Bernoulli **30**(4), 2024, 2721–2746
https://doi.org/10.3150/23-BEJ1692

Shrinkage estimation of higher-order Bochner integrals

SAITEJA UTPALA[1,a] and BHARATH K. SRIPERUMBUDUR[2,b]

[1]*Wadhwani AI, New Delhi, India,* [a]*saitejautpala@gmail.com*
[2]*Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA,* [b]*bks18@psu.edu*

We consider shrinkage estimation of higher-order Hilbert space-valued Bochner integrals in a non-parametric setting. We propose estimators that shrink the U-statistic estimator of the Bochner integral towards a pre-specified target element in the Hilbert space. Depending on the degeneracy of the kernel of the U-statistic, we construct consistent shrinkage estimators and develop oracle inequalities comparing the risks of the U-statistic estimator and its shrinkage version. Surprisingly, we show that the shrinkage estimator designed by assuming complete degeneracy of the kernel of the U-statistic is a consistent estimator even when the kernel is not completely degenerate. This work subsumes and improves upon Muandet et al. (*J. Mach. Learn. Res.* **17** (2016) 48) and Zhou, Chen and Huang (*J. Multivariate Anal.* **169** (2019) 166–178), which only handle mean element and covariance operator estimation in a reproducing kernel Hilbert space. We also specialize our results to normal mean estimation and show that for $d \geq 3$, the proposed estimator strictly improves upon the sample mean in terms of the mean squared error.

Keywords: Bernstein's inequality; Bochner integral; completely degenerate; James-Stein estimator; shrinkage estimation; SURE; U-statistics

1. Introduction

Let $\mathcal{X}$ be a separable topological space and $\mathcal{H}$ be a separable Hilbert space. For a Bochner measurable function—for example, continuous functions are Bochner measurable—$r : \mathcal{X}^k \to \mathcal{H}$, where $k \in \mathbb{N}$, define the Bochner integral (Dinculeanu, 2000) with respect to the k-fold product measure $\mathbb{P}^k := \mathbb{P} \times \overset{k}{\cdots} \times \mathbb{P}$ as

$$C = \int_{\mathcal{X}^k} r(x_1, \ldots, x_k)\, d\mathbb{P}^k(x_1, \ldots, x_k) = \int_{\mathcal{X}^k} r(x_1, \ldots, x_k) \prod_{i=1}^{k} d\mathbb{P}(x_i).$$

Given $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$, the goal of this paper is to construct and analyze shrinkage estimators of C, of the form

$$\check{C} := (1 - \hat{\alpha})\hat{C} + \hat{\alpha} f^* = (1 - \hat{\alpha})(\hat{C} - f^*) + f^*, \tag{1}$$

where $0 < \hat{\alpha} < 1$ is a random variable that depends on $(X_i)_{i=1}^n$, f^* is a fixed target in $\mathcal{H}$ towards which $\hat{C}$ is shrunk to, and $\hat{C}$ is the U-statistic estimator of C given by

$$\hat{C} = \frac{1}{{}^n\mathrm{C}_k} \sum_{J_k^n} r(X_{i_1}, \ldots, X_{i_k}),$$

with $J_k^n = \{(i_1, \ldots, i_k) : 1 < i_1 < i_2 < \cdots < i_k < n\}$. Without loss of generality, we assume that r is *symmetric* (see Section 2 for the definition).

Traditionally, shrinkage estimators of the form in (1) are studied for $k = 1$ and $r(x) = x$, $x \in \mathbb{R}^d$, which in fact corresponds to shrinking the empirical mean, $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$, towards a fixed vector

$f^* \in \mathbb{R}^d$. For $\mathbb{P} = N(\mu, \sigma^2 I)$ where σ^2 is known, James and Stein (1960), Stein (1956) constructed a shrinkage estimator, $\check{\mu}$ of μ of the form in (1), given by

$$\check{\mu} = \left(1 - \frac{(d-2)\sigma^2}{n\|\bar{X} - f^*\|_2^2}\right)\bar{X} + \frac{(d-2)\sigma^2}{n\|\bar{X} - f^*\|_2^2} f^*,$$

and showed that for $d \geq 3$, the shrinkage estimator, $\check{\mu}$ improves upon $\bar{X}$ in terms of the mean-squared error, i.e.,

$$\mathbb{E}\|\check{\mu} - \mu\|_2^2 < \mathbb{E}\|\bar{X} - \mu\|_2^2, \forall\, \mu \in \mathbb{R}^d. \tag{2}$$

When σ^2 is unknown, it can be replaced by its estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \|X_i - \bar{X}\|_2^2$ in $\check{\mu}$ while still maintaining (2) for $d \geq 3$. Similar types of results have been established for location families of spherically symmetric distributions (see Brandwein and Strawderman, 1990, 2012 and references therein).

For $k = 2$ and $r(x_1, x_2) = \frac{1}{2}(x_1 - x_2)(x_1 - x_2)^\top$, $x_1, x_2 \in \mathbb{R}^d$, (1) reduces to the covariance matrix associated with $\mathbb{P}$. Starting with Stein (1975), a lot of work has been carried out on the shrinkage estimation of covariance matrices under the parametric setting of samples being observed from a multivariate normal distribution. Under different losses (e.g., Frobenius loss, Stein loss) and under different settings of $d \leq n$, $d > n$, d growing to infinity with n, the shrinkage estimator has been shown to strictly improve upon the sample covariance matrix (e.g., see Chen et al., 2010, Fisher and Sun, 2011, Ledoit and Wolf, 2018 and references therein). In the non-parametric setting where no specific parametric assumption is made on $\mathbb{P}$, consistent shrinkage estimators of the sample covariance matrix have been developed in the high-dimensional setting (Ledoit and Wolf, 2004, Touloumis, 2015).

While most of the above-mentioned works deal with parametric families of distributions, recently, Muandet et al. (2016) proposed shrinkage estimators for C in the non-parametric setting without making parametric assumptions on $\mathbb{P}$, with $k = 1$ and $r(x) = K(\cdot, x)$, where K is the *reproducing kernel* (i.e., a positive definite kernel) of a *reproducing kernel Hilbert space* (RKHS)—see Section 2 for the definition. This corresponds to the shrinkage estimation of the mean element, which is an infinite-dimensional object if the RKHS is infinite-dimensional. This is in sharp contrast to the above-mentioned works where the parameter is finite-dimensional or its dimension grows with the sample size. Extending this idea, Zhou, Chen and Huang (2019) proposed shrinkage estimators for C when $k = 2$ and $r(x_1, x_2) = \frac{1}{2}(K(\cdot, x_1) - K(\cdot, X_2)) \otimes_{\mathcal{H}} (K(\cdot, x_1) - K(\cdot, x_2))$, which corresponds to the covariance operator on an RKHS with reproducing kernel, K. The mean element and covariance operator has been widely used in nonparametric goodness-of-fit testing (Balasubramanian, Li and Yuan, 2021), two-sample testing (Gretton et al., 2012), independence testing (Gretton et al., 2007), supervised dimensionality reduction (Fukumizu, Bach and Jordan, 2004), feature selection (Song et al., 2012), etc., and therefore their shrunk versions are also useful in these applications. Of course, the choice of $K(\cdot, x) = x$, $x \in \mathbb{R}^d$, results in the mean and covariance matrix of $\mathbb{P}$ with $\mathcal{H} = \mathbb{R}^d$.

One of the key ideas in constructing a shrinkage estimator is based on minimizing an unbiased estimator of the risk, referred to as Stein Unbiased Shrinkage Estimation (SURE). Formally, suppose $\Delta = \mathbb{E}\|\hat{C} - C\|_{\mathcal{H}}^2$ is the mean squared error (i.e., risk) of the empirical estimator $\hat{C}$. Define $\Delta_\alpha = \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2$, where $\hat{C}_\alpha \in \mathcal{C} = \{(1-\alpha)\hat{C} + \alpha f^* : \alpha \in \mathbb{R}\}$. Note that $(\Delta_\alpha)_\alpha$ corresponds to the family of risks associated with the estimators in $\mathcal{C}$. $\check{C}$ is constructed as $\hat{C}_{\hat{\alpha}}$, where $\hat{\alpha} = \arg\min_\alpha \hat{\Delta}_\alpha$, which means $\check{C} = (1 - \hat{\alpha})\hat{C} + \hat{\alpha} f^*$. It can be shown that $\hat{\alpha} = \hat{\Delta}_u / \|\hat{C} - f^*\|_{\mathcal{H}}^2$, so that the shrinkage estimator of C based on SURE is given by

$$\check{C} = \left(1 - \frac{\hat{\Delta}_u}{\|\hat{C} - f^*\|_{\mathcal{H}}^2}\right)\hat{C} + \frac{\hat{\Delta}_u}{\|\hat{C} - f^*\|_{\mathcal{H}}^2} f^*,$$

where $\hat{\Delta}_u$ is an unbiased estimator of Δ.

Another approach to find $\hat{\alpha}$ is based on the observation that $\Delta_\alpha < \Delta$ if and only if $\alpha \in \left(0, \frac{2\Delta}{\Delta + \|C - f^*\|_{\mathcal{H}}^2}\right)$ with $|\Delta_\alpha - \Delta|$ maximized at

$$\alpha_* = \frac{\Delta}{\Delta + \|C - f^*\|_{\mathcal{H}}^2}, \tag{3}$$

which corresponds to the midpoint of the above interval. α_* can be estimated as

$$\tilde{\alpha} = \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{C} - f^*\|_{\mathcal{H}}^2} \tag{4}$$

so that

$$\check{C} = \left(1 - \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{C} - f^*\|_{\mathcal{H}}^2}\right) \hat{C} + \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{C} - f^*\|_{\mathcal{H}}^2} f^*, \tag{5}$$

where $\hat{\Delta}$ is some estimator (not necessarily unbiased) of Δ. This means, the SURE approach first estimates the risk and then minimizes it to find $\hat{\alpha}$ while the latter approach first finds the optimal α (in population) which is then estimated to find $\hat{\alpha}$. The difference in these approaches is an additional term of $\hat{\Delta}$ in the denominator of $\tilde{\alpha}$ compared to that of $\hat{\alpha}$ obtained from SURE.

Muandet et al. (2016) and Zhou, Chen and Huang (2019) considered the latter approach to construct a shrinkage estimator of C and showed the oracle bound

$$\Delta_{\alpha_*} < \Delta_{\tilde{\alpha}} \leq \Delta_{\alpha_*} + O(n^{-3/2}), \text{ as } n \to \infty, \tag{6}$$

which holds for all $\mathbb{P}$ that satisfy certain moment conditions, and also showed $\check{C}$ to be a $\sqrt{n}$-consistent estimator of C. A motivation to consider this approach is as follows: For $f^* = 0$ and $r(x) = K(\cdot, x)$, we have

$$\|\hat{C}\|_{\mathcal{H}}^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} \langle K(\cdot, X_i), K(\cdot, X_j) \rangle_{\mathcal{H}} = \frac{1}{n^2} \sum_{i,j=1}^{n} K(X_i, X_j) = \frac{1}{n^2} \mathbf{1}^\top \mathbf{K} \mathbf{1},$$

where $\mathbf{1} = (1, \overset{n}{\ldots}, 1)^\top$ and $[\mathbf{K}]_{i,j} = K(X_i, X_j)$, $i, j = 1, \ldots, n$. If K is not strictly positive definite, then there exists $(X_1, \ldots, X_n)$ such that $\mathbf{1}^\top \mathbf{K} \mathbf{1} = 0$, which means $\|\hat{C}\|_{\mathcal{H}}^2 = 0$ resulting in an invalid estimator.

1.1. Contributions

In this work, we first generalize and improve the results of (Muandet et al., 2016) and (Zhou, Chen and Huang, 2019) to any k and any separable Hilbert space $\mathcal{H}$ (that is not necessarily an RKHS) without making any parametric assumptions on $\mathbb{P}$. Using the variance decomposition of the U-statistics, we construct an unbiased estimator, $\hat{\Delta}_{\text{general}}$ of Δ, which is used in (5) to construct the shrinkage estimator, $\check{C} = \hat{C}_{\tilde{\alpha}_{\text{general}}}$, where $\tilde{\alpha}_{\text{general}}$ is obtained by replacing $\hat{\Delta}$ by $\hat{\Delta}_{\text{general}}$ in (4). In Theorem 2, we show this estimator to be a $\sqrt{n}$-consistent estimator of C and improve on the oracle bound in (6) by showing

$$\Delta_{\alpha_*} < \Delta_{\tilde{\alpha}_{\text{general}}} \leq \Delta_{\alpha_*} + O(n^{-2}), \text{ as } n \to \infty. \tag{7}$$

Next, we present our key contributions in Theorems 3–6, which are detailed below. For $k \geq 2$, if $r - C$ is $\mathbb{P}$-*complete degenerate* (see Section 2 for the definition), again using the variance decomposition of degenerate U-statistics, we obtain an alternate estimator of Δ, i.e., $\hat{\Delta}_{\text{degen}}$, using which we show (see

Theorem 3) the resulting estimator $\check{C} = \hat{C}_{\tilde{\alpha}_{\text{degen}}}$ (obtained by using $\hat{\Delta}_{\text{degen}}$ in (5)) to be $n^{k/2}$-consistent estimator of C along with significantly faster error rates in the oracle bound:

$$\Delta_{\alpha_*} < \Delta_{\tilde{\alpha}_{\text{degen}}} \leq \Delta_{\alpha_*} + O(n^{-(3k+1)/2}), \text{ as } n \to \infty, \tag{8}$$

where $\tilde{\alpha}_{\text{degen}}$ is obtained by replacing $\hat{\Delta}$ by $\hat{\Delta}_{\text{degen}}$ in (4). Note that in these results (Theorems 2 and 3), the estimator is constructed based on the knowledge of whether $r - C$ is $\mathbb{P}$-complete degenerate or not. However, since it is not easy to verify the $\mathbb{P}$-complete degeneracy of $r - C$, in Theorem 4, we analyze the scenario of using $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$ as an estimator of C irrespective of whether $r - C$ is $\mathbb{P}$-complete degenerate or not—of course, the situation of $r - C$ being $\mathbb{P}$-complete degenerate is handled by Theorem 3. Moreover, this scenario is practically interesting because $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$ is computationally simpler than $\hat{C}_{\tilde{\alpha}_{\text{general}}}$. We show in Theorem 4 that for $k \geq 2$, $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$ is also a $\sqrt{n}$-consistent estimator of C—a surprising result—and satisfies the oracle bound:

$$\Delta_{\alpha_*} < \Delta_{\tilde{\alpha}_{\text{general}}} \leq \Delta_{\alpha_*} + O_{\mathbb{P}}(n^{-3/2}), \text{ as } n \to \infty,$$

without assuming the $\mathbb{P}$-complete degeneracy of $r - C$. This means, $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$ has a slightly weaker oracle bound than the one in (7) but the bound improves significantly to (8) if $r - C$ is $\mathbb{P}$-complete degenerate. To the best of our knowledge, we are not aware of any results in the literature similar to Theorems 3 and 4. All these results are based on Bernstein-type inequalities for unbounded, Hilbert space-valued random elements. For the degenerate case, we extended Bernstein's inequality of Arcones and Giné (1993, Proposition 2.3(c)) and de la Peña and Giné (2012, Theorem 4.1.12(a)) to unbounded Hilbert space-valued random elements (see Theorem A.5 of the Supplementary Material (Utpala and Sriperumbudur, 2024)—from now on referred to as the Supplement), which is of independent interest.

Since all the above-mentioned results are obtained in the non-parametric setting, we are not able to show the exact improvement of the shrinkage estimator over $\hat{C}$ but only show oracle bounds that include an additional error term. In order to understand the behavior of the proposed estimator in the parametric setting, in Section 4, we specialize and analyze our estimator $\hat{C}_{\tilde{\alpha}_{\text{general}}}$ in the well-studied normal mean estimation problem. In other words, we use $k = 1$, $r(x) = x$, $x \in \mathbb{R}^d$ and $\mathbb{P} = N(\mu, \sigma^2 I)$, where μ is the parameter of interest and $\sigma^2 > 0$ may not be known. In this setting with $f^* = 0$, it is easy to verify that

$$\hat{C}_{\tilde{\alpha}_{\text{general}}} = \hat{C}_{\tilde{\alpha}_{\text{degen}}} = \frac{\|\bar{X}\|_2^2}{\frac{S^2}{n} + \|\bar{X}\|_2^2} \bar{X} = \left(1 - \frac{\frac{S^2}{n}}{\frac{S^2}{n} + \|\bar{X}\|_2^2}\right) \bar{X},$$

where $S^2 := \frac{1}{n-1} \sum_{i=1}^n \|X_i - \bar{X}\|_2^2$. In Theorem 5, we show $\hat{C}_{\tilde{\alpha}_{\text{general}}}$ to strictly improve upon $\bar{X}$ in terms of the mean squared error for all $\mu \in \mathbb{R}^d$ if $n \geq 2$ and $d \geq 4 + \frac{2}{n-1}$. A small modification to this estimator, i.e.,

$$\left(1 - \frac{2n-2}{3n-1} \cdot \frac{\frac{S^2}{n}}{\frac{S^2}{n} + \|\bar{X}\|_2^2}\right) \bar{X}$$

yields that for all $d \geq 3$, the above modified estimator strictly improves upon $\bar{X}$ for all $\mu \in \mathbb{R}^d$ (see Theorem 6)—a result similar to that of the James-Stein estimator. The proofs of these results are provided in Section 5 and additional results are provided in the Supplement.

2. Definitions and notation

For $a \triangleq (a_1,\ldots,a_d) \in \mathbb{R}^d$, $b \triangleq (b_1,\ldots,b_d) \in \mathbb{R}^d$, $\|a\|_2 \triangleq \sqrt{\sum_{i=1}^d a_i^2}$ and $\langle a,b\rangle_2 = \sum_{i=1}^d a_i b_i$. ${}^n\mathrm{C}_i = \frac{n!}{(n-i)!i!}$, ${}^n\mathrm{P}_i = \frac{n!}{(n-i)!}$ and S_n denotes the symmetric group on $\{1,\ldots,n\}$ with $\sigma \in S_n$ being a permutation. $\mathrm{U}_k^n(r) = \frac{1}{{}^n\mathrm{P}_k}\sum_{I_k^n} r(X_{i_1},\ldots,X_{i_k})$ denotes a U-statistic with kernel r of order k computed with n variables, where $I_k^n = \{(i_1,\ldots,i_k) : i_1 \neq i_2 \neq \cdots \neq i_k\}$. A function $r : \mathcal{X}^k \to \mathcal{H}$ is said to be *symmetric* if it does not depend on the order of its inputs, i.e., $r(x_1,\ldots,x_k) = r(x_{\sigma(1)},\ldots,x_{\sigma(k)})$, $\forall \sigma \in S_k$. When r is symmetric, $\mathrm{U}_k^n(r)$ reduces to $\mathrm{U}_k^n(r) = \frac{1}{{}^n\mathrm{C}_k}\sum_{J_k^n} r(X_{i_1},\ldots,X_{i_k})$, where $J_k^n = \{(i_1,\ldots,i_k) : 1 < i_1 < i_2 < \cdots < i_k < n\}$. For a symmetric function $r : \mathcal{X}^k \to \mathcal{H}$ and a probability measure $\mathbb{P}$ on $\mathcal{X}$, the *canonical function of order i with respect to* $\mathbb{P}$, denoted as $r_i : \mathcal{X}^i \to \mathcal{H}$, is defined as

$$r_i(x_1,\ldots,x_i) = \int_{\mathcal{X}^{k-i}} r(x_1,\ldots,x_k) \prod_{j=i+1}^{k} d\mathbb{P}(x_j),$$

with the convention $r_0 := \int_{\mathcal{X}^k} r(x_1,\ldots,x_k)\prod_{j=1}^k d\mathbb{P}(x_j)$ and $r_k := r(x_1,\ldots,x_k)$. A symmetric function $r : \mathcal{X}^k \to \mathcal{H}$ is $\mathbb{P}$-*complete degenerate* if *(i)* $\forall i \in \{0,1,\ldots,k-1\}$ and $\forall x_1,\ldots,x_i \in \mathcal{X}$, $r_i(x_1,\ldots,x_i) = 0$; and *(ii)* r_k is not a constant function.

A real-valued symmetric function $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is called a positive definite (pd) kernel if, for all $n \in \mathbb{N}$, $\{\alpha_i\}_{i=1}^n \in \mathbb{R}$ and $\{x_i\}_{i=1}^n \in \mathcal{X}$, we have $\sum_{i,j=1}^n \alpha_i\alpha_j K(x_i,x_j) \geq 0$. A function $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, $(x,y) \mapsto K(x,y)$ is a *reproducing kernel* of the Hilbert space $(\mathscr{H}_K, \langle\cdot,\cdot\rangle_{\mathscr{H}_K})$ of functions if and only if *(i)* $\forall x \in \mathcal{X}$, $K(\cdot,x) \in \mathscr{H}_K$ and *(ii)* $\forall x \in \mathcal{X}$, $\forall f \in \mathscr{H}_K$, $\langle K(\cdot,x), f\rangle_{\mathscr{H}_K} = f(x)$ hold. If such a K exists, then $\mathscr{H}_K$ is called a *reproducing kernel Hilbert space*.

3. Main results

In this section, we present our main results related to the consistency of the shrinkage estimator and oracle bounds for the mean-squared error. Theorem 2 deals with r being a symmetric function while Theorem 3 considers the case of when $r - C$ is $\mathbb{P}$-complete degenerate. We show that the shrinkage estimator has a faster rate of convergence when $r - C$ is $\mathbb{P}$-complete degenerate (see Theorem 3) in contrast to r being simply symmetric (see Theorem 2). We would like to mention that the shrinkage estimators considered in Theorems 2 and 3 are different as their construction is based on whether $r - C$ is $\mathbb{P}$-complete degenerate or not. In Theorem 4, we show that the shrinkage estimator of Theorem 3, i.e., the $\mathbb{P}$-complete degenerate case, is still a $\sqrt{n}$-consistent estimator with a slightly slow error rate in the oracle bound, even if $r - C$ is not $\mathbb{P}$-complete degenerate but only symmetric. This result is interesting as the estimator in the degenerate case is simpler to compute than the estimator in the symmetric case.

Before we present our results, we state the following result, which provides the motivation for the estimator proposed in Theorem 2. This result is a simple extension of (Lee, 2019, Theorem 3) and the claim in the proof of Theorem 2 of Lee (2019) to Hilbert space-valued random elements.

Theorem 1. *Let* $\hat{C} = \frac{1}{{}^n\mathrm{C}_k}\sum_{J_k^n} r(X_{i_1},\ldots,X_{i_k})$ *be a U-statistics estimator of*

$$C = \int_{\mathcal{X}^k} r(x_1,\ldots,x_k)\prod_{i=1}^{k} d\mathbb{P}(x_i),$$

where $r : \mathcal{X}^k \to \mathcal{H}$ is a symmetric function. Let

$$\kappa_{2k-i}(X_1,\ldots,X_{2k-i}) = \langle r(X_1,\ldots,X_k), r(X_1,\ldots,X_i,X_{k+1},\ldots,X_{2k-i})\rangle_{\mathcal{H}}$$

for each $i \in \{0,1,\ldots,k\}$. Then,

$$\mathbb{E}_{X_1,\ldots,X_{2k-i}}\left[\kappa_{2k-i}(X_1,\ldots,X_{2k-i})\right] = \mathbb{E}_{X_1,\ldots,X_i}\|r_i(X_1,\ldots,X_i)\|^2_{\mathcal{H}}. \tag{9}$$

Further,

$$\Delta = \mathbb{E}\|\hat{C}\|^2_{\mathcal{H}} - \|C\|^2_{\mathcal{H}} = \frac{1}{{}^n\mathrm{C}_k}\sum_{i=1}^{k} {}^k\mathrm{C}_i\ {}^{n-k}\mathrm{C}_{k-i}\,\sigma_i^2, \tag{10}$$

where $\sigma_i^2 = \mathbb{E}\|r_i(X_1,\ldots,X_i)\|^2_{\mathcal{H}} - \|\mathbb{E}[r(X_1,\ldots,X_k)]\|^2_{\mathcal{H}}$, with r_i being the canonical function of order i with respect $\mathbb{P}$.

Combining (9) with the observation that

$$\|\mathbb{E}[r(X_1,\ldots,X_k)]\|^2_{\mathcal{H}} = \mathbb{E}\left[\kappa_{2k}(X_1,\ldots,X_{2k})\right]$$

yields

$$\sigma_i^2 = \mathbb{E}_{X_1,\ldots,X_{2k-i}}\left[\kappa_{2k-i}(X_1,\ldots,X_{2k-i})\right] - \mathbb{E}\left[\kappa_{2k}(X_1,\ldots,X_{2k})\right], \tag{11}$$

which therefore can be estimated as

$$\hat{\sigma}_i^2 = \mathrm{U}^n_{2k-i}\left[\kappa_{2k-i}(X_1,\ldots,X_{2k-i})\right] - \mathrm{U}^n_{2k}\left[\kappa_{2k}(X_1,\ldots,X_{2k})\right],$$

resulting in an estimator for Δ as

$$\hat{\Delta}_{\text{general}} = \sum_{i=1}^{k} \frac{{}^k\mathrm{C}_i\ {}^{n-k}\mathrm{C}_{k-i}}{{}^n\mathrm{C}_k}\hat{\sigma}_i^2.$$

Note that $\kappa_{2k-i}(X_1,\ldots,X_{2k-i})$ and $\kappa_{2k}(X_1,\ldots,X_{2k})$ need not be symmetric for any $i \in \{1,\ldots,k\}$ and $k \geq 1$, and therefore, U^n_{2k-i} and U^n_{2k} uses the permutation definition as mentioned in Section 2. Based on the above, a shrinkage estimator of C can be defined as

$$\hat{C}_{\tilde{\alpha}_{\text{general}}} = (1-\tilde{\alpha}_{\text{general}})\hat{C} + \tilde{\alpha}_{\text{general}} f^*, \tag{12}$$

where

$$\tilde{\alpha}_{\text{general}} = \frac{\hat{\Delta}_{\text{general}}}{\hat{\Delta}_{\text{general}} + \|\hat{C} - f^*\|^2_{\mathcal{H}}}.$$

The following result (proved in Section 5.1) analyzes the consistency and mean-squared error of $\hat{C}_{\tilde{\alpha}_{\text{general}}}$.

Theorem 2. *Let $n \geq 2k$, and $r : \mathcal{X}^k \to \mathcal{H}$ be a symmetric function such that $\mathbb{E}\|r(X_1,\ldots,X_k)\|_{\mathcal{H}} < \infty$, where $\mathcal{X}$ is a separable topological space and $\mathcal{H}$ is a separable Hilbert space. Define*

$$\hat{\Delta}_{\text{general}} = \sum_{i=1}^{k} \frac{{}^k\mathrm{C}_i\ {}^{n-k}\mathrm{C}_{k-i}}{{}^n\mathrm{C}_k}\left(\mathrm{U}^n_{2k-i}\left[\kappa_{2k-i}(X_1,\ldots,X_{2k-i})\right] - \mathrm{U}^n_{2k}\left[\kappa_{2k}(X_1,\ldots,X_{2k})\right]\right).$$

Suppose for all $m \geq 2$ *and all* $i \in \{0, 1, \ldots, k\}$,

$$\mathbb{E}\|r(X_1, \ldots, X_k) - C\|_{\mathcal{H}}^m \leq \frac{m!}{2}\beta^2\theta^{m-2}, \tag{13}$$

and

$$\mathbb{E}|\kappa_{2k-i}(X_1, \ldots, X_{2k-i}) - \mathbb{E}[\kappa_{2k-i}(X_1, \ldots, X_{2k-i})]|^m \leq \frac{m!}{2}\beta_i\theta_i^{m-2}, \tag{14}$$

for some finite positive constants $\beta, \theta, \{\beta_i\}_{i=0}^k$, *and* $\{\theta_i\}_{i=0}^k$. *Then, as* $n \to \infty$, *the following hold:*

(i) $|\tilde{\alpha}_{\text{general}} - \alpha_*| = O_{\mathbb{P}}(n^{-\frac{3}{2}})$;

(ii) $|\|\hat{C}_{\tilde{\alpha}_{\text{general}}} - C\|_{\mathcal{H}} - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}| = O_{\mathbb{P}}(n^{-\frac{3}{2}})$;

(iii) $\hat{C}_{\tilde{\alpha}_{\text{general}}}$ *is a* $\sqrt{n}$*-consistent estimator of* C;

(iv) $\min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 \leq \mathbb{E}\|\hat{C}_{\tilde{\alpha}_{\text{general}}} - C\|_{\mathcal{H}}^2 \leq \min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 + O(n^{-2})$,

where $\hat{C}_{\tilde{\alpha}_{\text{general}}}$ *is defined in* (12), α_* *is defined in* (3), *and* $\hat{C}_\alpha = (1-\alpha)\hat{C} + \alpha f^*$.

Remark 1.

(i) It follows from Theorem 2(*iv*) that $\Delta_{\tilde{\alpha}_{\text{general}}} \leq \Delta_{\alpha^*} + O(n^{-2})$ as $n \to \infty$, which when combined with $\Delta_{\alpha^*} < \Delta$, yields $\Delta_{\tilde{\alpha}_{\text{general}}} < \Delta + O(n^{-2})$ as $n \to \infty$, for all $\mathbb{P}$ that satisfy the moment conditions.

(ii) Muandet et al. (2016) considered $k = 1$, $\mathcal{H}$ to be a reproducing kernel Hilbert space (RKHS), $\mathscr{H}_K$, with a continuous reproducing kernel, K, $f^* = 0$ and $r(X) = K(\cdot, X) \in \mathscr{H}_K$, resulting in the problem of shrinkage estimation of the mean element. (Muandet et al., 2016, Theorem 7) provides an oracle bound

$$\min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 \leq \mathbb{E}\|\hat{C}_{\tilde{\alpha}_{\text{general}}} - C\|_{\mathcal{H}}^2 \leq \min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 + O(n^{-3/2}),\ n \to \infty, \tag{15}$$

which Theorem 2(*iv*) improves by a providing an improved error rate of n^{-2}.

(iii) With $k = 2$, $f^* = 0$ and $r(X, Y) = \frac{1}{2}(K(\cdot, X) - K(\cdot, Y)) \otimes_{\mathscr{H}} (K(\cdot, X) - K(\cdot, Y))$, i.e., the shrinkage estimation of the covariance operator on $\mathscr{H}_K$ with $\mathcal{H}$ being the space of Hilbert-Schmidt operators on $\mathscr{H}_K$, (Zhou, Chen and Huang, 2019, Theorem 2) showed (15), which is again improved by Theorem 2. Here $\otimes_{\mathscr{H}_K}$ denotes the tensor product on $\mathscr{H}_K$.

(iv) Clearly the moment conditions of Theorem 2 are satisfied if r is bounded. If r is unbounded, then the moment conditions are quite stringent as they require all the higher moment conditions to exist. However, by only requiring (13) to hold for $m = 2$, i.e., r has a finite central second moment, all the results of Theorem 2 can be obtained but at the cost of achieving polynomial concentration instead of sub-Gaussian concentration in Theorem 2(*i*,*ii*). This claim can be proved by using Theorem A.9 (Chebyshev inequality for U-statistics) in the proof of Theorem 2 instead of Theorem A.4 (Bernstein inequality for U-statistics) of the Supplement.

(v) Suppose there exists a constant $\zeta > 0$ such that $\mathbb{P}(\|r(X_1, \ldots, X_k) - C\|_2 \geq t) \leq 2e^{-t^2/\zeta^2}$ for all $t \geq 0$, i.e., $r(X_1, \ldots, X_k)$ is sub-Gaussian, which implies $r(X_1, \ldots, X_k)$ satisfies the moment condition,

$$\mathbb{E}\|r(X_1, \ldots, X_k) - C\|_{\mathcal{H}}^m \leq m\zeta^m\Gamma(m/2),\ \forall\ m \geq 1. \tag{16}$$

In the following, we show that (16) implies the moment conditions of Theorem 2. To this end, clearly (13) holds since $\mathbb{E}\|r(X_1,\ldots,X_k)-C\|_{\mathcal{H}}^m \le m\zeta^m\Gamma(m/2) \le m\zeta^m\Gamma(m) \le m!\zeta^m$ for all $m \ge 1$. Since

$$\kappa_{2k-i}(X_1,\ldots,X_{2k-i}) = \langle r(X_1,\ldots,X_k), r(X_1,\ldots,X_i,X_{k+1},\ldots,X_{2k-i})\rangle_{\mathcal{H}},$$

for $i \in \{0,\ldots,k\}$, we have

$$\begin{aligned}
&\mathbb{E}|\kappa_{2k-i}(X_1,\ldots,X_{2k-i}) - \mathbb{E}[\kappa_{2k-i}(X_1,\ldots,X_{2k-i})]|^m \\
&= \mathbb{E}\,|\langle r(X_1,\ldots,X_k), r(X_1,\ldots,X_i,X_{k+1},\ldots,X_{2k-i})\rangle_{\mathcal{H}} \\
&\qquad -\mathbb{E}[\langle r(X_1,\ldots,X_k), r(X_1,\ldots,X_i,X_{k+1},\ldots,X_{2k-i})\rangle_{\mathcal{H}}]|^m \\
&\le 2^{m-1}\mathbb{E}\,|\langle r(X_1,\ldots,X_k), r(X_1,\ldots,X_i,X_{k+1},\ldots,X_{2k-i})\rangle_{\mathcal{H}}|^m \\
&\qquad + 2^{m-1}\,|\mathbb{E}[\langle r(X_1,\ldots,X_k), r(X_1,\ldots,X_i,X_{k+1},\ldots,X_{2k-i})\rangle_{\mathcal{H}}]|^m \\
&\le 2^m\mathbb{E}\,|\langle r(X_1,\ldots,X_k), r(X_1,\ldots,X_i,X_{k+1},\ldots,X_{2k-i})\rangle_{\mathcal{H}}|^m \\
&\le 2^m\mathbb{E}\left[\|r(X_1,\ldots,X_k)\|_{\mathcal{H}}^m \|r(X_1,\ldots,X_i,X_{k+1},\ldots,X_{2k-i})\|_{\mathcal{H}}^m\right] \le 2^m\mathbb{E}\,\|r(X_1,\ldots,X_k)\|_{\mathcal{H}}^{2m} \\
&\le 2^m\mathbb{E}\,\|r(X_1,\ldots,X_k)-C+C\|_{\mathcal{H}}^{2m} \le 2^{3m-1}\mathbb{E}\,\|r(X_1,\ldots,X_k)-C\|_{\mathcal{H}}^{2m} + 2^{3m-1}\|C\|_{\mathcal{H}}^{2m} \\
&\le 2(\max(8\zeta^2, 8\|C\|_{\mathcal{H}}^2))^m m!,
\end{aligned}$$

implying that (14) holds. This means, when $k=1$ and $r(x)=x$, $x \in \mathbb{R}^d$, these moment conditions hold if X is sub-Gaussian.

The following examples specialize the proposed shrinkage estimator for the mean element and covariance operator on a Hilbert space.

Example 1 (Mean element, moment generating function and Weierstrass transform). Suppose $k=1$. Then

$$\hat{\Delta}_{\text{general}} = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^n \langle r(X_i), r(X_i)\rangle_{\mathcal{H}} - \frac{1}{{}^nC_2}\sum_{i<j}^n \langle r(X_i), r(X_j)\rangle_{\mathcal{H}}\right]$$

and

$$\|\hat{C}-f^*\|_{\mathcal{H}}^2 = \left\|\frac{1}{n}\sum_{i=1}^n (r(X_i)-f^*)\right\|_{\mathcal{H}}^2 = \frac{1}{n^2}\sum_{i,j}\langle r(X_i), r(X_j)\rangle_{\mathcal{H}} - \frac{2}{n}\sum_{i=1}^n \langle r(X_i), f^*\rangle_{\mathcal{H}} + \|f^*\|_{\mathcal{H}}^2.$$

Define $K(x,y) = \langle r(x), r(y)\rangle_{\mathcal{H}}$, $x,y \in \mathcal{H}$. It is easy to verify that K is a positive definite kernel and therefore a reproducing kernel (Aronszajn, 1950) of some reproducing kernel Hilbert space (RKHS), $\mathscr{H}_K$ so that $K(x,y) = \langle K(\cdot,x), K(\cdot,y)\rangle_{\mathscr{H}_K}$. Note that these quantities match those proposed in (Muandet et al., 2016), where $r(x) = K(\cdot,x)$ and $f^* = 0$, resulting in a mean element of $\mathbb{P}$ in $\mathscr{H}_K$. When $\mathcal{X} = \mathbb{R}^d$ and $r(x) = x$ for $x \in \mathbb{R}^d$, $\mathbb{E}[r(X)]$ corresponds to the mean vector in $\mathbb{R}^d$ and $K(x,y) = \langle x, y\rangle_2$ is the linear kernel. We analyze this scenario in detail in Section 4 when $\mathbb{P}$ is a Gaussian distribution.

The choice of $r(x) = e^{\langle\cdot,x\rangle_2}$ with $\mathcal{H}$ being an RKHS of the exponential kernel, i.e., $K(x,y) = e^{\langle x,y\rangle_2} = \langle r(x), r(y)\rangle_{\mathcal{H}}$, $x,y \in \mathbb{R}^d$, results in a shrinkage estimator for the moment generating function. Equiva-

lently, this choice can be interpreted as

$$r(x) = \left(1, (x_i)_{i=1}^d, (x_{i_1}x_{i_2}/\sqrt{2!})_{i_1,i_2=1}^d, \ldots, \left(\prod_{j=1}^m x_{i_j}/\sqrt{m!}\right)_{i_1,\ldots,i_m=1}^d, \ldots\right)$$

with $\mathcal{H} = \ell^2(\mathbb{N})$. Similarly, the choice of $r(x) = e^{\|\cdot - x\|_2^2}$ with $\mathcal{H}$ being an RKHS of a Gaussian kernel, i.e., $K(x,y) = e^{\|x-y\|_2^2}$, $x, y \in \mathbb{R}^d$, results in a shrinkage estimator for the Weierstrass transform of $\mathbb{P}$.

Example 2 (Covariance operator). Let $\mathcal{H}$ be the space of Hilbert-Schmidt operators defined on a reproducing kernel Hilbert space $\mathscr{H}_K$ with $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, defined on a topological space $\mathcal{X}$. Choosing $k = 2$ and

$$r(X,Y) = \frac{1}{2}(K(\cdot,X) - K(\cdot,Y)) \otimes_{\mathscr{H}_K} (K(\cdot,X) - K(\cdot,Y))$$

yields the covariance operator on $\mathscr{H}_K$. Note that

$$\begin{aligned}
4\langle r(X,Y), r(U,V)\rangle_{\mathcal{H}} &= \big\langle (K(\cdot,X) - K(\cdot,Y)) \otimes_{\mathscr{H}_K} (K(\cdot,X) - K(\cdot,Y)), \\
&\qquad (K(\cdot,U) - K(\cdot,V)) \otimes_{\mathscr{H}_K} (K(\cdot,U) - K(\cdot,V))\big\rangle_{\mathcal{H}} \\
&= \Big\langle K(\cdot,X) - K(\cdot,Y), K(\cdot,U) - K(\cdot,V)\Big\rangle_{\mathscr{H}_K}^2 \\
&= [K(X,U) - K(X,V) - K(Y,U) + K(Y,V)]^2 .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\hat{\Delta}_{\text{general}} &= \frac{2n-4}{{}^n\mathrm{C}_2}\mathrm{U}_3^n[\kappa_3(X_1,X_2,X_3)] - \frac{2n-3}{{}^n\mathrm{C}_2}\mathrm{U}_4^n[\kappa_4(X_1,X_2,X_3,X_4)] + \frac{1}{{}^n\mathrm{C}_2}\mathrm{U}_2^n[\kappa_2(X_1,X_2)] \\
&= \frac{2n-4}{{}^n\mathrm{C}_2}\mathrm{U}_3^n\left[\langle r(X_1,X_2), r(X_1,X_3)\rangle_{\mathcal{H}}\right] + \frac{1}{{}^n\mathrm{C}_2}\mathrm{U}_2^n\left[\langle r(X_1,X_2), r(X_1,X_2)\rangle_{\mathcal{H}}\right] \\
&\qquad - \frac{2n-3}{{}^n\mathrm{C}_2}\mathrm{U}_4^n\left[\langle r(X_1,X_2), r(X_3,X_4)\rangle_{\mathcal{H}}\right] \\
&= \frac{2n-4}{{}^n\mathrm{C}_2 \cdot {}^n\mathrm{P}_3}\sum_{i\neq j\neq l}\big\langle r(X_i,X_j), r(X_i,X_l)\big\rangle_{\mathcal{H}} + \frac{1}{{}^n\mathrm{C}_2 \cdot {}^n\mathrm{P}_2}\sum_{i\neq j}\big\langle r(X_i,X_j), r(X_i,X_j)\big\rangle_{\mathcal{H}} \\
&\qquad - \frac{2n-3}{{}^n\mathrm{C}_2 \cdot {}^n\mathrm{P}_4}\sum_{i\neq j\neq l\neq m}\big\langle r(X_i,X_j), r(X_l,X_m)\big\rangle_{\mathcal{H}} \\
&= \frac{2n-4}{4\cdot {}^n\mathrm{C}_2 \cdot {}^n\mathrm{P}_3}\sum_{i\neq j\neq l}\big[K(X_i,X_i) - K(X_i,X_l) - K(X_i,X_j) + K(X_j,X_l)\big]^2 \\
&\qquad + \frac{1}{4\cdot {}^n\mathrm{C}_2 \cdot {}^n\mathrm{P}_2}\sum_{i\neq j}\big[K(X_i,X_i) - 2K(X_i,X_j) + K(X_j,X_j)\big]^2 \\
&\qquad\qquad - \frac{2n-3}{4\cdot {}^n\mathrm{C}_2 \cdot {}^n\mathrm{P}_4}\sum_{i\neq j\neq l\neq m}\big[K(X_i,X_l) - K(X_i,X_m) - K(X_j,X_l) + K(X_j,X_m)\big]^2 .
\end{aligned}$$

Also for any $f^* \in \mathcal{H}$,

$$\begin{aligned}
\|\hat{C} - f^*\|_{\mathcal{H}}^2 &= \left\| \frac{1}{{}^n\mathrm{P}_2} \sum_{i \neq j} \left(r(X_i, X_j) - f^* \right) \right\|_{\mathcal{H}}^2 \\
&= \frac{1}{{}^n\mathrm{P}_2 \cdot {}^n\mathrm{P}_2} \sum_{i \neq j} \sum_{l \neq m} \left\langle r(X_i, X_j), r(X_l, X_m) \right\rangle_{\mathcal{H}} - \frac{2}{{}^n\mathrm{P}_2} \sum_{i \neq j} \left\langle r(X_i, X_j), f^* \right\rangle_{\mathcal{H}} + \|f^*\|_{\mathcal{H}}^2 \\
&= \frac{1}{4 \cdot {}^n\mathrm{P}_2 \cdot {}^n\mathrm{P}_2} \sum_{i \neq j} \sum_{l \neq m} \left[K(X_i, X_l) - K(X_i, X_m) - K(X_j, X_l) + K(X_j, X_m) \right]^2 \\
&\quad - \frac{1}{{}^n\mathrm{P}_2} \sum_{i \neq j} \left\langle K(\cdot, X_i) - K(\cdot, X_j), f^* \left(K(\cdot, X_i) - K(\cdot, X_j) \right) \right\rangle_{\mathscr{H}_K} + \|f^*\|_{\mathcal{H}}^2.
\end{aligned}$$

We would like to highlight that the expressions provided in (Zhou, Chen and Huang, 2019) for the above quantities are only asymptotically equivalent to ours when $f^* = 0$ because of the approximations the authors employed to simplify their asymptotic analysis.

For $K(x, y) = \langle x, y \rangle_2$, $x, y \in \mathbb{R}^d$ and $f^* = I_d$ (the $d \times d$ identity matrix), it can be shown that (see Proposition B.2 of the Supplement)

$$\hat{\Delta}_{\text{general}} = \frac{1}{(n-2)(n-3)} \sum_{i=1}^{n} \|\tilde{X}_i\|_2^4 - \frac{n(n+1)}{(n-1)^2(n-3)} \mathrm{Tr}[\hat{\Sigma}^2] - \frac{n}{(n-1)(n-2)(n-3)} \mathrm{Tr}^2[\hat{\Sigma}], \text{ and}$$

$$\|\hat{C} - I\|_F^2 = \frac{n^2}{(n-1)^2} \mathrm{Tr}[\hat{\Sigma}^2] - \frac{2n}{n-1} \mathrm{Tr}[\hat{\Sigma}] + d,$$

where $\tilde{X}_i = X_i - \bar{X}$, $i = 1, \ldots, n$, $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{X}_i^\top$, and $\hat{C} = \frac{1}{{}^n\mathrm{C}_2} \sum_{i<j} \frac{(X_i - X_j)(X_i - X_j)^\top}{2}$, with $\|\cdot\|_F$ being the Frobenius norm.

Theorem 2 is based on Bernstein's inequality for Hilbert space-valued U-statistics, which guarantees that $\hat{C}$ and $\hat{C}_{\hat{\alpha}_{\text{general}}}$ are $\sqrt{n}$-consistent estimators of C. However, if $r - C$ is bounded, real-valued, symmetric, $\mathbb{P}$-complete degenerate of $k \geq 2$ variables, Arcones and Giné (1993, Proposition 2.3(c)) and de la Peña and Giné (2012, Theorem 4.1.12(a)) showed that there exist finite positive constants c_1, c_2, c_3 depending only on k such that for all $\delta \in (0, 1)$,

$$\mathbb{P}\left\{ |\mathrm{U}_k^n(r) - C| \geq \sigma \left(\frac{\log \frac{c_1}{\delta}}{c_2 n} \right)^{\frac{k}{2}} + \|r\|_\infty \left(\frac{\log \frac{c_1}{\delta}}{c_3 n} \right)^{\frac{k+1}{2}} \right\} \leq \delta, \tag{17}$$

where $\|r\|_\infty = \sup_{x_1, \ldots, x_k} |r(x_1, \ldots, x_k)|$ and $\sigma^2 = \mathbb{E}(r(X_1, \ldots, X_k) - C)^2$ denotes the variance. We would like to mention that while Arcones and Giné (1993, Proposition 2.3(c)) and de la Peña and Giné (2012, Theorem 4.1.12(a)) presented (17) as an exponential concentration inequality, we wrote it in the confidence interval form in (17) to obtain the rate of convergence (see Appendix C of the Supplement for details). For $k = 2$, (17) implies a rate of n^{-1} to estimate C using $\mathrm{U}_k^n(r)$, which is significantly faster than the usual $n^{-1/2}$-rate that is obtained by Bernstein's inequality that does not take into account the complete degeneracy of $r - C$. Joly and Lugosi (2016) showed a similar result for the median-of-means estimator with the motivation of robust mean estimation in the presence of heavy

tails. In Theorem A.5 of the supplement, we generalize this result to unbounded, $\mathcal{H}$-valued, $\mathbb{P}$-complete degenerate U-statistics using the ideas from (de la Peña and Giné, 2012). Using this result, we devise an estimator of α denoted as $\tilde{\alpha}_{\text{degen}}$, using which we show $\hat{C}_{\tilde{\alpha}_{\text{degen}}} = (1 - \tilde{\alpha}_{\text{degen}})\hat{C} + \tilde{\alpha}_{\text{degen}} f^*$ to be $n^{k/2}$-consistent estimator of C if $r - C$ is $\mathbb{P}$-complete degenerate. Further, we provide improved error bound rates in the oracle inequality associated with $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$.

Our design of $\tilde{\alpha}_{\text{degen}}$ is based on the variance decomposition of U-statistics (see Theorem 1) and the definition of degeneracy. First if $r - C$ is $\mathbb{P}$-complete degenerate we have that $\forall i \in \{0, 1, \ldots, k-1\}$ and $\forall x_1, \ldots, x_i \in \mathcal{X}$, $r_i(x_1, \ldots, x_i) - C = 0$, which implies that $\sigma_i^2 = 0$. It therefore follows from (10) and (11) that

$$\Delta = \frac{1}{{}^n\mathrm{C}_k} \left[\mathbb{E}\left[\kappa_k(X_1, \ldots, X_k)\right] - \mathbb{E}\left[\kappa_{2k}(X_1, \ldots, X_{2k})\right] \right].$$

Using this observation, we consider the following estimator for Δ,

$$\hat{\Delta}_{\text{degen}} = \frac{1}{{}^n\mathrm{C}_k} \left[\mathrm{U}_k^n\left[\kappa_k(X_1, \ldots, X_k)\right] - \mathrm{U}_{2k}^n\left[\kappa_{2k}(X_1, \ldots, X_{2k})\right] \right]$$

so that

$$\tilde{\alpha}_{\text{degen}} = \frac{\hat{\Delta}_{\text{degen}}}{\hat{\Delta}_{\text{degen}} + \|\hat{C}\|_{\mathcal{H}}^2}. \tag{18}$$

Note that $\hat{\Delta}_{\text{general}} = \hat{\Delta}_{\text{degen}}$ when $k = 1$. The following result (proved in Section 5.2) presents the statistical behavior of $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$.

Theorem 3. *Let $n \geq 2k$, $k \geq 2$, $r : \mathcal{X}^k \to \mathcal{H}$ be a symmetric function such that $\mathbb{E}\|r(X_1, \ldots, X_k)\|_{\mathcal{H}}^2 < \infty$ and $r - C$ is $\mathbb{P}$-complete degenerate, where $\mathcal{X}$ is a separable topological space and $\mathcal{H}$ is a separable Hilbert space. Suppose there exist positive constants M, σ_1, σ_2 and $\theta, \theta_1, \theta_2$, such that $\forall p \geq 2$,*

$$\mathbb{E}\left| \|r(X_1, \ldots, X_k) - C\|_{\mathcal{H}}^2 - \mathbb{E}\|r(X_1, \ldots, X_k) - C\|_{\mathcal{H}}^2 \right|^p \leq \frac{p!}{2}\theta^2 M^{p-2}, \tag{19}$$

$$\mathbb{E}\left|\kappa_k(X_1, \ldots, X_k) - \mathbb{E}[\kappa_k(X_1, \ldots, X_k)]\right|^p \leq \frac{p!}{2}\sigma_1^2\theta_1^{p-2}, \text{ and} \tag{20}$$

$$\mathbb{E}\left|\kappa_{2k}(X_1, \ldots, X_{2k}) - \mathbb{E}[\kappa_{2k}(X_1, \ldots, X_{2k})]\right|^p \leq \frac{p!}{2}\sigma_2^2\theta_2^{p-2}. \tag{21}$$

Then, as $n \to \infty$, the following hold:

(i) $\left|\tilde{\alpha}_{\text{degen}} - \alpha_*\right| = O_{\mathbb{P}}(n^{-(2k+1)/2})$;

(ii) $\left| \|\hat{C}_{\tilde{\alpha}_{\text{degen}}} - C\|_{\mathcal{H}} - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}} \right| = O_{\mathbb{P}}(n^{-(2k+1)/2})$;

(iii) $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$ *is a $n^{k/2}$-consistent estimator of C;*

(iv) $\min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 \leq \mathbb{E}\|\hat{C}_{\tilde{\alpha}_{\text{degen}}} - C\|_{\mathcal{H}}^2 \leq \min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 + O(n^{-(3k+1)/2})$,

where α_* is defined in (3), $\tilde{\alpha}_{\text{degen}}$ is defined in (18), and $\hat{C}_\alpha = (1 - \alpha)\hat{C} + \alpha f^*$.

Now, inspired by our analysis of the completely degenerate case, we show that $\tilde{\alpha}_{\text{degen}}$ is a good estimator of α_* even if $r - C$ is not $\mathbb{P}$-complete degenerate. Specifically, we show that without any assumption of degeneracy, $\left|\tilde{\alpha}_{\text{degen}} - \alpha_*\right| = O_{\mathbb{P}}(n^{-1})$ (compared to $O_{\mathbb{P}}(n^{-3/2})$ with $\tilde{\alpha}_{\text{general}}$), $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$ is

a $\sqrt{n}$-consistent estimator of C and more importantly that $\mathbb{E}\|\hat{C}_{\tilde{\alpha}_{\text{degen}}} - C\|_{\mathcal{H}}^2 \leq \min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 + O(n^{-3/2})$ (in contrast to $O(n^{-2})$) as $n \to \infty$. This is surprising because the number of terms in $\hat{\Delta}_{\text{degen}}$ remains constant with k whereas the number of terms in $\hat{\Delta}_{\text{general}}$ grows linearly with k. This means $\hat{\Delta}_{\text{degen}}$ is computationally efficient than $\hat{\Delta}_{\text{general}}$ and therefore is $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$ over $\hat{C}_{\tilde{\alpha}_{\text{general}}}$. These are captured in the following result, which is proved in Section 5.3.

Theorem 4. *Let $n \geq 2k$, $k \geq 2$, $r : \mathcal{X}^k \to \mathcal{H}$ be a symmetric function such that $\mathbb{E}\|r(X_1,\ldots,X_k)\|_{\mathcal{H}}^2 < \infty$, where $\mathcal{X}$ is a separable topological space and $\mathcal{H}$ is a separable Hilbert space. Suppose there exist positive constants $\sigma, \sigma_1, \sigma_2$ and $\theta, \theta_1, \theta_2$ such that $\forall p \geq 2$,*

$$\mathbb{E}\|r(X_1,\ldots,X_k) - C\|_{\mathcal{H}}^p \leq \frac{p!}{2}\sigma^2\theta^{p-2},$$

$$\mathbb{E}\Big|\kappa_k(X_1,\ldots,X_k) - \mathbb{E}[\kappa_k(X_1,\ldots,X_k)]\Big|^p \leq \frac{p!}{2}\sigma_1^2\theta_1^{p-2}, \text{ and}$$

$$\mathbb{E}\Big|\kappa_{2k}(X_1,\ldots,X_{2k}) - \mathbb{E}[\kappa_{2k}(X_1,\ldots,X_{2k})]\Big|^p \leq \frac{p!}{2}\sigma_2^2\theta_2^{p-2}.$$

Then, as $n \to \infty$, the following hold:

(i) $\left|\tilde{\alpha}_{\text{degen}} - \alpha_*\right| = O_{\mathbb{P}}(n^{-1})$;

(ii) $\left|\|\hat{C}_{\tilde{\alpha}_{\text{degen}}} - C\|_{\mathcal{H}} - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}\right| = O_{\mathbb{P}}(n^{-1})$;

(iii) $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$ *is a $\sqrt{n}$-consistent estimator of C;*

(iv) $\min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 \leq \mathbb{E}\|\hat{C}_{\tilde{\alpha}_{\text{degen}}} - C\|_{\mathcal{H}}^2 \leq \min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 + O(n^{-3/2})$.

The reason for assuming $k > 1$ in Theorem 4 is that, when $k = 1$, we have $\hat{\Delta}_{\text{general}} = \hat{\Delta}_{\text{degen}}$, and the claims follow from Theorem 2. Furthermore, the observation in Remark 1(iv) is also valid for Theorems 3 and 4, and can be shown by using Theorem A.9 instead of Theorems A.5 and A.4 (of the Supplement), respectively in the proofs of Theorems 3 and 4.

Remark 2.

(i) Based on Remark 1(v), we would like to highlight that the moment conditions in Theorems 3 and 4 are satisfied if $r(X_1,\ldots,X_k)$ satisfies the moment condition,

$$\mathbb{E}\|r(X_1,\ldots,X_k) - C\|_{\mathcal{H}}^p \leq p\zeta^p\Gamma(p/2), \ \forall\, p \geq 1,$$

which in turn is implied if $r(X_1,\ldots,X_k)$ is sub-Gaussian.

(ii) Interestingly, as a converse, we show below that (19) and

$$\mathbb{E}\|r(X_1,\ldots,X_k) - C\|_{\mathcal{H}}^2 \leq \sigma^2 \tag{22}$$

combinedly imply that there exist constants σ_3 and θ_3 such that $\mathbb{E}\|r(X_1,\ldots,X_k) - C\|_{\mathcal{H}}^p \leq \frac{p!}{2}\sigma_3^2\theta_3^{p-2}$ for all $p \geq 2$, i.e., $r(X_1,\ldots,X_k)$ is sub-Gaussian. Similarly, (20) and (21) (which also appear in Theorem 4) along with (22) combinedly imply that there exist constants σ_4 and θ_4 such that $\mathbb{E}\|r(X_1,\ldots,X_k) - C\|_{\mathcal{H}}^p \leq \frac{p!}{2}\sigma_4^2\theta_4^{p-2}$ for all $p \geq 2$. This means, (19)–(21) along with (22) imply that $r(X_1,\ldots,X_k)$ is sub-Gaussian, which in combination with Remark 2(i) implies the the

equivalence of these conditions to the sub-Gaussianity of $r(X_1,\dots,X_k)$. To prove the first claim, for $p \geq 2$, we have

$$\begin{aligned}
&\mathbb{E}\|r(X_1,\dots,X_k) - C\|_{\mathcal{H}}^p = \mathbb{E}\left[\|r(X_1,\dots,X_k) - C\|_{\mathcal{H}}^2\right]^{p/2} \\
&\leq 2^{\frac{p}{2}-1}\mathbb{E}\left|\|r(X_1,\dots,X_k) - C\|_{\mathcal{H}}^2 - \mathbb{E}\|r(X_1,\dots,X_k) - C\|_{\mathcal{H}}^2\right|^{p/2} \\
&\qquad + 2^{\frac{p}{2}-1}\left(\mathbb{E}\|r(X_1,\dots,X_k) - C\|_{\mathcal{H}}^2\right)^{p/2} \\
&\overset{(22)}{\leq} 2^{\frac{p}{2}-1}\sqrt{\mathbb{E}\left|\|r(X_1,\dots,X_k) - C\|_{\mathcal{H}}^2 - \mathbb{E}\|r(X_1,\dots,X_k) - C\|_{\mathcal{H}}^2\right|^p} + \frac{1}{2}(2\sigma^2)^{\frac{p}{2}} \\
&\overset{(19)}{\leq} \sqrt{p!}\theta M^{\frac{p-2}{2}}2^{\frac{p-3}{2}} + \frac{1}{2}(2\sigma^2)^{\frac{p}{2}} \\
&\leq \frac{p!}{2}\max(2\sqrt{2}\theta, 4\sigma^2)\max\left(\sqrt{2M}, \sqrt{2\sigma^2}\right)^{p-2}.
\end{aligned}$$

A similar calculation involving (20)–(22) is provided in Appendix D of the Supplement.

Example 3 (Covariance operator). For the same setting as in Example 2, we obtain

$$\begin{aligned}
\hat{\Delta}_{\text{degen}} &= \frac{1}{{}^n\mathrm{C}_2}\mathrm{U}_2^n\left[\kappa_2(X_1,X_2)\right] - \frac{1}{{}^n\mathrm{C}_2}\mathrm{U}_4^n\left[\kappa_4(X_1,X_2,X_3,X_4)\right] \\
&= \frac{1}{4\cdot {}^n\mathrm{C}_2\cdot {}^n\mathrm{P}_2}\sum_{i\neq j}\left[K(X_i,X_i) - 2K(X_i,X_j) + K(X_j,X_j)\right]^2 \\
&\quad - \frac{1}{4\cdot {}^n\mathrm{C}_2\cdot {}^n\mathrm{P}_4}\sum_{i\neq j\neq l\neq m}\left[K(X_i,X_l) - K(X_i,X_m) - K(X_j,X_l) + K(X_j,X_m)\right]^2,
\end{aligned}$$

which reduces to

$$\hat{\Delta}_{\text{degen}} = \frac{n(n^2-3n+4)}{2\cdot {}^n\mathrm{C}_2\cdot {}^n\mathrm{P}_4}\sum_{i=1}^n\|\tilde{X}_i\|_2^4 - \frac{2n^2(n-2)}{{}^n\mathrm{C}_2\cdot {}^n\mathrm{P}_4}\mathrm{Tr}[\hat{\Sigma}^2] + \frac{n^2(n^2-5n+4)}{2\cdot {}^n\mathrm{C}_2\cdot {}^n\mathrm{P}_4}\mathrm{Tr}^2[\hat{\Sigma}],$$

when $K(x,y) = \langle x,y\rangle_2$, $x,y\in\mathbb{R}^d$. See Proposition B.2 of the Supplement for details.

The proposed shrinkage estimators $\hat{C}_{\tilde{\alpha}_{\text{general}}}$ and $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$ can be shown to be solutions to regularized minimization problems. Since

$$\hat{C}_\alpha = \arg\inf_{g\in\mathcal{H}}\frac{1}{{}^n\mathrm{C}_k}\sum_{(i_1,\dots,i_k)\in J_k^n}\left\|r(X_{i_1},\dots,X_{i_k}) - g\right\|_{\mathcal{H}}^2 + \frac{\alpha}{1-\alpha}\|g - f^*\|_{\mathcal{H}}^2,$$

where $\frac{\alpha}{1-\alpha}$, $0<\alpha<1$ acts as the regularization parameter, it follows that the choice of $\frac{\tilde{\alpha}_{\text{general}}}{1-\tilde{\alpha}_{\text{general}}}$ and $\frac{\tilde{\alpha}_{\text{degen}}}{1-\tilde{\alpha}_{\text{degen}}}$ as regularization parameters yield $\hat{C}_{\tilde{\alpha}_{\text{general}}}$ and $\hat{C}_{\tilde{\alpha}_{\text{degen}}}$, respectively. This demonstrates the regularization effect of shrinkage estimators. A similar result was shown in (Muandet et al., 2016) when $f^*=0$, $\mathcal{H}=\mathscr{H}_K$, $k=1$ and $r(x)=K(\cdot,x)$.

4. Normal mean estimation

In Section 3, we only established oracle bounds on the mean squared error that include an error term, since no parametric assumptions were made on $\mathbb{P}$. In this section, we study the estimator $\hat{C}_{\tilde{\alpha}_{\text{general}}}$ when $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \mathbb{R}^d$, $r(x) = x$ and $\mathbb{P}$ is a normal distribution, i.e., the shrinkage estimation of normal mean. Note that the degenerate case is not applicable in this setting as $k = 1$. This is the classical setting studied heavily in the literature (Brandwein and Strawderman, 2012). Since $\mathbb{P}$ is Gaussian, we show that concrete results can be obtained on the mean-squared error of $\hat{C}_{\tilde{\alpha}_{\text{general}}}$, in contrast to oracle inequalities of the previous section.

Define $C = \int_{\mathcal{X}} r(x)\, d\mathbb{P}(x) = \int x\, d\mathbb{P}(x) =: \mu$ and $\hat{C} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X} =: \hat{\mu}$. In this setting with $f^* = 0$, it is easy to verify that

$$\begin{aligned}
\hat{\Delta}_{\text{general}} &= \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^n \|X_i\|_2^2 - \frac{1}{n(n-1)}\sum_{i\neq j}\langle X_i, X_j\rangle_2\right] \\
&= \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^n \|X_i\|_2^2 - \frac{1}{n(n-1)}\sum_{i,j=1}^n \langle X_i, X_j\rangle_2 + \frac{1}{n(n-1)}\sum_{i=1}^n \|X_i\|_2^2\right] \\
&= \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^n \|X_i\|_2^2 - \frac{n}{n-1}\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_2^2\right] \\
&= \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^n \|X_i\|_2^2 - \frac{n}{n-1}\|\bar{X}\|_2^2\right] = \frac{1}{n(n-1)}\sum_{i=1}^n \|X_i - \bar{X}\|_2^2 =: \frac{S^2}{n},
\end{aligned}$$

and

$$\hat{C}_{\tilde{\alpha}_{\text{general}}} =: \breve{\mu} = \frac{\|\bar{X}\|_2^2}{\frac{S^2}{n} + \|\bar{X}\|_2^2}\bar{X} = \left(1 - \frac{\frac{S^2}{n}}{\frac{S^2}{n} + \|\bar{X}\|_2^2}\right)\bar{X}.$$

The following result (proved in Section 5.4) shows that the shrinkage estimator, $\breve{\mu}$ has strictly smaller mean squared error compared to $\hat{\mu}$ when $d \geq 4 + \frac{2}{n-1}$.

Theorem 5. *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N_d(\mu, \sigma^2 I)$. For $n \geq 2$ and $d \geq 4 + \frac{2}{n-1}$,*

$$\mathbb{E}\,\|\breve{\mu} - \mu\|_2^2 < \mathbb{E}\|\hat{\mu} - \mu\|_2^2$$

for all $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$.

When $n = 2$, $\breve{\mu}$ improves upon $\hat{\mu}$ for $d \geq 6$. For all $n \geq 3$, the improvement phenomenon occurs for $d \geq 5$. By slightly modifying the estimator $\breve{\mu}$, the following result (proved in Section 5.5) shows improvement over $\hat{\mu}$ when $d \geq 3$.

Theorem 6. *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N_d(\mu, \sigma^2 I)$. For $n \geq 2$, $c \in (0,2)$ and $d \geq \frac{4}{2-c} + \frac{2c}{(n-1)(2-c)}$,*

$$\mathbb{E}\|\breve{\mu}_c - \mu\|_2^2 < \mathbb{E}\|\hat{\mu} - \mu\|_2^2 \tag{23}$$

for all $\mu \in \mathbb{R}^d$ *and* $\sigma^2 > 0$ *where* $\breve{\mu}_c = (1 - c\tilde{\alpha}_{\text{general}})\hat{\mu}$ *with* $\tilde{\alpha}_{\text{general}} = \frac{\frac{S^2}{n}}{\frac{S^2}{n} + \|\bar{X}\|_2^2}$. *In particular, if* $c = \frac{2n-2}{3n-1}$, *then (23) holds for all* $d \geq 3$.

It is interesting to note that the estimator $\breve{\mu}_c$ with $c = \frac{2n-2}{3n-1}$ behaves similar to that of the James-Stein estimator in showing improvement over $\hat{\mu}$ for $d \geq 3$ but with important differences. $\breve{\mu}$ has an additional term of $\frac{S^2}{n}$ in the denominator and c depends only on n instead of d—James-Stein estimator has $c = d - 2$. Because of this additional term in the denominator, establishing Theorem 6 is far more tedious than proving such a result for the James-Stein estimator. In fact, because of this additional term in the denominator, we are not able to establish concrete results in the non-spherical Gaussian scenario and it remains an open question.

5. Proofs

The following is a master theorem, which we will repeatedly use to prove the results of Section 3.

Theorem 7. *Let* $\hat{C}$ *and* $\hat{\Delta}$ *be estimators of* C *and* Δ, *respectively, where* $\Delta = \mathbb{E}\|\hat{C} - C\|_{\mathcal{H}}^2$. *For* $\tau > 0$, *suppose there exist positive constants* $a, b, c_1, c_2, c_3, d_1, d_2$ *that do not depend on* τ *and* n *such that the following statements hold with probability at least* $1 - c_3 e^{-\tau}$:

$$\|\hat{C} - C\|_{\mathcal{H}} \leq c_1 \left(\frac{1+\tau}{n}\right)^{a/2} + c_2 \left(\frac{1+\tau}{n}\right)^{(a+1)/2}, \tag{24}$$

$$|\hat{\Delta} - \Delta| \leq d_1 \left(\frac{1+\tau}{n}\right)^{b/2} + d_2 \left(\frac{1+\tau}{n}\right)^{(b+1)/2}.$$

Define $\alpha_* = \frac{\Delta}{\Delta + \|C - f^*\|_{\mathcal{H}}^2}$ *and* $\hat{C}_{\tilde{\alpha}} = (1 - \tilde{\alpha})\hat{C} + \tilde{\alpha} f^*$ *as an estimator of* C *where* $\tilde{\alpha} = \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{C} - f^*\|_{\mathcal{H}}^2}$. *Then as* $n \to \infty$, *the following hold:*

(i) $|\tilde{\alpha} - \alpha_*| = O_{\mathbb{P}}\left(n^{-\min\{3a,b\}/2}\right)$;

(ii) $|\|\hat{C}_{\tilde{\alpha}} - C\|_{\mathcal{H}} - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}| = O_{\mathbb{P}}\left(n^{-\min\{3a,b\}/2}\right)$;

(iii) $\hat{C}_{\tilde{\alpha}}$ *is a* $n^{\min\{a,b\}/2}$*- consistent estimator of* C;

(iv) $\min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 \leq \mathbb{E}\|\hat{C}_{\tilde{\alpha}} - C\|_{\mathcal{H}}^2 \leq \min_\alpha \mathbb{E}\|\hat{C}_\alpha - C\|_{\mathcal{H}}^2 + O(n^{-\min\{4a,(a+b),2b\}/2})$.

Proof. Consider

$$\begin{aligned}
\alpha_* - \tilde{\alpha} &= \frac{\Delta}{\Delta + \|C - f^*\|_{\mathcal{H}}^2} - \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{C} - f^*\|_{\mathcal{H}}^2} = \frac{\Delta\|\hat{C} - f^*\|_{\mathcal{H}}^2 - \hat{\Delta}\|C - f^*\|_{\mathcal{H}}^2}{(\Delta + \|C - f^*\|_{\mathcal{H}}^2)(\hat{\Delta} + \|\hat{C} - f^*\|_{\mathcal{H}}^2)} \\
&= \frac{\Delta\left(\|\hat{C} - f^*\|_{\mathcal{H}}^2 - \|C - f^*\|_{\mathcal{H}}^2\right) + \|C - f^*\|_{\mathcal{H}}^2\left(\Delta - \hat{\Delta}\right)}{(\Delta + \|C - f^*\|_{\mathcal{H}}^2)(\hat{\Delta} + \|\hat{C} - f^*\|_{\mathcal{H}}^2)} \\
&= \frac{\alpha_*\left(\|\hat{C} - f^*\|_{\mathcal{H}}^2 - \|C - f^*\|_{\mathcal{H}}^2\right) + (1 - \alpha_*)\left(\Delta - \hat{\Delta}\right)}{\hat{\Delta} + \|\hat{C} - f^*\|_{\mathcal{H}}^2}
\end{aligned}$$

$$= \frac{\alpha_* \left(\|\hat{C} - f^*\|_{\mathcal{H}}^2 - \|C - f^*\|_{\mathcal{H}}^2 \right) + (1 - \alpha_*) \left(\Delta - \hat{\Delta} \right)}{\Delta + \|C - f^*\|_{\mathcal{H}}^2 - \left(\|C - f^*\|_{\mathcal{H}}^2 - \|\hat{C} - f^*\|_{\mathcal{H}}^2 \right) + \left(\hat{\Delta} - \Delta \right)}$$

from which we have

$$|\tilde{\alpha} - \alpha_*| \leq \frac{\alpha_* \left| \|C - f^*\|_{\mathcal{H}}^2 - \|\hat{C} - f^*\|_{\mathcal{H}}^2 \right| + (1 - \alpha_*) \left| \hat{\Delta} - \Delta \right|}{\Delta + \|C - f^*\|_{\mathcal{H}}^2 - \left| \|C - f^*\|_{\mathcal{H}}^2 - \|\hat{C} - f^*\|_{\mathcal{H}}^2 \right| - \left| \hat{\Delta} - \Delta \right|} \tag{25}$$

if

$$\Delta + \|C - f^*\|_{\mathcal{H}}^2 > \left| \|C - f^*\|_{\mathcal{H}}^2 - \|\hat{C} - f^*\|_{\mathcal{H}}^2 \right| + \left| \hat{\Delta} - \Delta \right|. \tag{26}$$

(i) Consider

$$\|\hat{C} - C\|_{\mathcal{H}}^2 \overset{(*)}{\leq} \left[c_1 \left(\frac{1+\tau}{n} \right)^{a/2} + c_2 \left(\frac{1+\tau}{n} \right)^{(a+1)/2} \right]^2 \overset{(**)}{\leq} d \left(\frac{1+\tau}{n} \right)^a, \tag{27}$$

for some constant d that doesn't depend on τ, n and we used (24) in $(*)$ and assume $\frac{1+\tau}{n} \leq 1$ in $(**)$. Using Lemma A.1 (see the Supplement) for (27) yields $\Delta \leq e_1 n^{-a}$, which implies that,

$$\alpha_* = \frac{\Delta}{\Delta + \|C - f^*\|_{\mathcal{H}}^2} \leq \frac{\Delta}{\|C - f^*\|_{\mathcal{H}}^2} \leq \frac{e_2}{n^a}, \tag{28}$$

for some positive constants e_1, e_2 that does not depend on τ and n. Next, $|\|C - f^*\|_{\mathcal{H}}^2 - \|\hat{C} - f^*\|_{\mathcal{H}}^2|$ can be bounded as

$$\begin{aligned} &|\|C - f^*\|_{\mathcal{H}}^2 - \|\hat{C} - f^*\|_{\mathcal{H}}^2| \leq \|\hat{C} - C\|_{\mathcal{H}}^2 + 2 \|C - f^*\|_{\mathcal{H}} \|\hat{C} - C\|_{\mathcal{H}} \\ &\overset{(*)}{\leq} d \left(\frac{1+\tau}{n} \right)^a + 2 \|C - f^*\|_{\mathcal{H}} \left(c_1 \left(\frac{1+\tau}{n} \right)^{a/2} + c_2 \left(\frac{1+\tau}{n} \right)^{(a+1)/2} \right) \leq f \left(\frac{1+\tau}{n} \right)^{a/2}, \end{aligned} \tag{29}$$

for some positive constant f that does not depend on τ and n, and we used (24) and (27) in $(*)$ along with the assumption that $n \geq \tau + 1$. Also, note that there exists a constant g such that

$$|\hat{\Delta} - \Delta| \leq d_1 \left(\frac{1+\tau}{n} \right)^{b/2} + d_2 \left(\frac{1+\tau}{n} \right)^{(b+1)/2} \leq g \left(\frac{1+\tau}{n} \right)^{b/2}. \tag{30}$$

If $n \geq \max \left\{ 1, \left(\frac{4f}{\|C - f^*\|_{\mathcal{H}}^2} \right)^{2/a}, \left(\frac{4g}{\|C - f^*\|_{\mathcal{H}}^2} \right)^{2/b} \right\} (1 + \tau)$, the denominator in (25) can be bounded as

$$\begin{aligned} &\Delta + \|C - f^*\|_{\mathcal{H}}^2 - \left| \|C - f^*\|_{\mathcal{H}}^2 - \|\hat{C} - f^*\|_{\mathcal{H}}^2 \right| - \left| \hat{\Delta} - \Delta \right| \\ &\geq \|C - f^*\|_{\mathcal{H}}^2 - f \left(\frac{1+\tau}{n} \right)^{\frac{a}{2}} - g \left(\frac{1+\tau}{n} \right)^{\frac{b}{2}} \\ &\geq \|C - f^*\|_{\mathcal{H}}^2 - \frac{1}{4} \|C - f^*\|_{\mathcal{H}}^2 - \frac{1}{4} \|C - f^*\|_{\mathcal{H}}^2 \geq \frac{1}{2} \|C - f^*\|_{\mathcal{H}}^2. \end{aligned} \tag{31}$$

Therefore, using (28)–(31) in (25), we obtain

$$|\tilde{\alpha} - \alpha_*| \leq h\left(\frac{1+\tau}{n}\right)^{\min\{3a,b\}/2}, \tag{32}$$

where h is a constant that does not depend on τ and n, thereby yielding the result.

(ii) We now bound $|\|\hat{C}_{\tilde{\alpha}} - C\|_{\mathcal{H}} - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}|$ as

$$\begin{aligned}
&|\|\hat{C}_{\tilde{\alpha}} - C\|_{\mathcal{H}} - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}| \\
&\leq \|\hat{C}_{\alpha_*} - \hat{C}_{\tilde{\alpha}}\|_{\mathcal{H}} \leq |\tilde{\alpha} - \alpha_*|\|\hat{C} - C\|_{\mathcal{H}} + |\tilde{\alpha} - \alpha_*|\,\|C - f^*\|_{\mathcal{H}} \\
&\leq |\tilde{\alpha} - \alpha_*|\left[\|\hat{C} - C\|_{\mathcal{H}} + \|C - f^*\|_{\mathcal{H}}\right] \\
&\overset{(32)}{\leq} h\left(\frac{1+\tau}{n}\right)^{\min\{3a,b\}/2}\left[c_1\left(\frac{1+\tau}{n}\right)^{a/2} + c_2\left(\frac{1+\tau}{n}\right)^{(a+1)/2} + \|C - f^*\|_{\mathcal{H}}\right] \\
&\leq p\left(\frac{1+\tau}{n}\right)^{\min\{3a,b\}/2},
\end{aligned} \tag{33}$$

where p is constant that does not depend on τ and n and the result follows.

(iii) $\|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}$ can be bounded as

$$\begin{aligned}
\|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}} &= \|(1-\alpha_*)(\hat{C} - C) + \alpha_* f^* - \alpha_* C\|_{\mathcal{H}} \\
&\leq (1-\alpha_*)\|\hat{C} - C\|_{\mathcal{H}} + \alpha_*\,\|C - f^*\|_{\mathcal{H}} \\
&\leq c_1\left(\frac{1+\tau}{n}\right)^{a/2} + c_2\left(\frac{1+\tau}{n}\right)^{(a+1)/2} + \frac{e_2}{n^a}\|C - f^*\|_{\mathcal{H}} \leq q\left(\frac{1+\tau}{n}\right)^{a/2},
\end{aligned} \tag{34}$$

where q is constant that does not depend on τ and n. The result follows from (33) and (34) by noting that

$$\|\hat{C}_{\tilde{\alpha}} - C\|_{\mathcal{H}} \leq \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}} + O_{\mathbb{P}}(n^{-\min\{3a,b\}/2}) \leq O_{\mathbb{P}}(n^{-a/2}) + O_{\mathbb{P}}(n^{-\min\{3a,b\}/2})$$

as $n \to \infty$.

(iv) We now bound $\|\hat{C}_{\tilde{\alpha}} - C\|_{\mathcal{H}}^2 - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}^2$ as

$$\begin{aligned}
&\|\hat{C}_{\tilde{\alpha}} - C\|_{\mathcal{H}}^2 - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}^2 \\
&= \left(\|\hat{C}_{\tilde{\alpha}} - C\|_{\mathcal{H}} - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}\right)^2 + 2\|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}\left(\|\hat{C}_{\tilde{\alpha}} - C\|_{\mathcal{H}} - \|\hat{C}_{\alpha_*} - C\|_{\mathcal{H}}\right), \\
&\leq \left(p\left(\frac{1+\tau}{n}\right)^{\min\{3a,b\}/2}\right)^2 + 2q\left(\frac{1+\tau}{n}\right)^{a/2}\left(p\left(\frac{1+\tau}{n}\right)^{\min\{3a,b\}/2}\right) \leq s\left(\frac{1+\tau}{n}\right)^{\min\{4a,a+b,2b\}/2},
\end{aligned}$$

where s is a constant that does not depend on τ and n. The result therefore follows by using Lemma A.1 of the Supplement. Finally, note that the assumptions on n and the condition in (26) hold as $n \to \infty$. □

5.1. Proof of Theorem 2

Note that

$$\|\hat{C} - C\|_{\mathcal{H}} = \left\| \mathrm{U}_k^n \left[r(X_1, \ldots, X_k) - \mathbb{E}\left(r(X_1, \ldots, X_k)\right) \right] \right\|_{\mathcal{H}}.$$

Using Theorem A.4 of the Supplement on $r(X_1, \ldots, X_k) - \mathbb{E}(r(X_1, \ldots, X_k))$, we get that with probability at least $1 - \exp(-\tau)$,

$$\|\hat{C} - C\|_{\mathcal{H}} \leq 4\beta\sqrt{k}\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}} + 4\theta k\left(\frac{1+\tau}{n}\right) = c_1\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}} + c_2\left(\frac{1+\tau}{n}\right), \tag{35}$$

where $c_1, c_2 > 0$ are constants that do not depend on τ and n. Now consider

$$\begin{aligned}
&|\hat{\Delta}_{\text{general}} - \Delta| \\
&= \left| \sum_{i=1}^{k} \frac{{}^{k}\mathrm{C}_i \, {}^{n-k}\mathrm{C}_{k-i}}{{}^{n}\mathrm{C}_k} \left(\mathrm{U}_{2k-i}^n \left[\kappa_{2k-i}(X_1, \ldots, X_{2k-i}) \right] - \mathrm{U}_{2k}^n \left[\kappa_{2k}(X_1, \ldots, X_{2k}) \right] \right) \right. \\
&\qquad \left. - \sum_{i=1}^{k} \frac{{}^{k}\mathrm{C}_i \, {}^{n-k}\mathrm{C}_{k-i}}{{}^{n}\mathrm{C}_k} \left(\mathbb{E}\left[\kappa_{2k-i}(X_1, \ldots, X_{2k-i}) \right] - \mathbb{E}\left[\kappa_{2k}(X_1, \ldots, X_{2k}) \right] \right) \right| \\
&\overset{(*)}{\leq} \spadesuit,
\end{aligned}$$

where we used Vandermonde's identity in $(*)$ and

$$\begin{aligned}
\spadesuit := &\sum_{i=1}^{k} \frac{{}^{k}\mathrm{C}_i \, {}^{n-k}\mathrm{C}_{k-i}}{{}^{n}\mathrm{C}_k} \left| \mathrm{U}_{2k-i}^n \left[\kappa_{2k-i}(X_1, \ldots, X_{2k-i}) - \mathbb{E}\left[\kappa_{2k-i}(X_1, \ldots, X_{2k-i}) \right] \right] \right| \\
&+ \left| \frac{{}^{n-k}\mathrm{C}_k}{{}^{n}\mathrm{C}_k} - 1 \right| \left| \mathrm{U}_{2k}^n \left[\kappa_{2k}(X_1, \ldots, X_{2k}) - \mathbb{E}\left[\kappa_{2k}(X_1, \ldots, X_{2k}) \right] \right] \right|.
\end{aligned}$$

Now applying Theorem A.4 of the Supplement to

$$\kappa_{2k-i}(X_1, \ldots, X_{2k-i}) - \mathbb{E}\left[\kappa_{2k-i}(X_1, \ldots, X_{2k-i}) \right]$$

for each $i \in \{0, 1, \ldots, k\}$, we have that with probability at least $1 - (k+1)\exp(-\tau)$,

$$\begin{aligned}
\spadesuit &\leq \sum_{i=1}^{k} \frac{{}^{k}\mathrm{C}_i \, {}^{n-k}\mathrm{C}_{k-i}}{{}^{n}\mathrm{C}_k} \left[4\beta_i\sqrt{2k-i}\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}} + 4\theta_i(2k-i)\left(\frac{1+\tau}{n}\right) \right] \\
&\quad + \left| \frac{{}^{n-k}\mathrm{C}_k}{{}^{n}\mathrm{C}_k} - 1 \right| \left[4\beta_{2k}\sqrt{2k}\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}} + 4\theta_{2k}(2k)\left(\frac{1+\tau}{n}\right) \right] \\
&\leq c_3 \left[\sqrt{2k}\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}} + 2k\left(\frac{1+\tau}{n}\right) \right] \left[\sum_{i=1}^{k} \frac{{}^{k}\mathrm{C}_i \, {}^{n-k}\mathrm{C}_{k-i}}{{}^{n}\mathrm{C}_k} + \left| \frac{{}^{n-k}\mathrm{C}_k}{{}^{n}\mathrm{C}_k} - 1 \right| \right]
\end{aligned}$$

$$\overset{(*)}{=} c_3 \left[\sqrt{2k} \left(\frac{1+\tau}{n} \right)^{\frac{1}{2}} + 2k \left(\frac{1+\tau}{n} \right) \right] \left[1 - \frac{{}^{n-k}\mathrm{C}_k}{{}^{n}\mathrm{C}_k} + \left| \frac{{}^{n-k}\mathrm{C}_k}{{}^{n}\mathrm{C}_k} - 1 \right| \right]$$

$$\overset{(**)}{=} 2c_3 \left[\sqrt{2k} \left(\frac{1+\tau}{n} \right)^{\frac{1}{2}} + 2k \left(\frac{1+\tau}{n} \right) \right] \left[1 - \frac{{}^{n-k}\mathrm{C}_k}{{}^{n}\mathrm{C}_k} \right]$$

$$\overset{(\dagger)}{\leq} 2c_3 \left[\sqrt{2k} \left(\frac{1+\tau}{n} \right)^{\frac{1}{2}} + 2k \left(\frac{1+\tau}{n} \right) \right] \left[\frac{n^k - (n-2k)^k}{n^k} \right]$$

$$\overset{(\ddagger)}{\leq} 2c_3 \left[\sqrt{2k} \left(\frac{1+\tau}{n} \right)^{\frac{1}{2}} + 2k \left(\frac{1+\tau}{n} \right) \right] \left[\frac{2n^{k-1}k^2}{n^k} \right]$$

$$\leq c_4 \left(\frac{1+\tau}{n} \right)^{3/2} + c_5 \left(\frac{1+\tau}{n} \right)^{2}, \tag{36}$$

where we used Vandermonde's identity that $\sum_{i=0}^{k} \frac{{}^{k}\mathrm{C}_i \, {}^{n-k}\mathrm{C}_{k-i}}{{}^{n}\mathrm{C}_k} = 1$ in $(*)$, $\frac{{}^{n-k}\mathrm{C}_k}{{}^{n}\mathrm{C}_k} < 1$ in $(**)$, and $\frac{{}^{n-k}\mathrm{C}_k}{{}^{n}\mathrm{C}_k} \geq \frac{(n-2k)^k}{n^k}$ in $(\dagger)$. In $(\ddagger)$, we used $0 < b < a \implies a^k - b^k \leq k a^{k-1}(a-b)$. Now applying Theorem 7 with $a = 1$ (see (35)) and $b = 3$ (see (36)), the result follows.

5.2. Proof of Theorem 3

Using Theorem A.5 of the Supplement on $r(X_1, \ldots, X_k) - \mathbb{E}(r(X_1, \ldots, X_k))$ yields that with probability at least $1 - \tilde{a}\exp(-\tau)$,

$$\|\hat{C} - C\|_{\mathcal{H}} \leq q k^k \left(\frac{\tau}{na'} \right)^{k/2} + M k^k \left(\frac{\tau}{na''} \right)^{(k+1)/2}, \tag{37}$$

where $\tilde{a}, a'$ and a'' are positive constants, and $q = (\theta + \sigma^2 + \theta^2 M^{-1})$ with $\sigma^2 = \mathbb{E}\|r(X_1, \ldots, X_k)\|_{\mathcal{H}}^2 - \|C\|_{\mathcal{H}}^2$. Therefore,

$$\begin{aligned}
|\hat{\Delta}_{\text{degen}} - \Delta| &= \Bigg| ({}^{n}\mathrm{C}_k)^{-1} \Big[\mathrm{U}_k^n \big[\kappa_k(X_1, \ldots, X_k) \big] - \mathrm{U}_{2k}^n \big[\kappa_{2k}(X_1, \ldots, X_{2k}) \big] \Big] \\
&\qquad - ({}^{n}\mathrm{C}_k)^{-1} \Big[\mathbb{E}\big[\kappa_k(X_1, \ldots, X_k) \big] - \mathbb{E}\big[\kappa_{2k}(X_1, \ldots, X_{2k}) \big] \Big] \Bigg| \\
&\leq \Big| ({}^{n}\mathrm{C}_k)^{-1} \mathrm{U}_k^n \big[\kappa_k(X_1, \ldots, X_k) - \mathbb{E}\left[\kappa_k(X_1, \ldots, X_k) \right] \big] \Big| \\
&\qquad + \Big| ({}^{n}\mathrm{C}_k)^{-1} \mathrm{U}_{2k}^n \big[\kappa_{2k}(X_1, \ldots, X_{2k}) - \mathbb{E}\left[\kappa_{2k}(X_1, \ldots, X_{2k}) \right] \big] \Big| \\
&=: \spadesuit.
\end{aligned}$$

Now, using Theorem A.4 of the Supplement for

$$\kappa_k(X_1, \ldots, X_k) - \mathbb{E}\left[\kappa_k(X_1, \ldots, X_k) \right] \text{ and } \kappa_{2k}(X_1, \ldots, X_{2k}) - \mathbb{E}\left[\kappa_{2k}(X_1, \ldots, X_{2k}) \right],$$

we obtain that with probability at least $1 - 2e^{-\tau}$,

$$\spadesuit \leq ({}^{n}\mathrm{C}_{k})^{-1}\left[4(\sigma_1+\sqrt{2}\sigma_2)\sqrt{k}\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}} + 4(\theta_1+2\theta_2)k\left(\frac{1+\tau}{n}\right)\right]$$
$$\leq c_1\left(\frac{1+\tau}{n}\right)^{\frac{2k+1}{2}} + c_2\left(\frac{1+\tau}{n}\right)^{\frac{2k+2}{2}}, \tag{38}$$

where c_1 and c_2 are positive constants that do not depend on τ and n. Now applying Theorem 7 with $a = k$ (see (37)) and $b = 2k+1$ (see (38)) and noting that $\min\{3a, b\} = \min\{3k, 2k+1\} = 2k+1$, $\min\{2b, (a+b), 4a\} = \min\{4k+2, 3k+1, 4k\} = 3k+1$, yields the result.

5.3. Proof of Theorem 4

Applying Theorem A.4 of the Supplement on $r(X_1,\ldots,X_k) - \mathbb{E}(r(X_1,\ldots,X_k))$, yields that with probability at least $1 - \exp(-\tau)$,

$$\|\hat{C} - C\|_{\mathcal{H}} \leq 4\sigma\sqrt{k}\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}} + 4\theta k\left(\frac{1+\tau}{n}\right) \leq c_1\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}},$$

where the second inequality holds for $n \geq \tau + 1$. Hence, it follows from Lemma A.1 of the Supplement that

$$\Delta = \mathbb{E}\|\hat{C} - C\|_{\mathcal{H}}^2 \leq \frac{e_1}{n}, \tag{39}$$

for some positive constant e_1. Therefore,

$$\begin{aligned}
&|\Delta - \hat{\Delta}_{\mathrm{degen}}| \\
&= \left|\Delta - ({}^{n}\mathrm{C}_{k})^{-1}\left[\mathrm{U}_k^n\left[\kappa_k(X_1,\ldots,X_k)\right] - \mathrm{U}_{2k}^n\left[\kappa_{2k}(X_1,\ldots,X_{2k})\right]\right]\right| \\
&= \Bigg|\Delta - ({}^{n}\mathrm{C}_{k})^{-1}\sigma_k^2 + ({}^{n}\mathrm{C}_{k})^{-1}\sigma_k^2 \\
&\qquad - ({}^{n}\mathrm{C}_{k})^{-1}\left[\mathrm{U}_k^n\left[\kappa_k(X_1,\ldots,X_k)\right] - \mathrm{U}_{2k}^n\left[\kappa_{2k}(X_1,\ldots,X_{2k})\right]\right]\Bigg| \\
&\leq \left|\Delta - ({}^{n}\mathrm{C}_{k})^{-1}\sigma_k^2\right| + ({}^{n}\mathrm{C}_{k})^{-1}\left|\mathrm{U}_k^n\left[\kappa_k(X_1,\ldots,X_k) - \mathbb{E}[\kappa_k(X_1,\ldots,X_k)]\right]\right| \\
&\qquad + ({}^{n}\mathrm{C}_{k})^{-1}\left|\mathrm{U}_{2k}^n\left[\kappa_{2k}(X_1,\ldots,X_{2k}) - \mathbb{E}[\kappa_{2k}(X_1,\ldots,X_{2k})]\right]\right| \\
&=: \spadesuit.
\end{aligned}$$

Now, using Theorem A.4 of the Supplement on

$$\kappa_k(X_1,\ldots,X_k) - \mathbb{E}\left[\kappa_k(X_1,\ldots,X_k)\right], \text{ and } \kappa_{2k}(X_1,\ldots,X_{2k}) - \mathbb{E}\left[\kappa_{2k}(X_1,\ldots,X_{2k})\right],$$

we obtain that with probability at least $1 - 2e^{-\tau}$,

$$\begin{aligned}
\spadesuit &\le \left|\Delta - ({}^n\mathrm{C}_k)^{-1}\sigma_k^2\right| + ({}^n\mathrm{C}_k)^{-1}\left[4\sigma_1\sqrt{k}\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}} + 4\theta_1 k\left(\frac{1+\tau}{n}\right)\right] \\
&\quad + ({}^n\mathrm{C}_k)^{-1}\left[4\sigma_2\sqrt{2k}\left(\frac{1+\tau}{n}\right)^{\frac{1}{2}} + 8\theta_2 k\left(\frac{1+\tau}{n}\right)\right] \\
&\le \left|\Delta - ({}^n\mathrm{C}_k)^{-1}\sigma_k^2\right| + c'({}^n\mathrm{C}_k)^{-1}\left(\frac{1+\tau}{n}\right)^{1/2} \\
&\le \left|\Delta - ({}^n\mathrm{C}_k)^{-1}\sigma_k^2\right| + c'\left(\frac{k}{n}\right)^k\left(\frac{1+\tau}{n}\right)^{1/2} \\
&\le \begin{cases} c'\left(\frac{1}{n}\right)\left(\frac{1+\tau}{n}\right)^{1/2}, & k = 1 \\ \max\{\Delta, ({}^n\mathrm{C}_k)^{-1}\sigma_k^2\} + c'\left(\frac{k}{n}\right)^k\left(\frac{1+\tau}{n}\right)^{1/2}, & k > 1 \end{cases} \\
&\le \begin{cases} c'\left(\frac{1+\tau}{n}\right)^{3/2}, & k = 1 \\ c''\left(\frac{1+\tau}{n}\right), & k > 1 \end{cases},
\end{aligned}$$

where $c', c'' > 0$ are constants that do not depend on τ and n, and we used (39) in the above inequality. Now applying Theorem 7 with $a = 1$ and $b = 2$ (for $k > 1$) yields the result.

5.4. Proof of Theorem 5

Define $\hat{\alpha} := \frac{\frac{S^2}{n}}{\frac{S^2}{n} + \|\bar{X}\|_2^2}$ so that $\breve{\mu} = (1 - \hat{\alpha})\bar{X}$. Define $W := \bar{X} \sim N(\mu, \frac{\sigma^2}{n}I)$ and $U := \frac{S^2}{n}$. Consider

$$\begin{aligned}
\mathbb{E}\|\hat{\mu} - \mu\|_2^2 - \mathbb{E}\|\breve{\mu} - \mu\|_2^2 &= \mathbb{E}\left[\|\bar{X} - \mu\|_2^2 - \|(\bar{X} - \mu) - \hat{\alpha}\bar{X}\|_2^2\right] = \mathbb{E}\left[2\hat{\alpha}\left\langle \bar{X} - \mu, \bar{X}\right\rangle_2 - \|\hat{\alpha}\bar{X}\|_2^2\right] \\
&= \mathbb{E}\left[2\left\langle W - \mu, \frac{UW}{U + \|W\|_2^2}\right\rangle_2 - \left\|\frac{UW}{U + \|W\|_2^2}\right\|_2^2\right]. \qquad (40)
\end{aligned}$$

Note that

$$\left\langle W - \mu, \frac{UW}{U + \|W\|_2^2}\right\rangle_2 = \sum_{i=1}^d (W_i - \mu_i)\left(\frac{UW_i}{U + \sum_i W_i^2}\right)$$

where $W_i \sim N(\mu_i, \sigma^2/n)$. By partial integration, we have

$$\mathbb{E}\left[(W_i - \mu_i)\left(\frac{UW_i}{U + \sum_i W_i^2}\right)\right] = \frac{\sigma^2}{n}\mathbb{E}\left[\frac{d}{dW_i}\left(\frac{UW_i}{U + \sum_i W_i^2}\right)\right]$$

$$= \frac{\sigma^2}{n}\mathbb{E}\left[\frac{U}{U+\|W\|_2^2} - \frac{2UW_i^2}{(U+\|W\|_2^2)^2}\right]$$

and therefore

$$\mathbb{E}\left[(W-\mu)^T\left(\frac{UW}{U+\|W\|_2^2}\right)\right] = \frac{\sigma^2}{n}\mathbb{E}\left[\frac{dU}{U+\|W\|_2^2} - \frac{2U\|W\|_2^2}{(U+\|W\|_2^2)^2}\right]. \tag{41}$$

Using (41) in (40), we have

$$\mathbb{E}\|\hat{\mu}-\mu\|_2^2 - \mathbb{E}\|\breve{\mu}-\mu\|_2^2 = \mathbb{E}\left[\frac{2d\sigma^2U}{n(U+\|W\|_2^2)} - \frac{4\sigma^2U\|W\|_2^2}{n(U+\|W\|_2^2)^2} - \frac{U^2\|W\|_2^2}{(U+\|W\|_2^2)^2}\right]. \tag{42}$$

Note that $\|W\|_2^2 \sim \frac{\sigma^2}{n}\chi_d^2(\lambda)$ where $\lambda = \frac{n\|\mu\|_2^2}{\sigma^2}$ with $\chi_d^2(\lambda)$ denoting a non-central χ^2 distribution with d degrees of freedom and λ being the non-centrality parameter. Also note that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)d}$ with S^2 being independent of W. Define $Z := \frac{n\|W\|_2^2}{\sigma^2} \sim \chi_d^2(\lambda)$ and $Y := \frac{n(n-1)U}{\sigma^2} \sim \chi^2_{(n-1)d}$ where Y and Z are independent. Then (42) reduces to

$$\begin{aligned}\mathbb{E}\|\hat{\mu}-\mu\|_2^2 - \mathbb{E}\|\breve{\mu}-\mu\|_2^2 &= \frac{\sigma^2}{n}\mathbb{E}\left[\frac{2dY}{Y+(n-1)Z} - \frac{4(n-1)YZ}{(Y+(n-1)Z)^2} - \frac{Y^2Z}{(Y+(n-1)Z)^2}\right] \\ &= \frac{\sigma^2}{n}\mathbb{E}\left[\frac{(2d-Z)Y^2 + YZ(n-1)(2d-4)}{(Y+(n-1)Z)^2}\right].\end{aligned} \tag{43}$$

Using the fact that $\int_0^\infty te^{-at}\,dt = \frac{1}{a^2}$ for $a > 0$ and employing Fubini's theorem, we have

$$\mathbb{E}\left[\frac{YZ}{(Y+(n-1)Z)^2}\right] = \int_0^\infty t\mathbb{E}\left[Ye^{-tY}\right]\mathbb{E}\left[Ze^{-t(n-1)Z}\right]\,dt \tag{44}$$

and

$$\mathbb{E}\left[\frac{Y^2(2d-Z)}{(Y+(n-1)Z)^2}\right] = \int_0^\infty t\mathbb{E}\left[Y^2e^{-tY}\right]\mathbb{E}\left[(2d-Z)e^{-t(n-1)Z}\right]\,dt. \tag{45}$$

To compute the above expectations, we require the following: for any $t > 0$,

- $\mathbb{E}\left[e^{-tY}\right] = (1+2t)^{-\frac{(n-1)d}{2}}$,
- $\mathbb{E}\left[Ye^{-tY}\right] = -\frac{d}{dt}\mathbb{E}\left[e^{-tY}\right] = \frac{(n-1)d}{1+2t}E\left[e^{-tY}\right]$,
- $$\begin{aligned}\mathbb{E}\left[Y^2e^{-tY}\right] = \frac{d^2}{dt^2}\mathbb{E}\left[e^{-tY}\right] &= \frac{(n-1)d}{1+2t}\left(\mathbb{E}\left[Ye^{-tY}\right] + \frac{2\mathbb{E}\left[e^{-tY}\right]}{1+2t}\right) \\ &= \frac{(n-1)^2d^2 + 2(n-1)d}{(1+2t)^2}\mathbb{E}\left[e^{-tY}\right],\end{aligned}$$
- $\mathbb{E}\left[e^{-t(n-1)Z}\right] = (1+2(n-1)t)^{-\frac{d}{2}}\exp\left(-\frac{\lambda(n-1)t}{1+2(n-1)t}\right)$,

•

$$\mathbb{E}\left[Ze^{-t(n-1)Z}\right] = -\frac{1}{n-1}\frac{d}{dt}\mathbb{E}\left[e^{-t(n-1)Z}\right]$$
$$= \left(\frac{d}{1+2(n-1)t} + \frac{\lambda}{(1+2(n-1)t)^2}\right)\mathbb{E}\left[e^{-t(n-1)Z}\right].$$

Therefore, (44) and (45) reduce to

$$\begin{aligned}
&\mathbb{E}\left[\frac{YZ}{(Y+(n-1)Z)^2}\right] \\
&= \int_0^\infty t\mathbb{E}\left[Ye^{-tY}\right]\mathbb{E}\left[Ze^{-t(n-1)Z}\right]\,dt \\
&= (n-1)d\int_0^\infty \frac{t}{1+2t}\left(\frac{d}{1+2(n-1)t} + \frac{\lambda}{(1+2(n-1)t)^2}\right)\mathbb{E}\left[e^{-tY}\right]\mathbb{E}\left[e^{-t(n-1)Z}\right]\,dt \\
&= d\int_0^\infty \frac{a}{n-1+2a}\left(\frac{d}{1+2a} + \frac{\lambda}{(1+2a)^2}\right)\mathbb{E}\left[e^{-\frac{aY}{n-1}}\right]\mathbb{E}\left[e^{-aZ}\right]\,da
\end{aligned} \tag{46}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\frac{Y^2(2d-Z)}{(Y+(n-1)Z)^2}\right] = \int_0^\infty t\mathbb{E}\left[Y^2e^{-tY}\right]\mathbb{E}\left[(2d-Z)e^{-t(n-1)Z}\right]\,dt \\
&= d(n-1)((n-1)d+2)\int_0^\infty \frac{t}{(1+2t)^2}\left(2d - \frac{d}{1+2(n-1)t} - \frac{\lambda}{(1+2(n-1)t)^2}\right) \\
&\qquad\qquad \times\mathbb{E}\left[e^{-tY}\right]\mathbb{E}\left[e^{-t(n-1)Z}\right]\,dt \\
&= d(n-1)((n-1)d+2)\int_0^\infty \frac{a}{(n-1+2a)^2}\left(2d - \frac{d}{1+2a} - \frac{\lambda}{(1+2a)^2}\right) \\
&\qquad\qquad \times\mathbb{E}\left[e^{-\frac{aY}{n-1}}\right]\mathbb{E}\left[e^{-aZ}\right]\,da.
\end{aligned} \tag{47}$$

Using (46) and (47) in (43), we obtain

$$\begin{aligned}
&\mathbb{E}\|\hat{\mu}-\mu\|_2^2 - \mathbb{E}\|\breve{\mu}-\mu\|_2^2 \\
&= \frac{d(n-1)\sigma^2}{n}\left[\int_0^\infty\left(\frac{((n-1)d+2)a}{(2a+n-1)^2}\left(2d - \frac{d}{1+2a} - \frac{\lambda}{(1+2a)^2}\right)\right.\right. \\
&\qquad \left.\left. + \frac{(2d-4)a}{2a+n-1}\left(\frac{d}{1+2a} + \frac{\lambda}{(1+2a)^2}\right)\right)\mathbb{E}\left[e^{-\frac{aY}{n-1}}\right]\mathbb{E}\left[e^{-aZ}\right]\,da\right] \\
&= \int_0^\infty \frac{d(n-1)\sigma^2 a}{n(2a+n-1)^2}\mathcal{B}(a,\lambda)(1+2a)^{-\frac{d}{2}-2}\mathbb{E}\left[e^{-\frac{aY}{n-1}}\right]\,da,
\end{aligned}$$

with

$$\mathcal{B}(a,\lambda) := \left((nd-d+2)\left(2d(1+2a)^2 - d(1+2a) - \lambda\right) + (2d-4)(2a+n-1)(d+2ad+\lambda)\right)e^{-\frac{a\lambda}{1+2a}}$$

$$= \left((nd-d+2)(d(8a^2+6a+1)-\lambda)+(2d-4)(2a+n-1)(d+2ad+\lambda)\right)e^{-\frac{a\lambda}{1+2a}}$$
$$=: (\theta_1+\theta_2\lambda)e^{-\frac{a\lambda}{1+2a}},$$

where for all $a \in [0,\infty)$,

$$\theta_1 := d(nd-d+2)(8a^2+6a+1)+d(2d-4)(2a+n-1)(1+2a) > 0 \text{ for } d \geq 2,$$

and

$$\theta_2 := (2d-4)(2a+n-1)-(n-1)d-2 = 4a(d-2)+(n-1)(d-4)-2 \geq 0$$

if $d \geq \sup_a \frac{8a+2+4(n-1)}{4a+n-1} = 4+\frac{2}{n-1}$. This means for $d \geq 4+\frac{2}{n-1}$, $n \geq 2$, $\mathcal{B}(a,\lambda) > 0$ for all λ and $a \in [0,\infty)$ and the result follows.

5.5. Proof of Theorem 6

Proceeding as in the proof of Theorem 5, we obtain

$$\mathbb{E}\|\hat{\mu}-\mu\|_2^2 - \mathbb{E}\|\check{\mu}_c-\mu\|_2^2 = \frac{c\sigma^2}{n}\mathbb{E}\left[\frac{(2d-cZ)Y^2+YZ(n-1)(2d-4)}{(Y+(n-1)Z)^2}\right]$$
$$= \int_0^\infty \frac{dc(n-1)\sigma^2 a}{n(2a+n-1)^2}\mathcal{A}(a,\lambda)(1+2a)^{-\frac{d}{2}-2}\mathbb{E}\left[e^{-\frac{aY}{n-1}}\right]\,da,$$

where

$$\mathcal{A}(a,\lambda) := \left((nd-d+2)\left(2d(1+2a)^2-cd(1+2a)-c\lambda\right)+(2d-4)(2a+n-1)(d+2ad+\lambda)\right)e^{-\frac{a\lambda}{1+2a}}$$
$$= \left((nd-d+2)(d(8a^2+8a-2ac+2-c)-c\lambda)+(2d-4)(2a+n-1)(d+2ad+\lambda)\right)e^{-\frac{a\lambda}{1+2a}}$$
$$=: (\theta_3+\theta_4\lambda)e^{-\frac{a\lambda}{1+2a}}$$

with

$$\theta_3 := d(nd-d+2)(8a^2+8a-2ac+2-c)+d(2d-4)(2a+n-1)(1+2a) > 0$$

for $d \geq 2$, $c \in [0,2)$ and all $a \in [0,\infty)$, and

$$\theta_4 := (2d-4)(2a+n-1)-(n-1)dc-2c = 4a(d-2)+(n-1)(2d-4-dc)-2c \geq 0$$

for $d \geq \frac{4}{2-c}+\frac{2c}{(n-1)(2-c)}$, $n \geq 2$, $c \in (0,2)$ and all $a \in [0,\infty)$. This means, that for the choice of n, c, and d in the statement of Theorem 6, the result follows.

Acknowledgements

The authors thank the editor, associate editor, and two reviewers for their detailed comments, which helped to fix some minor errors and improve the presentation. The authors particularly thank the reviewer who pointed out a mistake in the proof of Theorem A.5, and for providing detailed comments that led to the discussion in Remarks 1(v) and 2.

Funding

BKS thanks Donald Richards for helpful comments on the proof of Theorem 5. BKS is partially supported by the NSF DMS CAREER Award #1945396.

Supplementary Material

Supplement to "Shrinkage estimation of higher-order Bochner integrals" (DOI: 10.3150/23-BEJ1692SUPP; .pdf). Additional technical results are provided in the Supplement.

References

Arcones, M.A. and Giné, E. (1993). Limit theorems for U-processes. *Ann. Probab.* **21** 1494–1542. MR1235426

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437 https://doi.org/10.2307/1990404

Balasubramanian, K., Li, T. and Yuan, M. (2021). On the optimality of kernel-embedding based goodness-of-fit tests. *J. Mach. Learn. Res.* **22** Paper No. 1. MR4253694

Brandwein, A.C. and Strawderman, W.E. (1990). Stein estimation: The spherically symmetric case. *Statist. Sci.* **5** 356–369. MR1080957

Brandwein, A.C. and Strawderman, W.E. (2012). Stein estimation for spherically symmetric distributions: Recent developments. *Statist. Sci.* **27** 11–23. MR2953492 https://doi.org/10.1214/10-STS323

Chen, Y., Wiesel, A., Eldar, Y.C. and Hero, A.O. (2010). Shrinkage algorithms for MMSE covariance estimation. *IEEE Trans. Signal Process.* **58** 5016–5029. MR2722661 https://doi.org/10.1109/TSP.2010.2053029

de la Peña, V.H. and Giné, E. (2012). *Decoupling: From Dependence to Independence*. New York: Springer Science & Business Media. MR1666908 https://doi.org/10.1007/978-1-4612-0537-1

Dinculeanu, N. (2000). *Vector Integration and Stochastic Integration in Banach Spaces*. *Pure and Applied Mathematics (New York)*. New York: Wiley Interscience. MR1782432 https://doi.org/10.1002/9781118033012

Fisher, T.J. and Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Comput. Statist. Data Anal.* **55** 1909–1918. MR2765053 https://doi.org/10.1016/j.csda.2010.12.006

Fukumizu, K., Bach, F.R. and Jordan, M.I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **5** 73–99. MR2247974 https://doi.org/10.1162/153244303768966111

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B. and Smola, A. (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* (J. Platt, D. Koller, Y. Singer and S. Roweis, eds.) **20**. Curran Associates.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

James, W. and Stein, C. (1960). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Berkeley-Los Angeles, CA: Univ. California Press. MR0133191

Joly, E. and Lugosi, G. (2016). Robust estimation of U-statistics. *Stochastic Process. Appl.* **126** 3760–3773. MR3565476 https://doi.org/10.1016/j.spa.2016.04.021

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 https://doi.org/10.1016/S0047-259X(03)00096-4

Ledoit, O. and Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli* **24** 3791–3832. MR3788189 https://doi.org/10.3150/17-BEJ979

Lee, A.J. (2019). *U-Statistics: Theory and Practice*. Routledge.

Muandet, K., Sriperumbudur, B., Fukumizu, K., Gretton, A. and Schölkopf, B. (2016). Kernel mean shrinkage estimators. *J. Mach. Learn. Res.* **17** Paper No. 48. MR3504608

Song, L., Smola, A., Gretton, A., Bedo, J. and Borgwardt, K. (2012). Feature selection via dependence maximization. *J. Mach. Learn. Res.* **13** 1393–1434. MR2930643

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I* 197–206. Berkeley-Los Angeles, CA: Univ. California Press. MR0084922

Stein, C. (1975). Estimation of a covariance matrix. Rietz Lecture, 39th Annual Meeting, Atlanta, GA.

Touloumis, A. (2015). Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Comput. Statist. Data Anal.* **83** 251–261. MR3281809 https://doi.org/10.1016/j.csda.2014.10.018

Utpala, S. and Sriperumbudur, B.K. (2024). Supplement to "Shrinkage estimation of higher-order Bochner integrals." https://doi.org/10.3150/23-BEJ1692SUPP

Zhou, Y., Chen, D.-R. and Huang, W. (2019). A class of optimal estimators for the covariance operator in reproducing kernel Hilbert spaces. *J. Multivariate Anal.* **169** 166–178. MR3875593 https://doi.org/10.1016/j.jmva.2018.09.003

Received July 2022 and revised October 2023