Multivariate Central Limit Theorems for Random Clique Complexes
Tadas Temčinas, Vidit Nanda, and Gesine Reinert

Abstract. Motivated by open problems in applied and computational algebraic topology, we establish multivariate normal approximation theorems for three random vectors which arise organically in the study of random clique complexes. These are:
(1) the vector of critical simplex counts attained by a lexicographical Morse matching,
(2) the vector of simplex counts in the link of a fixed simplex, and
(3) the vector of total simplex counts.
The first of these random vectors forms a cornerstone of modern homology algorithms, while the second one provides a natural generalisation for the notion of vertex degree, and the third one may be viewed from the perspective of U-statistics. To obtain distributional approximations for these random vectors, we extend the notion of dissociated sums to a multivariate setting and prove a new central limit theorem for such sums using Stein’s method.

Keywords: Stein’s method, multivariate normal approximation, discrete Morse theory, random graphs, random simplicial complexes

MSC: 60F05 Central limit and other weak theorems, 60D05 Geometric probability, stochastic geometry, random sets, 05C80 Random graphs (graph-theoretic aspects).

1. Introduction

Methods from applied and computational algebraic topology have recently found substantial applications in the analysis of nonlinear and unstructured datasets [21, 10]. The modus operandi of topological data analysis is to first build a nested family of simplicial complexes around the elements of a dataset, and to then compute the associated persistent homology barcodes [14]. Of central interest, when testing hypotheses under this paradigm, is the question of what homology groups to expect when the input data are randomly generated. Significant efforts have therefore been devoted to answering this question for various models of noise, giving rise to the field of stochastic topology [26, 8, 25, 1, 13]. Our work here is a contribution to this area at the interface between probability theory and algebraic topology.

A cornerstone for statistical inference, beyond providing expectations, are distributional approximations. This paper establishes the first multivariate normal approximations to three important counting problems in stochastic topology; as these approximations are based on Stein’s method, explicit bounds on the approximation errors are provided. Our starting point is the ubiquitous graph model \( G(n, p) \); a graph \( G \) chosen from this model has as its vertex set \( [n] = \{1, 2, \ldots, n\} \), and each of its possible \( \binom{n}{2} \) edges is included independently with probability \( p \in [0, 1] \). It was established by Erdős and Rényi in [16] that \( p = \log(n) / n \) is a sharp threshold for connectivity in \( G(n, p) \), in the sense that the following assertions hold for any random graph \( G \sim G(n, p) \) and every arbitrarily small \( \epsilon > 0 \): if \( p \) exceeds \((1 + \epsilon) \cdot \log(n) / n\), then \( G \) is connected with high probability. Conversely, if \( p \) is smaller than \((1 - \epsilon) \cdot \log(n) / n\), then \( G \) is disconnected with high probability.
A natural higher-order generalisation of $G(n, p)$ is furnished by the random clique complex model $X(n, p)$, whose constituent complexes $\mathcal{L}$ are constructed as follows. One first selects an underlying graph $G \sim G(n, p)$, and then deterministically fills out all $k$-cliques in $G$ with $(k - 1)$-dimensional simplices for $k \geq 3$. Higher connectivity is now measured by the Betti numbers $\beta_k(\mathcal{L})$, which are ranks of rational homology groups $H_k(\mathcal{L}; \mathbb{Q})$ — in particular, $\beta_0(\mathcal{L})$ equals the number of connected components of the underlying random graph $G$. In [27], Kahle proved the following far-reaching generalisation of the Erdős-Rényi connectivity result: for each $k \geq 1$ and $\epsilon > 0$,

\begin{align*}
(1) \text{ if } & \quad p \geq \left[ \left( \frac{k}{2} + 1 + \epsilon \right) \cdot \frac{\log(n)}{n} \right]^{1/(k+1)}, \\
& \text{then } \beta_k(\mathcal{L}) = 0 \text{ with high probability; and moreover,} \\
(2) \text{ if } & \quad \left[ \frac{k + 1 + \epsilon}{n} \right]^{1/k} \leq p \leq \left[ \left( \frac{k}{2} + 1 - \epsilon \right) \cdot \frac{\log(n)}{n} \right]^{1/(k+1)}, \\
& \text{then } \beta_k(\mathcal{L}) \neq 0 \text{ with high probability.}
\end{align*}

With this result in mind, we motivate and describe three random vectors pertaining to $\mathcal{L} \sim X(n, p)$; the normal approximation of these three random vectors will be our focus in this paper. All three are denoted $T = (T_1, \ldots, T_d)$ for an integer $d > 0$.

**Random Vector 1: Critical Simplex Counts.** The computation of Betti numbers $\beta_k(\mathcal{L})$ begins with the chain complex

$$
\cdots \xrightarrow{d_{k+1}} C_k \xrightarrow{d_k} C_{k-1} \xrightarrow{d_{k-1}} \cdots \xrightarrow{d_2} C_1 \xrightarrow{d_1} C_0.
$$

Here $C_k$ is a vector space whose dimension equals the number of $k$-simplices in $\mathcal{L}$, while $d_k : C_k \to C_{k-1}$ is an incidence matrix encoding which $(k - 1)$-simplices lie in the boundary of a given $k$-simplex. These matrices satisfy the property that every successive composite $d_{k+1} \circ d_k$ equals zero, and $\beta_k(\mathcal{L})$ is the dimension of the quotient vector space $\ker d_k / \text{im} d_{k+1}$. Thus, one is required to diagonalise the matrices $\{d_k : C_k \to C_{k-1}\}$ via row and column operations, which is a straightforward task in principle. Unfortunately, Gaussian elimination on an $m \times m$ matrix incurs an $O(m^3)$ cost, which becomes prohibitive when facing simplicial complexes built around large data sets [38]. The standard remedy is to construct a much smaller chain complex which has the same homology groups, and by far the most fruitful mechanism for achieving such homology-preserving reductions is **discrete Morse theory** [18, 36, 22, 31].

The key structure here is that of an *acyclic partial matching*, which pairs together certain adjacent simplices of $\mathcal{L}$; and the homology groups of $\mathcal{L}$ may be recovered from a chain complex whose vector spaces are spanned by unpaired, or *critical*, simplices. One naturally seeks an optimal acyclic partial matching on $\mathcal{L}$ which admits the fewest possible critical simplices. Unfortunately, the optimal matching problem is computationally intractable to solve [24] even approximately [5] for large $\mathcal{L}$. Our third random vector is obtained by letting $T_k$ equal the number of critical $k$-simplices for a specific type of acyclic partial matching on $\mathcal{L}$, called the *lexicographical* matching. Knowledge of this random vector serves to simultaneously quantify the benefit of using discrete Morse theoretic reductions on random simplicial complexes and to provide a robust null model by which to measure their efficacy on general (i.e., not necessarily random) simplicial complexes.
Random Vector 2: Link Simplex Counts. The link of a simplex $t$ in $\mathcal{L}$, denoted $L_k(t)$, consists of all simplices $s$ for which the union $s \cup t$ is also a simplex in $\mathcal{L}$ and the intersection $s \cap t$ is empty. The link of $t$ forms a simplicial complex in its own right; and if we restrict attention to the underlying random graph $G$, then the link of a vertex is precisely the collection of its neighbours. Therefore, the Betti numbers $\beta_k(L_k(t))$ generalise the degree distribution for vertices of random graphs in two different ways — one can study neighbourhoods of higher-dimensional simplices by increasing the dimension of $t$, and one can examine higher-order connectivity properties by increasing the homological dimension $k$. The second random vector of interest to us here is obtained by letting $T_k$ equal the number of $k$-simplices lying in the link of a fixed simplex $t$ in $\mathcal{L}$, given that $t$ indeed is a simplex in the random complex. As far as we are aware, ours is the first work that studies this random vector. A different conditional distribution, which follows directly from results on subgraph counts in $G(n, p)$, has been studied before, see Remark 5.1.

There are compelling reasons to better understand the combinatorics and topology of such links from a probabilistic viewpoint. For instance, the fact that the link of a $k$-simplex in a triangulated $n$-manifold is always a triangulated sphere of dimension $(n - k - 1)$ has been exploited to produce canonical stratifications of simplicial complexes into homology manifolds [2, 37]. Knowledge of simplex counts (and hence, Betti numbers) of links would therefore form an essential first step in any systematic study involving canonical stratifications of random clique complexes.

Random Vector 3: Total Simplex Counts. The strategy employed in Kahle’s proof of the second assertion above involves first checking that the expected number of $k$-simplices in $\mathcal{L} \sim X(n, p)$ is much larger than the expected number of simplices of dimensions $k \pm 1$ whenever $p$ lies in the range indicated by (2). Therefore, one may combine the Morse inequalities with the linearity of expectation in order to guarantee that the expected $\beta_k(\mathcal{L})$ is nonzero — see [27, Section 4] for details. To facilitate more refined analysis and estimates of this sort, the first random vector we study in this paper is obtained by letting $T_k$ equal the total number of $k$-dimensional simplices in $\mathcal{L}$.

Since $T_k$ is precisely the number of $(k + 1)$-cliques in $G \sim G(n, p)$, this random vector falls within the purview of generalised $U$-statistics. We extend results from [23] to show not only distributional convergence asymptotically but a stronger result, detailing explicit non-asymptotic bounds on the approximation. Several interesting problems can be seen as special cases — these include classical $U$-statistics [32, 30], monochromatic subgraph counts of inhomogeneous random graphs with independent random vertex colours, and the number of overlapping patterns in a sequence of independent Bernoulli trials. To the best of our knowledge, this is the first multivariate normal approximation result with explicit bounds where the sizes of the subgraphs are permitted to increase with $n$.

Main Results. The central contributions of this work are multivariate normal approximations for all three random vectors $T$ described above. For the purposes of these introductory remarks, we will restrict attention to the case where $T$ is the vector of critical simplex counts. Letting $\{Y_{i,j}\}_{1 \leq i < j \leq n}$ be a sequence of i.i.d. Bernoulli variables, the $k$-th component is

$$T_k = \sum_{s \subseteq [n]} \prod_{|s| = k} Y_{i,j} \left[ \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s} Y_{i,j} \right) - \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s} Y_{i,j} \right) \right].$$

This variable, which we discuss in Section 4, arises naturally in stochastic topology [9, 8] but has been poorly studied from a distributional approximation perspective. To the best of our knowledge, only the expected value of a closely-related random variable has been calculated (see [6, Section 8]). While there is no shortage of multivariate normal approximation theorems [17, 40, 35, 12], the existing ones are not sufficiently fine-grained for our purposes. We therefore return to the pioneering work of Barbour, Karoński, and Ruciński [4], who proved a univariate central limit theorem (CLT) for a decomposable
sum of random variables using Stein’s method, treating the case of dissociated sums as a special case. Our approximation result, described below, forms a new extension of their ideas to the multivariate setting, and may be of independent interest.

Let $n$ and $d$ be positive integers. For each $i \in [d] := \{1, 2, \ldots, d\}$, we fix an index set $I_i \subset [n] \times \{i\}$ and consider the union of disjoint sets $I := \bigcup_{i \in [d]} I_i$. Associate to each such $s = (k, i) \in I$ a real centered random variable $X_s$ and form for each $i \in [d]$ the sum

$$W_i := \sum_{s \in I_i} X_s.$$ 

Consider the resulting random vector $W = (W_1, \ldots, W_d) \in \mathbb{R}^d$. The following notion is a natural multivariate generalisation of the dissociated sum from [34]; see also [4].

**Definition 1.1.** We call $W$ a vector of dissociated sums if for each $s \in I$ and $j \in [d]$ there exists a dependency neighbourhood $\mathcal{D}_j(s) \subset I_j$ satisfying three criteria:

1. The difference $\left(W_j - \sum_{u \in \mathcal{D}_j(s)} X_u\right)$ is independent of $X_s$.
2. For each $t \in I$, the quantity $\left(W_j - \sum_{u \in \mathcal{D}_j(s)} X_u - \sum_{v \in \mathcal{D}_j(t) \setminus \mathcal{D}_j(s)} X_v\right)$ is independent of the pair $(X_s, X_t)$; and finally,
3. $X_s$ and $X_t$ are independent if $t \notin \bigcup_j \mathcal{D}_j(s)$.

Let $W$ be a vector of dissociated sums as defined above. For each $s \in I$, by construction, the sets $\mathcal{D}_j(s), j \in [d]$ are disjoint (although for $s \neq t$, the sets $\mathcal{D}_j(s)$ and $\mathcal{D}_j(t)$ may not be disjoint). We write $\mathcal{D}(s) = \bigcup_{j \in [d]} \mathcal{D}_j(s)$ for the disjoint union of these of dependency neighbourhoods. With this preamble in place, we state our main result.

**Theorem 1.2.** Let $h : \mathbb{R}^d \to \mathbb{R}$ be any three times continuously differentiable function whose third partial derivatives are Lipschitz continuous and bounded. Consider a standard $d$-dimensional Gaussian vector $Z \sim \text{MVN}(0, \text{Id}_{d \times d})$. Assume that for all $s \in I$, we have $\mathbb{E}\left\{X_s\right\} = 0$ and $\mathbb{E}\left|X_s^3\right| < \infty$. Then, for any vector of dissociated sums $W \in \mathbb{R}^d$ with a positive semi-definite covariance matrix $\Sigma$,

$$\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{1/2}Z)\right| \leq B_{1.2} \sup_{i,j,k \in [d]} \left\| \frac{\partial^3 h}{\partial x_i \partial x_j \partial x_k} \right\|_\infty,$$

where $B_{1.2} = B_{1.2.1} + B_{1.2.2}$ is the sum given by

$$B_{1.2.1} := \frac{1}{3} \sum_{s \in I} \sum_{t,u \in \mathcal{D}(s)} \left( \frac{1}{2} \mathbb{E}\left|X_s X_t X_u\right| + \mathbb{E}\left|X_s X_t\right| \mathbb{E}\left|X_u\right| \right),$$

$$B_{1.2.2} := \frac{1}{3} \sum_{s \in I} \sum_{t \in \mathcal{D}(s)} \sum_{v \in \mathcal{D}(t) \setminus \mathcal{D}(s)} \left( \mathbb{E}\left|X_s X_t X_v\right| + \mathbb{E}\left|X_s X_t\right| \mathbb{E}\left|X_v\right| \right).$$

In the special case where each $W_k$ is a sum of an equal number of i.i.d. random variables and each i.i.d. sequence is independent, the bound in Theorem 1.2 is optimal with respect to the size $n$ of the sum in each component. However, compared to the CLT from [17], the bound is not optimal in the length $d$ of the vector $W$. In any event, the desired CLT for critical simplex counts follows as a corollary to Theorem 1.2. We state a simplified version of this result here and note that the full statement and proof have been recorded as Theorem 4.6 below. In the statement below, $W \in \mathbb{R}^d$ is an appropriately scaled and centered vector whose $k$-th component counts the number of critical simplices of dimension $k$ for the lexicographical acyclic partial matching on $X \sim \text{X}(n, p)$.

**Theorem 1.3.** Let $Z \sim \text{MVN}(0, \text{Id}_{d \times d})$ and $\Sigma$ be the covariance matrix of $W$. Let $h : \mathbb{R}^d \to \mathbb{R}$ be three times partially differentiable function whose third partial derivatives are Lipschitz continuous and bounded.
Then there is a constant $B_{1.3} > 0$ independent of $n$ and a natural number $N_{1.3}$ such that for any $n \geq N_{1.3}$ we have

$$
|E_h(W) - E_h(\Sigma^1 Z)| \leq B_{1.3} \sup_{i,j,k \in [d]} \left\| \frac{\partial^3 h}{\partial x_i \partial x_j \partial x_k} \right\| \infty n^{-1}.
$$

En route to proving Theorem 1.3, we also establish the following properties, which are of direct interest in computational topology. Here we assume that $p \in (0, 1)$ and $k \in \{1, 2, \ldots \}$ are constants.

1. The expected number of critical $k$-simplices is one order of $n$ smaller that the expected number of total $k$-simplices; see Lemma 4.2.
2. The variance of the number of critical $k$-simplices is at least of the order $n^{2k}$, as shown in Lemma 4.4. An upper bound of the same order can be proved similarly. The variance of the total number of $k$-simplices is also of the same order.
3. Knowing the expected value and the variance one can prove concentration results using different concentration inequalities, for example, Chebyshev’s inequality. This would show that not only the expected value of critical simplices is smaller compared to all simplices but also that large deviations from the mean are unlikely, hence implying that the substantial improvement of one order of $n$ is not only expected but also likely.
4. For counting critical simplices to high accuracy in probability, it is not necessary to check every simplex. Certain simplices have a very small chance of being critical, and can be safely ignored. The probability of this omission causing an error is vanishingly small asymptotically; see Proposition 4.5.

More details are provided in Remark 4.7.

Related Work. Theorem 1.2 is not the first generalisation of the results in [4] to a multivariate setting, see for example [17, 40]. The key advantage of our approach is that it allows for bounds which are non-uniform in each component of the vector $W$. This is useful when, for example, the number of summands in each component are of different order or when the sizes of dependency neighbourhoods in each component are of different order. The applications considered here are precisely of this type, where the non-uniformity of the bounds is crucial. Moreover, we do not require the covariance matrix $\Sigma$ to be invertible, and can therefore accommodate degenerate multivariate normal distributions.

Another multivariate central limit theorem for centered subgraph counts in the more general setting of a random graph associated to a graphon can be found in [29]. That proof is based on Stein’s method via a Stein coupling. Translating this result for uncentered subgraph counts would yield an approximation by a function of a multivariate normal. In [41], an exchangeable pair coupling led to [41, Proposition 2] which can be specialised to joint counts of edges and triangles; our approximation significantly generalises this result beyond the case where $k \in \{1, 2\}$. Several univariate normal approximation theorems for subgraph counts are available; recent developments in this area include [39], which uses Malliavin calculus together with Stein’s method, and [15], which uses the Stein-Tikhomirov method.

Organisation. In Section 2 we prove our main approximation theorem using smooth test functions and extend the result to non-smooth test functions using a smoothing technique from [19]. In Section 3, we recall concepts from the theory of simplicial complexes, which we later use. In Section 4 we prove an approximation theorem for critical simplex counts of lexicographical matchings. Two technical computations required in this Section have been consigned to the Appendix. In Section 5 we prove an approximation theorem for count variables of simplices that are in the link of a fixed simplex. In Section 6 we introduce a slight generalisation of generalised $U$-statistics for which Theorem 1.2 gives a CLT with explicit bounds. We then apply the CLT to simplex counts in the random clique complex.
**Acknowledgements.** TT acknowledges funding from EPSRC studentship 2275810. VN is supported by the EPSRC grant EP/R018472/1. GR is funded in part by the EPSRC grants EP/T018445/1 and EP/R018472/1. The authors would like to thank Xiao Fang and Matthew Kahle for helpful discussions.

## 2. A Multivariate CLT for Dissociated Sums

Throughout this paper we use the following notation. Given positive integers \( n, m \) we write \([m, n]\) for the set \( \{m, m+1, \ldots, n\} \) and \([n]\) for the set \([1, n]\). Given a set \( X \) we write \(|X|\) for its cardinality, \( \mathcal{P}(X) \) for its powerset, and given a positive integer \( k \) we write \( C_k = \{ t \in \mathcal{P}([n]) \mid |t| = k \} \) for the collection of subsets of \([n]\) which are of size \( k \). For a function \( f : \mathbb{R}^d \to \mathbb{R} \) we write \( \partial_i f = \frac{\partial f}{\partial x_i} \) and \( \partial_{ijk} f = \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k} \). Also, we write \( |f|_k = \sup_{i_1, \ldots, i_k} \| \partial_{i_1 \ldots i_k} f \|_\infty \) for any integer \( k \geq 1 \), as long as the quantities exist. Here \( \| \cdot \|_\infty \) denotes the supremum norm while \( \| \cdot \|_2 \) denotes the Euclidean norm.

The notation \( \nabla \) denotes the gradient operator in \( \mathbb{R}^d \). For a positive integer \( d \) we define a class of test functions \( h : \mathbb{R}^d \to \mathbb{R} \), as follows. We say \( h \in \mathcal{H}_d \) iff \( h \) is three times partially differentiable with third partial derivatives being Lipschitz and \( |h|_3 < \infty \). The notation \( \text{Id}_{d \times d} \) denotes the \( d \times d \) identity matrix. The vertex set of all graphs and simplicial complexes is assumed to be \([n]\). If \( s = (x, i) \in \mathbb{I} \) is an element of the index set in Definition 1.1, then we denote the second component of the tuple by \( |s| \), that is \( |s| := i \). We also use Bachmann-Landau asymptotic notation: we say \( f(n) = \Omega(g(n)) \) iff \( \lim \sup_{n \to \infty} \frac{|f(n)|}{g(n)} < \infty \) and \( f(n) = \Omega(g(n)) \) iff \( \lim \inf_{n \to \infty} \frac{|f(n)|}{g(n)} > 0 \).

Throughout this section, \( W \in \mathbb{R}^d \) is a vector of dissociated sums in the sense of Definition 1.1, with covariance matrix whose entries are \( \Sigma_{ij} = \text{Cov}(W_i, W_j) \) for \((i, j) \in [d]^2 \). For each \( s \in \mathbb{I} \) we denote by \( \mathbb{D}(s) \subset \mathbb{I} \) the disjoint union \( \bigcup_{j=1}^d \mathbb{D}_j(s) \). For each triple \((s, t, j) \in \mathbb{R}^2 \times [d] \) we write the set-difference \( \mathbb{D}_j(t) \setminus \mathbb{D}_j(s) \) as \( \mathbb{D}_j(t; s) \), with \( \mathbb{D}(t; s) \subset \mathbb{I} \) denoting the disjoint union of such differences over \( j \in [d] \).

### 2.1. Smooth Test Functions.

To prove Theorem 1.2 we use Stein’s method for multivariate normal distributions; for details see for example Chapter 12 in [12]. Our proof of Theorem 1.2 is based on the Stein characterization of the multivariate normal distribution: \( Z \in \mathbb{R}^d \) is a multivariate normal \( \text{MVN}(0, \Sigma) \) if and only if the identity

\[
\mathbb{E} \left\{ \nabla^T \Sigma \nabla f(Z) - Z^T \nabla f(Z) \right\} = 0
\]

holds for all twice continuously differentiable \( f : \mathbb{R}^d \to \mathbb{R} \) for which the expectation exists. In particular, we will use the following result based on [35, Lemma 1 and Lemma 2]. As Lemma 1 and Lemma 2 in [35] are stated there only for infinitely differentiable test functions, we give the proof here for completeness.

**Lemma 2.1 (Lemma 1 and Lemma 2 in [35]).** Fix \( n \geq 2 \). Let \( h : \mathbb{R}^d \to \mathbb{R} \) be \( n \) times continuously differentiable with \( n \)-th partial derivatives being Lipschitz and \( Z \sim \text{MVN}(0, \text{Id}_{d \times d}) \). Then, if \( \Sigma \in \mathbb{R}^{d \times d} \) is symmetric positive semidefinite, there exists a solution \( f : \mathbb{R}^d \to \mathbb{R} \) to the equation

\[
\nabla^T \Sigma \nabla f(w) - w^T \nabla f(w) = h(w) - \mathbb{E} h \left( \Sigma^{1/2} Z \right), \quad w \in \mathbb{R}^d,
\]

such that \( f \) is \( n \) times continuously differentiable and we have for every \( k = 1, \ldots, n \):

\[
|f|_k \leq \frac{1}{k} |h|_k.
\]

**Proof.** Let \( h \) be as in the assertion. It is shown in Lemma 2.1 in [11], which is based on a reformulation of Eq. (2.20) in [3], that a solution of (2.2) for \( h \) is given by \( f(x) = f_h(x) = \int_0^1 \frac{1}{t} \mathbb{E} \{ h(Z_{x,t}) \} \, dt, \)
with \( Z_{x,t} = \sqrt{t}x + \sqrt{1-t}\Sigma^{1/2}Z \). As \( h \) has \( n \)-th partial derivatives being Lipschitz and hence for differentiating \( f \) we can bring the derivative inside the integral, it is straightforward to see that the solution \( f \) is \( n \) times continuously differentiable.

The bound on \( |f|_k \) is a consequence of
\[
\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}(x) = \int_0^1 (2t)^{-1}t^{k/2}E \left\{ \frac{\partial^k h}{\partial x_{i_1} \cdots \partial x_{i_k}}(Z_{x,t}) \right\} dt
\]
for any \( i_1, i_2, \ldots, i_k \); see, for example, Equation (10) in [35]. Taking the sup-norm on both sides and bounding the right hand side of the equation gives
\[
\left\| \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}} \right\|_\infty \leq \left\| \frac{\partial^k h}{\partial x_{i_1} \cdots \partial x_{i_k}} \right\|_\infty \int_0^1 (2t)^{-1}t^{k/2}dt \leq \frac{1}{k} |h|_k.
\]

Note that neither Lemma 2.1 nor indeed Theorem 1.2 require the covariance matrix \( \Sigma \) to be invertible.

**Proof of Theorem 1.2.** To prove Theorem 1.2, we replace \( w \) by \( W \) in Equation (2.2) and take the expected value on both sides. As a result, we aim to bound the expression
\[
|E \left\{ \nabla^T \Sigma \nabla f(W) - W^T \nabla f(W) \right\} | = |E \left\{ \sum_{i,j=1}^d \partial_{ij} f(W) \Sigma_{ij} - \sum_{i=1}^d W_i \partial_i f(W) \right\} |
\]
where \( f \) is a solution to the Stein equation (2.2) for the test function \( h \). Since the variables \( \{X_s \mid s \in \mathcal{I}\} \) are centered and as \( X_t \) is independent of \( X_s \) if \( t \not\in D(s) \), for each \( (i,j) \in [d]^2 \) we have
\[
\Sigma_{ij} = \text{Cov}(W_j, W_j) = \sum_{s \in \mathcal{I}} \sum_{t \in D_j(s)} E \{X_s X_t\}.
\]

We now use the decomposition of \( \Sigma_{ij} \) from (2.4) in the expression (2.3). For each pair \( (s,j) \in \mathcal{I} \times [d] \) and \( t \in D(s) \) we set \( D_j(t; s) = D_j(t) \setminus D_j(s) \) and
\[
U_j^s := \sum_{u \in D_j(s)} X_u, \quad W_j := W_j - U_j^s, \quad \text{and} \quad V_j^{s,t} := \sum_{v \in D_j(t; s)} X_v, \quad W_j^{s,t} := W_j - V_j^{s,t}.
\]

By Definition 1.1, \( W_j^s \) is independent of \( X_s \), while \( W_j^{s,t} \) is independent of the pair \( (X_s, X_t) \).

Next we decompose the r.h.s. of (2.3);
\[
|E \left\{ \sum_{i=1}^d W_i \partial_i f(W) - \sum_{i,j=1}^d \partial_{ij} f(W) \Sigma_{ij} \right\} | = |R_1 + R_2 + R_3|;
\]
with
\[
R_1 = \sum_{i=1}^d E \{W_i \partial_i f(W)\} - \sum_{s \in \mathcal{I}} \sum_{j=1}^d E \{X_s U_j^s \partial_{sj} f(W^s)\},
\]
\[
R_2 = \sum_{s \in \mathcal{I}} \sum_{j=1}^d E \{X_s U_j^s \partial_{sj} f(W^s)\} - \sum_{s \in \mathcal{I}} \sum_{t \in D_j(s)} E \{X_s X_t\} E \partial_{|s|t} f(W^{s,t}), \text{ and}
\]
\[
R_3 = \sum_{s \in \mathcal{I}} \sum_{t \in D(s)} E \{X_s X_t\} (E \partial_{|s|t} f(W^{s,t}) - E \partial_{|s|t} f(W)) .
\]

Here we recall that if \( s = (k,i) \) then \( |s| = i \in [d] \).
As with the vector of dissociated sums $W \in \mathbb{R}^d$ itself, we can assemble these differences into random vectors. Thus, $W^s \in \mathbb{R}^d$ is $(W^s_1, \ldots, W^s_d)$, and similarly $W^{s,t} = (W^{s,t}_1, \ldots, W^{s,t}_d)$. In the next three claims, we provide bounds on $R$, for $t \in [3]$.

**Claim 2.2.** The absolute value of the expression $R_1$ from (2.6) is bounded above by

$$|R_1| \leq \left( \frac{1}{2} \sum_{s \in \mathcal{D}(s)} \sum_{a \in \mathcal{D}(s)} \mathbb{E}\left| X_s X_t X_u \right| \right) |f|_3.$$  

**Proof.** Note that

$$R_1 = \sum_{i=1}^{d} \sum_{s \in \mathcal{I}_i} \mathbb{E}\{ X_s \partial_i f(W) \} - \sum_{s \in \mathcal{I}_i} \sum_{j=1}^{d} \mathbb{E}\{ X_s U^j_i \partial_{ij} f(W^s) \}$$

$$= \sum_{i=1}^{d} \sum_{s \in \mathcal{I}_i} \left( \mathbb{E}\{ X_s \partial_i f(W) \} - \sum_{j=1}^{d} \mathbb{E}\{ X_s U^j_i \partial_{ij} f(W^s) \} \right).$$

For each $s \in \mathcal{I}_i$, it follows from (2.5) that $W = U^s + W^s$. Using the Lagrange form of the remainder term in Taylor’s theorem, we obtain

$$\partial_i f(W) = \sum_{j=1}^{d} \partial_{ij} f(W^s) U^s_j + \frac{1}{2} \sum_{j,k=1}^{d} \partial_{ijk} f(W^s + \theta_s U^s) U^s_j U^s_k$$

for some random $\theta_s \in (0,1)$. Using this Taylor expansion in the expression for $R_1$, we get the following four-term summand $S_{i,s}$ for each $i \in [d]$ and $s \in \mathcal{I}_i$:

$$S_{i,s} = \mathbb{E}\{ X_s \partial_i f(W^s) \} + \sum_{j=1}^{d} \mathbb{E}\{ X_s \partial_{ij} f(W^s) U^s_j \}$$

$$+ \frac{1}{2} \sum_{j,k=1}^{d} \mathbb{E}\{ X_s \partial_{ijk} f(W^s + \theta_s U^s) U^s_j U^s_k \} - \sum_{j=1}^{d} \mathbb{E}\{ X_s \partial_{ij} f(W^s) U^s_j \}. $$

The second and fourth terms cancel each other. Recalling that $X_s$ is centered by definition and independent of $W^s$ by Definition 1.1, the third term also vanishes and

$$R_1 = \sum_{i=1}^{d} \sum_{s \in \mathcal{I}_i} S_{i,s} = \frac{1}{2} \sum_{i=1}^{d} \sum_{s \in \mathcal{I}_i} \mathbb{E}\{ X_s \partial_{ij} f(W^s + \theta_s U^s) U^s_j U^s_k \}. $$

Recalling that $\| \partial_{ijk} f \|_\infty \leq \| f \|_3$ and that $U^s_j = \sum_{t \in \mathcal{D}(s)} X_t$, we have:

$$|R_1| \leq \frac{1}{2} \sum_{i=1}^{d} \sum_{s \in \mathcal{I}_i} \sum_{j,k=1}^{d} \mathbb{E}\left| X_s \partial_{ijk} f(W^s + \theta_s U^s) U^s_j U^s_k \right|$$

$$\leq \frac{|f|_3}{2} \sum_{i,j,k=1}^{d} \sum_{s \in \mathcal{I}_i} \sum_{t \in \mathcal{D}(s)} \mathbb{E}\left| X_s X_t X_u \right|$$

$$\leq \frac{|f|_3}{2} \sum_{s \in \mathcal{I}_i} \sum_{t \in \mathcal{D}(s)} \sum_{u \in \mathcal{D}(s)} \mathbb{E}\left| X_s X_t X_u \right|,$$

as desired. 

**Claim 2.3.** The absolute value of the expression $R_2$ from (2.7) is bounded above by

$$|R_2| \leq \left( \sum_{s \in \mathcal{I}_i} \sum_{t \in \mathcal{D}(s)} \sum_{u \in \mathcal{D}(t,s)} \mathbb{E}\left| X_s X_t X_u \right| \right) |f|_3.$$
PROOF. Recalling that $U_i^s = \sum_{t \in D_i(s)} X_t$ and $D(s) = \bigcup_{j=1}^d D_j(s)$,

$$R_2 = \sum_{s \in I} \sum_{t \in D(s)} \left\{ \mathbb{E} \left\{ X_s X_t \partial_{|s||t|} f(W^s) \right\} - \mathbb{E} \left\{ X_s X_t \right\} \mathbb{E} \left\{ \partial_{|s||t|} f(W^s) \right\} \right\}.$$

Fix $s \in I$ and $t \in D_j(s)$. Recall that by (2.5), $W^s = W^{s,t} + V^{s,t}$. Using the Lagrange form of the remainder term in Taylor’s theorem, we obtain:

$$\partial_{|s||t|} f(W^s) = \partial_{|s||t|} f(W^{s,t}) + \sum_{k=1}^d \partial_{|s||t||k|} f(W^{s,t} + \theta_{s,t} V^{s,t}) V_k^{s,t}$$

for some random $\theta_{s,t} \in (0,1)$. Using this Taylor expansion in the expression for $R_2$, we get the following three-term summand $S_{s,t}$ for each pair $(s, t) \in I \times D_j(s)$:

$$S_{s,t} = \mathbb{E} \left\{ X_s X_t \partial_{|s||t|} f(W^{s,t}) \right\} + \sum_{k=1}^d \mathbb{E} \left\{ X_s X_t \partial_{|s||t||k|} f(W^{s,t} + \theta_{s,t} V^{s,t}) V_k^{s,t} \right\} - \mathbb{E} \left\{ X_s X_t \right\} \mathbb{E} \left\{ \partial_{|s||t|} f(W^s) \right\}.$$

Recalling that $W^{s,t}$ is independent of the pair $(X_s, X_t)$ the first and the last terms cancel each other and only the sum over $k$ is left:

$$R_2 = \sum_{s \in I} \sum_{t \in D(s)} S_{s,t} = \sum_{s \in I} \sum_{t \in D(s)} \sum_{k=1}^d \mathbb{E} \left\{ X_s X_t \partial_{|s||t||k|} f(W^{s,t} + \theta_{s,t} V^{s,t}) V_k^{s,t} \right\}.$$

Recalling that $\|\partial_{ijk} f\|_\infty \leq |f|_3$ and that $V_k^{s,t} = \sum_{\nu \in D_k(t,s)} X_{\nu}$ we have:

$$|R_2| \leq \sum_{s \in I} \sum_{t \in D(s)} \sum_{k=1}^d \sum_{\nu \in D_k(t,s)} \mathbb{E} \left| X_s X_t X_{\nu} \partial_{|s||t||k|} f(W^{s,t} + \theta_{s,t} V^{s,t}) \right|$$

$$\leq |f|_3 \sum_{s \in I} \sum_{t \in D(s)} \sum_{u \in D(t,s)} \mathbb{E} \left| X_s X_t X_u \right|,$$

as required.

\[\square\]

CLAIM 2.4.

$$|R_3| \leq \left( \sum_{s \in I} \sum_{t \in D(s)} \left\{ \sum_{u \in D(s)} \mathbb{E} \left| X_s X_t \right| \mathbb{E} \left| X_u \right| + \sum_{u \in D(t,s)} \mathbb{E} \left| X_s X_t \right| \mathbb{E} \left| X_u \right| \right\} \right) |f|_3.$$ 

PROOF. Fix $(s, t) \in I \times D_j(s)$. Recall that by (2.5), $W^{s,t} = W - U^s - V^{s,t}$. Using the Lagrange form of the remainder term in Taylor’s theorem, we obtain:

$$\partial_{|s||t|} f(W^{s,t}) = \partial_{|s||t|} f(W) - \sum_{k=1}^d \partial_{|s||t||k|} f(W - \rho_{s,t} (U^s + V^{s,t})) (U_k^s + V_k^{s,t})$$

for some random $\rho_{s,t} \in (0,1)$. Recalling that $U_k^s = \sum_{t \in D_k(s)} X_t$ and $V_k^{s,t} = \sum_{\nu \in D_k(t,s)} X_{\nu}$,

$$R_3 = -\sum_{s \in I} \sum_{t \in D(s)} \sum_{k=1}^d \mathbb{E} \left\{ X_s X_t \right\} \mathbb{E} \left\{ \partial_{|s||t||k|} f(W - \rho_{s,t} (U^s + V^{s,t})) (U_k^s + V_k^{s,t}) \right\}$$

$$= -\sum_{s \in I} \sum_{t \in D(s)} \sum_{u \in D(s)} \mathbb{E} \left\{ X_s X_t \right\} \mathbb{E} \left\{ X_u \partial_{|s||t||k|} f(W - \rho_{s,t} (U^s + V^{s,t})) \right\}$$

$$-\sum_{s \in I} \sum_{t \in D(s)} \sum_{u \in D(t,s)} \mathbb{E} \left\{ X_s X_t \right\} \mathbb{E} \left\{ X_u \partial_{|s||t||k|} f(W - \rho_{s,t} (U^s + V^{s,t})) \right\}.$$
Recalling that \( \| \partial_{ijk} f \|_{\infty} \leq |f|_3 \) we bound:
\[
|R_3| \leq |f|_3 \sum_{s \in I} \sum_{t \in \mathcal{D}(s)} \sum_{u \in \mathcal{D}(s)} \mathbb{E} |X_s X_t| \mathbb{E} |X_u| + |f|_3 \sum_{s \in I} \sum_{t \in \mathcal{D}(s)} \sum_{u \in \mathcal{D}(t,s)} \mathbb{E} |X_s X_t| \mathbb{E} |X_u|,
\]
as required. \( \square \)

Take any \( h \in \mathcal{X}_d \). Let \( f : \mathbb{R}^d \to \mathbb{R} \) be the associated solution from Lemma 2.1. Combining Claims 2.1 - 2.4 and using Lemma 2.1 we have:
\[
\left| \mathbb{E} h(W) - \mathbb{E} (\Sigma^{1/2} Z) \right| \\
\leq \left| \mathbb{E} \left\{ \nabla^T \Sigma \nabla f(W) - W^T \nabla f(W) \right\} \right| \\
\leq |f|_3 \sum_{s \in I} \sum_{t \in \mathcal{D}(s)} \sum_{u \in \mathcal{D}(s)} \frac{1}{2} \mathbb{E} |X_s X_t X_u| + \mathbb{E} |X_s X_t| \mathbb{E} |X_u| \\
+ |f|_3 \sum_{s \in I} \sum_{t \in \mathcal{D}(s)} \sum_{u \in \mathcal{D}(t,s)} (\mathbb{E} |X_s X_t X_u| + \mathbb{E} |X_s X_t| \mathbb{E} |X_u|) \\
\leq \frac{1}{3} |h|_3 B_{1,2}.
\]
\( \square \)

In most of our applications, the variables \( X_s \) are centered and rescaled Bernoulli random variables. Hence, the following lemma is useful.

**Lemma 2.5.** Let \( \xi_1, \xi_2, \xi_3 \) be Bernoulli random variables with expected values \( \mu_1, \mu_2, \mu_3 \) respectively. Let \( c_1, c_2, c_3 > 0 \) be any constants. Consider variables \( X_i := c_i (\xi_i - \mu_i) \) for \( i = 1, 2, 3 \). Then we have
\[
\mathbb{E} |X_1 X_2 X_3| \leq c_1 c_2 c_3 \{ \mu_1 \mu_2 (1 - \mu_1)(1 - \mu_2) \}^{1/2} ; \\
\mathbb{E} |X_1 X_2| \mathbb{E} |X_3| \leq c_1 c_2 c_3 \{ \mu_1 \mu_2 (1 - \mu_1)(1 - \mu_2) \}^{1/2}.
\]

**Proof.** Note that \( X_3 \) can take two values: \( -c_3 \mu_3 \) or \( c_3 (1 - \mu_3) \). As \( 0 \leq \mu_3 \leq 1 \), we have
\[
\mathbb{E} |X_1 X_2| \mathbb{E} |X_3| \leq c_3 \mathbb{E} |X_1 X_2| ; \\
\mathbb{E} |X_1 X_2 X_3| \leq c_3 \mathbb{E} |X_1 X_2|.
\]
Applying the Cauchy-Schwarz inequality and direct calculation of the second moments gives
\[
\mathbb{E} |X_1 X_2| \leq \{ \mathbb{E} \{ X_1^2 \} \ \mathbb{E} \{ X_2^2 \} \}^{1/2} = c_1 c_2 \{ \mu_1 \mu_2 (1 - \mu_1)(1 - \mu_2) \}^{1/2},
\]
which finishes the proof. \( \square \)

**2.2. Non-smooth Test Functions.** Here we follow [29, Section 5.3] very closely to derive a bound on the convex set distance between a vector of dissociated sums \( W \in \mathbb{R}^d \) with covariance matrix \( \Sigma \) and a target multivariate normal distribution \( \Sigma^{1/2} Z \), where \( Z \sim \text{MVN}(0, \text{Id}_{d \times d}) \). The smoothing technique used here is introduced in [19]. However, a better (polylogarithmic) dependence on \( d \) could potentially be achieved using a recent result [20, Proposition 2.6], at the expense of larger constants. The recursive approach from [42, 28] usually yields better dependence on \( n \); however, this requires the target normal distribution to have an invertible covariance matrix. Since this property does not always hold in our applications of interest, we do not use the recursive approach here. To state our next result, let \( \mathcal{X} \) be a class of convex sets in \( \mathbb{R}^d \).
THEOREM 2.6. Consider a standard $d$-dimensional Gaussian vector $Z \sim \text{MVN}(0, \text{Id}_{d \times d})$. For any centered vector of dissociated sums $W \in \mathbb{R}^d$ with a positive semi-definite covariance matrix $\Sigma$ and finite third absolute moments we have

$$\sup_{A \in \mathcal{X}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{1/2} Z \in A)| \leq 2^{2d-3} 3^{-d} \frac{1}{d!} B_{1,2}^d,$$

where the quantity $B_{1,2}$ as in Theorem 1.2.

PROOF. Fix $A \in \mathcal{X}$, $\epsilon > 0$ and define

$$A^\epsilon = \left\{ y \in \mathbb{R}^d : d(y, A) < \epsilon \right\}, \quad \text{and} \quad A^{-\epsilon} = \left\{ y \in \mathbb{R}^d : B(y; \epsilon) \subseteq A \right\},$$

where $d(y, A) = \inf_{x \in A} \|x - y\|_2$ and $B(y; \epsilon) = \{ z \in \mathbb{R}^d | \|y - z\|_2 \leq \epsilon \}$.

Let $\mathcal{H}_{*,A} := \{ h_{*,A} : \mathbb{R}^d \to [0, 1]; A \in \mathcal{X} \}$ be a class of functions such that $h_{*,A}(x) = 1$ for $x \in A$ and 0 for $x \notin A^\epsilon$. Then, by [7, Lemma 2.1] as well as inequalities (1.2) and (1.4) from [7], for any $\epsilon > 0$ we have:

$$\sup_{A \in \mathcal{X}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{1/2} Z \in A)| \leq 4d^{\frac{1}{2}} \epsilon + \sup_{A \in \mathcal{X}} \left| \mathbb{E}h_{*,A}(W) - \mathbb{E}h_{*,A}(\Sigma^{1/2} Z) \right|$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a bounded Lebesgue measurable function, and for $\delta > 0$ let

$$(S_{\delta} f)(x) = \frac{1}{(2\delta)^d} \int_{x_1 - \delta}^{x_1 + \delta} \cdots \int_{x_d - \delta}^{x_d + \delta} f(z) dz_1 \cdots dz_d.$$ 

Set $\delta = \frac{\epsilon}{16 \sqrt{d}}$ and $h_{*,A} = S_{\delta}^4 I_A / 4$, where $I_A$ is the indicator function of the subset $A \subseteq \mathbb{R}^d$. By [19, Lemma 3.9] we have that $h_{*,A}$ is bounded and has three continuous bounded partial derivatives and its third partials are Lipschitz. Moreover, the following bounds hold:

$$\|h_{*,A}\|_\infty \leq 1, \quad |h_{*,A}|_2 \leq \frac{1}{e^2}, \quad |h_{*,A}|_3 \leq \frac{1}{e^3}.$$ 

Note that $h_{*,A} = S_{\delta}^4 I_A / 4 \in \mathcal{H}_{*,A}$ and hence [7, Lemma 2.1] applies. Using this with Theorem 1.2 we get:

$$\sup_{A \in \mathcal{X}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{1/2} Z \in A)| \leq 4d^{\frac{1}{2}} \epsilon + \sup_{A \in \mathcal{X}} \left| \mathbb{E}h_{*,A}(W) - \mathbb{E}h_{*,A}(\Sigma^{1/2} Z) \right| \leq 4d^{\frac{1}{2}} \epsilon + \frac{1}{3e^3} B_{1,2}^d.$$ 

Since this bound works for every $\epsilon > 0$, we minimise it by using $\epsilon = \left( \frac{3B_{1,2}^d}{4d^{\frac{1}{2}}} \right)^{\frac{1}{4}}$.

\[ \square \]

The next result provides a simplification of Theorems 1.2 and 2.6 under the assumption that one uses bounds that are uniform in $s, t, u \in \mathcal{I}$. Its proof follows immediately from writing the sum over $\sum_{s \in \mathcal{I}} \sum_{u \in \mathcal{D}(s)}$ as the sum over $\sum_{i \in [d]} \sum_{j \in [d]} \sum_{k \in [d]} \sum_{s \in \mathcal{I}} \sum_{d \in \mathcal{D}(s)} \sum_{u \in \mathcal{D}(s)}$.

COROLLARY 2.7. We have the following two bounds:

1. \textbf{Under the assumptions of Theorem 1.2,}

$$\left| \mathbb{E}h(W) - \mathbb{E}h(\Sigma^{1/2} Z) \right| \leq B_{2,7} |h|_3.$$
(2) Assuming the hypotheses of Theorem 2.6,

\[
\sup_{A \in \mathcal{A}} |P(W \in A) - P(\Sigma^1 Z \in A)| \leq 2^{\frac{2}{3}} 3^{-\frac{3}{4}} d \pi B_{2,7}^{1/3}.
\]

Here \(B_{2,7}\) is a sum over \((i, j, k) \in [d]^3\) of the form

\[
B_{2,7} := \frac{1}{3} \sum_{(i, j, k)} |\mathbf{I}_i| \alpha_{ij} \left( \frac{3\alpha_{ik}}{2} + 2\alpha_{jk} \right) \beta_{ijk},
\]

and \(\alpha_{ij}\) is the largest value attained by \(|\mathbf{D}_j(s)|\) over \(s \in \mathbf{I}_i\), and

\[
\beta_{ijk} = \max_{s,t,u} \left( E |X_s X_t X_u|, E |X_s X_i| E |X_u| \right)
\]

as \((s, t, u)\) range over \(\mathbf{I}_i \times \mathbf{I}_j \times \mathbf{I}_k\).

3. Simplicial Complex Preliminaries

3.1. First definitions. Firstly, we recall the notion of a simplicial complex \([43, \text{Ch 3.1}]\); these provide higher-dimensional generalisations of a graph and constitute data structures of interest across algebraic topology in general as well as applied and computational topology in particular.

A simplicial complex \(\mathcal{L}\) on a vertex set \(V\) is a set of nonempty subsets of \(V\) (i.e. \(\emptyset \notin \mathcal{L} \subseteq \mathcal{P}(V)\)) such that the following properties are satisfied:

1. for each \(v \in V\) the singleton \(\{v\}\) lies in \(\mathcal{L}\), and
2. if \(t \in \mathcal{L}\) and \(s \subset t\) then \(s \in \mathcal{L}\).

The \textbf{dimension} of a simplicial complex \(\mathcal{L}\) is \(\max_{s \in \mathcal{L}} |s| - 1\). Elements of a simplicial complex are called \textbf{simplices}. If \(s\) is a simplex, then its dimension is \(|s| - 1\). A simplex of dimension \(k\) can be called a \(k\)-simplex. Note that the notion of one-dimensional simplicial complex is equivalent to the notion of a graph, with the vertex set \(V\) and edges as subsets.

Given a graph \(G = (V, E)\) the clique complex \(\mathcal{X}\) of \(G\) is a simplicial complex on \(V\) such that

\[
t \in \mathcal{X} \iff \forall u, v \in t, \{u, v\} \in E.
\]

Recall that \(G(n, p)\) is a random graph on \(n\) vertices where each pair of vertices is connected with probability \(p\), independently of any other pair. The \(X(n, p)\) random simplicial complex is the clique complex of the \(G(n, p)\) random graph, which is a random model studied in stochastic topology \([25, 27]\). Note that \(t \in \mathcal{X}\) if and only if the vertices of \(t\) span a clique in \(G\). Thus, elements in \(X(n, p)\) are cliques in \(G(n, p)\).

3.2. Links. The link of a simplex \(t\) in a simplicial complex \(\mathcal{L}\) is the subcomplex

\[
\text{lk}(t) = \{ s \in \mathcal{L} \mid s \cup t \in \mathcal{L} \text{ and } t \cap s = \emptyset \}.
\]

\textbf{Example 3.1.} If we look at a graph as a one dimensional simplicial complex, then the vertices are sets of the form \(\{i\}\) and edges are sets of the form \(\{i, j\}\). For a vertex \(t = \{v\}\), the edges of the form \(s = \{v, u\}\) will not be in the link of \(t\) because \(t \cap s = \emptyset\) is not satisfied. If we pick \(s = \{i, j\}\) and \(v \notin s\), then \(s \cup t \in \mathcal{L}\) is not satisfied. So there will be no edges in the link. However, if \(s = \{u\}\) and \(u\) is a neighbour of \(v\), then \(s \cup t \in \mathcal{L}\) and \(s \cap t = \emptyset\). Hence the link of a vertex will be precisely the other vertices that the vertex is connected to; the notion of the link generalises the idea of a neighbourhood in a graph.
FIGURE 1. Left: the link (highlighted in blue) of the vertex 1 (highlighted in red). Right: the link (highlighted in blue) of the edge \(\{1, 2\}\) (highlighted in red). The two-dimensional simplices are shaded in grey.

EXAMPLE 3.2. Now consider the simplicial complex depicted in Figure 1: it has 8 vertices, 12 edges and 3 two-dimensional simplices that are shaded in grey. On the left hand side of the figure we see highlighted in blue the link of the vertex 1, which is highlighted in red. So \(\text{lk}(\{1\}) = \{2, 3, 5, 6, 8, 2, 3, 2, 8, 5, 6\}\). On the right hand side of the figure we see highlighted in blue the link of the edge \(\{1, 2\}\), which is highlighted in red. That is, \(\text{lk}(\{1, 2\}) = \{3, 8\}\).

3.3. Discrete Morse theory. A partial matching on a simplicial complex \(L\) is a collection \(\Sigma = \{(s, t) \mid s \subseteq t \in L \text{ and } |t| - |s| = 1\}\) such that every simplex appears in at most one pair of \(\Sigma\). A \(\Sigma\)-path (of length \(k \geq 1\)) is a sequence of distinct simplices of \(L\) of the following form:

\[
(s_1 \subseteq t_1 \supseteq s_2 \subseteq t_2 \supseteq \ldots \supseteq s_k \subseteq t_k)
\]

such that \((s_i, t_i) \in \Sigma\) and \(|t_i| - |s_{i+1}| = 1\) for all \(i \in [k]\). A \(\Sigma\)-path is called a gradient path if \(k = 1\) or \(s_1\) is not a subset of \(t_k\). A partial matching \(\Sigma\) on \(L\) is called acyclic iff every \(\Sigma\)-path is a gradient path. Given a partial matching \(\Sigma\) on \(L\), we say that a simplex \(t \in L\) is critical iff \(t\) does not appear in any pair of \(\Sigma\).

For a one-dimensional simplicial complex, viewed as a graph, a partial matching \(\Sigma\) is comprised of elements \((v; \{u, v\})\) with \(v\) a vertex and \(\{u, v\}\) an edge. A \(\Sigma\)-path is then a sequence of distinct vertices and edges

\[
v_1, \{v_1, v_2\}, v_2, \{v_2, v_3\}, \ldots, v_k, \{v_k, v_{k+1}\}
\]

where each consecutive pair of the form \((v_i, \{v_i, v_{i+1}\})\) is constrained to lie in \(\Sigma\).

We refer the interested reader to [18] for an introduction to discrete Morse theory and to [36] for seeing how it is used to reduce computations in the persistent homology algorithm. In this work we aim to understand how much improvement one would likely get on a random input when using a specific type of acyclic partial matching, defined below.

DEFINITION 3.3. Let \(L\) be a simplicial complex and assume that the vertices are ordered by \([n] = \{1, \ldots, n\}\). For each simplex \(s \in L\) define

\[
I_L(s) := \{j \in [n] \mid j < \min(s) \text{ and } s \cup \{j\} \in L\}.
\]

Now consider the pairings

\[
s \leftrightarrow s \cup \{i\},
\]

where \(i = \min I_L(s)\) is the smallest element in the set \(I_L(s)\), defined whenever \(I_L(s) \neq \emptyset\). We call this the lexicographical matching.
Due to the min \( I_\mathcal{L}(s) \) construction in the lexicographical matching, the indices are decreasing along any path and hence it will be a gradient path, showing that the lexicographical matching is indeed an acyclic partial matching on \( \mathcal{L} \).

**Example 3.4.** Consider the simplicial complex \( \mathcal{L} \) depicted in Figure 2. The complex has 5 vertices, 6 edges and one two-dimensional simplex that is shaded in grey. The red arrows show the lexicographical matching on this simplicial complex: there is an arrow from a simplex \( s \) to \( t \) iff the pair \( (s, t) \) is part of the matching. More explicitly, the lexicographical matching on \( \mathcal{L} \) is

\[
\Sigma = \{ (\{2\}, \{1, 2\}), (\{3\}, \{2, 3\}), (\{4\}, \{1, 4\}), (\{5\}, \{3, 5\}), (\{4, 5\}, \{3, 4, 5\}) \}.
\]

Note that \( \{3, 4\} \) cannot be matched because the set \( I_\mathcal{L}(\{3, 4\}) \) is empty. Also, in any lexicographical matching \( \{1\} \) is always critical as there are no vertices with a smaller label and hence the set \( I_\mathcal{L}(\{1\}) \) is empty. So under this matching there are two critical simplices: \( \{1\} \) and \( \{3, 4\} \), highlighted in blue in the figure. Hence, if we were computing the homology of this complex, considering only two simplices would be sufficient instead of all 12 which are in \( \mathcal{L} \) - a significant improvement.

**4. Critical Simplex Counts for Lexicographical Morse Matchings**

Now we attend to our motivating problem, critical simplex counts. Consider the random simplicial complex \( X(n, p) \). In this section we study the joint distribution of critical simplices in different dimensions with respect to the lexicographical matching on \( X(n, p) \). We start with the following lemma, which is an immediate consequence of Definition 3.3, allowing us to write down the variables of interest in terms of the edge indicators.

**Lemma 4.1.** Let \( \mathcal{L} \) be a simplicial complex. Consider the lexicographical matching on \( \mathcal{L} \). Then \( t \in \mathcal{L} \) matches with one of its cofaces (i.e. \( s \in \mathcal{L} \) with \( |s| - |t| = 1 \) and \( t \subset s \)) iff it is not the case that for all \( j < \min(t) \) we have \( t \cup \{j\} \notin \mathcal{L} \). Also, \( t \in \mathcal{L} \) matches with one of its faces (i.e. \( s \in \mathcal{L} \) with \( |t| - |s| = 1 \) and \( s \subset t \)) iff for all \( j < \min(t) \) we have \( t \setminus \{\min(t)\} \cup \{j\} \notin \mathcal{L} \).

For any pair of integers \( 1 \leq i < j \leq n \) let \( Y_{ij} := \mathbb{1} \{ \{i, j\} \in X(n, p) \} \) be the edge indicator. Fix \( s \in C_k \). Define the variables \( X^+_s = \mathbb{1} \{ s \text{ matches with its coface given it is a simplex} \} \) and \( X^-_s = \mathbb{1} \{ s \text{ matches with its face given it is a simplex} \} \). The events that the two variables indicate are disjoint. By Lemma 4.1 we can see that \( X^+_s = 1 - \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s} Y_{ij} \right) \) and \( X^-_s = \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s} Y_{ij} \right) \), where \( s_+ := s \setminus \{\min(s)\} \). Hence, for any \( s \in X(n, p) \),

\[
\mathbb{1} \{ s \text{ is a critical simplex} \} = \mathbb{1} \{ s \in X(n, p) \} \left( 1 - (X^+_s + X^-_s) \right)
\]

\[
= \prod_{i \neq j \in s} Y_{ij} \left[ \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s} Y_{ij} \right) - \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s} Y_{ij} \right) \right].
\]
Thus, the random variable of interest, counting the number of \((k - 1)\)-simplices that are critical under the lexicographical matching, is

\[
T_k = \sum_{s \in C_k} \prod_{i \neq j \in s} Y_{ij} \left[ \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s}^{} Y_{ij} \right) - \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s}^{} Y_{ij} \right) \right].
\]  

(4.1)

Note that this random variable does not fit into the framework of generalised \(U\)-statistics, which we will discuss in Section 6, because the summands in \(T_k\) depend not only on the variables that are indexed by the subset \(s\).

### 4.1. Moments.

**Lemma 4.2.** For any \(1 \leq k \leq n - 1\) we have:

\[
p^{k+1 \choose 2} + k \left( \frac{n-2}{k} \right) (1 - p) \leq \mathbb{E}\{T_{k+1}\} \leq p^{k+1 \choose 2} \left( \frac{n-1}{k} \right) (1 - p).
\]

**Proof.**

\[
\mathbb{E}\{T_{k+1}\} = \sum_{l=1}^{n-k} \sum_{s \in C_{k+1}} \mathbb{E}\left\{ \prod_{i \neq j \in s} Y_{ij} \left[ \prod_{i=1}^{l-1} \left( 1 - \prod_{j \in s}^{} Y_{ij} \right) - \prod_{i=1}^{l-1} \left( 1 - \prod_{j \in s}^{} Y_{ij} \right) \right] \right\}
\]

\[
= p^{k+1 \choose 2} \frac{n-k}{l} \sum_{s \in C_{k+1}} \left\{ (1 - p^{k+1 \choose l-1} - (1 - p^{k \choose l-1}) \right\}
\]

\[
= p^{k+1 \choose 2} \frac{n-k-1}{l} \left( \frac{n-l}{k} \right) \left\{ (1 - p^{k+1 \choose l} - (1 - p^{k \choose l}) \right\}
\]

\[
\leq p^{k+1 \choose 2} \left( \frac{n-1}{k} \right) \sum_{l=0}^{\infty} \left\{ (1 - p^{k+1 \choose l} - (1 - p^{k \choose l}) \right\}
\]

\[
= p^{k+1 \choose 2} \left( \frac{n-1}{k} \right) (1 - p).
\]

Moreover,

\[
\mathbb{E}\{T_{k+1}\} = p^{k+1 \choose 2} \frac{n-k-1}{l} \left( \frac{n-l-1}{k} \right) \left\{ (1 - p^{k+1 \choose l+1} - (1 - p^{k \choose l+1}) \right\}
\]

\[
\geq p^{k+1 \choose 2} \left( \frac{n-2}{k} \right) \left\{ (1 - p^{k+1 \choose l+1} - (1 - p^{k \choose l+1}) \right\}
\]

\[
= p^{k+1 \choose 2} \left( \frac{n-2}{k} \right) (1 - p).
\]

\[\square\]

In this example, bounding the variance is not immediate. The proof of the following Lemmas 4.3 and 4.4 are long (and not particularly insightful) calculations, which are deferred to the Appendix.

**Lemma 4.3.** For any integer \(1 \leq k \leq n - 1\) we have:

\[
\text{Var}\{T_{k+1}\} = 2p^{2k+1 \choose 2} V_1 + 2p^{2k+1 \choose 2} V_2 + p^{2k+1 \choose 2} V_3 + p^{k+1 \choose 2} V_4,
\]

where
\[ V_1 = \sum_{i<j} \sum_{m=1}^{k} \sum_{q=1}^{\min(k+1,j-i)} \left( \begin{array}{c} n-j \\ 2k+1-m-q \\ k \\ m-1 \\ q-1 \end{array} \right) \left( 2k+1-m-q \right) \left( k \right) \left( j-i+1 \right) \]

\[ \left\{ \theta(i,j,q,m,1) \left[ (1-2p^{k+1} + p^{2k+2-m})^{i-1} - (1-p^{k+1} - p^{k} + p^{2k+2-m})^{i-1} \right] \right. \]

\[ \left. + \theta(i,j,q,m,0) \left[ (1-2p^{k} + p^{2k+2-m})^{i-1} - (1-p^{k+1} - p^{k} + p^{2k+2-m})^{i-1} \right] - \eta(i) \eta(j) \right\}; \]

\[ V_2 = \sum_{i<j} \sum_{m=1}^{k} \sum_{q=1}^{\min(k+1,j-i)} \left( \begin{array}{c} n-j \\ 2k+1-m-q \\ k \\ m-1 \\ q-1 \end{array} \right) \left( 2k+1-m-q \right) \left( k \right) \left( j-i+1 \right) \]

\[ \left\{ \theta(i,j,q,m,1) \left[ (1-2p^{k+1} + p^{2k+2-m})^{i-1} - (1-p^{k+1} - p^{k} + p^{2k+2-m})^{i-1} \right] \right. \]

\[ \left. + \theta(i,j,q,m,0) \left[ (1-2p^{k} + p^{2k+2-m})^{i-1} - (1-p^{k+1} - p^{k} + p^{2k+2-m})^{i-1} \right] - \eta(i) \eta(j) \right\}; \]

\[ V_3 = \sum_{i=1}^{n-k} \sum_{m=1}^{k} \left( \begin{array}{c} n-i \\ 2k+1-m \end{array} \right) \left( 2k+1-m \right) \left( k \right) \left( m-1 \right) \left( 1 - 2p^{k} + p^{2k-1-m} \right) \]

\[ \leq \sum_{i=1}^{n-k} \left( \begin{array}{c} n-i \\ k \end{array} \right) \left\{ \eta(i) - \left( p^{k+1} \right)^{\frac{1}{2}} \eta(i)^{2} \right\}. \]

Here we have used the following notation:

\[ \eta(a) := (1-p^{k+1})^{a-1} - (1-p^{k})^{a-1}; \]

\[ \theta(i,j,q,m,\delta) := \left[ (1-p^{k+1} + p^{2k+2-m})^{i-1} \right] \left[ (1-p^{k} + p^{2k+2-m})^{j-1} \right] - \eta(i) \eta(j). \]

Also, \( \sum_{i<j}^{n-k} \) stands for \( \sum_{i=1}^{n-k} \sum_{j=i+1}^{n-k} \).

**Lemma 4.4.** For a fixed integer \( 1 \leq k \leq n-1 \) and \( p \in (0,1) \) there is a constant \( C_{p,k} > 0 \) independent of \( n \) and a natural number \( N_{p,k} \) such that for any \( n \geq N_{p,k} \):

\[ \text{Var}(T_{k+1}) \geq C_{p,k} n^{2k}. \]

In Lemma 4.4 the constant could have been made explicit at the expense of an even longer calculation.

Just knowing the expectation and the variance can already give us some information about the variable. For example, we obtain the following proposition. This proposition shows that considering only a subset of the simplices already gives a good approximation for the critical simplex counts. We recall the notation that \( f(n) = \omega(g(n)) \) indicates that \( \lim_{n \to \infty} \frac{f(n)}{g(n)} = \infty. \)

**Proposition 4.5.** Fix \( k \in [n] \). Let \( K \leq n-k \) and set the random variable:

\[ T_{k+1}^{K} := \sum_{s \in C_{k+1}} \prod_{i \neq j \in s} \left( \min(s) \right)^{-1} \left( 1 - \prod_{j \in s} Y_{i,j} \right) - \prod_{i=1}^{\min(s)-1} \left( Y_{i,j} \right). \]

If \( K = K(n) = \omega(\ln^{1+\epsilon}(n)) \) for any \( \epsilon > 0 \), then the variable \( T_{k+1}^{K} \) vanishes with high probability, provided that \( p \) and \( k \) stay constant.
PROOF. A similar calculation to that for Lemma 4.2 shows that:
\[
\mathbb{E} \left\{ T_{k+1} - N_{k+1}^T \right\} = n-k \sum_{i=k+1}^{n-k} \binom{n-i}{k} p^{(i+1)} \left\{ (1 - p^{k+1})^{i-1} - (1 - p^i)^{i-1} \right\}
\leq \binom{n}{k} p^{(i+1)} (1 - p^{k+1}) K \sum_{i=0}^{\infty} (1 - p^{k+1})^i \leq \binom{n}{k} p^{(i+1)} - k \binom{n}{k} (1 - p^{k+1}) K.
\]

Using Markov’s inequality, we get:
\[
\mathbb{P}(T_{k+1} - N_{k+1}^T \geq 1) \leq \binom{n}{k} p^{(i+1)} - k \binom{n}{k} (1 - p^{k+1}) K,
\]
which asymptotically vanishes as long as \( K = o(\ln^{1+\epsilon}(n)) \). \qed

4.2. Approximation theorem. For \( i \in [d] \), recall a random variable counting \( i \)-simplices in \( X(n, p) \) that are critical under the lexicographical matching, as given in (4.1). We write for the \( i \)-th index set \( \mathbb{I}_i := C_{i+1} \times \{ i \} \). For \( s = (\phi, i) \in \mathbb{I}_i \), we write
\[
\mu_s = p^{(i+1)} \left( (1 - p^{i+1})^{\min(\phi) - 1} - (1 - p^i)^{\min(\phi) - 1} \right)
\]
and \( \sigma_i = \sqrt{\text{Var}(T_{i+1})} \). Let
\[
X_s = \sigma_i^{-1} \left\{ \prod_{j \in \phi} Y_{i,j} \left[ \prod_{i=1}^{\min(\phi) - 1} \left( 1 - \prod_{j \in \phi} Y_{i,j} \right) - \prod_{i=1}^{\min(\phi) - 1} \left( 1 - \prod_{j \in \phi} Y_{i,j} \right) - \mu_s \right] \right\}.
\]

Let \( W_i = \sum_{s \in \mathbb{I}_i} X_s \) and \( W = (W_1, W_2, \ldots, W_d) \in \mathbb{R}^d \). For bounds that asymptotically go to zero for this example, we use Theorems 1.2 and 2.6 directly: the uniform bounds from Corollary 2.7 are not fine enough here.

THEOREM 4.6. Let \( Z \sim \text{MVN}(0, \text{Id}_{d \times d}) \) and \( \Sigma \) be the covariance matrix of \( W \).

1. Let \( h \in \mathcal{H}_i \). Then there is a constant \( B_{4.6.1} > 0 \) independent of \( n \) and a natural number \( N_{4.6.1} \) such that for any \( n \geq N_{4.6.1} \) we have
\[
\left| \mathbb{E} h(W) - \mathbb{E} h(\Sigma^{1/2} Z) \right| \leq B_{4.6.1} |h|_2 n^{-1}.
\]

2. Let \( \mathcal{K} \) be the class of convex sets in \( \mathbb{R}^d \). Then there is a constant \( B_{4.6.2} > 0 \) independent of \( n \) and a natural number \( N_{4.6.2} \) such that for any \( n \geq N_{4.6.2} \) we have
\[
\sup_{A \in \mathcal{K}} \left| \mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{1/2} Z \in A) \right| \leq B_{4.6.2} n^{-\frac{1}{2}}.
\]

PROOF. It is clear that \( W \) satisfies the conditions of Theorems 1.2 and 2.6 for any \( s = (\phi, i) \in \mathbb{I}_i \) setting
\[
\mathcal{D}_i(s) = \{ (\phi, j) \in \mathbb{I}_j | \phi \cap \phi \geq 1 \}.
\]
We apply Theorems 1.2 and 2.6. For the bounds on the quantity \( B_{1.2} \) from Theorems 1.2 and 2.6 we use Lemma 2.5 and Lemma 4.4. We write \( C \) for an unspecified positive constant that does not depend on \( n \). Also, we assume here that \( n \) is large enough for the bound in Lemma 4.4 to apply. Let \( \mu(i, a) = p^{(i+1)} \left( (1 - p^{i+1})^{a-1} - (1 - p^i)^{a-1} \right) \). Then we have:
\[
B_{1.2} \leq \frac{1}{3} \sum_{i,j,k=1}^{d} \sum_{a=1}^{n-i} \sum_{\phi \in C_{i+1}} \sum_{b=1}^{n-j} \sum_{(\phi,j) \in \mathcal{D}_i((\phi,i))} \sum_{\min(\phi) = a} \sum_{\min(\phi) = b} \mu(i, a).
\]
\[
\left\{ \sum_{r \in \mathbb{D}_k((\phi,j))} \frac{3}{2} (\sigma_i \sigma_j \sigma_k)^{-1} \{\mu(i,a)\mu(j,b)(1 - \mu(i,a))(1 - \mu(j,b))\}^{1/2} \right. \\
+ \sum_{r \in \mathbb{D}_k((\phi,j))} (\sigma_i \sigma_j \sigma_k)^{-1} \{\mu(i,a)\mu(j,b)(1 - \mu(i,a))(1 - \mu(j,b))\}^{1/2} \right\} \\
\leq \sum_{i,j,k=1}^{d} \sum_{a=1}^{n-i} \sum_{b=1}^{n-j} Cn^{i+j-1}n^k n^{-i-j-k} \left\{ (1 - p^{i+1})^{a-1}(1 - p^{j+1})^{b-1} + (1 - p^{i+1})^{b-1}(1 - p^{j+1})^{a-1} \right\}^{1/2} \\
\leq Cn^{-1} \sum_{i,j,k=1}^{d} \sum_{a=1}^{\infty} \sum_{b=1}^{\infty} \left[ (1 - p^{i+1})^{a-1}(1 - p^{j+1})^{b-1} \right]^{1/2} + \left[ (1 - p^{i+1})^{b-1}(1 - p^{j+1})^{a-1} \right]^{1/2} \\
\leq Cn^{-1} d^3 \left( \frac{1}{(1 - \sqrt{1 - p^{i+1}})^2} + \frac{1}{(1 - \sqrt{1 - p^{j+1}})^2} \right) \leq Cn^{-1}.
\]

\[\square\]

**Remark 4.7.** The relevance of understanding the number of critical simplices in the context of applied and computational topology is as follows. We assume that \(p \in (0,1)\) and \(k \in \{1,2,\ldots\}\) are constants.

1. As seen in Lemma 4.2, the expected number of critical \(k\)-simplices under the lexicographical matching is one power of \(n\) smaller than the total number of \(k\)-simplices in \(X(n,p)\).
2. In light of our approximation Theorem 4.6 we also know that the (rescaled) deviations from the mean are approximately normal and the bounds are of the same order of \(n\) compared to the approximation of all simplex counts in \(X(n,p)\) as given in Theorem 6.6. Knowing the expectation and the variance from Lemmas 4.3 and 4.2, one can apply concentration inequalities, for example, Chebyshev’s inequality, to show that the number of critical simplices concentrates around its mean. Hence, because of the concentration of measure, the computational improvements as a result of lexicographical matching are likely substantial in \(X(n,p)\).
3. From Proposition 4.5, it is very likely that in \(X(n,p)\) all \(k\)-simplices \(s \in X(n,p)\) with \(\min(s) = \omega(\ln^{1+\varepsilon}(n))\) for any fixed \(\varepsilon > 0\) are not critical.

### 5. Simplex Counts in Links

Consider a random simplicial complex \(X(n,p)\). For \(1 \leq i < j \leq n\) define the edge indicator \(Y_{ij} := 1\ (\{i,j\} \in X(n,p))\). In this section we study the count of \((k-1)\)-simplices that would be in the link of a fixed subset \(t \subseteq [n]\) if the subset spanned a simplex in \(X(n,p)\). Given that \(t\) is a simplex, the variable counts the number of \((k-1)\)-simplices in \(\text{lk}(t)\). Thus, the random variable of interest is

\[T^*_k = \sum_{s \in C_k} \left\{ \mathbbm{1}(t \cap s = \emptyset) \prod_{i \neq j} Y_{ij} \prod_{i \in s, j \not\in t} Y_{ij} \right\} \tag{5.1} \]

Note that the product \(\prod_{i \in s, j \not\in t} Y_{ij}\) ensures that \(t \cup s\) is a simplex if \(t\) spans a simplex.

**Remark 5.1.** The random variable \(T^*_k\) does not fit into the framework of generalised \(U\)-statistics, which we will discuss in Section 6, because the summands depend not only on the variables that are indexed by the subset \(s\) and we do not sum over all subsets \(s\) but rather only the ones that do not intersect \(t\).
Moreover, note that given the number of vertices of the link of a simplex $t$, the conditional distribution of the link of $t$ is again $X(n', p)$, where $n'$ is a random variable equal to the number of vertices in the link. If we are interested in such a conditional distribution, the results of Section 6 apply. However, in this section we study the number of simplices in the link of $t$ given that $t$ is a simplex rather than given the number of vertices of the link of $t$. Such a random variable behaves differently from the simplex counts in $X(n, p)$ which are studied in Section 6. For example, the summands of $T_k^t$ have a different dependence structure compared to the summands of $T_k$ from Equation 6.4. As a result, the approximation bounds are of different order.

It is natural to ask whether the results obtained in this section follow from those of Section 6 below. This might well be the case, but the answer is not straightforward. One could derive an approximation for the number of simplices in $1k(t)$ given the number of vertices in the link; the variable $T_k^t$ could then be approximated by a mixture, induced by the distribution of the number of vertices in the link (which is binomial). However, applying this approach naively yields bounds that do not converge to zero. While it is certainly possible that a different approach would succeed, we prefer not to rely on Section 6 and prove the approximation directly.

5.1. Moments. It is easy to see that for any positive integer $k$ and $t \subseteq [n],$
\[
\mathbb{E}\{T_{k+1}^t\} = \binom{n - |t|}{k+1} p^{|t|+1} =: \binom{n - |t|}{k+1} \mu_{k+1}^t.
\]
since there are $\binom{n - |t|}{k+1}$ choices for $s \in C_{k+1}$ such that $s \cap t = \emptyset$. Next we derive a lower bound on the variance.

**Lemma 5.2.** For any fixed $1 \leq k \leq n-1$ and $t \subseteq [n]$ we have:
\[
\text{Var}(T_{k+1}^t) \geq (k+1) \binom{n - |t|}{k+1} \binom{2k+1}{k} \left( \frac{1}{2k+1} \right)^2 \left( p^{-|t|} - 1 \right).
\]

**Proof.** First let us calculate $\text{Cov}(T_{k+1}^t, T_{l+1}^u)$. For fixed subsets $s \in C_{k+1}$ and $u \in C_{l+1}$ if $|s \cap u| = 0$, then the corresponding variables $\prod_{i \notin s} Y_{ij} \prod_{i \in s, j \in t} Y_{ij}$ and $\prod_{i \notin u} Y_{ij} \prod_{i \in u, j \in t} Y_{ij}$ are independent and so have zero covariance.

For $1 \leq m \leq l+1$, the number of pairs of subsets $s \in C_{k+1}$ and $u \in C_{l+1}$ such that $s \cap t = u = u \cap t$ and $|s \cap u| = m$ is $\binom{n - |t|}{k+1} \binom{n - |t| - k - 1}{l+1-m}$. Since each summand is non-negative, we lower bound by the $m = 1$ summand and get (with $\binom{m}{2} := 0$)
\[
\text{Cov}(T_{k+1}^t, T_{l+1}^u) = \sum_{m=1}^{l+1} \binom{n - |t|}{k+1} \binom{n - |t| - k - 1}{l+1-m} \mu_{k+1}^t \mu_{l+1}^u p^{-|t|} p^{-|u|} - \mu_{k+1}^t \mu_{l+1}^u \mu_{k+1}^l \mu_{l+1}^u \left\{ p^{-|t|} - 1 \right\}.
\]

Taking $l = k$ completes the proof. \hfill $\Box$

5.2. Approximation theorem. For a multivariate normal approximation of counts given in Equation (5.1), we write $\sigma_i = \sqrt{\text{Var}(T_{i+1})}$ and $C_{i+1} = \{ \phi \in C_{i+1} \mid \phi \cap t = \emptyset \}$, as well as $I_i := C_{i+1} \times \{ i \}$. For $s = (\phi, i) \in I_i$ define
\[
X_s = \sigma_i^{-1} \left( \prod_{i \notin \phi} Y_{ij} \prod_{i \in \phi, j \in t} Y_{ij} - \mu_{i+1}^t \right).
\]
It is clear that $\mathbb{E} \{X_s\} = 0$. Let $W'_l = \sum_{s \in I_l} X_s$ and $W^l = (W'_1, W'_2, \ldots, W'_{d^l}) \in \mathbb{R}^{d^l}$. Then we have the following approximation theorem.

**Theorem 5.3.** Let $Z \sim \text{MVN}(0, \text{Id}_{d \times d})$ and $\Sigma$ be the covariance matrix of $W^l$.

1. Let $h \in \mathcal{K}_{d^l}$. Then

$$\left| \mathbb{E} h(W^{l^t}) - \mathbb{E} h(\Sigma^{1/2} Z) \right| \leq |h|_3 B_{5.3}(n - |t|)^{-1/2}.$$

2. Let $\mathcal{K}$ be the class of convex sets in $\mathbb{R}^d$. Then

$$\sup_{A \in \mathcal{K}} |\mathbb{P}(W^l \in A) - \mathbb{P}(\Sigma^{1/2} Z \in A)| \leq 2^{2d+4} d^{3/2} B_{5.3}(n - |t|)^{-1/2}.$$

Here

$$B_{5.3} = \frac{7}{6} (2d + 1)^{3d+17} (p^{|t|} - 1)^{-1/2} p^{-d(d+2|t|)}.$$

**Proof.** It is clear that $W^l$ satisfies the conditions of Corollary 2.7 with the dependency neighborhood $D_j(s) = \{ (\phi, i) \in I_j \mid |\phi \cap \psi| \geq 1 \}$ for any $s = (\phi, i) \in I_i$. So we aim to bound the quantity $B_{2.7}$ from the corollary.

Given $\phi \in C_{i+1}^l$ and $m \leq \min(i, j, j + 1)$ there are $\binom{i+1}{m} \binom{n-|t|-i-1}{j+1-m}$ subsets $\psi \in C_{j+1}^l$ such that $|\phi \cap \psi| = m$. Therefore, for any $i, j \in [d]$ and $s \in I_i$, we have

$$|D_j(s)| = \sum_{m=1}^{\min(i,j)+1} \binom{i+1}{m} \binom{n-|t|-i-1}{j+1-m} \leq (i + 1)^{\min(i,j)+2}(n - |t|)^j \leq (d + 1)^{d+2}(n - |t|)^j \tag{5.2}$$

giving a bound for $\alpha_{ij}$. For a bound on $\beta_{ijk}$, applying Lemma 2.5, for any $i, j, k \in [d]$ and $s \in I_i, u \in I_j, v \in I_k$ we get

$$\mathbb{E} |X_sX_uX_v| \leq (\sigma_i \sigma_j \sigma_k)^{-1} \left\{ \mu_{i+1}^l \mu_{j+1}^l (1 - \mu_{i+1}^l)(1 - \mu_{j+1}^l) \right\}^{1/2}; \tag{5.3}$$

$$\mathbb{E} |X_sX_u| \mathbb{E} |X_v| \leq (\sigma_i \sigma_j \sigma_k)^{-1} \left\{ \mu_{i+1}^l \mu_{j+1}^l (1 - \mu_{i+1}^l)(1 - \mu_{j+1}^l) \right\}^{1/2}. \tag{5.4}$$

Now we apply Corollary 5.2 and get

$$\sigma_i^2 \geq (i + 1) \left( \binom{n-|t|}{2i+1} \right) \left( \binom{n-|t|}{i} \right) (\mu_{k+1}^l)^2 \left\{ p^{|t|} - 1 \right\} \geq \frac{(n - |t|)^{2i+1}}{(2d + 1)^{d+1} d^2} (\mu_{k+1}^l)^2 \left\{ p^{|t|} - 1 \right\}. \tag{5.5}$$

Taking both sides of the inequality to the power of $-\frac{1}{2}$ we get for any $i \in [d]$

$$\sigma_i^{-1} \leq (n - |t|)^{-i-\frac{1}{2}} (2d + 1)^{\frac{d+1}{2}} d^2 (\mu_{k+1}^l)^{-1} \left\{ p^{|t|} - 1 \right\}^{-\frac{1}{2}}. \tag{5.5}$$

Using Equations (5.2) - (5.5) to bound $B_{2.7}$ from Corollary 2.7 we get:

$$B_{2.7} \leq \frac{7}{6} \sum_{i,j,k=1}^{d} \binom{n-|t|}{i+1} (d + 1)^{2d+4} (n - |t|)^{i+j+k} (\sigma_i \sigma_j \sigma_k)^{-1} \left\{ \mu_{i+1}^l \mu_{j+1}^l (1 - \mu_{i+1}^l)(1 - \mu_{j+1}^l) \right\}^{1/2}.$$
\[
\leq \frac{7}{6} \sum_{i,j,k=1}^{d} (n - |t|)^{i+j+k+1} (d+1)^{2d+4} (n - |t|)^{-i-j-k-\frac{1}{2}} (2d+1)^{\frac{3d+1}{2} d^2} \\
(p^{-|t|} - 1)^{-\frac{3}{2}} (\mu_k^{-1})^{i+j=1} \left\{ (\mu_{k+1}^{t_{j+1}} - 1) (1 - \mu_{j+1}^{t_{j=1}} (1 - \mu_{j+1}^{t_{j=1}})) \right\}^{\frac{1}{2}} \\
\leq (n - |t|)^{-\frac{7}{6}} (2d+1)^{5d+\frac{7}{2}} (p^{-|t|} - 1)^{-\frac{3}{2}} \sum_{i,j,k=1}^{d} \left( (\mu_{i+1}^{t_{j+1}})^{-1} (\mu_{k+1}^{t_{j=1}})^{-2} \right)^{\frac{1}{2}} \\
\leq \left\{ \frac{7}{6} (2d+1)^{5d+\frac{7}{2}} (p^{-|t|} - 1)^{-\frac{3}{2}} p^{-(d+1)(d+2|t|)} \right\} (n - |t|)^{-\frac{7}{2}}.
\]

\[\square\]

**Remark 5.4.** Recall that \(E \{ T_k^{T+1} \} = \binom{n-|t|}{k+1} \binom{\binom{t+1}{k+1} + |t|(k+1)}{\binom{t+1}{k+1}} \). By Stirling’s approximation, if \(p \in (0,1)\) is a constant, then \(\max(k, |t|) = \Omega(\ln^{1+\varepsilon}(n))\) for any positive \(\varepsilon\) forces the expectation to go to 0 asymptotically. Hence, by Markov’s inequality, with high probability there are no \(k\)-simplices in the link of \(t\) as long as \(\max(k, |t|) = \Omega(\ln^{1+\varepsilon}(n))\) or larger for any \(\varepsilon > 0\) for a constant \(p\).

Recall that in Theorem 5.3 we count all simplices up to dimension \(d\) in the link of \(t\). Note that if \(\max(d^2, d|t|) = O(\ln^{-1+\varepsilon}(n))\) for any \(\varepsilon > 0\), then the bounds in Theorem 5.3 tend to 0 as \(n\) tends to infinity as long as \(p \in (0,1)\) stays constant. In particular, if \(d\) is a constant, Theorem 5.3 gives an approximation for all sizes of \(t\) for which the approximation is needed.

### 6. Simplex Counts in \(X(n, p)\)

In this section we study the simplex counts in \(X(n, p)\) or, equivalently, the clique counts in \(G(n, p)\). In order to do that, we prove a multivariate normal approximation theorem for generalised \(U\)-statistics, which might be of independent interest. The approximation theorem for simplex counts in \(X(n, p)\) then follows as a special case.

Here we consider generalised \(U\)-statistics, which were first introduced in [23]. We expand the notion slightly by considering independent but not necessarily identically distributed variables instead of i.i.d. variables.

Let \(\{\xi_i\}_{1 \leq i \leq n}\) be a sequence of independent random variables taking values in a measurable set \(\mathcal{X}\) and let \(\{Y_{ij}\}_{1 \leq i, j \leq n}\) be an array of independent random variables taking values in a measurable set \(\mathcal{Y}\) which is independent of \(\{\xi_i\}_{1 \leq i \leq n}\). We use the convention that \(Y_{i,j} = Y_{j,i}\) for any \(i \neq j\). For example, one can think of \(X_i\) as a random label of a vertex \(i\) in a random graph where \(Y_{i,j}\) is the indicator for the edge connecting \(i\) and \(j\). Given a subset \(s \subseteq [n]\) of size \(m\), write \(s = \{s_1, s_2, \ldots, s_m\}\) such that \(s_1 < s_2 < \ldots < s_m\) and set \(\mathcal{X}_s = (\xi_{s_1}, \xi_{s_2}, \ldots, \xi_{s_m})\) and \(\mathcal{Y}_s = (Y_{s_1, s_2}, Y_{s_1, s_3}, \ldots, Y_{s_{m-1}, s_m})\). Recall that \(C_k\) denotes the set of subsets of \([n]\) which are of size \(k\).

**Definition 6.1.** Given \(1 \leq k \leq n\) and a measurable function \(f : \mathcal{X}^k \times \mathcal{Y}^k \to \mathbb{R}\) define the associated generalised \(U\)-statistic by

\[
S_{n,k}(f) = \sum_{s \in C_k} f(\mathcal{X}_s, \mathcal{Y}_s).
\]

**6.1. The First Approximation Theorem.** Let \(\{k_i\}_{i \in [d]}\) be a collection of positive integers, each being at most \(n\), and for each \(i \in [d]\) let \(f_i : \mathcal{X}^{k_i} \times \mathcal{Y}^{k_i} \to \mathbb{R}\) be a measurable function. We are interested in the joint distribution of \(S_{n,k_1}(f_1), S_{n,k_2}(f_2), \ldots, S_{n,k_d}(f_d)\).

Fix \(i \in [d]\). For \(s \in \mathbb{I}_i := \mathbb{C}_{k_i} \times \{i\}\) define \(X_s = \sigma_i^{-1} (f_i(\mathcal{X}_s, \mathcal{Y}_s) - \mu_s)\), where \(\mu_s = E \{ f_i(\mathcal{X}_s, \mathcal{Y}_s) \}\) and \(\sigma_i^2 = \text{Var}(S_{n,k_i}(f_i))\). Now let \(W_i = \sum_{s \in \mathbb{I}_i} X_s\) be a random variable and write \(W = (W_1, W_2, \ldots, W_d) \in \mathbb{R}^d\).

By construction, \(W_i\) has mean 0 and variance 1.
ASSUMPTION 6.2. We assume that

(1) For any \( i \in [d] \) there is some \( \alpha_i > 0 \) such that for all \( s,t \in I_i \), the variables \( f_i(\mathcal{X}_s, \mathcal{Y}_t) \) are either independent or \( \text{Cov}(f_i(\mathcal{X}_s, \mathcal{Y}_t), f_i(\mathcal{X}_t, \mathcal{Y}_s)) > \alpha_i \).

(2) There is \( \beta \geq 0 \) such that for any \( i,j,l \in [d] \) and any \( s \in I_i, t \in I_j, u \in I_l \) we have
\[
\mathbb{E} \left\{ f_i(\mathcal{X}_s, \mathcal{Y}_t) - \mu_s \right\} \left\{ f_j(\mathcal{X}_t, \mathcal{Y}_l) - \mu_t \right\} \left\{ f_l(\mathcal{X}_u, \mathcal{Y}_u) - \mu_u \right\} \leq \beta
\]
as well as
\[
\mathbb{E} \left\{ f_i(\mathcal{X}_s, \mathcal{Y}_t) - \mu_s \right\} \left\{ f_j(\mathcal{X}_t, \mathcal{Y}_l) - \mu_t \right\} \left| \mathbb{E} \left\{ f_l(\mathcal{X}_u, \mathcal{Y}_u) - \mu_u \right\} \leq \beta.
\]

(3) The random variables \( X_s \) have finite absolute third moments.

The first assumption is not necessary but very convenient and we use it to derive a lower bound for the variance \( \sigma_i^2 \). It holds in a variety of settings, for example, subgraph counts in a random graph. A normal approximation theorem can be proven in our framework when the assumption does not hold and a sufficiently large lower bound for the variance is acquired in a different way. Similarly, we use the second assumption to get a convenient bound on mixed moments. However, depending on a particular question at hand, one might want to use a bound on mixed moments, which is not uniform in \( i,j,l \) and sometimes even one that is not uniform in \( s,u,v \). We will discuss such an example (which does not fit into the framework of generalised U-statistics) in Section 4. In this section, in order to maintain the generality and simplicity of the proofs, we work under Assumption 6.2. In [23, Theorem 6] it is assumed that all summands in the generalised U-statistic have finite second moment as well as that the sums admit a particular decomposition which is not easily translatable to our framework. In contrast to [23] we obtain a non-asymptotic bound on the normal approximation, as follows.

THEOREM 6.3. Let \( Z \sim \text{MVN}(0, \text{Id}_{d \times d}) \) and let \( W \) with covariance matrix \( \Sigma \) satisfy Assumption 6.2.

(1) Let \( h \in \mathcal{M}_d \). Then
\[
\left| \mathbb{E} h(W) - \mathbb{E} h(\Sigma^{1/2}Z) \right| \leq |h|_3 B_{6.3} n^{-\frac{3}{2}}.
\]

(2) Let \( \mathcal{M} \) be a class of convex sets in \( \mathbb{R}^d \). Then
\[
\sup_{A \in \mathcal{M}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{1/2}Z \in A)| \leq 2^{\frac{d}{2}} 3^{-\frac{d}{2}} \pi B_{6.3} n^{-\frac{3}{2}}.
\]

Here,
\[
B_{6.3} = \frac{2 \psi}{3} \sum_{i,j=1}^d k_i^{\min(k_i,k_j)+1} k_j^{\min(k_i,k_j)+1} \left( k_i^{\min(k_i,k_j)+1} + k_j^{\min(k_i,k_j)+1} \right) K_i K_j K_l
\]
and
\[
K_i = (2k_i^2 - 2)^{-\frac{3}{2}} + 4.
\]

PROOF. Note that if \( s = (\phi, i) \in I_i \) and \( u = (\psi, j) \in I_j \) are chosen such that \( \phi \cap \psi = \emptyset \), then the corresponding variables \( X_s \) and \( X_u \) are independent since \( f_i(\mathcal{X}_s, \mathcal{Y}_j) \) and \( f_j(\mathcal{X}_u, \mathcal{Y}_u) \) do not share any random variables from the sets \( \{\xi_i\}_{1 \leq i \leq n} \) and \( \{\chi_j\}_{1 \leq j \leq n} \). Hence, if for any \( s = (\phi, i) \in I_i \) we set \( D_i(s) = \{ (\phi, j) \in I_j \mid |\phi \cap \psi| \geq 1 \} \), then \( W \) satisfies the assumptions of Corollary 2.7. It remains to bound the quantity \( B_{2.7} \).

First, to find \( a_{ij} \) as in Corollary 2.7, given \( \phi \in C_{k_i} \) and if \( k_i, k_j \geq m \) then there are \( \binom{k_i}{m} \binom{n-k_i}{k_j-m} \) subsets \( \psi \in C_{k_j} \) such that \(|\phi \cap \psi| = m \). Therefore, we have for any \( i,j \in [d] \) and \( s \in I_i \)
\[
|D_j(s)| = \sum_{m=1}^{\min(k_i,k_j)} \binom{k_i}{m} \binom{n-k_i}{k_j-m} = a_{ij} \leq k_i^{\min(k_i,k_j)+1} (n-k_i)^{k_j-1}.
\]
Note that
\[ E[XX_1X_u] = (\sigma_i \sigma_j \sigma_k)^{-1} E \left[ \left\{ f_i(X_s, Y_s) - \mu_s \right\} \left\{ f_j(X_t, Y_t) - \mu_t \right\} \left\{ f_l(X_u, Y_u) - \mu_u \right\} \right] \]
as well as
\[ E[XX_1] E[X_u] = (\sigma_i \sigma_j \sigma_k)^{-1} E \left[ \left\{ f_i(X_s, Y_s) - \mu_s \right\} \left\{ f_j(X_t, Y_t) - \mu_t \right\} \right] E \left[ f_l(X_u, Y_u) - \mu_u \right]. \]
Using Assumption 6.2, for any \( i, j, l \in [d] \) and \( s \in \mathcal{I}, t \in \mathcal{I}_j, u \in \mathcal{I}_l \)
\[ E[XX_1X_u] \leq (\sigma_i \sigma_j \sigma_k)^{-1} \beta \] and \( E[XX_1] E[X_u] \leq (\sigma_i \sigma_j \sigma_k)^{-1} \beta. \) (6.2)

To take care of the variance terms, we lower bound the variance using Assumption 6.2;
\[
\text{Var}(S_{n,k_i}(f_i)) = \sum_{s \in C_i} \sum_{t \in \mathcal{D}_s(s)} \text{Cov}(f_i(X_s, Y_s), f_i(X_t, Y_t))
\]
\[
\geq \sum_{m=1}^{k_i} \sum_{s \in C_i} \sum_{t \in C_j} \text{Cov}(f_i(X_s, Y_s), f_i(X_t, Y_t))
\]
\[
\geq \alpha_i \kappa_i \left( \frac{n}{2k_i - 1} \right) \left( \frac{2k_i - 1}{k_i} \right) \geq \alpha_i \kappa_i \frac{n^{2k_i - 1}}{(2k_i - 1)^{2k_i - 1}} \frac{2k_i - 1}{k_i}. \]
Here the second-to-last inequality follows by taking only the term for \( m = 1 \). Now we take both sides of the inequality to the power of \(-\frac{1}{2}\) to get that for any \( i \in [d] \)
\[
\sigma_i^{-1} \leq n^{-k_i + \frac{1}{2}} \alpha_i^{-\frac{1}{2}} \left( 2k_i^2 - k_i \right)^{-\frac{1}{2} + \frac{1}{2}}. \] (6.3)

Using Equations (6.1) - (6.3) to bound the quantity \( B_{2.7} \) from Corollary 2.7 we get
\[
B_{2.7} \leq \frac{2}{3} \sum_{i,j,l=1}^{d} (\sigma_i \sigma_j \sigma_k)^{-1} \beta \left( \frac{n}{k_i} \right)^{\min(k_i,k_j)+1} (n - k_i)^{k_j-1}
\]
\[
\left\{ k_i^{\min(k_i,k_j)+1}(n - k_i)^{k_j-1} + k_j^{\min(k_i,k_j)+1}(n - k_i)^{k_i-1} \right\}
\]
\[
\leq \frac{2}{3} \sum_{i,j,l=1}^{d} n^{k_i+k_j-k_l-2} \frac{k_j^{\min(k_i,k_j)+1}}{k_i!} \left\{ k_i^{\min(k_i,k_j)+1} + k_j^{\min(k_i,k_j)+1} \right\} \beta
\]
\[
\left( n^{-k_i-k_j+k_l} \alpha_i \alpha_j \alpha_l \right)^{-\frac{1}{2}} \left( 2k_i^2 - k_i \right)^{-\frac{1}{2} + \frac{1}{2}} \left( 2k_j^2 - k_j \right)^{-\frac{1}{2} + \frac{1}{2}} \left( 2k_l^2 - k_l \right)^{-\frac{1}{2} + \frac{1}{2}}
\]
\[
\leq \frac{2\beta}{3} \sum_{i,j,l=1}^{d} \frac{k_i^{\min(k_i,k_j)+1}}{k_i! \sqrt{\alpha_i \alpha_j \alpha_l}} \left\{ k_i^{\min(k_i,k_j)+1} + k_j^{\min(k_i,k_j)+1} \right\} K_i K_j K_l \right\} n^{-\frac{1}{2}}.
\]

\[ \square \]

6.2. Approximation Theorem with no Variables \( \{ \xi_i \}_{i \in [n]} \). Next we consider the special case that the functions in Definition 6.1 only depend on the second component, so that the sequence \( \{ \xi_i \}_{i \in [n]} \) can be ignored. Continuing to use the same notation, we want to understand the joint distribution of \( S_{n,k_1}(f_1), S_{n,k_2}(f_2), \ldots S_{n,k_d}(f_d) \). However, we add an additional assumption.
ASSUMPTION 6.4. We assume that the functions $f_i$ only depend on the variables $\{Y_{ij}\}$ for $1 \leq i < j \leq n$. That is, we can write $f_i : \mathcal{G}^k \rightarrow \mathbb{R}$.

Such functions appear naturally, for example, when counting subgraphs in an inhomogeneous Bernoulli random graph. A detailed example of such generalised $U$-statistic is worked out in Section 6.3.

In this case, we can adapt the previous theorems slightly and get improved bounds. We still work under Assumption 6.2. The key difference in this case is that the dependency neighbourhoods become smaller: now the subsets need to overlap in at least 2 elements for the corresponding summands to share at least one variable $Y_{ij}$ and hence become dependent. This makes both the variance and the size of dependency neighbourhoods smaller. In the context of Theorem 1.2, the trade-off works out in our favour to give smaller bounds, as follows. For any $s = (\phi, i) \in \mathcal{I}_i$ we set $\mathcal{D}_i(s) = \{ (\psi, j) \in \mathcal{I}_j | |\phi \cap \psi| \geq 2 \}$, so that $W$, under the additional Assumption 6.4, satisfies the assumptions of Corollary 2.7.

In this case, we can adjust Equations (6.1) and (6.3). The proofs are exactly the same as previously, with the only difference being that when we sum over $m$, we start at $m = 2$ as opposed to $m = 1$.

THEOREM 6.5. Consider $W$ that satisfies Assumption 6.4. Let $Z \sim \text{MVN}(0, \text{Id}_{d \times d})$ and $\Sigma$ be the covariance matrix of $W$.

1. Let $h \in \mathcal{H}_d$. Then

$$\left| \mathbb{E} h(W) - \mathbb{E} h(\Sigma^{\frac{1}{2}}Z) \right| \leq |h|_3 B_{6.5} n^{-1};$$

2. Let $\mathcal{K}$ be a class of convex sets in $\mathbb{R}^d$. Then

$$\sup_{A \in \mathcal{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}}Z \in A)| \leq 2^2 3^{-\frac{3}{2}} d^\frac{3}{2} B_{6.5}^{\frac{3}{2}} n^{-\frac{1}{2}}.$$

Here

$$B_{6.5} = \frac{16\beta}{3} \sum_{i,j,l=1}^{d} \frac{k_i^{\min(k_i,k_j)+1}}{k_i! \sqrt{k_i k_j k_l}} \left( k_i^{\min(k_i,k_l)+1} + k_j^{\min(k_j,k_l)+1} \right) K_i K_j K_l$$

and

$$K_i = (2k_i^2 - k_i)^{-\frac{k_i}{2} + \frac{1}{2}}.$$

PROOF. Equation (6.1) becomes

$$|\mathcal{D}_i(s)| \leq k_i^{\min(k_i,k_j)+1} (n - k_i) k_j^{-2}.$$

Equation (6.3) becomes

$$\sigma_i^{-1} \leq 2n^{-k_i+1} \alpha_i^{-\frac{1}{2}} (2k_i^2 - 2k_i)^{-\frac{k_i}{2} + 1}.$$

Using the adjusted bounds in Corollary 2.7 gives the result.

6.3. Approximation Theorem for Simplex Counts. In this section we apply Theorem 6.5 to approximate simplex counts. Consider $G \sim \mathcal{G}(n, p)$. For $1 \leq x < y \leq n$ let $Y_{x,y} := 1 (x \sim y)$ be the edge indicator. In this section we are interested in the $(i+1)$-clique count in $\mathcal{G}(n, p)$ or, equivalently, the $i$-simplex count in $\mathcal{X}(n, p)$, given by

$$T_{i+1} = \sum_{s \subseteq C_{i+1}} \prod_{1 \neq y \in s} Y_{x,y}.$$ (6.4)
Let \( \mathcal{Y}^{i+1} = \{0,1\}^{i+1} \) and let \( f_i : \mathcal{Y}^{i+1} \to \mathbb{R} \) be the function
\[
    f_i(\mathcal{Y}) = \prod_{Y_{xy} \in \mathcal{Y}} Y_{xy}.
\]
Then the associated generalised U-statistic \( S_{n,i+1}(f_i) \) equals the \((i+1)\)-clique count \( T_{i+1} \), as given by Equation (6.4). To apply Theorem 6.5 we need to center and rescale our variables. It is easy to see that \( \mathbb{E}(f_i(\mathcal{Y})) = p^{(i+1)/2} \), if \( \phi \in C_{i+1} \) Just like in Section 6.1, we let \( I_i := C_{i+1} \times \{i\} \) and for \( s = (\phi,i) \in I_i \), we define \( X_s := \sigma^{-1}\left(f_i(\mathcal{Y}) - p^{(i+1)/2}\right) \) and \( W_s = \sum_{s \in I_i} X_s \). Now the vector of interest is \( W = (W_1, W_2, \ldots, W_d) \in \mathbb{R}^d \). This brings us to the next approximation theorem.

**COROLLARY 6.6.** Let \( Z \sim \text{MVN}(0, \text{Id}_{d \times d}) \) and \( \Sigma \) be the covariance matrix of \( W \).

1. Let \( h \in \mathcal{X}_d \). Then
\[
    \left| \mathbb{E}(h(W)) - \mathbb{E}(\Sigma^{1/2} Z) \right| \leq |h| \|B_{6.6} n^{-1}.
\]

2. Let \( \mathcal{X} \) be a class of convex sets in \( \mathbb{R}^d \). Then
\[
    \sup_{A \in \mathcal{X}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{1/2} Z \in A)| \leq 2^{2 + 3 - 3d} B_{6.6}^{1/2} n^{-2}.
\]

**Proof.** Firstly, observe that for any \( \phi, \psi \in C_{i+1} \) for which \( |\phi \cap \psi| \leq 1 \) the covariance vanishes, while if \( |\phi \cap \psi| \geq 2 \) the covariance is non-zero, and we have
\[
    \text{Cov}(f_i(\mathcal{Y}_\phi), f_i(\mathcal{Y}_\psi)) = p^{(i+1)/(i+2)} - p^{(i+1)/2} \leq p^{(i+1)/2}(p^{-1} - 1).
\]

For \( s = (\phi,i) \in I_i \) write \( \hat{X}_s = f_i(\mathcal{Y}_\phi) - p^{(i+1)/2} \). Then by Lemma 2.5 we get:
\[
    \mathbb{E}|\hat{X}_s| \leq \left\{ p^{(i+1)/(i+2)}(1 - p^{(i+1)/2})(1 - p^{(i+1)/2}) \right\}^{1/2};
\]
\[
    \mathbb{E}|\hat{X}_s \hat{X}_u| \leq \left\{ p^{(i+1)/(i+2)}(1 - p^{(i+1)/2})(1 - p^{(i+1)/2}) \right\}^{1/2}.
\]

Since \( \left\{ p^{(i+1)/(i+2)}(1 - p^{(i+1)/2})(1 - p^{(i+1)/2}) \right\}^{1/2} = p(1 - p^{(i+1)/2}) \), we see that Assumption 6.2 holds. Assumption 6.4 also holds and therefore we can apply Theorem 6.5 with \( k_i = i + 1, K_i = (2(i + 1)^2 - 2(i + 1) - 3(i+1)+1, A_i = p^{(i+1)/(i+2)}(p^{-1} - 1), \) and \( B = p(1 - p^{(i+1)/2}) \). Using the bounds \( K_i \leq 1 \) as well as \( 2 \leq k_i^{\min(k_i,k_i)} + 1 = d+1 \), and \( \sqrt{\alpha_i} \geq p^{(i+1)/2} \sqrt{p^{-1} - 1} \) finishes the proof.

**Remark 6.7.** It is easy to show that with high probability there are no large cliques in \( G(n,p) \) for \( p < 1 \) constant. To see this, the expectation of the number of \( k \)-cliques is \( \binom{n}{k} p^k \binom{k}{i} \). By Stirling’s approximation, \( k = \Omega(\ln^{1+\epsilon}(n)) \) for any positive \( \epsilon \) forces the expectation to go to 0 asymptotically. Hence, by Markov’s inequality, with high probability there are no cliques of order \( \ln^{1+\epsilon}(n) \) or larger for any \( \epsilon > 0 \). For cliques of order larger than \( \ln^{1/2}(n) \) and fixed \( p \), a Poisson approximation might be more suitable.

Recall that in Corollary 6.6 the size of the maximal clique we count is \( d+1 \). Note that if \( d = O(\ln^{1/2-\epsilon}(n)) \) for any \( \epsilon > 0 \), then the bounds in Corollary 6.6 tend to 0 as \( n \) tends to infinity as long as \( p \in (0,1) \) stays constant. This might seem quite small but in the light there not being any cliques of order \( \ln^{1+\epsilon}(n) \) with high probability, this is meaningfully large.
Remark 6.8. Note that in Corollary 6.6 we use multivariate normal distribution with covariance $\Sigma$, which is the covariance of $W$ when $n$ is finite and it differs from the limiting covariance, as mentioned in [41]. To approximate $W$ with the limiting distribution, one could proceed in the spirit of [41, Proposition 3] in two steps: use the existing theorems to approximate $W$ with $\Sigma Z$ and then approximate $\Sigma Z$ with $\Sigma_L Z$ where $\Sigma_L$ is the limiting covariance, which is non-invertible, as observed in [23].

Remark 6.9. Corollary 6.6 generalises the result [41, Proposition 2] beyond the case when $d = 2$ and we get a bound of the same order of $n$. [29, Theorem 3.1] considers centered subgraph counts in a random graph associated to a graphon. If we take the graphon to be constant, the associated random graph is just $G(n, p)$. Compared to [29, Theorem 3.1] we place weaker smoothness conditions on our test functions. However, we make use of the special structure of cliques whereas [29, Theorem 3.1] applies to any centered subgraph counts. Translating [29, Theorem 3.1] into a result for uncentered subgraph counts, as we provide here in the special case of clique counts, is not trivial for general $d$.

However, it should be possible to extend our results, using the same abstract approximation theorem, beyond the random clique complex to Linial-Meshulam random complexes [33] or even more general multiparameter Costa-Farber random complexes [13]. We shall consider this conjecture in future work.

References

[1] Robert J Adler, Omer Bobrowski, and Shmuel Weinberger. “Crackle: the homology of noise”. In: Discrete and Computational Geometry (2014), pp. 680–704.

[2] Ryo Asai and Jay Shah. “Algorithmic canonical stratifications of simplicial complexes”. In: Journal of Pure and Applied Algebra 226.9 (2022), p. 107051.

[3] Andrew D Barbour. “Stein’s method for diffusion approximations”. In: Probability Theory and Related Fields 84.3 (1990), pp. 297–322.

[4] Andrew D Barbour, Michal Karoński, and Andrzej Ruciński. “A central limit theorem for decomposable random variables with applications to random graphs”. In: Journal of Combinatorial Theory, Series B 47.2 (1989), pp. 125–145.

[5] Ulrich Bauer and Abhishek Rathod. “Hardness of approximation for Morse matching”. In: SODA ’19: Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms (2019), pp. 2663–2674.

[6] Ulrich Bauer and Abhishek Rathod. “Parameterized inapproximability of Morse matching”. In: arXiv: 2109.04529v1 (2021).

[7] Vidmantas Bentkus. “On the dependence of the Berry–Esseen bound on dimension”. In: Journal of Statistical Planning and Inference 113.2 (2003), pp. 385–402.

[8] Omer Bobrowski and Matthew Kahle. “Topology of random geometric complexes: a survey”. In: Journal of Applied and Computational Topology 1.3 (2018), pp. 331–364.

[9] Omer Bobrowski and Dmitri Krioukov. “Random simplicial complexes: models and phenomena”. In: Higher-Order Systems. Springer, 2022, pp. 59–96.

[10] Gunnar Carlsson et al. “Persistence barcodes for shapes”. In: International Journal of Shape Modeling 11.02 (2005), pp. 149–187.

[11] Sourav Chatterjee and Elizabeth Meckes. “Multivariate normal approximation using exchangeable pairs”. In: Alea 4 (2008), pp. 257–283.

[12] Louis HY Chen, Larry Goldstein, and Qi-Man Shao. Normal Approximation by Stein’s Method. Springer, 2011.

[13] A Costa and M Farber. “Large random simplicial complexes, I”. In: Journal of Topology and Analysis 8.03 (2016), pp. 399–429.
[14] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. Providence, USA: American Mathematical Society, 2010.

[15] Peter Eichelsbacher and Benedikt Rednoß. “Kolmogorov bounds for decomposable random variables and subgraph counting by the Stein-Tikhomirov method”. In: *arXiv:2107.03775* (2021).

[16] Paul Erdős and Alfréd Rényi. “On random graphs I”. In: *Publ. math. debrecen* 6.290-297 (1959), p. 18.

[17] Xiao Fang. “A multivariate CLT for bounded decomposable random vectors with the best known rate”. In: *Journal of Theoretical Probability* 29.4 (2016), pp. 1510–1523.

[18] Robin Forman. “A user’s guide to discrete Morse theory”. In: *Séminaire Lotharingien de Combinatoire* 48 (2002), B48c.

[19] Han L Gan, Adrian Röllin, and Nathan Ross. “Dirichlet approximation of equilibrium distributions in Cannings models with mutation”. In: *Advances in Applied Probability* 49.3 (2017), pp. 927–959.

[20] Robert E Gaunt and Siqi Li. “Bounding Kolmogorov distances through Wasserstein and related integral probability metrics”. In: *arXiv preprint arXiv:2201.12087* (2022).

[21] Robert Ghrist. “Barcodes: the persistent topology of data”. In: *Bull. Amer. Math. Soc. (N.S.)* 45.1 (2008), pp. 61–75.

[22] Gregory Henselman-Petrusek and Robert Ghrist. “Matroid Filtrations and Computational Persistent Homology”. In: *arXiv:1606.00199* (2016).

[23] Svante Janson and Krzysztof Nowicki. “The asymptotic distributions of generalized U-statistics with applications to random graphs”. In: *Probability Theory and Related Fields* 90.3 (1991), pp. 341–375.

[24] Michael Joswig and Marc E Pfetsch. “Computing optimal Morse matchings”. In: *SIAM Journal on Discrete Mathematics* 20.1 (2006), pp. 11–25.

[25] Matthew Kahle. “Topology of random clique complexes”. In: *Discrete mathematics* 309.6 (2009), pp. 1658–1671.

[26] Matthew Kahle. “Random geometric complexes”. In: *Discrete & Computational Geometry* 45.3 (2011), pp. 553–573.

[27] Matthew Kahle. “Sharp vanishing thresholds for cohomology of random flag complexes”. In: *Annals of Mathematics* (2014), pp. 1085–1107.

[28] Mikolaj J Kaspzrak and Giovanni Peccati. “Vector-valued statistics of binomial processes: Berry-Esseen bounds in the convex distance”. In: *arXiv preprint arXiv:2203.13137* (2022).

[29] Gursharn Kaur and Adrian Röllin. “Higher-order fluctuations in dense random graph models”. In: *Electronic Journal of Probability* 26 (2021), pp. 1–36.

[30] Vladimir S Korolyuk and Yu V Borovskich. *Theory of U-statistics*. Vol. 273. Springer Science & Business Media, 2013.

[31] Leon Lampret. “Chain complex reduction via fast digraph traversal”. In: *arXiv:1903.00783* (2019).

[32] J Lee. *U-statistics: Theory and Practice*. Taylor & Francis, 1990.

[33] Nathan Linial and Roy Meshulam. “Homological connectivity of random 2-complexes”. In: *Combinatorica* 26.4 (2006), pp. 475–487.

[34] WG McGinley and Robin Sibson. “Dissociated random variables”. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 77. 1. Cambridge University Press. 1975, pp. 185–188.

[35] Elizabeth Meckes. “On Stein’s method for multivariate normal approximation”. In: *High dimensional probability V: the Luminy volume*. Institute of Mathematical Statistics, 2009, pp. 153–178.

[36] Konstantin Mischaikow and Vidit Nanda. “Morse theory for filtrations and efficient computation of persistent homology”. In: *Discrete & Computational Geometry* 50.2 (2013), pp. 330–353.

[37] Vidit Nanda. “Local Cohomology and Stratification”. In: *Foundations of Computational Mathematics* 20 (2020), pp. 195–222.
Appendix A. Proofs of Lemmas 4.3 and 4.4

Proof of Lemma 4.3. For \( s \in C_{k+1} \) recall that \( s_\_ = s \setminus \{ \min(s) \} \). We write:

\[
Y_s^+ = \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s} Y_{i,j} \right), \quad Y_s^- = \prod_{i=1}^{\min(s)-1} \left( 1 - \prod_{j \in s_\_} Y_{i,j} \right),
\]

\[
Z_s = \prod_{i \notin j \in s} Y_{i,j}, \quad Y_s = Y_s^+ - Y_s^-.
\]

Then \( Z_s \) and \( Y_s \) are independent and \( T_{k+1} = \sum_{s \in C_{k+1}} Z_s Y_s \). Consider the variance:

\[
\text{Var}(T_{k+1}) = \sum_{s \in C_{k+1}} \text{Var}(Z_s Y_s) + \sum_{s \notin t \in C_{k+1}} \text{Cov}(Z_s Y_s, Z_t Y_t)
\]

\[+ \sum_{s \notin t \in C_{k+1}} \text{Cov}(Z_s Y_s, Z_t Y_t). \tag{A.1} \]

For the first term in (A.1), writing

\[
\mathbb{P}(Z_s Y_s = 1) = \mu(i) := p^{(k+1)/2}((1 - p^{k+1})^{i-1} - (1 - p^k)^{i-1})
\]

we see:

\[
\sum_{s \in C_{k+1}} \text{Var}(Z_s Y_s) = \sum_{i=1}^{n} \sum_{s \in C_{k+1}} \left( \mathbb{E} \left\{ (Z_s Y_s)^2 \right\} - \mathbb{E} \left\{ Z_s Y_s \right\}^2 \right)
\]

\[= \sum_{i=1}^{n-k} \binom{n-i}{k} \left( \mathbb{P}(Z_s Y_s = 1) - \mathbb{P}(Z_s Y_s = 1)^2 \right)
\]

\[= \sum_{i=1}^{n-k} \binom{n-i}{k} \left( \mu(i) - \mu(i)^2 \right) = p^{(k+1)/2} V_4.
\]

Now consider the covariance terms in (A.1), the expansion of the variance. Note that for any \( s, t \in C_{k+1} \) if \( s \cap t = \emptyset \), then the variables \( Z_s Y_s \) and \( Z_t Y_t \) can be written as functions of two disjoint sets of independent edge indicators and hence have zero covariance.

Fix \( s, t \in C_{k+1} \) and assume \( |s \cap t| = m \) where \( 1 \leq m \leq k \). Note that because \( m \neq k+1 \), we have \( s \neq t \). There are \( 2^{(k+1)/2} - \binom{m}{2} \) distinct edges in \( s \) and \( t \) combined and hence \( \mathbb{P}(Z_s Z_t = 1) = p^{2^{(k+1)/2} - \binom{m}{2}} \).
Also, $Y_sY_t=Y_s^+Y_t^+ + Y_s^-Y_t^- - Y_s^+Y_t^- - Y_s^-Y_t^+$. For the rest of the proof when calculating probabilities we assume w.l.o.g. that $\min(s) \leq \min(t)$. Then we have for $Y_s^+Y_t^+$:

$$
Y_s^+Y_t^+ = \prod_{i=1}^{\min(t)-1} (1 - \prod_{j \in t} Y_{ij}) \prod_{i=1}^{\min(s)-1} (1 - \prod_{j \in s} Y_{ij})
$$

$$
= \prod_{i=1}^{\min(s)-1} (1 - \prod_{j \in s} Y_{ij}) (1 - \prod_{j \in t} Y_{ij}) \prod_{i=\min(s)}^{\min(t)-1} (1 - \prod_{j \in t} Y_{ij}) \prod_{i=\min(s)}^{\min(s)} (1 - \prod_{j \in s} Y_{ij}).
$$

Fix $i \in [\min(s)-1]$. Then with \(\neg\) denoting the complement

$$
\mathbb{P}[(1 - \prod_{j \in s} Y_{ij}) (1 - \prod_{j \in t} Y_{ij}) = 1] = \mathbb{P}[\neg(\prod_{j \in s} Y_{ij} = 1 \cup \prod_{j \in t} Y_{ij} = 1)]
$$

$$
= 1 - \left\{ \mathbb{P}(\prod_{j \in s} Y_{ij} = 1) + \mathbb{P}(\prod_{j \in t} Y_{ij} = 1) - \mathbb{P}(\prod_{j \in s} Y_{ij} = 1 \cap \prod_{j \in t} Y_{ij} = 1) \right\}
$$

$$
= 1 - (2p^{k+1} - p^{2k+2-m}).
$$

Moreover, $\prod_{i=1}^{\min(s)-1} (1 - \prod_{j \in s} Y_{ij})(1 - \prod_{j \in t} Y_{ij})$ and $\prod_{i=\min(s)}^{\min(t)-1} (1 - \prod_{j \in t} Y_{ij})$ are independent of $Z_sZ_t$.

Recall the notation $[a, b] = \{a, a+1, \ldots, b\}$ for two positive integers $a \leq b$. Setting $q_{s,t} := |s \cap [\min(s), \min(t)-1]|$,

$$
\mathbb{P} \left( Y_s^+Y_t^+ = 1 | Z_sZ_t = 1 \right)
$$

$$
= \mathbb{P} \left( \prod_{i=1}^{\min(s)-1} (1 - \prod_{j \in s} Y_{ij}) (1 - \prod_{j \in t} Y_{ij}) = 1 \right) \mathbb{P} \left( \prod_{i=\min(s)}^{\min(t)-1} (1 - \prod_{j \in t} Y_{ij}) = 1 \right)
$$

$$
= \mathbb{P} \left( \prod_{i=\min(s)}^{\min(t)-1} (1 - \prod_{j \in t} Y_{ij}) = 1 \prod_{i \not= \min(s)} (1 - \prod_{j \in s} Y_{ij}) \prod_{i \not= \min(s)} (1 - \prod_{j \in t} Y_{ij}) = 1 \right)
$$

$$
= (1 - 2p^{k+1} + p^{2k+2-m})^{\min(s)-1} (1 - p^{k+1})^{\min(t)-\min(s)-q_{s,t}} (1 - p^{k+1-m})^{q_{s,t}}.
$$

This strategy of splitting the product $Y_s^+Y_t^+$ into three products of independent variables, only one of which is dependent on $Z_sZ_t$, works exactly in the same way for the variables $Y_s^-Y_t^+, Y_s^+Y_t^-, Y_s^-Y_t^-$. We write $i = \min(s)$, $j = \min(t)$, and $q$ instead of $q_{s,t}$. Also, we set

$$
\pi(i, j, a, b, d_1, d_2, q) := (1 - p^a - p^b + p^{a+b-d_1})^{i-1} (1 - p^a)^{j-i} (1 - p^{a-d_2})^q.
$$

Using the described strategy we get:

$$
\mathbb{P} \left( Y_s^-Y_t^- = 1 | Z_sZ_t = 1 \right) = \pi(i, j, k, k, |s \cap t_-, |s \cap t_-, q)
$$

$$
\mathbb{P} \left( Y_s^+Y_t^- = 1 | Z_sZ_t = 1 \right) = \pi(i, j, k, k+1, |s \cap t_-, |s \cap t_-, q)
$$

$$
\mathbb{P} \left( Y_s^-Y_t^+ = 1 | Z_sZ_t = 1 \right) = \pi(i, j, k+1, k, |s \cap t_+, |m, q).
$$

Now we are ready to calculate the covariance:

$$
\text{Cov}(Z_sY_s, Z_tY_t) = \mathbb{E} \left\{ Z_sZ_tY_s^+Y_t^+ \right\} + \mathbb{E} \left\{ Z_sZ_tY_s^-Y_t^- \right\} - \mathbb{E} \left\{ Z_sZ_tY_s^+Y_t^- \right\}
$$
\[-\mathbb{E} \{ Z_store Y_s Y_t \} - \mathbb{E} \{ Z_s Y_t \} \mathbb{E} \{ Z_t Y_t \} \]
\[= \mathbb{P}(Z_s Z_t = 1) \left\{ \mathbb{P}(Y_s^+ Y_t^- = 1|Z_s Z_t = 1) + \mathbb{P}(Y_s^- Y_t^+ = 1|Z_s Z_t = 1) \right\} - \mathbb{P}(Z_s Y_s = 1) \mathbb{P}(Z_t Y_t = 1) \]
\[= p^{2(k+1)-(2)}(\pi(i, j, k + 1, k + 1, m, m, q) + \pi(i, j, k, k, |s_\cap t_\cap|, |s_\cap t_\setminus|, q)
- \pi(i, j, k, k + 1, |s_\cap t_\setminus|, |s_\cap t_\setminus|, q) - \pi(i, j, k + 1, k, |s_\cap t_\cap, m, q)) \right) - \mu(i)\mu(j). \]

Next we consider the two covariance sums in (A.1) separately. First let us assume that min(s) \neq min(t). Given i, j \in [n - k], m \in [k], and q \in [min(k + 1, |j - i|)] define the set
\[\Gamma_{k+1}(i, j, m, q) = \{ (s, t) | s, t \in C_{k+1}, \min(s) = i, \min(t) = j, |s_\cap t| = m, \max(q_{s,t}, q_{t,s}) = q \} \]
as well as
\[\Gamma_{k+1}^+(i, j, m, q) = \{ (s, t) \in \Gamma_{k+1}(i, j, m, q) | \min(t) \in s \} \]
and
\[\Gamma_{k+1}^-(i, j, m, q) = \{ (s, t) \in \Gamma_{k+1}(i, j, m, q) | \min(t) \notin s \} . \]
Next we argue that
\[|\Gamma_{k+1}^+(i, j, m, q)| = \left( \frac{n - j}{2k + 1 - m - q} \right) \left( \frac{2k + 1 - m - q}{k} \right) \left( \frac{k}{m - 1} \right) \left( \frac{j - i + 1}{q - 1} \right). \]
To see this, assume i \lt j. Note that to pick a pair (s, t) \in \Gamma_{k+1}^+(i, j, m, q) with min(s) = i and min(t) = j we need to pick the 2k - m vertices in s \cup t. Firstly, we pick the vertices that are not included in s \cap [\min(s), \min(t) - 1] = s \setminus [i, j - 1]. Since min(s) \in s \cap [\min(s), \min(t) - 1], this amounts to choosing 2k - m - (q - 1) vertices out of n - j. Then we decide which of the vertices that we have just picked will lie in t. This means we further need to choose k out of 2k + 1 - m - q vertices. Then we choose m - 1 out of k vertices of t to lie in s \cap t (under the assumption that we already have min(t) \in s). Finally, we choose the set s \cap [\min(s), \min(t) - 1], which amounts to picking q - 1 vertices out of j - i + 1 possible choices. If any of the binomial coefficients are negative, we set them to 0. The case j \lt i is analogous.

An analogous argument shows that
\[|\Gamma_{k+1}^-(i, j, m, q)| = \left( \frac{n - j}{2k + 1 - m - q} \right) \left( \frac{2k + 1 - m - q}{k} \right) \left( \frac{k}{m} \right) \left( \frac{j - i + 1}{q - 1} \right). \]

Now using the covariance expression we have just derived, we get
\[\sum_{s \neq t \in C_{k+1}} \text{Cov}(Z_s Y_s, Z_t Y_t) \]
\[= \sum_{i=1}^{n-k} \sum_{j=i+1}^{n-k} \sum_{m=1}^{k} \sum_{q=1}^{\min(k+1,j-i)} \sum_{(s,t) \in \Gamma_{k+1}^+(i, j, m, q)} \text{Cov}(Z_s Y_s, Z_t Y_t) \]
\[+ \sum_{i=1}^{n-k} \sum_{j=i+1}^{n-k} \sum_{m=1}^{k} \sum_{q=1}^{\min(k+1,j-i)} \sum_{(s,t) \in \Gamma_{k+1}^-(i, j, m, q)} \text{Cov}(Z_s Y_s, Z_t Y_t) \]
\[+ \sum_{j=1}^{n-k} \sum_{i=j+1}^{n-k} \sum_{m=1}^{k} \sum_{q=1}^{\min(k+1,j-i)} \sum_{(s,t) \in \Gamma_{k+1}^+(j, i, m, q)} \text{Cov}(Z_s Y_s, Z_t Y_t) \]
\[+ \sum_{j=1}^{n-k} \sum_{i=j+1}^{n-k} \sum_{m=1}^{k} \sum_{q=1}^{\min(k+1,j-i)} \sum_{(s,t) \in \Gamma_{k+1}^-(j, i, m, q)} \text{Cov}(Z_s Y_s, Z_t Y_t) \]
\[+ \sum_{j=1}^{n-k} \sum_{i=j+1}^{n-k} \sum_{m=1}^{k} \sum_{q=1}^{\min(k+1,j-i)} \sum_{(s,t) \in \Gamma_{k+1}^+(j, i, m, q)} \text{Cov}(Z_s Y_s, Z_t Y_t) \]
\[+ \sum_{j=1}^{n-k} \sum_{i=j+1}^{n-k} \sum_{m=1}^{k} \sum_{q=1}^{\min(k+1,j-i)} \sum_{(s,t) \in \Gamma_{k+1}^-(j, i, m, q)} \text{Cov}(Z_s Y_s, Z_t Y_t) \]

REFERENCES
\[
\sum_{i=1}^{n-k} \sum_{j=i+1}^{n-k} \sum_{m=1}^{k \min(k+1-j,i)} \Gamma_{k+1}^{+}(i, j, m, q) \left\{ p^{2(k+1)} - \left( \begin{array}{c} m \\ q \end{array} \right) \right\} \left( \pi(i, j, k + 1, m, m, q) - \pi(i, j, k, m, m, q) \right) - \mu(i) \mu(j) \\
+ \sum_{i=1}^{n-k} \sum_{j=i+1}^{n-k} \sum_{m=1}^{k \min(k+1-j,i)} \sum_{q=1}^{\min(k+1-j-i)} \Gamma_{k+1}^{-}(i, j, m, q) \left\{ p^{2(k+1)} - \left( \begin{array}{c} m \\ q \end{array} \right) \right\} \left( \pi(i, j, k + 1, m, m, q) - \pi(i, j, k, m, m, q) \right) - \mu(i) \mu(j) \\
+ \sum_{j=1}^{n-k} \sum_{i=j+1}^{n-k} \sum_{m=1}^{k \min(k+1-j,i)} \sum_{q=1}^{\min(k+1-j-i)} \Gamma_{k+1}^{+}(j, i, m, q) \left\{ p^{2(k+1)} - \left( \begin{array}{c} m \\ q \end{array} \right) \right\} \left( \pi(j, i, k + 1, m, m, q) - \pi(j, i, k, m, m, q) \right) - \mu(i) \mu(j) \\
+ \sum_{j=1}^{n-k} \sum_{i=j+1}^{n-k} \sum_{m=1}^{k \min(k+1-j,i)} \sum_{q=1}^{\min(k+1-j-i)} \Gamma_{k+1}^{-}(j, i, m, q) \left\{ p^{2(k+1)} - \left( \begin{array}{c} m \\ q \end{array} \right) \right\} \left( \pi(j, i, k + 1, m, m, q) - \pi(j, i, k, m, m, q) \right) - \mu(i) \mu(j) \\
= 2p^{2(k+1)}V_1 + 2p^{2(k+1)}V_2.
\]

Similarly, we calculate the remaining term in the expansion of the variance (A.1). We notice that if \( i = j \), then \( q = 0 \) and we have \( \Gamma_{k+1}(i, i, m, 0) = \Gamma_{k+1}^{+}(i, i, m, 0) \). Hence, \( |\Gamma_{k+1}(i, i, m, 0)| = \left( \begin{array}{c} n-i \\ 2k+1-m \end{array} \right) \left( \begin{array}{c} k \\ m-1 \end{array} \right) \), and

\[
\sum_{s \neq i \in \Gamma_{k+1}} \sum_{m=1}^{k} \sum_{(s,t) \in \Gamma_{k+1}(i,i,m,0)} \text{Cov}(Z_s Y_s, Z_t Y_t) = \sum_{i=1}^{n-k} \sum_{m=1}^{k} \sum_{(s,t) \in \Gamma_{k+1}(i,i,m,0)} \text{Cov}(Z_s Y_s, Z_t Y_t) \\
= \sum_{i=1}^{n-k} \sum_{m=1}^{k} \left( \begin{array}{c} n-i \\ 2k+1-m \end{array} \right) \left( \begin{array}{c} k \\ m-1 \end{array} \right) \left\{ p^{2(k+1)} \left[ (1-2p^{k+1}+p^{2k+2-m})^{i-1} + (1-2p^k+p^{2k+1-m})^{i-1} \\
- 2(1-p^k-p^{k+1}+p^{2k+2-m})^{i-1} \right] - (1-p^{k+1})^{i-1} - (1-p^k)^{i-1})^2 \right\} \\
= p^{2(k+1)}V_3.
\]

\[\square\]

**Proof of Lemma 4.4.** Fix \( 1 \leq k \leq n-1 \) and \( p \in (0,1) \), and consider the variance. From Lemma 4.3 we have \( \text{Var}\{T_{k+1}\} = 2p^{2(k+1)}V_1 + 2p^{2(k+1)}V_2 + p^{2(k+1)}V_3 + p^{2(k+1)}V_4 \). First we lower bound \( V_1 \) and \( V_2 \) by just the negative part of the sum:

\[
V_1 \geq -\sum_{i<j}^{n-k} \sum_{m=1}^{k \min(k+1-j-i)} \sum_{q=1}^{\min(k+1-j-i)} \left( \begin{array}{c} n-j \\ 2k+1-m-q \end{array} \right) \left( \begin{array}{c} 2k+1-m-q \\ k \end{array} \right) \left( \begin{array}{c} k \\ m-1 \end{array} \right) \left( \begin{array}{c} j-i+1 \\ q-1 \end{array} \right)
\]
\[
\left\{ (1 - p^{k+1})^{i+j-2} + (1 - p^k)^{i+j-2} \\
+ p^{-q}(1 - p^{k+1})^{i-1-q}(1 - p^{k+1-m})^q(1 - p^k + p^{2k+1-m})^{i-1} \\
+ p^{-q}(1 - p^k)^{i-1-q}(1 - p^{k-m})^q(1 - p^k + p^{2k+2-m})^{i-1}\right\};
\]
\[
V_2 \geq -\sum_{i<j} \sum_{m=1}^{n-k} \sum_{q=1}^{\min(k+1,j-i)} \left( \frac{n-j}{2k+1-m-q} \right) \left( \frac{2k+1-m}{k} \right) \left( \frac{k}{m} \right) \left( \frac{j-i+1}{q-1} \right)
\]
\[
\left\{ (1 - p^{k+1})^{i+j-2} + (1 - p^k)^{i+j-2} \\
+ p^{-q}(1 - p^{k+1})^{i-1-q}(1 - p^{k+1-m})^q(1 - p^k + p^{2k+1-m})^{i-1} \\
+ p^{-q}(1 - p^k)^{i-1-q}(1 - p^{k-m})^q(1 - p^k + p^{2k+2-m})^{i-1}\right\}.
\]

Now using that \((k^m + (k_{-m}) = (k+1)^m\) and \((1 - p^{k+1}) \geq (1 - p^{k-m})\) for \(m \geq 0\) it is easy to see that \(V_1 + V_2 \geq -4R_1 - 4R_2\), where
\[
R_1 := \sum_{i<j} \sum_{m=1}^{n-k} \sum_{q=1}^{\min(k+1,j-i)} \left( \frac{n-j}{2k+1-m-q} \right) \left( \frac{2k+1-m}{k} \right) \left( \frac{k+1}{m} \right) \left( \frac{j-i+1}{q-1} \right)
\]
\[
(1 - p^{k+1})^{i+j-2};
\]
\[
R_2 := \sum_{i<j} \sum_{m=1}^{n-k} \sum_{q=1}^{\min(k+1,j-i)} \left( \frac{n-j}{2k+1-m-q} \right) \left( \frac{2k+1-m}{k} \right) \left( \frac{k+1}{m} \right) \left( \frac{j-i+1}{q-1} \right)
\]
\[
p^{-q}(1 - p^{k+1})^{j-1}(1 - p^{k+1} - p^k + p^{2k+1-m})^{i-1}.
\]

For \(V_3\) we lower bound by terms with \(m = 1\) and the negative parts of the other terms:
\[
V_3 \geq \sum_{i=1}^{n-k} \left( \frac{n-i}{2k} \right) \left( \frac{2k}{k} \right) \left\{ (1 - 2p^{k+1} + p^{2k+1})^{i-1} - (1 - p^{k+1})^{2i-2} \right\}
\]
\[
- \sum_{i=1}^{n-k} \sum_{m=2}^{k} \left( \frac{n-i}{2k+1-m} \right) \left( \frac{2k+1-m}{k} \right) \left( \frac{k}{m-1} \right)
\]
\[
2p^{-q}(1 - p^k - p^{k+1} + p^{2k+2-m})^{i-1} + 2(1 - p^{k+1})^{2i-2}
\]
\[
= R_4 - R_3;
\]

here we call the positive part of the lower bound \(R_4\) and the negative part \(R_3\). For \(V_4\) we use the trivial lower bound \(V_4 \geq 0\). Hence, we have:
\[
\text{Var}(T_{k+1}) \geq p^{2(k+1)}(R_4 - 8R_1 - 8R_2 - R_3).
\]

Let us now upper bound \(R_1\):
\[
R_1 \leq \sum_{i<j} \sum_{m=1}^{n-k} \sum_{q=1}^{\min(k+1,j-i)} \left( \frac{n-j}{2k+1-m-q} \right) \left( \frac{2k+1-m}{k} \right) \left( \frac{k+1}{m} \right) \left( \frac{j-i+1}{q-1} \right)
\]
\[
\left( 1 - p^{k+1} \right)^{i+j-2}
\]
\[
\leq \sum_{i<j} \sum_{q=1}^{n-k \min(k+1,j-i)} \left( \frac{n^{2k-q}}{k} \right) \left( \frac{2k-1}{k} \right) \left( \frac{k+1}{k} \right) \left( \frac{n^{q-2}}{1} \right) \left( 1 - p^{k+1} \right)^{i+j-2}
\]

for \(1 - p^{k+1} \geq 0\).
$$\leq (k + 1)^{k+1} \frac{n^{2k-1}}{(k-1)!} (2k - 1)^k \sum_{i<j}^\infty (1 - p^{k+1})^{i+j-2}$$

$$= n^{2k-1}(2k - 1)^k (k+1)^{k+1} \frac{1 - p^{k+1}}{(k-1)!} \frac{1}{(2 - p^{k+1}) p^{2k^2+2}}.$$ 

Noting that $(1 - p^{k+1})^{i-i}(1 - p^{k+1} - p^k + p^{2k+1-m})^{i-1} \leq (1 - p^{k+1})^{i-1}$, we can bound $R_2$ in an identical way:

$$R_2 \leq \frac{n^{2k-1}(2k - 1)^k (k+1)^{k+1}}{(k-1)!} p^{-\binom{i}{2}} \sum_{i<j}^\infty (1 - p^{k+1})^{i-1}$$

$$= \frac{n^{2k-1}(2k - 1)^k (k+1)^{k+1}}{(k-1)!} p^{-\binom{i}{2}} \frac{1 - p^{k+1}}{p^{2k^2+2}}.$$ 

Noting that $(1 - p^k - p^{k+1} + p^{2k+2-m})^{i-1} \leq (1 - p^{k+1})^{i-1}$ and $(1 - p^{k+1})^{2i-2} \leq (1 - p^{k+1})^{i-1}$ we proceed to bound $R_3$:

$$R_3 \leq \sum_{i=1}^{n-k} \sum_{m=2}^k \left( \frac{n-i}{2k+1-m} \right) \left( \frac{2k+1-m}{k} \right) \left( \frac{k}{m-1} \right) 2(p^{-\binom{i}{2}} + 1)(1 - p^{k+1})^{i-1}$$

$$\leq \sum_{i=1}^{n-k} \sum_{m=2}^k \frac{n^{2k+1-m}(2k+1-m)^{k} k^{m-1}}{k!} 2(p^{-\binom{i}{2}} + 1)(1 - p^{k+1})^{i-1}$$

$$\leq \sum_{i=1}^{n-k} \frac{k n^{2k+1-m}(2k+1-m)^{k} k^{m-1}}{k!} 2(p^{-\binom{i}{2}} + 1) \sum_{i=1}^\infty (1 - p^{k+1})^{i-1}$$

$$= \frac{n^{2k-1}(2k - 1)^k k^k}{k!} 2(p^{-\binom{i}{2}} + 1) p^{-k-1}.$$ 

To lower bound $R_4$ we just take the $i = 2$ term:

$$R_4 \geq \left( \frac{n-2}{2k} \right) \left( \frac{2k}{k} \right) \left( \frac{k}{k} \right) \left( 1 - 2p^{k+1} + p^{2k+1} \right) - \left( 1 - 2p^{k+1} + p^{2k+2} \right)$$

$$\geq \left( \frac{n-2}{2k} \right) \left( \frac{2k}{k} \right) p^{2k+1}(1 - p).$$ 

Since $R_1, R_2, R_3$ are all at most of the order $n^{2k-1}$ and $R_2$ is at least of the order $n^{2k}$, we have that for any fixed $k \geq 1$ and $p \in (0, 1)$ there exists a constant $C_{p,k} > 0$ independent of $n$ and a natural number $N_{p,k}$ such that for any $n \geq N_{p,k}$:

$$\text{Var}(T_{k+1}) \geq p^{2(k+1)}(R_4 - 8R_1 - 8R_2 - R_3) \geq C_{p,k} n^{2k}.$$ 

$\square$