal eigenbasis of functions and eigenvalues are used to prove asymptotic results (see Section 5.5.2 of Serfling (1980)) and finite-sample bounds (Bentkus and Götze, 1999; Götze and Tikhomirov, 2005; Yanushkevichiene, 2012). In fact, these finite-sample bounds are dependent on the specific $\lambda_k$'s, making the results hard to apply. Instead, we forgo orthonormality at the cost of a convergence slightly stronger than $L_2$. This allows for a much more flexible choice of $\{\phi_k, \lambda_k\}_{k=1}^\infty$ and is particularly well-suited for a kernel-based setting; see Remark 17 for a discussion.*

Before stating the results, we introduce some more notations. For every $K \in \mathbb{N}$, we define a diagonal matrix of the first $K$ "eigenvalues" and a concatenation of the first $K$ "eigenfunctions" by

$$\Lambda^K \;\coloneqq\; \text{diag}\{\lambda_1, \ldots, \lambda_K\} \;\in \mathbb{R}^{K \times K} \;, \qquad \phi^K(x) \;\coloneqq\; (\phi_1(x), \ldots, \phi_K(x))^\top \;\in \mathbb{R}^K \;. \tag{4}$$

We denote the mean and variance of $\phi^K(\mathbf{X}_1)$ by $\mu^K \coloneqq \mathbb{E}[\phi^K(\mathbf{X}_1)]$ and $\Sigma^K \coloneqq \text{Cov}[\phi^K(\mathbf{X}_1)]$.

### 3.1. Result for the general case

Let $\eta_i^K$, with $i, K \in \mathbb{N}$, be i.i.d. standard Gaussian vectors in $\mathbb{R}^K$. In the general case, the limiting distribution is given in terms of a quadratic form of Gaussians, defined by

$$U_n^K \coloneqq \frac{1}{n(n-1)} \textstyle\sum_{1 \leq i \neq j \leq n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K + \frac{2}{n} \sum_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K + D.$$

We also denote the dominating moment terms by

$$\sigma_{\max} \;\coloneqq\; \max\{\sigma_{\text{full}}, (n-1)^{1/2}\sigma_{\text{cond}}\} \;, \qquad M_{\max;\nu} \;\coloneqq\; \max\{M_{\text{full};\nu}, (n-1)^{1/2} M_{\text{cond};\nu}\} \;.$$

We are ready to state our main result – a finite-sample error bound that compares $D_n$ to the limiting distribution of $U_n^K$, where the error is given in terms of $n$ and the moment terms.

**Theorem 2** *There exists a constant $C > 0$ such that, for all $u$, $R$, $d$ and $n$, if $\nu \in (2, 3]$ satisfies Assumption 2, then the following holds:*

$$\sup_{t\in\mathbb{R}} \Big|\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}} D_n > t\Big) - \lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}} U_n^K > t\Big)\Big|$$
$$\leq C\, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{(M_{\text{full};\nu})^\nu}{\sigma_{\max}^\nu} + \frac{((n-1)^{1/2}\, M_{\text{cond};\nu})^\nu}{\sigma_{\max}^\nu} \Big)^{\frac{1}{2\nu+1}} \;\leq\; 2^{\frac{1}{2\nu+1}} C\, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{M_{\max;\nu}}{\sigma_{\max}} \Big)^{\frac{\nu}{2\nu+1}} \;.$$

**Remark 3** *If $\nu = 3$, the RHS is given by $2^{3/7} C n^{-\frac{1}{14}} \big(\frac{M_{\max;3}}{\sigma_{\max}}\big)^{6/7}$. If Assumption 1 holds for $\nu$, the RHS can be replaced by $C' n^{-\frac{\nu-2}{4\nu+2}}$ for some constant $C'$ and is dimension-independent.*

**Remark 4** *At first sight, one may be tempted to move $\lim_{K\to\infty}$ inside $\mathbb{P}$ such that, instead of the cumbersome expression of $W_n^K$ with finite $K$, one may deal with random quantities in a Hilbert space. The reason to stick with $W_n^K$ is that in Assumption 2, convergence of the infinite sum is required only in $L_\nu$ and not almost surely. This makes verification of the assumption substantially simpler in practice: In Appendix A, we illustrate how this assumption holds via a simple Taylor-expansion argument coupled with suitable tail behaviour of the data to control error terms. The same argument is not applicable if we instead require an almost sure convergence.*

Theorem 2 immediately implies a convergence theorem:

**Corollary 5** *Let the dimension $d_n$ depend on $n$. Suppose Assumptions 1 and 2 hold for some $\nu > 2$ and the sequential distribution limit $\bar{U} = \lim_{n\to\infty} \lim_{K\to\infty} \frac{\sqrt{n(n-1)}}{\sigma_{\max}} (U_n^K - D)$ exists. Then*

$$\frac{\sqrt{n(n-1)}}{\sigma_{\max}} (D_n - D) \xrightarrow{d} \bar{U} \qquad\qquad \text{as } \; n \to \infty \;.$$

$U_n^K$ is a quadratic form of Gaussians, which does not admit a closed-form c.d.f. in general and whose limiting behaviour depends heavily on $\lambda_k$ and $\phi_k$. Nevertheless, the presence of Gaussianity still allows us to obtain crude bounds that reflect how dimension $d$ affects its distribution. By combining such bounds with Theorem 2, we can bound the c.d.f. of the original U-statistic $D_n$.

**Proposition 6** *There exists constants $C_1, C_2, C_3 > 0$ such that, for all $u$, $R$, $d$, $n$ and $K$, if $\nu \in (2, 3]$ satisfies Assumption 2, then for all $\epsilon > 0$,*

$$\mathbb{P}(|D_n - D| > \epsilon) \;\geq\; 1 - C_1 \Big( \frac{\sqrt{n(n-1)}}{\sigma_{\max}} \Big)^{1/2} \epsilon^{1/2} - C_2\, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{M_{\max;\nu}}{\sigma_{\max}} \Big)^{\frac{\nu}{2\nu+1}} \;,$$
$$\mathbb{P}(|D_n - D| > \epsilon) \;\leq\; C_3 \epsilon^{-2} \Big( \frac{\sigma_{\max}}{\sqrt{n(n-1)}} \Big)^2 \;.$$

**Remark 7** *The second bound is a concentration inequality directly available via Markov's inequality, whereas the first bound is an anti-concentration result. Anti-concentration results are generally available only for random variables from known distribution families, and we obtain such a result by comparing $D_n$ to $U_n^K$. The error bounds are free of any dependence on $K$ and specific choices of $\phi_k$ and $\lambda_k$. The trailing error term involving $M_{\max;\nu}/\sigma_{\max}$ is inherited from Theorem 2 and is negligible, whereas the other error term is directly related to the inverse of the Markov error term.*

Proposition 6 provide two-sided bounds on how likely it is for $D_n$ to be far from $D$. The next corollary provides a more explicit statement.

**Corollary 8** *Let the dimension $d_n$ depend on $n$ and fix $\epsilon > 0$. Suppose Assumptions 1 and 2 hold for some $\nu \in (2, 3]$. As $n \to \infty$, we have that $\mathbb{P}(|D_n - D| > \epsilon) \to 1$ if $\sigma_{\max} = \omega(n)$ and $\mathbb{P}(|D_n - D| > \epsilon) \to 0$ if $\sigma_{\max} = o(n)$.*

Another way of formulating the bounds in Proposition 6 is the following: Similar to the intuition for a Gaussian, when $n$ is large (with $d_n$ depending on $n$), the distribution of $D_n$ is not only concentrated in an interval around $D$ with width being a multiple of $\frac{\sigma_{\max}}{n}$, but also "well spread-out" within the interval. The probability mass gets concentrated around $D$ when $\sigma_{\max} = o(n)$, but spreads out along the whole real line when $\sigma_{\max} = \omega(n)$; the latter only happens in a high dimensional regime.

To have a more precise understanding of the limiting behaviour of $D_n$, we need a better knowledge of $U_n^K$. By a closer examination of $U_n^K$, we see that it is a sum of three terms: A sum of weighted chi-squares with variance of the order $n^{-1}(n-1)^{-1}\sigma_{\text{full}}^2$, a Gaussian with variance of the order $n^{-1}\sigma_{\text{cond}}^2$, and a constant $D$. The first term closely resembles the limit for degenerate U-statistics when $d$ is fixed, while the second term corresponds exactly to the Gaussian limit for non-degenerate U-statistics. It turns out that, unless we are at the boundary case where $\rho_d = \Theta(n^{1/2})$, we can always approximate $U_n^K$ by ignoring either the first or the second term. Ignoring the first term gives exactly the Gaussian limit, where a well-established result has already been provided in (3). Ignoring the second term gives an infinite sum of weighted chi-squares, which is discussed next.

### 3.2. The case $\rho_d = \omega(n^{1/2})$

Let $\{\xi_k\}_{k=1}^\infty$ be a sequence of i.i.d. standard Gaussians in 1d, and for $K \in \mathbb{N}$, let $\{\tau_{k;d}\}_{k=1}^K$ be the eigenvalues of $(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}$. The limiting distribution we consider is given in terms of

$$W_n^K \;:=\; \frac{1}{\sqrt{n(n-1)}} \sum\nolimits_{k=1}^K \tau_{k;d}(\xi_k^2 - 1) + D\;. \tag{5}$$

Note that in this case, $\sigma_{\max} = \sigma_{\text{full}}$. The next result adapts Theorem 2 by replacing $U_n^K$ with $W_n^K$:

**Proposition 9** *There exists a constant $C > 0$ such that, for all $u$, $R$, $d$, $n$ and $K$, if $\nu \in (2, 3]$ satisfies Assumption 2, then the following holds:*

$$\begin{aligned}\sup\nolimits_{t\in\mathbb{R}}\Big|\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}D_n > t\Big) - \lim_{K\to\infty}\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}W_n^K > t\Big)\Big| \\ \leq\; C\Big(\frac{1}{(n-1)^{1/5}} + \Big(\frac{\sqrt{n-1}\,\sigma_{\text{cond}}}{\sigma_{\text{full}}}\Big)^{2/5} + n^{-\frac{\nu-2}{4\nu+2}}\Big(\frac{(M_{\text{full};\nu})^\nu}{\sigma_{\text{full}}^\nu} + \frac{((n-1)^{1/2}M_{\text{cond};\nu})^\nu}{\sigma_{\text{full}}^\nu}\Big)^{\frac{1}{2\nu+1}}\Big)\;.\end{aligned}$$

**Remark 10** *In the case $\nu = 3$, the error term above becomes*

$$C\Big(\frac{1}{(n-1)^{1/5}} + \Big(\frac{\sqrt{n-1}\,\sigma_{\text{cond}}}{\sigma_{\text{full}}}\Big)^{2/5} + n^{-\frac{1}{14}}\Big(\frac{(M_{\text{full};3})^3}{\sigma_{\text{full}}^3} + \frac{\big((n-1)^{1/2}M_{\text{cond};3}\big)^3}{\sigma_{\text{full}}^3}\Big)^{\frac{1}{7}}\Big)\;.$$

*In the case when Assumption 1 holds for $\nu$, the error term is $\Theta\big(\big(\frac{n-1}{\rho_d^2}\big)^{1/5} + n^{-\frac{\nu-2}{4\nu+2}}\big)$.*

**Remark 11** *Proposition 9 agrees with the classical results for degenerate U-statistics. In those results, $\{\phi_k\}_{k=1}^{\infty}$ are chosen such that they are orthonormal in $L_2(\mathbb{R}^d, R)$ and $\mathbb{E}[\phi_k(\mathbf{X}_1)] = 0$. This corresponds to $\Sigma^K$ being a diagonal matrix and the expression for $\tau_{k;d}$ can be simplified.*

We seek to obtain a better understanding of the limiting distribution of $D_n$ in the case $\rho_d = \omega(n^{1/2})$. Write $W_n := \lim_{K\to\infty} W_n^K$ as the distributional limit of $W_n^K$ as $K \to \infty$. Provided that $W_n$ exists, Proposition 9 gives the convergence of $D_n$ to $W_n$ in the Kolmogorov metric. The next lemma guarantees the existence of $W_n$.

**Proposition 12** *Fix $n, d$. If Assumption 2 holds for some $\nu \geq 2$ and $|D|, \sigma_{\text{full}} < \infty$, $W_n$ exists.*

While $W_n^K$ is a sum of chi-squares, the distributional limit $W_\infty := \lim_{n\to\infty} \lim_{K\to\infty} W_n^K$ may actually be Gaussian. The crucial subtlety lies in the fact that the weights of $W_n^K$ may depend on $K$ and also on $n$ (through $d \equiv d_n$). In what is well-known in the probability literature as the "fourth moment phenomenon" (Nualart and Peccati, 2005), the necessary and sufficient condition for Gaussianity of $W_\infty$ is that the limiting excess kurtosis is zero. In our case, the limiting moments can be computed easily when Assumption 2 holds for $\nu \geq 4$, as they depend only on moments of the original function $u$ and *not* on specific values of the intractable weights $\tau_{k;d}$. Lemma 33 in the appendix shows that $\mathbb{E}[W_n^K] = D$ for every $K \in \mathbb{N}$, $\lim_{K\to\infty} \text{Var}[W_n^K] = \frac{2}{n(n-1)}\sigma_{\text{full}}^2$ and

$$\lim_{K\to\infty} \mathbb{E}\big[(W_n^K - D)^4\big] = \frac{12(4\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)u(\mathbf{X}_2, \mathbf{X}_3)u(\mathbf{X}_3, \mathbf{X}_4)u(\mathbf{X}_4, \mathbf{X}_1)] + \sigma_{\text{full}}^4)}{n^2(n-1)^2},$$

provided that Assumption 2 holds for $\nu \geq 1$, $\nu \geq 2$ and $\nu \geq 4$ respectively. If the excess kurtosis is indeed zero, Gaussian is still the correct limiting distribution for $D_n$, but now with a *larger* variance (characterized by $\sigma_{\text{full}}$) than the one naively predicted by the Gaussian CLT limit for non-degenerate U-statistics. Meanwhile, when the excess kurtosis is not zero, the limiting distribution is an infinite sum of weighted chi-squares. A naive example is the following:

**Lemma 13** *Suppose there exists a finite $K_*$ such that $\lambda_k = 0$ for all $k > K_*$. Then $W_n = W_n^{K^*}$, which is a weighted sum of chi-squares.*

A weighted sum of chi-squares does not admit a closed-form distribution function. Fortunately in the case when $\tau_{k;d} \geq 0$ for all $k$, many numerical approximation schemes are available and used widely. These methods generally rely on matching the moments of $W_n$, which can be computed easily due to Proposition 12. The simplest example is the Welch-Satterthwaite method, which approximates the distribution of $W_n$ by a gamma distribution with the same mean and variance. We refer readers to Bodenham and Adams (2016) and Duchesne and De Micheaux (2010) for a review of other moment-matching methods.

### 3.3. Proof overview

The proof for Theorem 2 consists of three main steps:

(i) **"Spectral" approximation.** We first use Assumption 2 to replace $u(\mathbf{X}_i, \mathbf{X}_j)$ with the truncated sum $\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_i)\phi_k(\mathbf{X}_j)$, which gives a truncation error that vanishes as $K \to \infty$;

(ii) **Gaussian approximation.** The truncated sum is a simple quadratic form of i.i.d. vectors in $\mathbb{R}^d$, each of which can be approximated by a Gaussian vector. This is done by following Chatterjee (2006)'s adaptataion of Lindeberg's telescoping sum argument. Similar proof ideas have been used to develop new convergence results in statistics and machine learning; examples include empirical risk (Montanari and Saeed, 2022) and bootstrap for non-asymptotically normal estimators (Austern and Syrgkanis, 2020). This step introduces errors in terms of moment terms of $U_n^K$, which are then related to those of $D_n$;

(iii) **Bound the distribution of $U_n^K$.** Step (ii) introduces errors in terms of the distribution of $U_n^K$, a quadratic form of Gaussians, over a short interval. These errors are then controlled by the distribution bounds from Carbery and Wright (2001).

The proof for Proposition 9 is similar, except that we use an additional Markov-type argument to remove the linear sum from $U_n^K$ and obtain the limit in terms of $W_n^K$.

## 4. Kernel-based testing in high dimensions

Given two probability measures $P$ and $Q$ on $\mathbb{R}^d$, we consider the problem of testing $H_0 : P = Q$ against $H_1 : P \neq Q$ through some measure of discrepancy between $P$ and $Q$. We focus on *Maximum Mean Discrepancy* (MMD) and *(Langevin) Kernelized Stein Discrepancy* (KSD), two kernel-based methods that use a U-statistic $D_n$ as the test statistic. It is well-known that $\sigma_{\text{cond}} = 0$ under $H_0$ and the limit of $D_n$ is a weighted sum of chi-squares (see Gretton et al. (2012) for MMD and Liu et al. (2016) for KSD). Instead, we are interested in quantifying the power of $D_n$ given as $\mathbb{P}_{H_1}(D_n > t)$. The test threshold $t$ is often chosen adaptively in practice, but we assume $t$ to be fixed for simplicity of analysis. The results in Section 3 offer two key insights to this problem:

(i) $D_n$ may have different limiting distributions depending on $\rho_d$. In the non-Gaussian case, the confidence interval and thereby the distribution curve can be wider than what a Berry-Esséen bound predicts, and there may be potential asymmetry;

(ii) We can completely characterise the high-dimensional behaviour of the power in terms of $\rho_d$, which in turn depends on the hyperparameters and the set of alternatives considered.

In this section, we first show that our results naturally apply to MMD and KSD. We then investigate their high-dimensional behaviours in an example of Gaussian mean-shift under simple kernels. Throughout, $\|\bullet\|_2$ denotes the vector Euclidean norm, which is not to be confused with $\|\bullet\|_{L_2}$.

### 4.1. Notations

We follow the kernel definition from Steinwart and Scovel (2012) as below:

**Definition 14** *A function $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a* kernel *on $\mathbb{R}^d$ if there exists a Hilbert space $(\mathcal{H}, \langle \bullet, \bullet \rangle_{\mathcal{H}})$ and a map $\phi : \mathbb{R}^d \to \mathcal{H}$ such that $\kappa(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{H}$.*

We give the minimal definitions of MMD and KSD, and refer interested readers to Gretton et al. (2012) and Gorham and Mackey (2017) for further reading. Throughout, we let $\{\mathbf{Y}_j\}_{j=1}^n$ be i.i.d. samples from $P$ and $\{\mathbf{X}_i\}_{i=1}^n$ be i.i.d. samples from $Q$. We also write $\mathbf{Z}_i \coloneqq (\mathbf{X}_i, \mathbf{Y}_i)$ and assume that $\kappa$ is measurable. MMD with respect to $\kappa$ is defined by

$$D^{\text{MMD}}(Q, P) \coloneqq \mathbb{E}_{\mathbf{Y},\mathbf{Y}' \sim P}[\kappa(\mathbf{Y}, \mathbf{Y}')] - 2\mathbb{E}_{\mathbf{Y} \sim P, \mathbf{X} \sim Q}[\kappa(\mathbf{Y}, \mathbf{X})] + \mathbb{E}_{\mathbf{X},\mathbf{X}' \sim Q}[\kappa(\mathbf{X}, \mathbf{X}')] .$$

A popular unbiased estimator for $D^{\text{MMD}}$ is exactly a U-statistic:

$$D_n^{\text{MMD}} \coloneqq \frac{1}{n(n-1)} \sum\nolimits_{1 \leq i \neq j \leq n} u^{\text{MMD}}(\mathbf{Z}_i, \mathbf{Z}_j) ,$$

where the summand is given by $u^{\rm MMD}\big((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')\big) \coloneqq \kappa(\mathbf{x},\mathbf{x}')+\kappa(\mathbf{y},\mathbf{y}')-\kappa(\mathbf{x},\mathbf{y}')-\kappa(\mathbf{x}',\mathbf{y})$. To define KSD, we assume that $\kappa$ is continuously differentiable with respect to both arguments, and $P$ admits a continuously differentiable, positive Lebesgue density $p$. The following formulation of KSD is due to Theorem 2.1 of Chwialkowski et al. (2016):

$$D^{\rm KSD}(Q,P) \;\coloneqq\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[u_P^{\rm KSD}(\mathbf{X},\mathbf{X}')]\;,$$

where we assume $\mathbb{E}_{\mathbf{X}\sim Q}[u_P^{\rm KSD}(\mathbf{X},\mathbf{X})] < \infty$ and the function $u_P^{\rm KSD}: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is given by

$$\begin{aligned} u_P^{\rm KSD}(\mathbf{x},\mathbf{x}') = \big(\nabla \log p(\mathbf{x})\big)^\top \big(\nabla \log p(\mathbf{x}')\big)\kappa(\mathbf{x},\mathbf{x}') \;+\; \big(\nabla \log p(\mathbf{x})\big)^\top \nabla_2 \kappa(\mathbf{x},\mathbf{x}') \\ +\; \big(\nabla \log p(\mathbf{x}')\big)^\top \nabla_1 \kappa(\mathbf{x},\mathbf{x}') \;+\; {\rm Tr}(\nabla_1\nabla_2\kappa(\mathbf{x},\mathbf{x}'))\;. \end{aligned}$$

$\nabla_1$ and $\nabla_2$ are the differential operators with respect to the first and second arguments of $\kappa$ respectively. The estimator is again a U-statistic, given by $D_n^{\rm KSD} \coloneqq \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n} u_P^{\rm KSD}(\mathbf{X}_i,\mathbf{X}_j)$.

### 4.2. General results

We show that a kernel structure allows Assumption 2 to be fulfilled under some natural conditions. Let $\mathbf{V}_1,\mathbf{V}_2 \overset{i.i.d.}{\sim} R$ for some probability measure $R$ on $\mathbb{R}^b$ and $\kappa^*$ be a measurable kernel on $\mathbb{R}^b$. A sequence of functions $\{\phi_k\}_{k=1}^\infty$ in $L_2(\mathbb{R}^b,R)$ and a sequence of non-negative values $\{\lambda_k\}_{k=1}^\infty$ with $\lim_{k\to\infty}\lambda_k = 0$ is called a *weak Mercer representation* if

$$\big|\sum_{k=1}^K \lambda_k\phi_k(\mathbf{V}_1)\phi_k(\mathbf{V}_2) - \kappa^*(\mathbf{V}_1,\mathbf{V}_2)\big| \to 0 \quad \text{almost surely} \qquad\qquad \text{as } K\to\infty\;.$$

Steinwart and Scovel (2012) show that such a representation exists if $\mathbb{E}[\kappa^*(\mathbf{V}_1,\mathbf{V}_1)] < \infty$, whose result is summarised in Lemma 37 in the appendix. To deduce from this the $L_\nu$ convergence of Assumption 2, we need the following assumptions on the kernel $\kappa^*$:

**Assumption 3** *Fix $\nu > 2$. Assume $\mathbb{E}[\kappa^*(\mathbf{V}_1,\mathbf{V}_1)] < \infty$ and let $\{\lambda_k\}_{k=1}^\infty$ and $\{\phi_k\}_{k=1}^\infty$ be a weak Mercer representation of $\kappa^*$ under $R$. Also assume that for some $\nu^* > \nu$, $\|\kappa^*(\mathbf{V}_1,\mathbf{V}_2)\|_{L_{\nu^*}} < \infty$ and $\sup_{K\geq 1}\|\sum_{k=1}^K\lambda_k\phi_k(\mathbf{V}_1)\phi_k(\mathbf{V}_2)\|_{L_{\nu^*}} < \infty$.*

For MMD, we can use the weak Mercer representation of $u^{\rm MMD}$ to show that our results apply:

**Lemma 15** *$u^{\rm MMD}$ defines a kernel on $\mathbb{R}^{2d}$. Moreover, if Assumption 3 holds for $\kappa^* = u^{\rm MMD}$ under $P\otimes Q$ for some $\nu > 2$, then Assumption 2 holds for $\min\{\nu,3\}$ with $u = u^{\rm MMD}$ and $R = P\otimes Q$.*

In the case of KSD, we use the representation of $\kappa$ directly. We require some additional assumptions for the score function $\nabla\log p(\mathbf{x})$ to be well-behaved and the differential operation on $\kappa$ to behave well under the representation.

**Assumption 4** *Fix $n$, $d$ and $\nu > 2$. Assume that Assumption 3 holds with $\nu$ for $\kappa$ under $Q$, with $\{\lambda_k\}_{k=1}^\infty$ and $\{\phi_k\}_{k=1}^\infty$ as the weak Mercer representation of $\kappa$ under $Q$ and $\nu^*$ being defined as in Assumption 3. Further assume that (i) $\|\|\nabla\log p(\mathbf{X}_1)\|_2\|_{L_{2\nu^{**}}} < \infty$ for $\nu^{**} = \frac{\nu(\nu+\nu^*)}{\nu^*-\nu}$; (ii) $\sup_{k\in\mathbb{N}}\|\phi_k(\mathbf{X}_1)\|_{L_{2\nu}} < \infty$; (iii) $\phi_k$'s are differentiable with $\sup_{k\in\mathbb{N}}\|\|\nabla\phi_k(\mathbf{X}_1)\|_2\|_{L_\nu} < \infty$; (iv) As $K\to\infty$, we have the convergence $\big\|\big\|\sum_{k=1}^K\lambda_k(\nabla\phi_k(\mathbf{X}_1))\phi_k(\mathbf{X}_2) - \nabla_1\kappa(\mathbf{X}_1,\mathbf{X}_2)\big\|_2\big\|_{L_{2\nu}} \to 0$ as well as the convergence $\big\|\sum_{k=1}^K\lambda_k(\nabla\phi_k(\mathbf{X}_1))^\top(\nabla\phi_k(\mathbf{X}_2)) - {\rm Tr}(\nabla_1\nabla_2\kappa(\mathbf{X}_1,\mathbf{X}_2))\big\|_{L_\nu} \to 0$.*

We can now form a decomposition of $u_P^{\mathrm{KSD}}$. Given $\{\lambda_k\}_{k=1}^\infty$ and $\{\phi_k\}_{k=1}^\infty$ from Assumption 4 and any fixed $d \in \mathbb{N}$, define the sequences $\{\alpha_k\}_{k=1}^\infty$ and $\{\psi_k\}_{k=1}^\infty$ as, for $1 \leq l \leq d$ and $k' \in \mathbb{N}$,

$$\alpha_{(k'-1)d+l} \ \coloneqq \lambda_{k'} \qquad \text{and} \qquad \psi_{(k'-1)d+l}(\mathbf{x}) \ \coloneqq (\partial_{x_l} \log p(\mathbf{x}))\phi_{k'}(\mathbf{x}) + \partial_{x_l}\phi_{k'}(\mathbf{x}) \ . \tag{6}$$

**Lemma 16** *If Assumption 4 holds for some $\nu > 2$, then Assumption 2 holds for $\min\{\nu, 3\}$ with $u = u_P^{\mathrm{KSD}}$, $R = Q$, $\lambda_k = \alpha_k$ and $\phi_k = \psi_k$ .*

**Remark 17** *The benefits of formulating our results in terms of Assumption 2 are now clear: By forgoing orthonormality, we can choose a functional decomposition e.g. in terms of the Mercer representation of a kernel, which is already widely considered in this literature. The non-negative eigenvalues from Lemma 37 also allow moment-matching methods discussed in Section 3.2 to be considered. In fact, a Mercer representation is not even necessary: In Appendix A.1, we construct a simple decomposition for the setup in Section 4.3 such that Assumption 2 can be verified easily.*

### 4.3. Gaussian mean-shift examples

We study KSD and MMD under Gaussian mean-shift, where $P = \mathcal{N}(0, \Sigma)$ and $Q = \mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$ to be specified. Two simple kernels are considered in this section, namely the *RBF* kernel and the *linear* kernel.

**RBF kernel.** We consider the RBF kernel $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/(2\gamma))$, where $\gamma = \gamma(d)$ is a bandwidth potentially depending on $d$. A common strategy to choose $\gamma$ is the *median heuristic*:

$$\gamma_{\text{med}} \ \coloneqq \ \text{Median}\left\{\|\mathbf{V} - \mathbf{V}'\|_2^2 : \mathbf{V}, \mathbf{V}' \in \mathcal{V}\,,\ \mathbf{V} \neq \mathbf{V}'\right\} \ ,$$

where the samples $\mathcal{V} = \{\mathbf{X}_i\}_{i=1}^n$ for KSD and $\mathcal{V} = \{\mathbf{X}_i\}_{i=1}^n \cup \{\mathbf{Y}_i\}_{i=1}^n$ for MMD. In Appendix A, we include a further discussion of this setup as well as verification of Assumption 1 and Assumption 2.

We focus on $\Sigma = I_d$, where the $d$-dependence of the moment ratio $\rho_d$ can be explicitly studied for both KSD and MMD. Importantly, we give bounds in terms of the bandwidth $\gamma$ and the scale of mean shift $\|\mu\|_2^2$, which reveal their effects on $\rho_d$ and thereby on the behaviour of the test power. The assumptions on $\gamma$ and $\|\mu\|_2^2$ in both propositions are for simplicity rather than necessity.

**Proposition 18 (KSD-RBF moment ratio)** *Assume $\gamma = \omega(1)$ and $\|\mu\|_2^2 = \Omega(1)$. Under the Gaussian mean-shift setup with $\Sigma = I_d$, the KSD U-statistic satisfies that*

*(i) If $\gamma = o(d^{1/2})$, then $\rho_d = \exp\left(\frac{3d}{4\gamma^2} + o\left(\frac{d}{\gamma^2}\right)\right)\Theta\left(\frac{d}{\gamma\|\mu\|_2^2} + \frac{d^{1/2}}{\gamma^{1/2}\|\mu\|_2} + 1\right)$ ;*

*(ii) If $\gamma = \omega(d^{1/2})$, then $\rho_d = \Theta\left(\frac{d^{1/2}(1+\gamma^{-1/2}\|\mu\|_2)}{\|\mu\|_2\,(1+\gamma^{-1}d^{1/2}\|\mu\|_2)} + 1\right)$ ;*

*(iii) If $\gamma = \Theta(d^{1/2})$, then $\rho_d = \Theta\left(\frac{d^{1/2}}{\|\mu\|_2^2} + \frac{d^{1/4}}{\|\mu\|_2} + 1\right)$ .*

**Proposition 19 (MMD-RBF moment ratio)** *Consider the Gaussian mean-shift setup with $\Sigma = I_d$ and assume $\gamma = \omega(1)$ and $\|\mu\|_2^2 = \Omega(1)$. For the MMD U-statistic, if $\gamma = o(\|\mu\|_2^2)$ and $\gamma = o(d^{1/2})$, then $\rho_d = \Theta\left(\exp\left(\frac{3d}{4\gamma^2} + o\left(\frac{d}{\gamma^2}\right)\right)\right)$. If instead $\gamma = \omega(\|\mu\|_2^2)$, then*

*(i) For $\gamma = o(d^{1/2})$, we have $\rho_d = \Theta\left(\frac{\gamma}{\|\mu\|_2^2}\exp\left(\frac{3d}{4\gamma^2} + o\left(\frac{d}{\gamma^2}\right)\right)\right)$ ;*

*(ii) For $\gamma = \omega(d^{1/2})$, we have $\rho_d = \Theta\left(\frac{\|\mu\|_2 + d^{1/2}}{\|\mu\|_2 + \gamma^{-1}d^{1/2}\|\mu\|_2^2}\right)$ ;*

*(iii) For $\gamma = \Theta(d^{1/2})$, we have $\rho_d = O\left(\frac{d^{1/2}}{\|\mu\|_2^2}\right)$ .*

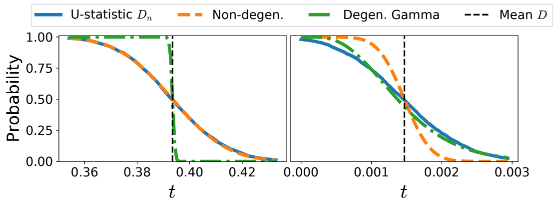


Figure 2: Behaviour of $\mathbb{P}(X > t)$ for $X = D_n^{\text{MMD}}$ with the RBF kernel versus $X$ being the theoretical limits. *Left.* $n = 1000$ and $d = 2$. *Right.* $n = 50$ and $d = 1000$.

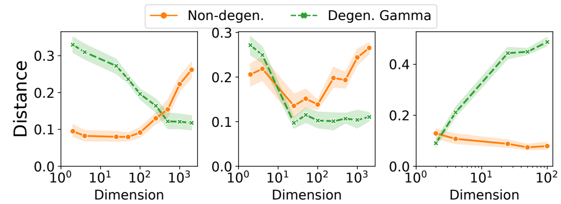


Figure 3: $L_\infty$ distance between the c.d.f. of $D_n^{\text{KSD}}$ with RBF and those of the theoretical limits as $d$ varies. *Left.* $n = 50$ fixed (high dimensions). *Middle.* $n = \Theta(d^{1/2})$ (high dimensions). *Right:* $n = \Theta(d^2)$ (low dimensions).

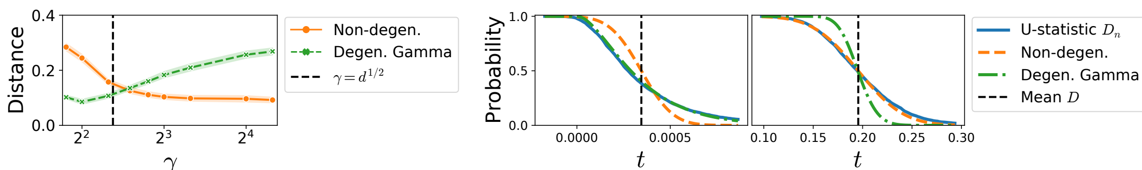


Figure 4: Behaviour of $\mathbb{P}(D_n^{\text{KSD}} > t)$ with RBF as $\gamma$ varies for $n = 50$ and $d = 27$. *Left.* $L_\infty$ distance between the c.d.f. of $D_n^{\text{KSD}}$ and the theoretical limits. *Middle.* Distribution curves at $\gamma = 4$. *Right.* Distribution curves at $\gamma = 16$.

The case $\|\mu\|_2 = \Omega(\|\Sigma\|_2) = \Omega(d^{1/2})$ is not very interesting, as it means that the signal-to-noise ratio (SNR) is high and can even increase with $d$. WLOG we focus on a low SNR setting with $\|\mu\|_2 = \Theta(1)$. In this case, it has been shown that the median-heuristic bandwith scales as $\gamma_{\text{med}} = \Theta(d)$ (Reddi et al., 2015; Ramdas et al., 2015; Wynne and Duncan, 2022). While Propositions 18 and 19 do not directly address the case $\gamma = \gamma_{\text{med}}$ due to its data dependence, they do show that $\rho_d = \Theta(d^{1/2})$ for both KSD and MMD with a data-*independent* bandwidth $\gamma = \Theta(d)$[†]. In this case, the asymptotic distributions of $D_n^{\text{KSD}}$ and $D_n^{\text{MMD}}$ are *(i)* the non-degenerate Gaussian limit predicted by (3) when $d = o(n)$ and *(ii)* the degenerate limit from Proposition 9 when $d = \omega(n)$.

Intriguingly, in both results, different regimes arise based on how $\gamma$ compares with the noise scale $\|\Sigma\|_2 = d^{1/2}$. In fact, a phase transition as $\gamma$ drops from $\omega(d^{1/2})$ to $o(d^{1/2})$ has been reported in Ramdas et al. (2015) but with no further comments[‡§]. Our results offer one explanation: Such transitions may happen due to a change in the dependence of $\rho_d$ on $\gamma$, $\|\mu\|_2$ and $d$. Fig. 4 shows a transition across different limits as $\gamma$ varies, where the transition occurs at around $\gamma \sim d^{1/2}$.

**Linear kernel.** Section 3.2 discussed that the limit of $D_n$ can be non-Gaussian. This is true for MMD with a *linear* kernel $\kappa(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$: It satisfies Lemma 13 with $K_* = d$ and the limit is a shifted-and-rescaled chi-square. Fig. 1 verifies this for some $\Sigma \neq I_d$ by showing an asymmetric distribution curve close to the chi-square limit. We remark that a linear kernel, while not commonly used, is a valid choice here since $D^{\text{MMD}} = 0$ iff $P = Q$ under our setup.

**Simulations.** We set $\mu = (2, 0, \ldots, 0)^\top \in \mathbb{R}^d$, $\Sigma = I_d$ and $\gamma = \gamma_{\text{med}}$ for KSD with RBF and MMD with RBF. The exact setup for MMD with linear kernel is described in Appendix A.4. The

†. In our experiments, the data-independent choice $\gamma = d$ and the data-dependent $\gamma = \gamma_{\text{med}}$ yield almost identical plots.

‡. Their bandwidth $\gamma_{\text{Ramdas}}$ is defined to equal our $\sqrt{2\gamma}$. Phase transition occurs at $\gamma_{\text{Ramdas}} = d^{1/4}$ in their Figure 1. While their figure is for MMD with threshold chosen by a permutation test, ours is for KSD with a fixed threshold.

§. This was investigated in Ramdas (2015, Section 10.4) in a special case when $\gamma = \omega(\|\mu\|_2^2 + d)$ (case *(ii)* of Theorem 19) and $n = o(d^{5/2})$, where the author derived the test power of the RBF-kernel MMD for different SNRs.

limits for comparison are the non-degenerate Gaussian limit in (3) ("Non-degen.") and Gamma / shifted-and-rescaled chi-square ("Degen. Gamma" / "Degen. Chi-square") distributions that match the degenerate limit in Proposition 9 by mean and variance. Fig. 1 plots the distribution curves for KSD with RBF and MMD with linear kernel. Fig. 2 plots the same quantity for MMD with RBF. Fig. 3 and Fig. 4 examine the behaviour of KSD with RBF as $d$ or $\gamma$ varies (as a data-independent function of $d$, similar to Ramdas et al. (2015)). Results involving $D_n$ are averaged over 30 random seeds, and shaded regions are $95\%$ confidence intervals[¶]. Code for reproducing all experiments can be found at `github.com/XingLLiu/u-stat-high-dim.git`.


## Acknowledgments

KHH is supported by the Gatsby Charitable Foundation. XL is supported by the President's PhD Scholarships of Imperial College London and the EPSRC StatML CDT programme EP/S023151/1. ABD is supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1 and EPSRC Grant EP/W006022/1, particularly the "Ecosystems of Digital Twins" theme within those grants & The Alan Turing Institute. We thank Antonin Schrab, Heishiro Kanagawa and Arthur Gretton for their helpful comments.


¶. The shaded regions are not visible for $\mathbb{P}(D_n > t)$ in Fig. 1, 2 and 4 as the confidence intervals are very narrow.

## References

Morgane Austern and Vasilis Syrgkanis. Asymptotics of the empirical bootstrap method beyond asymptotic normality. *arXiv preprint arXiv:2011.11248*, 2020.

Vidmantas Bentkus and Friedrich Götze. Optimal bounds in non-gaussian limit theorems for U-statistics. *The Annals of Probability*, 27(1):454–521, 1999.

Bhaswar B Bhattacharya, Sayan Das, Somabha Mukherjee, and Sumit Mukherjee. Asymptotic distribution of random quadratic forms. *arXiv preprint arXiv:2203.02850*, 2022.

Dean A. Bodenham and Niall M. Adams. A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing*, 26(4):917–928, 2016.

Donald Lyman Burkholder. Martingale transforms. *The Annals of Mathematical Statistics*, 37(6): 1494–1504, 1966.

Anthony Carbery and James Wright. Distributional and $l^q$ norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001.

Sourav Chatterjee. A generalization of the Lindeberg principle. *The Annals of Probability*, 34(6): 2061–2076, 2006.

Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*, volume 2. Springer, 2011.

Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808 – 835, 2010.

Xiaohui Chen. Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *The Annals of Statistics*, 46(2):642 – 678, 2018.

Xiaohui Chen and Kengo Kato. Randomized incomplete U-statistics in high dimensions. *The Annals of Statistics*, 47(6):3127 – 3156, 2019.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2606–2615, 2016.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2009.

S. W. Dharmadhikari, V. Fabian, and K. Jogdeo. Bounds on the moments of martingales. *The Annals of Mathematical Statistics*, pages 1719–1723, 1968.

Pierre Duchesne and Pierre Lafaye De Micheaux. Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862, 2010.

Fabrice Gamboa, Pierre Gremaud, Thierry Klein, and Agnès Lagnoux. Global sensitivity analysis: A novel generation of mighty estimators based on rank statistics. *Bernoulli*, 28(4):2345–2374, 2022.

Hanjia Gao and Xiaofeng Shao. Two sample testing in high dimension via maximum mean discrepancy. *arXiv preprint arXiv:2109.14913*, 2021.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

Friedrich Götze and AN Tikhomirov. Asymptotic expansions in non-central limit theorems for quadratic forms. *Journal of Theoretical Probability*, 18(4):757–811, 2005.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Ilmun Kim and Aaditya Ramdas. Dimension-agnostic inference using cross u-statistics. *arXiv preprint arXiv:2011.05068*, 2020.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 276–284, 2016.

Jan R. Magnus. The moments of products of quadratic forms in normal variables. *Statistica Neerlandica*, 32:201–210, 1978.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

David Nualart and Giovanni Peccati. Central limit theorems for sequences of multiple stochastic integrals. 2005.

Wei Peng, Tim Coleman, and Lucas Mentch. Rates of convergence for random forests via generalized U-statistics. *Electronic Journal of Statistics*, 16(1):232–292, 2022.

Aaditya Ramdas. *Computational and statistical advances in testing and learning*. PhD thesis, Carnegie Mellon University, 2015.

Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Sashank Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial Intelligence and Statistics*, pages 772–780. PMLR, 2015.

Robert J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

Shubhanshu Shekhar, Ilmun Kim, and Aaditya Ramdas. A permutation-free kernel two-sample test. In *Advances in Neural Information Processing Systems*, 2022.

Yanglei Song, Xiaohui Chen, and Kengo Kato. Approximating high-dimensional infinite-order U-statistics: Statistical and computational guarantees. *Electronic Journal of Statistics*, 2019.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417, 2012.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, 1≤r≤ 2. *The Annals of Mathematical Statistics*, pages 299–303, 1965.

Lan Wang, Bo Peng, and Runze Li. A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, 110(512):1658–1669, 2015.

Runmin Wang, Changbo Zhu, Stanislav Volgushev, and Xiaofeng Shao. Inference for change points in high-dimensional data via self-normalization. *The Annals of Statistics*, 50(2):781–806, 2022.

George Wynne and Andrew B Duncan. A kernel two-sample test for functional data. *Journal of Machine Learning Research*, 23(73):1–51, 2022.

Jian Yan and Xianyang Zhang. Kernel two-sample tests in high dimension: Interplay between moment discrepancy and dimension-and-sample orders. *Biometrika*, 2022. ISSN 1464-3510.

O Yanushkevichiene. On bounds in limit theorems for some U-statistics. *Theory of Probability & Its Applications*, 56(4):660–673, 2012.

The appendix is organised as follows. The first few appendices provide additional content:
**Appendix A** states additional results for Section 4.3 including moment computations and verification of assumptions.
**Appendix B** presents auxiliary tools used in subsequent proofs.

The remaining appendices consist of proofs:
**Appendix C** proves our main theorem. Appendix C.1 provides a list of intermediate lemmas that extends the proof overview in Section 3.3.
**Appendix D** proves the remaining results in Section 3.
**Appendix E** proves the results in Section 4.
**Appendices F** and **G** present proofs for the results in Appendices A and B respectively.

Throughout the appendix, we say that $C$ is an absolute constant whenever we mean that it is a number independent of all variables involved, including $u$, $R$, $d$, $n$ and $K$.

## Appendix A. Additional results for Gaussian mean-shift

In this section, we consider the Gaussian mean-shift setup defined in Section 4.3, where $Q = \mathcal{N}(\mu, \Sigma)$ and $P = \mathcal{N}(0, \Sigma)$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. We derive analytical expressions of the moments of U-statistics for (i) KSD with RBF, (ii) MMD with RBF and (iii) MMD with linear kernel. We also verify Assumption 2 for the three cases, which confirm that our error bounds apply.

**Remark on verification of Assumption 1.** Recall that Assumption 1, which controls the moment ratios $M_{\text{full};\nu}/\sigma_{\text{full}}$ and $M_{\text{cond};\nu}/\sigma_{\text{cond}}$ for some $\nu \in (2, 3]$, is required for our bounds to imply a convergence theorem. As discussed in the main text, this issue is not specific to our theorem and is also relevant to e.g. Berry-Esséen bounds for sample averages of $\{f(\mathbf{X}_i)\}_{i=1}^n$ for $f : \mathbb{R}^d \to \mathbb{R}$ and $d$ large. A detailed verification requires a careful calculation to control the order of $M_{\text{full};\nu}, \sigma_{\text{full}}, M_{\text{cond};\nu}$ and $\sigma_{\text{cond}}$. For KSD and MMD with the RBF kernel, a careful control of $\sigma_{\text{full}}$ and $\sigma_{\text{cond}}$ has already been done in the proof of Proposition 18 and Proposition 19, which involves examining multiple cases depending on the relative sizes of $\gamma$, $\|\mu\|_2$ and $d$ followed by an elaborate calculation. To perform this verification for all cases in full generality, in principle, one may expand on those calculations and follow a similar tedious argument. In the sections below, we perform this verification only for the setup in Fig. 1, i.e. KSD with the RBF kernel in the case $\|\mu\|_2 = \Theta(1)$ and $\gamma = \Omega(d)$ and MMD with the linear kernel in the general case. For MMD with the RBF kernel, we discuss the relevance of this verification to Gao and Shao (2021), who has done a verification of similar quantities but also in a special case. In Figure 6, we also include simulations verifying Assumption 1 under the setups considered in Figure 1-3, where we demonstrate that the moment ratios stay around 1 as the dimension varies from 1 to 2000.

### A.1. A decomposition of the RBF kernel

For both MMD and KSD, the key in verifying the assumptions for the RBF kernel is a functional decomposition. The usual Mercer representation of the RBF kernel is available only with respect to a univariate zero-mean Gaussian measure and involves some cumbersome Hermite polynomials. Since we do not actually require orthogonality of the functions in Assumption 2, we opt for a simpler functional representation as given below. We also assume WLOG that the bandwidth $\gamma > 8$, since we only consider the case $\gamma = \omega(1)$ in our setup.

**Lemma 20** *Assume that $\gamma > 8$. Consider two independent $d$-dimensional Gaussian vectors $\mathbf{U} \sim \mathcal{N}(\mu_1, I_d)$ and $\mathbf{V} \sim \mathcal{N}(\mu_2, I_d)$ for some mean vectors $\mu_1, \mu_2 \in \mathbb{R}^d$. Then, for any $\nu \in (2, 4]$ and $\mu_1, \mu_2 \in \mathbb{R}^d$, we have that*

$$\mathbb{E}\Big[\Big| \exp\Big(-\tfrac{1}{2\gamma}\|\mathbf{U}-\mathbf{V}\|_2^2\Big) - \prod_{j=1}^d \Big(\sum_{k=0}^K \lambda_k^* \phi_k^*(U_j)\phi_k^*(V_j)\Big)\Big|^\nu\Big] \xrightarrow{K\to\infty} 0 \,.$$

*where $\phi_k^*(x) := x^k e^{-x^2/(2\gamma)}$ and $\lambda_k^* := \frac{1}{k!\,\gamma^k}$ for each $k \in \mathbb{N} \cup \{0\}$.*

To see that Lemma 20 indeed gives the functional decomposition we want in Assumption 2, we need to rewrite the product of sums into a sum. To this end, let $g_d$ be the $d$-tuple generalisation of the Cantor pairing function from $\mathbb{N}$ to $(\mathbb{N} \cup \{0\})^d$ and $[g_d(k)]_l$ be the $l$-th element of $g_d(k)$. Given $\{\lambda_l^*\}_{l=0}^\infty$ and $\{\phi_l^*\}_{l=0}^\infty$ from Lemma 20, we define, for every $k \in \mathbb{N}$ and $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$\alpha_k := \prod_{l=1}^d \lambda^*_{[g_d(k)]_l} \qquad \text{and} \qquad \psi_k(\mathbf{x}) := \prod_{l=1}^d \phi^*_{[g_d(k)]_l}(x_l) \,. \tag{7}$$

With this construction, for each $K \in \mathbb{N}$, we can now write

$$\begin{aligned}\prod_{j=1}^d \Big(\sum_{k=0}^K \lambda_k^* \phi_k^*(U_j)\phi_k^*(V_j)\Big) &= \sum_{k_1,\ldots,k_d=0}^K (\lambda_{k_1}^* \ldots \lambda_{k_d}^*)(\phi_{k_1}^*(U_1)\ldots\phi_{k_d}^*(U_d))(\phi_{k_1}^*(V_1)\ldots\phi_{k_d}(V_d)) \\ &= \sum_{k_1,\ldots,k_d=0}^K \alpha_{g_d^{-1}(k_1,\ldots,k_d)}\, \psi_{g_d^{-1}(k_1,\ldots,k_d)}(\mathbf{U})\, \psi_{g_d^{-1}(k_1,\ldots,k_d)}(\mathbf{V}) \,.\end{aligned}$$

Since the Cantor pairing function is such that $\min_{l\leq d}[g_d(K)]_l \to \infty$ as $K \to \infty$, Lemma 20 indeed gives a functional decomposition in terms of $\{\alpha_k\}_{k=1}^\infty$ and $\{\psi_k\}_{k=1}^\infty$ as

$$\mathbb{E}\Big[\Big| \exp\Big(-\tfrac{1}{2\gamma}\|\mathbf{U}-\mathbf{V}\|_2^2\Big) - \sum_{k=1}^K \alpha_k \psi_k(\mathbf{U})\psi_k(\mathbf{V})\Big|^\nu\Big] \xrightarrow{K\to\infty} 0 \,. \tag{8}$$

We remark that both $\alpha_k$ and $\psi_k$ are independent of the mean vectors $\mu_1$ and $\mu_2$, which makes this representation useful for a generic mean-shift setting.

### A.2. KSD U-statistic with RBF kernel

Under the Gaussian mean-shift setup with an identity covariance matrix, gradient of the log-density is given by $\nabla \log p(\mathbf{x}) = -\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^d$ and the U-statistic for the RBF-kernel KSD is

$$\begin{aligned}u_P^{\text{KSD}}(\mathbf{x}, \mathbf{x}') &= \big(\nabla \log p(\mathbf{x})\big)^\top \big(\nabla \log p(\mathbf{x}')\big)\kappa(\mathbf{x},\mathbf{x}') + \big(\nabla \log p(\mathbf{x})\big)^\top \nabla_2 \kappa(\mathbf{x},\mathbf{x}') \\ &\quad + \big(\nabla \log p(\mathbf{x}')\big)^\top \nabla_1 \kappa(\mathbf{x},\mathbf{x}') + \mathrm{Tr}(\nabla_1 \nabla_2 \kappa(\mathbf{x},\mathbf{x}')) \\ &= \exp\Big(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\gamma}\Big)\Big(\mathbf{x}^\top\mathbf{x}' + \frac{1}{\gamma}\mathbf{x}^\top(\mathbf{x}'-\mathbf{x}) + \frac{1}{\gamma}(\mathbf{x}')^\top(\mathbf{x}-\mathbf{x}') + \Big(\frac{d}{\gamma} - \frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\gamma^2}\Big)\Big) \\ &= \exp\Big(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\gamma}\Big)\Big(\mathbf{x}^\top\mathbf{x}' - \frac{\gamma+1}{\gamma^2}\|\mathbf{x}-\mathbf{x}'\|_2^2 + \frac{d}{\gamma}\Big) \,. \end{aligned} \tag{9}$$

We first verify that Assumption 2 holds by adapting $\{\alpha_k\}_{k=1}^\infty$ and $\{\psi_k\}_{k=1}^\infty$ from Appendix A.1.

**Lemma 21** *Assume that $\gamma > 24$. For $k' \in \mathbb{N}$, consider*

$$\begin{aligned} \lambda_{(k'-1)(d+3)+1} &= -\frac{\gamma+1}{\gamma^2}\alpha_{k'} \,, & \phi_{(k'-1)(d+3)+1}(\mathbf{x}) &= \psi_{k'}(\mathbf{x})(\|\mathbf{x}\|_2^2+1) \,, \\ \lambda_{(k'-1)(d+3)+2} &= \frac{\gamma+1}{\gamma^2}\alpha_{k'} \,, & \phi_{(k'-1)(d+3)+2}(\mathbf{x}) &= \psi_{k'}(\mathbf{x})\|\mathbf{x}\|_2^2 \,, \\ \lambda_{(k'-1)(d+3)+3} &= \left(\frac{d}{\gamma}+\frac{\gamma+1}{\gamma^2}\right)\alpha_{k'} \,, & \phi_{(k'-1)(d+3)+3}(\mathbf{x}) &= \psi_{k'}(\mathbf{x}) \,, \end{aligned}$$

*and for $l = 1, \dots, d$, define*

$$\lambda_{(k'-1)(d+3)+3+l} = \frac{\gamma^2+2\gamma+2}{\gamma^2}\alpha_{k'} \,, \qquad \phi_{(k'-1)(d+3)+3+l}(\mathbf{x}) = \psi_{k'}(\mathbf{x})x_l \,.$$

*Then Assumption 2 holds with any $\nu \in (2, 3]$ for $u = u_P^{\mathrm{KSD}}$, $\{\lambda_k\}_{k=1}^{\infty}$ and $\{\phi_k\}_{k=1}^{\infty}$ defined above.*

The following result (proved in Appendix F.2) provides analytical forms or upper bounds for the moments of KSD U-statistic.

**Lemma 22 (KSD moments)** *Let $\kappa$ be a RBF kernel with bandwidth $\gamma = \omega(1)$, and let $\mathbf{X}, \mathbf{X}'$ be independent draws from $Q$. Under the mean-shift setup with an identity covariance matrix, it follows that*

*(i) For every $\mathbf{x} \in \mathbb{R}^d$,*

$$\begin{aligned} g^{\mathrm{KSD}}(\mathbf{x}) &\coloneqq \mathbb{E}[u_P^{\mathrm{KSD}}(\mathbf{x}, \mathbf{X}')] \\ &= \left(\frac{\gamma}{\gamma+1}\right)^{d/2} \exp\left(-\frac{1}{2(\gamma+1)}\|\mathbf{x}-\mu\|_2^2\right)\left(\frac{2+\gamma}{1+\gamma}\mu^\top\mathbf{x} - \frac{1}{1+\gamma}\|\mu\|_2^2\right) \;; \end{aligned}$$

*(ii) The mean is given by $D^{\mathrm{KSD}}(Q, P) = \left(\frac{\gamma}{\gamma+2}\right)^{d/2} \|\mu\|_2^2$;*

*(iii) The variance of the conditional expectation $g^{\mathrm{KSD}}(\mathbf{X})$ is given by*

$$\sigma_{\mathrm{cond}}^2 = \left(\frac{\gamma^2}{(1+\gamma)(3+\gamma)}\right)^{d/2}\left(\frac{(2+\gamma)^2}{(1+\gamma)(3+\gamma)}\|\mu\|_2^2 + \left(1 - \left(\frac{(1+\gamma)(3+\gamma)}{(2+\gamma)^2}\right)^{d/2}\right)\|\mu\|_2^4\right) \;;$$

*(iv) The variance of $u_P^{\mathrm{KSD}}(\mathbf{X}, \mathbf{X}')$ is given by*

$$\begin{aligned} \sigma_{\mathrm{full}}^2 = \left(\frac{\gamma}{4+\gamma}\right)^{d/2}\Bigg(& d + \frac{d^2}{\gamma^2} + \frac{2d\|\mu\|_2^2}{\gamma} + 2\|\mu\|_2^2 + \left(1 - \left(\frac{\gamma(4+\gamma)}{(2+\gamma)^2}\right)^{d/2}\right)\|\mu\|_2^4 \\ &+ o\left(d + \frac{d^2}{\gamma^2} + \frac{d\|\mu\|_2^2}{\gamma} + \|\mu\|_2^2\right)\Bigg) \;; \end{aligned}$$

*(v) For any $\nu > 2$, there exist positive constants $C_1, C_2$ depending only on $\nu$ such that the $\nu$-th absolute moment of the conditional expectation satisfies*

$$\mathbb{E}[|g^{\mathrm{KSD}}(\mathbf{X})|^\nu] \le \left(\frac{\gamma}{1+\gamma}\right)^{\nu d/2}\left(\frac{1+\gamma}{1+\nu+\gamma}\right)^{d/2}(C_1\|\mu\|_2^\nu + C_2\|\mu\|_2^{2\nu}) \,.$$

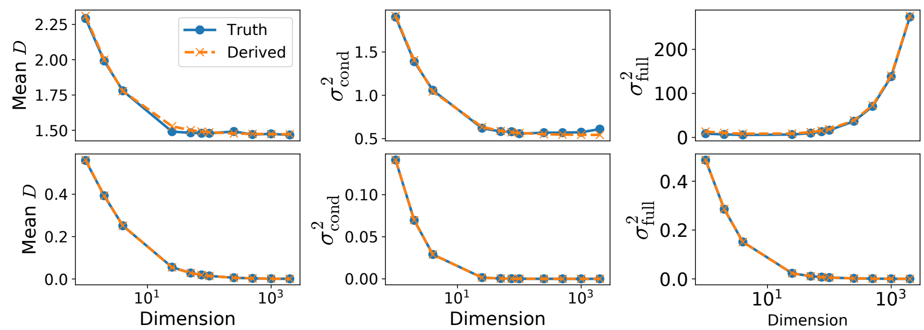


Figure 5: Verifying the analytical expressions for the first two moments of KSD and MMD. *Top.* KSD moments derived in Lemma 22. *Bottom.* MMD moments derived in Lemma 24. The ground truth is estimated using $n = 4000$ samples for KSD and $n = 10000$ samples for MMD, respectively, and the reported results are averaged over 5 random seeds.

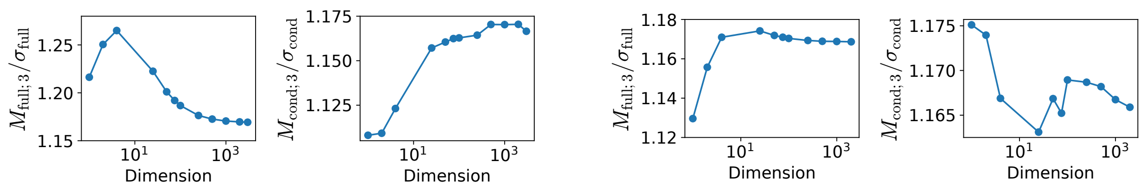


Figure 6: Verifying Assumption 1 for $\nu = 3$ for KSD and MMD with RBF kernels. All moment ratios appear to be bounded by a dimension-independent constant. *Left and middle-left.* KSD moment ratios. *Middle-right and right.* MMD moment ratios. The reported results are averaged over 5 random seeds.

*(vi) For any $\nu > 2$, there exist positive constants $C_3, C_4, C_5, C_6$ depending only on $\nu$ such that the $\nu$-th absolute moment of $u_P^{\text{KSD}}(\mathbf{X}, \mathbf{X}')$ satisfies*

$$\mathbb{E}[|u_P^{\text{KSD}}(\mathbf{X}, \mathbf{X}')|^\nu] \leq \left(\frac{\gamma}{2\nu + \gamma}\right)^{d/2} \left(C_3 d^{\nu/2} + C_4 \left(\frac{d}{\gamma}\right)^\nu + C_5 \|\mu\|_2^\nu + C_6 \|\mu\|_2^{2\nu} + o\left(d^{\nu/2} + \frac{d^\nu}{\gamma^\nu} + \frac{\|\mu\|_2^{2\nu}}{\gamma^\nu}\right)\right).$$

*In particular, when $\|\mu\|_2 = \Theta(1)$ and $\gamma = \Omega(d)$, Assumption 1 holds with any $\nu > 2$ for $u_P^{\text{KSD}}$ under $Q$.*

### A.3. MMD U-statistic with RBF kernel

Under the Gaussian mean-shift setup with an identity covariance matrix, the MMD U-statistic with a RBF kernel has the form

$$\begin{aligned} u^{\text{MMD}}(\mathbf{z}, \mathbf{z}') &= \kappa(\mathbf{x}, \mathbf{x}') + \kappa(\mathbf{y}, \mathbf{y}') - \kappa(\mathbf{x}, \mathbf{y}') - \kappa(\mathbf{x}', \mathbf{y}) \\ &= \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\gamma}\right) + \exp\left(-\frac{\|\mathbf{y} - \mathbf{y}'\|_2^2}{2\gamma}\right) - \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}'\|_2^2}{2\gamma}\right) - \exp\left(-\frac{\|\mathbf{x}' - \mathbf{y}\|_2^2}{2\gamma}\right), \end{aligned} \tag{10}$$

for $\mathbf{z} := (\mathbf{x}, \mathbf{y}), \mathbf{z}' := (\mathbf{x}', \mathbf{y}') \in \mathbb{R}^{2d}$. We first verify that Assumption 2 holds again by adapting $\{\alpha_k\}_{k=1}^\infty$ and $\{\psi_k\}_{k=1}^\infty$ from Appendix A.1.

**Lemma 23** *Assume that $\gamma > 8$. Then Assumption 2 holds with any value $\nu \in (2, 3]$ and the function $u((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) = u^{\mathrm{MMD}}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}'))$ for $\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}' \in \mathbb{R}^d$, with the sequences of values and functions given for each $k \in \mathbb{N}$ as $\gamma_k = \alpha_k$ and $\phi_k(\mathbf{x}, \mathbf{y}) = \psi_k(\mathbf{x}) - \psi_k(\mathbf{y})$.*

We next compute the moments. The analytical form of the population MMD (i.e. the expectation) has been derived in previous works under both the Gaussian mean-shift setup with a general covariance matrix $\Sigma$ (Wynne and Duncan (2022, Proposition 2, Corollary 19); (Ramdas et al., 2015, Proposition 1)) and an expression up to the learning term was also derived under a more general mean-shift setup (Reddi et al., 2015, Lemma 1). We only consider the Gaussian mean-shift case with $\Sigma = I_d$ but provide expressions for the second moments and a generic moment bound, while making minimal assumptions on the kernel bandwidth compared to Reddi et al. (2015).

**Lemma 24 (RBF-MMD moments)** *Let $\kappa$ be a RBF kernel and let $\mathbf{X}, \mathbf{X}' \sim Q$ and $\mathbf{Y}, \mathbf{Y}' \sim P$ be mutually independent draws. Under the mean-shift setup with an identity covariance matrix, it follows that*

*(i) For every $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2d}$,*

$$\begin{aligned} g^{\mathrm{mmd}}(\mathbf{z}) &:= \mathbb{E}[u^{\mathrm{MMD}}(\mathbf{z}, \mathbf{Z}')] \\ &= \left(\frac{\gamma}{1+\gamma}\right)^{d/2} \left(e^{-\frac{1}{2(1+\gamma)}\|\mathbf{x}-\mu\|_2^2} + e^{-\frac{1}{2(1+\gamma)}\|\mathbf{y}\|_2^2} - e^{-\frac{1}{2(1+\gamma)}\|\mathbf{x}\|_2^2} - e^{-\frac{1}{2(1+\gamma)}\|\mathbf{y}-\mu\|_2^2}\right); \end{aligned}$$

*(ii) The mean is given by $D^{\mathrm{MMD}}(Q, P) = 2\left(\frac{\gamma}{2+\gamma}\right)^{d/2}\left(1 - \exp\left(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\right)\right)$;*

*(iii) The variance of the conditional expectation is given by*

$$\begin{aligned} \sigma_{\mathrm{cond}}^2 &= 2\left(\frac{\gamma}{1+\gamma}\right)^{d/2}\left(\frac{\gamma}{3+\gamma}\right)^{d/2} \\ &\times \Bigg(1 + \exp\left(-\frac{1}{3+\gamma}\|\mu\|_2^2\right) + 2\left(\frac{3+\gamma}{2+\gamma}\right)^{d/2}\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2}\exp\left(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\right) \\ &\quad - 2\exp\left(-\frac{7+5\gamma}{4(1+\gamma)(3+\gamma)}\|\mu\|_2^2\right) - \left(\frac{3+\gamma}{2+\gamma}\right)^{d/2}\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2} \\ &\quad - \left(\frac{3+\gamma}{2+\gamma}\right)^{d/2}\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2}\exp\left(-\frac{1}{2+\gamma}\|\mu\|_2^2\right)\Bigg); \end{aligned}$$

*(iv) The variance is given by*

$$\begin{aligned} \sigma_{\mathrm{full}}^2 &= 2\left(\frac{\gamma}{4+\gamma}\right)^{d/2}\left(1 + \exp\left(-\frac{1}{4+\gamma}\|\mu\|_2^2\right)\right) - 2\left(\frac{\gamma}{2+\gamma}\right)^{d} \\ &\quad - 8\left(\frac{\gamma}{3+\gamma}\right)^{d/2}\left(\frac{\gamma}{1+\gamma}\right)^{d/2}\exp\left(-\frac{2+\gamma}{2(1+\gamma)(3+\gamma)}\|\mu\|_2^2\right) \\ &\quad - 2\left(\frac{\gamma}{2+\gamma}\right)^{d}\exp\left(-\frac{1}{2+\gamma}\|\mu\|_2^2\right) + 8\left(\frac{\gamma}{2+\gamma}\right)^{d}\exp\left(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\right). \end{aligned}$$

While we do not verify Assumption 1 here, we remark that Gao and Shao (2021) also encounter similar moment ratios when deriving finite-sample bounds for MMD with a studentised version of U-statistic (see e.g. their Theorem 13). They show that those ratios are controlled under an elaborate list of assumptions; in particular, those assumptions hold for the RBF kernel under a condition that amounts to choosing $\gamma = \Theta(d)$ in our Gaussian mean-shift setup. For our case, as discussed, a rigorous verification of Assumption 1 can be done by following the proofs of Propositions 18 and 19 to control $M_{\mathrm{cond};\nu}$ and $M_{\mathrm{full};\nu}$ for any $\nu > 2$. Fig. 6 also verifies Assumption 1 by simulation.

### A.4. MMD U-statistic with linear kernel

In this section, we consider the mean-shift setup with a general covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, i.e., $Q = \mathcal{N}(\mu, \Sigma)$ and $P = \mathcal{N}(0, \Sigma)$. The MMD with a linear kernel $\kappa(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$ has the form

$$u^{\mathrm{MMD}}(\mathbf{z}, \mathbf{z}') \ = \mathbf{x}^\top \mathbf{x}' + \mathbf{y}^\top \mathbf{y}' - \mathbf{x}^\top \mathbf{y}' - \mathbf{y}^\top \mathbf{x}' \ ,$$

where $\mathbf{z} \coloneqq (\mathbf{x}, \mathbf{y}), \mathbf{z}' \coloneqq (\mathbf{x}', \mathbf{y}') \in \mathbb{R}^{2d}$. In this case, Assumption 2 holds directly because we can represent $u^{\mathrm{MMD}}$ as

$$u^{\mathrm{MMD}}(\mathbf{z}, \mathbf{z}') \ = \ (\mathbf{x} - \mathbf{y})^\top (\mathbf{x}' - \mathbf{y}') \ = \ \textstyle\sum_{l=1}^d (x_l - y_l)(x'_l - y'_l) \ = \ \textstyle\sum_{l=1}^d \gamma_l \psi_l(\mathbf{z}) \psi_l(\mathbf{z}') \ , \quad (11)$$

where $\gamma_l = 1$, $\psi_l(\mathbf{z}) = x_l - y_l$ and $\psi_l(\mathbf{z}') = x'_l - y'_l$.

We next compute the moment terms and verify that Assumption 1 holds. The next result, proved in Appendix F.4, gives the analytical expressions of the first two moments of the linear-kernel MMD.

**Lemma 25 (Linear-MMD moments)** *Let $\kappa$ be a linear kernel, and let $\mathbf{X}, \mathbf{X}' \sim Q$ and $\mathbf{Y}, \mathbf{Y}' \sim P$ be mutually independent draws. Write $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ and $\mathbf{Z}' = (\mathbf{X}', \mathbf{Y}')$. Under the mean-shift setup, it follows that*

*(i) For every $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2d}$, we have $g^{\mathrm{mmd}}(\mathbf{z}) \coloneqq \mathbb{E}[u^{\mathrm{MMD}}(\mathbf{z}, \mathbf{Z}')] = \mu^\top \mathbf{y} - \mu^\top \mathbf{x}$ ;*

*(ii) The mean is given by $D^{\mathrm{MMD}}(Q, P) = \|\mu\|_2^2$ ;*

*(iii) The variance of the conditional expectation $g^{\mathrm{mmd}}(\mathbf{Z})$ is given by $\sigma^2_{\text{cond}} = 2\mu^\top \Sigma \mu$ ;*

*(iv) The variance of $u^{\mathrm{MMD}}(\mathbf{Z}, \mathbf{Z}')$ is given by $\sigma^2_{\text{full}} = 4\mathrm{Tr}(\Sigma^2) + 4\mu^\top \Sigma \mu$ ;*

*(v) The third central moment of $g^{\mathrm{mmd}}(\mathbf{Z})$ satisfies $M^3_{\text{cond};3} \leq C(\mu^\top \Sigma \mu)^{3/2}$ for some absolute constant $C$ ;*

*(vi) The third central moment of $u^{\mathrm{MMD}}(\mathbf{Z}, \mathbf{Z}')$ satisfies $M^3_{\text{full};3} \leq C\big(\mathrm{Tr}(\Sigma^2) + \mu^\top \Sigma \mu\big)^{3/2}$ for some absolute constant $C$ .*

*In particular, Assumption 1 holds with $\nu = 3$ for $u^{\mathrm{MMD}}$ defined in (11) under $Q$.*

In the last example in Section 4.3, we chose $\mu = (0, 10, \ldots, 0) \in \mathbb{R}^d$ and a diagonal $\Sigma$ with $\Sigma_{11} = 0.5(d + 1)$, $\Sigma_{ii} = 0.5$ for $i > 1$ and $\Sigma_{ij} = 0$ otherwise. Note that by the invariance of Gaussian distributions under orthogonal transformation, this is equivalent to choosing $\Sigma$ as $0.5\mathbf{I}_d + 0.5\mathbf{J}_d$, where $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix, $\mathbf{J}_d \in \mathbb{R}^{d\times d}$ is the all-one matrix and $\mu$ is transformed by an appropriate orthogonal matrix of eigenvectors. Notably, this choice ensures the limit of $u^{\mathrm{MMD}}$ remains non-Gaussian. Indeed, when $Q$ and $P$ are Gaussian, the statistic $D_n^{\mathrm{MMD}}$ can be written as a sum of shifted-and-rescaled chi-squares, where the scaling factors are $0.5(d + 1), 0.5, \ldots, 0.5$, the eigenvalues of $\Sigma$. As $d$ grows, the eigenvalue $0.5(d + 1)$ dominates, and the limiting distribution is then dominated by the first summand, thereby yielding a chi-square limit up to shifting and rescaling. This is numerically demonstrated in the right figure of Fig. 1. As a remark, we do not expect this exact setting to occur in practice; it should instead be treated as a toy setup to demonstrate the possibility of non-Gaussianity and convey an intuition of when this may occur.

## Appendix B. Auxiliary tools

### B.1. Generic moment bounds

We first present two-sided bounds on the moments of a martingale, which are useful in bounding $\nu$-th moment terms of different statistics. The original result is due to Burkholder (1966), and the constant $C_\nu$ is provided by von Bahr and Esseen (1965) and Dharmadhikari et al. (1968).

**Lemma 26** *Fix $\nu > 1$. For a martingale difference sequence $Y_1, \dots, Y_n$ taking values in $\mathbb{R}$,*

$$c_\nu \, n^{\min\{0,\nu/2-1\}} \sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] \;\leq\; \mathbb{E}\Big[\Big|\sum_{i=1}^n Y_i\Big|^\nu\Big] \;\leq\; C_\nu \, n^{\max\{0,\nu/2-1\}} \sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] \;,$$

*for $C_\nu := \max\big\{2, (8(\nu-1)\max\{1, 2^{\nu-3}\})^\nu\big\}$ and some absolute constant $c_\nu > 0$ that depends only on $\nu$.*

The next moment computation for a quadratic form of Gaussians is used throughout the proof:

**Lemma 27 (Lemma 2.3, Magnus (1978))** *Given a standard Gaussian vector $\eta$ in $\mathbb{R}^m$ and a symmetric $m \times m$ matrix $A$, we have that $\mathbb{E}[\eta^\top A \eta] = \mathrm{Tr}(A)$ and*

$$\mathbb{E}[(\eta^\top A \eta)^2] \;=\; \mathrm{Tr}(A)^2 + 2\mathrm{Tr}(A^2) \;, \quad \mathbb{E}[(\eta^\top A \eta)^3] \;=\; \mathrm{Tr}(A)^3 + 6\mathrm{Tr}(A)\mathrm{Tr}(A^2) + 8\mathrm{Tr}(A^3) \;.$$

The next two results are used for the moment computation involving an RBF kernel.

**Lemma 28** *Fix $\mathbf{m}_i \in \mathbb{R}^d$ and $a_i > 0$ for $i = 1, 2$. Let $\mathbf{X} \sim \mathcal{N}(\mathbf{m}_1, a_1^2 I_d)$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be a deterministic function such that $\mathbb{E}[|f(\mathbf{X})|] < \infty$. It follows that*

$$\mathbb{E}\Big[f(\mathbf{X}) \exp\Big(-\frac{\|\mathbf{X} - \mathbf{m}_2\|_2^2}{2a_2^2}\Big)\Big] \;=\; \Big(\frac{a_2^2}{a_1^2 + a_2^2}\Big)^{d/2} \exp\Big(-\frac{\|\mathbf{m}_1 - \mathbf{m}_2\|_2^2}{2(a_1^2 + a_2^2)}\Big)\mathbb{E}[f(\mathbf{W})] \;,$$

*where $\mathbf{W} \sim \mathcal{N}(\mathbf{m}, a^2 I_d)$ with $\mathbf{m} := \frac{a_1^2 a_2^2}{a_1^2 + a_2^2}\big(\frac{1}{a_1^2}\mathbf{m}_1 + \frac{1}{a_2^2}\mathbf{m}_2\big)$ and $a^2 := \frac{a_1^2 a_2^2}{a_1^2 + a_2^2}$ .*

**Lemma 29** *Fix $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^d$ and $a_i > 0$ for $i = 1, 2, 3$. Let $\mathbf{X} \sim \mathcal{N}(\mathbf{m}_1, a_1^2 I_d)$ and $\mathbf{X}' \sim \mathcal{N}(\mathbf{m}_2, a_2^2 I_d)$, and let $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a deterministic function with $\mathbb{E}[|f(\mathbf{X}, \mathbf{X}')|] < \infty$. Then*

$$\mathbb{E}\Big[f(\mathbf{X}, \mathbf{X}') \exp\Big(-\frac{\|\mathbf{X} - \mathbf{X}'\|_2^2}{2a_3^2}\Big)\Big]$$
$$= \Big(\frac{a_3^2}{a_1^2 + a_2^2 + a_3^2}\Big)^{d/2} \exp\Big(-\frac{\|\mathbf{m}_1 - \mathbf{m}_2\|_2^2}{2(a_1^2 + a_2^2 + a_3^2)}\Big)\mathbb{E}\Big[f\Big(\mathbf{W}, \mathbf{W}' + \frac{a_2^2}{a_2^2 + a_3^2}\mathbf{W}\Big)\Big] \;,$$

*where $\mathbf{W} \sim \mathcal{N}\big(\mathbf{m}, a^2 I_d\big)$ and $\mathbf{W}' \sim \mathcal{N}\big(\mathbf{m}', (a')^2 I_d\big)$ are independent with*

$$\mathbf{m} \;:=\; \frac{a_1^2(a_2^2 + a_3^2)}{a_1^2 + a_2^2 + a_3^2}\Big(\frac{1}{a_1^2}\mathbf{m}_1 + \frac{1}{a_2^2 + a_3^2}\mathbf{m}_2\Big) \;, \qquad a^2 \;:=\; \frac{a_1^2(a_2^2 + a_3^2)}{a_1^2 + a_2^2 + a_3^2} \;,$$
$$\mathbf{m}' \;:=\; \frac{a_3^2}{a_2^2 + a_3^2}\mathbf{m}_2 \;, \qquad (a')^2 \;:=\; \frac{a_2^2 a_3^2}{a_2^2 + a_3^2} \;.$$

### B.2. Moment bounds for U-statistics

We first present a result that bounds the moments of a U-statistic $D_n$ defined as in (1).

**Lemma 30** *Fix $n \geq 2$ and $\nu \geq 2$. Then, there exist absolute constants $c_\nu, C_\nu > 0$ depending only on $\nu$ such that*

$$\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] \;\leq\; C_\nu \, n^{\nu/2}(n-1)^{-\nu} M_{\mathrm{cond};\nu}^\nu + C_\nu \,(n-1)^{-\nu} M_{\mathrm{full};\nu}^\nu \;,$$
$$\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] \;\geq\; c_\nu n(n-1)^{-\nu} M_{\mathrm{cond};\nu}^\nu + c_\nu n^{-(\nu-1)}(n-1)^{-(\nu-1)} M_{\mathrm{full};\nu}^\nu \;.$$

*In other words,*

$$\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] \;=\; O(n^{-\nu/2} M_{\mathrm{cond};\nu}^\nu + n^{-\nu} M_{\mathrm{full};\nu}^\nu) \;,$$
$$\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] \;=\; \Omega(n^{-(\nu-1)} M_{\mathrm{cond};\nu}^\nu + n^{-2(\nu-1)} M_{\mathrm{full};\nu}^\nu) \;.$$

The next two results summarise how the moments of variables under the functional decomposition in Assumption 2 interact with the moments of the original statistic $u$ under $R$:

**Lemma 31** *Let $\{\phi_k\}_{k=1}^{\infty}$, $\{\lambda_k\}_{k=1}^{\infty}$ and $\varepsilon_{K;\nu}$ be defined as in Assumption 2. For $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d.}{\sim} R$, write $\mu_k := \mathbb{E}[\phi_k(\mathbf{X}_1)]$ and let the moment terms $D, M_{\text{cond};\nu}, M_{\text{full};\nu}$ be defined as in Section 2.1. Then we have the following:*

*(i)* $\big|\sum_{k=1}^K \lambda_k \mu_k^2 - D\big| \leq \varepsilon_{K;1}$;

*(ii) for any $\nu \in [1,3]$, we have that*

$$\tfrac{1}{4}(M_{\text{cond};\nu})^\nu - \varepsilon_{K;\nu}^\nu \;\leq\; \mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\mu_k\Big|^\nu\Big] \;\leq\; 4((M_{\text{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu) \;;$$

*(iii) there exist some absolute constants $c, C > 0$ such that*

$$\mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \leq 4C(M_{\text{full};\nu})^\nu - \frac{1}{2}(M_{\text{cond};\nu})^\nu + (4C+2)\varepsilon_{K;\nu}^\nu \;,$$
$$\mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \geq \frac{c}{4}(M_{\text{full};\nu})^\nu - 8(M_{\text{cond};\nu})^\nu - (c+8)\varepsilon_{K;\nu}^\nu \;.$$

The next result assumes the notations of Lemma 31, and additionally denotes

$$\Lambda^K \;\coloneqq\; \text{diag}\{\lambda_1, \ldots, \lambda_K\} \;\in \mathbb{R}^{K\times K} \;, \qquad \phi^K(x) \;\coloneqq\; (\phi_1(x), \ldots, \phi_K(x))^\top \;\in \mathbb{R}^K \;.$$

**Lemma 32** *For $\mu^K \coloneqq \mathbb{E}[\phi^K(\mathbf{X}_1)]$ and $\Sigma^K \coloneqq \text{Cov}[\phi^K(\mathbf{X}_1)]$, we have*

$$\sigma_{\text{cond}}^2 - 4\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2}^2 \;\leq (\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) \;\leq\; (\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2 \;.$$
$$(\sigma_{\text{full}} - \varepsilon_{K;2})^2 \;\leq \text{Tr}((\Lambda^K\Sigma^K)^2) \;\leq\; (\sigma_{\text{full}} + \varepsilon_{K;2})^2 \;.$$

*In particular, for $\nu \in [1,3]$ and two i.i.d. zero-mean Gaussian vector $\mathbf{Z}_1$ and $\mathbf{Z}_2$ in $\mathbb{R}^K$ with variance $\Sigma^K$, there exists some absolute constant $C > 0$ such that*

$$\mathbb{E}[|(\mu^K)^\top \Lambda^K \mathbf{Z}_1|^\nu] \;\leq\; 7\big(\sigma_{\text{cond}}^\nu + 8\varepsilon_{K;2}^\nu\big) \;, \qquad \mathbb{E}[|\mathbf{Z}_1^\top \Lambda^K \mathbf{Z}_2|^\nu] \;\leq\; 6\big(\sigma_{\text{full}}^\nu + \varepsilon_{K;2}^\nu\big) \;,$$
$$\mathbb{E}\big[|(\phi^K(\mathbf{X}_1) - \mu^K)^\top \Lambda^K \mathbf{Z}_1|^\nu\big] \;\leq\; 8C(M_{\text{full};\nu})^\nu - (M_{\text{cond};\nu})^\nu + (8C+4)\varepsilon_{K;\nu}^\nu \;.$$

The next lemma gives an equivalent expression for $W_n^K$ defined in (5) and also controls the moments of $W_n^K$.

**Lemma 33** *Let $\{\eta_i^K\}_{i=1}^n$ be a sequence of i.i.d. standard Gaussian vectors in $\mathbb{R}^K$. Then*

*(i) the distribution of $W_n^K$ satisfies*

$$W_n^K \;\overset{d}{=}\; \frac{1}{n^{3/2}(n-1)^{1/2}}\Big(\sum_{i,j=1}^n (\eta_i^K)^\top (\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K - n\text{Tr}(\Sigma^K\Lambda^K)\Big) + D \;;$$

*(ii) the mean satisfies $\mathbb{E}[W_n^K] = D$ for every $K \in \mathbb{N}$;*

*(iii) the variance is controlled as*

$$\frac{2}{n(n-1)}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 \;\leq \text{Var}[W_n^K] \;\leq\; \frac{2}{n(n-1)}(\sigma_{\text{full}} + \varepsilon_{K;2})^2 \;;$$

*(iv) the third central moment is controlled as*

$$\mathbb{E}\big[(W_n^K - D)^3\big] \;\leq \frac{8\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] - M_{\text{full};3}^3 + (M_{\text{full};3} + \varepsilon_{K;3})^3\big)}{n^{3/2}(n-1)^{3/2}} \;,$$
$$\mathbb{E}\big[(W_n^K - D)^3\big] \;\geq \frac{8\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] + M_{\text{full};3}^3 - (M_{\text{full};3} + \varepsilon_{K;3})^3\big)}{n^{3/2}(n-1)^{3/2}} \;;$$

*(v) the fourth central moment is controlled as*

$$\begin{aligned}\mathbb{E}\big[(W_n^K - D)^4\big] \;\leq\; & \frac{12}{n^2(n-1)^2}\Big(4\,\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] \\ & - 4M_{\text{full};4}^4 + 4(M_{\text{full};4} + \varepsilon_{K;4})^4 + (\sigma_{\text{full}} + \varepsilon_{K;2})^4\Big) \;, \\ \mathbb{E}\big[(W_n^K - D)^4\big] \;\geq\; & \frac{12}{n^2(n-1)^2}\Big(4\,\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] \\ & + 4M_{\text{full};4}^4 - 4(M_{\text{full};4} + \varepsilon_{K;4})^4 + (\sigma_{\text{full}} - \varepsilon_{K;2})^4\Big) \;;\end{aligned}$$

*(vi) we also have a generic moment bound: For $m \in \mathbb{N}$, there exists some absolute constant $C_m > 0$ depending only on $m$ such that*

$$\mathbb{E}\big[(W_n^K)^{2m}\big] \;\leq \frac{C_m}{n^m(n-1)^m}(\sigma_{\text{full}} + \varepsilon_{K;2})^{2m} + C_m\, D^{2m} \;;$$

*(vii) if Assumption 2 holds for some $\nu \geq 2$ then $\lim_{K\to\infty} \text{Var}[W_n^K] = \frac{2}{n(n-1)}\sigma_{\text{full}}^2$. If Assumption 2 holds for some $\nu \geq 3$, then*

$$\lim_{K\to\infty} \mathbb{E}\big[(W_n^K - D)^3\big] \;= \frac{8\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)]}{n^{3/2}(n-1)^{3/2}} \;,$$

*and if Assumption 2 holds for some $\nu \geq 4$, then*

$$\lim_{K\to\infty} \mathbb{E}\big[(W_n^K - D)^4\big] \;= \frac{12(4\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] + \sigma_{\text{full}}^4)}{n^2(n-1)^2} \;.$$

### B.3. Distribution bounds

The following is a standard approximation of an indicator function for bounding the probability of a given event; see e.g. the proof of Theorem 3.3, Chen et al. (2011).

**Lemma 34** *Fix any $m \in \mathbb{N}\cup\{0\}$ and $\tau, \delta \in \mathbb{R}$. Then there exists an $m$-times differentiable $\mathbb{R} \to \mathbb{R}$ function $h_{m;\tau;\delta}$ such that $h_{m;\tau+\delta;\delta}(x) \;\leq\; \mathbb{I}_{\{x>\tau\}} \leq h_{m;\tau;\delta}(x)$. For $0 \leq r \leq m$, the $r$-th derivative $h^{(r)}_{m;\tau;\delta}$ is continuous and bounded above by $\delta^{-r}$. Moreover, for every $\epsilon \in [0,1]$, $h^{(m)}$ satisfies that*

$$|h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \;\leq\; C_{m,\epsilon}\,\delta^{-(m+\epsilon)}\,|x-y|^{\epsilon} \;,$$

*with respect to the constant $C_{m,\epsilon} = \binom{m}{\lfloor m/2 \rfloor}(m+1)^{m+\epsilon}$.*

The next bound is useful for approximating the distribution of a sum of two (possibly correlated) random variables $X$ and $Y$ by the distribution of $X$ alone, provided that the influence of $Y$ is small.

**Lemma 35** *For two real-valued random variables $X$ and $Y$, any $a, b \in \mathbb{R}$ and $\epsilon > 0$, we have*

$$\mathbb{P}(a \leq X + Y \leq b) \;\leq\; \mathbb{P}(a - \epsilon \leq X \leq b + \epsilon) + \mathbb{P}(|Y| \geq \epsilon) \;,$$
$$\mathbb{P}(a \leq X + Y \leq b) \;\geq\; \mathbb{P}(a + \epsilon \leq X \leq b - \epsilon) - \mathbb{P}(|Y| \geq \epsilon) \;.$$

Theorem 8 of Carbery and Wright (2001) gives a general anti-concentration result for a polynomial of random variables drawn from a log-concave density. The next lemma restates the result in the case of a quadratic form of a $K$-dimensional standard Gaussian vector $\eta$.

**Lemma 36** *Let $p(\mathbf{x})$ be a degree-two polynomial of $\mathbf{x} \in \mathbb{R}^K$ taking values in $\mathbb{R}$. Then there exists an absolute constant $C$ independent of $p$ and $\eta$ such that, for every $t \in \mathbb{R}$,*

$$\mathbb{P}\big(|p(\eta)| \leq t\big) \;\leq\; Ct^{1/2}(\mathbb{E}[|p(\eta)|^2])^{-1/4} \;\leq\; Ct^{1/2}(\text{Var}[p(\eta)])^{-1/4} \;.$$

### B.4. Weak Mercer representation

In Section 4.2, we have used the *weak Mercer representation* from Steinwart and Scovel (2012). We summarise their result below, which combines their Lemma 2.3, Lemma 2.12 and Corollary 3.2:

**Lemma 37** *Consider a probability measure $R$ on $\mathbb{R}^b$, $\mathbf{V}_1, \mathbf{V}_2 \overset{i.i.d.}{\sim} R$ and a measurable kernel $\kappa^*$ on $\mathbb{R}^b$. If $\mathbb{E}[\kappa^*(\mathbf{V}_1, \mathbf{V}_1)] < \infty$, there exists a sequence of functions $\{\phi_k\}_{k=1}^\infty$ in $L_2(\mathbb{R}^b, R)$ and a bounded sequence of non-negative values $\{\lambda_k\}_{k=1}^\infty$ with $\lim_{k\to\infty} \lambda_k = 0$, such that as $K$ grows, $\big|\sum_{k=1}^K \lambda_k \phi_k(\mathbf{V}_1)\phi_k(\mathbf{V}_2) - \kappa^*(\mathbf{V}_1, \mathbf{V}_2)\big| \to 0$. The series converges $R \otimes R$ almost surely.*

## Appendix C. Proof of the main result

In this section, we prove Theorem 2. The proof is necessarily tedious as we seek to control "spectral" approximation errors (i.e. the error from a truncated functional decomposition) and multiple stochastic approximation errors at the same time. The section is organised as follows:

- In Appendix C.1, we list notations and key lemmas that formalise the steps in the proof outline in Section 3.3;
- In Appendix C.2, we present the proof body of Theorem 2, which directly combines results from the different lemmas;
- In Appendix C.3, C.4, Appendix C.5 and C.6, we present the proof of the key lemmas. Each section starts with an informal sketch of proof ideas followed by the actual proof of the result.

### C.1. Auxiliary lemmas

Recall that our goal is to study the distribution of

$$D_n \;\coloneqq\; \tfrac{1}{n(n-1)} \textstyle\sum_{1\leq i\neq j\leq n} u(\mathbf{X}_i, \mathbf{X}_j) \;.$$

The three results in this section form the key steps of the proof. We fix $\sigma > 0$ to be some normalisation constant to be chosen later.

*1. "Spectral" approximation.* For $K \in \mathbb{N}$, we define the truncated version of $D_n$ by

$$D_n^K \coloneqq \frac{1}{n(n-1)} \sum\nolimits_{1 \leq i \neq j \leq n} \sum\nolimits_{k=1}^K \lambda_k \phi_k(\mathbf{X}_i) \phi_k(\mathbf{X}_j) = \frac{1}{n(n-1)} \sum\nolimits_{1 \leq i \neq j \leq n} (\phi^K(\mathbf{X}_i))^\top \Lambda^K \phi^K(\mathbf{X}_j) \;.$$

We also denote the rescaled statistics for convenience as

$$\tilde{D}_n \;\coloneqq\; \frac{\sqrt{n(n-1)}}{\sigma} D_n \;, \qquad\qquad \tilde{D}_n^K \;\coloneqq\; \frac{\sqrt{n(n-1)}}{\sigma} D_n^K \;.$$

The first lemma allows us to study the distribution of $D_n^K$ in lieu of that of $D_n$ up to some approximation error that vanishes as $K$ grows.

**Lemma 38** *Fix $\delta, \sigma > 0$, $K \in \mathbb{N}$ and $t \in \mathbb{R}$. Then*

$$\mathbb{P}(\tilde{D}_n^K > t + \delta) - \varepsilon'_K \leq \mathbb{P}(\tilde{D}_n > t) \leq \mathbb{P}(\tilde{D}_n^K > t - \delta) + \varepsilon'_K \;, \quad \varepsilon'_K \coloneqq \frac{3 n^{1/4} (n-1)^{1/4} \varepsilon_{K;1}^{1/2}}{\sigma^{1/2} \delta^{1/2}} \;.$$

*2. Gaussian approximation via Lindeberg's technique.* The distribution of $D_n^K$ is easier to handle, as it is a double sum of a simple quadratic form of $K$-dimensional random vectors. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be i.i.d. Gaussian random vectors in $\mathbb{R}^K$ with mean and variance matching those of $\phi^K(\mathbf{X}_1)$, and denote $Z_{ik}$ as the $k$-th coordinate of $\mathbf{Z}_i$. The goal is to replace $D_n^K$ by the random variable

$$D_Z^K \;\coloneqq\; \frac{1}{n(n-1)} \sum\nolimits_{1 \leq i \neq j \leq n} \mathbf{Z}_i^\top \Lambda^K \mathbf{Z}_j \;=\; \frac{1}{n(n-1)} \sum\nolimits_{1 \leq i \neq j \leq n} \sum\nolimits_{k=1}^K \lambda_k Z_{ik} Z_{jk} \;.$$

Notice that $D_Z^K$ takes the same form as $D_n^K$ except that each $\phi^K(\mathbf{X}_i)$ is replaced by $\mathbf{Z}_i$. Analogous to $\tilde{D}_n$ and $\tilde{D}_n^K$, we also define a rescaled version as

$$\tilde{D}_Z^K \;\coloneqq\; \frac{\sqrt{n(n-1)}\, D_Z^K}{\sigma} \;.$$

The second lemma replaces the distribution $\tilde{D}_n^K$ by that of $\tilde{D}_Z^K$, up to some approximation error that vanishes as $n$ grows:

**Lemma 39** *Fix $\delta, \sigma > 0$, $K \in \mathbb{N}$, $t \in \mathbb{R}$ and any $\nu \in (2, 3]$. Then*

$$\mathbb{P}(\tilde{D}_n^K > t - \delta) \;\leq\; \mathbb{P}(\tilde{D}_Z^K > t - 2\delta) + E_{\delta;K} \;, \quad \mathbb{P}(\tilde{D}_n^K > t + \delta) \;\geq\; \mathbb{P}(\tilde{D}_Z^K > t + 2\delta) - E_{\delta;K} \;,$$

*where the approximation error is defined as, for some absolute constant $C > 0$,*

$$E_{\delta;K} \;\coloneqq\; \frac{C}{\delta^\nu n^{\nu/2-1}} \Big( \frac{(M_{\mathrm{full};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{\sigma^\nu} + \frac{(M_{\mathrm{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{(n-1)^{-\nu/2}\, \sigma^\nu} \Big) \;.$$

*3. Replace $D_Z^K$ by $U_n^K$.* As in the statement of Theorem 2, let $\{\eta_i^K\}_{i=1}^n$ be the i.i.d. standard normal vectors in $\mathbb{R}^K$, and recall the notations $\mu^K := \mathbb{E}[\phi^K(\mathbf{X}_1)]$ and $\Sigma^K := \text{Cov}[\phi^K(\mathbf{X}_1)]$. We can then express $D_Z^K$ as

$$\begin{aligned} D_Z^K &= \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \big((\Sigma^K)^{1/2}\eta_i^K + \mu^K\big)^\top \Lambda^K \big((\Sigma^K)^{1/2}\eta_j^K + \mu^K\big) \\ &= \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K + \frac{2}{n} \sum_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K \\ &\quad + (\mu^K)^\top \Lambda^K \mu^K \,. \end{aligned}$$

This is similar to the desired variable $U_n^K$ except for the third term:

$$U_n^K = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K + \frac{2}{n} \sum_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K + D \,.$$

As before, we denote $\tilde{U}_n^K := \frac{\sqrt{n(n-1)}U_n^K}{\sigma}$. The next lemma shows that the distribution of $\tilde{D}_Z^K$ can be approximated by that of $\tilde{U}_n^K$, up to some approximation error that vanishes as $K \to \infty$.

**Lemma 40** *For any $a, b \in \mathbb{R}$ and $\epsilon > 0$, we have that*

$$\begin{aligned} \mathbb{P}(a \leq \tilde{D}_Z^K \leq b) \;&\leq\; \mathbb{P}\big(a - \epsilon \leq \tilde{U}_n^K \leq b + \epsilon\big) + \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \,, \\ \mathbb{P}(a \leq \tilde{D}_Z^K \leq b) \;&\geq\; \mathbb{P}(a + \epsilon \leq \tilde{U}_n^K \leq b - \epsilon) - \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \,. \end{aligned}$$

*4. Bound the distribution of $\tilde{U}_n^K$ over a short interval.* If we are to use Lemma 38 and Lemma 39 directly, we would end up comparing $\mathbb{P}(\tilde{D}_n > t)$ against the probabilities $\mathbb{P}(\tilde{U}_n^K > t + 2\delta)$ and $\mathbb{P}(\tilde{U}_n^K > t - 2\delta)$ for some small $\delta > 0$. It turns out these are not too different from $\mathbb{P}(\tilde{U}_n^K > t)$: As $\tilde{U}_n^K$ is a quadratic form of Gaussians, we can ensure it is "well spread-out" such that the probability mass of $\tilde{U}_n^K$ within a small interval $(t - 2\delta, t + 2\delta)$ is not too large. This is ascertained by the following lemma:

**Lemma 41** *For $a \leq b \in \mathbb{R}$, there exists some absolute constant $C$ such that*

$$\mathbb{P}(a \leq \tilde{U}_n^K \leq b) \;\leq C(b-a)^{1/2} \Big( \frac{1}{\sigma^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma^2}(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2}) \Big)^{-1/4} \,.$$

### C.2. Proof body of Theorem 2

Fix $\delta, \sigma > 0$, $K \in \mathbb{N}$ and $t \in \mathbb{R}$. By Lemma 38, we have that

$$\mathbb{P}(\tilde{D}_n^K > t + \delta) - \varepsilon'_K \leq \mathbb{P}(\tilde{D}_n > t) \leq \mathbb{P}(\tilde{D}_n^K > t - \delta) + \varepsilon'_K \,, \quad \varepsilon'_K := \frac{3n^{1/4}(n-1)^{1/4}\varepsilon_{K;1}^{1/2}}{\sigma^{1/2}\delta^{1/2}} \,.$$

By Lemma 39, we have

$$\mathbb{P}(\tilde{D}_n^K > t - \delta) \;\leq\; \mathbb{P}(\tilde{D}_Z^K > t - 2\delta) + E_{\delta;K} \,, \quad \mathbb{P}(\tilde{D}_n^K > t + \delta) \;\geq\; \mathbb{P}(\tilde{D}_Z^K > t + 2\delta) - E_{\delta;K} \,,$$

where the error term is defined as, for some absolute constant $C' > 0$,

$$E_{\delta;K} \;\coloneqq\; \frac{C'}{\delta^\nu n^{\nu/2-1}} \Big( \frac{(M_{\text{full};\nu})^\nu + \varepsilon^\nu_{K;\nu}}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu + \varepsilon^\nu_{K;\nu}}{(n-1)^{-\nu/2}\,\sigma^\nu} \Big) \;.$$

To combine the two bounds, we consider the following decomposition:

$$\begin{aligned} \mathbb{P}(\tilde{D}^K_Z > t - 2\delta) \;&=\; \mathbb{P}(\tilde{D}^K_Z > t) + \mathbb{P}(t - 2\delta < \tilde{D}^K_Z \leq t) \;, \\ \mathbb{P}(\tilde{D}^K_Z > t + 2\delta) \;&=\; \mathbb{P}(\tilde{D}^K_Z > t) - \mathbb{P}(t < \tilde{D}^K_Z \leq t + 2\delta) \;. \end{aligned} \tag{12}$$

This allows us to combine the earlier two bounds as

$$\big|\mathbb{P}(\tilde{D}_n > t) - \mathbb{P}(\tilde{D}^K_Z > t)\big| \;\leq\; \max\{\mathbb{P}(t - 2\delta \leq \tilde{D}^K_Z < t) \;, \mathbb{P}(t < \tilde{D}^K_Z \leq t + 2\delta)\} + E_{\delta;K} + \varepsilon'_K \;,$$

which gives the error of approximating the c.d.f. of $\tilde{D}_n$ by that of $\tilde{D}^K_Z$. Now fix some $\epsilon > 0$. By applying Lemma 40 and taking appropriate limits of the endpoints to change $\leq$ to $<$, $\geq$ to $>$ and taking the right endpoint to positive infinity, we can now approximate the c.d.f. of $\tilde{D}^K_Z$ by that of $\tilde{U}^K_n$:

$$\begin{aligned} \mathbb{P}(t - 2\delta \leq \tilde{D}^K_Z < t) \;&\leq\; \mathbb{P}\big(t - 2\delta - \epsilon \leq \tilde{U}^K_n < t + \epsilon\big) + \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \;, \\ \mathbb{P}(t \leq \tilde{D}^K_Z < t + 2\delta) \;&\leq\; \mathbb{P}(t - \epsilon \leq \tilde{U}^K_n < t + 2\delta + \epsilon) + \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \;, \\ \mathbb{P}(\tilde{D}^K_Z > t) \;&\leq\; \mathbb{P}\big(\tilde{U}^K_n > t - \epsilon\big) + \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \;, \\ \mathbb{P}(\tilde{D}^K_Z > t) \;&\geq\; \mathbb{P}(\tilde{U}^K_n > t + \epsilon) - \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \;. \end{aligned}$$

Substituting the bounds into the earlier bound and using a similar decomposition to (12), we get that the error of approximating the c.d.f. of $\tilde{D}_n$ by that of $\tilde{U}^K_n$ is

$$\begin{aligned} \big|\mathbb{P}(\tilde{D}_n > t) - \mathbb{P}(\tilde{U}^K_n > t)\big| \;\leq\;& \max\{\mathbb{P}(t - \epsilon \leq \tilde{U}^K_n < t) \;, \mathbb{P}(t < \tilde{U}^K_n \leq t + \epsilon)\} \\ &+ \max\{\mathbb{P}(t - 2\delta - \epsilon \leq \tilde{U}^K_n < t + \epsilon) \;, \mathbb{P}(t - \epsilon < \tilde{U}^K_n \leq t + 2\delta + \epsilon)\} \\ &+ E_{\delta;K} + \varepsilon'_K + \frac{4\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \;. \end{aligned}$$

To bound the maxima, we recall that by Lemma 41, there exists some absolute constant $C''$ such that for any $a \leq b \in \mathbb{R}$,

$$\mathbb{P}(a \leq \tilde{U}^K_n \leq b) \;\leq C''(b-a)^{1/2} \Big( \frac{1}{\sigma^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma^2}(\sigma^2_{\text{cond}} - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon) \Big)^{-1/4} \;.$$

Substituting this into the above bound while noting $(2\delta + 2\epsilon)^{1/2} \leq 2\delta^{1/2} + 2\epsilon^{1/2}$, we get that

$$\begin{aligned} &\big|\mathbb{P}(\tilde{D}_n > t) - \mathbb{P}(\tilde{U}^K_n > t)\big| \\ &\quad\leq C''\big(6\epsilon^{1/2} + 4\delta^{1/2}\big) \Big( \frac{1}{\sigma^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma^2}(\sigma^2_{\text{cond}} - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2}) \Big)^{-1/4} \\ &\qquad + E_{\delta;K} + \varepsilon'_K + \frac{4\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \;. \end{aligned}$$

We now take $K \to \infty$. By Assumption 2, $\varepsilon_{K;2} \to 0$ in the first term and the two trailing error terms vanish. The second error term becomes

$$E_{\delta;K} \;\to\; \frac{C'}{\delta^\nu n^{\nu/2-1}} \Big( \frac{(M_{\text{full};\nu})^\nu}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu} \Big) \;.$$

By additionally taking $\epsilon \to 0$ in the first term and taking a supremum over $t$ on both sides, we then obtain

$$\begin{aligned}\sup_{t \in \mathbb{R}} \Big| \mathbb{P}(\tilde{D}_n > t) - \lim_{K \to \infty} \mathbb{P}(\tilde{U}_n^K > t) \Big| \;&\leq 4C'' \delta^{1/2} \Big( \frac{\sigma^2_{\text{full}}}{\sigma^2} + \frac{\sigma^2_{\text{cond}}}{(n-1)^{-1}\sigma^2} \Big)^{-1/4} \\ &\quad + \frac{C'}{\delta^\nu n^{\nu/2-1}} \Big( \frac{(M_{\text{full};\nu})^\nu}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu} \Big) \;.\end{aligned}$$

Finally, we choose

$$\delta \;=\; n^{-\frac{\nu-2}{2\nu+1}} \Big( \frac{(M_{\text{full};\nu})^\nu}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu} \Big)^{\frac{2}{2\nu+1}}$$

and $\sigma = \sigma_{\max} \coloneqq \max\{\sigma_{\text{full}}, (n-1)^{1/2}\sigma_{\text{cond}}\}$. Then $\big(\frac{\sigma^2_{\text{full}}}{\sigma^2} + \frac{\sigma^2_{\text{cond}}}{(n-1)^{-1}\sigma^2}\big)^{-1/4} \leq 1$, and by redefining constants, we get that there exists some absolute constant $C > 0$ such that

$$\begin{aligned} &\sup_{t \in \mathbb{R}} \Big| \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\max}} D_n > t \Big) - \lim_{K \to \infty} \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\max}} U_n^K > t \Big) \Big| \\ &\leq\; C\, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{(M_{\text{full};\nu})^\nu}{\sigma^\nu_{\max}} + \frac{(M_{\text{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu_{\max}} \Big)^{\frac{1}{2\nu+1}} \\ &\leq\; 2^{\frac{1}{2\nu+1}} C\, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{M_{\max;\nu}}{\sigma_{\max}} \Big)^{\frac{\nu}{2\nu+1}} \;, \end{aligned} \tag{13}$$

where we have recalled that $M_{\max;\nu} \coloneqq \max\{M_{\text{full};\nu}, (n-1)^{1/2} M_{\text{cond};\nu}\}$. This finishes the proof.

### C.3. Proof of Lemma 38

*Proof overview.* The proof idea is reminiscent of the standard technique for proving that convergence in probability implies weak convergence. We first approximate each probability by the expectation of a $\delta^{-1}$ Lipschitz function $h$ that is uniformly bounded by 1. This introduces an approximation error of $\delta$, while replaces the difference in probability by the difference $\mathbb{E}[h(\tilde{D}_n) - h(\tilde{D}_n^K)]$. The expectation can be further split by the events $\{|\tilde{D}_n - \tilde{D}_n^K| < \epsilon\}$ and $\{|\tilde{D}_n - \tilde{D}_n^K| \geq \epsilon\}$. In the first case, the expectation can be bounded by a Lipschitz argument; in the second case, we can use the boundedness of $h$ to bound the expectation by $2\mathbb{P}(|\tilde{D}_n - \tilde{D}_n^K| \geq \epsilon)$, which is in turn bounded by a Markov argument to give the "spectral" approximation error. Choosing $\epsilon$ appropriately gives the above error term.

**Proof of Lemma 38** For any $\tau \in \mathbb{R}$ and $\delta > 0$, let $h_{\tau;\delta}$ be the function defined in Lemma 34 with $m = 0$, which satisfies

$$h_{\tau+\delta;\delta}(x) \;\leq\; \mathbb{I}_{\{x > \tau\}} \;\leq\; h_{\tau;\delta}(x) \;.$$

By applying the above bounds with $\tau$ set to $t$ and $t - \delta$, we get that

$$\mathbb{P}(\tilde{D}_n > t) - \mathbb{P}\big(\tilde{D}_n^K > t - \delta\big) \;= \mathbb{E}[\mathbb{I}_{\{\tilde{D}_n > t\}} - \mathbb{I}_{\{\tilde{D}_n^K > t-\delta\}}] \;\leq\; \mathbb{E}[h_{t;\delta}(\tilde{D}_n) - h_{t;\delta}(\tilde{D}_n^K)] \;,$$

and similarly

$$\mathbb{P}\big(\tilde{D}_n^K > t+\delta\big) - \mathbb{P}(\tilde{D}_n > t) \;\leq\; \mathbb{E}[h_{t+\delta;\delta}(\tilde{D}_n^K) - h_{t+\delta;\delta}(\tilde{D}_n)] \;.$$

Therefore, defining $\xi_\tau := |\mathbb{E}[h_{\tau;\delta}(\tilde{D}_n) - h_{\tau;\delta}(\tilde{D}_n^K)]|$, we get that

$$\mathbb{P}\big(\tilde{D}_n^K > t+\delta\big) - \xi_{t+\delta} \;\leq\; \mathbb{P}(\tilde{D}_n > t) \;\leq\; \mathbb{P}\big(\tilde{D}_n^K > t-\delta\big) + \xi_t \;.$$

To bound quantities of the form $\xi_\tau$, fix any $\epsilon > 0$ and write $\xi_\tau = \xi_{\tau,1} + \xi_{\tau,2}$ where

$$\begin{aligned}
\xi_{\tau,1} \;&:=\; \Big|\mathbb{E}\Big[\big(h_{\tau;\delta}(\tilde{D}_n) - h_{\tau;\delta}(\tilde{D}_n^K)\big)\mathbb{I}_{\{|\tilde{D}_n - \tilde{D}_n^K| \leq \epsilon\}}\Big]\Big| \;, \\
\xi_{\tau,2} \;&:=\; \Big|\mathbb{E}\Big[\big(h_{\tau;\delta}(\tilde{D}_n) - h_{\tau;\delta}(\tilde{D}_n^K)\big)\mathbb{I}_{\{|\tilde{D}_n - \tilde{D}_n^K| > \epsilon\}}\Big]\Big| \;.
\end{aligned}$$

The first term can be bounded by recalling from Lemma 34 that $h_{\tau;\delta}$ is $\delta^{-1}$-Lipschitz:

$$\xi_{\tau,1} \;\leq \delta^{-1}\mathbb{E}\big[\big|\tilde{D}_n - \tilde{D}_n^K\big|\mathbb{I}_{\{|\tilde{D}_n - \tilde{D}_n^K| \leq \epsilon\}}\big] \;\leq\; \delta^{-1}\epsilon\,\mathbb{P}\big(|\tilde{D}_n - \tilde{D}_n^K| \leq \epsilon\big) \;\leq\; \delta^{-1}\epsilon \;.$$

The second term can be bounded by noting that $h_{\tau;\delta}$ is uniformly bounded above by $1$ and applying Markov's inequality:

$$\xi_{\tau,2} \;\leq\; 2\mathbb{E}[\mathbb{I}_{\{|\tilde{D}_n - \tilde{D}_n^K| > \epsilon\}}] \;=\; 2\mathbb{P}(|\tilde{D}_n - \tilde{D}_n^K| > \epsilon) \;\leq\; 2\epsilon^{-1}\mathbb{E}\big[|\tilde{D}_n - \tilde{D}_n^K|\big] \;.$$

By the definition of $\tilde{D}_n$ and $\tilde{D}_n^K$, a triangle inequality and noting that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d., the absolute moment term can be bounded as

$$\begin{aligned}
\mathbb{E}\big[|\tilde{D}_n - \tilde{D}_n^K|\big] \;&=\; \frac{\sqrt{n(n-1)}}{\sigma}\,\mathbb{E}\big[|D_n - D_n^K|\big] \\
&= \frac{1}{\sigma\sqrt{n(n-1)}}\Big\|\textstyle\sum_{1\leq i\neq j\leq n}\big(u(\mathbf{X}_i,\mathbf{X}_j) - \sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_i)\phi_k(\mathbf{X}_j)\big)\Big\|_{L_1} \\
&\leq \frac{\sqrt{n(n-1)}}{\sigma}\Big\|u(\mathbf{X}_1,\mathbf{X}_2) - \textstyle\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\Big\|_{L_1} \\
&= \sigma^{-1}\sqrt{n(n-1)}\,\varepsilon_{K;1} \;.
\end{aligned}$$

Combining the bounds on $\xi_{\tau,1}, \xi_{\tau,2}$ and $\mathbb{E}[|\tilde{D}_n - \tilde{D}_n^K|]$ and choosing $\epsilon = \big(\sqrt{n(n-1)}\sigma^{-1}\delta\varepsilon_{K;1}\big)^{1/2}$, we get that

$$\xi_\tau \;\leq \delta^{-1}\epsilon + 2\sqrt{n(n-1)}\,\epsilon^{-1}\sigma^{-1}\varepsilon_{K;1} \;=\; \frac{3n^{1/4}(n-1)^{1/4}\varepsilon_{K;1}^{1/2}}{\sigma^{1/2}\delta^{1/2}} \;=:\; \varepsilon'_K \;,$$

which yields the desired bound

$$\mathbb{P}(\tilde{D}_n^K > t+\delta) - \varepsilon'_K \;\leq\; \mathbb{P}(\tilde{D}_n > t) \;\leq\; \mathbb{P}(\tilde{D}_n^K > t-\delta) + \varepsilon'_K \;.$$

■

### C.4. Proof of Lemma 39

For convenience, we denote $\mathbf{V}_i \coloneqq \phi^K(\mathbf{X}_i)$ throughout this section.

*Proof overview.* The key idea in the proof rests on Lindeberg's telescoping sum argument for central limit theorem. We follow Chatterjee (2006)'s adaptataion of the Lindeberg idea for statistics that are not asymptotically normal. As before, the difference in probability is first approximated by a difference in expectation $\mathbb{E}[h(\tilde{D}_n^K) - h(\tilde{D}_Z^K)]$ with respect to some function $h$, which introduces a further approximation error $\delta$. The next step is to note that both $\tilde{D}_n^K$ and $\tilde{D}_Z^K$ can be expressed in terms of some common function $\tilde{f}$, such that

$$\tilde{D}_n^K \;=\; \tilde{f}(\mathbf{V}_1, \ldots, \mathbf{V}_n)\;, \qquad\qquad \tilde{D}_Z^K \;=\; \tilde{f}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)\;.$$

Denoting $g = h \circ \tilde{f}$, we can then write the difference in expectation in terms of Lindeberg's telescoping sum as

$$\begin{aligned}\mathbb{E}[h(\tilde{D}_n^K) - h(\tilde{D}_Z^K)] &= \mathbb{E}[g(\mathbf{V}_1, \ldots, \mathbf{V}_1) - g(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)] \\ &= \textstyle\sum_{i=1}^n \big(\mathbb{E}[g(\mathbf{V}_1, \ldots, \mathbf{V}_{i-1}, \mathbf{V}_i, \mathbf{Z}_{i+1}, \ldots, \mathbf{Z}_n) \\ &\qquad\qquad - g(\mathbf{V}_1, \ldots, \mathbf{V}_{i-1},\ \mathbf{Z}_i,\ \mathbf{Z}_{i+1}, \ldots, \mathbf{Z}_n)]\big)\;.\end{aligned}$$

Since each summand differs only in the $i$-th argument, we can perform a second-order Taylor expansion about the $i$-th argument provided that the function $h$ such that $h$ is twice-differentiable. The second-order remainder term is further "Taylor-expanded" to an additional $\epsilon$-order for any $\epsilon \in [0, 1]$ by choosing $h''$ to be $\epsilon$-Hölder. Write $D_i$ as the differential operator with respect to the $i$-th argument and denote $\tilde{f}_i(\mathbf{x}) \coloneqq \tilde{f}(\mathbf{V}_1, \ldots, \mathbf{V}_{i-1}, \mathbf{x}, \mathbf{Z}_{i+1}, \ldots, \mathbf{Z}_n)$. Then informally speaking, the Taylor expansion argument amounts to bounding each summand as

$$\begin{aligned}\big|(\text{summand})_i\big| \;&\leq \mathbb{E}[D_i(h \circ \tilde{f}_i)(0)(\mathbf{V}_i - \mathbf{Z}_i)] + \frac{1}{2}\mathbb{E}[D_i^2(h \circ \tilde{f}_i)(0)(\mathbf{V}_i^2 - \mathbf{Z}_i^2)] \\ &\quad + \frac{1}{6}\big(\text{Hölder constant of } h''\big) \times \mathbb{E}\big[\big|D_i\tilde{f}_i(0)\mathbf{V}_i\big|^{2+\epsilon} + \big|D_i\tilde{f}_i(0)\mathbf{Z}_i\big|^{2+\epsilon}\big]\;,\end{aligned}$$

where we have used the fact that $\tilde{f}_i$ is a linear function in expressing the last quantity. The first two terms vanish because $h \circ \tilde{f}_i$ is independent of $\big(\mathbf{V}_i, \mathbf{Z}_i\big)$ and the first two moments of $\mathbf{V}_i$ and $\mathbf{Z}_i$ match. The third term is bounded carefully by noting the moment structure of $\mathbf{V}_i$ and $\mathbf{Z}_i$ to give the error term $\frac{1}{n}E_{\delta;K}$. Summing the errors over $1 \leq i \leq n$ then gives the Gaussian approximation error bound in Lemma 39.

**Proof of Lemma 39** For any $\tau \in \mathbb{R}$ and $\delta > 0$, let $h_{\tau;\delta}$ be the twice-differentiable function defined in Lemma 34 (i.e. $m = 2$), which satisfies

$$h_{\tau+\delta;\delta}(x) \;\leq\; \mathbb{I}_{\{x>\tau\}} \;\leq\; h_{\tau;\delta}(x)\;.$$

By applying the above bounds with $\tau$ set to $t - \delta$ and $t - 2\delta$, we get that

$$\begin{aligned}\mathbb{P}(\tilde{D}_n^K > t - \delta) - \mathbb{P}(\tilde{D}_Z^K > t - 2\delta) \;&= \mathbb{E}[\mathbb{I}_{\{\tilde{D}_n^K > t-\delta\}} - \mathbb{I}_{\{\tilde{D}_Z^K > t-2\delta\}}] \\ &\leq \mathbb{E}[h_{t-\delta;\delta}(\tilde{D}_n^K) - h_{t-\delta;\delta}(\tilde{D}_Z^K)]\;,\end{aligned}$$

and similarly

$$\begin{aligned}\mathbb{P}(\tilde{D}_Z^K > t+2\delta) - \mathbb{P}(\tilde{D}_n^K > t+\delta) &= \mathbb{E}[\mathbb{I}_{\{\tilde{D}_Z^K > t+2\delta\}} - \mathbb{I}_{\{\tilde{D}_n^K > t+\delta\}}] \\ &\leq \mathbb{E}[h_{t+2\delta;\delta}(\tilde{D}_Z^K) - h_{t+2\delta;\delta}(\tilde{D}_n^K)] \;.\end{aligned}$$

Therefore, we obtain that

$$\mathbb{P}(\tilde{D}_n^K > t-\delta) \;\leq\; \mathbb{P}(\tilde{D}_Z^K > t-2\delta) + E'_{\delta;K} \;, \quad \mathbb{P}(\tilde{D}_n^K > t+\delta) \;\geq\; \mathbb{P}(\tilde{D}_Z^K > t+2\delta) - E'_{\delta;K} \;, \tag{14}$$

where $E'_{\delta;K} \coloneqq \sup_{\tau\in\mathbb{R}} |\mathbb{E}[h_{\tau;\delta}(\tilde{D}_n^K) - h_{\tau;\delta}(\tilde{D}_Z^K)]|$. The next step is to bound $E'_{\delta;K}$, to which we apply Lindeberg's technique for proving central limit theorem. We denote the scaled mean as

$$\tilde{\mu} \;\coloneqq\; \frac{\mathbb{E}[\mathbf{V}_1]}{\sigma^{1/2}(n(n-1))^{1/4}} \;=\; \frac{\mathbb{E}[\mathbf{Z}_1]}{\sigma^{1/2}(n(n-1))^{1/4}} \;,$$

and define the centred and scaled versions of $\mathbf{V}_i$ and $\mathbf{Z}_i$ respectively as

$$\tilde{\mathbf{V}}_i \;\coloneqq\; \frac{\mathbf{V}_i}{\sigma^{1/2}(n(n-1))^{1/4}} - \tilde{\mu} \;, \qquad\qquad \tilde{\mathbf{Z}}_i \;\coloneqq\; \frac{\mathbf{Z}_i}{\sigma^{1/2}(n(n-1))^{1/4}} - \tilde{\mu} \;.$$

We also define the function $f : (\mathbb{R}^K)^n \to \mathbb{R}$ by

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) \coloneqq \textstyle\sum_{1\leq i\neq j\leq n} (\mathbf{v}_i + \tilde{\mu})^\top \Lambda^K (\mathbf{v}_j + \tilde{\mu}) \;, \quad \text{where we recall } \Lambda^K \coloneqq \text{diag}\{\lambda_1, \ldots, \lambda_K\} \;.$$

This allows us to express the random quantities in (14) as

$$\tilde{D}_n^K \;=\; f(\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_n) \;, \qquad\qquad \tilde{D}_Z^K \;=\; f(\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n) \;.$$

By defining the random function

$$F_i(\mathbf{v}) \;\coloneqq\; f(\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_{i-1}, \mathbf{v}, \tilde{\mathbf{Z}}_{i+1}, \ldots, \tilde{\mathbf{Z}}_n) \qquad\quad \text{for } \mathbf{v} \in \mathbb{R}^K \text{ and } 1 \leq i \leq n \;,$$

we can write $E'_{\delta;K}$ into Lindeberg's telescoping sum as

$$\begin{aligned}E'_{\delta;K} \;&=\; \sup_{\tau\in\mathbb{R}} |\mathbb{E}[h_{\tau;\delta} \circ f(\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_n) - h_{\tau;\delta} \circ f(\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n)]| \\ &=\; \sup_{\tau\in\mathbb{R}} \left| \textstyle\sum_{i=1}^n \mathbb{E}[h_{\tau;\delta}(F_i(\tilde{\mathbf{V}}_i) - h_{\tau;\delta}(F_i(\tilde{\mathbf{Z}}_i))] \right| \\ &\leq\; \sup_{\tau\in\mathbb{R}} \textstyle\sum_{i=1}^n |\mathbb{E}[h_{\tau;\delta} \circ F_i(\tilde{\mathbf{V}}_i) - h_{\tau;\delta} \circ F_i(\tilde{\mathbf{Z}}_i)]| \;.\end{aligned}$$

Since $h_{\tau;\delta} \circ f$ is twice-differentiable, by a second-order Taylor expansion around $\mathbf{0} \in \mathbb{R}^K$, there exists random values $\theta_V, \theta_Z \in (0,1)$ almost surely such that

$$\begin{aligned}h_{\tau;\delta} \circ F_i(\tilde{\mathbf{V}}_i) \;&=\; \frac{\partial h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{0}} \tilde{\mathbf{V}}_i + \frac{1}{2} \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_V \tilde{\mathbf{V}}_i} \tilde{\mathbf{V}}_i^{\otimes 2} \;, \\ h_{\tau;\delta} \circ F_i(\tilde{\mathbf{Z}}_i) \;&=\; \frac{\partial h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{0}} \tilde{\mathbf{Z}}_i + \frac{1}{2} \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_Z \tilde{\mathbf{Z}}_i} \tilde{\mathbf{Z}}_i^{\otimes 2} \;.\end{aligned}$$

Substituting this into the sum above gives

$$E'_{\delta;K} \;\leq\; \sup_{\tau \in \mathbb{R}} \Big( \sum_{i=1}^n \Big| \mathbb{E}\Big[ \frac{\partial h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\mathbf{0}} (\tilde{\mathbf{V}}_i - \tilde{\mathbf{Z}}_i) \Big] \Big| + \frac{1}{2} \sum_{i=1}^n \Big| \mathbb{E}\Big[ \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\theta_V \tilde{\mathbf{V}}_i} \tilde{\mathbf{V}}_i^{\otimes 2} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\theta_Z \tilde{\mathbf{Z}}_i} \tilde{\mathbf{Z}}_i^{\otimes 2} \Big] \Big| \Big) .$$

The first sum vanishes because the only randomness of the derivative comes from $F_i$, who is independent of $(\tilde{\mathbf{V}}_i, \tilde{\mathbf{Z}}_i)$, and the mean of $\tilde{\mathbf{V}}_i$ and $\tilde{\mathbf{Z}}_i$ match. To handle the second sum, we make use of independence again and the fact that the second moment of $\tilde{\mathbf{V}}_i$ and $\tilde{\mathbf{Z}}_i$ also match: By subtracting and adding the term

$$\mathbb{E}\Big[ \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\mathbf{0}} (\tilde{\mathbf{V}}_i)^{\otimes 2} \Big] \;=\; \mathbb{E}\Big[ \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\mathbf{0}} (\tilde{\mathbf{Z}}_i)^{\otimes 2} \Big] ,$$

we can apply a triangle inequality to get that

$$E'_{\delta;K} \;\leq\; \frac{1}{2} \sup_{\tau \in \mathbb{R}} \Big( \sum_{i=1}^n \Big| \mathbb{E}\Big[ \Big( \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\theta_V \tilde{\mathbf{V}}_i} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\mathbf{0}} \Big) \tilde{\mathbf{V}}_i^{\otimes 2} \Big] \Big| + \sum_{i=1}^n \Big| \mathbb{E}\Big[ \Big( \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\theta_Z \tilde{\mathbf{Z}}_i} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\mathbf{0}} \Big) \tilde{\mathbf{Z}}_i^{\otimes 2} \Big] \Big| \Big) . \quad (15)$$

The final step is to bound the two sums by exploiting the derivative structure of $h_{\tau;\delta}$ and $F_i$. Note that $F_i$ is a linear function: its first derivative is given by

$$\partial F_i(\mathbf{x}) \;=\; 2 \sum_{1 \leq j < i} \Lambda^K \tilde{\mathbf{V}}_j + 2 \sum_{i < j \leq n} \Lambda^K \tilde{\mathbf{Z}}_j + 2(n-1) \Lambda^K \tilde{\mu} \;\in\; \mathbb{R}^K ,$$

which is independent of $\mathbf{x}$, while its higher derivatives vanish. By a second-order chain rule, this implies that almost surely

$$\begin{aligned} &\Big| \Big( \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\theta_V \tilde{\mathbf{V}}_i} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\mathbf{0}} \Big) \tilde{\mathbf{V}}_i^{\otimes 2} \Big| \\ &= \Big| \big( \partial^2 h_{\tau;\delta}\big(F_i(\theta_V \tilde{\mathbf{V}}_i)\big) - \partial^2 h_{\tau;\delta}\big(F_i(\mathbf{0})\big) \big) \big( \partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i \big)^2 \Big| \\ &\leq \big| \partial^2 h_{\tau;\delta}\big(F_i(\theta_V \tilde{\mathbf{V}}_i)\big) - \partial^2 h_{\tau;\delta}\big(F_i(\mathbf{0})\big) \big| \, \big| \partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i \big|^2 . \end{aligned}$$

For $\nu \in (2, 3]$, by the Hölder property of $\partial^2 h_{\tau;\delta}$ from Lemma 34, we get that almost surely,

$$\begin{aligned} \big| \partial^2 h_{\tau;\delta}(F_i(\theta_V \tilde{\mathbf{V}}_i)) - \partial^2 h_{\tau;\delta}(F_i(\mathbf{0})) \big| \;&\leq 18 \times 3^{\nu-2} \delta^{-\nu} |F_i(\theta_V \tilde{\mathbf{V}}_i) - F_i(\mathbf{0})|^{\nu-2} \\ &= 18 \times 3^{\nu-2} \delta^{-\nu} |\partial F_i(\mathbf{0})^\top (\theta_V \tilde{\mathbf{V}}_i)|^{\nu-2} \\ &\leq 54 \delta^{-\nu} |\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^{\nu-2} . \end{aligned}$$

In the last inequality, we have used that $\theta_V$ takes value in $[0, 1]$. Combining the results, we get that each summand in the first sum in (15) can be bounded as

$$\Big| \mathbb{E}\Big[ \Big( \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\theta_V \tilde{\mathbf{V}}_i} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\mathbf{0}} \Big) \tilde{\mathbf{V}}_i^{\otimes 2} \Big] \Big| \;\leq 54 \delta^{-\nu} \mathbb{E}\big[ |\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^{\nu} \big] .$$

The exact same argument applies to the summands of the second sum to give

$$\Big| \mathbb{E}\Big[ \Big( \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\theta_Z \tilde{\mathbf{Z}}_i} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\mathbf{0}} \Big) \tilde{\mathbf{Z}}_i^{\otimes 2} \Big] \Big| \;\leq 54 \delta^{-\nu} \mathbb{E}\big[ |\partial F_i(\mathbf{0})^\top \tilde{\mathbf{Z}}_i|^{\nu} \big] ,$$

so a substitution back into (15) gives

$$E'_{\delta;K} \leq 27\delta^{-\nu} \sum_{i=1}^{n} \Big( \mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu\big] + \mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu\big] \Big) .$$

We defer to Lemma 42 to show that there exists an absolute constant $C' > 0$ such that the moment terms can be bounded as

$$\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu\big] + \mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{Z}}_i|^\nu\big] \leq \frac{C'}{n^{\nu/2}} \Big( \frac{(M_{\text{full};\nu})^\nu + \varepsilon^\nu_{K;\nu}}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu + \varepsilon^\nu_{K;\nu}}{(n-1)^{-\nu/2}\,\sigma^\nu} \Big) . \tag{16}$$

Combining with (14) and defining $E_{\delta;K}$ to be the upper bound for $E'_{\delta;K}$, we get that

$$\mathbb{P}(\tilde{D}^K_n > t - \delta) \leq \mathbb{P}(\tilde{D}^K_Z > t - 2\delta) + E_{\delta;K} , \quad \mathbb{P}(\tilde{D}^K_n > t + \delta) \geq \mathbb{P}(\tilde{D}^K_Z > t + 2\delta) - E_{\delta;K} ,$$

where we have made the $K$-dependence explicit and define, for $C := 27C'$,

$$E_{\delta;K} := \frac{C}{\delta^\nu n^{\nu/2-1}} \Big( \frac{(M_{\text{full};\nu})^\nu + \varepsilon^\nu_{K;\nu}}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu + \varepsilon^\nu_{K;\nu}}{(n-1)^{-\nu/2}\,\sigma^\nu} \Big) .$$

■

**Lemma 42** (16) *holds.*

**Proof of Lemma 42** We seek to bound $\mathbb{E}[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu] + \mathbb{E}[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{Z}}_i|^\nu]$ for $\nu \in (2, 3]$ and

$$\partial F_i(\mathbf{0}) = 2 \sum_{1 \leq j < i} \Lambda^K \tilde{\mathbf{V}}_j + 2 \sum_{i < j \leq n} \Lambda^K \tilde{\mathbf{Z}}_j + 2(n-1) \Lambda^K \tilde{\mu} \in \mathbb{R}^K .$$

We first focus on bounding the first expectation. By convexity of the function $x \mapsto |x|^\nu$, we can apply Jensen's inequality to bound

$$\begin{aligned}
\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu\big] &= \mathbb{E}\Big[\Big|2 \sum_{j<i} \tilde{\mathbf{V}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i + 2 \sum_{j>i} \tilde{\mathbf{Z}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i + 2(n-1)\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{V}}_i\Big|^\nu\Big] \\
&\leq \tfrac{1}{3}\mathbb{E}\big[|6 \sum_{j<i} \tilde{\mathbf{V}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i|^\nu\big] + \tfrac{1}{3}\mathbb{E}\big[|6 \sum_{j>i} \tilde{\mathbf{Z}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i|^\nu\big] + \tfrac{1}{3}\mathbb{E}\big[|6(n-1)\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu\big] \\
&\leq 72\big(\mathbb{E}\big[|\sum_{j<i} \tilde{\mathbf{V}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i|^\nu\big] + \mathbb{E}\big[|\sum_{j>i} \tilde{\mathbf{Z}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i|^\nu\big] + \mathbb{E}\big[|(n-1)\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu\big]\big) ,
\end{aligned}$$

where we have noted that $\nu \leq 3$. Since $\tilde{\mathbf{V}}_i$'s are i.i.d., $\tilde{\mathbf{Z}}_i$'s are i.i.d. and all variables involved are zero-mean, $(\tilde{\mathbf{V}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i)_{j=1}^{i-1}$ forms a martingale difference sequence with respect to the filtration $\sigma(\tilde{\mathbf{V}}_i, \tilde{\mathbf{V}}_1), \ldots, \sigma(\tilde{\mathbf{V}}_i, \tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_{i-1})$, and so is $(\tilde{\mathbf{Z}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i)_{j=i+1}^{n}$ with respect to the filtration $\sigma(\tilde{\mathbf{V}}_i, \tilde{\mathbf{Z}}_{i+1}), \ldots, \sigma(\tilde{\mathbf{V}}_i, \tilde{\mathbf{Z}}_{i+1}, \ldots, \tilde{\mathbf{Z}}_n)$. This allows the above two moments of sums to be bounded via the martigale moment inequality from Lemma 26: There exists an absolute constant $C_0 > 0$ such that

$$\begin{aligned}
&\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu\big] \\
&\quad\leq C_0 \Big( (i-1)^{\nu/2-1} \sum_{j=1}^{i-1} \mathbb{E}[|\tilde{\mathbf{V}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i|^\nu] + (n-i)^{\nu/2-1} \sum_{j=i+1}^{n} \mathbb{E}[|\tilde{\mathbf{Z}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i|^\nu] \\
&\qquad\quad + (n-1)^\nu\, \mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] \Big) \\
&\quad\leq C_0 (n-1)^{\nu/2} \Big( \mathbb{E}[|\tilde{\mathbf{V}}_1^\top \Lambda^K \tilde{\mathbf{V}}_2|^\nu] + \mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] + (n-1)^{\nu/2} \mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] \Big) .
\end{aligned}$$

By the exact same argument, the other expectation we want to bound can also be controlled as

$$\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{Z}}_i|^\nu\big] \leq C_0(n-1)^{\nu/2}\Big(\mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{Z}}_2|^\nu] + \mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] + (n-1)^{\nu/2}\mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{Z}}_1|^\nu]\Big) .$$

Finally, we relate these moments terms to moments of $u(\mathbf{X}_1, \mathbf{X}_2)$, up to error terms that vanish as $K \to \infty$: Denoting $\mu_k := \mathbb{E}[\phi_k(\mathbf{X}_1)]$, we have that by Lemma 31,

$$\mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] = \frac{1}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}}\mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\mu_k\Big|^\nu\Big] \leq \frac{4((M_{\text{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu)}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} ,$$

and for some absolute constant $C_1 > 0$,

$$\begin{aligned}\mathbb{E}[|\tilde{\mathbf{V}}_1^\top \Lambda^K \tilde{\mathbf{V}}_2|^\nu] &= \frac{1}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}}\mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \\ &\leq \frac{4C_1(M_{\text{full};\nu})^\nu - \frac{1}{2}(M_{\text{cond};\nu})^\nu + (4C_1+2)\varepsilon_{K;\nu}^\nu}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \leq \frac{4C_1(M_{\text{full};\nu})^\nu + (4C_1+2)\varepsilon_{K;\nu}^\nu}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} .\end{aligned}$$

For the moment terms involving the Gaussians $\tilde{\mathbf{Z}}_1$ and $\tilde{\mathbf{Z}}_2$, we apply Lemma 32 to show that

$$\begin{aligned}\mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{Z}}_1|^\nu] &= \frac{\mathbb{E}[|(\mathbb{E}[\mathbf{V}_1])^\top \Lambda^K \mathbf{Z}_1|^\nu]}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \leq \frac{7(\sigma_{\text{cond}}^\nu + 8\varepsilon_{K;2}^\nu)}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \leq \frac{7((M_{\text{cond};\nu})^\nu + 8\varepsilon_{K;\nu}^\nu)}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} ,\\ \mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{Z}}_2|^\nu] &= \frac{\mathbb{E}[|\mathbf{Z}_1^\top \Lambda^K \mathbf{Z}_2|^\nu]}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \leq \frac{6(\sigma_{\text{full}}^\nu + \varepsilon_{K;2}^\nu)}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \leq \frac{6((M_{\text{full};\nu})^\nu + \varepsilon_{K;\nu}^\nu)}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} .\end{aligned}$$

In the last inequalities for both bounds, we have noted that $L_2$ norm is dominated by $L_\nu$ norm since $\nu > 2$. Meanwhile by Lemma 32 again, there exists some absolute constant $C_2 > 0$ such that

$$\mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] = \frac{\mathbb{E}[|(\mathbf{V}_1 - \mathbb{E}[\mathbf{V}_1])^\top \Lambda^K \mathbf{Z}_1|^\nu]}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \leq \frac{8C_2(M_{\text{full};\nu})^\nu + (8C_2+4)\varepsilon_{K;\nu}^\nu}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} .$$

Substituting the five moment bounds into the earlier bounds on $\mathbb{E}[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu]$ and $\mathbb{E}[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{Z}}_i|^\nu]$ and combining the constant terms, we get that there exists an absolute constant $C > 0$ such that

$$\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu\big] + \mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{Z}}_i|^\nu\big] \leq \frac{C}{n^{\nu/2}}\Big(\frac{(M_{\text{full};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu}\Big) .$$

■

### C.5. Proof of Lemma 40

*Proof overview.* For convenience, we write

$$U_0 := \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}(\eta_i^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K + \frac{2}{n}\sum_{i=1}^n(\mu^K)^\top\Lambda^K(\Sigma^K)^{1/2}\eta_i^K ,$$

so that

$$\tilde{D}_Z^K = \frac{\sqrt{n(n-1)}}{\sigma}U_0 + \frac{\sqrt{n(n-1)}}{\sigma}(\mu^K)^\top\Lambda^K\mu^K , \quad \tilde{U}_n^K = \frac{\sqrt{n(n-1)}}{\sigma}U_0 + \frac{\sqrt{n(n-1)}}{\sigma}D .$$

To approximate the distribution of $\tilde{D}_Z^K$ by that of $\tilde{U}_n^K$, the proof boils down to replacing $(\mu^K)^\top \Lambda^K \mu^K$ by $D$. We use a Markov-type argument so that we obtain an error term that is separate from the distribution terms.

**Proof of Lemma 40** Recall that Lemma 35 allows us to approximate the distribution of a sum of two random variables by a single one provided that the other is negligible. Writing

$$\tilde{D}_Z^K \;=\; \tilde{U}_n^K + (\tilde{D}_Z^K - \tilde{U}_n^K) \;=\; \tilde{U}_n^K + \frac{\sqrt{n(n-1)}}{\sigma}\big((\mu^K)^\top \Lambda^K \mu^K - D\big) \;,$$

we can apply Lemma 35 to obtain that for any $a, b \in \mathbb{R}$ and $\epsilon > 0$,

$$\mathbb{P}(a \leq \tilde{D}_Z^K \leq b) \;\leq \mathbb{P}\big(a - \epsilon \leq \tilde{U}_n^K \leq b + \epsilon\big) + \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma}\big|(\mu^K)^\top \Lambda^K \mu^K - D\big| \geq \epsilon\Big) \;,$$
$$\mathbb{P}(a \leq \tilde{D}_Z^K \leq b) \;\geq \mathbb{P}(a + \epsilon \leq \tilde{U}_n^K \leq b - \epsilon) - \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma}\big|(\mu^K)^\top \Lambda^K \mu^K - D\big| \geq \epsilon\Big) \;.$$

Note that $|(\mu^K)^\top \Lambda^K \mu^K - D|$ is deterministic. By a Markov inequality and the bound from Lemma 31, we get that

$$\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma}\big|(\mu^K)^\top \Lambda^K \mu^K - D\big| \geq \epsilon\Big) \;\leq\; \frac{\sqrt{n(n-1)}}{\epsilon\sigma}\mathbb{E}\Big[\big|(\mu^K)^\top \Lambda^K \mu^K - D\big|\Big]$$
$$=\; \frac{|\sum_{k=1}^K \lambda_K \mu_k^2 - D|}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \;\leq\; \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \;.$$

Combining the two results gives the desired bounds. ■

## C.6. Proof of Lemma 41

*Proof overview.* The key ingredient of the proof is Theorem 8 of Carbery and Wright (2001), which gives an anti-concentration bound for the distribution of a polynomial of Gaussians in terms of its variance. In Lemma 36, we have rewritten the result in the special case of a degree-two polynomial, which allows us to control the distribution of $\tilde{U}_n^K$ in terms of its variance.

We introduce some matrix shorthands: For any $m \in \mathbb{N}$, denote $O_m$ as the zero matrix in $\mathbb{R}^{m\times m}$, $J_m$ as the all-one matrix in $\mathbb{R}^{m\times m}$ and $I_m$ as the identity matrix in $\mathbb{R}^{m\times m}$. Define the $nK \times nK$ matrix $M$ as

$$M \;\coloneqq\; \begin{pmatrix} O_K & \Lambda^K & \dots & \Lambda^K \\ \Lambda^K & O_K & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Lambda^K \\ \Lambda^K & \dots & \Lambda^K & O_K \end{pmatrix} \;=\; \Lambda^K \otimes (J_n - I_n) \;,$$

as well as

$$\mu \coloneqq \big((\mu^K)^\top, \dots, (\mu^K)^\top\big)^\top \in \mathbb{R}^{nK} \;, \quad \Sigma \coloneqq \Sigma^K \otimes I_n \in \mathbb{R}^{nK\times nK} \;, \quad \Lambda \coloneqq \Lambda^K \otimes I_n \in \mathbb{R}^{nK\times nK} \;.$$

We also consider the concatenated $nK$-dimensional standard Gaussian vector

$$\eta \;\coloneqq\; \big((\eta_1^K)^\top, \dots, (\eta_n^K)^\top\big)^\top \;.$$

**Proof of Lemma 41** The goal is to bound the distribution function between $a \leq b \in \mathbb{R}$ of

$$\begin{aligned}\tilde{U}_n^K \;=\; \frac{\sqrt{n(n-1)}}{\sigma} U_n^K \;&=\; \frac{1}{\sigma\sqrt{n(n-1)}} \sum\nolimits_{1 \leq i \neq j \leq n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K \\ &\quad + \frac{2\sqrt{n-1}}{\sigma\sqrt{n}} \sum\nolimits_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K + \frac{\sqrt{n(n-1)}}{\sigma} D \\ &=\; \frac{1}{\sigma\sqrt{n(n-1)}} \eta^\top \Sigma^{1/2} M \Sigma^{1/2} \eta + \frac{2\sqrt{n-1}}{\sigma\sqrt{n}} \mu^\top \Lambda \Sigma^{1/2} \eta + \frac{\sqrt{n(n-1)}}{\sigma} D \;.\end{aligned}$$

For convenience, define

$$Q_1 \;\coloneqq\; \eta^\top \Sigma^{1/2} M \Sigma^{1/2} \eta \;, \qquad Q_2 \;\coloneqq\; \mu^\top \Lambda \Sigma^{1/2} \eta \;, \qquad \tilde{U}_0 \;\coloneqq\; \frac{1}{\sigma\sqrt{n(n-1)}} Q_1 + \frac{2\sqrt{n-1}}{\sigma\sqrt{n}} Q_2 \;.$$

Denote $\alpha \coloneqq \frac{b-a}{2}$ and $\beta \coloneqq \frac{a+b}{2}$. Rewriting the probability in terms of $\tilde{U}_0$, $\alpha$ and $\beta$, we get that

$$\begin{aligned}\mathbb{P}(a \leq \tilde{U}_n^K \leq b) \;&=\; \mathbb{P}\Big( (\beta - \alpha) \;\leq \tilde{U}_0 + \frac{\sqrt{n(n-1)}}{\sigma} D \leq \; (\beta + \alpha) \Big) \\ &=\; \mathbb{P}\Big( \Big| \tilde{U}_0 + \frac{\sqrt{n(n-1)}}{\sigma} D - \beta \Big| \;\leq\; \alpha \Big) \;.\end{aligned}$$

Since $\tilde{U}_0 + \frac{\sqrt{n(n-1)}}{\sigma} D - \beta$ is a degree-two polynomial of $\eta$, we can apply Lemma 36 to bound the above probability: For an absolute constant $C'$, we have

$$\mathbb{P}(a \leq \tilde{U}_n^K \leq b) \;\leq C' \alpha^{1/2} \big( \mathrm{Var}[\tilde{U}_0] \big)^{-1/4} \;, \tag{17}$$

where the variance term can be expanded as

$$\mathrm{Var}\big[\tilde{U}_0\big] \;=\; \frac{1}{n(n-1)\sigma^2} \mathrm{Var}[Q_1] + \frac{4(n-1)}{n\sigma^2} \mathrm{Var}[Q_2] + \frac{4}{n\,\sigma^2} \,\mathrm{Cov}[Q_1, Q_2] \;.$$

We now provide bound the individual terms in the variance. By noting that each summand in $Q_1$ is zero-mean when $i \neq j$ and that each summand in $Q_2$ is zero-mean, the covariance term can be computed as

$$\begin{aligned}\mathrm{Cov}[Q_1, Q_2] \;&=\; \sum\nolimits_{1 \leq i \neq j \leq n} \sum\nolimits_{l=1}^n \mathbb{E}\Big[ (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K \times (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_l^K \Big] \\ &=\; \tfrac{1}{2} \mathbb{E}\Big[ (\eta_1^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_1^K \times (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_1^K \Big] \;.\end{aligned}$$

Denote $\xi_k$ as the $k$-th coordinate of $\eta_1^K$. Then the above expectation is taken over a linear combination of terms of the form $\xi_{k_1} \xi_{k_2} \xi_{k_3}$. If any of $k_1, k_2, k_3$ is distinct from the other two indices, the expectation is zero; if $k_1 = k_2 = k_3$, the expectation is again zero by property of a standard Gauassian. Therefore, we have

$$\mathrm{Cov}[Q_1, Q_2] \;=\; 0 \;.$$

On the other hand, the first variance can be computed by using the moment formula for a quadratic form of Gaussian from Lemma 27 and the cyclic property of trace:

$$\begin{aligned}
\mathrm{Var}[Q_1] &= 2\mathrm{Tr}\big((\Sigma^{1/2}M\Sigma^{1/2})^2\big) \;=\; 2\mathrm{Tr}\big((\Sigma M)^2\big) \\
&= 2\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2 \otimes (J_n - I_n)^2\big) \\
&= 2\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2 \otimes J_n^2\big) - 4\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2 \otimes J_n\big) + 2\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2 \otimes I_n\big) \\
&= \big(2n^2 - 4n + 2n\big)\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2\big) \\
&= 2n(n-1)\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big) \\
&\geq 2n(n-1)(\sigma_{\mathrm{full}} - \varepsilon_{K;2})^2 \;.
\end{aligned}$$

In the last inequality, we have used the bound from Lemma 32 on $\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big)$. The second variance is on a Gaussian random variable and can be bounded by Lemma 32 again as

$$\mathrm{Var}[Q_2] \;= \mu^\top \Lambda\Sigma\Lambda\mu \;=\; n(\mu^K)^\top \Lambda^K\Sigma^K\Lambda^K\mu^K \;\geq\; n(\sigma^2_{\mathrm{cond}} - 2\sigma_{\mathrm{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2}) \;.$$

This implies that

$$\mathrm{Var}\big[\tilde{U}_0\big] \;\geq\; \frac{2}{\sigma^2}(\sigma_{\mathrm{full}} - \varepsilon_{K;2})^2 + \frac{4(n-1)}{\sigma^2}(\sigma^2_{\mathrm{cond}} - 2\sigma_{\mathrm{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2}) \;.$$

Substituting this into (17) and redefining the constants, we get that there exists an absolute constant $C$ such that

$$\mathbb{P}(a \leq \tilde{U}^K_n \leq b) \;\leq C(b-a)^{1/2}\Big(\frac{1}{\sigma^2}(\sigma_{\mathrm{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma^2}(\sigma^2_{\mathrm{cond}} - 2\sigma_{\mathrm{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\Big)^{-1/4} \;.$$

■

## Appendix D. Proofs for the remaining results in Section 3

### D.1. Proofs for variants and corollaries of the main result

The upper bound in Proposition 6 is a concentration inequality and is obtained by a standard argument via Chebyshev's inequality. The lower bound is a combination of the anti-concentration bound for a Gaussian quadratic form from Lemma 41 and Theorem 2.

**Proof of Proposition 6** Denote $\tilde{U}^K_n := \frac{\sqrt{n(n-1)}U^K_n}{\sigma_{\max}}$. In Lemma 41, we have shown that for any $a, b \in \mathbb{R}$ with $a \leq b$, there exists some absolute constant $C'$ such that

$$\mathbb{P}(a \leq \tilde{U}^K_n \leq b) \leq C'(b-a)^{1/2}\Big(\frac{1}{\sigma^2_{\max}}(\sigma_{\mathrm{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma^2_{\max}}(\sigma^2_{\mathrm{cond}} - 2\sigma_{\mathrm{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\Big)^{-1/4} \;.$$

Take $K \to \infty$ and using Assumption 2 for $\nu \geq 2$, we get that $\varepsilon_{K;2} \to 0$. For a fixed $\epsilon > 0$, set $a = \frac{\sqrt{n(n-1)}}{\sigma_{\max}}D - \epsilon$ and $b = \frac{\sqrt{n(n-1)}}{\sigma_{\max}}D + \epsilon$, we get that

$$\lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}|U^K_n - D| \leq \epsilon\Big) \;\leq \sqrt{2}\,C'\,\epsilon^{1/2}\Big(\frac{\sigma^2_{\mathrm{full}}}{\sigma^2_{\max}} + \frac{(n-1)\sigma^2_{\mathrm{cond}}}{\sigma^2_{\max}}\Big)^{-1/4} \;\leq\; \sqrt{2}\,C'\,\epsilon^{1/2} \;.$$

Now by Theorem 2, there exists an absolute constant $C''$ such that

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\max}} D_n > t \Big) - \lim_{K\to\infty} \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\max}} U_n^K > t \Big) \right| \leq C'' \, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{M_{\max;\nu}}{\sigma_{\max}} \Big)^{\frac{\nu}{2\nu+1}} .$$

By a triangle inequality, we get that

$$\begin{aligned} \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\max}} |D_n - D| > \epsilon \Big) &\geq \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\max}} |U_n^K - D| > \epsilon \Big) - 2C'' \, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{M_{\max;\nu}}{\sigma_{\max}} \Big)^{\frac{\nu}{2\nu+1}} \\ &\geq 1 - \sqrt{2}\, C' \epsilon^{1/2} - 2C'' \, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{M_{\max;\nu}}{\sigma_{\max}} \Big)^{\frac{\nu}{2\nu+1}} . \end{aligned}$$

By replacing $\epsilon$ with $\frac{\sqrt{n(n-1)}}{\sigma_{\max}}\epsilon$ and redefining constants, we get the desired lower bound that there exists absolute constants $C_1, C_2 > 0$ such that

$$\mathbb{P}(|D_n - D| > \epsilon) \geq 1 - C_1 \Big( \frac{\sqrt{n(n-1)}}{\sigma_{\max}} \Big)^{1/2} \epsilon^{1/2} - C_2 \, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{M_{\max;\nu}}{\sigma_{\max}} \Big)^{\frac{\nu}{2\nu+1}} .$$

For the upper bound, we apply a Chebyshev inequality directly to $D_n$ and bound the variance by Lemma 30: There exists some absolute constant $C_3' > 0$ such that

$$\begin{aligned} \mathbb{P}(|D_n - D| > \epsilon) \leq \epsilon^{-2} \mathrm{Var}[D_n] &\leq C_3' \epsilon^{-2} \Big( \frac{\sigma^2_{\text{cond}}}{n^{-1}(n-1)^2} + \frac{\sigma^2_{\text{full}}}{(n-1)^2} \Big) \\ &\leq C_3' \epsilon^{-2} \Big( \frac{\sigma_{\max}}{n-1} \Big)^2 \leq C_3 \epsilon^{-2} \Big( \frac{\sigma_{\max}}{\sqrt{n(n-1)}} \Big)^2 . \end{aligned}$$

In the last inequality, we have noted that $\frac{1}{n-1} \leq \frac{2}{n}$ for $n \geq 2$ and defined $C_3 = 2C_3'$. This finishes the proof. ■

Theorem 2 provides an approximation of the distribution of $D_n$ by that of a Gaussian quadratic form. Proposition 9 combines Theorem 2 with a Markov argument, which makes a further approximation of the Gaussian quadratic form by a weighted sum of chi-squares $U_n^K$. The approximation error introduced vanishes as $n, d$ grow provided that $\rho_d = \omega(n^{1/2})$, i.e. $n^{-1/2}\sigma_{\text{full}} = \omega(\sigma_{\text{cond}})$.

**Proof of Proposition 9** We first seek to compare $W_n^K$ to the distribution of

$$U_n^K = \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K + \frac{2}{n} \sum_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K + D ,$$

where $\{\eta_i^K\}_{i=1}^n$ are i.i.d. standard Gaussian vectors in $\mathbb{R}^K$. The first step is to write

$$U_n^K = \frac{\sqrt{n-1}}{\sqrt{n}} W_0 + D + \Big( 1 - \frac{\sqrt{n-1}}{\sqrt{n}} \Big) W_0 + W_1 + W_2 ,$$

where we have defined the zero-mean random variables

$$\begin{aligned} W_0 &\coloneqq \frac{1}{n(n-1)} \Big( \sum_{i,j=1}^n (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K - n\mathrm{Tr}(\Sigma^K \Lambda^K) \Big) , \\ W_1 &\coloneqq \frac{1}{n(n-1)} \Big( \sum_{i=1}^n (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_i^K - n\mathrm{Tr}(\Sigma^K \Lambda^K) \Big) , \\ W_2 &\coloneqq \frac{2}{n} \sum_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K . \end{aligned}$$

Fix $\epsilon_0, \epsilon_1, \epsilon_2 > 0$. We first use the bound from Lemma 35: For any $a, b \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{P}\Big(a \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}\Big(\frac{\sqrt{n-1}}{\sqrt{n}}W_0 + D\Big) \leq b\Big) \\
&\leq \mathbb{P}\Big(a - \epsilon_0 - \epsilon_1 - \epsilon_2 \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \leq b + \epsilon_0 + \epsilon_1 + \epsilon_2\Big) \\
&\quad + \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}\Big(1 - \frac{\sqrt{n-1}}{\sqrt{n}}\Big)|W_0| \geq \epsilon_0\Big) + \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_1| \geq \epsilon_1\Big) \\
&\quad + \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_2| \geq \epsilon_2\Big)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\Big(a \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}\Big(\frac{\sqrt{n-1}}{\sqrt{n}}W_0 + D\Big) \leq b\Big) \\
&\geq \mathbb{P}\Big(a + \epsilon_0 + \epsilon_1 + \epsilon_2 \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \leq b - \epsilon_0 - \epsilon_1 - \epsilon_2\Big) \\
&\quad - \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}\Big(1 - \frac{\sqrt{n-1}}{\sqrt{n}}\Big)|W_0| \geq \epsilon_0\Big) - \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_1| \geq \epsilon_1\Big) \\
&\quad - \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_2| \geq \epsilon_2\Big) .
\end{aligned}$$

We now bound the error terms. By the Chebyshev's inequality, the variance formula of a quadratic form of Gaussians from Lemma 27 and the bound from Lemma 32, we get that

$$\begin{aligned}
\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_1| \geq \epsilon_1\Big) \;\leq\; \epsilon_1^{-2}\text{Var}\Big[\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}W_1\Big] \;&=\; \frac{2}{\epsilon_1^2(n-1)\sigma_{\text{full}}^2}\text{Tr}\big((\Lambda^K\Sigma^K)^2\big) \\
&\leq \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{\epsilon_1^2(n-1)\sigma_{\text{full}}^2} .
\end{aligned}$$

Similarly, by the Chebyshev's inequality, the variance formula of a Gaussian and the bound from Lemma 32, we get that

$$\begin{aligned}
\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_2| \geq \epsilon\Big) \;\leq\; \epsilon_2^{-2}\text{Var}\Big[\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}W_2\Big] \;&=\; \frac{4(n-1)}{\epsilon_2^2\sigma_{\text{full}}^2}\mathbb{E}\big[(\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K\mu^K\big] \\
&\leq \frac{4(n-1)(\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2}{\epsilon_2^2\sigma_{\text{full}}^2} .
\end{aligned}$$

By Lemma 33, we can replace $W_0$ by using the following equality in distribution:

$$\begin{aligned}
\frac{\sqrt{n-1}}{\sqrt{n}}W_0 \;&=\; \frac{1}{n^{3/2}(n-1)^{1/2}}\Big(\textstyle\sum_{i,j=1}^n (\eta_i^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K - n\text{Tr}(\Sigma^K\Lambda^K)\Big) \\
&\overset{d}{=} W_n^K - D .
\end{aligned}$$

Finally, using a Chebyshev's inequality together with the moment bound in Lemma 33, we get that

$$\begin{aligned}
\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}\Big(1 - \frac{\sqrt{n-1}}{\sqrt{n}}\Big)|W_0| \geq \epsilon_0\Big) \;&\leq\; \frac{n(n-1)}{\epsilon_0^2\sigma_{\text{full}}^2}\Big(1 - \frac{\sqrt{n-1}}{\sqrt{n}}\Big)^2\text{Var}\big[W_0\big] \\
&= \frac{n^2}{\epsilon_0^2\sigma_{\text{full}}^2}\Big(1 - \frac{\sqrt{n-1}}{\sqrt{n}}\Big)^2\text{Var}\big[W_n^K\big] \\
&\leq \frac{2n(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{\epsilon_0^2(n-1)\sigma_{\text{full}}^2}\Big(1 - \frac{\sqrt{n-1}}{\sqrt{n}}\Big)^2 \\
&\leq \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{\epsilon_0^2(n-1)\sigma_{\text{full}}^2} .
\end{aligned}$$

In the last inequality, we have noted that $\sqrt{n} - \sqrt{n-1} \leq 1$. Combining the above bounds, we get that

$$\begin{aligned}
\mathbb{P}\Big( a \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K \leq b \Big) \;&\leq \mathbb{P}\Big( a - \epsilon_0 - \epsilon_1 - \epsilon_2 \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \leq b + \epsilon_0 + \epsilon_1 + \epsilon_2 \Big) \\
&\quad + \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{(n-1)\sigma_{\text{full}}^2} \big( \epsilon_0^{-2} + \epsilon_1^{-2} \big) + \frac{4(n-1)(\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2}{\epsilon_2^2 \sigma_{\text{full}}^2} \;, \\
\mathbb{P}\Big( a \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K \leq b \Big) \;&\geq \mathbb{P}\Big( a + \epsilon_0 + \epsilon_1 + \epsilon_2 \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \leq b - \epsilon_0 - \epsilon_1 - \epsilon_2 \Big) \\
&\quad - \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{(n-1)\sigma_{\text{full}}^2} \big( \epsilon_0^{-2} + \epsilon_1^{-2} \big) - \frac{4(n-1)(\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2}{\epsilon_2^2 \sigma_{\text{full}}^2} \;.
\end{aligned}$$

Taking $b \to \infty$ and $a \to t$ from the right, we get that

$$\begin{aligned}
&\Big| \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K > t \Big) - \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K > t \Big) \Big| \\
&\leq \max \Big\{ \mathbb{P}\Big( t - \epsilon_0 - \epsilon_1 - \epsilon_2 \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \leq t \Big) \;,\; \mathbb{P}\Big( t \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \leq \epsilon_0 + \epsilon_1 + \epsilon_2 \Big) \Big\} \\
&\quad + \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{(n-1)\sigma_{\text{full}}^2} \big( \epsilon_0^{-2} + \epsilon_1^{-2} \big) + \frac{4(n-1)(\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2}{\epsilon_2^2 \sigma_{\text{full}}^2} \;.
\end{aligned}$$

This allows us to follow a similar argument to the proof of Theorem 2 to approximate $W_n^K$ by $U_n^K$. To bound the maxima, we apply Lemma 41 with $\sigma = \sigma_{\text{full}}$: There exists some absolute constant $C'$ such that for any $a \leq b \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}\Big( a \leq \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \leq b \Big) \\
&\qquad \leq C'(b-a)^{1/2} \Big( \frac{1}{\sigma_{\text{full}}^2} (\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma_{\text{full}}^2} (\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2}) \Big)^{-1/4} \;.
\end{aligned}$$

By additionally noting that $(\epsilon_0 + \epsilon_1 + \epsilon_2)^{1/2} \leq \sqrt{\epsilon_0} + \sqrt{\epsilon_1} + \sqrt{\epsilon_2}$, we get that

$$\begin{aligned}
&\Big| \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K > t \Big) - \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K > t \Big) \Big| \\
&\leq C'(\sqrt{\epsilon_0} + \sqrt{\epsilon_1} + \sqrt{\epsilon_2}) \Big( \frac{1}{\sigma_{\text{full}}^2} (\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma_{\text{full}}^2} (\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2}) \Big)^{-1/4} \\
&\quad + \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{(n-1)\sigma_{\text{full}}^2} \big( \epsilon_0^{-2} + \epsilon_1^{-2} \big) + \frac{4(n-1)(\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2}{\epsilon_2^2 \sigma_{\text{full}}^2} \;.
\end{aligned}$$

Taking $K \to \infty$ on both sides, the inequality becomes

$$\begin{aligned}
&\Big| \lim_{K \to \infty} \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K > t \Big) - \lim_{K \to \infty} \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K > t \Big) \Big| \\
&\qquad \leq C'(\sqrt{\epsilon_0} + \sqrt{\epsilon_1} + \sqrt{\epsilon_2}) \Big( 1 + \frac{(n-1)\sigma_{\text{cond}}^2}{\sigma_{\text{full}}^2} \Big)^{-1/4} + \frac{2}{n-1} \big( \epsilon_0^{-2} + \epsilon_1^{-2} \big) + \frac{4(n-1)\sigma_{\text{cond}}^2}{\epsilon_2^2 \sigma_{\text{full}}^2} \\
&\qquad \leq C'(\sqrt{\epsilon_0} + \sqrt{\epsilon_1} + \sqrt{\epsilon_2}) + \frac{2}{n-1} \big( \epsilon_0^{-2} + \epsilon_1^{-2} \big) + \frac{4(n-1)\sigma_{\text{cond}}^2}{\epsilon_2^2 \sigma_{\text{full}}^2} \;.
\end{aligned}$$

Choosing $\epsilon_0 = \epsilon_1 = (n-1)^{-2/5}$ and $\epsilon_2 = \big((n-1)\sigma^2_{\rm cond}/\sigma^2_{\rm full}\big)^{2/5}$, redefining constants and taking a supremum over $t \in \mathbb{R}$, we get that there exists some absolute constant $C'' > 0$ such that

$$\sup_{t\in\mathbb{R}} \Big| \lim_{K\to\infty} \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\rm full}} W_n^K > t \Big) - \lim_{K\to\infty} \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\rm full}} U_n^K > t \Big) \Big| \leq C'' \Big( \frac{1}{(n-1)^{1/5}} + \Big( \frac{\sqrt{n-1}\,\sigma_{\rm cond}}{\sigma_{\rm full}} \Big)^{2/5} \Big) .$$

The final step is to relate this bound to $D_n$. Consider the last step (13) of the proof of Theorem 2 in Appendix C.2. If we set $\sigma = \sigma_{\rm full}$ instead of $\sigma_{\max}$, we get that there exists some absolute constant $C''' > 0$ such that

$$\sup_{t\in\mathbb{R}} \Big| \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\rm full}} D_n > t \Big) - \lim_{K\to\infty} \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\rm full}} U_n^K > t \Big) \Big| \leq C'''\, n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{(M_{{\rm full};\nu})^\nu}{\sigma^\nu_{\rm full}} + \frac{(M_{{\rm cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu_{\rm full}} \Big)^{\frac{1}{2\nu+1}} .$$

Setting $C = \max\{C'', C'''\}$ and using a triangle inequality, we get the desired bound that

$$\sup_{t\in\mathbb{R}} \Big| \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\rm full}} D_n > t \Big) - \lim_{K\to\infty} \mathbb{P}\Big( \frac{\sqrt{n(n-1)}}{\sigma_{\rm full}} W_n^K > t \Big) \Big| \leq C \Big( \frac{1}{(n-1)^{1/5}} + \Big( \frac{\sqrt{n-1}\,\sigma_{\rm cond}}{\sigma_{\rm full}} \Big)^{2/5} + n^{-\frac{\nu-2}{4\nu+2}} \Big( \frac{(M_{{\rm full};\nu})^\nu}{\sigma^\nu_{\rm full}} + \frac{(M_{{\rm cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu_{\rm full}} \Big)^{\frac{1}{2\nu+1}} \Big) .$$

■

### D.2. Proofs for results on $W_n$

**Proof of Proposition 12** To prove the existence of distribution, we seek to apply Lévy's continuity theorem. We first verify that there exists a sufficiently large $K^*$ such that the sequence $(W_n^K)_{K \geq K^*}$ is tight. Since Assumption 2 holds for some $\nu \geq 2$, we get that as $K \to \infty$,

$$\varepsilon_{K;2} \;\coloneqq\; \mathbb{E}\Big[ \big| \textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1) \phi_k(\mathbf{X}_2) - u(\mathbf{X}_1, \mathbf{X}_2) \big|^2 \Big]^{1/2} \to 0 .$$

In particular, there exists some sufficiently large $K^*$ such that $\varepsilon_{K;2} \leq 1$ for all $K \geq K^*$. By Lemma 33, we have that for all $K \geq K^*$,

$$\text{Var}[W_n^K] \;\leq\; \frac{2}{n(n-1)} (\sigma_{\rm full} + \varepsilon_{K;2})^2 \;\leq\; \frac{2}{n(n-1)} (\sigma_{\rm full} + 1)^2 .$$

Note that by assumption, we have $|D|, \sigma_{\rm full} < \infty$. This implies that the sequence $(W_n^K)_{K \geq K^*}$ is tight by a Markov inequality:

$$\lim_{x\to\infty} \Big( \sup_{K\geq K^*} \mathbb{P}\big( |W_n^K| > x \big) \Big) \;\leq\; \lim_{x\to\infty} \Big( x^{-2} \sup_{K\geq K^*} \mathbb{E}[(W_n^K)^2] \Big) \;\leq\; \lim_{x\to\infty} \frac{2n^{-1}(n-1)^{-1}(\sigma_{\rm full}+1)^2 + D^2}{x^2} \;=\; 0 .$$

We defer to Lemma 43 to show that the characteristic function of $(W_n^K - D)$ converges pointwise as $K \to \infty$. This allows us to apply Lévy's continuity theorem and obtain that $W_n$ exists. ■

**Proof of Lemma 13** The result holds by noting that for all $k > K^*$, $W_n^K = W_n^{K^*}$ almost surely, and the latter random variable does not depend on $K$. ■

**Lemma 43** *The characteristic function of* $(W_n^K - D)$ *converges pointwise as* $K \to \infty$.

**Proof of Lemma 43** Define $a_k := \frac{1}{\sqrt{n(n-1)}}\tau_{k;d}$ and $T_k := a_k(\xi_k^2 - 1)$, which allows us to write

$$W_n^K \;=\; \frac{1}{\sqrt{n(n-1)}} \sum_{k=1}^K \tau_{k;d}(\xi_k^2 - 1) + D \;=\; \sum_{k=1}^K T_k + D\;.$$

Denote $i = \sqrt{-1}$ as the imaginary unit and $Y$ as a chi-squared random variable with degree 1. Since each $T_k$ is a scaled and shifted chi-squared random variable with degree 1, it has the characteristic function

$$\psi_{T_k}(t) \;=\; \mathbb{E}[\exp(it\,T_k)] \;=\; \mathbb{E}[\exp(ia_kYt)]\exp(-ia_kt) \;=(1-2ia_kt)^{-1/2}\exp(-ia_kt)\;.$$

Since $T_k$'s are independent, by the convolution theorem, the characteristic function of $W_n^K - D$ is given by

$$\psi_{W_n^K - D}(t) \;=\; \exp\Big(-i\sum_{k=1}^K a_kt\Big)\prod_{k=1}^K(1-2ia_kt)^{-1/2}\;.$$

We want to prove that for every $t \in \mathbb{R}$, $\psi_{W_n^K-D}(t)$ converges to some function as limit $K \to \infty$. By taking the principal-valued complex logarithm (i.e. discontinuity along negative real axis), we get that

$$\log\psi_{W_n^K-D}(t) \;=\; \sum_{k=1}^K\Big(-ia_kt - \frac{1}{2}\log(1-2ia_kt)\Big) + 2im_K\pi \;=:\; S_K + 2im_K\pi\;, \quad (18)$$

for some $m_K \in \mathbb{N}$ for each $K$ that adjusts for values at discontinuity. Now consider the real part of the logarithm:

$$\begin{aligned}\mathrm{Re}\big(\log\psi_{W_n^K-D}(t)\big) \;=\; \mathrm{Re}(S_K) \;&=\; -\frac{1}{2}\sum_{k=1}^K\log|1-2ia_kt| \\ &=\; -\frac{1}{2}\sum_{k=1}^K\log\sqrt{1+4a_k^2t^2} \;=\; -\frac{1}{4}\sum_{k=1}^K\log(1+4a_k^2t^2)\;.\end{aligned}$$

Recall by Lemma 32 that

$$\sum_{k=1}^K a_k^2 \;=\; \frac{1}{n(n-1)}\sum_{k=1}^K\tau_{k;d}^2 \;=\; \mathrm{Tr}((\Sigma^K\Lambda^K)^2) \xrightarrow{K\to\infty} \sigma_{\text{full}}^2\;. \quad (19)$$

Fix $\epsilon > 0$. The above implies that there exists a sufficiently large $K^*$ such that for all $K_1, K_2 \geq K^*$, $\sum_{k=K_1}^{K_2} a_k^2 < \epsilon$. Then for all $K_1, K_2 \geq K^*$, we have

$$0 \;\leq\; \sum_{k=K_1}^{K_2}\log(1+4a_k^2t^2) \;\leq\; 4t^2\sum_{k=K_1}^{K_2}a_k^2 \;\leq\; 4t^2\epsilon\;.$$

This implies that $(\mathrm{Re}(S_K))_{K\in\mathbb{N}}$ is a Cauchy sequence and therefore converges. Now we handle the imaginary part. First let $m'_K \in \mathbb{Z}$ be such that

$$\mathrm{Im}\Big(\sum_{k=1}^K\log(1-2ia_kt)\Big) \;=\; \sum_{k=1}^K\arctan(-2a_kt) + 2m'_K\pi\;.$$

Then we have

$$\mathrm{Im}(S_K) \;=\; \sum_{k=1}^{K} \Big( -a_k t + \frac{1}{2}\arctan(2a_k t)\Big) - m'_K \pi \;=:\; I_K - m'_K\pi\;. \tag{20}$$

To show that $I_K$ converges, we first note that by a third-order Taylor expansion, we have that $\arctan(x) = x + \frac{6(x_*)^2-2}{6(x_*^2+1)^3}x^3$ for some $x_* \in [0,x]$ (we use this to denote $[0,x]$ for $x \geq 0$ as well as $[x,0]$ for $x < 0$, with an abuse of notation). This implies that for all $K_1, K_2 \geq K^*$, where $K^*$ is defined as before,

$$\begin{aligned}
\Big| \sum_{k=K_1}^{K_2} \Big( -a_k t + \frac{1}{2}\arctan(2a_k t)\Big)\Big| \;&=\; \Big| \sum_{k=K_1}^{K_2} \Big( -a_k t + \frac{1}{2}\arctan(2a_k t)\Big)\Big| \\
&\leq\; \sum_{k=K_1}^{K_2} \sup_{b_k\in[0,a_k]} \Big| \frac{1}{2}\frac{24b_k^2t^2-2}{6(4b_k^2t^2+1)^3} 8a_k^3t^3\Big| \\
&= 4t^3 \sum_{k=K_1}^{K_2} |a_k|^3 \Big( \sup_{b_k\in[0,a_k]} \Big| \frac{24b_k^2t^2+6-8}{6(4b_k^2t^2+1)^3}\Big|\Big) \\
&= 4t^3 \sum_{k=K_1}^{K_2} |a_k|^3 \Big( \sup_{b_k\in[0,a_k]} \Big| \frac{1}{(4b_k^2t^2+1)^2} - \frac{4}{3(4b_k^2t^2+1)^2}\Big|\Big) \\
&\leq 20t^3 \sum_{k=K_1}^{K_2} |a_k|^3 \;\leq\; 20t^3 \Big( \sum_{k=K_1}^{K_2} (a_k)^2\Big)^{3/2} \;\leq\; 20t^3\epsilon^{3/2}\;,
\end{aligned}$$

where, in the last line, we have used the relative sizes of $l_p$ norms. This implies that $I_K$ converges. To show that Eq. (20) converges, we need to show that $m_K$ in Eq. (20) is eventually constant. By using Eq. (20) and a triangle inequality, we have that

$$\begin{aligned}
\pi|m'_{K+1} - m'_K| \;&\leq |I_{K+1} - I_K| + \Big|\mathrm{Im}(S_{K+1}) - \mathrm{Im}(S_K)\Big| \\
&= |I_{K+1} - I_K| + \Big| a_{K+1}t + \frac{1}{2}\log(1 - 2ia_{K+1}t)\Big|\;.
\end{aligned}$$

The first term converges to zero, since we have shown that $I_K$ converges. Since $a_K \to 0$ by Eq. (19) and the complex logarithm we use is continuous outside $\{z : \mathrm{Re}(z) > 0\}$, the second term above also converges to zero. Therefore $|m'_{K+1} - m'_K| \to 0$, and since $(m'_K)_{K\in\mathbb{N}}$ is an integer sequence, $(m'_K)_{K\in\mathbb{N}}$ converges. By Eq. (20), this implies that $\mathrm{Im}(S_K)$ converges, and since we have shown $\mathrm{Re}(S_K)$ converges, we get that $S_K$ converges. Finally, to show that $\psi_{W_n^K - D}(t)$ converges, since $\mathrm{Re}(S_K) = \mathrm{Re}\big(\psi_{W_n^K-D}(t)\big)$, we only need to show that $\mathrm{Im}\big(\psi_{W_n^K-D}(t)\big)$ converges. By Eq. (18), this again reduces to showing that $m_K$ is eventually constant. As before, by a triangle inequality,

$$\begin{aligned}
2\pi|m_{K+1} - m_K| \;&\leq |\mathrm{Im}(S_{K+1}) - \mathrm{Im}(S_K)| + \big|\mathrm{Im}(\log\psi_{W_n^{K+1}-D}(t)) - \mathrm{Im}(\log\psi_{W_n^K-D}(t))\big| \\
&= |\mathrm{Im}(S_{K+1}) - \mathrm{Im}(S_K)| + \Big| a_{K+1}t + \frac{1}{2}\log(1 - 2ia_{K+1}t)\Big| \xrightarrow{K\to\infty} 0\;,
\end{aligned}$$

where the convergence of both terms has been shown earlier. This proves that the characteristic function $\psi_{W_n^K-D}(t)$ converges for every $t \in \mathbb{R}$. ■

## Appendix E. Proofs for Section 4

### E.1. Proofs for the general results

**Proof of Lemma 15** To prove the first result, note that since $\kappa$ is a kernel, there exists a RKHS $\mathcal{H}$ and a map $\Phi : \mathbb{R}^d \to \mathcal{H}$ such that we can write

$$\begin{aligned}
u^{\mathrm{MMD}}\big((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')\big) \;&= \langle\Phi(\mathbf{x}),\Phi(\mathbf{y}')\rangle_{\mathcal{H}} + \langle\Phi(\mathbf{y}),\Phi(\mathbf{y}')\rangle_{\mathcal{H}} - \langle\Phi(\mathbf{x}),\Phi(\mathbf{y}')\rangle_{\mathcal{H}} - \langle\Phi(\mathbf{x}'),\Phi(\mathbf{y})\rangle_{\mathcal{H}} \\
&= \langle\Phi(\mathbf{x}) - \Phi(\mathbf{y}), \Phi(\mathbf{x}') - \Phi(\mathbf{y}')\rangle_{\mathcal{H}}\;.
\end{aligned}$$

Defining $\Phi_*\big((\mathbf{x},\mathbf{y})\big) \coloneqq \Phi(\mathbf{x}) - \Phi(\mathbf{y})$ proves that $u^{\rm MMD}$ is a kernel. To prove the second result, note that by the definition of a weak Mercer representation, we have that almost surely

$$\big| \sum_{k=1}^K \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2) - u^{\rm MMD}(\mathbf{Z}_1,\mathbf{Z}_2)\big| \xrightarrow{K\to\infty} 0 \;,$$

which in particular implies convergence in probability. The argument uses the Vitali convergence theorem. By Assumption 3, there exists some $\nu^* > \nu$ such that $\sup_{K\geq 1}\mathbb{E}[| \sum_{k=1}^K \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)|^{\nu^*}] < \infty$ and $\mathbb{E}[|u^{\rm MMD}(\mathbf{Z}_1,\mathbf{Z}_2)|^{\nu^*}] < \infty$. By a triangle inequality and a Jensen's inequality, we have

$$\begin{aligned}
&\sup_{K\geq 1}\mathbb{E}\Big[\Big| \sum_{k=1}^K \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2) - u^{\rm MMD}(\mathbf{Z}_1,\mathbf{Z}_2)\Big|^{\nu^*}\Big] \\
&\leq \sup_{K\geq 1}\mathbb{E}\Big[\Big| \big|\sum_{k=1}^K \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\big| + \big|u^{\rm MMD}(\mathbf{Z}_1,\mathbf{Z}_2)\big|\Big|^{\nu^*}\Big] \\
&\leq 2^{\nu^*-1}\sup_{K\geq 1}\mathbb{E}\Big[\big| \sum_{k=1}^K \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\big|^{\nu^*}\Big] + 2^{\nu^*-1}\mathbb{E}\Big[\big|u^{\rm MMD}(\mathbf{Z}_1,\mathbf{Z}_2)\big|^{\nu^*}\Big] \;<\; \infty \;.
\end{aligned}$$

This implies for any $\nu \in (2,\nu^*)$, the sequence $\big(\big( \sum_{k=1}^K \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2) - u^{\rm MMD}(\mathbf{Z}_1,\mathbf{Z}_2)\big)^\nu\big)_{K\in\mathbb{N}}$ is uniformly integrable, and therefore converges to zero in $L_1(\mathbb{R}^{2d}, P\otimes Q)$ by the Vitali convergence theorem. Since convergence in $L_\nu$ implies convergence in $L_{\min\{\nu,3\}}$, we get that Assumption 2 holds for $\min\{\nu,3\}$. ■

Before we prove the next result, recall that $\{\lambda_k\}_{k=1}^\infty$ and $\{\phi_k\}_{k=1}^\infty$ are defined as the weak Mercer representation for the kernel $\kappa$ under $Q$, and we have assumed that $\phi_k$'s are differentiable. We have also defined the sequence of values $\{\alpha_k\}_{k=1}^\infty$ and the sequence of functions $\{\psi_k\}_{k=1}^\infty$ in (6) as

$$\alpha_{(k'-1)d+l} \;\coloneqq \lambda_{k'} \qquad \text{and} \qquad \psi_{(k'-1)d+l}(\mathbf{x}) \;\coloneqq (\partial_{x_l}\log p(\mathbf{x}))\phi_{k'}(\mathbf{x}) + \partial_{x_l}\phi_{k'}(\mathbf{x}) \;,$$

for $1\leq l\leq d$ and $k'\in\mathbb{N}$. For convenience, we denote $\psi_{k';l} \coloneqq \psi_{(k'-1)d+l}$ in the proof below.

**Proof of Lemma 16** Recall that $\psi_{k';l}(\mathbf{x}) \coloneqq (\partial_{x_l}\log p(\mathbf{x}))\phi_{k'}(\mathbf{x}) + \partial_{x_l}\phi_{k'}(\mathbf{x})$. Write $\tilde\psi_{k'}(\mathbf{x}) \coloneqq (\psi_{k';1}(\mathbf{x}),\ldots,\psi_{k';n}(\mathbf{x}))^\top$. We first consider the error term with $dK'$ summands for some $K'\in\mathbb{N}$:

$$\begin{aligned}
&\mathbb{E}\Big[\Big| \sum_{k=1}^{dK'} \alpha_k\psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2) - u_P^{\rm KSD}(\mathbf{X}_1,\mathbf{X}_2)\Big|^\nu\Big] \\
&= \mathbb{E}\Big[\Big| \sum_{l=1}^{d}\sum_{k'=1}^{K'} \lambda_{k'}\psi_{k';l}(\mathbf{X}_1)\psi_{k';l}(\mathbf{X}_2) - u_P^{\rm KSD}(\mathbf{X}_1,\mathbf{X}_2)\Big|^\nu\Big] \\
&= \mathbb{E}\Big[\Big| \sum_{k'=1}^{K'} \lambda_{k'}\big(\tilde\psi_{k'}(\mathbf{X}_1)\big)^\top\big(\tilde\psi_{k'}(\mathbf{X}_2)\big) - u_P^{\rm KSD}(\mathbf{X}_1,\mathbf{X}_2)\Big|^\nu\Big] \\
&= \mathbb{E}\big[\big|T_1+T_2+T_3+T_4 - u_P^{\rm KSD}(\mathbf{X}_1,\mathbf{X}_2)\big|^\nu\big] \;,
\end{aligned}$$

where the random quantities are defined in terms of $\mathbf{X}_1,\mathbf{X}_2 \overset{i.i.d.}{\sim} Q$:

$$\begin{aligned}
T_1 \;&\coloneqq \big(\nabla\log p(\mathbf{X}_1)\big)^\top\big(\nabla\log p(\mathbf{X}_2)\big)\sum_{k'=1}^{K'}\lambda_{k'}\phi_{k'}(\mathbf{X}_1)\phi_{k'}(\mathbf{X}_2) \;,\\
T_2 \;&\coloneqq \big(\nabla\log p(\mathbf{X}_1)\big)^\top\Big(\sum_{k'=1}^{K'}\lambda_{k'}\big(\nabla\phi_{k'}(\mathbf{X}_2)\big)\phi_{k'}(\mathbf{X}_1)\Big) \;,\\
T_3 \;&\coloneqq \big(\nabla\log p(\mathbf{X}_2)\big)^\top\Big(\sum_{k'=1}^{K'}\lambda_{k'}\big(\nabla\phi_{k'}(\mathbf{X}_1)\big)\phi_{k'}(\mathbf{X}_2)\Big) \;,\\
T_4 \;&\coloneqq \sum_{k'=1}^{K'}\lambda_{k'}\big(\nabla\phi_{k'}(\mathbf{X}_1)\big)^\top\big(\nabla\phi_{k'}(\mathbf{X}_2)\big) \;.
\end{aligned}$$

Recall that by Assumption 3, there exists some $\nu^* > \nu$ such that we have $\|\kappa^*(\mathbf{Z}_1, \mathbf{Z}_2)\|_{L_{\nu^*}} < \infty$ and $\sup_{K \geq 1} \| \sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\|_{L_{\nu^*}} < \infty$. By using the proof of the second part of Lemma 15 above, for $\nu^{\Delta} := \frac{\nu+\nu^*}{2} \in (\nu, \nu^*)$, we have

$$\mathbb{E}\Big[\big| \textstyle\sum_{k'=1}^{K'} \lambda_{k'}\phi_{k'}(\mathbf{X}_1)\phi_{k'}(\mathbf{X}_2) - u(\mathbf{X}_1, \mathbf{X}_2)\big|^{\nu^{\Delta}}\Big] \xrightarrow{K' \to \infty} 0 \;.$$

Meanwhile by Assumption 4, $\big\|\|\nabla \log p(\mathbf{X}_1)\|_2\big\|_{L_{2\nu^{**}}} < \infty$, where

$$\nu^{**} \;=\; \frac{\nu(\nu+\nu^*)}{\nu^* - \nu} \;=\; \Big(\frac{1}{\nu} - \frac{2}{\nu+\nu^*}\Big)^{-1} \;=\; \Big(\frac{1}{\nu} - \frac{1}{\nu^{\Delta}}\Big)^{-1} \;>\; \nu \;.$$

By a Cauchy-Schwarz inequality and a Hölder's inequality, we have that

$$\big\|\big(\nabla \log p(\mathbf{X}_1)\big)^\top \big(\nabla \log p(\mathbf{X}_2)\big)\big\|_{L_{\nu^{**}}} \;\leq \big\| \,\|\nabla \log p(\mathbf{X}_1)\|_2 \,\big\|_{L_{2\nu^{**}}} < \infty \;.$$

Now by a Hölder's inequality and noting that $(\nu^{**})^{-1} + (\nu^{\Delta})^{-1} = \nu^{-1}$, we can now bound the error of using $T_1$ to approximate the first term of $u_P^{\mathrm{KSD}}$ as

$$\begin{aligned}
&\mathbb{E}[|E_1|^{\nu}] \;:=\; \mathbb{E}\big[\big|T_1 - \big(\nabla \log p(\mathbf{X}_1)\big)^\top \big(\nabla \log p(\mathbf{X}_2)\big)\, u(\mathbf{X}_1, \mathbf{X}_2)\big|^{\nu}\big] \\
&=\; \big\|T_1 - \big(\nabla \log p(\mathbf{X}_1)\big)^\top \big(\nabla \log p(\mathbf{X}_2)\big)\, u(\mathbf{X}_1, \mathbf{X}_2)\big\|_{L_\nu}^{\nu} \\
&\leq\; \big\|\big(\nabla \log p(\mathbf{X}_1)\big)^\top \big(\nabla \log p(\mathbf{X}_2)\big)\big\|_{L_{\nu^{**}}}^{\nu} \,\big\| \textstyle\sum_{k'=1}^{K'} \lambda_{k'}\phi_{k'}(\mathbf{X}_1)\phi_{k'}(\mathbf{X}_2) - u(\mathbf{X}_1, \mathbf{X}_2)\big\|_{L_{\nu^{\Delta}}}^{\nu} \\
&\xrightarrow{K' \to \infty}\; 0 \;.
\end{aligned}$$

For $T_2$, we consider a similar approximation error quantity and apply a Cauchy-Schwarz inequality:

$$\begin{aligned}
\mathbb{E}[|E_2|^{\nu}] \;&:= \mathbb{E}\big[\big|T_2 - \big(\nabla \log p(\mathbf{X}_1)\big)^\top \nabla_2 \kappa(\mathbf{X}_1, \mathbf{X}_2)\big|^{\nu}\big] \\
&=\; \mathbb{E}\Big[\Big|\big(\nabla \log p(\mathbf{X}_1)\big)^\top \Big(\textstyle\sum_{k'=1}^{K'} \lambda_{k'}\big(\nabla \phi_{k'}(\mathbf{X}_2)\big)\phi_{k'}(\mathbf{X}_1) - \nabla_2 \kappa(\mathbf{X}_1, \mathbf{X}_2)\Big)\Big|^{\nu}\Big] \\
&\leq\; \| \,\|\nabla \log p(\mathbf{X}_1)\|_2 \,\|_{L_{2\nu}}^{\nu} \Big\|\Big\| \textstyle\sum_{k'=1}^{K'} \lambda_{k'}\big(\nabla \phi_{k'}(\mathbf{X}_2)\big)\phi_{k'}(\mathbf{X}_1) - \nabla_2 \kappa(\mathbf{X}_1, \mathbf{X}_2)\Big\|_2\Big\|_{L_{2\nu}}^{\nu} \\
&\xrightarrow{K' \to \infty}\; 0 \;,
\end{aligned}$$

where we have noted that the first term is bounded since $2\nu < 2\nu^{**}$ and used Assumption 4(iv). By symmetry of $\kappa$ and the fact that $\mathbf{X}_1$ and $\mathbf{X}_2$ are exchangeable, we have the same result for $T_3$:

$$\mathbb{E}[|E_3|^{\nu}] \;:= \mathbb{E}\big[\big|T_3 - \big(\nabla \log p(\mathbf{X}_2)\big)^\top \nabla_1 \kappa(\mathbf{X}_1, \mathbf{X}_2)\big|^{\nu}\big] \;\xrightarrow{K' \to \infty}\; 0 \;.$$

Meanwhile, the second condition of Assumption 4(iv) directly says that

$$\mathbb{E}[|E_4|^{\nu}] \;:= \mathbb{E}\big[\big|T_4 - \mathrm{Tr}\big(\nabla_1 \nabla_2 \kappa(\mathbf{X}_1, \mathbf{X}_2)\big)\big|^{\nu}\big] \;\xrightarrow{K' \to \infty}\; 0 \;.$$

Combining the results and applying a Jensen's inequality to the convex function $x \mapsto |x|^{\nu}$, we have

$$\begin{aligned}
\mathbb{E}\Big[\Big| \textstyle\sum_{k=1}^{dK'} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^{\nu}\Big] \;&=\; \mathbb{E}\big[\big|E_1 + E_2 + E_3 + E_4\big|^{\nu}\big] \\
&\leq \mathbb{E}\big[\big|\tfrac{1}{4}(4E_1) + \tfrac{1}{4}(4E_2) + \tfrac{1}{4}(4E_3) + \tfrac{1}{4}(4E_4)\big|^{\nu}\big] \\
&\leq 4^{\nu-1}\big(\mathbb{E}[|E_1|^{\nu}] + \mathbb{E}[|E_2|^{\nu}] + \mathbb{E}[|E_3|^{\nu}] + \mathbb{E}[|E_4|^{\nu}]\big) \;\xrightarrow{K' \to \infty}\; 0 \;.
\end{aligned}$$

Now consider $K \in \mathbb{N}$ that is not necessarily divisible by $d$, and let $K'$ be the greatest integer such that $K \geq dK'$. Then by a triangle inequality and a similar Jensen's inequality as above, we get

$$\begin{aligned} &\mathbb{E}\Big[\Big| \textstyle\sum_{k=1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^\nu\Big] \\ &\leq 2^{\nu-1}\mathbb{E}\Big[\Big| \textstyle\sum_{k=1}^{dK'} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^\nu\Big] \\ &\quad + 2^{\nu-1}\mathbb{E}\Big[\Big| \textstyle\sum_{k=dK'+1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)\Big|^\nu\Big] . \end{aligned} \tag{21}$$

The first term is $o(1)$ as $K \to \infty$ by the previous argument, so we only need to focus on the second term. The expectation can be bounded by noting that $\alpha_k = \lambda_{K'+1} \geq 0$ for all $dK' + 1 \leq k \leq K$ and using a triangle inequality followed by a Jensen's inequality:

$$\begin{aligned} &\mathbb{E}\Big[\Big| \textstyle\sum_{k=dK'+1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)\Big|^\nu\Big] \\ &\leq (\lambda_{K'+1})^\nu \mathbb{E}\Big[\Big(\frac{1}{K-dK'} \textstyle\sum_{k=dK'+1}^{K} (K-dK')|\psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)|\Big)^\nu\Big] \\ &\leq (\lambda_{K'+1})^\nu (K-dK')^{\nu-1} \textstyle\sum_{k=dK'+1}^{K} \mathbb{E}[|\psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)|^\nu] \\ &\leq (\lambda_{K'+1})^\nu d^\nu \sup_{k\in\{dK'+1,\ldots,dK'+d\}} \mathbb{E}[|\psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)|^\nu] \\ &= (\lambda_{K'+1})^\nu d^\nu \sup_{1\leq l\leq d} \mathbb{E}[|\psi_{dK'+l}(\mathbf{X}_1)|^\nu]^2 . \end{aligned}$$

In the last equality, we have noted that $\mathbf{X}_1$ and $\mathbf{X}_2$ are identically distributed. Now by the definition of $\psi_k$, another Jensen's inequality on $x \mapsto |x|^\nu$ and a Cauchy-Schwarz inequality, we have

$$\begin{aligned} \mathbb{E}[|\psi_{dK'+l}(\mathbf{X}_1)|^\nu] &= \mathbb{E}[|(\partial_{x_l} \log p(\mathbf{X}_1))\phi_{K'+1}(\mathbf{X}_1) + \partial_{x_l}\phi_{K'+1}(\mathbf{X}_1)|^\nu] \\ &\leq 2^{\nu-1}\mathbb{E}[|(\partial_{x_l} \log p(\mathbf{X}_1))\phi_{K'+1}(\mathbf{X}_1)|^\nu] + 2^{\nu-1}\mathbb{E}[|\partial_{x_l}\phi_{K'+1}(\mathbf{X}_1)|^\nu] \\ &\leq 2^{\nu-1}\mathbb{E}[|\partial_{x_l} \log p(\mathbf{X}_1)|^{2\nu}]^{1/2}\,\mathbb{E}[|\phi_{K'+1}(\mathbf{X}_1)|^{2\nu}]^{1/2} + 2^{\nu-1}\mathbb{E}[|\partial_{x_l}\phi_{K'+1}(\mathbf{X}_1)|^\nu] \\ &\leq 2^{\nu-1}\mathbb{E}[\|\nabla \log p(\mathbf{X}_1)\|_2^{2\nu}]^{1/2}\,\mathbb{E}[|\phi_{K'+1}(\mathbf{X}_1)|^{2\nu}]^{1/2} + 2^{\nu-1}\mathbb{E}[\|\nabla\phi_{K'+1}(\mathbf{X}_1)\|_2^\nu] \\ &= 2^{\nu-1}\|\|\nabla \log p(\mathbf{X}_1)\|_2\|_{L_{2\nu}}^\nu\, \|\phi_{K'+1}(\mathbf{X}_1)\|_{L_{2\nu}}^\nu + 2^{\nu-1}\|\|\nabla\phi_{K'+1}(\mathbf{X}_1)\|_2\|_{L_\nu}^\nu . \end{aligned}$$

By Assumption 4(i), (ii) and (iii), all three norms are bounded, so $\mathbb{E}[|\psi_{dK'+l}(\mathbf{X}_1)|^\nu] < \infty$. By the definition of $\lambda_k$ from the weak Mercer representation, as $K \to \infty$ and therefore $K' \to \infty$, $\lambda_{K'+1} \to 0$, which implies

$$\mathbb{E}\Big[\Big| \textstyle\sum_{k=dK'+1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)\Big|^\nu\Big] = o(1) .$$

This means that both terms in (21) converge to 0 as $K \to \infty$. In other words,

$$\mathbb{E}\Big[\Big| \textstyle\sum_{k=1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^\nu\Big] \xrightarrow{K\to\infty} 0 .$$

Since $L_\nu$-convergence implies $L_{\min\{\nu,3\}}$-convergence and we have assumed that $\nu > 2$, we get that Assumption 2 holds for $\min\{\nu, 3\}$ with respect to the $u_P^{\mathrm{KSD}}$, $\alpha_k$ and $\psi_k$. ■

### E.2. Proof of Proposition 18

From Lemma 22, we can write the variance ratio as

$$\frac{\sigma^2_{\text{full}}}{\sigma^2_{\text{cond}}} \;=\; \Big(\frac{\gamma}{4+\gamma}\Big)^{d/2} \Big(\frac{(1+\gamma)(3+\gamma)}{\gamma^2}\Big)^{d/2} \frac{B}{A} \;=\; C \times \frac{B}{A} \;,$$

where

$$\begin{aligned} A \;&:= \frac{(2+\gamma)^2}{(1+\gamma)(3+\gamma)} \|\mu\|_2^2 + \bigg(1 - \Big(\frac{(1+\gamma)(3+\gamma)}{(2+\gamma)^2}\Big)^{d/2}\bigg) \|\mu\|_2^4 \\ &= \; (1+o(1))\,\|\mu\|_2^2 + \Big(1 - (1-\alpha)^{d/2}\Big)\|\mu\|_2^4 \\ B \;&:= d + \frac{d^2}{\gamma^2} + \frac{2d\|\mu\|_2^2}{\gamma} + 2\|\mu\|_2^2 + \bigg(1 - \Big(\frac{\gamma(4+\gamma)}{(2+\gamma)^2}\Big)^{d/2}\bigg)\|\mu\|_2^4 + o\Big(d + \frac{d^2}{\gamma^2} + \frac{d\|\mu\|_2^2}{\gamma} + \|\mu\|_2^2\Big) \\ &= d + \frac{d^2}{\gamma^2} + \frac{2d\|\mu\|_2^2}{\gamma} + 2\|\mu\|_2^2 + \Big(1 - (1-\delta)^{d/2}\Big)\|\mu\|_2^4 + o\Big(d + \frac{d^2}{\gamma^2} + \frac{d\|\mu\|_2^2}{\gamma} + \|\mu\|_2^2\Big) \\ C \;&:= \Big(\frac{\gamma}{4+\gamma}\Big)^{d/2}\Big(\frac{(1+\gamma)(3+\gamma)}{\gamma^2}\Big)^{d/2} \;=\; \Big(\frac{(1+\gamma)(3+\gamma)}{\gamma(4+\gamma)}\Big)^{d/2} \;, \end{aligned}$$

and we have written $\frac{(1+\gamma)(3+\gamma)}{(2+\gamma)^2} = 1-\alpha$ with $\alpha := \frac{1}{(2+\gamma)^2}$ and $\frac{\gamma(4+\gamma)}{(2+\gamma)^2} = 1-\delta$ with $\delta := \frac{4}{(2+\gamma)^2}$. To simplify $A$ and $B$, we first rewrite

$$\begin{aligned} 1-(1-\alpha)^{d/2} \;&=\; 1 - \exp\Big(-\frac{d}{2}\log(1-\alpha)\Big) \overset{(a)}{=} 1 - \exp\Big(-\frac{d}{2}\Big(\frac{1}{(2+\gamma)^2} + O\Big(\frac{1}{\gamma^4}\Big)\Big)\Big) \\ &= 1 - \exp\Big(-\frac{d}{2(2+\gamma)^2} + O\Big(\frac{d}{\gamma^4}\Big)\Big) \;. \end{aligned} \tag{22}$$

In $(a)$, we have used a Taylor expansion by noting that $\gamma$ is small by the stated assumption $\gamma = \omega(1)$. Similarly we can obtain

$$\begin{aligned} 1-(1-\delta)^{d/2} \;&=\; 1 - \exp\Big(\frac{d}{2}\log(1-\delta)\Big) \;=\; 1 - \exp\Big(\frac{d}{2}\Big(-\frac{4}{(2+\gamma)^2} + O\Big(\frac{1}{\gamma^4}\Big)\Big)\Big) \\ &= \; 1 - \exp\Big(-\frac{2d}{(2+\gamma)^2} + O\Big(\frac{d}{\gamma^4}\Big)\Big) \;, \end{aligned} \tag{23}$$

$$\begin{aligned} C \;&=\; \exp\Big(\frac{d}{2}\log\Big(1 + \frac{3}{\gamma(4+\gamma)}\Big)\Big) \;=\; \exp\Big(\frac{d}{2}\Big(\frac{3}{\gamma(4+\gamma)} + O\Big(\frac{1}{\gamma^4}\Big)\Big)\Big) \\ &= \; \exp\Big(\frac{3d}{2\gamma(4+\gamma)} + O\Big(\frac{d}{\gamma^4}\Big)\Big) \;. \end{aligned} \tag{24}$$

Therefore, the terms $(1-\alpha)^{d/2}$, $(1-\gamma)^{d/2}$ and $C$ can be small, large or close to a constant, depending on whether $\gamma^2$ grows faster than, lower than, or at the same rate as $d$. We now consider the three cases individually.

**Case 1:** $\gamma = o(d^{1/2})$**.** In this case, $\frac{d}{(2+\gamma)^2} = \omega(1)$, so we have $(1-\alpha)^{d/2} = o(1)$ and $(1-\delta)^{d/2} = o(1)$. Therefore

$$A \;= (1+o(1))\|\mu\|_2^2 + (1+o(1))\|\mu\|_2^4 \;=\; \Theta(\|\mu\|_2^2 + \|\mu\|_2^4) \;=\; \Theta(\|\mu\|_2^4) \;,$$

and

$$\begin{aligned} B \;&= d + \frac{d^2}{\gamma^2} + \frac{2d\|\mu\|_2^2}{\gamma} + 2\|\mu\|_2^2 + \|\mu\|_2^4 + o\left(d + \frac{d^2}{\gamma^2} + \frac{d\|\mu\|_2^2}{\gamma} + \|\mu\|_2^2\right) \\ &= \Theta\left(d + \frac{d^2}{\gamma^2} + \frac{d\|\mu\|_2^2}{\gamma} + \|\mu\|_2^4\right) \;. \end{aligned}$$

Combining with the previous expressions for $A$, $B$ and $C$ yields

$$\rho_d = \frac{\sigma_{\text{full}}}{\sigma_{\text{cond}}} = \sqrt{C} \times \frac{\sqrt{B}}{\sqrt{A}} = \exp\left(\frac{3d}{4\gamma(4+\gamma)} + O\left(\frac{d}{\gamma^4}\right)\right) \Theta\left(\sqrt{\frac{d}{\|\mu\|_2^4} + \frac{d^2}{\gamma^2\|\mu\|_2^4} + \frac{d}{\gamma\|\mu\|_2^2} + 1}\right)$$
$$\stackrel{(a)}{=} \exp\left(\frac{3d}{4\gamma^2} + o\left(\frac{d}{\gamma^2}\right)\right) \Theta\left(\sqrt{\frac{d^2}{\gamma^2\|\mu\|_2^4} + \frac{d}{\gamma\|\mu\|_2^2} + 1}\right)$$
$$\stackrel{(b)}{=} \exp\left(\frac{3d}{4\gamma^2} + o\left(\frac{d}{\gamma^2}\right)\right) \Theta\left(\frac{d}{\gamma\|\mu\|_2^2} + \frac{d^{1/2}}{\gamma^{1/2}\|\mu\|_2} + 1\right),$$

where in $(a)$ we have used the fact that $\gamma = o(d^{1/2})$, and in $(b)$ we have noted that for $a, b, c > 0$, $\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$ and by a Jensen's inequality, $\sqrt{a+b+c} \geq \frac{1}{\sqrt{3}}(\sqrt{a} + \sqrt{b} + \sqrt{c})$.

**Case 2:** $\gamma = \omega(d^{1/2})$**.** Since in this case $\frac{d}{\gamma^2}$ is small, we can use Taylor expansion to approximate the exponential term in (22) to get

$$1 - (1-\alpha)^{d/2} = 1 - \exp\left(-\frac{d}{2(2+\gamma)^2} + O\left(\frac{d}{\gamma^4}\right)\right) = 1 - \left(1 - \frac{d}{2(2+\gamma)^2} + O\left(\frac{d^2}{\gamma^4}\right)\right)$$
$$= \frac{d}{2(2+\gamma)^2} + o\left(\frac{d}{\gamma^2}\right).$$

Using a similar argument applied to (23), we have

$$1 - (1-\delta)^{d/2} = 1 - \exp\left(-\frac{2d}{(2+\gamma)^2} + O\left(\frac{d}{\gamma^4}\right)\right) = \frac{2d}{(2+\gamma)^2} + o\left(\frac{d}{\gamma^2}\right),$$

and (24) yields

$$C = \exp\left(\frac{3d}{2\gamma(4+\gamma)} + O\left(\frac{d}{\gamma^4}\right)\right) = 1 + \frac{3d}{2\gamma(4+\gamma)} + O\left(\frac{d^2}{\gamma^4}\right) = 1 + o(1).$$

We therefore conclude that

$$A = (1+o(1))\|\mu\|_2^2 + \left(\frac{d}{2(2+\gamma)^2} + o\left(\frac{d}{\gamma^2}\right)\right)\|\mu\|_2^4 = \Theta\left(\|\mu\|_2^2 + \frac{d}{\gamma^2}\|\mu\|_2^4\right),$$

A similar argument shows that

$$B = d + \frac{d^2}{\gamma^2} + \frac{2d\|\mu\|_2^2}{\gamma} + 2\|\mu\|_2^2 + \|\mu\|_2^4\left(\frac{2d}{(2+\gamma)^2} + o\left(\frac{d}{\gamma^2}\right)\right) + o\left(d + \frac{d^2}{\gamma^2} + \frac{d\|\mu\|_2^2}{\gamma} + \|\mu\|_2^2\right)$$
$$= \Theta\left(d + \frac{d\|\mu\|_2^2}{\gamma} + \|\mu\|_2^2 + \frac{d\|\mu\|_2^4}{\gamma^2}\right),$$

where in the last line we noted that $\gamma = \omega(d^{1/2})$ implies $\frac{d^2}{\gamma^2} = o(d)$. Combining the results gives

$$\rho_d = \frac{\sigma_{\text{full}}}{\sigma_{\text{cond}}} = \sqrt{C} \times \frac{\sqrt{B}}{\sqrt{A}}$$
$$= \sqrt{1+o(1)}\ \Theta\left(\frac{d^{1/2} + \gamma^{-1/2}d^{1/2}\|\mu\|_2 + \|\mu\|_2 + \gamma^{-1}d^{1/2}\|\mu\|_2^2}{\|\mu\|_2 + \gamma^{-1}d^{1/2}\|\mu\|_2^2}\right)$$
$$= \Theta\left(\frac{d^{1/2} + \gamma^{-1/2}d^{1/2}\|\mu\|_2}{\|\mu\|_2 + \gamma^{-1}d^{1/2}\|\mu\|_2^2} + 1\right) = \Theta\left(\frac{d^{1/2}(1 + \gamma^{-1/2}\|\mu\|_2)}{\|\mu\|_2\,(1 + \gamma^{-1}d^{1/2}\|\mu\|_2)} + 1\right).$$

**Case 3: $\gamma = \Theta(d^{1/2})$.** Since in this case $\frac{d}{\gamma^4}$ is small, we have that $\exp\big(O\big(\frac{d}{\gamma^4}\big)\big) = 1 + O\big(\frac{d}{\gamma^4}\big)$ by a Taylor expansion. Substituting this into (22), we have

$$\begin{aligned} 0 \;\leq\; 1-(1-\alpha)^{d/2} \;&= 1 - \exp\Big(-\frac{d}{2(2+\gamma)^2}\Big)\Big(1 + O\Big(\frac{d}{\gamma^4}\Big)\Big) \\ &= 1 - \exp\Big(-\frac{d}{2(2+\gamma)^2}\Big) + O\Big(\frac{d}{\gamma^4}\Big) \;=\; \Theta(1)\;, \end{aligned}$$

where the last line holds as $1 - \exp\big(-\frac{d}{2(2+\gamma)^2}\big) = \Theta(1)$. A similar argument applied to (23) and (24) gives

$$\begin{aligned} 0 \;\leq\; 1-(1-\delta)^{d/2} \;&= 1 - \exp\Big(-\frac{2d}{(2+\gamma)^2} + O\Big(\frac{d}{\gamma^4}\Big)\Big) \;=\; \Theta(1)\;, \\ 0 \;\leq\; C \;&=\; \exp\Big(\frac{3d}{2\gamma(4+\gamma)} + O\Big(\frac{d}{\gamma^4}\Big)\Big) \;=\; \Theta(1)\;. \end{aligned}$$

Combining the above derivations yields

$$\begin{aligned} A \;&= \Theta(\|\mu\|_2^2 + \|\mu\|_2^4) \;=\; \Theta(\|\mu\|_2^4)\;, \\ B \;&= \Theta\Big(d + \frac{d^2}{\gamma^2} + \frac{2d\|\mu\|_2^2}{\gamma} + 2\|\mu\|_2^2 + \|\mu\|_2^4\Big) \;=\; \Theta\Big(d + d^{1/2}\|\mu\|_2^2 + \|\mu\|_2^4\Big)\;, \end{aligned}$$

where in the equality for $B$ we have used the fact that $\|\mu\|_2^2 = \Omega(1)$ implies $\|\mu\|_2^2 = O(\|\mu\|_2^4)$ and that $\gamma = \Theta(d^{1/2})$ implies $\frac{d}{\gamma} = \Theta(d^{1/2})$. Therefore,

$$\rho_d \;=\; \frac{\sigma_{\text{full}}}{\sigma_{\text{cond}}} \;=\; \sqrt{C} \times \frac{\sqrt{B}}{\sqrt{A}} \;=\Theta\left(\sqrt{\frac{d}{\|\mu\|_2^4} + \frac{d^{1/2}}{\|\mu\|_2^2} + 1}\right) \;=\; \Theta\left(\frac{d^{1/2}}{\|\mu\|_2^2} + \frac{d^{1/4}}{\|\mu\|_2} + 1\right)\;.$$

This completes the proof.

### E.3. Proof of Proposition 19

Recall the expressions of $\sigma^2_{\text{cond}}$ and $\sigma^2_{\text{full}}$ for MMD-RBF from Lemma 24, which allow us to rewrite $\sigma^2_{\text{cond}} = CA$ and $\sigma^2_{\text{full}} = CB$, where

$$\begin{aligned} A \;&:= 1 + \exp\Big(-\frac{1}{3+\gamma}\|\mu\|_2^2\Big) + 2\Big(\frac{3+\gamma}{2+\gamma}\Big)^{d/2}\Big(\frac{1+\gamma}{2+\gamma}\Big)^{d/2}\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big) \\ &\quad - 2\exp\Big(-\frac{2+\gamma}{2(1+\gamma)(3+\gamma)}\|\mu\|_2^2\Big) - \Big(\frac{3+\gamma}{2+\gamma}\Big)^{d/2}\Big(\frac{1+\gamma}{2+\gamma}\Big)^{d/2} \\ &\quad - \Big(\frac{3+\gamma}{2+\gamma}\Big)^{d/2}\Big(\frac{1+\gamma}{2+\gamma}\Big)^{d/2}\exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big) \\ B \;&:= \Big(\frac{3+\gamma}{4+\gamma}\Big)^{d/2}\Big(\frac{1+\gamma}{\gamma}\Big)^{d/2}\Big(1 + \exp\Big(-\frac{1}{4+\gamma}\|\mu\|_2^2\Big)\Big) - \Big(\frac{3+\gamma}{2+\gamma}\Big)^{d/2}\Big(\frac{1+\gamma}{2+\gamma}\Big)^{d/2} \\ &\quad - 4\exp\Big(-\frac{2+\gamma}{2(1+\gamma)(3+\gamma)}\|\mu\|_2^2\Big) - \Big(\frac{3+\gamma}{2+\gamma}\Big)^{d/2}\Big(\frac{1+\gamma}{2+\gamma}\Big)^{d/2}\exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big) \\ &\quad + 4\Big(\frac{3+\gamma}{2+\gamma}\Big)^{d/2}\Big(\frac{1+\gamma}{2+\gamma}\Big)^{d/2}\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big) \\ C \;&:= 2\Big(\frac{\gamma}{1+\gamma}\Big)^{d/2}\Big(\frac{\gamma}{3+\gamma}\Big)^{d/2}\;. \end{aligned}$$

This implies that $\sigma^2_{\text{full}}/\sigma^2_{\text{cond}} = B/A$, so it suffices to calculate the leading terms in $A$ and $B$, respectively. We first write $\frac{(3+\gamma)(1+\gamma)}{(2+\gamma)^2} = 1 - \frac{1}{(2+\gamma)^2} =: 1-\alpha$ and $\frac{(3+\gamma)(1+\gamma)}{(4+\gamma)\gamma} = 1 + \frac{3}{\gamma(4+\gamma)} =: 1+\beta$, where $\alpha$ and $\beta$ are small as $\gamma = \omega(1)$ by assumption. Rearranging $A$ gives

$$\begin{aligned} A &= 1 + \exp\Big(-\frac{1}{3+\gamma}\|\mu\|_2^2\Big) - 2\exp\Big(-\frac{2+\gamma}{2(1+\gamma)(3+\gamma)}\|\mu\|_2^2\Big) \\ &\quad + (1-\alpha)^{d/2}\Big(2\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big) - 1 - \exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big)\Big) \\ &=: A_1 + (1-\alpha)^{d/2} A_2 \;, \end{aligned}$$

and similarly,

$$\begin{aligned} B &= (1+\beta)^{d/2}\Big(1 + \exp\Big(-\frac{1}{4+\gamma}\|\mu\|_2^2\Big)\Big) - 4\exp\Big(-\frac{2+\gamma}{2(1+\gamma)(3+\gamma)}\|\mu\|_2^2\Big) \\ &\quad - (1-\alpha)^{d/2}\Big(1 + \exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big) - 4\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big)\Big) \\ &=: (1+\beta)^{d/2} B_1 - B_2 - (1-\alpha)^{d/2} B_3 \;. \end{aligned}$$

These expressions can be simplified further depending on the relative growth rates of $d, \gamma$ and $\|\mu\|_2^2$; we consider these cases individually.

**Case 1: $\gamma = o(d^{1/2})$ and $\gamma = o(\|\mu\|_2^2)$.** Since $\gamma = o(\|\mu\|_2^2)$, all exponential terms of the form $\exp\Big(-\frac{1}{a(b+\gamma)}\|\mu\|_2^2\Big)$, for any positive constants $a, b$, are $o(1)$. Moreover, since we have assumed that $\gamma = \omega(1)$, we can apply a Taylor expansion to yield

$$\begin{aligned} (1-\alpha)^{d/2} &= \exp\Big(\frac{d}{2}\log\Big(1 - \frac{1}{(2+\gamma)^2}\Big)\Big) \\ &= \exp\Big(\frac{d}{2}\Big(-\frac{1}{(2+\gamma)^2} + O\Big(\frac{1}{\gamma^4}\Big)\Big)\Big) = \exp\Big(-\frac{d}{2\gamma^2} + o\Big(\frac{d}{\gamma^2}\Big)\Big) \;. \end{aligned} \tag{25}$$

Therefore, when $\gamma = o(d^{1/2})$, we have $(1-\alpha)^{d/2} = o(1)$. Thus the dominating term in $A$ is the leading constant 1 and

$$A = 1 + o(1) \;.$$

To control $B$, we first consider a similar Taylor expansion by noting that $\gamma = \omega(1)$:

$$\begin{aligned} (1+\beta)^{d/2} &= \exp\Big(\frac{d}{2}\log\Big(1+\frac{3}{\gamma(4+\gamma)}\Big)\Big) = \exp\Big(\frac{d}{2}\Big(\frac{3}{\gamma(4+\gamma)} + O\Big(\frac{1}{\gamma^4}\Big)\Big)\Big) \\ &= \exp\Big(\frac{3d}{2\gamma(4+\gamma)} + O\Big(\frac{d}{\gamma^4}\Big)\Big) = \exp\Big(\frac{3d}{2\gamma^2} + o\Big(\frac{d}{\gamma^2}\Big)\Big) \;. \end{aligned} \tag{26}$$

Since $\gamma = o(d^{1/2})$, we have that $(1+\beta)^{d/2} = \omega(1)$. All exponential terms and $(1-\alpha)^{d/2}$ are $o(1)$ by the calculations above, so

$$B = (1+\beta)^{d/2} + o((1+\beta)^{d/2}) = \Theta\Big(\exp\Big(\frac{3d}{2\gamma^2} + o\Big(\frac{d}{\gamma^2}\Big)\Big)\Big) \;.$$

Combining the results for $A$ and $B$ gives

$$\rho_d = \frac{\sigma_{\text{full}}}{\sigma_{\text{cond}}} = \frac{\sqrt{B}}{\sqrt{A}} = \Theta\Big(\exp\Big(\frac{3d}{4\gamma^2} + o\Big(\frac{d}{\gamma^2}\Big)\Big)\Big) \;.$$

**Case 2: $\gamma = o(d^{1/2})$ and $\gamma = \omega(\|\mu\|_2^2)$.** Since $\gamma = \omega(\|\mu\|_2^2)$, we can bound $A_1$ by first extracting an exponential factor and then applying two second-order Taylor expansions:

$$\begin{aligned}
A_1 \; &= \; 1 + \exp\Big(-\frac{1}{3+\gamma}\|\mu\|_2^2\Big) - 2\exp\Big(-\frac{2+\gamma}{2(1+\gamma)(3+\gamma)}\|\mu\|_2^2\Big) \\
&= 1 + \exp\Big(-\frac{1}{3+\gamma}\|\mu\|_2^2\Big)\Big(1 - 2\exp\Big(\frac{\gamma}{2(1+\gamma)(3+\gamma)}\|\mu\|_2^2\Big)\Big) \\
&= 1 + \Big(1 - \frac{\|\mu\|_2^2}{3+\gamma} + \frac{\|\mu\|_2^4}{2(3+\gamma)^2} + O\Big(\frac{\|\mu\|_2^6}{\gamma^3}\Big)\Big) \\
&\qquad\qquad \times \Big(-1 - \frac{\gamma\|\mu\|_2^2}{(1+\gamma)(3+\gamma)} - \frac{\gamma^2\|\mu\|_2^4}{4(1+\gamma)^2(3+\gamma)^2} + O\Big(\frac{\|\mu\|_2^6}{\gamma^3}\Big)\Big) \\
&= 1 - 1 + \Big(\frac{1}{3+\gamma} - \frac{\gamma}{(1+\gamma)(3+\gamma)}\Big)\|\mu\|_2^2 \\
&\quad + \Big(-\frac{1}{2(3+\gamma)^2} - \frac{\gamma^2}{4(1+\gamma)^2(3+\gamma)^2} + \frac{\gamma}{(1+\gamma)(3+\gamma)^2}\Big)\|\mu\|_2^4 + O\Big(\frac{\|\mu\|_2^6}{\gamma^3}\Big) \\
&= \frac{1}{(1+\gamma)(3+\gamma)}\|\mu\|_2^2 + \frac{-2+\gamma^2}{4(3+4\gamma+\gamma^2)^2}\|\mu\|_2^4 + O\Big(\frac{\|\mu\|_2^6}{\gamma^3}\Big) \,.
\end{aligned}$$

Note that the first term is on the order $\gamma^{-2}\|\mu\|_2^2$, the second term is on the order $\gamma^{-2}\|\mu\|_2^4$ and the third term is on the order $\gamma^{-3}\|\mu\|_2^6$. Since $\gamma^{-1}\|\mu\|_2^2 = o(1)$ and $\|\mu\|_2^2 = \Omega(1)$, the second term dominates and we get that

$$A_1 \; = \; \frac{\|\mu\|_2^4}{4\gamma^2} + o\Big(\frac{\|\mu\|_2^4}{\gamma^2}\Big) \,. \tag{27}$$

To control $A_2$, we use a similar Taylor expansion to get that

$$\begin{aligned}
A_2 \; &= 2\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big) - 1 - \exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big) \\
&= \; -1 + \exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big)\Big(2\exp\Big(\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big) - 1\Big) \\
&= \; -1 + \Big(1 - \frac{\|\mu\|_2^2}{2+\gamma} + \frac{\|\mu\|_2^4}{2(2+\gamma)^2} + O\Big(\frac{\|\mu\|_2^6}{\gamma^3}\Big)\Big)\Big(1 + \frac{\|\mu\|_2^2}{2+\gamma} + \frac{\|\mu\|_2^4}{4(2+\gamma)^2} + O\Big(\frac{\|\mu\|_2^6}{\gamma^3}\Big)\Big) \\
&= \Big(\frac{1}{4(2+\gamma)^2} + \frac{1}{2(2+\gamma)^2} - \frac{1}{(2+\gamma)^2}\Big)\|\mu\|_2^4 + O\Big(\frac{\|\mu\|_2^6}{\gamma^3}\Big) \\
&= \; -\frac{\|\mu\|_2^4}{4(2+\gamma)^2} + O\Big(\frac{\|\mu\|_2^6}{\gamma^3}\Big) \,.
\end{aligned} \tag{28}$$

In particular, we have $A_2 = O(\gamma^{-2}\|\mu\|_2^4) = O(A_1)$. Since $\gamma = o(d^{1/2})$, we have $(1-\alpha)^{d/2} = o(1)$ as before, which implies

$$A \; = \; A_1 + (1-\alpha)^{d/2}A_2 \; = \; \frac{\|\mu\|_2^4}{4\gamma^2} + o\Big(\frac{\|\mu\|_2^4}{\gamma^2}\Big) \,.$$

To control $B$, recall we have shown in Case 1 that $(1+\beta)^{d/2} = \omega(1)$ and $(1-\alpha)^{d/2} = o(1)$ for $\gamma = o(d^{1/2})$. All exponential terms are $O(1)$ and $B_1 = 2 + O(\gamma^{-1}\|\mu\|_2^2)$ by a Taylor expansion. By (26), we obtain that

$$\begin{aligned}
B \; = \; (1+\beta)^{d/2}B_1 - B_2 - (1-\alpha)^{d/2}B_3 \; &= 2(1+\beta)^{d/2} + o((1+\beta)^{d/2}) \\
&= \Theta\Big(\exp\Big(\frac{3d}{2\gamma^2} + o\Big(\frac{d}{\gamma^2}\Big)\Big)\Big) \,.
\end{aligned}$$

We hence conclude that

$$\rho_d \;=\; \frac{\sigma_{\text{full}}}{\sigma_{\text{cond}}} \;=\; \frac{\sqrt{B}}{\sqrt{A}} \;= \Theta\Big(\frac{\gamma}{\|\mu\|_2^2}\exp\Big(\frac{3d}{4\gamma^2}+o\Big(\frac{d}{\gamma^2}\Big)\Big)\Big)\;.$$

**Case 3:** $\gamma = \omega(\|\mu\|_2^2)$ **and** $\gamma = \omega(d^{1/2})$**.** We first rewrite the expressions of $A$ and $B$ as

$$A \;= (A_1 + A_2) - (1-(1-\alpha)^{d/2})A_2\;, \tag{29}$$

$$B \;= (B_1 - B_2 - B_3) + ((1+\beta)^{d/2}-1)B_1 + (1-(1-\alpha)^{d/2})B_3\;. \tag{30}$$

Since $\gamma = \omega(d^{1/2})$, we can perform a further Taylor expansion on the expressions in (25) and (26):

$$(1-\alpha)^{d/2} \;=\; \exp\Big(-\frac{d}{2(2+\gamma)^2}+O\Big(\frac{d}{\gamma^4}\Big)\Big) \;=\; 1-\frac{d}{2(2+\gamma)^2}+O\Big(\frac{d}{\gamma^4}\Big)\;, \tag{31}$$

$$(1+\beta)^{d/2} \;=\; \exp\Big(\frac{3d}{2\gamma(4+\gamma)}+O\Big(\frac{d}{\gamma^4}\Big)\Big) \;=\; 1+\frac{3d}{2\gamma(4+\gamma)}+O\Big(\frac{d}{\gamma^4}\Big)\;, \tag{32}$$

On the other hand, since $\gamma^{-1}\|\mu\|_2^2$ is small, we can consider performing Taylor expansions on each exponential. By grouping the terms and extracting an appropriate exponential, we get that

$$\begin{aligned}
A_1 + A_2 \;&=\; \exp\Big(-\frac{1}{3+\gamma}\|\mu\|_2^2\Big) - \exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big)\\
&\quad - 2\exp\Big(-\frac{2+\gamma}{2(1+\gamma)(3+\gamma)}\|\mu\|_2^2\Big) + 2\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big)\\
&=\; \exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big)\Big(\exp\Big(\frac{1}{(3+\gamma)(2+\gamma)}\|\mu\|_2^2\Big)-1\Big)\\
&\quad - 2\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big)\Big(-\exp\Big(-\frac{1}{2(6+11\gamma+6\gamma^2+\gamma^3)}\|\mu\|_2^2\Big)+1\Big)\\
&=\; \frac{1}{(3+\gamma)(2+\gamma)}\|\mu\|_2^2 + o\Big(\frac{1}{(3+\gamma)(2+\gamma)}\|\mu\|_2^2\Big) + O\Big(\frac{\|\mu\|_2^2}{\gamma^3}\Big)\\
&=\; \frac{\|\mu\|_2^2}{(3+\gamma)(2+\gamma)} + o\Big(\frac{\|\mu\|_2^2}{\gamma^2}\Big)\;.
\end{aligned} \tag{33}$$

In the last line, we have used that the dominating term is of the order $\|\mu\|_2^2/\gamma^2$. For $A_2$, we recall from (28) that $A_2 = -\frac{\|\mu\|_2^4}{4\gamma^2} + o\big(\frac{\|\mu\|_2^4}{\gamma^2}\big)$. Substituting the computations into (29) and using (31), we obtain that

$$\begin{aligned}
A \;&= (A_1 + A_2) - (1-(1-\alpha)^{d/2})A_2\\
&= \frac{\|\mu\|_2^2}{(3+\gamma)(2+\gamma)} + o\Big(\frac{\|\mu\|_2^2}{\gamma^2}\Big) + \Big(\frac{d}{2(2+\gamma)^2}+O\Big(\frac{d}{\gamma^4}\Big)\Big)\Big(\frac{\|\mu\|_2^4}{4(2+\gamma)^2}+o\Big(\frac{\|\mu\|_2^4}{\gamma^2}\Big)\Big)\\
&= \Theta\Big(\frac{\|\mu\|_2^2}{\gamma^2}+\frac{d\|\mu\|_2^4}{\gamma^4}\Big)\;.
\end{aligned}$$

We use a similar argument to compute $B$. By grouping terms appropriately and performing Taylor expansions, we have

$$\begin{aligned} B_1 - B_2 - B_3 &= \exp\Big(-\frac{1}{4+\gamma}\|\mu\|_2^2\Big) - \exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big) \\ &\quad - 4\exp\Big(-\frac{2+\gamma}{2(1+\gamma)(3+\gamma)}\|\mu\|_2^2\Big) + 4\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big) \\ &= \exp\Big(-\frac{1}{4+\gamma}\|\mu\|_2^2\Big)\Big(1 - \exp\Big(-\frac{2}{(4+\gamma)(2+\gamma)}\|\mu\|_2^2\Big)\Big) \\ &\quad - 4\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big)\Big(-\exp\Big(-\frac{1}{2(6+11\gamma+6\gamma^2+\gamma^3)}\|\mu\|_2^2\Big) + 1\Big), \\ &= \frac{2\|\mu\|_2^2}{(4+\gamma)(2+\gamma)} + o\Big(\frac{2\|\mu\|_2^2}{(4+\gamma)(2+\gamma)}\Big) + O\left(\frac{\|\mu\|_2^2}{\gamma^3}\right) \\ &= \frac{2\|\mu\|_2^2}{(4+\gamma)(2+\gamma)} + o\Big(\frac{\|\mu\|_2^2}{\gamma^2}\Big). \end{aligned} \tag{34}$$

By performing Taylor expansions again, we can control $B_1$ and $B_3$ as

$$B_1 = 1 + \exp\Big(-\frac{1}{4+\gamma}\|\mu\|_2^2\Big) = 2 + o(1), \tag{35}$$

$$B_3 = 1 + \exp\Big(-\frac{1}{2+\gamma}\|\mu\|_2^2\Big) - 4\exp\Big(-\frac{1}{2(2+\gamma)}\|\mu\|_2^2\Big) = -2 + o(1). \tag{36}$$

Substituting the bounds into (30) and using the bounds in (31) and (32), we obtain that

$$\begin{aligned} B &= (B_1 - B_2 - B_3) + ((1+\beta)^{d/2} - 1)B_1 + (1 - (1-\alpha)^{d/2})B_3 \\ &= \frac{2\|\mu\|_2^2}{(4+\gamma)(2+\gamma)} + o\left(\frac{\|\mu\|_2^2}{\gamma^2}\right) + \left(\frac{3d}{2\gamma(4+\gamma)} + O\left(\frac{d}{\gamma^4}\right)\right)(2 + o(1)) \\ &\quad + \left(\frac{d}{2(2+\gamma)^2} + O\left(\frac{d}{\gamma^4}\right)\right)(-2 + o(1)) \\ &= \frac{2\|\mu\|_2^2}{(4+\gamma)(2+\gamma)} + o\left(\frac{\|\mu\|_2^2}{\gamma^2}\right) + \frac{2d(6+4\gamma+\gamma^2)}{\gamma(2+\gamma)^2(4+\gamma)} + o\left(\frac{d}{\gamma^2}\right) = \Theta\Big(\frac{\|\mu\|_2^2}{\gamma^2} + \frac{d}{\gamma^2}\Big). \end{aligned}$$

The variance ratio can therefore be bounded as

$$\begin{aligned} \rho_d = \frac{\sigma_{\text{full}}}{\sigma_{\text{cond}}} = \frac{\sqrt{B}}{\sqrt{A}} = \Theta\Big(\Big(\frac{\gamma^{-2}\|\mu\|_2^2 + \gamma^{-2}d}{\gamma^{-2}\|\mu\|_2^2 + \gamma^{-4}d\|\mu\|_2^4}\Big)^{1/2}\Big) &= \Theta\Big(\frac{(\|\mu\|_2^2 + d)^{1/2}}{(\|\mu\|_2^2 + \gamma^{-2}d\|\mu\|_2^4)^{1/2}}\Big) \\ &\overset{(a)}{=} \Theta\Big(\frac{\|\mu\|_2 + d^{1/2}}{\|\mu\|_2 + \gamma^{-1}d^{1/2}\|\mu\|_2^2}\Big). \end{aligned}$$

In $(a)$, we have noted that for $a, b > 0$, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and, by the concavity of the square-root function, $\sqrt{a+b} = \sqrt{\frac{1}{2}(2a) + \frac{1}{2}(2b)} \geq \frac{1}{\sqrt{2}}(\sqrt{a} + \sqrt{b})$.

**Case 4:** $\gamma = \omega(\|\mu\|_2^2)$ **and** $\gamma = \Theta(d^{1/2})$**.** We can directly make use of the computations from Case 2 and 3 except that we control $(1-\alpha)^{d/2}$ and $(1+\beta)^{d/2}$ differently. Since

$$0 \leq (1-\alpha)^{d/2} = (1 - \frac{1}{(2+\gamma)^2})^{d/2} \leq 1,$$

we see that $A = (A_1 + A_2) - (1-(1-\alpha)^{d/2})A_2$ takes value between $A_1 + A_2$ and $A_1$, whose Taylor expansions under $\gamma = \omega(\|\mu\|_2^2)$ have been obtained in (33) and (27) respectively. Therefore,

$$A \;=\; \Theta\Big((A_1 + A_2) + A_1\Big) \;=\; \Theta\Big(\frac{\|\mu\|_2^4}{\gamma^2}\Big) \;.$$

To compute $B$, we first recall the Taylor expansion from (26) using $\gamma = \omega(1)$ and additionally make use of $\gamma = \Theta(d^{1/2})$ to get

$$(1+\beta)^{d/2} \;=\; \exp\left(\frac{3d}{2\gamma^2} + o\left(\frac{d}{\gamma^2}\right)\right) \;=\; \exp\big(\Theta(1)\big) \;=\; O(1) \;.$$

By using the expressions from (34), (35) and (36), we get that

$$\begin{aligned} B \;&= (B_1 - B_2 - B_3) + ((1+\beta)^{d/2} - 1)B_1 + (1-(1-\alpha)^{d/2})B_3 \\ &= \frac{2\|\mu\|_2^2}{(4+\gamma)(2+\gamma)} + o\Big(\frac{\|\mu\|_2^2}{\gamma^2}\Big) + ((1+\beta)^{d/2} - 1)(2 + o(1)) + (1-(1-\alpha)^{d/2})(-2 + o(1)) \\ &= \Theta\Big(\frac{\|\mu\|_2^2}{\gamma^2} + ((1+\beta)^{d/2} + (1-\alpha)^{d/2} - 2)\Big) \\ &= O(\gamma^{-2}\|\mu\|_2^2 + 1) \;=\; O(1) \;. \end{aligned}$$

In the last equality, we have noted that $\gamma^{-2}\|\mu\|_2^2 = o(\gamma^{-1}) = o(1)$ by assumption. By additionally noting that $\gamma = \Theta(d^{1/2})$, the variance ratio can therefore be bounded as

$$\rho_d = \frac{\sigma_{\text{full}}}{\sigma_{\text{cond}}} = \frac{\sqrt{B}}{\sqrt{A}} = O\Big(\frac{1}{\gamma^{-1}\|\mu\|_2^2}\Big) = O\Big(\frac{d^{1/2}}{\|\mu\|_2^2}\Big) \;.$$

This completes the proof.

## Appendix F. Proofs for Appendix A

### F.1. Proofs for RBF decomposition and verifying Assumption 2

In this section, we prove Lemma 20, Lemma 21 and Lemma 23.

**Proof of Lemma 20** We first focus on the one-dimensional RBF kernel, denoted as $\kappa_1$, which can be expressed for $x, x' \in \mathbb{R}$ as

$$|\kappa_1(x,x')| \;=\; \big|\exp(-(x-x')^2/(2\gamma))\big| \;=\; \left|\exp\left(\frac{xx'}{\gamma}\right)e^{-x^2/(2\gamma)}e^{-(x')^2/(2\gamma)}\right| \;.$$

By applying a Taylor expansion around 0 to the infinitely differentiable function $z \mapsto \exp(\frac{z}{\gamma})$ for $z \in \mathbb{R}$, we obtain that for any $K \in \mathbb{N}$ and every $x, x' \in \mathbb{R}$.

$$\begin{aligned} &\Big|\kappa_1(x,x') - \textstyle\sum_{k=0}^K \frac{1}{k!}\Big(\frac{xx'}{\gamma}\Big)^k e^{-x^2/(2\gamma)}e^{-(x')^2/(2\gamma)}\Big| \\ &\qquad\qquad \le \sup_{z\in[0,xx']}\Big|\frac{1}{(K+1)!}\Big(\frac{xx'}{\gamma}\Big)^{K+1}e^{z/\gamma}\Big|\; e^{-x^2/(2\gamma)}e^{-(x')^2/(2\gamma)} \;. \end{aligned}$$

Fix $\nu \in (2,4]$. Consider two independent normal random variables $U \sim \mathcal{N}(b_1, 1)$ and $V \sim \mathcal{N}(b_2, 1)$ for some $b_1, b_2 \in \mathbb{R}$, and recall that $\phi_k^*(x) := x^k e^{-x^2/(2\gamma)}$ and $\lambda_k^* := \frac{1}{k!\,\gamma^k}$. The above then implies that

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\kappa_1(U,V) - \textstyle\sum_{k=0}^{K} \lambda_k^* \phi_k^*(U)\phi_k^*(V)\Big|^{\nu}\Big] \\
&\qquad \le \mathbb{E}\Big[\sup_{z\in[0,UV]} \Big|\frac{1}{(K+1)!}\Big(\frac{UV}{\gamma}\Big)^{K+1} e^{z/\gamma}\Big|^{\nu}\; e^{-\nu U^2/(2\gamma)} e^{-\nu V^2/(2\gamma)}\Big] \\
&\qquad = \frac{1}{((K+1)!\,\gamma^{K+1})^{\nu}}\, \mathbb{E}\big[|UV|^{\nu(K+1)} e^{-\nu U^2/(2\gamma) - \nu V^2/(2\gamma) + \sup_{z\in[0,UV]} \nu z/\gamma}\big] \\
&\qquad \le \frac{1}{((K+1)!\,\gamma^{K+1})^{\nu}}\, \mathbb{E}\big[|UV|^{\nu(K+1)} e^{-\nu(|U|-|V|)^2/(2\gamma)}\big] \\
&\qquad \le \frac{1}{((K+1)!\,\gamma^{K+1})^{\nu}}\, \mathbb{E}\big[|U|^{\nu(K+1)}\big]\, \mathbb{E}\big[|V|^{\nu(K+1)}\big] \;.
\end{aligned}$$

In the last inequality, we have noted that $U$ and $V$ are independent and bounded the exponential term from above by 1. By the formula of absolute moments of a Gaussian, we get that

$$\mathbb{E}\big[|U-b_1|^{\nu(K+1)}\big] \;=\; \mathbb{E}\big[|V-b_2|^{\nu(K+1)}\big] \;=\; \frac{2^{(\nu K)/2}}{\sqrt{\pi}}\Gamma\Big(\frac{\nu K+1}{2}\Big) \;.$$

By a Jensen's inequality applied to the convex function $x \mapsto |x|^{\nu(K+1)}$, we get that

$$\begin{aligned}
&\mathbb{E}\big[|U|^{\nu(K+1)}\big] \;=\; \mathbb{E}\big[|b_1 + (U-b_1)|^{\nu(K+1)}\big] \;=\; \mathbb{E}\Big[\Big|\frac{1}{2}(2b_1) + \frac{1}{2}(2(U-b_1))\Big|^{\nu(K+1)}\Big] \\
&\le 2^{\nu(K+1)-1}\big(b^{\nu(K+1)} + \mathbb{E}[|U-b_1|^{\nu(K+1)}]\big) \;=\; \frac{(2b_1)^{\nu(K+1)}}{2} + \frac{2^{\frac{3}{2}\nu(K+1)}}{2\sqrt{\pi}}\Gamma\Big(\frac{\nu(K+1)+1}{2}\Big) \;.
\end{aligned}$$

Similarly, we get that

$$\mathbb{E}\big[|V|^{\nu(K+1)}\big] \;\le\; \frac{(2b_2)^{\nu(K+1)}}{2} + \frac{2^{\frac{3}{2}\nu(K+1)}}{2\sqrt{\pi}}\Gamma\Big(\frac{\nu(K+1)+1}{2}\Big) \;. \tag{37}$$

Substituting these moment bounds and noting that $(K+1)! = \Gamma(K+2)$, we get that

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\kappa_1(U,V) - \textstyle\sum_{k=0}^{K} \lambda_k^* \phi_k^*(U)\phi_k^*(V)\Big|^{\nu}\Big] \\
&\qquad \le \frac{1}{\gamma^{\nu(K+1)}\big(\Gamma(K+2)\big)^{\nu}}\Big(\frac{(2b_1)^{\nu(K+1)}}{2} + \frac{2^{\frac{3}{2}\nu(K+1)}}{2\sqrt{\pi}}\Gamma\Big(\frac{\nu(K+1)+1}{2}\Big)\Big) \\
&\qquad\quad \times \Big(\frac{(2b_2)^{\nu(K+1)}}{2} + \frac{2^{\frac{3}{2}\nu(K+1)}}{2\sqrt{\pi}}\Gamma\Big(\frac{\nu(K+1)+1}{2}\Big)\Big) \\
&\qquad =: T\,(A_1+B)(A_2+B) \;.
\end{aligned}$$

As $K$ grows, the dominating terms are the Gamma functions, so we only need to control their ratios. By Stirling's formula for the gamma function, we have $\Gamma(x) = \sqrt{2\pi}\, x^{x-1/2} e^{-x}\big(1 + O(x^{-1})\big)$ for $x > 0$. This immediately implies that

$$T A_1 A_2 \;=\; \Theta\Big(\frac{(4b_1b_2/\gamma)^{\nu(K+1)}}{(K+2)^{\nu(K+3/2)} e^{-\nu(K+2)}}\Big) \;=\; o(1)$$

as $K \to \infty$. Meanwhile,

$$\frac{\Gamma\big(\frac{\nu(K+1)+1}{2}\big)}{\big(\Gamma(K+2)\big)^{\nu}} \;=\Theta\Big(\frac{K^{\nu K/2}}{K^{\nu K}}\Big) \;=\; \Theta\Big(K^{-\nu K/2}\Big) \;,$$

which implies that

$$TA_1B \;=\; \Theta\Big((4\sqrt{2}b_1/\gamma)^{\nu K} K^{-\nu K/2}\Big) \;=\; o(1)\;,$$

since the dominating term is $K^{-\nu K/2}$. Similarly, $TA_2B = o(1)$. On the other hand, another application of Stirling's formula gives that

$$\begin{aligned}\frac{\big(\Gamma\big(\frac{\nu(K+1)+1}{2}\big)\big)^2}{\big(\Gamma(K+2)\big)^{\nu}} &= (2\pi)^{-(\nu-2)/2}\,\frac{\big(\frac{\nu(K+1)+1}{2}\big)^{\nu(K+1)}}{(K+2)^{\nu(K+3/2)}}\,e^{-\nu(K+1)-1+\nu(K+2)}\,\frac{\big(1+O(K^{-1})\big)^2}{\big(1+O(K^{-1})\big)^{\nu}}\\ &= \Theta\Big(\frac{(\nu/2)^{\nu K}K^{\nu K}}{K^{\nu(K+3/2)}}\Big) \;=\; \Theta\big((\nu/2)^{\nu K}K^{-3\nu/2}\big)\;.\end{aligned}$$

This implies that

$$TB^2 \;=\; \Theta\big((8/\gamma)^{\nu K}(\nu/2)^{\nu K}K^{-3\nu/2}\big) \;=\; \Theta\big((2\nu/\gamma)^{\nu K}K^{-3\nu/2}\big) \;=\; o(1)\;,$$

where we have recalled that $\nu \leq 4$ and used the assumption that $\gamma > 8$. In summary, we have proved that for $\nu \in (2,4]$ and *any* fixed $b_1, b_2 \in \mathbb{R}$,

$$\mathbb{E}\Big[\Big|\kappa_1(U,V) - \textstyle\sum_{k=0}^{K}\lambda_k^*\phi_k^*(U)\phi_k^*(V)\Big|^{\nu}\Big] \;\leq\; T(A_1+B)(A_2+B) \;\xrightarrow{K\to\infty}\; 0\;.$$

To extend this to multiple dimensions, we note that for the vectors $\mathbf{x} = (x_1,\ldots,x_d) \in \mathbb{R}^d$ and $\mathbf{x}' = (x_1,\ldots,x_d) \in \mathbb{R}^d$, the multi-dimensional RBF kernel can then be expressed as

$$\kappa(\mathbf{x},\mathbf{x}') \;=\; \exp\big(-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\gamma)\big) \;=\; \textstyle\prod_{l=1}^{d}\exp\big(-(x_l-x_l')^2/(2\gamma)\big) \;=\; \prod_{l=1}^{d}\kappa_1(x_l,x_l')\;.$$

Recall that we have defined the independent normal vectors $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, I_d)$ and $\mathbf{V} \sim \mathcal{N}(\mu, I_d)$. Let $U_1,\ldots,U_d$ be the coordinates of $\mathbf{U}$ and $V_1,\ldots,V_d$ be those of $\mathbf{V}$, which are all independent since the covariance matrices are $I_d$. For $0 \leq l \leq d$ and $K \in \mathbb{N}$, define the random quantities

$$S_{j;K} \;\coloneqq\; \textstyle\sum_{k=0}^{K}\lambda_k^*\phi_k^*(U_j)\phi_k^*(V_j) \quad\text{ and }\quad W_{l;K} \;\coloneqq\; \Big(\prod_{j=1}^{l}\kappa_1(U_j,V_j)\Big)\Big(\prod_{j=l+1}^{d}S_{j;K}\Big)\;.$$

In particular $\kappa(\mathbf{U},\mathbf{V}) = W_{d;K}$. Now by expanding a telescoping sum and applying a triangle inequality followed by a Jensen's inequality, we have

$$\begin{aligned}&\mathbb{E}\big[|\kappa(\mathbf{U},\mathbf{V}) - W_{0;K}|^{\nu}\big] \;=\; \mathbb{E}\Big[\Big|\textstyle\sum_{l=1}^{d}(W_{l;K}-W_{l-1;K})\Big|^{\nu}\Big]\\ &\leq \mathbb{E}\Big[\Big(\textstyle\sum_{l=1}^{d}|W_{l;K}-W_{l-1;K}|\Big)^{\nu}\Big]\\ &\leq d^{\nu-1}\textstyle\sum_{l=1}^{d}\mathbb{E}[|W_{l;K}-W_{l-1;K}|^{\nu}]\\ &= d^{\nu-1}\textstyle\sum_{l=1}^{d}\Big(\prod_{j=1}^{l-1}\mathbb{E}[|\kappa_1(U_j,V_j)|^{\nu}]\Big)\mathbb{E}[|\kappa_1(U_l,V_l)-S_{l;K}|^{\nu}]\Big(\prod_{j=l+1}^{d}\mathbb{E}[|S_{j;K}|^{\nu}]\Big)\;.\end{aligned}$$

In the last equality, we have used the independence of $U_j$'s and $V_j$'s. To bound the summands, we first note that $\kappa_1$ is uniformly bounded in norm by 1, which implies that $\mathbb{E}[|\kappa_1(U_j,V_j)|^{\nu}] \leq 1$. By the previous result, $\mathbb{E}[|\kappa_1(U_l,V_l) - S_{l;K}|^{\nu}] = o(1)$ as $K \to \infty$. By a triangle inequality and a Jensen's inequality, we have that

$$\begin{aligned}\mathbb{E}[|S_{j;K}|^{\nu}] \;&\leq \mathbb{E}\big[\big||\kappa_1(U_j,V_j)| + |S_{j;K}-\kappa_1(U_j,V_j)|\big|^{\nu}\big]\\ &\leq 2^{\nu-1}\mathbb{E}\big[|\kappa_1(U_j,V_j)|^{\nu}\big] + 2^{\nu-1}\mathbb{E}\big[|S_{j;K}-\kappa_1(U_j,V_j)|^{\nu}\big] \;\leq\; 2^{\nu-1}+o(1)\;.\end{aligned}$$

This implies that each summand satisfies

$$\Big(\prod_{j=1}^{l-1} \mathbb{E}[|\kappa_1(U_j,V_j)|^\nu]\Big)\mathbb{E}[|\kappa_1(U_l,V_l)-S_{l;K}|^\nu]\Big(\prod_{j=l+1}^{d}\mathbb{E}[|S_{j;K}|^\nu]\Big) \;=\; o(1)$$

as $K \to \infty$. Since $d$ is not affected by $K$, we have shown the desired result

$$\mathbb{E}\Big[\Big|\kappa(\mathbf{U},\mathbf{V}) - \prod_{j=1}^{d}\Big(\sum_{k=0}^{K}\lambda_k^*\phi_k^*(U_j)\phi_k^*(V_j)\Big)\Big|^\nu\Big] \;=\; \mathbb{E}\big[|\kappa(\mathbf{U},\mathbf{V}) - W_{0;K}|^\nu\big] \;\xrightarrow{K\to\infty}\; 0\;.$$

■

**Proof of Lemma 21** We first rewrite $u_P^{\mathrm{KSD}}$ as

$$\begin{aligned}
u_P^{\mathrm{KSD}}(\mathbf{x},\mathbf{x}') &= e^{-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\gamma)}\Big(\mathbf{x}^\top\mathbf{x}' - \frac{\gamma+1}{\gamma^2}\|\mathbf{x}-\mathbf{x}'\|_2^2 + \frac{d}{\gamma}\Big)\\
&= e^{-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\gamma)}\Big(-\frac{\gamma+1}{\gamma^2}(\|\mathbf{x}\|_2^2+\|\mathbf{x}'\|_2^2) + \frac{\gamma^2+2\gamma+2}{\gamma^2}\mathbf{x}^\top\mathbf{x}' + \frac{d}{\gamma}\Big)\\
&= e^{-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\gamma)}\Big(-\frac{\gamma+1}{\gamma^2}(\|\mathbf{x}\|_2^2+1)(\|\mathbf{x}'\|_2^2+1) + \frac{\gamma+1}{\gamma^2}\|\mathbf{x}\|_2^2\|\mathbf{x}'\|_2^2\\
&\qquad + \frac{\gamma^2+2\gamma+2}{\gamma^2}\sum_{l=1}^{d} x_l x_l' + \Big(\frac{d}{\gamma}+\frac{\gamma+1}{\gamma^2}\Big)\Big)\;.
\end{aligned}$$

For $K' \in \mathbb{N}$, write $S_{K'} \coloneqq \sum_{k'=1}^{K'} \alpha_{k'}\psi_{k'}(\mathbf{X}_1)\psi_{k'}(\mathbf{X}_2)$, and define the following random variables comparing each set of eigenvalue and eigenfunction to the corresponding term in $u_P^{\mathrm{KSD}}$:

$$\begin{aligned}
T_{K';1} &= \sum_{k'=1}^{K'}\lambda_{(k'-1)(d+3)+1}\,\phi_{(k'-1)(d+3)+1}(\mathbf{X}_1)\,\phi_{(k'-1)(d+3)+1}(\mathbf{X}_1)\\
&\qquad - e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)}\Big(-\frac{\gamma+1}{\gamma^2}(\|\mathbf{X}_1\|_2^2+1)(\|\mathbf{X}_2\|_2^2+1)\Big)\\
&= -\frac{\gamma+1}{\gamma^2}(\|\mathbf{X}_1\|_2^2+1)(\|\mathbf{X}_2\|_2^2+1)S_{K'}\;,\\
T_{K';2} &= \sum_{k'=1}^{K'}\lambda_{(k'-1)(d+3)+2}\,\phi_{(k'-1)(d+3)+2}(\mathbf{X}_1)\,\phi_{(k'-1)(d+3)+2}(\mathbf{X}_1)\\
&\qquad - e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)}\Big(\frac{\gamma+1}{\gamma^2}\|\mathbf{X}_1\|_2^2\|\mathbf{X}_2\|_2^2\Big)\\
&= \frac{\gamma+1}{\gamma^2}\|\mathbf{X}_1\|_2^2\|\mathbf{X}_2\|_2^2\,S_{K'}\;,
\end{aligned}$$

$$\begin{aligned}
T_{K';3} &= \sum_{k'=1}^{K'}\lambda_{(k'-1)(d+3)+3}\,\phi_{(k'-1)(d+3)+3}(\mathbf{X}_1)\,\phi_{(k'-1)(d+3)+3}(\mathbf{X}_1)\\
&\qquad - e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)}\Big(\frac{d}{\gamma}+\frac{\gamma+1}{\gamma^2}\Big)\\
&= \Big(\frac{d}{\gamma}+\frac{\gamma+1}{\gamma^2}\Big)S_{K'}\;,\\
T_{K';3+l} &= \sum_{k'=1}^{K'}\lambda_{(k'-1)(d+3)+3+l}\,\phi_{(k'-1)(d+3)+3+l}(\mathbf{X}_1)\,\phi_{(k'-1)(d+3)+3+l}(\mathbf{X}_1)\\
&\qquad - e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)}\Big(\frac{\gamma^2+2\gamma+2}{\gamma^2}(\mathbf{X}_1)_l(\mathbf{X}_2)_l\Big)\\
&= \Big(\frac{\gamma^2+2\gamma+2}{\gamma^2}(\mathbf{X}_1)_l(\mathbf{X}_2)_l\Big)\,S_{K'}
\end{aligned}$$

for $l = 1, \dots, d$, where we have denoted the $l$-th coordinates of $\mathbf{X}_1$ and $\mathbf{X}_2$ by $(\mathbf{X}_1)_l$ and $(\mathbf{X}_2)_l$ respectively. We now bound the approximation error with $(d+3)K'$ summands for $K' \in \mathbb{N}$ and $\nu \in (2, 3]$. Fix some $\nu_1 \in (\nu, 4]$ and let $\nu_2 = 1/(\nu^{-1} - \nu_1^{-1})$. By using the quantites defined above, a Jensen's inequality to the convex function $x \mapsto |x|^\nu$ and a Hölder's inequality to each $\mathbb{E}[|T_{K';l}|^\nu]$, we have

$$\begin{aligned}
&\mathbb{E}\big[\big|\textstyle\sum_{k=1}^{(d+3)K'} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\big|^\nu\big] \\
&= \mathbb{E}\big[\big|\textstyle\sum_{l=1}^{d+3} T_{K';l}\big|^\nu\big] \\
&\leq (d+3)^{\nu-1} \textstyle\sum_{l=1}^{d+3} \mathbb{E}[|T_{K';l}|^\nu] \\
&\leq (d+3)^{\nu-1} \mathbb{E}[|S_{K'}|^{\nu_1}]^{\nu/\nu_1} \Big( \Big(\frac{\gamma+1}{\gamma^2}\Big)^\nu \mathbb{E}[(\|\mathbf{X}_1\|_2^2 + 1)^{\nu_2}]^{\nu/\nu_2}\, \mathbb{E}[(\|\mathbf{X}_2\|_2^2 + 1)^{\nu_2}]^{\nu/\nu_2} \\
&\qquad + \Big(\frac{\gamma+1}{\gamma^2}\Big)^\nu \mathbb{E}\big[\|\mathbf{X}_1\|_2^{2\nu_2}\big]^{\nu/\nu_2} \mathbb{E}\big[\|\mathbf{X}_2\|_2^{2\nu_2}\big]^{\nu/\nu_2} + \Big(\frac{d}{\gamma} + \frac{\gamma+1}{\gamma^2}\Big)^\nu \\
&\qquad + \textstyle\sum_{l=1}^{d} \Big(\frac{\gamma^2 + 2\gamma + 2}{\gamma^2}\Big)^\nu \mathbb{E}\big[|(\mathbf{X}_1)_l|^{\nu_2}\big]^{\nu/\nu_2} \mathbb{E}\big[|(\mathbf{X}_2)_l|^{\nu_2}\big]^{\nu/\nu_2} \Big) .
\end{aligned}$$

The only $K'$-dependence above comes from $\mathbb{E}[|S_{K'}|^{\nu_1}]^{\nu/\nu_1} = \|S_{K'}\|_{L_{\nu_1}}^\nu$, which converges to 0 as $K'$ grows by Lemma 20. Therefore

$$\mathbb{E}\big[\big|\textstyle\sum_{k=1}^{(d+3)K'} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\big|^\nu\big] \xrightarrow{K' \to \infty} 0 .$$

Now for $K \in \mathbb{N}$ not necessarily divisible by $d+3$, we let $K'$ be the largest integer such that $dK' \leq K$. By a triangle inequality and a Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\big[\big|\textstyle\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\big|^\nu\big] \\
&\leq \mathbb{E}\Big[\Big(\big|\textstyle\sum_{k=1}^{(d+3)K'} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\big| + \big|\textstyle\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|\Big)^\nu\Big] \\
&\leq 2^{\nu-1} \mathbb{E}\big[\big|\textstyle\sum_{k=1}^{(d+3)K'} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\big|^\nu\big] \\
&\quad + 2^{\nu-1} \mathbb{E}\big[\big|\textstyle\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|^\nu\big] .
\end{aligned}$$

The goal is to show that the bound converges to 0 as $K$ grows. We have already shown that the first term is $o(1)$, so we focus on the second term. The expectation in the second term can be bounded using a Jensen's inequality as

$$\begin{aligned}
\mathbb{E}\big[\big|\textstyle\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|^\nu\big] \;&\leq\; \mathbb{E}\big[\big(\textstyle\sum_{k=(d+3)K'+1}^{K} |\lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)|\big)^\nu\big] \\
&\leq (K - (d+3)K')^{\nu-1} \textstyle\sum_{k=(d+3)K'+1}^{K} \mathbb{E}\big[\big(\lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big)^\nu\big] \\
&\leq d^\nu \sup_{k \in \{(d+3)K'+1, \dots, (d+3)K'+(d+3)\}} \mathbb{E}\big[\big(\lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big)^\nu\big] \\
&= d^\nu \sup_{1 \leq l \leq d+3} \mathbb{E}\big[\big(\lambda_{(d+3)K'+l} \phi_{(d+3)K'+l}(\mathbf{X}_1)\phi_{(d+3)K'+l}(\mathbf{X}_2)\big)^\nu\big] .
\end{aligned}$$

By observing the formula for $\lambda_k$ and $\phi_k$, we see that there exists some $K$-independent constant $C_{d,\gamma}$ such that for $1 \leq l \leq d+3$,

$$|\lambda_{(d+3)K'+l}| \;\leq\; C_{d,\gamma} \alpha_{K'+1} \qquad \text{and} \qquad |\phi_{(d+3)K'+l}| \;\leq\; C_{d,\gamma} \psi_{K'+1}(\mathbf{x})(\|\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2 + 1) .$$

This allows us to obtain the bound

$$
\begin{aligned}
&\mathbb{E}\big[\big|\textstyle\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|^\nu\big] \\
&\leq d^\nu C_{d,\gamma}^2 \alpha_{K'+1}^\nu \mathbb{E}[(\psi_{K'+1}(\mathbf{X}_1)\psi_{K'+1}(\mathbf{X}_2))^\nu\,(\|\mathbf{X}_1\|_2^2+\|\mathbf{X}_1\|_2+1)^\nu(\|\mathbf{X}_2\|_2^2+\|\mathbf{X}_2\|_2+1)^\nu] \\
&\overset{(a)}{=} d^\nu C'_{d,\gamma}\Big(\textstyle\prod_{l=1}^d \lambda^*_{[g_d(K'+1)]_l}\Big)^\nu \mathbb{E}\Big[\Big(\textstyle\prod_{l=1}^d \phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_1)_l\big)\,\phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_2)_l\big)\big)\Big)^\nu \\
&\qquad\qquad\qquad\qquad\qquad \times (\|\mathbf{X}_1\|_2^2+\|\mathbf{X}_1\|_2+1)^\nu(\|\mathbf{X}_2\|_2^2+\|\mathbf{X}_2\|_2+1)^\nu\Big] \\
&\overset{(b)}{\leq} d^\nu C'_{d,\gamma}\mathbb{E}\big[(\|\mathbf{X}_1\|_2^2+\|\mathbf{X}_1\|_2+1)^{2\nu}(\|\mathbf{X}_2\|_2^2+\|\mathbf{X}_2\|_2+1)^{2\nu}\big]^{1/2} \\
&\quad\times\Big(\textstyle\prod_{l=1}^d \lambda^*_{[g_d(K'+1)]_l}\Big)^\nu \mathbb{E}\Big[\Big(\textstyle\prod_{l=1}^d \phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_1)_l\big)\,\phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_2)_l\big)\big)\Big)^{2\nu}\Big]^{1/2} \\
&\overset{(c)}{=} d^\nu C'_{d,\gamma}\mathbb{E}\big[(\|\mathbf{X}_1\|_2^2+\|\mathbf{X}_1\|_2+1)^{2\nu}(\|\mathbf{X}_2\|_2^2+\|\mathbf{X}_2\|_2+1)^{2\nu}\big]^{1/2} \\
&\quad\times\textstyle\prod_{l=1}^d \big(\lambda^*_{[g_d(K'+1)]_l}\big)^\nu\,\Big(\textstyle\prod_{l=1}^d \mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_1)_l\big)\big)^{2\nu}\big]\mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_2)_l\big)\big)^{2\nu}\big]\Big)^{1/2} \\
&\overset{(d)}{=} d^\nu C'_{d,\gamma}\mathbb{E}\big[(\|\mathbf{X}_1\|_2^2+\|\mathbf{X}_1\|_2+1)^{2\nu}(\|\mathbf{X}_2\|_2^2+\|\mathbf{X}_2\|_2+1)^{2\nu}\big]^{1/2} \\
&\quad\times\textstyle\prod_{l=1}^d \Big(\big(\lambda^*_{[g_d(K'+1)]_l}\big)^\nu\,\mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_1)_l\big)\big)^{2\nu}\big]\Big)\,,
\end{aligned}
$$

where we have used the definitions of $\alpha_k$ and $\psi_k$ from (7) in $(a)$, a Cauchy-Schwarz inequality in $(b)$, the independence of $(\mathbf{X}_1)_l$ and $(\mathbf{X}_2)_l$ for $1 \leq l \leq d$ due to the identity covariance matrix in $(c)$ and finally the fact that $\mathbf{X}_1$ and $\mathbf{X}_2$ are identically distributed in $(d)$. The only quantity that depends on $K'$ now is

$$
\big(\lambda^*_{[g_d(K'+1)]_l}\big)^\nu\,\mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_1)_l\big)\big)^{2\nu}\big]
$$

for $1 \leq l \leq d$. We now seek to bound this quantity. Recall from Lemma 20 that $\lambda^*_k := \frac{1}{k!\,\gamma^k}$, and for $V \sim \mathcal{N}(b,1)$, we have

$$
\mathbb{E}[(\phi^*_k(U))^{2\nu}] \;=\; \mathbb{E}\big[|U|^{2\nu k}e^{-\nu U^2/\gamma}\big] \;\leq\; \mathbb{E}\big[|U|^{2\nu k}\big] \;\leq\; \frac{(2b)^{2\nu k}}{2} + \frac{2^{3\nu k}}{2\sqrt{\pi}}\,\Gamma\Big(\frac{2\nu k+1}{2}\Big)\,.
$$

where we have used a bound similar to (37) in the proof of Lemma 20. By Stirling's formula for the gamma function, we have $\Gamma(x) = \sqrt{2\pi}\,x^{x-1/2}e^{-x}\big(1+O(x^{-1})\big)$ for $x > 0$, which implies

$$
\begin{aligned}
(\lambda^*_k)^\nu\,\mathbb{E}[(\phi^*_k(U))^{2\nu}] \;&\leq \frac{1}{(k!)^\nu\,\gamma^{\nu k}}\Big(\frac{(2b)^{2\nu k}}{2} + \frac{2^{3\nu k}}{2\sqrt{\pi}}\,\Gamma\Big(\frac{2\nu k+1}{2}\Big)\Big) \\
&= O\Big(\Big(\frac{8}{\gamma}\Big)^{\nu k}\frac{(\nu k)^{\nu k}e^{-\nu k}}{(k+1)^{\nu(k+1/2)}e^{-\nu(k+1)}}\Big) \\
&= O\Big(\Big(\frac{8\nu}{\gamma}\Big)^{\nu k}\Big) \;=\; O\Big(\Big(\frac{24}{\gamma}\Big)^{\nu k}\Big) \;=\; o(1)
\end{aligned}
$$

as $k \to \infty$, where we have used the assumption that $\gamma > 24$. By construction of $g_d$ in (7), as $K' \to \infty$, $\min_{1 \leq l \leq d}[g_d(K'+1)]_l \to \infty$, which implies that

$$
\big(\lambda^*_{[g_d(K'+1)]_l}\big)^\nu\,\mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_1)_l\big)\big)^{2\nu}\big] \;\xrightarrow{K'\to\infty}\; 0\,.
$$

Therefore

$$\mathbb{E}\big[\big|\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|^\nu\big] \xrightarrow{K\to\infty} 0 ,$$

which finishes the proof that

$$\mathbb{E}\big[\big|\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1,\mathbf{X}_2)\big|^\nu\big] \xrightarrow{K\to\infty} 0 .$$

In other words, Assumption 2 holds. ■

**Proof of Lemma 23** Fix $\nu \in (2,3]$. Consider the independent Gaussian vectors $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d.}{\sim} P \equiv \mathcal{N}(\mathbf{0}, I_d)$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d.}{\sim} Q \equiv \mathcal{N}(\mu, I_d)$. Write $\mathbf{Z}_1 = (\mathbf{X}_1, \mathbf{Y}_1)$, $\mathbf{Z}_2 = (\mathbf{X}_2, \mathbf{Y}_2)$ and

$$T_K(\mathbf{x}, \mathbf{x}') \;:=\; e^{-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\gamma)} - \sum_{k=1}^{K} \alpha_k \psi_k(\mathbf{x})\psi_k(\mathbf{x}')$$

for $K \in \mathbb{N}$, and recall that

$$u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2) \;=\; e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)} - e^{-\|\mathbf{X}_1-\mathbf{Y}_2\|_2^2/(2\gamma)} - e^{-\|\mathbf{X}_2-\mathbf{Y}_1\|_2^2/(2\gamma)} + e^{-\|\mathbf{Y}_1-\mathbf{Y}_2\|_2^2/(2\gamma)} .$$

Then by a triangle inequality and Jensen's inequality, we get that

$$\begin{aligned}
&\mathbb{E}\Big[\Big|u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2) - \sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\Big|^\nu\Big] \\
&= \mathbb{E}\Big[\Big|u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2) - \sum_{k=1}^{K} \alpha_k \big(\psi_k(\mathbf{X}_1) - \psi_k(\mathbf{Y}_1)\big)\big(\psi_k(\mathbf{X}_2) - \psi_k(\mathbf{Y}_2)\big)\Big|^\nu\Big] \\
&= \mathbb{E}\big[\big|T_K(\mathbf{X}_1,\mathbf{X}_2) - T_K(\mathbf{X}_1,\mathbf{Y}_2) - T_K(\mathbf{X}_2,\mathbf{Y}_1) + T_K(\mathbf{X}_2,\mathbf{Y}_2)\big|^\nu\big] \\
&\leq 4^{\nu-1}\big(\mathbb{E}[|T_K(\mathbf{X}_1,\mathbf{X}_2)|^\nu] + \mathbb{E}[|T_K(\mathbf{X}_1,\mathbf{Y}_2)|^\nu] + \mathbb{E}[|T_K(\mathbf{X}_2,\mathbf{Y}_1)|^\nu] + \mathbb{E}[|T_K(\mathbf{Y}_1,\mathbf{Y}_2)|^\nu]\big) .
\end{aligned}$$

Since each expectation is taken with respect to a product of two Gaussian distributions with identity covariance matrices, by Lemma 20 and (8), they all decay to 0 as $K \to \infty$. This proves that

$$\mathbb{E}\Big[\Big|u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2) - \sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\Big|^\nu\Big] \xrightarrow{K\to\infty} 0 ,$$

and therefore Assumption 2 holds. ■

### F.2. Proof for Lemma 22

We restate the KSD U-statistic for RBF under our Gaussian mean-shift setup from (9):

$$u_P^{\mathrm{KSD}}(\mathbf{x},\mathbf{x}') \;=\; \exp\Big(-\frac{1}{2\gamma}\|\mathbf{x}-\mathbf{x}'\|_2^2\Big)\Big(\mathbf{x}^\top\mathbf{x}' - \frac{\gamma+1}{\gamma^2}\|\mathbf{x}-\mathbf{x}'\|_2^2 + \frac{d}{\gamma}\Big) . \tag{38}$$

#### F.2.1. PROOF FOR $g(\mathbf{x})$

Fix $\mathbf{x} \in \mathbb{R}^d$. Taking expectation of $u_P^{\mathrm{KSD}}(\mathbf{x}, \mathbf{X}')$ with respect to the distribution of $\mathbf{X}'$,

$$\begin{aligned}
g(\mathbf{x}) &= \mathbb{E}[u_P^{\mathrm{KSD}}(\mathbf{x},\mathbf{X}')] \\
&= \mathbb{E}\Big[\exp\Big(-\frac{1}{2\gamma}\|\mathbf{x}-\mathbf{X}'\|_2^2\Big)\Big(\mathbf{x}^\top\mathbf{X}' - \frac{1+\gamma}{\gamma^2}\|\mathbf{x}-\mathbf{X}'\|_2^2 + \frac{d}{\gamma}\Big)\Big] \\
&= \Big(\frac{\gamma}{1+\gamma}\Big)^{d/2} \exp\Big(-\frac{1}{2(1+\gamma)}\|\mathbf{x}-\mu\|_2^2\Big)\,\mathbb{E}\Big[\mathbf{x}^\top\mathbf{W}' - \frac{1+\gamma}{\gamma^2}\|\mathbf{x}-\mathbf{W}'\|_2^2 + \frac{d}{\gamma}\Big] .
\end{aligned}$$

where the third line follows by applying Lemma 28, and $\mathbf{W}' \sim \mathcal{N}\left(\frac{1}{1+\gamma}\left(\mu + \frac{1}{\gamma}\mathbf{x}\right), \frac{\gamma}{1+\gamma} I_d\right)$. The proof is completed by calculating the expectation as

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{x}^\top \mathbf{W}' - \frac{1+\gamma}{\gamma^2}\|\mathbf{x} - \mathbf{W}'\|_2^2 + \frac{d}{\gamma}\right] \\
&= \mathbb{E}\left[\mathbf{x}^\top \mathbf{W}' - \frac{1+\gamma}{\gamma^2}\left(\|\mathbf{W}'\|_2^2 - 2\mathbf{x}^\top \mathbf{W}' + \|\mathbf{x}\|_2^2\right) + \frac{d}{\gamma}\right] \\
&= \frac{\gamma}{1+\gamma}\left(\mu + \frac{1}{\gamma}\mathbf{x}\right)^\top \mathbf{x} - \frac{\gamma+1}{\gamma^2}\left(\frac{\gamma d}{1+\gamma} + \frac{\gamma^2}{(1+\gamma)^2}\left\|\mu + \frac{1}{\gamma}\mathbf{x}\right\|_2^2 - \frac{\gamma}{1+\gamma}\left(\mu + \frac{1}{\gamma}\mathbf{x}\right)^\top \mathbf{x} + \|\mathbf{x}\|_2^2\right) + \frac{d}{\gamma} \\
&= \frac{2+\gamma}{1+\gamma}\mu^\top \mathbf{x} - \frac{1}{1+\gamma}\|\mu\|_2^2 \,.
\end{aligned}$$

### F.2.2. PROOF FOR $D^{\mathrm{KSD}}(Q, P)$

Noting that $D^{\mathrm{KSD}}(Q, P) = \mathbb{E}[g^{\mathrm{KSD}}(\mathbf{x})]$, we can apply Lemma 28 again to yield

$$\begin{aligned}
D^{\mathrm{KSD}}(Q, P) &= \left(\frac{\gamma}{1+\gamma}\right)^{d/2} \mathbb{E}\left[\exp\left(-\frac{1}{2(1+\gamma)}\|\mathbf{X} - \mu\|_2^2\right)\left(\frac{2+\gamma}{1+\gamma}\mu^\top \mathbf{X} - \frac{1}{1+\gamma}\|\mu\|_2^2\right)\right] \\
&= \left(\frac{\gamma}{1+\gamma}\right)^{d/2}\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2} \mathbb{E}\left[\frac{2+\gamma}{1+\gamma}\mu^\top \mathbf{W} - \frac{1}{1+\gamma}\|\mu\|_2^2\right] ,
\end{aligned}$$

where $W \sim \mathcal{N}\left(\frac{1+\gamma}{2+\gamma}\left(\mu + \frac{1}{1+\gamma}\mu\right), \frac{1+\gamma}{2+\gamma} I_d\right)$. We then have

$$\begin{aligned}
D^{\mathrm{KSD}}(Q, P) &= \left(\frac{\gamma}{1+\gamma}\right)^{d/2}\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2}\left(\mu^\top\left(\mu + \frac{1}{1+\gamma}\mu\right) - \frac{1}{1+\gamma}\|\mu\|_2^2\right) \\
&= \left(\frac{\gamma}{2+\gamma}\right)^{d/2}\|\mu\|_2^2 ,
\end{aligned}$$

as required.

### F.2.3. PROOF FOR $\sigma^2_{\mathrm{cond}}$

We first calculate the second moment as

$$\begin{aligned}
\mathbb{E}[g(\mathbf{X})^2] &= \left(\frac{\gamma}{1+\gamma}\right)^{d} \mathbb{E}\left[\exp\left(-\frac{1}{1+\gamma}\|\mathbf{X} - \mu\|_2^2\right)\left(\frac{2+\gamma}{1+\gamma}\mu^\top \mathbf{X} - \frac{1}{1+\gamma}\|\mu\|_2^2\right)^2\right] \\
&= \left(\frac{\gamma}{1+\gamma}\right)^{d}\left(\frac{1+\gamma}{3+\gamma}\right)^{d/2} \mathbb{E}\left[\left(\frac{2+\gamma}{1+\gamma}\mu^\top \mathbf{W} - \frac{1}{1+\gamma}\|\mu\|_2^2\right)^2\right]
\end{aligned}$$

where in the last line we have applied Lemma 28 while setting $\mathbf{W} \sim \mathcal{N}\left(\mathbf{m}, \frac{1+\gamma}{3+\gamma} I_d\right)$ and $\mathbf{m} := \frac{1+\gamma}{3+\gamma}\left(\mu + \frac{2}{1+\gamma}\mu\right) = \mu$. This gives

$$\begin{aligned}
\mathbb{E}[g(\mathbf{X})^2] &= \left(\frac{\gamma}{1+\gamma}\right)^{d}\left(\frac{1+\gamma}{3+\gamma}\right)^{d/2} \mathbb{E}\left[\left(\frac{2+\gamma}{1+\gamma}\right)^2(\mu^\top \mathbf{W})^2 + \frac{1}{(1+\gamma)^2}\|\mu\|_2^4 - \frac{2(2+\gamma)}{(1+\gamma)^2}\|\mu\|_2^2 \mu^\top \mathbf{W}\right] \\
&= \left(\frac{\gamma}{1+\gamma}\right)^{d}\left(\frac{1+\gamma}{3+\gamma}\right)^{d/2}\left(\left(\frac{2+\gamma}{1+\gamma}\right)^2 \mu^\top\left(\frac{1+\gamma}{3+\gamma} I_d + \mu\mu^\top\right)\mu + \frac{1}{(1+\gamma)^2}\|\mu\|_2^4 \right. \\
&\qquad \left. - \frac{2(2+\gamma)}{(1+\gamma)^2}\|\mu\|_2^4\right) \\
&= \left(\frac{\gamma^2}{(1+\gamma)(3+\gamma)}\right)^{d/2}\left(\frac{(2+\gamma)^2}{(1+\gamma)(3+\gamma)}\|\mu\|_2^2 + \|\mu\|_2^4\right) .
\end{aligned}$$

We hence obtain

$$\begin{aligned}\sigma^2_{\text{cond}} &= \mathbb{E}[g(\mathbf{X})^2] - D^{\text{KSD}}(Q,P)^2 \\ &= \left(\frac{\gamma^2}{(1+\gamma)(3+\gamma)}\right)^{d/2}\left(\frac{(2+\gamma)^2}{(1+\gamma)(3+\gamma)}\|\mu\|_2^2 + \|\mu\|_2^4\right) - \left(\frac{\gamma}{2+\gamma}\right)^d \|\mu\|_2^4 \\ &= \left(\frac{\gamma^2}{(1+\gamma)(3+\gamma)}\right)^{d/2}\left(\frac{(2+\gamma)^2}{(1+\gamma)(3+\gamma)}\|\mu\|_2^2 + \left(1 - \left(\frac{(1+\gamma)(3+\gamma)}{(2+\gamma)^2}\right)^{d/2}\right)\|\mu\|_2^4\right).\end{aligned}$$

F.2.4. PROOF FOR $\sigma^2_{\text{full}}$

For simplicity, we define $\mathbf{Z} \coloneqq \mathbf{X} - \mu$ and $\mathbf{Z}' \coloneqq \mathbf{X}'z - \mu$ so that $\mathbf{Z}, \mathbf{Z}'$ are independent copies from $\mathcal{N}(0, I_d)$. By (38), the second moment can be simplified as

$$\begin{aligned}&\mathbb{E}[u_P^{\text{KSD}}(\mathbf{X},\mathbf{X}')^2] \\ &= \mathbb{E}\left[\exp\left(-\frac{1}{\gamma}\|\mathbf{X}-\mathbf{X}'\|_2^2\right)\left(\mathbf{X}^\top\mathbf{X}' - \frac{\gamma+1}{\gamma^2}\|\mathbf{X}-\mathbf{X}'\|_2^2 + \frac{d}{\gamma}\right)^2\right] \\ &= \mathbb{E}\left[\exp\left(-\frac{1}{\gamma}\|\mathbf{Z}-\mathbf{Z}'\|_2^2\right)\left((\mathbf{Z}+\mu)^\top(\mathbf{Z}'+\mu) - \frac{\gamma+1}{\gamma^2}\|\mathbf{Z}-\mathbf{Z}'\|_2^2 + \frac{d}{\gamma}\right)^2\right] \\ &= \left(\frac{\gamma}{4+\gamma}\right)^{d/2} \underbrace{\mathbb{E}\left[\left((\mathbf{W}+\mu)^\top(\mathbf{W}'+\alpha_1\mathbf{W}+\mu) - \alpha_2\|(1-\alpha_1)\mathbf{W}-\mathbf{W}'\|_2^2 + \frac{d}{\gamma}\right)^2\right]}_{=:T},\end{aligned}$$

where in the last line we have applied Lemma 29, and

$$\mathbf{W} \sim \mathcal{N}\left(0,\ \frac{1+\gamma/2}{2+\gamma/2}I_d\right), \qquad \mathbf{W}' \sim \mathcal{N}\left(0,\ \frac{\gamma}{2+\gamma}I_d\right), \qquad \alpha_1 \coloneqq \frac{1}{1+\gamma/2}, \qquad \alpha_2 \coloneqq \frac{\gamma+1}{\gamma^2}.$$

We now aim to compute the expectation $T$ by first taking an expectation over $\mathbf{W}'$:

$$\begin{aligned}T &= \mathbb{E}\left[\left((\mathbf{W}+\mu)^\top(\mathbf{W}'+\alpha_1\mathbf{W}+\mu) - \alpha_2\|(1-\alpha_1)\mathbf{W}-\mathbf{W}'\|_2^2 + \frac{d}{\gamma}\right)^2\right] \\ &= \mathbb{E}\Bigg[\Big(-\alpha_2\big\|\mathbf{W}'\big\|_2^2 + \big(\mathbf{W}+\mu+2\alpha_2(1-\alpha_1)\mathbf{W}\big)^\top\mathbf{W}' \\ &\qquad + \underbrace{(\mathbf{W}+\mu)^\top(\alpha_1\mathbf{W}+\mu) - \alpha_2(1-\alpha_1)^2\|\mathbf{W}\|_2^2 + \frac{d}{\gamma}}_{=:\beta_{\mathbf{W}}}\Big)^2\Bigg] \\ &= \mathbb{E}\Big[\alpha_2^2\big\|\mathbf{W}'\big\|_2^4 + \left((\mathbf{W}+\mu+2\alpha_2(1-\alpha_1)\mathbf{W})^\top\mathbf{W}'\right)^2 + \beta_{\mathbf{W}}^2 \\ &\qquad - 2\alpha_2\big(\mathbf{W}+\mu+2\alpha_2(1-\alpha_1)\mathbf{W}\big)^\top\mathbf{W}'\|\mathbf{W}'\|_2^2 - 2\alpha_2\beta_{\mathbf{W}}\big\|\mathbf{W}'\big\|_2^2 \\ &\qquad + 2\beta_{\mathbf{W}}(\mathbf{W}+\mu+2\alpha_2(1-\alpha_1)\mathbf{W})^\top\mathbf{W}'\Big].\end{aligned}$$

Since $\mathbf{W}'$ is zero-mean, independent of $\mathbf{W}$ and follows a distribution symmetric around zero, $\mathbb{E}[\mathbf{W}'] = \mathbb{E}[\mathbf{W}'\|\mathbf{W}'\|_2^2] = \mathbf{0}$. Since $\|\mathbf{W}'\|_2^2 \sim \frac{\gamma}{2+\gamma}\chi_d^2$ where $\chi_d^2$ is a chi-squared distribution with $d$ degrees of freedom, we have $\mathbb{E}[\|\mathbf{W}'\|_2^2] = \frac{\gamma}{2+\gamma}d$ and $\mathbb{E}[\|\mathbf{W}'\|_2^4] = \frac{\gamma^2}{(2+\gamma)^2}(2d+d^2)$. We also have $\mathbb{E}[\mathbf{W}'\mathbf{W}'^\top] = \frac{\gamma}{2+\gamma}\mathbf{I}_d$. Thus

$$\begin{aligned}
T &= \mathbb{E}\left[\alpha_2^2\|\mathbf{W}'\|_2^4 + \left((\mathbf{W}+\mu+2\alpha_2(1-\alpha_1)\mathbf{W})^\top\mathbf{W}'\right)^2 + \beta_{\mathbf{W}}^2 - 2\alpha_2\beta_{\mathbf{w}}\|\mathbf{W}'\|_2^2\right] \\
&= \mathbb{E}\left[\alpha_2^2\frac{\gamma^2}{(2+\gamma)^2}(2d+d^2) + \frac{\gamma}{2+\gamma}\|2\alpha_2(1-\alpha_1)\mathbf{W}+\mathbf{W}+\mu\|_2^2 + \beta_{\mathbf{W}}^2 - 2\alpha_2\beta_{\mathbf{W}}\frac{\gamma}{2+\gamma}d\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\frac{(1+\gamma)^2}{\gamma^2(2+\gamma)^2}(2d+d^2) + \frac{\gamma}{2+\gamma}\left\|\frac{2(1+\gamma)}{\gamma(2+\gamma)}\mathbf{W}+\mathbf{W}+\mu\right\|_2^2 + \beta_{\mathbf{W}}^2 - 2\beta_{\mathbf{W}}\frac{1+\gamma}{\gamma(2+\gamma)}d\right] \\
&= \mathbb{E}\left[\frac{(1+\gamma)^2}{\gamma^2(2+\gamma)^2}(2d+d^2) + \frac{\gamma}{2+\gamma}\left\|\frac{2+4\gamma+\gamma^2}{\gamma(2+\gamma)}\mathbf{W}+\mu\right\|_2^2 + \beta_{\mathbf{W}}^2 - 2\beta_{\mathbf{W}}\frac{1+\gamma}{\gamma(2+\gamma)}d\right] \\
&\overset{(b)}{=} \frac{(1+\gamma)^2}{\gamma^2(2+\gamma)^2}(2d+d^2) + \frac{(2+4\gamma+\gamma^2)^2}{\gamma(2+\gamma)^2(4+\gamma)}d + \frac{\gamma}{2+\gamma}\|\mu\|_2^2 + \mathbb{E}\Big[\beta_{\mathbf{W}}^2 - 2\beta_{\mathbf{W}}\frac{1+\gamma}{\gamma(2+\gamma)}d\Big] \\
&= \frac{(1+\gamma)^2}{\gamma^2(2+\gamma)^2}d^2 + \left(\frac{2(1+\gamma)^2}{\gamma^2(2+\gamma)^2} + \frac{(2+4\gamma+\gamma^2)^2}{\gamma(2+\gamma)^2(4+\gamma)}\right)d + \frac{\gamma}{2+\gamma}\|\mu\|_2^2 + \mathbb{E}\Big[\beta_{\mathbf{W}}^2 - 2\beta_{\mathbf{W}}\frac{1+\gamma}{\gamma(2+\gamma)}d\Big]\ .
\end{aligned}$$

In $(a)$, we have substituted in $\alpha_1 = 2/(2+\gamma)$ and $\alpha_2 = (\gamma+1)/\gamma^2$, and in $(b)$ we have taken the expectation of the second term. Now re-express $\beta_{\mathbf{W}}$ as

$$\begin{aligned}
\beta_{\mathbf{W}} &= \frac{2}{2+\gamma}\|\mathbf{W}\|_2^2 + \left(\frac{2}{2+\gamma}+1\right)\mu^\top\mathbf{W} + \|\mu\|_2^2 - \frac{\gamma+1}{\gamma^2}\left(1-\frac{2}{2+\gamma}\right)^2\|\mathbf{W}\|_2^2 + \frac{d}{\gamma} \\
&= \left(\frac{2}{2+\gamma} - \frac{\gamma+1}{(2+\gamma)^2}\right)\|\mathbf{W}\|_2^2 + \frac{4+\gamma}{2+\gamma}\mu^\top\mathbf{W} + \|\mu\|_2^2 + \frac{d}{\gamma} \\
&= \frac{\gamma+3}{(2+\gamma)^2}\|\mathbf{W}\|_2^2 + \frac{4+\gamma}{2+\gamma}\mu^\top\mathbf{W} + \|\mu\|_2^2 + \frac{d}{\gamma}\ .
\end{aligned}$$

By noting that odd moments of $\mathbf{W}$ vanish, we get that

$$\begin{aligned}
\mathbb{E}[-2\beta_{\mathbf{W}}\frac{1+\gamma}{\gamma(2+\gamma)}d] &= -\frac{2(1+\gamma)}{\gamma(2+\gamma)}d\Big(\frac{\gamma+3}{(2+\gamma)(4+\gamma)}d + \|\mu\|_2^2 + \frac{d}{\gamma}\Big) \\
&= -\frac{2(1+\gamma)}{\gamma(2+\gamma)}\Big(\frac{\gamma+3}{(2+\gamma)(4+\gamma)} + \frac{1}{\gamma}\Big)d^2 - \frac{2(1+\gamma)}{\gamma(2+\gamma)}d\|\mu\|_2^2\ ,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\beta_{\mathbf{W}}^2] &= \frac{(\gamma+3)^2}{(2+\gamma)^2(4+\gamma)^2}(2d+d^2) + \frac{4+\gamma}{2+\gamma}\|\mu\|_2^2 + \left(\|\mu\|_2^2 + \frac{d}{\gamma}\right)^2 + \frac{2(\gamma+3)}{(2+\gamma)(4+\gamma)}\left(\|\mu\|_2^2 + \frac{d}{\gamma}\right)d \\
&= \left(\frac{(\gamma+3)^2}{(2+\gamma)^2(4+\gamma)^2} + \frac{1}{\gamma^2} + \frac{2(\gamma+3)}{(2+\gamma)(4+\gamma)\gamma}\right)d^2 + \left(\frac{2(\gamma+3)^2}{(2+\gamma)^2(4+\gamma)^2}\right)d \\
&\quad + \left(\frac{2}{\gamma} + \frac{2(\gamma+3)}{(2+\gamma)(4+\gamma)}\right)d\|\mu\|_2^2 + \frac{4+\gamma}{2+\gamma}\|\mu\|_2^2 + \|\mu\|_2^4\ .
\end{aligned}$$

The coefficient of $d^2$ in $T$ can then be computed by noting $\gamma = \omega(1)$ as

$$\begin{aligned}
&\frac{(1+\gamma)^2}{\gamma^2(2+\gamma)^2} - \frac{2(1+\gamma)}{\gamma(2+\gamma)}\Big(\frac{\gamma+3}{(2+\gamma)(4+\gamma)} + \frac{1}{\gamma}\Big) + \Big(\frac{(\gamma+3)^2}{(2+\gamma)^2(4+\gamma)^2} + \frac{1}{\gamma^2} + \frac{2(\gamma+3)}{(2+\gamma)(4+\gamma)\gamma}\Big) \\
&= \frac{(2+\gamma)^2}{\gamma^2(4+\gamma)^2} = \frac{1}{\gamma^2} + o\Big(\frac{1}{\gamma^2}\Big)\ .
\end{aligned}$$

Similarly, the coefficient of $d$ in $T$ can be computed as

$$\Big(\frac{2(1+\gamma)^2}{\gamma^2(2+\gamma)^2} + \frac{(2+4\gamma+\gamma^2)^2}{\gamma(2+\gamma)^2(4+\gamma)}\Big) + \Big(\frac{2(\gamma+3)^2}{(2+\gamma)^2(4+\gamma)^2}\Big) = 1 + o(1)\ ,$$

the coefficient of $d\|\mu\|_2^2$ in $T$ can be computed as

$$-\frac{2(1+\gamma)}{\gamma(2+\gamma)} + \Big(\frac{2}{\gamma} + \frac{2(\gamma+3)}{(2+\gamma)(4+\gamma)}\Big) \;=\; \frac{2(2+\gamma)}{\gamma(4+\gamma)} \;=\; \frac{2}{\gamma} + o\Big(\frac{1}{\gamma}\Big)\;,$$

the coefficient of $\|\mu\|_2^2$ in $T$ can be computed as

$$\frac{\gamma}{2+\gamma} + \frac{4+\gamma}{2+\gamma} \;=\; 2 + o(1)\;,$$

and finally the coefficient of $\|\mu\|_2^4$ in $T$ is 1. Combining the five computations of coefficients, we get that

$$T \;=\; 4d + \frac{d^2}{\gamma^2} + \frac{2d\|\mu\|_2^2}{\gamma} + 2\|\mu\|_2^2 + \|\mu\|_2^4 + o\Big(d + \frac{d^2}{\gamma^2} + \frac{d\|\mu\|_2^2}{\gamma} + +\|\mu\|_2^2\Big)\;,$$

and therefore the desired quantity is given as

$$\begin{aligned}
\sigma^2_{\text{full}} &= \mathbb{E}[u_P^{\text{KSD}}(\mathbf{X},\mathbf{X}')^2] - D^{\text{KSD}}(Q,P)^2 \\
&= \Big(\frac{\gamma}{4+\gamma}\Big)^{d/2} T - \Big(\frac{\gamma}{2+\gamma}\Big)^{d} \|\mu\|_2^4 \\
&= \Big(\frac{\gamma}{4+\gamma}\Big)^{d/2} T - \Big(\frac{\gamma}{4+\gamma}\Big)^{d/2} \Big(\frac{\gamma(4+\gamma)}{(2+\gamma)^2}\Big)^{d/2} \|\mu\|_2^4 \\
&= \Big(\frac{\gamma}{4+\gamma}\Big)^{d/2} \Bigg(d + \frac{d^2}{\gamma^2} + \frac{2d\|\mu\|_2^2}{\gamma} + 2\|\mu\|_2^2 + \bigg(1 - \Big(\frac{\gamma(4+\gamma)}{(2+\gamma)^2}\Big)^{d/2}\bigg) \|\mu\|_2^4 \\
&\qquad\qquad + o\bigg(d + \frac{d^2}{\gamma^2} + \frac{d\|\mu\|_2^2}{\gamma} + \|\mu\|_2^2\bigg)\Bigg)\;.
\end{aligned}$$

### F.2.5. Proof for upper bound on $\mathbb{E}[|g^{\text{KSD}}(\mathbf{X})|^\nu]$

Fix $\nu > 2$. We can apply Lemma 28 to rewrite the $\nu$-th moment of $g^{\text{KSD}}(\mathbf{Z})$ as

$$\begin{aligned}
\mathbb{E}[|g^{\text{KSD}}(\mathbf{X})|^\nu] &= \Big(\frac{\gamma}{1+\gamma}\Big)^{\nu d/2} \mathbb{E}\Big[\exp\Big(-\frac{\nu}{2(1+\gamma)}\|\mathbf{X}-\mu\|_2^2\Big)\Big|\frac{2+\gamma}{1+\gamma}\mu^\top\mathbf{X} - \frac{1}{1+\gamma}\|\mu\|_2^2\Big|^\nu\Big] \\
&= \Big(\frac{\gamma}{1+\gamma}\Big)^{\nu d/2} \Big(\frac{(1+\gamma)/\nu}{1+(1+\gamma)/\nu}\Big)^{d/2} \underbrace{\mathbb{E}\Big[\Big|\frac{2+\gamma}{1+\gamma}\mu^\top\mathbf{W} - \frac{1}{1+\gamma}\|\mu\|_2^2\Big|^\nu\Big]}_{=:T}\;,
\end{aligned}$$

where $\mathbf{W} \sim \mathcal{N}(\mathbf{m}, a^2 I_d)$ with $\mathbf{m} := \frac{(1+\gamma)/\nu}{1+(1+\gamma)/\nu}\big(\mu + \frac{\nu}{1+\gamma}\mu\big) + \mu$ and $a^2 := \frac{(1+\gamma)/\nu}{1+(1+\gamma)/\nu} = \frac{1+\gamma}{1+\nu=\gamma}$. Defining $\mathbf{V} := W - \mu$ so that $\mathbf{V} \sim \mathcal{N}(0, a^2 I_d)$, we have

$$\begin{aligned}
T &= \mathbb{E}\Big[\Big|\frac{2+\gamma}{1+\gamma}\mu^\top(\mathbf{V}+\mu) - \frac{1}{1+\gamma}\|\mu\|_2^2\Big|^\nu\Big] \;=\; \mathbb{E}\Big[\Big|\frac{2+\gamma}{1+\gamma}\mu^\top\mathbf{V} - \|\mu\|_2^2\Big|^\nu\Big] \\
&\overset{(i)}{\leq} 2^{\nu-1}\Big(\Big(\frac{2+\gamma}{1+\gamma}\Big)^\nu \mathbb{E}[|\mu^\top\mathbf{V}|^\nu] + \|\mu\|_2^{2\nu}\Big) \\
&= 2^{\nu-1}\Big(C_\nu\Big(\frac{2+\gamma}{1+\gamma}\Big)^\nu a^\nu \|\mu\|_2^\nu + \|\mu\|_2^{2\nu}\Big)\;,
\end{aligned}$$

where *(i)* follows by the fact that $|u+v|^\nu \leq 2^{\nu-1}(|u|^\nu + |v|^\nu)$ for any $u, v \in \mathbb{R}$, and in the last line we have computed the expectation by noting that $\mu^\top\mathbf{V}$ follows a univariate Gaussian distribution

$\mathcal{N}(0, a^2\|\mu\|_2^2)$ and using its moment formula to yield $\mathbb{E}[|\mu^\top \mathbf{V}|^\nu \leq C_\nu a^\nu \|\mu\|_2^\nu$ for some constant $C_\nu$ that depends only on $\nu$. Combining these and substituting the definition of $a^2$ gives

$$\begin{aligned}
&\mathbb{E}[|g^{\mathrm{KSD}}(\mathbf{X})|^\nu] \\
&\leq 2^{\nu-1}\left(\frac{\gamma}{1+\gamma}\right)^{\nu d/2}\left(\frac{(1+\gamma)/\nu}{1+(1+\gamma)/\nu}\right)^{d/2}\left(C_\nu\left(\frac{2+\gamma}{1+\gamma}\right)^\nu\left(\frac{1+\gamma}{1+\nu+\gamma}\right)^{\nu/2}\|\mu\|_2^\nu+\|\mu\|_2^{2\nu}\right) \\
&\leq \left(\frac{\gamma}{1+\gamma}\right)^{\nu d/2}\left(\frac{1+\gamma}{1+\nu+\gamma}\right)^{d/2}\left(2^{3\nu/2-1}C_\nu\|\mu\|_2^\nu+2^{\nu-1}\|\mu\|_2^{2\nu}\right) ,
\end{aligned}$$

where in the last line we have used the assumption that $\nu > 2$ to yield the inequality

$$\left(\frac{2+\gamma}{1+\gamma}\right)^\nu\left(\frac{1+\gamma}{1+\nu+\gamma}\right)^{\nu/2} = \left(\frac{2+\gamma}{1+\gamma}\right)^\nu\left(\frac{2+\gamma}{1+\nu+\gamma}\right)^{\nu/2} = 2^{\nu/2}\times 1 = 2^{\nu/2} .$$

Defining the constants $C_1 \coloneqq 2^{3\nu/2-1}C_\nu$ and $C_2 \coloneqq 2^{\nu-1}$ completes the proof.

### F.2.6. PROOF FOR UPPER BOUND ON $\mathbb{E}[|u_P^{\mathrm{KSD}}(\mathbf{X},\mathbf{X}')|^\nu]$

Fix $\nu > 2$. Define $\mathbf{Z} \coloneqq \mathbf{X}-\mu$ and $\mathbf{Z}' \coloneqq \mathbf{X}'-\mu$ so that $\mathbf{Z}, \mathbf{Z}'$ are independent draws from $\mathcal{N}(0, I_d)$. Using (9), we can write the $\nu$-th central moment as

$$\begin{aligned}
&\mathbb{E}[|u_P^{\mathrm{KSD}}(\mathbf{X},\mathbf{X}')|^\nu] \\
&= \mathbb{E}\left[\left|\exp\left(-\frac{1}{2\gamma}\|\mathbf{X}-\mathbf{X}'\|_2^2\right)\left(\mathbf{X}^\top\mathbf{X}'-\frac{\gamma+1}{\gamma^2}\|\mathbf{X}-\mathbf{X}'\|_2^2+\frac{d}{\gamma}\right)\right|^\nu\right] \\
&= \mathbb{E}\left[\exp\left(-\frac{\nu}{2\gamma}\|\mathbf{Z}-\mathbf{Z}'\|_2^2\right)\left|(\mathbf{Z}+\mu)^\top(\mathbf{Z}'+\mu)-\frac{\gamma+1}{\gamma^2}\|\mathbf{Z}-\mathbf{Z}'\|_2^2+\frac{d}{\gamma}\right|^\nu\right] \\
&= \left(\frac{\gamma/\nu}{2+\gamma/\nu}\right)^{d/2}\underbrace{\mathbb{E}\left[\left|(\mathbf{W}+\mu)^\top(\mathbf{W}'+(1-\alpha_2)\mathbf{W}+\mu)-\alpha_1\|\mathbf{W}-\mathbf{W}'-(1-\alpha_2)\mathbf{W}\|_2^2+\frac{d}{\gamma}\right|^\nu\right]}_{=:T} ,
\end{aligned} \tag{39}$$

where the last line follows by using Lemma 29 and defining the quantities $\alpha_1 \coloneqq \frac{\gamma+1}{\gamma^2}$, $\alpha_2 \coloneqq \frac{\gamma}{\nu+\gamma}$, $\alpha_3 \coloneqq \frac{\gamma}{2\nu+\gamma}$, and

$$\mathbf{W}' \sim \mathcal{N}\left(\mathbf{0},\ \frac{\gamma/\nu}{1+\gamma/\nu}I_d\right) = \mathcal{N}(\mathbf{0},\ \alpha_2 I_d) , \qquad \mathbf{W} \sim \mathcal{N}\left(\mathbf{0},\ \frac{\gamma/\nu}{2+\gamma/\nu}I_d\right) = \mathcal{N}(\mathbf{0},\ \alpha_3 I_d) ,$$

while also noting that $1-\alpha_2 = \frac{\gamma/\nu}{1+\gamma/\nu}\times\frac{\nu}{\gamma} = \frac{\nu}{\nu+\gamma} = 1-\alpha_2$. By a Jensen's inequality, we get that

$$\begin{aligned}
T &= \mathbb{E}\left[\left|(\mathbf{W}+\mu)^\top(\mathbf{W}'+(1-\alpha_2)\mathbf{W}+\mu)-\alpha_1\left\|\alpha_2\mathbf{W}-\mathbf{W}'\right\|_2^2+\frac{d}{\gamma}\right|^\nu\right] \\
&= \mathbb{E}\left[\left|(\mathbf{W}+\mu)^\top\mathbf{W}'+(1-\alpha_2)\|\mathbf{W}+\mu\|_2^2+\alpha_2(\mathbf{W}+\mu)^\top\mu-\alpha_1\|\alpha_2\mathbf{W}-\mathbf{W}'\|_2^2+\frac{d}{\gamma}\right|^\nu\right] \\
&\leq 5^{\nu-1}\mathbb{E}\Big[\left|(\mathbf{W}+\mu)^\top\mathbf{W}'\right|^\nu+|1-\alpha_2|^\nu\|\mathbf{W}+\mu\|_2^{2\nu}+\alpha_2^\nu\left|(\mathbf{W}+\mu)^\top\mu\right|^\nu \\
&\qquad +\alpha_1^\nu\|\alpha_2\mathbf{W}-\mathbf{W}'\|_2^{2\nu}+\left(\frac{d}{\gamma}\right)^\nu\Big] ,
\end{aligned}$$

where the last line follows from a Jensen's inequality applied to the convex function $x \mapsto |x|^\nu$.

We next seek to bound the expectation of each term individually. To bound $\mathbb{E}\big[\big|\big(\mathbf{W}+\mu\big)^\top\mathbf{W}'\big|^\nu\big]$, we note that $(\mathbf{W}-\mu)^\top\mathbf{W}'$ conditioning on $\mathbf{W}$ follows a normal distribution $\mathcal{N}(0,\alpha_2\|\mathbf{W}-\mu\|_2^2)$. Hence, using the moment formula of univariate Gaussians, we have

$$\mathbb{E}\big[\big|\big(\mathbf{W}+\mu\big)^\top\mathbf{W}'\big|^\nu\big] = \mathbb{E}\big[\mathbb{E}\big[\big|\big(\mathbf{W}+\mu\big)^\top\mathbf{W}'\big|^\nu\big|\mathbf{W}\big]\big] = \mathbb{E}\big[C_\nu\alpha_2^\nu\|\mathbf{W}+\mu\|_2^\nu\big] ,$$

for some constant $C_\nu$ constant depending only on $\nu$. By the convexity of the function $\mathbf{x}\mapsto\|\mathbf{x}\|_2^\nu$, we can bound the above term as

$$\begin{aligned}\mathbb{E}\big[C_\nu\alpha_2^\nu\|\mathbf{W}+\mu\|_2^\nu\big] &\leq 2^{\nu-1}C_\nu\alpha_2^\nu\big(\mathbb{E}\big[\|\mathbf{W}\|_2^\nu\big]+\|\mu\|_2^\nu\big)\\ &\overset{(i)}{\leq} C_\nu\alpha_2^\nu\alpha_3^\nu\big(\mathbb{E}\big[\|\mathbf{W}\|_2^{2\nu}\big]\big)^{1/2}+C_\nu\alpha_2^\nu\|\mu\|_2^\nu\\ &\overset{(ii)}{=} C_\nu\alpha_2^\nu\alpha_3^\nu(d^{\nu/2}+o(d^{\nu/2}))+C_\nu\alpha_2^\nu\|\mu\|_2^\nu\\ &\overset{(iii)}{\leq} C_\nu d^{\nu/2}+o(d^{\nu/2})+C_\nu\|\mu\|_2^\nu ,\end{aligned}$$

where *(i)* holds by a Jensen's inequality, in *(ii)* we have noted that $\alpha_3^{-1}\|\mathbf{W}\|_2^2$ follows a chi-squared distribution with $d$ degrees of freedom and used the formula for its $\nu$-th moment, and *(iii)* follows since $\alpha_2=\frac{\gamma}{\nu+\gamma}<1$ and $\alpha_3=\frac{\gamma}{2\nu+\gamma}<1$. The expectation of the second term can be bounded using a similar argument as

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{W}+\mu\|_2^{2\nu}\big] \leq 2^{\nu-1}\big(\mathbb{E}\big[\|\mathbf{W}\|_2^{2\nu}\big]+\|\mu\|_2^{2\nu}\big) &= 2^{\nu-1}\big(\alpha_3^\nu d^\nu+o(d^\nu)+\|\mu\|_2^{2\nu}\big)\\ &\leq 2^{\nu-1}d^\nu+o(d^\nu)+2^{\nu-1}\|\mu\|_2^{2\nu} .\end{aligned}$$

The expectation of the third term is

$$\begin{aligned}\mathbb{E}\big[\big|(\mathbf{W}+\mu)^\top\mu\big|^\nu\big] = \mathbb{E}\big[\big|\mathbf{W}^\top\mu+\|\mu\|_2^2\big|^\nu\big] &\leq 2^{\nu-1}\big(\mathbb{E}\big[\big|\mathbf{W}^\top\mu\big|^\nu\big]+\|\mu\|_2^{2\nu}\big)\\ &= 2^{\nu-1}\big(\alpha_3^{\nu/2}\|\mu\|_2^\nu+\|\mu\|_2^{2\nu}\big)\\ &\leq 2^{\nu-1}\|\mu\|_2^\nu+2^{\nu-1}\|\mu\|_2^{2\nu} ,\end{aligned}$$

where the second last line holds as $\mu^\top\mathbf{W}$ is a univariate Gaussian with zero-mean and variance $\alpha_3\|\mu\|_2^2$, and the last line holds again as $\alpha_3\leq1$. It then remains to bound $\mathbb{E}\big[\|\alpha_2\mathbf{W}-\mathbf{W}'\|_2^{2\nu}\big]$. Noting that $\alpha_2\mathbf{W}-\mathbf{W}'\sim\mathcal{N}(0,\alpha_2(\alpha_3+1)I_d)$, the random variable $\alpha_2^{-1}(\alpha_3+1)^{-1}\|\alpha_2\mathbf{W}-\mathbf{W}'\|_2^2$ follows a chi-squared distribution with $d$ degrees of freedom. A similar argument as before gives

$$\mathbb{E}\big[\|\alpha_2\mathbf{W}-\mathbf{W}'\|_2^{2\nu}\big] \leq \alpha_2^\nu(\alpha_3+1)^\nu\big(d^\nu+o(d^\nu)\big) \leq 2^\nu d^\nu+o(d^\nu) ,$$

where in the last inequality we have used the fact that $\alpha_2(\alpha_3+1)<2$. Combining these terms, we can bound $T$ as

$$\begin{aligned}T \leq 5^{\nu-1}\Big[&\big(C_\nu d^{\nu/2}+o(d^{\nu/2})+C_\nu\|\mu\|_2^\nu\big)+|1-\alpha_2|^\nu2^{\nu-1}\big(d^\nu+o(d^\nu)+\|\mu\|_2^{2\nu}\big)\\ &+2^{\nu-1}\big(\|\mu\|_2^\nu+\|\mu\|_2^{2\nu}\big)+\alpha_1^\nu2^\nu\big(d^\nu+o(d^\nu)\big)+\Big(\frac{d}{\gamma}\Big)^\nu\Big] .\end{aligned}$$

To proceed, we note that $\alpha_1^\nu=\big(\frac{\gamma+1}{\gamma^2}\big)^\nu=\big(\frac{1}{\gamma}+\frac{1}{\gamma^2}\big)^\nu=\frac{1}{\gamma^\nu}+o\big(\frac{1}{\gamma^\nu}\big)$ and that $(1-\alpha_2)^\nu=\big(\frac{\nu}{\nu+\gamma}\big)^\nu=\frac{1}{\gamma^\nu}+o\big(\frac{1}{\gamma^\nu}\big)$, since $\gamma=\omega(1)$ by assumption. Therefore,

$$|1-\alpha_2|^\nu2^{\nu-1}\big(d^\nu+o(d^\nu)+\|\mu\|_2^{2\nu}\big) = 2^\nu\frac{d^\nu}{\gamma^\nu}+2^\nu\frac{\|\mu\|_2^{2\nu}}{\gamma^\nu}+o\Big(\frac{d^\nu}{\gamma^\nu}+\frac{\|\mu\|_2^{2\nu}}{\gamma^\nu}\Big) ,$$

and

$$\alpha_1^\nu 2^\nu \big(d^\nu + o(d^\nu)\big) \;=\; \frac{d^\nu}{\gamma^\nu} + o\Big(\frac{d^\nu}{\gamma^\nu}\Big) \;.$$

It then follows by grouping and rearranging that $T$ can be bounded as

$$\begin{aligned} T \;&\leq 5^{\nu-1}\Big[C_\nu d^{\nu/2} + o(d^{\nu/2}) + C_\nu \|\mu\|_2^\nu + 2^\nu \frac{d^\nu}{\gamma^\nu} + 2^\nu \frac{\|\mu\|_2^{2\nu}}{\gamma^\nu} + o\Big(\frac{d^\nu}{\gamma^\nu} + \frac{\|\mu\|_2^{2\nu}}{\gamma^\nu}\Big) \\ &\qquad + 2^{\nu-1}\big(\|\mu\|_2^\nu + \|\mu\|_2^{2\nu}\big) + \frac{d^\nu}{\gamma^\nu} + o\Big(\frac{d^\nu}{\gamma^\nu}\Big) + \Big(\frac{d}{\gamma}\Big)^\nu\Big] \\ &= 5^{\nu-1}\Big[C_\nu d^{\nu/2} + (2+2^\nu)\Big(\frac{d}{\gamma}\Big)^\nu + (C_\nu + 2^{\nu-1})\|\mu\|_2^\nu + 2^{\nu-1}\|\mu\|_2^{2\nu} \\ &\qquad + o\Big(d^{\nu/2} + \frac{d^\nu}{\gamma^\nu} + \frac{\|\mu\|_2^{2\nu}}{\gamma^\nu}\Big)\Big] \\ &= C_3 d^{\nu/2} + C_4\Big(\frac{d}{\gamma}\Big)^\nu + C_5\|\mu\|_2^\nu + C_6\|\mu\|_2^{2\nu} + o\Big(d^{\nu/2} + \frac{d^\nu}{\gamma^\nu} + \frac{\|\mu\|_2^{2\nu}}{\gamma^\nu}\Big) \;, \end{aligned}$$

where in the last line we have redefined the constants: $C_3 \coloneqq 5^{\nu-1}C_\nu$, $C_4 \coloneqq 5^{\nu-1}(2+2^\nu)$, $C_5 \coloneqq 5^{\nu-1}(C_\nu + 2^{\nu-1})$ and $C_6 \coloneqq 10^{\nu-1}$. The proof is finished by substituting this bound into (39) to yield

$$\begin{aligned} \mathbb{E}[|u_P^{\mathrm{KSD}}(\mathbf{X},\mathbf{X}')|^\nu] \;\leq\; \Big(\frac{\gamma}{2\nu+\gamma}\Big)^{d/2} \Big(&C_3 d^{\nu/2} + C_4\Big(\frac{d}{\gamma}\Big)^\nu + C_5\|\mu\|_2^\nu + C_6\|\mu\|_2^{2\nu} \\ &+ o\Big(d^{\nu/2} + \frac{d^\nu}{\gamma^\nu} + \frac{\|\mu\|_2^{2\nu}}{\gamma^\nu}\Big)\Big) \;. \end{aligned}$$

#### F.2.7. PROOF FOR VERIFYING ASSUMPTION 1

First note that when $\gamma = \Omega(d)$, for any fixed $a, b, c > 0$, we have that by a Taylor expansion,

$$\begin{aligned} \Big(\frac{a+\gamma}{b+\gamma}\Big)^{d/c} \;&=\; \Big(1 + \frac{a-b}{b+\gamma}\Big)^{d/c} \;=\; \exp\Big(\frac{d}{c}\log\Big(1 + \frac{a-b}{b+\gamma}\Big)\Big) \;=\; \exp\Big(\frac{d(a-b)}{c(b+\gamma)} + o\Big(\frac{d}{\gamma^2}\Big)\Big) \\ &=\; \exp\Big(\frac{d(a-b)}{c(b+\gamma)}\Big)\Big(1 + o\Big(\frac{d}{\gamma^2}\Big)\Big) \;=\; \Theta(1) \;. \end{aligned}$$

Using this together with the assumption $\|\mu\|_2 = \Theta(1)$ and the moment bounds in Lemma 22(iii)-(vi), we get that

$$\begin{aligned} \sigma^2_{\text{cond}} \;&= \Theta\bigg(\Big(\frac{\gamma^2}{(1+\gamma)(3+\gamma)}\Big)^{d/2}\Big(\frac{(2+\gamma)^2}{(1+\gamma)(3+\gamma)} + \Big(1 - \Big(\frac{(1+\gamma)(3+\gamma)}{(2+\gamma)^2}\Big)^{d/2}\Big)\Big)\bigg) \;=\; \Theta(1) \;, \\ \sigma^2_{\text{full}} \;&= \Theta\bigg(\Big(\frac{\gamma}{4+\gamma}\Big)^{d/2}\Big(d + \frac{d^2}{\gamma^2} + \frac{2d}{\gamma} + \Big(1 - \Big(\frac{\gamma(4+\gamma)}{(2+\gamma)^2}\Big)^{d/2}\Big) + o\Big(d + \frac{d^2}{\gamma^2} + \frac{d}{\gamma}\Big)\Big)\bigg) \\ &= \Theta\Big(d + \frac{d^2}{\gamma^2} + \frac{2d}{\gamma}\Big) \;=\; \Theta\Big(d + \frac{d^2}{\gamma^2}\Big) \;, \end{aligned}$$

and for $\nu \in (2,3]$,

$$\begin{aligned} M^\nu_{\text{cond};\nu} \;&\leq \mathbb{E}[|g^{\mathrm{KSD}}(\mathbf{X})|^\nu] \;=\; O\bigg(\Big(\frac{\gamma}{1+\gamma}\Big)^{\nu d/2}\Big(\frac{1+\gamma}{1+\nu+\gamma}\Big)^{d/2}\bigg) \;=\; O(1) \;, \\ M^\nu_{\text{full};\nu} \;&\leq \mathbb{E}[|u_P^{\mathrm{KSD}}(\mathbf{X},\mathbf{X}')|^\nu] \;=\; O\bigg(\Big(\frac{\gamma}{2\nu+\gamma}\Big)^{d/2}\Big(d^{\nu/2} + \Big(\frac{d}{\gamma}\Big)^\nu\Big)\bigg) \;=\; O\Big(d^{\nu/2} + \frac{d^\nu}{\gamma^\nu}\Big) \;. \end{aligned}$$

This implies that

$$\frac{M_{\text{cond};\nu}}{\sigma_{\text{cond}}} \;=\; O(1)\;, \qquad \text{and} \qquad \frac{M_{\text{full};\nu}}{\sigma_{\text{full}}} \;=\; O\bigg(\Big(d^{1/2}+\frac{d}{\gamma}\Big)^{-1}\Big(d^{1/2}+\frac{d}{\gamma}\Big)\bigg) \;=\; O(1)\;.$$

In other words, $\frac{M_{\text{cond};\nu}}{\sigma_{\text{cond}}}$ and $\frac{M_{\text{full};\nu}}{\sigma_{\text{full}}}$ are both bounded by finite, $d$-independent constants, which verifies Assumption 1.

## F.3. Proof for Lemma 24

### F.3.1. Proof for $g^{\text{mmd}}(\mathbf{z})$

Recall that the MMD U-statistic is $u^{\text{MMD}}(\mathbf{z},\mathbf{z}') = \kappa(\mathbf{x},\mathbf{x}') + \kappa(\mathbf{y},\mathbf{y}') - \kappa(\mathbf{x},\mathbf{y}') - \kappa(\mathbf{x}',\mathbf{y})$ for $\mathbf{z} := (\mathbf{x},\mathbf{y})$ and $\mathbf{z}' = (\mathbf{x}',\mathbf{y}')$. Taking expectation with respect to the second argument, we have

$$\begin{aligned} g^{\text{mmd}}(\mathbf{z}) \;&:=\; \mathbb{E}[u^{\text{MMD}}(z,Z')] \;=\; \mathbb{E}[\kappa(\mathbf{x},\mathbf{X}') + \kappa(\mathbf{y},\mathbf{Y}') - \kappa(\mathbf{x},\mathbf{Y}') - \kappa(\mathbf{X}',\mathbf{y})] \\ &= \mathbb{E}\Big[\exp\Big(-\frac{\|\mathbf{x}-\mathbf{X}'\|_2^2}{2\gamma}\Big) + \exp\Big(-\frac{\|\mathbf{y}-\mathbf{Y}'\|_2^2}{2\gamma}\Big) - \exp\Big(-\frac{\|\mathbf{x}-\mathbf{Y}'\|_2^2}{2\gamma}\Big) - \exp\Big(-\frac{\|\mathbf{X}'-\mathbf{y}\|_2^2}{2\gamma}\Big)\Big]\;. \end{aligned}$$

We can apply Lemma 28 to compute each term. For example, setting $\mathbf{a}_1 = 1, \mathbf{a}_2 = \gamma, \mathbf{m}_1 = \mu$ and $\mathbf{m}_2 = \mathbf{x}$ in Lemma 28, the first term simplifies to

$$\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{x}-\mathbf{X}'\|_2^2}{2\gamma}\right)\right] \;=\; \Big(\frac{\gamma}{1+\gamma}\Big)^{d/2}\exp\Big(-\frac{1}{2(1+\gamma)}\|\mathbf{x}-\mu\|_2^2\Big)\;.$$

Computing similarly the other terms yields the desired result:

$$g^{\text{mmd}}(\mathbf{z}) \;=\; \Big(\frac{\gamma}{1+\gamma}\Big)^{d/2}\Big[e^{-\frac{1}{2(1+\gamma)}\|\mathbf{x}-\mu\|_2^2} + e^{-\frac{1}{2(1+\gamma)}\|\mathbf{y}\|_2^2} - e^{-\frac{1}{2(1+\gamma)}\|\mathbf{x}\|_2^2} - e^{-\frac{1}{2(1+\gamma)}\|\mathbf{y}-\mu\|_2^2}\Big]\;.$$

### F.3.2. Proof for $D^{\text{MMD}}(Q,P)$

This is a special case of Ramdas et al. (2015, Proposition 1) with $\mu_1 = 0$, $\mu_2 = \mu$ and $\Sigma = I_d$. Alternatively, applying Lemma 28 to compute each term in $\mathbb{E}[g(Z)]$ yields the same result.

### F.3.3. Proof for $\sigma^2_{\text{cond}}$

For $\mathbf{Z} = (\mathbf{X},\mathbf{Y})$, the second moment of $g(\mathbf{Z})$ is

$$\begin{aligned} \mathbb{E}[g(\mathbf{Z})^2] \;&=\; \Big(\frac{\gamma}{1+\gamma}\Big)^{d}\mathbb{E}\Big[\Big(e^{-\frac{1}{2(1+\gamma)}\|\mathbf{x}-\mu\|_2^2} + e^{-\frac{1}{2(1+\gamma)}\|\mathbf{y}\|_2^2} - e^{-\frac{1}{2(1+\gamma)}\|\mathbf{x}\|_2^2} - e^{-\frac{1}{2(1+\gamma)}\|\mathbf{y}-\mu\|_2^2}\Big)^2\Big] \\ &= \Big(\frac{\gamma}{1+\gamma}\Big)^{d}\mathbb{E}\Big[\exp\Big(-\frac{\|\mathbf{X}-\mu\|_2^2}{1+\gamma}\Big) + \exp\Big(-\frac{\|\mathbf{Y}\|_2^2}{1+\gamma}\Big) + \exp\Big(-\frac{\|\mathbf{X}\|_2^2}{1+\gamma}\Big) \\ &\qquad + \exp\Big(-\frac{\|\mathbf{Y}-\mu\|_2^2}{1+\gamma}\Big) + 2\exp\Big(-\frac{\|\mathbf{X}-\mu\|_2^2}{2(1+\gamma)}\Big)\exp\Big(-\frac{\|\mathbf{Y}\|_2^2}{2(1+\gamma)}\Big) \\ &\qquad - 2\exp\Big(-\frac{\|\mathbf{X}-\mu\|_2^2+\|\mathbf{X}\|_2^2}{2(1+\gamma)}\Big) - 2\exp\Big(-\frac{\|\mathbf{X}-\mu\|_2^2}{2(1+\gamma)}\Big)\exp\Big(-\frac{\|\mathbf{Y}-\mu\|_2^2}{2(1+\gamma)}\Big) \\ &\qquad - 2\exp\Big(-\frac{\|\mathbf{X}\|_2^2}{2(1+\gamma)}\Big)\exp\Big(-\frac{\|\mathbf{Y}\|_2^2}{2(1+\gamma)}\Big) - 2\exp\Big(-\frac{\|\mathbf{Y}\|_2^2+\|\mathbf{Y}-\mu\|_2^2}{2(1+\gamma)}\Big) \\ &\qquad + 2\exp\Big(-\frac{1}{2(1+\gamma)}\|\mathbf{X}\|_2^2\Big)\exp\Big(-\frac{1}{2(1+\gamma)}\|\mathbf{Y}-\mu\|_2^2\Big)\Big]\;. \end{aligned}$$

We can compute each term by applying Lemma 28. Noting that $\mathbf{Y} - \mu$ and $\mathbf{X}$ are equal in distribution, and also that Lemma 28 depends on $\mathbf{m}_1$ and $\mathbf{m}_2$ only through their difference, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{X}-\mu\|_2^2}{1+\gamma}\right)\right] &= \mathbb{E}\left[\exp\left(-\frac{\|\mathbf{Y}\|_2^2}{1+\gamma}\right)\right] = \left(\frac{1+\gamma}{3+\gamma}\right)^{d/2}, \\
\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{X}\|_2^2}{1+\gamma}\right)\right] &= \mathbb{E}\left[\exp\left(-\frac{\|\mathbf{Y}-\mu\|_2^2}{1+\gamma}\right)\right] = \left(\frac{1+\gamma}{3+\gamma}\right)^{d/2} \exp\left(-\frac{\|\mu\|_2^2}{3+\gamma}\right), \\
\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{X}-\mu\|_2^2}{2(1+\gamma)}\right)\right] &= \mathbb{E}\left[\exp\left(-\frac{\|\mathbf{Y}\|_2^2}{2(1+\gamma)}\right)\right] = \left(\frac{1+\gamma}{2+\gamma}\right)^{d/2}, \\
\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{X}\|_2^2}{2(1+\gamma)}\right)\right] &= \mathbb{E}\left[\exp\left(-\frac{\|\mathbf{Y}-\mu\|_2^2}{2(1+\gamma)}\right)\right] = \left(\frac{1+\gamma}{2+\gamma}\right)^{d/2} \exp\left(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\right).
\end{aligned}$$

It remains to calculate the expectations of the sixth and ninth terms, which involve two differently centred quadratic forms of $\mathbf{X}$ and $\mathbf{Y}$ respectively. The sixth term simplifies to

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{X}-\mu\|_2^2+\|\mathbf{X}\|_2^2}{2(1+\gamma)}\right)\right] &= \mathbb{E}\left[\exp\left(-\frac{1}{2(1+\gamma)}\left(2\|\mathbf{X}\|_2^2 - 2\mu^\top\mathbf{X} + \|\mu\|_2^2\right)\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\frac{1}{1+\gamma}\left\|\mathbf{X}-\frac{\mu}{2}\right\|_2^2\right)\right]\exp\left(-\frac{\|\mu\|_2^2}{4(1+\gamma)}\right) \\
&= \left(\frac{1+\gamma}{3+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mu\|_2^2}{4(3+\gamma)}\right)\exp\left(-\frac{\|\mu\|_2^2}{4(1+\gamma)}\right) \\
&= \left(\frac{1+\gamma}{3+\gamma}\right)^{d/2}\exp\left(-\frac{2+\gamma}{(2(3+\gamma)(1+\gamma)}\|\mu\|_2^2\right), \qquad (40)
\end{aligned}$$

and a similar calculation gives,

$$\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{Y}\|_2^2+\|\mathbf{Y}-\mu\|_2^2}{2(1+\gamma)}\right)\right] = \left(\frac{1+\gamma}{3+\gamma}\right)^{d/2}\exp\left(-\frac{2+\gamma}{(2(3+\gamma)(1+\gamma)}\|\mu\|_2^2\right).$$

Combining the above identities yields

$$\begin{aligned}
\mathbb{E}[g(Z)^2] = &\left(\frac{\gamma}{1+\gamma}\right)^d \left(2\left(\frac{1+\gamma}{3+\gamma}\right)^{d/2} + 2\left(\frac{1+\gamma}{3+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mu\|_2^2}{3+\gamma}\right)\right. \\
&+ 2\left(\frac{1+\gamma}{2+\gamma}\right)^d - 4\left(\frac{1+\gamma}{3+\gamma}\right)^{d/2}\exp\left(-\frac{2+\gamma}{2(3+\gamma)(1+\gamma)}\|\mu\|_2^2\right) \\
&\left. - 4\left(\frac{1+\gamma}{2+\gamma}\right)^d\exp\left(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\right) + 2\left(\frac{1+\gamma}{2+\gamma}\right)^d\exp\left(-\frac{\|\mu\|_2^2}{2+\gamma}\right)\right) \\
= &\, 2\left(\frac{\gamma}{1+\gamma}\right)^{d/2}\left(\frac{\gamma}{3+\gamma}\right)^{d/2} \\
&\times\left(1+\exp\left(-\frac{\|\mu\|_2^2}{3+\gamma}\right) + \left(\frac{3+\gamma}{2+\gamma}\right)^{d/2}\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2} - 2\exp\left(-\frac{(2+\gamma)\|\mu\|_2^2}{2(3+\gamma)(1+\gamma)}\right)\right. \\
&\left. - 2\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\right) + \left(\frac{1+\gamma}{2+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mu\|_2^2}{2+\gamma}\right)\right).
\end{aligned}$$

By noting that $D^{\mathrm{MMD}}(Q,P)^2 = 4\big(\frac{\gamma}{2+\gamma}\big)^d\Big(1-\exp\big(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\big)\Big)^2$, we hence obtain

$$\begin{aligned}
\sigma^2_{\mathrm{cond}} &= \mathbb{E}[g(Z)^2] - D^{\mathrm{MMD}}(Q,P)^2 \;=\; \mathbb{E}[g(Z)^2] - 4\left(\frac{\gamma}{2+\gamma}\right)^d \left[1-\exp\left(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\right)\right]^2 \\
&= 2\left(\frac{\gamma}{1+\gamma}\right)^{d/2}\left(\frac{\gamma}{3+\gamma}\right)^{d/2} \\
&\quad \times \left(1+\exp\left(-\frac{\|\mu\|_2^2}{3+\gamma}\right) + 2\left(\frac{3+\gamma}{2+\gamma}\right)^{d/2}\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\right)\right. \\
&\qquad - 2\exp\left(-\frac{7+5\gamma}{4(1+\gamma)(3+\gamma)}\|\mu\|_2^2\right) - \left(\frac{3+\gamma}{2+\gamma}\right)^{d/2}\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2} \\
&\qquad \left. - \left(\frac{3+\gamma}{2+\gamma}\right)^{d/2}\left(\frac{1+\gamma}{2+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mu\|_2^2}{2+\gamma}\right)\right),
\end{aligned}$$

as required.

#### F.3.4. PROOF FOR $\sigma^2_{\mathrm{full}}$

The second moment is

$$\begin{aligned}
&\mathbb{E}[u^{\mathrm{MMD}}(Z,Z')^2] \\
&= \mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{\gamma}\right) + \exp\left(-\frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{\gamma}\right) + \exp\left(-\frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{\gamma}\right) + \exp\left(-\frac{\|\mathbf{X}'-\mathbf{Y}\|_2^2}{\gamma}\right) \\
&\quad + 2\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{2\gamma} - \frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{2\gamma}\right) - 2\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{2\gamma} - \frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma}\right) \\
&\quad - 2\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{2\gamma} - \frac{\|\mathbf{X}'-\mathbf{Y}\|_2^2}{2\gamma}\right) - 2\exp\left(-\frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{2\gamma} - \frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma}\right) \\
&\quad - 2\exp\left(-\frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{2\gamma} - \frac{\|\mathbf{X}'-\mathbf{Y}\|_2^2}{2\gamma}\right) + 2\exp\left(-\frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma} - \frac{\|\mathbf{X}'-\mathbf{Y}\|_2^2}{2\gamma}\right)\bigg] \\
&= \mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{\gamma}\right) + \exp\left(-\frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{\gamma}\right)\bigg] + 2\mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{\gamma}\right)\bigg] \\
&\quad + 2\mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{2\gamma} - \frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{2\gamma}\right)\bigg] - 4\mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{2\gamma} - \frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma}\right)\bigg] \\
&\quad - 4\mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{2\gamma} - \frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma}\right)\bigg] + 2\left(\mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma}\right)\bigg]\right)^2,
\end{aligned}$$

where the second equality follows by the fact that $X, X'$ and $Y, Y'$ are respectively independent copies from $Q$ and $P$. To calculate each term, we apply Lemma 29 to yield

$$\begin{aligned}
\mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{\gamma}\right)\bigg] &= \mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{\gamma}\right)\bigg] = \left(\frac{\gamma/2}{2+\gamma/2}\right)^{d/2} = \left(\frac{\gamma}{4+\gamma}\right)^{d/2}, \\
\mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{2\gamma}\right)\bigg] &= \mathbb{E}\bigg[\exp\left(-\frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{2\gamma}\right)\bigg] = \left(\frac{\gamma}{2+\gamma}\right)^{d/2}, \\
\mathbb{E}\bigg[e^{-\frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{\gamma}}\bigg] &= \left(\frac{\gamma}{4+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mu\|_2^2}{4+\gamma}\right), \quad \mathbb{E}\bigg[e^{-\frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma}}\bigg] = \left(\frac{\gamma}{2+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\right).
\end{aligned}$$

Let $T_1 \coloneqq \mathbb{E}\Big[\exp\Big(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{2\gamma} - \frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma}\Big)\Big]$ and $T_2 \coloneqq \mathbb{E}\Big[\exp\Big(-\frac{\|\mathbf{Y}-\mathbf{Y}'\|_2^2}{2\gamma} - \frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma}\Big)\Big]$, which are the only remaining terms to compute. The first term can be simplified as

$$\begin{aligned}
T_1 &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}'\|_2^2}{2\gamma}\right)\Big|\mathbf{X}\right]\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{X}-\mathbf{Y}'\|_2^2}{2\gamma}\right)\Big|\mathbf{X}\right]\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\left(\frac{\gamma}{1+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mathbf{X}\|_2^2}{2(1+\gamma)}\right)\times\left(\frac{\gamma}{1+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mathbf{X}-\mu\|_2^2}{2(1+\gamma)}\right)\right] \\
&= \left(\frac{\gamma}{1+\gamma}\right)^{d}\mathbb{E}\left[\exp\left(-\frac{\|\mathbf{X}\|_2^2}{2(1+\gamma)}\right)\exp\left(-\frac{\|\mathbf{X}-\mu\|_2^2}{2(1+\gamma)}\right)\right] \\
&\overset{(b)}{=} \left(\frac{\gamma}{1+\gamma}\right)^{d}\left(\frac{1+\gamma}{3+\gamma}\right)^{d/2}\exp\left(-\frac{2+\gamma}{2(3+\gamma)(1+\gamma)}\|\mu\|_2^2\right) \\
&= \left(\frac{\gamma}{1+\gamma}\right)^{d/2}\left(\frac{\gamma}{3+\gamma}\right)^{d/2}\exp\left(-\frac{2+\gamma}{2(3+\gamma)(1+\gamma)}\|\mu\|_2^2\right) \ ,
\end{aligned}$$

where in $(a)$ we have applied Lemma 28 to compute the conditional expectations, and $(b)$ follows from substituting (40). The second term can be simplified with a similar calculation as,

$$T_2 = \left(\frac{\gamma}{1+\gamma}\right)^{d/2}\left(\frac{\gamma}{3+\gamma}\right)^{d/2}\exp\left(-\frac{2+\gamma}{2(3+\gamma)(1+\gamma)}\|\mu\|_2^2\right) \ .$$

Collecting all terms gives

$$\begin{aligned}
\mathbb{E}[u^{\mathrm{MMD}}(Z,Z')^2] &= 2\left(\frac{\gamma}{4+\gamma}\right)^{d/2} + 2\left(\frac{\gamma}{4+\gamma}\right)^{d/2}\exp\left(-\frac{\|\mu\|_2^2}{4+\gamma}\right) + 2\left(\frac{\gamma}{2+\gamma}\right)^{d} \\
&\quad - 8\left(\frac{\gamma}{1+\gamma}\right)^{d/2}\left(\frac{\gamma}{3+\gamma}\right)^{d/2}\exp\left(-\frac{2+\gamma}{2(3+\gamma)(1+\gamma)}\|\mu\|_2^2\right) \\
&\quad + 2\left(\frac{\gamma}{2+\gamma}\right)^{d}\exp\left(-\frac{\|\mu\|_2^2}{2+\gamma}\right) \ .
\end{aligned}$$

By noting that

$$\begin{aligned}
D^{\mathrm{MMD}}(Q,P)^2 &= 4\Big(\frac{\gamma}{2+\gamma}\Big)^{d}\Big(1-\exp\big(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\big)\Big)^2 \\
&= 4\Big(\frac{\gamma}{2+\gamma}\Big)^{d}\Big(1-\exp\big(-\frac{\|\mu\|_2^2}{2+\gamma}\big) + 2\exp\big(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\big)\Big) \ ,
\end{aligned}$$

the variance takes the following form after subtracting $D^{\mathrm{MMD}}(Q,P)^2$ and collecting similar terms

$$\begin{aligned}
\sigma_{\text{full}}^2 &= \mathbb{E}[u^{\mathrm{MMD}}(Z,Z')^2] - D^{\mathrm{MMD}}(Q,P)^2 \\
&= 2\Big(\frac{\gamma}{4+\gamma}\Big)^{d/2}\Big(1+\exp\big(-\frac{\|\mu\|_2^2}{4+\gamma}\big)\Big) - 2\Big(\frac{\gamma}{2+\gamma}\Big)^{d} \\
&\quad - 8\Big(\frac{\gamma}{1+\gamma}\Big)^{d/2}\Big(\frac{\gamma}{3+\gamma}\Big)^{d/2}\exp\Big(-\frac{2+\gamma}{2(3+\gamma)(1+\gamma)}\|\mu\|_2^2\Big) \\
&\quad - 2\Big(\frac{\gamma}{2+\gamma}\Big)^{d}\exp\Big(-\frac{\|\mu\|_2^2}{2+\gamma}\Big) + 8\Big(\frac{\gamma}{2+\gamma}\Big)^{d}\exp\Big(-\frac{\|\mu\|_2^2}{2(2+\gamma)}\Big) \ ,
\end{aligned}$$

which completes the proof.

### F.4. Proof of Lemma 25

With a linear kernel and under the stated assumption, the MMD statistic is

$$u^{\mathrm{MMD}}(\mathbf{z}, \mathbf{z}') \;=\; \mathbf{x}^\top \mathbf{x}' + \mathbf{y}^\top \mathbf{y}' - \mathbf{x}^\top \mathbf{y}' - \mathbf{y}^\top \mathbf{x}' \;, \qquad \text{where} \quad \mathbf{z} = (\mathbf{x}, \mathbf{y}), \mathbf{z}' = (\mathbf{x}', \mathbf{y}') \in \mathbb{R}^{2d} \;.$$

#### F.4.1. PROOF FOR $g^{\mathrm{mmd}}(\mathbf{z})$ AND $D^{\mathrm{MMD}}(Q, P)$

The expression for $g^{\mathrm{mmd}}$ can be computed as

$$g^{\mathrm{mmd}}(\mathbf{z}) \;= \mathbb{E}[u^{\mathrm{MMD}}(\mathbf{z}, \mathbf{Z}')] \;=\; \mathbb{E}[\mathbf{x}^\top \mathbf{X}' + \mathbf{y}^\top \mathbf{Y}' - \mathbf{x}^\top \mathbf{Y}' - \mathbf{y}^\top \mathbf{X}'] \;=\; \mu^\top \mathbf{x} - \mu^\top \mathbf{y} \;.$$

The formula for $D^{\mathrm{MMD}}(Q, P)$ then follows as

$$D^{\mathrm{MMD}}(Q, P) \;=\; \mathbb{E}[u^{\mathrm{MMD}}(\mathbf{Z}, \mathbf{Z}')] \;=\; \mathbb{E}[g^{\mathrm{mmd}}(\mathbf{Z})] \;=\; \mathbb{E}[\mu^\top \mathbf{X} - \mu^\top \mathbf{Y}] \;=\; \mu^\top \mu \;=\; \|\mu\|_2^2 \;.$$

#### F.4.2. PROOF FOR $\sigma^2_{\mathrm{cond}}$

A direct computation gives

$$\begin{aligned} \mathbb{E}[g^{\mathrm{mmd}}(\mathbf{Z})^2] \;&= \mathbb{E}\Big[(\mu^\top \mathbf{X})^2 + (\mu^\top \mathbf{Y})^2 - 2\mu^\top \mathbf{X} \mu^\top \mathbf{Y}\Big] \\ &= \mu^\top (\Sigma + \mu\mu^\top)\mu + \mu^\top \Sigma \mu \;=\; 2\mu^\top \Sigma \mu + \|\mu\|_2^4 \;. \end{aligned}$$

Therefore, $\sigma^2_{\mathrm{cond}} = \mathbb{E}[g^{\mathrm{mmd}}(\mathbf{Z})^2] - D^{\mathrm{MMD}}(Q, P)^2 = 2\mu^\top \Sigma \mu$, as required.

#### F.4.3. PROOF FOR $\sigma^2_{\mathrm{full}}$

The second moment is

$$\begin{aligned} \mathbb{E}[u^{\mathrm{MMD}}(\mathbf{Z}, \mathbf{Z}')^2] \;&= \mathbb{E}\big[(\mathbf{X}^\top \mathbf{X}')^2 + (\mathbf{Y}^\top \mathbf{Y}')^2 + (\mathbf{X}^\top \mathbf{Y}')^2 + (\mathbf{Y}^\top \mathbf{X}')^2 \\ &\qquad + 2\mathbf{X}^\top \mathbf{X}' \mathbf{Y}^\top \mathbf{Y}' - 2\mathbf{X}^\top \mathbf{X}' \mathbf{X}^\top \mathbf{Y}' - 2\mathbf{X}^\top \mathbf{X}' \mathbf{Y}^\top \mathbf{X}' \\ &\qquad - 2\mathbf{Y}^\top \mathbf{Y}' \mathbf{X}^\top \mathbf{Y}' - 2\mathbf{Y}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{X}' + 2\mathbf{X}^\top \mathbf{Y}' \mathbf{Y}^\top \mathbf{X}'\big] \\ &= \mathbb{E}\big[(\mathbf{X}^\top \mathbf{X}')^2 + (\mathbf{Y}^\top \mathbf{Y}')^2 + (\mathbf{X}^\top \mathbf{Y}')^2 + (\mathbf{Y}^\top \mathbf{X}')^2\big] \;. \end{aligned}$$

In the last equality, we have noted that the cross-terms vanish since $\mathbf{X}, \mathbf{X}', \mathbf{Y}$ and $\mathbf{Y}'$ are mutually independent and $\mathbf{X}, \mathbf{X}'$ are zero-mean. A direct computation gives

$$\begin{aligned} \mathbb{E}\big[(\mathbf{X}^\top \mathbf{X}')^2\big] \;&= \mathrm{Tr}\big(\mathbb{E}[\mathbf{X}\mathbf{X}^\top \mathbf{X}'(\mathbf{X}')^\top]\big) \;=\; \mathrm{Tr}(\Sigma^2) \;, \\ \mathbb{E}\big[(\mathbf{Y}^\top \mathbf{Y}')^2\big] \;&= \mathrm{Tr}\big(\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top \mathbf{Y}'(\mathbf{Y}')^\top]\big) \;=\; \mathrm{Tr}\big((\Sigma + \mu\mu^\top)^2\big) \;=\; \mathrm{Tr}(\Sigma^2) + 2\mu^\top \Sigma \mu^\top + \|\mu\|_2^4 \;, \\ \mathbb{E}\big[(\mathbf{X}^\top \mathbf{Y}')^2\big] \;&= \mathbb{E}\big[(\mathbf{Y}^\top \mathbf{X}')^2\big] \;=\; \mathrm{Tr}(\mathbb{E}[\mathbf{Y}(\mathbf{Y}')^\top \mathbf{X}(\mathbf{X}')^\top]) \;=\; \mathrm{Tr}(\Sigma^2) + \mu^\top \Sigma \mu \;. \end{aligned}$$

Therefore, $\mathbb{E}[u^{\mathrm{MMD}}(\mathbf{Z}, \mathbf{Z}')^2] = 4\mathrm{Tr}(\Sigma^2) + 4\mu^\top \Sigma \mu + \|\mu\|_2^4$, and

$$\sigma^2_{\mathrm{full}} \;=\; \mathbb{E}[u^{\mathrm{MMD}}(\mathbf{Z}, \mathbf{Z}')^2] - D^{\mathrm{MMD}}(Q, P)^2 \;=\; 4\mathrm{Tr}(\Sigma^2) + 4\mu^\top \Sigma \mu \;,$$

which completes the proof.

### F.4.4. PROOF FOR UPPER BOUND ON $M^3_{\mathrm{cond};3}$

The 3rd absolute centred moment of $g^{\mathrm{mmd}}(\mathbf{Z})$ satisfies

$$M^3_{\mathrm{cond};3} = \mathbb{E}[|g^{\mathrm{mmd}}(\mathbf{Z}) - \mathbb{E}[g^{\mathrm{mmd}}(\mathbf{Z})]|^3] = \mathbb{E}[|\mu^\top \mathbf{Y} - \mu^\top \mathbf{X} - \mu^\top \mu|^3] = \mathbb{E}[|\mu^\top \mathbf{Y} - \mu^\top \mathbf{V}|^3] ,$$

where we have defined $\mathbf{V} \coloneqq \mathbf{X} - \mu$ so that $\mathbf{V} \sim \mathcal{N}(0, \Sigma)$. Noting that $|a+b|^3 \leq 2^{3-1}(|a|^3 + |b|^3)$ for any $a, b \in \mathbb{R}$ by Jensen's inequality, we can bound the above as

$$M^3_{\mathrm{cond};3} = \mathbb{E}\big[\big|\mu^\top \mathbf{Y} - \mu^\top \mathbf{V}\big|^m\big] \leq 4\mathbb{E}[|\mu^\top \mathbf{Y}|^3 + |\mu^\top \mathbf{V}|^3] \overset{(a)}{=} 8C'(\mu^\top \Sigma \mu)^{3/2} \overset{(b)}{=} C(\mu^\top \Sigma\mu)^{3/2} .$$

In $(a)$ we have noted that the absolute 3rd moment of a univariate normal variable $\mathcal{N}(0, \sigma^2)$ is given as $C'\sigma^3$ for some absolute constant $C'$. In $(b)$, we have defined $C \coloneqq 8C'$.

### F.4.5. PROOF FOR UPPER BOUND ON $M^3_{\mathrm{full};3}$

For any $\mathbf{z} = (\mathbf{x}, \mathbf{y}), \mathbf{z}' = (\mathbf{x}', \mathbf{z}') \in \mathbb{R}^{2d}$ we have

$$u^{\mathrm{MMD}}(\mathbf{z}, \mathbf{z}') \ = \mathbf{x}^\top \mathbf{x}' + \mathbf{y}^\top \mathbf{y} - \mathbf{x}^\top \mathbf{y}' - \mathbf{y}^\top \mathbf{x}' \ = \ (\mathbf{x} - \mathbf{y})^\top (\mathbf{x}' - \mathbf{y}') .$$

Write $\mathbf{V} \coloneqq \mathbf{X} - \mu$ and $\mathbf{V}' \coloneqq \mathbf{X} - \mu$ so that $\mathbf{V}, \mathbf{V}' \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$. We can compute the 3rd absolute central moment as

$$\begin{aligned} M^3_{\mathrm{full};3} &= \mathbb{E}[|u^{\mathrm{MMD}}(\mathbf{Z}, \mathbf{Z}') - \mathbb{E}[u^{\mathrm{MMD}}(\mathbf{Z}, \mathbf{Z}')]|^3] \\ &= \mathbb{E}[|(\mathbf{X} - \mathbf{Y})^\top (\mathbf{X}' - \mathbf{Y}') - \mu^\top \mu|^3] \\ &= \mathbb{E}[|(\mathbf{V} + \mu - \mathbf{Y})^\top (\mathbf{V}' + \mu - \mathbf{Y}') - \mu^\top \mu|^3] \\ &= \mathbb{E}[|(\mathbf{V} - \mathbf{Y})^\top (\mathbf{V}' - \mathbf{Y}') + \mu^\top (\mathbf{V}' - \mathbf{Y}') + (\mathbf{V} - \mathbf{Y})^\top \mu|^3] . \end{aligned}$$

By a Jensen's inequality applied to the convex function $x \mapsto |x|^3$ and a Hölder's inequality, we get that

$$\begin{aligned} M^3_{\mathrm{full};3} &\leq 9\big(\mathbb{E}\big[|(\mathbf{V} - \mathbf{Y})^\top (\mathbf{V}' - \mathbf{Y}')|^3\big] + \mathbb{E}\big[|\mu^\top (\mathbf{V}' - \mathbf{Y}')|^3\big] + \mathbb{E}\big[|(\mathbf{V} - \mathbf{Y})^\top \mu|^3\big]\big) \\ &\leq 9\big(\mathbb{E}\big[|\mathbf{U}^\top \mathbf{U}'|^3\big] + 2\mathbb{E}\big[|\mathbf{U}^\top \mu|^3\big]\big) . \end{aligned}$$

In the last line, we have used that $\mathbf{U} \coloneqq \mathbf{V} - \mathbf{Y}$ and $\mathbf{U}' \coloneqq \mathbf{V}' - \mathbf{Y}'$ are identically distributed. In fact they are both $\mathcal{N}(\mathbf{0}, 2\Sigma)$. The second expectation can be computed by the formula for the absolute 3rd moment of a univariate Gaussian as

$$\mathbb{E}\big[|\mathbf{U}^\top \mu|^3\big] \ = \ C'(\mu^\top \Sigma \mu)^{3/2}$$

where $C'$ is some absolute constant. Similarly the first expectation can be computed first by noting that $\mathbf{U}^\top \mathbf{U}'$ conditioning on $\mathbf{U}$ is a univariate Gaussian and secondly by using the moment formula for a Gaussian quadratic form Lemma 27:

$$\begin{aligned} \mathbb{E}\big[|\mathbf{U}^\top \mathbf{U}'|^3\big] \ &= \ \mathbb{E}\big[\mathbb{E}[|\mathbf{U}^\top \mathbf{U}'|^3 | \mathbf{U}]\big] \ = \ C'\mathbb{E}\big[(\mathbf{U}^\top \Sigma \mathbf{U})^{3/2}\big] \\ &\leq C'\mathbb{E}\big[(\mathbf{U}^\top \Sigma \mathbf{U})^3\big]^{1/2} = C'\big(\mathrm{Tr}(\Sigma^2)^3 + 6\mathrm{Tr}(\Sigma^2)\mathrm{Tr}(\Sigma^4) + 8\mathrm{Tr}(\Sigma^6)\big) \leq 15C'\mathrm{Tr}(\Sigma^2)^3 . \end{aligned}$$

In the last line, we have noted that $\text{Tr}(A^m) \leq \text{Tr}(A)^m$ for $m \in \mathbb{N}$ and positive semi-definite matrix $A$, which holds by expressing each trace as a sum of eigenvalues and applying the Hölder's inequality. Combining the two computations and redefining constants, we get that for some constant $C$,

$$M^3_{\text{full};3} \leq C\big(\text{Tr}(\Sigma^2)^3 + (\mu^\top \Sigma \mu)^{3/2}\big) \;\leq\; C\big(\text{Tr}(\Sigma^2) + \mu^\top \Sigma \mu\big)^{3/2} .$$

F.4.6. PROOF FOR VERIFYING ASSUMPTION 1

By the bounds from *(iii)-(vi)*, there exists absolute constants $C_1, C_2$ such that

$$\frac{M_{\text{cond};3}}{\sigma_{\text{cond}}} \;\leq\; \frac{C_1^{1/3}(\mu^\top\Sigma\mu)^{1/2}}{2^{1/2}(\mu^\top\Sigma\mu)^{1/2}} \;=\; 2^{-1/2}C_1^{1/3} , \quad \frac{M_{\text{full};3}}{\sigma_{\text{full}}} \;\leq\; \frac{C_2^{1/3}\big(\text{Tr}(\Sigma^2)+\mu^\top\Sigma\mu\big)^{1/2}}{2\big(\text{Tr}(\Sigma^2)+\mu^\top\Sigma\mu\big)^{1/2}} \;=\; 2^{-1}C_2^{1/3} ,$$

which prove that Assumption 1 holds with $\nu = 3$.

## Appendix G. Proofs for Appendix B

### G.1. Proofs for Appendix B.1

The proof of Lemma 26 combines the following two results:

**Lemma 44 (Theorem 2, von Bahr and Esseen (1965))** *Fix $\nu \in [1,2]$. For a martingale difference sequence $Y_1, \ldots, Y_n$ taking values in $\mathbb{R}$,*

$$\mathbb{E}\big[\big|\textstyle\sum_{i=1}^n Y_i\big|^\nu\big] \;\leq\; 2\textstyle\sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] .$$

**Lemma 45 (Dharmadhikari et al. (1968))** *Fix $\nu \geq 2$. For a martingale difference sequence $Y_1, \ldots, Y_n$ taking values in $\mathbb{R}$,*

$$\mathbb{E}\big[\big|\textstyle\sum_{i=1}^n Y_i\big|^\nu\big] \;\leq\; C_\nu n^{\nu/2-1}\textstyle\sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] ,$$

*where $C_\nu = (8(\nu-1)\max\{1, 2^{\nu-3}\})^\nu$.*

**Proof of Lemma 26** We first consider the upper bound. For $\nu \in [1,2]$, the result follows directly from the Von Bahn-Esseen inequality as stated below in Lemma 44, and for $\nu > 1$, the result follows directly from Lemma 45. As for the lower bound, by Theorem 9 of Burkholder (1966), there exists an absolute constant $c_\nu > 0$ depending only on $\nu$ such that

$$\mathbb{E}\big[\big|\textstyle\sum_{i=1}^n Y_i\big|^\nu\big] \;\geq\; c_\nu\,\mathbb{E}\big[\big(\textstyle\sum_{i=1}^n Y_i^2\big)^{\nu/2}\big] .$$

For $\nu \in [1,2]$, by applying Jensen's inequality on the concave function $x \mapsto x^{\nu/2}$, we get that

$$\mathbb{E}\big[\big|\textstyle\sum_{i=1}^n Y_i\big|^\nu\big] \geq c_\nu\,\mathbb{E}\big[\big(\frac{1}{n}\textstyle\sum_{i=1}^n nY_i^2\big)^{\nu/2}\big] \;\geq\; c_\nu\, n^{\nu/2-1}\textstyle\sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] .$$

For $\nu > 2$, by noting that $(a+b)^{\nu/2} \geq a^{\nu/2} + b^{\nu/2}$ for $a, b \geq 0$, we get that

$$\mathbb{E}\big[\big|\textstyle\sum_{i=1}^n Y_i\big|^\nu\big] \geq c_\nu\,\mathbb{E}\big[\textstyle\sum_{i=1}^n \big(Y_i^2\big)^{\nu/2}\big] \;\geq\; c_\nu \textstyle\sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] .$$

Combining the two results above give the desired bound. ■

**Proof of Lemma 28** A direct computation gives

$$
\begin{aligned}
&\mathbb{E}\left[f(\mathbf{X})\exp\left(-\frac{1}{2a_2^2}\|\mathbf{X}-\mathbf{m}_2\|_2^2\right)\right] \\
=\;& \frac{1}{(2\pi)^{d/2}a_1^d}\int f(\mathbf{x})\exp\left(-\frac{1}{2a_2^2}\|\mathbf{x}-\mathbf{m}_2\|_2^2\right)\exp\left(-\frac{1}{2a_1^2}\|\mathbf{x}-\mathbf{m}_1\|_2^2\right)d\mathbf{x} \\
=\;& \frac{1}{(2\pi)^{d/2}a_1^d}\int f(\mathbf{x})\exp\Bigg(\underbrace{-\frac{1}{2}\left(\frac{\|\mathbf{x}\|_2^2}{a_2^2}+\frac{\|\mathbf{m}_2\|_2^2}{a_2^2}-\frac{2\mathbf{m}_2^\top\mathbf{x}}{a_2^2}+\frac{\|\mathbf{x}\|_2^2}{a_1^2}+\frac{\|\mathbf{m}_1\|_2^2}{a_1^2}-\frac{2\mathbf{m}_1^\top\mathbf{x}}{a_1^2}\right)}_{=:T}\Bigg)d\mathbf{x}\,.
\end{aligned}
$$

Simplifying $T$ by completing the square yields

$$
\begin{aligned}
T \;=\;& -\frac{1}{2}\left(\frac{\|\mathbf{x}\|_2^2}{a_2^2}+\frac{\|\mathbf{x}\|_2^2}{a_1^2}-\frac{2\mathbf{m}_2^\top\mathbf{x}}{a_2^2}-\frac{2\mathbf{m}_1^\top\mathbf{x}}{a_1^2}\right)-\frac{1}{2}\left(\frac{\|\mathbf{m}_2\|_2^2}{a_2^2}+\frac{\|\mathbf{m}_1\|_2^2}{a_1^2}\right) \\
=\;& -\frac{a_1^2+a_2^2}{2a_1^2a_2^2}\left(\|\mathbf{x}\|_2^2-\frac{2a_1^2a_2^2}{a_1^2+a_2^2}\left(\frac{\mathbf{m}_2}{a_2^2}+\frac{\mathbf{m}_1}{a_1^2}\right)^\top\mathbf{x}+\frac{a_1^4a_2^4}{(a_1^2+a_2^2)^2}\left\|\frac{\mathbf{m}_2}{a_2^2}+\frac{\mathbf{m}_1}{a_1^2}\right\|_2^2\right) \\
& -\frac{1}{2}\underbrace{\left(\frac{\|\mathbf{m}_2\|_2^2}{a_2^2}+\frac{\|\mathbf{m}_1\|_2^2}{a_1^2}-\frac{a_1^2a_2^2}{a_1^2+a_2^2}\left\|\frac{\mathbf{m}_2}{a_2^2}+\frac{\mathbf{m}_1}{a_1^2}\right\|_2^2\right)}_{=:T'} \\
=\;& -\frac{a_1^2+a_2^2}{2a_1^2a_2^2}\left\|\mathbf{x}-\frac{a_1^2a_2^2}{a_1^2+a_2^2}\left(\frac{\mathbf{m}_1}{a_1^2}+\frac{\mathbf{m}_2}{a_2^2}\right)\right\|_2^2-\frac{1}{2(a_1^2+a_2^2)}\|\mathbf{m}_1-\mathbf{m}_2\|_2^2\,,
\end{aligned}
$$

where we have simplified $T'$ as

$$
\begin{aligned}
T' \;=\;& \frac{\|\mathbf{m}_2\|_2^2}{a_2^2}+\frac{\|\mathbf{m}_1\|_2^2}{a_1^2}-\frac{a_1^2}{a_2^2(a_1^2+a_2^2)}\|\mathbf{m}_2\|_2^2-\frac{a_2^2}{a_1^2(a_1^2+a_2^2)}\|\mathbf{m}_1\|_2^2+\frac{2}{a_1^2+a_2^2}\mathbf{m}_1^\top\mathbf{m}_2 \\
=\;& \frac{1}{a_1^2+a_2^2}\|\mathbf{m}_1-\mathbf{m}_2\|_2^2\,.
\end{aligned}
$$

Substituting this into $\mathbb{E}\left[f(\mathbf{X})\exp\left(-\frac{1}{2a_2^2}\|\mathbf{X}-\mathbf{m}_2\|_2^2\right)\right]$, we have

$$
\begin{aligned}
&\mathbb{E}\left[f(\mathbf{X})\exp\left(-\frac{1}{2a_2^2}\|\mathbf{X}-\mathbf{m}_2\|_2^2\right)\right] \\
=\;& \frac{1}{(2\pi)^{d/2}a_1^d}\exp\left(-\frac{\|\mathbf{m}_1-\mathbf{m}_2\|_2^2}{2(a_1^2+a_2^2)}\right)\int f(x)\exp\left(-\frac{a_1^2+a_2^2}{2a_1^2a_2^2}\left\|\mathbf{x}-\frac{a_1^2a_2^2}{a_1^2+a_2^2}\left(\frac{\mathbf{m}_1}{a_1^2}+\frac{\mathbf{m}_2}{a_2^2}\right)\right\|_2^2\right)dx \\
=\;& \left(\frac{a_2^2}{a_1^2+a_2^2}\right)^{d/2}\exp\left(-\frac{\|\mathbf{m}_1-\mathbf{m}_2\|_2^2}{2(a_1^2+a_2^2)}\right)\mathbb{E}[f(\mathbf{W})]\,,
\end{aligned}
$$

where $\mathbf{W}\sim\mathcal{N}\left(\frac{a_1^2a_2^2}{a_1^2+a_2^2}\left(\frac{\mathbf{m}_1}{a_1^2}+\frac{\mathbf{m}_2}{a_2^2}\right),\frac{a_1^2a_2^2}{a_1^2+a_2^2}I_d\right)$, which completes the proof. ■

**Proof of Lemma 29** Rewriting by the tower rule,

$$\mathbb{E}\Big[f(\mathbf{X},\mathbf{X}')\exp\Big(-\frac{1}{2a_3^2}\|\mathbf{X}-\mathbf{X}'\|_2^2\Big)\Big] \;=\; \mathbb{E}\Big[\mathbb{E}\Big[f(\mathbf{X},\mathbf{X}')\exp\Big(-\frac{1}{2a_3^2}\|\mathbf{X}-\mathbf{X}'\|_2^2\Big)\Big|\mathbf{X}\Big]\Big]$$
$$= \mathbb{E}\left[\left(\frac{a_3^2}{a_2^2+a_3^2}\right)^{d/2}\exp\Big(-\frac{1}{2(a_2^2+a_3^2)}\|\mathbf{X}-\mathbf{m}_2\|_2^2\Big)\,\mathbb{E}\Big[f\Big(\mathbf{X},\mathbf{W}'+\frac{a_2^2}{a_2^2+a_3^2}\mathbf{X}\Big)\Big|\mathbf{X}\Big]\right] ,$$

where the last line follows by applying Lemma 28 to the inner expectation, and where $\mathbf{W}' \sim \mathcal{N}\Big(\frac{a_3^2}{a_2^2+a_3^2}\mathbf{m}_2, \frac{a_2^2a_3^2}{a_2^2+a_3^2}I_d\Big)$. Applying Lemma 28 again gives

$$\mathbb{E}\Big[f(\mathbf{X},\mathbf{X}')\exp\Big(-\frac{1}{2a_3^2}\|\mathbf{X}-\mathbf{X}'\|_2^2\Big)\Big]$$
$$= \left(\frac{a_3^2}{a_2^2+a_3^2}\right)^{d/2}\left(\frac{a_2^2+a_3^2}{a_1^2+a_2^2+a_3^2}\right)^{d/2}\exp\Big(-\frac{1}{2(a_1^2+a_2^2+a_3^2)}\|\mathbf{m}_1-\mathbf{m}_2\|_2^2\Big)$$
$$\times\,\mathbb{E}\Big[\mathbb{E}\Big[f(\mathbf{W},\mathbf{W}'+\frac{a_2^2}{a_2^2+a_3^2}\mathbf{W})\Big|\mathbf{W}\Big]\Big]$$
$$= \left(\frac{a_3^2}{a_1^2+a_2^2+a_3^2}\right)^{d/2}\exp\Big(-\frac{1}{2(a_1^2+a_2^2+a_3^2)}\|\mathbf{m}_1-\mathbf{m}_2\|_2^2\Big)\,\mathbb{E}\Big[f\Big(\mathbf{W},\mathbf{W}'+\frac{a_2^2}{a_2^2+a_3^2}\mathbf{W}\Big)\Big] ,$$

where $\mathbf{W}\sim\mathcal{N}\left(\frac{a_1^2(a_2^2+a_3^2)}{a_1^2+a_2^2+a_3^2}\Big(\frac{1}{a_1^2}\mathbf{m}_1+\frac{1}{a_2^2+a_3^2}\mathbf{m}_2\Big), \frac{a_1^2(a_2^2+a_3^2)}{a_1^2+a_2^2+a_3^2}I_d\right)$. ■

## G.2. Proofs for Appendix B.2

**Proof of Lemma 30** Consider the sequence of sigma algebras with $\mathcal{F}_0$ being the trivial sigma algebra and $\mathcal{F}_i := \sigma(X_1,\ldots,X_i)$ for $i=1,\ldots,n$. This allows us to define a martingale difference sequence: For $i=1,\ldots,n$, let

$$Y_i \;:=\; \mathbb{E}[D_n|\mathcal{F}_i]-\mathbb{E}[D_n|\mathcal{F}_{i-1}] \,.$$

This implies that $\mathbb{E}[|D_n-\mathbb{E}D_n|^\nu]=\mathbb{E}\big[\big|\sum_{i=1}^n Y_i\big|^\nu\big]$. By Lemma 26, we get that for some universal constants $c'_\nu, C'_\nu$,

$$c'_\nu\sum_{i=1}^n\mathbb{E}[|Y_i|^\nu] \;\leq\; \mathbb{E}[|D_n-\mathbb{E}D_n|^\nu] \;\leq\; C'_\nu\, n^{\nu/2-1}\sum_{i=1}^n\mathbb{E}[|Y_i|^\nu]\,. \tag{41}$$

To compute the $\nu$-th moment of $Y_i$, recall that $D_n=\frac{1}{n(n-1)}\sum_{j,l\in[n],j\neq l}u(\mathbf{X}_j,\mathbf{X}_l)$, which implies

$$\begin{aligned}\mathbb{E}[|Y_i|^\nu] &= \mathbb{E}\big[\big|\mathbb{E}[D_n|\mathcal{F}_i]-\mathbb{E}[D_n|\mathcal{F}_{i-1}]\big|^\nu\big]\\
&= \frac{1}{n^\nu(n-1)^\nu}\mathbb{E}\Big[\Big|\sum_{j,l\in[n],j\neq l}\big(\mathbb{E}[u(\mathbf{X}_j,\mathbf{X}_l)|\mathcal{F}_i]-\mathbb{E}[u(\mathbf{X}_j,\mathbf{X}_l)|\mathcal{F}_{i-1}]\big)\Big|^\nu\Big]\\
&\overset{(a)}{=} \frac{2}{n^\nu(n-1)^\nu}\mathbb{E}\Big[\Big|\sum_{j\in[n],j\neq i}\big(\mathbb{E}[u(\mathbf{X}_i,\mathbf{X}_j)|\mathcal{F}_i]-\mathbb{E}[u(\mathbf{X}_i,\mathbf{X}_j)|\mathcal{F}_{i-1}]\big)\Big|^\nu\Big]\\
&=: \frac{2}{n^\nu(n-1)^\nu}\,\mathbb{E}[|S_i|^\nu]\,.\end{aligned}$$

In $(a)$, we have used that each summand is zero if both $j$ and $l$ do not equal $i$, and that $u$ is symmetric. In the case $j < i$, we can compute each summand of $S_i$ as

$$\begin{aligned}\mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_j)|\mathcal{F}_i] - \mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_j)|\mathcal{F}_{i-1}] &= u(\mathbf{X}_i, \mathbf{X}_j) - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_j)|\mathbf{X}_j] \\ &= A_{ij} - B_j + B_i \;,\end{aligned}$$

where $A_{ij} := u(\mathbf{X}_i, \mathbf{X}_j) - \mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_1)|\mathbf{X}_i]$ and

$$B_i \;:=\; \mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_1)|\mathbf{X}_i] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)] \;=\; \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_i)|\mathbf{X}_i] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)]$$

by symmetry of $u$. In the case $j > i$, we can compute each summand as

$$\mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_j)|\mathcal{F}_i] - \mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_j)|\mathcal{F}_{i-1}] \;=\; \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_i)|\mathbf{X}_i] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)] \;=\; B_i \;.$$

Therefore

$$S_i \;=\; \textstyle\sum_{j<i}(A_{ij} - B_j) + nB_i \;.$$

Consider $R_1 := nB_i$ and $R_2 := \sum_{j<i}(A_{ij} - B_j)$, which forms a two-element martingale difference sequence with respect to the filtration $\sigma(\mathbf{X}_i) \subseteq \sigma(\mathbf{X}_i, \mathbf{X}_1 \ldots, \mathbf{X}_{i-1})$. By Lemma 26 again, there exist constants $c^*_\nu$ and $C^*_\nu$ depending only on $\nu$ such that

$$\begin{aligned}\mathbb{E}[|S_i|^\nu] \;&=\; \mathbb{E}\big[\big|\textstyle\sum_{l=1}^2 R_l\big|^\nu\big] \;\leq\; C^*_\nu\Big(\mathbb{E}[|nB_i|^\nu] + \mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big]\Big) \\ &\qquad\qquad\qquad\quad\;=\; C^*_\nu\big(n^\nu M^\nu_{\text{cond};\nu} + \mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big]\big) \;, \\ \mathbb{E}[|S_i|^\nu] \;&=\; \mathbb{E}\big[\big|\textstyle\sum_{l=1}^2 R_l\big|^\nu\big] \;\geq\; c^*_\nu\Big(\mathbb{E}[|nB_i|^\nu] + \mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big]\Big) \\ &\qquad\qquad\qquad\quad\;=\; c^*_\nu\big(n^\nu M^\nu_{\text{cond};\nu} + \mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big]\big) \;.\end{aligned}$$

Now consider $T_j := A_{ij} - B_j$ for $j = 1, \ldots, i-1$, which again forms a martingale difference sequence with respect to $\sigma(\mathbf{X}_i, \mathbf{X}_1), \ldots, \sigma(\mathbf{X}_i, \mathbf{X}_1 \ldots, \mathbf{X}_{i-1})$. Then by Lemma 26 again, there exist constants $c^\Delta_\nu$ and $C^\Delta_\nu$ depending only on $\nu$ such that

$$\begin{aligned}\mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big] \;&\leq\; C^\Delta_\nu\,(i-1)^{\nu/2-1}\textstyle\sum_{j=1}^{i-1}\mathbb{E}[|A_{ij} - B_j|^\nu] \;=\; C^\Delta_\nu\,(i-1)^{\nu/2}M^\nu_{\text{full};\nu} \;, \\ \mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big] \;&\geq\; c^\Delta_\nu\,\textstyle\sum_{j=1}^{i-1}\mathbb{E}[|A_{ij} - B_j|^\nu] \;=\; c^\Delta_\nu\,(i-1)M^\nu_{\text{full};\nu} \;.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}[|S_i|^\nu] \;&\leq C^*_\nu n^\nu M^\nu_{\text{cond};\nu} + C^*_\nu C^\Delta_\nu\,(i-1)^{\nu/2}M^\nu_{\text{full};\nu} \;, \\ \mathbb{E}[|S_i|^\nu] \;&\geq c^*_\nu n^\nu M^\nu_{\text{cond};\nu} + c^*_\nu c^\Delta_\nu\,(i-1)M^\nu_{\text{full};\nu} \;,\end{aligned}$$

which yield the following bounds on the $\nu$-th moment of $Y_i$:

$$\begin{aligned}\mathbb{E}[|Y_i|^\nu] \;&\leq\; 2C^*_\nu\big((n-1)^{-\nu}M^\nu_{\text{cond};\nu} + C^\Delta_\nu n^{-\nu}(n-1)^{-\nu}\,(i-1)^{\nu/2}M^\nu_{\text{full};\nu}\big) \;, \\ \mathbb{E}[|Y_i|^\nu] \;&\geq\; 2c^*_\nu\big((n-1)^{-\nu}M^\nu_{\text{cond};\nu} + c^\Delta_\nu n^{-\nu}(n-1)^{-\nu}\,(i-1)M^\nu_{\text{full};\nu}\big) \;,\end{aligned}$$

To sum these terms over $i = 1, \ldots, n$, we note that since $\nu/2 > 0$,

$$\sum_{i=1}^{n} (i-1)^{\nu/2} \;\leq\; \int_0^n x^{\nu/2} dx \;=\; \frac{n^{1+\nu/2}}{1+\nu/2} \;, \qquad \sum_{i=1}^{n} (i-1) \;=\; \frac{n(n-1)}{2} \;.$$

Define $C_\nu := \frac{2C'_\nu C^*_\nu \max\{1, C^\Delta_\nu\}}{1+\nu/2}$ and $c_\nu := c'_\nu c^*_\nu \min\{1, c^\Delta_\nu\}$. By summing the bounds on $\mathbb{E}[|Y_i|^\nu]$ and substituting into (41), we get the desired bounds

$$\begin{aligned}
\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] \;&\leq\; C_\nu\, n^{\nu/2-1} \big( n(n-1)^{-\nu} M^\nu_{\text{cond};\nu} + n^{-\nu}(n-1)^{-\nu} n^{1+\nu/2} M^\nu_{\text{full};\nu} \big) \\
&=\; C_\nu\, n^{\nu/2} (n-1)^{-\nu} M^\nu_{\text{cond};\nu} + C_\nu\, (n-1)^{-\nu} M^\nu_{\text{full};\nu} \;, \\
\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] \;&\geq\; c_\nu n (n-1)^{-\nu} M^\nu_{\text{cond};\nu} + c_\nu n^{-(\nu-1)} (n-1)^{-(\nu-1)} M^\nu_{\text{full};\nu} \;.
\end{aligned}$$

■

**Proof of Lemma 31** The first result is directly obtained from linearity of expectation and Jensen's inequality:

$$\begin{aligned}
\Big| D - \sum_{k=1}^{K} \lambda_k \mu_k^2 \Big| \;&=\; \Big| \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)] - \sum_{k=1}^{K} \lambda_k \mathbb{E}[\phi_k(\mathbf{X}_1)]\mathbb{E}[\phi_k(\mathbf{X}_2)] \Big| \\
&=\; \Big| \mathbb{E}\Big[ u(\mathbf{X}_1, \mathbf{X}_2) - \sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) \Big] \Big| \\
&\leq \mathbb{E}\Big| u(\mathbf{X}_1, \mathbf{X}_2) - \sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) \Big| \;=\; \varepsilon_{K;1} \;.
\end{aligned}$$

To prove the next few bounds, we first derive a useful inequality: For $a, b \in \mathbb{R}$ and $\nu \geq 1$, by Jensen's inequality, we have

$$|a+b|^\nu \;=\; \big| \tfrac{1}{2}(2a) + \tfrac{1}{2}(2b) \big|^\nu \;\leq\; \tfrac{1}{2}|2a|^\nu + \tfrac{1}{2}|2b|^\nu \;=\; 2^{\nu-1}(|a|^\nu + |b|^\nu) \;.$$

By a triangle inequality followed by applying the above inequality again with $a$ replaced by $|a| - |b|$ and $b$ replaced by $|b|$, we have

$$|a+b|^\nu \;\geq\; ||a| - |b||^\nu \;\geq\; 2^{-(\nu-1)}|a|^\nu - |b|^\nu \;.$$

Since $\nu \in [1, 3]$, we have $2^{\nu-1} \in [1, 4]$. Therefore

$$\frac{1}{4}|a|^\nu - |b|^\nu \;\leq\; |a+b|^\nu \;\leq\; 4(|a|^\nu + |b|^\nu) \;. \tag{42}$$

Now to prove the conditional bound, we make use of the fact that $\mathbf{X}_1, \mathbf{X}_2$ are i.i.d. to see that

$$\begin{aligned}
&\mathbb{E}\Big[ \Big| \sum_{k=1}^{K} \lambda_k (\phi_k(\mathbf{X}_1) - \mu_k)\mu_k \Big|^\nu \Big] \\
&= \mathbb{E}\Big[ \Big| \sum_{k=1}^{K} \lambda_k \big( \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)|\mathbf{X}_1] - \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)] \big) \Big|^\nu \Big] \\
&= \mathbb{E}\Big[ \Big| \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)] + \Delta_{K;1} - \Delta_{K;2} \Big|^\nu \Big] \;,
\end{aligned} \tag{43}$$

where

$$\begin{aligned}
\Delta_{K;1} \;&:=\; \sum_{k=1}^{K} \lambda_k \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)|\mathbf{X}_1] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1] \;, \\
\Delta_{K;2} \;&:=\; \sum_{k=1}^{K} \lambda_k \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)] \;.
\end{aligned}$$

Moments of the two error terms can be bounded by Jensen's inequality applied to $x \mapsto |x|^\nu$ with respect to the conditional expectation $\mathbb{E}[\,\bullet\,|\mathbf{X}_2]$ and the expectation $\mathbb{E}[\,\bullet\,]$:

$$\begin{aligned}\mathbb{E}[|\Delta_{K;1}|^\nu], \mathbb{E}[|\Delta_{K;2}|^\nu] &\leq \mathbb{E}\Big[\Big|u(\mathbf{X}_1,\mathbf{X}_2) - \textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\Big|^\nu\Big] \\ &= \Big\|u(\mathbf{X}_1,\mathbf{X}_2) - \textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\Big\|_{L_\nu}^\nu \;=\; \varepsilon_{K;\nu}^\nu \;.\end{aligned}$$

On the other hand,

$$(M_{\mathrm{cond};\nu})^\nu \;=\; \mathbb{E}\Big[\Big|\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_1] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)]\Big|^\nu\Big] \;.$$

Therefore applying (42) gives

$$\frac{1}{4}(M_{\mathrm{cond};\nu})^\nu - \varepsilon_{K;\nu}^\nu \;\leq\; \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\mu_k\Big|^\nu\Big] \;\leq\; 4((M_{\mathrm{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu)$$

For the last bound, we start by considering the following quantity, which can be thought of as the truncated version of $M_{\mathrm{full};\nu}^\nu$:

$$\begin{aligned} m_K &:= \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \mu_k^2)\Big|^\nu\Big] \\ &= \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\phi_k(\mathbf{X}_2) + \sum_{k=1}^K \lambda_k\mu_k(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \\ &=: \mathbb{E}[|T_2 + T_1|^\nu] \;.\end{aligned}$$

Since $\{T_1, T_2\}$ forms a two-element martingale difference sequence with respect to $\sigma(\mathbf{X}_2) \subseteq \sigma(\mathbf{X}_1,\mathbf{X}_2)$, by Lemma 26, there exists absolute constants $c'_\nu, C'_\nu > 0$ depending only on $\nu$ such that

$$c'_\nu\big(\mathbb{E}[|T_1|^\nu] + \mathbb{E}[|T_2|^\nu]\big) \;\leq\; m_K \;\leq\; C'_\nu\big(\mathbb{E}[|T_1|^\nu] + \mathbb{E}[|T_2|^\nu]\big) \;.$$

Similarly, by writing

$$\begin{aligned}\mathbb{E}[|T_2|^\nu] &= \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\phi_k(\mathbf{X}_2)\Big|^\nu\Big] \\ &= \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k) + \sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\mu_k\Big|^\nu\Big] \\ &= \mathbb{E}[|R_2 + R_1|^\nu] \;,\end{aligned}$$

and noting that $\{R_1, R_2\}$ forms a two-element martingale difference sequence with respect to $\sigma(\mathbf{X}_1) \subseteq \sigma(\mathbf{X}_1,\mathbf{X}_2)$, by Lemma 26, there exists absolute constants $c''_\nu, C''_\nu > 0$ depending only on $\nu$ such that

$$c''_\nu\big(\mathbb{E}[|R_1|^\nu] + \mathbb{E}[|R_2|^\nu]\big) \;\leq\; \mathbb{E}[|T_2|^\nu] \;\leq\; C''_\nu\big(\mathbb{E}[|R_1|^\nu] + \mathbb{E}[|R_2|^\nu]\big) \;.$$

Combining the results and setting $A = \sup_{\nu\in[1,3]} C'_\nu \max\{C''_\nu, 1\}$ and $a = \inf_{\nu\in[1,3]} c'_\nu \min\{c''_\nu, 1\}$, we have shown that

$$a\big(\mathbb{E}[|T_1|^\nu] + \mathbb{E}[|R_1|^\nu] + \mathbb{E}[|R_2|^\nu]\big) \;\leq\; m_K \;\leq A\big(\mathbb{E}[|T_1|^\nu] + \mathbb{E}[|R_1|^\nu] + \mathbb{E}[|R_2|^\nu]\big) \;.$$

Notice that the quantity we would like to control is exactly

$$\mathbb{E}[|R_2|^\nu] \;= \mathbb{E}\Big[\Big| \textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \;,$$

and that $\mathbb{E}[|T_1|^\nu] = \mathbb{E}[|R_1|^\nu]$. By setting $c = A^{-1}$ and $C = a^{-1}$, this allows us to obtain a bound about $\mathbb{E}[|R_2|^\nu]$ as

$$cm_K - 2\mathbb{E}[|T_1|^\nu] \leq \mathbb{E}\Big[\Big| \textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \leq Cm_K - 2\mathbb{E}[|T_1|^\nu] \;.$$

Now notice that

$$\mathbb{E}[|T_1|^\nu] \;=\; \mathbb{E}\Big[\Big| \textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\mu_k\Big|^\nu\Big] \;,$$

which has already been controlled by the second result of the lemma as

$$\tfrac{1}{4}(M_{\text{cond};\nu})^\nu - \varepsilon_{K;\nu}^\nu \;\leq\; \mathbb{E}[|T_1|^\nu] \;\leq\; 4((M_{\text{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu) \;.$$

On the other hand, we can use an exactly analogous argument by using (42) and applying Jensen's inequality to control the errors to show that

$$\frac{1}{4}(M_{\text{full};\nu})^\nu - \varepsilon_{K;\nu}^\nu \;\leq\; m_K \;\leq\; 4((M_{\text{full};\nu})^\nu + \varepsilon_{K;\nu}^\nu) \;.$$

Applying these two results to the previous bound gives the desired bounds:

$$\mathbb{E}\Big[\Big| \textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \leq 4C(M_{\text{full};\nu})^\nu - \frac{1}{2}(M_{\text{cond};\nu})^\nu + (4C+2)\varepsilon_{K;\nu}^\nu \;,$$
$$\mathbb{E}\Big[\Big| \textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \geq \tfrac{c}{4}(M_{\text{full};\nu})^\nu - 8(M_{\text{cond};\nu})^\nu - (c+8)\varepsilon_{K;\nu}^\nu \;.$$

■

**Proof of Lemma 32** To compute the first bound, we rewrite the expression of interest as a quantity that we have already considered in the proof of Lemma 31:

$$\begin{aligned}
(\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) \;&= (\mu^K)^\top \Lambda^K \mathbb{E}\Big[\big(\phi^K(\mathbf{X}_1) - \mu^K\big)\big(\phi^K(\mathbf{X}_1) - \mu^K\big)^\top\Big]\Lambda^K(\mu^K) \\
&= \mathbb{E}\Big[\Big(\big(\phi^K(\mathbf{X}_1) - \mu^K\big)^\top \Lambda^K \mu^K\Big)^2\Big] \\
&= \mathbb{E}\Big[\Big(\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\mu_k\Big)^2\Big] \\
&= \mathbb{E}\Big[\Big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_2] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)] + \Delta_{K;1} - \Delta_{K;2}\Big)^2\Big] \;,
\end{aligned}$$

where we have used the calculation in (43) with $\nu = 2$ and defined the same error terms

$$\begin{aligned}
\Delta_{K;1} \;&\coloneqq\; \textstyle\sum_{k=1}^K \lambda_k \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)|\mathbf{X}_2] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_2] \;, \\
\Delta_{K;2} \;&\coloneqq\; \textstyle\sum_{k=1}^K \lambda_k \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)] \;.
\end{aligned}$$

Since we are dealing with the second moment, we can provide a finer bound by expanding the square explicitly:

$$(\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) \; = \mathrm{Var}\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_2] + \mathbb{E}[(\Delta_{K;1} - \Delta_{K;2})^2] \\ + 2\mathbb{E}\big[\big(\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_2] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)]\big)(\Delta_{K;1} - \Delta_{K;2})^2\big] \; .$$

Then by a Cauchy-Schwartz inequality, we get that

$$\Big|(\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) - \mathrm{Var}\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_2]\Big| \\ = 2\big|\mathbb{E}\big[\big(\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_2] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)]\big)(\Delta_{K;1} - \Delta_{K;2})^2\big]\big| + \mathbb{E}[(\Delta_{K;1} - \Delta_{K;2})^2] \\ \leq 2\sqrt{\mathrm{Var}\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_2]}\sqrt{\mathbb{E}[(\Delta_{K;1} - \Delta_{K;2})^2]} + \mathbb{E}[(\Delta_{K;1} - \Delta_{K;2})^2] \; .$$

The variance term is exactly $\sigma^2_{\text{cond}}$. Since the individual error terms have already been bounded in the proof of Lemma 31 as $\mathbb{E}[\Delta^2_{K;1}], \mathbb{E}[\Delta^2_{K;2}] \leq \varepsilon^2_{K;2}$, by a triangle inequality and a Cauchy-Schwarz inequality, we have

$$|\mathbb{E}[(\Delta_{K;1} - \Delta_{K;2})^2]| \; = |\mathbb{E}[\Delta^2_{K;1}] - 2\mathbb{E}[\Delta_{K;1}\Delta_{K;2}] + \mathbb{E}[\Delta^2_{K;2}]| \\ \leq |\mathbb{E}[\Delta^2_{K;1}]| + 2\sqrt{|\mathbb{E}[\Delta^2_{K;1}]||\mathbb{E}[\Delta^2_{K;2}]|} + |\mathbb{E}[\Delta^2_{K;2}]| \; \leq \; 4\varepsilon^2_{K;2} \; .$$

Combining the bounds gives

$$\big|(\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) - (\sigma_{\text{cond}})^2\big| \; \leq \; 4\varepsilon^2_{K;2} + 4\sigma_{\text{cond}}\varepsilon_{K;2} \; ,$$

which rearranges to give

$$\sigma^2_{\text{cond}} - 4\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon^2_{K;2} \; \leq \; (\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) \; \leq \sigma^2_{\text{cond}} + 4\sigma_{\text{cond}}\varepsilon_{K;2} + 4\varepsilon^2_{K;2} \\ \leq (\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2 \; .$$

The second bound is obtained similarly by giving a finer control than the bound in Lemma 31. We first rewrite the expression of interest by using linearity of expectation and the cyclic property of trace:

$$\mathrm{Tr}((\Lambda^K \Sigma^K)^2) \; = \mathrm{Tr}\big(\Lambda^K \mathbb{E}\big[\phi^K(\mathbf{X}_1)\phi^K(\mathbf{X}_1)^\top\big]\Lambda^K \mathbb{E}\big[\phi^K(\mathbf{X}_2)\phi^K(\mathbf{X}_2)^\top\big]\big) \\ = \mathbb{E}\Big[\big(\phi^K(\mathbf{X}_1)^\top \Lambda^K \phi^K(\mathbf{X}_2)\big)^2\Big] \; = \; \mathbb{E}\Big[\Big(\textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\Big)^2\Big] \; .$$

Again by expanding the square explicitly, we get that

$$\mathrm{Tr}((\Lambda^K \Sigma^K)^2) \; = \mathbb{E}\Big[\Big(\textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \bar{u}(\mathbf{X}_1, \mathbf{X}_2) + \bar{u}(\mathbf{X}_1, \mathbf{X}_2)\Big)^2\Big] \\ = \mathbb{E}\Big[\Big(\textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \bar{u}(\mathbf{X}_1, \mathbf{X}_2)\Big)^2\Big] + \mathbb{E}\big[\bar{u}(\mathbf{X}_1, \mathbf{X}_2)^2\big] + 2\Delta_{K;3} \\ = \varepsilon^2_{K;2} + \sigma^2_{\text{full}} + 2\Delta_{K;3} \; ,$$

where we have defined the additional error term as

$$\Delta_{K;3} \; \coloneqq \; \mathbb{E}\Big[\Big(\textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \bar{u}(\mathbf{X}_1, \mathbf{X}_2)\Big)\bar{u}(\mathbf{X}_1, \mathbf{X}_2)\Big] \; .$$

By a Cauchy-Schwarz inequality, we get that

$$\begin{aligned}\big|\mathrm{Tr}((\Lambda^K\Sigma^K)^2) - \sigma_{\text{full}}^2 - \varepsilon_{K;2}^2\big| \;&=\; 2|\Delta_{K;3}| \\ \leq 2\sqrt{\mathbb{E}\big[\big(\textstyle\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \bar{u}(\mathbf{X}_1,\mathbf{X}_2)\big)^2\big]}&\sqrt{\mathbb{E}\big[\bar{u}(\mathbf{X}_1,\mathbf{X}_2)^2\big]} \;=\; 2\varepsilon_{K;2}\sigma_{\text{full}}\;.\end{aligned}$$

Combining the above two bounds yields the desired inequality that

$$(\sigma_{\text{full}} - \varepsilon_{K;2})^2 \;\leq\; \mathrm{Tr}((\Lambda^K\Sigma^K)^2) \;\leq\; (\sigma_{\text{full}} + \varepsilon_{K;2})^2\;.$$

To prove the third bound, note that $(\mu^K)^\top\Lambda^K\mathbf{Z}_1$ is a zero-mean normal random variable with variance given by $(\mu^K)^\top\Lambda^K\Sigma^K\mu^K$ , which is already bounded above. By applying the formula of the $\nu$-th absolute moment of a normal distribution and noting that $\nu \leq 3$, we obtain

$$\begin{aligned}\mathbb{E}[|(\mu^K)^\top\Lambda^K\mathbf{Z}_1|^\nu] \;&=\; \frac{2^{\nu/2}}{\sqrt{\pi}}\Gamma\Big(\frac{\nu+1}{2}\Big)\big((\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K\mu^K\big)^{\nu/2} \\ &\leq \frac{2^{\nu/2}}{\sqrt{\pi}}(\sigma_{\text{full}} + 2\varepsilon_{K;2})^\nu \;\overset{(a)}{\leq}\; \frac{2^{\nu/2}}{\sqrt{\pi}}\max\{1, 2^{\nu-1}\}\big(\sigma_{\text{cond}}^\nu + 2^\nu\varepsilon_{K;2}^\nu\big) \;\overset{(b)}{\leq}\; 7(\sigma_{\text{cond}}^\nu + 8\varepsilon_{K;2}^\nu)\;.\end{aligned}$$

In $(a)$, we have noted that given $a, b > 0$, for $\nu/2 \in (0,1]$, $(a+b)^{\nu/2} \leq a^{\nu/2} + b^{\nu/2}$ and for $\nu/2 > 1$, the bound follows from Jensen's inequality. In $(b)$, we have noted that $\nu \leq 3$. This finishes the proof for the third bound.

To prove the fourth bound, we can first condition on $\mathbf{Z}_2$:

$$\mathbb{E}[|\mathbf{Z}_1^\top\Lambda^K\mathbf{Z}_2|^\nu] \;=\; \mathbb{E}\big[\,\mathbb{E}[|\mathbf{Z}_1^\top\Lambda^K\mathbf{Z}_2|^\nu|\,\mathbf{Z}_2]\,\big]\;.$$

The inner expectation is again the $\nu$-th absolute moment of a conditionally Gaussian random variable with variance $\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2$, so again by the formula of the $\nu$-th absolute moment of a normal distribution, we get that

$$\mathbb{E}[|\mathbf{Z}_1^\top\Lambda^K\mathbf{Z}_2|^\nu] \;\leq\; \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big(\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2\big)^{\nu/2}\Big] \;\leq\; \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big(\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2\big)^2\Big]^{\nu/4}\;.$$

We have noted that $\nu \leq 3$ and used a Hölder's inequality. The remaining expectation is taken over a quadratic form of normal variables. Writing $\Sigma_* = (\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}$ for short, the second moment can be computed by the formula from Lemma 27 as

$$\mathbb{E}\Big[\big(\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2\big)^2\Big] \;=\mathrm{Tr}(\Sigma_*^2)^2 + 2\mathrm{Tr}\big(\Sigma_*^4\big) \;\overset{(a)}{\leq}\; 3\mathrm{Tr}(\Sigma_*^2)^2 \;=\; 3\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big)^2\;.$$

Note that in $(a)$, we have used the fact that the square of a symmetric matrix, $\Sigma_*^2$, has non-negative eigenvalues, and therefore $\mathrm{Tr}(\Sigma_*^4) \leq \mathrm{Tr}(\Sigma_*^2)^2$. Since we have already bounded $\mathrm{Tr}((\Lambda^K\Sigma^K)^2)$ earlier, substituting the above result into the previous bound, we get that

$$\begin{aligned}\mathbb{E}[|\mathbf{Z}_1^\top\Lambda^K\mathbf{Z}_2|^\nu] \;&\leq \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big(\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2\big)^2\Big]^{\nu/4} \leq \frac{2^{\nu/2}3^{\nu/4}}{\sqrt{\pi}}\,\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big)^{\nu/2} \\ &\leq \frac{2^{\nu/2}3^{\nu/4}}{\sqrt{\pi}}(\sigma_{\text{full}}^2 + \varepsilon_{K;2}^2)^{\nu/2} \\ &\leq \frac{2^{\nu/2}3^{\nu/4}}{\sqrt{\pi}}\max\{1, 2^{\nu/2-1}\}\big(\sigma_{\text{full}}^\nu + \varepsilon_{K;2}^\nu\big) \;\leq\; 6\big(\sigma_{\text{full}}^\nu + \varepsilon_{K;2}^\nu\big)\;.\end{aligned}$$

In the last two inequalities, we have used the same argument as in the proof for the third bound to expand the term with $\nu$-th power. This gives the desired bound.

To prove the final bound, we first condition on $\mathbf{X}_1$:

$$\mathbb{E}\big[\big|(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\mathbf{Z}_1\big|^\nu\big] \;=\; \mathbb{E}\big[\mathbb{E}\big[\big|(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\mathbf{Z}_1\big|^\nu\big|\mathbf{X}_1\big]\big]\;.$$

The inner expectation is the $\nu$-th absolute moment of a conditionally Gaussian random variable with variance $(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K(\phi^K(\mathbf{X}_1)-\mu^K)$, so by the formula of the $\nu$-th absolute moment of a normal distribution with $\nu \leq 3$, we get that

$$\begin{aligned}
&\mathbb{E}[|(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\mathbf{Z}_2|^\nu]\\
&\leq \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big((\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K(\phi^K(\mathbf{X}_1)-\mu^K)\big)^{\nu/2}\Big]\\
&= \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big((\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\mathbb{E}\big[(\phi^K(\mathbf{X}_2)-\mu^K)(\phi^K(\mathbf{X}_2)-\mu^K)^\top\big]\Lambda^K(\phi^K(\mathbf{X}_1)-\mu^K)\big)^{\nu/2}\Big]\\
&\overset{(a)}{\leq} \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big|(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K(\phi^K(\mathbf{X}_2)-\mu^K)\big|^\nu\Big]\\
&= \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big|\textstyle\sum_{k=1}^K\lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|^\nu\Big]\\
&\overset{(b)}{\leq} 8C(M_{\text{full};\nu})^\nu-(M_{\text{cond};\nu})^\nu+(8C+4)\varepsilon_{K;\nu}^\nu\;.
\end{aligned}$$

In $(a)$, we have applied Jensen's inequality to the convex function $x\mapsto|x|^{\nu/2}$ to move the inner expectation outside the norm. In $(b)$, we have applied the bound in Lemma 31 and noted that $\frac{2^{\nu/2}}{\sqrt{\pi}}<2$ for $\nu\in[1,3]$. This gives the desired result. ■

**Proof of Lemma 33** For the first equality in distribution, we recall that $\{\tau_{k;d}\}_{k=1}^K$ are the eigenvalues of $(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}$ and $\{\xi_k\}_{k=1}^K$ are a sequence of i.i.d. standard Gaussian variables. Let $\{\eta_{ik}\}_{i\in[n],k\in[K]}$ be a set of i.i.d. standard Gaussian variables. Since Gaussianity is preserved under orthogonal transformation, we have

$$\begin{aligned}
&\frac{1}{n^{3/2}(n-1)^{1/2}}\Big(\textstyle\sum_{i,j=1}^n(\eta_i^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K-n\mathrm{Tr}(\Sigma^K\Lambda^K)\Big)\\
&\overset{d}{=}\frac{1}{n^{3/2}(n-1)^{1/2}}\Big(\textstyle\sum_{k=1}^K\sum_{i,j=1}^n\tau_{k;d}\eta_{ik}\eta_{jk}-n\mathrm{Tr}\big((\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\big)\Big)\\
&=\frac{1}{n^{3/2}(n-1)^{1/2}}\textstyle\sum_{k=1}^K\tau_{k;d}\Big(\big(\sum_{i=1}^n\eta_{ik}\big)\big(\sum_{j=1}^n\eta_{jk}\big)-n\Big)\\
&\overset{d}{=}\frac{1}{n^{1/2}(n-1)^{1/2}}\textstyle\sum_{k=1}^K\tau_{k;d}(\xi_k^2-1)\;=\;W_n^K-D\;,
\end{aligned}$$

which proves the desired statement.

We now use the expression above for moment computation. The expectation is given by $\mathbb{E}[W_n^K]=D$ for every $K\in\mathbb{N}$. The variance can be computed by noting that the quantity is a quadratic form in Gaussian, applying Lemma 27 and using the cyclic property of trace:

$$\begin{aligned}
\mathrm{Var}[W_n^K]\;&=\frac{1}{n(n-1)}\mathrm{Var}\big[(\eta_1^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_1^K\big]\\
&=\frac{2}{n(n-1)}\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big)\;.
\end{aligned}$$

By Lemma 32, we get the desired bound that

$$\frac{2}{n(n-1)}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 \;\leq\; \text{Var}[W_n^K] \;\leq\; \frac{2}{n(n-1)}(\sigma_{\text{full}} + \varepsilon_{K;2})^2 \;.$$

The third central moment can be expanded using a binomial expansion and noting that each summand is zero-mean:

$$\begin{aligned}
\mathbb{E}\big[(W_n^K - D)^3\big] &= \frac{1}{n^{3/2}(n-1)^{3/2}}\mathbb{E}\Big[\Big(\sum_{k=1}^K \tau_{k;d}(\xi_k^2 - 1)\Big)^3\Big] \\
&= \frac{1}{n^{3/2}(n-1)^{3/2}}\mathbb{E}\big[\sum_{k=1}^K \tau_{k;d}^3(\xi_k^2 - 1)^3\big] \\
&= \frac{8}{n^{3/2}(n-1)^{3/2}}\sum_{k=1}^K \tau_{k;d}^3 \;.
\end{aligned}$$

Meanwhile, the sum can be further expressed as

$$\begin{aligned}
&\sum_{k=1}^K \tau_{k;d}^3 \\
&= \text{Tr}\Big(\big((\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\big)^3\Big) \;=\; \text{Tr}\Big(\big(\Sigma^K\Lambda^K\big)^3\Big) \\
&= \text{Tr}\Big(\big(\mathbb{E}\big[\phi^K(\mathbf{X}_1)(\phi^K(\mathbf{X}_1))^\top\big]\Lambda^K\big)^3\Big) \\
&= \mathbb{E}\Big[(\phi^K(\mathbf{X}_1))^\top\Lambda^K\phi^K(\mathbf{X}_2)\,(\phi^K(\mathbf{X}_2))^\top\Lambda^K\phi^K(\mathbf{X}_3)\,(\phi^K(\mathbf{X}_3))^\top\Lambda^K\phi^K(\mathbf{X}_1)\Big] \\
&= \mathbb{E}\Big[\big(\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big)\big(\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_2)\phi_k(\mathbf{X}_3)\big)\big(\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_3)\phi_k(\mathbf{X}_1)\big)\Big] \\
&=: \mathbb{E}[S_{12}S_{23}S_{31}] \;.
\end{aligned}$$

We now approximate each $S_{ij}$ term by $u(\mathbf{X}_i, \mathbf{X}_j)$. For convenience, denote $U_{ij} = u(\mathbf{X}_i, \mathbf{X}_j)$ and $\Delta_{ij} = S_{ij} - U_{ij}$. Then

$$\begin{aligned}
\sum_{k=1}^K \tau_{k;d}^3 &= \mathbb{E}\big[(U_{12} + \Delta_{12})(U_{23} + \Delta_{23})(U_{31} + \Delta_{31})\big] \\
&= \mathbb{E}[U_{12}U_{23}U_{31}] + \mathbb{E}[U_{12}U_{23}\Delta_{31}] + \mathbb{E}[U_{12}\Delta_{23}U_{31}] + \mathbb{E}[U_{12}\Delta_{23}\Delta_{31}] \\
&\quad + \mathbb{E}[\Delta_{12}U_{23}U_{31}] + \mathbb{E}[\Delta_{12}U_{23}\Delta_{31}] + \mathbb{E}[\Delta_{12}\Delta_{23}U_{31}] + \mathbb{E}[\Delta_{12}\Delta_{23}\Delta_{31}] \;.
\end{aligned}$$

Recall that $\varepsilon_{K;3} = \mathbb{E}[|\Delta_{ij}|^3]^{1/3}$ for $i \neq j$ by definition. Then by a triangle inequality followed by a Hölder's inequality, we get that

$$\begin{aligned}
&\Big|\sum_{k=1}^K \tau_{k;d}^3 - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)u(\mathbf{X}_2, \mathbf{X}_3)u(\mathbf{X}_3, \mathbf{X}_1)]\Big| \\
&\leq \big|\mathbb{E}[U_{12}U_{23}\Delta_{31}]\big| + \big|\mathbb{E}[U_{12}\Delta_{23}U_{31}]\big| + \big|\mathbb{E}[U_{12}\Delta_{23}\Delta_{31}]\big| \\
&\quad + \big|\mathbb{E}[\Delta_{12}U_{23}U_{31}]\big| + \big|\mathbb{E}[\Delta_{12}U_{23}\Delta_{31}]\big| + \big|\mathbb{E}[\Delta_{12}\Delta_{23}U_{31}]\big| + \big|\mathbb{E}[\Delta_{12}\Delta_{23}\Delta_{31}]\big| \\
&\leq 3\mathbb{E}[|u(\mathbf{X}_1, \mathbf{X}_2)|^3]^{2/3}\varepsilon_{K;3} + 3\mathbb{E}[|u(\mathbf{X}_1, \mathbf{X}_2)|^3]^{1/3}\varepsilon_{K;3}^2 + \varepsilon_{K;3}^3 \\
&= 3M_{\text{full};3}^2\varepsilon_{K;3} + 3M_{\text{full};3}\varepsilon_{K;3}^2 + \varepsilon_{K;3}^3 \;.
\end{aligned}$$

This implies that

$$\begin{aligned}
\sum_{k=1}^K \tau_{k;d}^3 &\leq \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)u(\mathbf{X}_2, \mathbf{X}_3)u(\mathbf{X}_3, \mathbf{X}_1)] - M_{\text{full};3}^3 + (M_{\text{full};3} + \varepsilon_{K;3})^3 \;, \\
\sum_{k=1}^K \tau_{k;d}^3 &\geq \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)u(\mathbf{X}_2, \mathbf{X}_3)u(\mathbf{X}_3, \mathbf{X}_1)] + M_{\text{full};3}^3 - (M_{\text{full};3} + \varepsilon_{K;3})^3 \;,
\end{aligned}$$

which gives the desired bounds:

$$\mathbb{E}\big[(W_n^K - D)^3\big] \;\leq\; \frac{8\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] - M_{\text{full};3}^3 + (M_{\text{full};3} + \varepsilon_{K;3})^3\big)}{n^{3/2}(n-1)^{3/2}} \;,$$
$$\mathbb{E}\big[(W_n^K - D)^3\big] \;\geq\; \frac{8\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] + M_{\text{full};3}^3 - (M_{\text{full};3} + \varepsilon_{K;3})^3\big)}{n^{3/2}(n-1)^{3/2}} \;.$$

The fourth central moment can again be expanded using a binomial expansion and noting that each summand is zero-mean:

$$\begin{aligned}
&\mathbb{E}\big[(W_n^K - D)^4\big] \\
&= \tfrac{1}{n^2(n-1)^2}\mathbb{E}\Big[\Big(\textstyle\sum_{k=1}^K \tau_{k;d}(\xi_k^2 - 1)\Big)^4\Big] \\
&= \tfrac{1}{n^2(n-1)^2}\Big(\mathbb{E}\Big[\textstyle\sum_{k=1}^K \tau_{k;d}^4(\xi_k^2-1)^4\Big] + 3\mathbb{E}\Big[\textstyle\sum_{1\leq k\neq k'\leq K} \tau_{k;d}^2(\xi_k^2-1)^2\tau_{k';d}^2(\xi_{k'}^2-1)^2\Big]\Big) \\
&= \tfrac{1}{n^2(n-1)^2}\Big(60\textstyle\sum_{k=1}^K \tau_{k;d}^4 + 12\textstyle\sum_{1\leq k\neq k'\leq K} \tau_{k;d}^2\tau_{k';d}^2\Big) \\
&= \tfrac{1}{n^2(n-1)^2}\Big(48\textstyle\sum_{k=1}^K \tau_{k;d}^4 + 12\textstyle\sum_{1\leq k, k'\leq K} \tau_{k;d}^2\tau_{k';d}^2\Big) \\
&= \tfrac{12}{n^2(n-1)^2}\Big(4\textstyle\sum_{k=1}^K \tau_{k;d}^4 + \big(\textstyle\sum_{k=1}^K \tau_{k;d}^2\big)^2\Big) \;.
\end{aligned}$$

Since we have already controlled $\sum_{k=1}^K \tau_{k;d}^2 = \mathrm{Tr}\big((\Sigma^K\Lambda^K)^2\big)$, we focus on bounding the first sum. Using notations from the third moment, we can express the sum as

$$\begin{aligned}
\textstyle\sum_{k=1}^K \tau_{k;d}^4 \;&= \mathbb{E}\Big[\big(\textstyle\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big)\big(\textstyle\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_2)\phi_k(\mathbf{X}_3)\big) \\
&\qquad\big(\textstyle\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_3)\phi_k(\mathbf{X}_4)\big)\big(\textstyle\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_4)\phi_k(\mathbf{X}_1)\big)\Big] \\
&= \mathbb{E}[S_{12}S_{23}S_{34}S_{41}] \\
&= \mathbb{E}\big[(U_{12}+\Delta_{12})(U_{23}+\Delta_{23})(U_{34}+\Delta_{34})(U_{41}+\Delta_{41})\big] \;.
\end{aligned}$$

A similar argument as before shows that

$$\begin{aligned}
&\Big|\textstyle\sum_{k=1}^K \tau_{k;d}^4 - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)]\Big| \\
&\qquad\qquad \leq 4M_{\text{full};4}^3\varepsilon_{K;4} + 6M_{\text{full};4}^2\varepsilon_{K;4}^2 + 4M_{\text{full};4}\varepsilon_{K;4}^3 + \varepsilon_{K;4}^4 \;.
\end{aligned}$$

This implies that

$$\textstyle\sum_{k=1}^K \tau_{k;d}^4 \;\leq \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] - M_{\text{full};4}^4 + (M_{\text{full};4}+\varepsilon_{K;4})^4 \;,$$
$$\textstyle\sum_{k=1}^K \tau_{k;d}^4 \;\geq \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] + M_{\text{full};4}^4 - (M_{\text{full};4}+\varepsilon_{K;4})^4 \;.$$

On the other hand, by Lemma 32, we have

$$(\sigma_{\text{full}} - \varepsilon_{K;2})^2 \;\leq\; \textstyle\sum_{k=1}^K \tau_{k;d}^2 \;=\; \mathrm{Tr}((\Lambda^K\Sigma^K)^2) \;\leq\; (\sigma_{\text{full}} + \varepsilon_{K;2})^2 \;.$$

Combining the results give the desired bounds:

$$\begin{aligned}\mathbb{E}\big[(W_n^K - D)^4\big] \;&\leq \tfrac{12}{n^2(n-1)^2}\Big(4\,\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] \\ &\qquad - 4M_{\text{full};4}^4 + 4(M_{\text{full};4}+\varepsilon_{K;4})^4 + (\sigma_{\text{full}}+\varepsilon_{K;2})^4\Big)\;, \\ \mathbb{E}\big[(W_n^K - D)^4\big] \;&\geq \tfrac{12}{n^2(n-1)^2}\Big(4\,\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] \\ &\qquad + 4M_{\text{full};4}^4 - 4(M_{\text{full};4}+\varepsilon_{K;4})^4 + (\sigma_{\text{full}}-\varepsilon_{K;2})^4\Big)\;.\end{aligned}$$

For the generic moment bound, we first use a Jensen's inequality to get that

$$\begin{aligned}\mathbb{E}\big[(W_n^K)^{2m}\big] \;&= \mathbb{E}\Big[\Big(\tfrac{1}{n^{1/2}(n-1)^{1/2}}\textstyle\sum_{k=1}^K \tau_{k;d}(\xi_k^2-1)+D\Big)^{2m}\Big] \\ &\leq \tfrac{2^{2m-1}}{n^m(n-1)^m}\mathbb{E}\Big[\Big(\textstyle\sum_{k=1}^K \tau_{k;d}(\xi_k^2-1)\Big)^{2m}\Big] + 2^{2m-1}\,D^{2m}\;.\end{aligned}$$

Denote the set of all possible orderings of a length-$2m$ sequence consisting of elements from $[K]$ by $\mathcal{P}(K,2m)$ and denote its elements by $p$. Consider the subset

$$\mathcal{P}'(K,2m) \;\coloneqq\; \{p \in \mathcal{P}(K,2m)\;:\; \text{every element in } p \text{ appears at least twice}\,\}\;.$$

By noting that $\xi_k - 1$ is zero-mean and $\{\xi_k\}_{k=1}^K$ are independent, we can re-express the sum first as a sum over $\mathcal{P}(K,2m)$ and then as a sum over $\mathcal{P}'(K,2m)$:

$$\begin{aligned}\mathbb{E}\Big[\Big(\textstyle\sum_{k=1}^K \tau_{k;d}(\xi_k^2-1)\Big)^{2m}\Big] \;&= \textstyle\sum_{p\in\mathcal{P}(K,2m)}\big(\prod_{k\in p}\tau_{k;d}\big)\mathbb{E}\big[\prod_{k\in p}(\xi_k^2-1)\big] \\ &= \textstyle\sum_{p\in\mathcal{P}'(K,2m)}\big(\prod_{k\in p}\tau_{k;d}\big)\mathbb{E}\big[\prod_{k\in p}(\xi_k^2-1)\big] \\ &\quad + \textstyle\sum_{p\in\big(\mathcal{P}(K,2m)\setminus\mathcal{P}'(K,2m)\big)}\big(\prod_{k\in p}\tau_{k;d}\big)\mathbb{E}\big[\prod_{k\in p}(\xi_k^2-1)\big] \\ &= \textstyle\sum_{p\in\mathcal{P}'(K,2m)}\big(\prod_{k\in p}\tau_{k;d}\big)\mathbb{E}\big[\prod_{k\in p}(\xi_k^2-1)\big]\;.\end{aligned}$$

Write $C'_m$ as the $2m$-th central moment of a chi-squared random variable with degree $1$, which depends only on $m$ and not on $K$ or $\tau_{k;d}$. By a Hölder's inequality and the bound from Lemma 32, we get that

$$\begin{aligned}\mathbb{E}\Big[\Big(\textstyle\sum_{k=1}^K \tau_{k;d}(\xi_k^2-1)\Big)^{2m}\Big] \;&\leq C'_m \textstyle\sum_{p\in\mathcal{P}'(K,2m)}\big(\prod_{k\in p}\tau_{k;d}\big) \\ &\leq C'_m\binom{2m}{m}\big(\textstyle\sum_{k=1}^K \tau_{k;d}^2\big)^m \\ &= C'_m\binom{2m}{m}\,\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big)^m \;\leq\; C'_m\binom{2m}{m}(\sigma_{\text{full}}+\varepsilon_{K;2})^{2m}\;.\end{aligned}$$

Writing $C_m \coloneqq 2^{2m-1}\max\{1, C'_m\binom{2m}{m}\}$, we get the desired bound that

$$\mathbb{E}\big[(W_n^K)^{2m}\big] \;\leq \tfrac{C_m}{n^m(n-1)^m}(\sigma_{\text{full}}+\varepsilon_{K;2})^{2m} + C_m\,D^{2m}\;.$$

Finally, if Assumption 2 is true for some $\nu \geq 2$, we have $\varepsilon_{K;2} \to 0$ as $K$ grows. Taking $K \to \infty$ in the bound for second moment gives

$$\lim_{K\to\infty} \mathrm{Var}[W_n^K] \;=\; \frac{2}{n(n-1)}\sigma^2_{\text{full}} \;.$$

If Assumption 2 holds for $\nu \geq 3$, similarly we have

$$\lim_{K\to\infty} \mathbb{E}\big[(W_n^K - D)^3\big] \;=\; \frac{8\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)]}{n^{3/2}(n-1)^{3/2}} \;.$$

If Assumption 2 holds for $\nu \geq 4$, we have

$$\lim_{K\to\infty} \mathbb{E}\big[(W_n^K - D)^4\big] \;=\; \frac{12(4\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] + \sigma^4_{\text{full}})}{n^2(n-1)^2} \;.$$

■

### G.3. Proofs for Appendix B.3

**Proof of Lemma 34** Write $\delta' := \delta/(m+1)$ for convenience. Define the $m$-times differentiable function

$$h_{m;\tau;\delta}(x) \;:=\; (\delta')^{-(m+1)} \int_x^{x+\delta'} \int_{y_1}^{y_1+\delta'} \ldots \int_{y_{m-1}}^{y_{m-1}+\delta'} \int_{y_m}^{y_m+\delta'} \mathbb{I}_{\{y>\tau\}}\, dy\, dy_m \ldots dy_1 \;.$$

In the case $m=0$, the function is $h_{0;\tau;\delta}(x) \;:=\; \delta^{-1}\int_x^{x+\delta} \mathbb{I}_{\{y>\tau\}}\, dy$. By construction, $h_{m;\tau;\delta}(x) = 0$ for $x \leq \tau-\delta$, $h_{m;\tau;\delta}(x) \in [0,1]$ for $x \in (\tau-\delta,\tau]$ and $h_{m;\tau;\delta}(x) = 1$ for $x > \tau$. This implies $\mathbb{I}_{\{x>\tau\}} \leq h_{m;\tau;\delta}(x) \leq \mathbb{I}_{\{x>\tau-\delta\}}$ and therefore the desired inequality

$$h_{m;\tau+\delta;\delta}(x) \;\leq\; \mathbb{I}_{\{x>\tau\}} \;\leq\; h_{m;\tau;\delta}(x) \;.$$

Next, we prove the properties of the derivatives of $h_{m;\tau;\delta}$. Denote recursively

$$J_{m+1}(x) \;:=\; \int_x^{x+\delta'} \mathbb{I}_{\{y>\tau\}} dy \;, \qquad J_r(x) \;:=\; \int_x^{x+\delta'} J_{r+1}(y)\, dy \quad \text{for } 0 \leq r \leq m \;.$$

Since $h_{m;\tau;\delta}(x) = (\delta')^{-(m+1)} J_0(x)$ and $\frac{\partial}{\partial x} J_i(x) = J_{i+1}(x+\delta') - J_{i+1}(x)$ for $0 \leq i \leq m$, by induction, we have that for $0 \leq r \leq m$,

$$h^{(r)}_{m;\tau;\delta}(x) \;= (\delta')^{-(m+1)} \frac{\partial^r}{\partial x^r} J_0(x) \;=\; (\delta')^{-(m+1)} \sum\nolimits_{i=0}^r \binom{r}{i} (-1)^i J_{r+1}\big(x + (r-i)\delta'\big) \;. \quad (44)$$

Note that $J_{m+1}$ is continuous, uniformly bounded above by $\delta'$, and satisfies that $J_{m+1}(x) = 0$ for $x$ outside $[\tau-\delta',\tau]$. By induction, we get that for $0 \leq r \leq m$, $J_{r+1}$ is continuous, bounded above by $(\delta')^{m+1-r}$ and satisfies that $J_{r+1}(x) = 0$ for $x$ outside $[\tau-(m+1-r)\delta',\tau]$. This shows that $h^{(r)}_{m;\tau;\delta}$ is continuous and $h^{(r)}_{m;\tau;\delta}(x) = 0$ for $x$ outside $[\tau-\delta,\tau]$, and the uniform bound

$$\big|h^{(r)}_{m;\tau;\delta}(x)\big| \;\leq\; (\delta')^{-r} \sum\nolimits_{i=0}^r \binom{r}{i} \;=\; \big(\tfrac{2}{m+1}\big)^r \delta^{-r} \;\leq\; \delta^{-r} \;.$$

Finally to prove the Hölder property of $h^{(m)}_{m;\tau;\delta}(x)$, we first note that $J_{m+1}$ is constant outside $x \in [\tau-\delta', \tau]$ and linear within the interval with Lipschitz constant 1. The formula in (44) suggests that $h^{(m)}_{m;\tau;\delta}(x)$ is piecewise linear and the Lipschitz constant in the interval $[\tau-(m-i+1)\delta', \tau-(m-i)\delta']$ is given by the Lipschitz constant of the $i$-th summand. Therefore, $h^{(m)}_{m;\tau;\delta}$ is also Lipschitz with Lipschitz constant

$$L_m \;\coloneqq\; (\delta')^{-(m+1)} \max_{0 \leq i \leq m} \binom{m}{i} \;=\; (\delta')^{-(m+1)} \binom{m}{\lfloor m/2 \rfloor} .$$

For $x, y \in [\tau-\delta, \tau]$, we then have

$$\begin{aligned} |h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \;\leq\; L_m|x-y| \;&= L_m\delta \left|\frac{x-y}{\delta}\right| \\ &\leq L_m\delta \left|\frac{x-y}{\delta}\right|^\epsilon \;=\; L_m\delta^{1-\epsilon}|x-y|^\epsilon , \end{aligned} \tag{45}$$

where we have noted that $\left|\frac{x-y}{\delta}\right| \leq 1$ and $\epsilon \in [0,1]$. (45) is trivially true for $x, y$ both outside $[\tau-\delta, \tau]$ since $h^{(m)}_{m;\tau;\delta}$ evaluates to zero. Now consider $x \in [\tau-\delta, \tau]$ and $y < \tau - \delta$. We have that

$$\begin{aligned} |h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \;=\; |h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(\tau-\delta)| \;&\overset{(45)}{\leq}\; L_m\delta^{1-\epsilon}(x-\tau+\delta)^\epsilon \\ &\leq L_m\delta^{1-\epsilon}|x-y|^\epsilon . \end{aligned}$$

Similarly for $x \in [\tau-\delta, \tau]$ and $y > \tau$, we have that

$$|h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \;=\; |h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(\tau)| \;\overset{(45)}{\leq}\; L_m\delta^{1-\epsilon}(\tau-x)^\epsilon \;\leq\; L_m\delta^{1-\epsilon}|x-y|^\epsilon .$$

Therefore (45) holds for all $x, y$. The proof for the derivative bound is complete by computing the constant explicitly as

$$L_m\delta^{1-\epsilon} \;=\; (\delta')^{-(m+\epsilon)} \binom{m}{\lfloor m/2 \rfloor} \;=\; \delta^{-(m+\epsilon)} \binom{m}{\lfloor m/2 \rfloor} (m+1)^{m+\epsilon} ,$$

and therefore

$$|h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \;\leq\; C_{m,\epsilon}\, \delta^{-(m+\epsilon)}\, |x-y|^\epsilon , \tag{46}$$

with respect to the constant $C_{m,\epsilon} = \binom{m}{\lfloor m/2 \rfloor}(m+1)^{m+\epsilon}$. ■

**Proof of Lemma 35** By conditioning on the size of $Y$, we have that for any $a, b \in \mathbb{R}$ and $\epsilon > 0$,

$$\begin{aligned} \mathbb{P}(a \leq X+Y \leq b) \;&= \mathbb{P}(a \leq X+Y \leq b\,,\, |Y| \leq \epsilon) + \mathbb{P}(a \leq X \leq b\,,\, |Y| \geq \epsilon) \\ &\leq \mathbb{P}(a-\epsilon \leq X \leq b+\epsilon) + \mathbb{P}(|Y| \geq \epsilon) , \end{aligned}$$

and by using the order of inclusion of events, we have the lower bound

$$\begin{aligned} \mathbb{P}(a \leq X+Y \leq b) \;&\geq \mathbb{P}(a+\epsilon \leq X \leq b-\epsilon\,,\, |Y| \leq \epsilon) \\ &= \mathbb{P}(a+\epsilon \leq X \leq b-\epsilon) - \mathbb{P}(|Y| \geq \epsilon) . \end{aligned}$$

■

### G.4. Proof for Appendix B.4

**Proof of Lemma 37** By Lemma 2.3 of Steinwart and Scovel (2012), the assumption that $\kappa^*$ is measurable and $\mathbb{E}[\kappa^*(\mathbf{V}_1, \mathbf{V}_1)] < \infty$ implies the RKHS $\mathcal{H}$ associated with $\kappa^*$ is compactly embedded into $L_2(\mathbb{R}^d, R)$. By Lemma 2.12 and Corollary 3.2 of Steinwart and Scovel (2012), for some index set $\mathcal{I} \subseteq \mathbb{N}$, there exists a sequence of non-negative, bounded values $\{\lambda_k\}_{k \in \mathcal{I}}$ that converges to 0 and a sequence of functions $\{\phi_k\}_{k \in \mathcal{I}}$ that form an orthonormal basis of $L_2(\mathbb{R}^d, R)$ such that

$$\sum\nolimits_{k \in \mathcal{I}} \lambda_k \psi_k(\mathbf{V}_1) \psi_k(\mathbf{V}_2) \;=\; \kappa^*(\mathbf{V}_1, \mathbf{V}_2) \;,$$

where the equality holds almost surely when $\mathcal{I}$ is finite and the convergence holds almost surely when $\mathcal{I}$ is infinite. We can extend $\mathcal{I}$ to $\mathbb{N}$ by adding zero values of $\lambda_k$ and $\phi_k$ whenever necessary and drop the requirement that $\{\phi_k\}_{k=1}^\infty$ forms a basis, which gives the desired statement. ■