On Concentration Inequalities for Sampling without Replacement

Shaojie Li[1], Bowei Zhu[2], Yong Liu[2,*]
[1]College of Design and Engineering, National University of Singapore, Singapore
[2]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
li_sj@nus.edu.sg, {bowei.zhu,liuyonggsai}@ruc.edu.cn

Abstract

This paper studies concentration inequalities for sampling without replacement. We introduce a general inequality that provides simple guidelines for proving concentration inequalities, offers a degree of unification of existing results, and is strong enough to yield sharper bounds. As applications, we revisit several standard statistics and either recover classical inequalities through streamlined proofs or strengthen them to tighter bounds, with rates optimal up to constants.

1 Introduction

Concentration inequalities have been considerably developed over the years [19, 29], with major impact across functional analysis, high-dimensional geometry, high-dimensional probability and statistics, information theory, machine learning [25, 26], statistical physics, stochastic analysis, and theoretical computer science [10, 9]. Motivated by their central role, this paper investigates concentration inequalities for sampling without replacement.

Sampling without replacement is a widely used scheme with numerous applications, including stochastic optimization [34, 30], online convex optimization [16], transductive learning [14, 3], confidence sequences [39, 40], and graph neural networks [27, 15]. Compared to the with-replacement scheme, which has been extensively studied, sampling without replacement is far less understood, yet in practice it is very common, often easier to implement, and usually performs better [5].

Existing work on the sampling without replacement can be grouped into Hoeffding-type inequalities [18, 1, 7, 28] and Serfling-type inequalities [33, 6, 32, 2, 36, 14, 12, 37, 8, 17, 41]. Hoeffding-type inequalities typically rely on Hoeffding's reduction lemma (see Lemma 8), which reduce the analysis of sample without replacement to the with-replacement case [18]; as a consequence, the resulting bounds mirror those in the with-replacement setting. By contrast, Serfling-type inequalities are found to yield tighter bounds by incorporating a *finite-sampling correction factor* (see Section 3), indicating that sampling without replacement often results in more concentrated bounds and performs better than the with-replacement case. For this reason, Serfling-type results are preferred in the without-replacement regime.

Within the Serfling line of work mentioned above, the most thoroughly studied case is the sum of real-valued random variables. [33] used a martingale approach to prove a better analogue of the classical Hoeffding's inequality [18], obtaining a Hoeffding-Serfling bound. Building on and refining this martingale analysis, [6] improved Serfling's result, yielding a Hoeffding-Serfling bound and, crucially, established a Bernstein-Serfling inequality, which is especially useful when the variance is small. Subsequently, [17] gave Serfling-type inequalities for hypergeometric distribution—an extreme case of without-replacement sampling—achieving interestingly sharper bounds. [2] provided a Hoeffding-Serfling bound for U-statistics under the martingale analysis of [6]. Recently, [41]

*Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

gave Serfling-type bounds for the suprema of empirical processes, which improves relevant bounds of [12, 37, 28]. Although these inequalities are valuable, focusing primarily on sums limits their applicability to more general functionals [23, 20].

Several works have treated general functions under without-replacement sampling [32, 36, 14, 12, 37, 8]. We will return to these in Section 2.1, as our general inequality also falls into this category. A careful reading of these related works highlights three main challenges. First, existing results for general functionals predominantly yield Hoeffding-type bounds that ignore variance; in many regimes the variance is much smaller than the bounded range underlying Hoeffding's method, motivating Bernstein-type inequalities. Second, one open question has been raised in the literature: [6] asked for Serfling-type variance estimates for sampling without replacement. Third, important cases—such as Bernstein-Serfling bounds for U-statistics—are still missing. In short, there is a continuing need for more powerful and more general theoretical tools.

1.1 Contributions and organization of the paper

This paper addresses these challenges through a general inequality and its applications to standard statistics under sampling without replacement. Our contributions are summarized as follows.

(1) General inequality (Section 2). We present a general concentration inequality for block-symmetric functions of random variables sampled without replacement. The inequality provides a simple guideline for deriving concentration bounds: by choosing a suitable surrogate, one can prove Hoeffding-Serfling type bounds in constant-surrogate cases and derive Bernstein-Serfling type bounds when non-constant surrogates are available. In this sense, it offers a degree of unification while also leading to sharper bounds in several applications, and the resulting rates are optimal up to absolute constants.

(2) Applications (Sections 3–6). In Section 3, we revisit the *sum* functional and derive Hoeffding- and Bernstein-Serfling inequalities, paralleling [18, 6], through a streamlined alternative to martingale-based proofs. In Section 4, we establish Serfling-type inequalities comparing the empirical variance and the population variance, addressing an open question posed by [6]. Combined with the Bernstein-Serfling bound for sums, this yields tighter *empirical Bernstein inequalities*. In Section 5, we derive Hoeffding- and Bernstein-Serfling inequalities for *U-statistics*; the Bernstein-Serfling bound does not follow by a direct extension of the existing martingale analysis (see a discussion in [2]), but follows naturally from our guideline. In Section 6, we obtain Serfling-type inequalities for the *suprema of empirical processes*, comparable to recent bounds of [41], while using a more direct route.

We discuss the optimality of the derived rates in Section 7 and conclude the paper in Section 8. The Appendix collects all proofs.

2 A General Inequality

Let $\mathcal{X} = \{x_1, ..., x_N\}$ be a finite population ($N \geq 2$) with range $[b_1, b_2]$, mean $\mu = \frac{1}{N}\sum_{i=1}^N x_i$, and variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \mu)^2$. We use capital letters for random variables and lower-case letters for their realizations. Sampling without replacement a list $(X_1, ..., X_m)$ of size m from $\mathcal{X}$ can be described sequentially as follows: let first $\mathcal{I}_1 = \{1, ..., N\}$, sample an integer I_1 uniformly on $\mathcal{I}_1$, and set X_1 to be x_{I_1}; then, for each $i = 2, ..., m$, sample I_i uniformly on the remaining indices $\mathcal{I}_i = \mathcal{I}_{i-1} \backslash \{I_{i-1}\}$.

Let S_N be the set of all permutations of the indices $\{1, ..., N\}$. Sampling without replacement is equivalently characterized by random permutations: draw a uniform random permutation $\Pi = \{\Pi_1, ..., \Pi_N\} \sim \text{Unif}(S_N)$ and set $(X_1, ..., X_m) = (x_{\Pi_1}, ..., x_{\Pi_m})$. Any functional of $(X_1, ..., X_m)$ can thus be written as a function of the permutation Π; e.g., $S_m = \sum_{i=1}^m X_i = \sum_{i=1}^m x_{\Pi_i}$. Throughout, probabilities $\mathbb{P}(\cdot)$ and expectations $\mathbb{E}[\cdot]$ are taken with respect to $\Pi \sim \text{Unif}(S_N)$. Unlike sampling with replacement, where the sampled observations are independent, sampling without replacement introduces dependence among the sampled observations. Although both sampling schemes have the same one-dimensional marginals, their joint distributions differ. This dependence is responsible for the *finite-population correction* that appears in Serfling-type concentration bounds.

We study concentration for general functions of samples drawn without replacement. Under the permutation representation, the first m positions form the sampled block and the remaining $u = N -$

m positions form the unsampled block. For $\pi = (\pi_1, \ldots, \pi_N) \in S_N$, the position k identifies a place in the permutation, π_k is the population index occupying that position, and x_{π_k} is the corresponding population observation. Thus, $(x_{\pi_1}, \ldots, x_{\pi_m})$ is the sampled block, while $(x_{\pi_{m+1}}, \ldots, x_{\pi_N})$ is the unsampled block.

Many statistics considered below are invariant to reorderings within each of these two blocks. Moreover, since the unsampled block is the complement of the sampled block, an (m, u)-symmetric function depends only on the sampled subset, or equivalently on its unordered m-element set of population indices, rather than on the ordering within either block. This motivates the following definition.

Definition 1. *Let $u = N - m$. A function $f : S_N \to \mathbb{R}$ is called (m, u)-symmetric if, for every $\pi \in S_N$ and every block-permutation τ that only reorders indices within $\{1, ..., m\}$ and/or within $\{m + 1, ..., N\}$, we have*

$$f(\pi) = f\big(\pi \circ \tau\big),$$

i.e., invariant to any reordering within the first m and/or within the last u coordinates of π.

For later use, let

$$I := \{1, \ldots, m\}, \qquad J := \{m + 1, \ldots, N\}$$

denote the sampled and unsampled position sets, respectively. For $i \in I$ and $j \in J$, let $\pi^{i,j}$ be obtained from π by swapping the entries at positions i and j, i.e.,

$$(\pi^{i,j})_i = \pi_j, \qquad (\pi^{i,j})_j = \pi_i,$$

with all other entries unchanged. At the observation level, this replaces the sampled observation x_{π_i} by the previously unsampled observation x_{π_j}, while x_{π_i} moves to the unsampled block. Thus, "swapping positions i and j" and "replacing one sampled observation by one unsampled observation" refer to the same cross-block operation. Note that (m, u)-symmetry only concerns reorderings within the two blocks; it does not imply invariance under such a cross-block transposition.

The notion of (m, u)-symmetric originates from [13] and is also used in [36] as a systematic review for inequalities of general functions. All instances of f considered in this paper satisfy this (m, u)-symmetry. Theorem 1 below gives a concentration inequality for general functions of random variables sampled without replacement.

Theorem 1. *Let f be an (m, u)-symmetric function, and for any $\pi \in S_N$ define*

$$V_\pi(f) = \sum_{i \in I} \sum_{j \in J} \big(f(\pi) - f(\pi^{i,j})\big)^2 \mathbf{1}\{f(\pi) \geq f(\pi^{i,j})\}.$$

Suppose that there exists a (m, u)-symmetric function $g : S_N \to \mathbb{R}_{\geq 0}$ and a constant $a \geq 0$ such that, for all $\pi \in S_N$,

$$V_\pi(f) \leq g(\pi) \quad \textit{and} \quad V_\pi(g) \leq a^2 g(\pi).$$

Let $\Pi \sim \mathrm{Unif}(S_N)$. Then, for any $t \geq 0$

$$\mathbb{P}\left\{f(\Pi) - \mathbb{E}[f(\Pi)] \geq t\right\} \leq \exp\left\{\frac{-Nt^2}{4\mathbb{E}[g(\Pi)] + 2at}\right\}.$$

Remark. Choice of the surrogate function. There are two directly verifiable choices of g. First, if $V_\pi(f) \leq C$ for all $\pi \in S_N$, one may take $g(\pi) = C$, for which $V_\pi(g) = 0$. Second, if $V_\pi(V_\pi(f)) \leq a^2 V_\pi(f)$ for all $\pi \in S_N$, one may take $g(\pi) = V_\pi(f)$. More generally, one can first compute or upper-bound the local variation $V_\pi(f)$, and then construct a tractable (m, u)-symmetric surrogate g satisfying $V_\pi(f) \leq g(\pi)$ and $V_\pi(g) \leq a^2 g(\pi)$. A useful surrogate should therefore satisfy three criteria: domination, tractability, and closure. A constant surrogate is easy to verify but typically yields a range-based bound, whereas a nonconstant surrogate can retain state-dependent variation and lead to variance-sensitive bounds.

2.1 Comparison with Known Concentration Inequalities

This section reviews related work on general functions of random variables, echoing Section 1. In the product-measure setting, a natural question is: which smoothness assumptions on a function f of variables $X_1, \ldots, X_m$ ensure concentration of $f(X_1, \ldots, X_m)$ around its mean or median [35]? In related work, [14] and [12] introduced a smoothness condition and provided the following result; we state the inequality from [12], since it slightly improves the constants over [14].

Theorem 2 ([12]). *Let f be a (m,u)-symmetric function, and suppose there exists a constant $\beta > 0$ such that $|f(\pi) - f(\pi^{i,j})| \le \beta$ for all π, $i \in \{1,\dots,m\}$, and $j \in \{m+1,\dots,N\}$. Let $\Pi \sim \mathrm{Unif}(S_N)$. Then, for any $t \ge 0$,*

$$\mathbb{P}\{\, f(\Pi) - \mathbb{E}[f(\Pi)] \ge t \,\} \;\le\; \exp\left\{ -\frac{t^2}{2m\beta^2} \left(\frac{N - 1/2}{u} \right) \left(1 - \frac{1}{2\max\{m,u\}} \right) \right\}.$$

This can be viewed as a without-replacement analogue of the bounded differences inequality (i.e., McDiarmid inequality). The condition in Theorem 2 is equivalent to that in [12], written in permutation notation; see Appendix F.

In comparison, [8] established concentration under a weaker smoothness condition:

Theorem 3 ([8]). *Let f be a (m,u)-symmetric function, and suppose there exists a constant $\alpha > 0$ such that $\sum_{i\in I}\sum_{j\in J}\big(f(\pi) - f(\pi^{i,j})\big)^2 \le \alpha$ for all π, $i \in \{1,\dots,m\}$, and $j \in \{m+1,\dots,N\}$. Let $\Pi \sim \mathrm{Unif}(S_N)$. Then, for any $t \ge 0$,*

$$\mathbb{P}\{\, f(\Pi) - \mathbb{E}[f(\Pi)] \ge t \,\} \;\le\; \exp\left\{ -\frac{(N+2)t^2}{4\alpha} \right\}.$$

Theorem 3 is the permutation representation of the concentration bound obtained from Bobkov's logarithmic-Sobolev inequality [8, Theorem 2.1]; see Appendix G for the exact correspondence. If the condition $|f(\pi) - f(\pi^{i,j})| \le \beta$ holds, then $\alpha = mu\beta^2$, and Bobkov's inequality becomes $\mathbb{P}\{\, f(\Pi) - \mathbb{E}[f(\Pi)] \ge t \,\} \;\le\; \exp\left\{ -\frac{(N+2)t^2}{4mu\beta^2} \right\}$. Ignoring constants, this reduces to Theorem 2.

As discussed in [36, Section 4], there are many applications where the bounds of [14, 12] are too weak to be useful, while [8] yield sufficiently strong results. To simplify control of the term $\sum_{i\in I}\sum_{j\in J}\big(f(\pi) - f(\pi^{i,j})\big)^2$, Bobkov's inequality was reformulated as a smoothness requirement on $V_\pi(f)$ in [37, 36] (with [36] providing a survey, not new results):

Theorem 4 ([36]). *Let f be a (m,u)-symmetric function, and suppose there exists a constant $\alpha > 0$ such that $V_\pi(f) \le \alpha$ for all π. Let $\Pi \sim \mathrm{Unif}(S_N)$. Then, for any $t \ge 0$,*

$$\mathbb{P}\{\, f(\Pi) - \mathbb{E}[f(\Pi)] \ge t \,\} \;\le\; \exp\left\{ -\frac{(N+2)t^2}{8\alpha} \right\}.$$

More recently, letting $\pi(m)$ denote the projection of π onto its first m coordinates, [32] proved a bound requiring smoothness only with respect to $\pi(m)$.

Taken together, Sections 1 and 2.1 provide an overview of prior results for sampling without replacement. Echoing to the summary in Section 1, existing concentration bounds in the without-replacement setting remain limited: for general functions they are largely Hoeffding-type, leave key questions unresolved, and fail to cover important cases. We next show how Theorem 1 addresses these limitations and the advantages it offers.

2.2 Merits of the General Concentration Inequality

At a high level, Theorem 1 yields a stronger and flexible concentration inequality under a broader smoothness condition; see below.

(1) A unifying recipe. Theorem 1 provides a simple three-step guideline for deriving concentration bounds in the without-replacement setting: (i) choose the target functional f and construct a surrogate g so that $V_\pi(f) \le g(\pi)$; (ii) bound the sensitivity of g via $V_\pi(g) \le a^2 g(\pi)$ and obtain a; (iii) compute $\mathbb{E}[g(\Pi)]$ and plug it into the theorem. This procedure is transparent and readily applicable across a wide range of problems. Subsequent sections instantiate this mechanism.

(2) Unification of existing results. Theorem 1 subsumes several known inequalities (up to constants). Setting $g(\pi) = \alpha$ gives $a = 0$ and recovers the bound of [37, 36]. Assuming $|f(\pi) - f(\pi^{i,j})| \le \beta$ for all $i \le m$, $j > m$ yields $g(\pi) = mu\beta^2$ and $a = 0$, recovering [14, 12]. Detailed, problem-specific comparisons appear in the sections that follow.

(3) Improvement over existing inequalities. Unlike prior analyses reviewed in Section 2.1, which encode smoothness via fixed parameters (α or β)—effectively treating g as a constant and thereby

suppressing problem-specific structure—the resulting bound depends on the *data-dependent* quantity $\mathbb{E}[g(\Pi)]$. When the concentrated term f admits a *smoother*, non-constant surrogate g with $V_\pi(f) \le g(\pi)$ and $V_\pi(g) \le a^2 g(\pi)$, the bound moves beyond Hoeffding-type control and becomes *Bernstein-type* (variance-sensitive). Crucially, $g(\pi)$ remains (m,u)-symmetric, so the without-replacement structure is preserved; in particular, the pair $(a, \mathbb{E}[g(\Pi)])$ yields *Serfling-type* finite-population improvements. This perspective retains the block-symmetric combinatorial structure of the problem, replaces crude global range parameters by the adaptive quantity $\mathbb{E}[g(\Pi)]$, and thereby delivers tighter bounds under broader smoothness—recovering range-based results as the special case $a = 0$ and strengthening them whenever $\mathbb{E}[g(\Pi)] \ll mu\beta^2$.

2.3 Technical Novelty and Comparison with [31]

Our result is closely related to the two-level concentration framework of [31]. Proposition 1.3 of [31] gives a representative bound. Assume that a probability measure P satisfies a modified log-Sobolev inequality associated with a difference operator Γ, with constant $\rho > 0$. Assume also that the operator satisfies $\Gamma(h^2) \le 2h\Gamma(h)$ for all positive functions h. If there exists a nonnegative auxiliary function q such that $\Gamma(f) \le q$ and $\Gamma(q) \le b$ for some constant $b > 0$, then for any $t \ge 0$,

$$P\{f - \mathbb{E}_P f \ge t\} \le \exp\left\{-\frac{1}{4\rho}\min\left(\frac{t^2}{2\mathbb{E}_P q^2}, \frac{t}{b}\right)\right\}. \tag{1}$$

At a high level, both [31] and our analysis use an entropy/Herbst-type mechanism: one controls a first-order carré-du-champ quantity and then controls a second-order quantity in order to close the concentration argument. The key distinction lies in the form of the second-order control. The work [31] uses a uniformly bounded worst-case control, $\Gamma(q) \le b$, on the auxiliary function q, whereas our approach uses the multiplicative, state-dependent condition $V_\pi(g) \le a^2 g(\pi)$. Under the uniformly bounded control, one can transfer the concentration problem from f to an auxiliary function q and close the argument through a stand-alone estimate for q. By contrast, our proof closes the multiplicative control internally: the entropy inequality is first decoupled into a residual entropy term and a log-Laplace term involving g; the condition $V_\pi(g) \le a^2 g(\pi)$ controls this log-Laplace term, and the residual entropy contribution is then absorbed before the final Herbst integration; see the proof of Theorem 1. A detailed proof roadmap of Theorem 1 and the complete derivation are given in Appendix A.1.

Thus, the technical novelty lies in retaining the state-dependent multiplicative control of local variation and closing the entropy argument through an internal bootstrap, rather than reducing the local variation to a deterministic worst-case bound. On the technical side, we also provide a sharper modified log-Sobolev inequality for cross-block swaps, improving the constant in [8, Theorem 2.1]; see Appendix A.2 for a self-contained proof.

At the level of the final bound, the refined multiplicative control yields sharper constants. Specifically, after the sampling-without-replacement specialization and surrogate alignment detailed in Appendix A.4, (1) gives the min-form constants $(16\mathbb{E}g, 8a)$, whereas Theorem 1 gives the exact Bernstein denominator $4\mathbb{E}g + 2at$, which implies the min-form constants $(8\mathbb{E}g, 4a)$. Thus, our bound improves the constants of [31] by a factor of two; see Appendix A.4. The same argument can also be adapted to many other settings beyond sampling without replacement; see Remark 2.

3 Sums of Random Variables

This section studies concentration for the sum functional $S_m = \sum_{i=1}^m X_i = \sum_{i=1}^m x_{\Pi_i}$, which is (m,u)-symmetric. The mean is $\mu = \frac{1}{N}\sum_{i=1}^N x_i$ and the variance is $\sigma^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \mu)^2$. For $i \in I$ and $j \in J$, a cross-block transposition replaces x_{π_i} in the sampled sum by x_{π_j}, and hence

$$S_m(\pi) - S_m(\pi^{i,j}) = x_{\pi_i} - x_{\pi_j}.$$

Theorem 5. *Suppose that $x_i \in [b_1, b_2]$ for all $i \in [N]$. Then, for any $t \ge 0$,*

$$\mathbb{P}\{S_m - \mathbb{E}S_m \ge t\} \le \exp\left(\frac{-Nt^2}{16um\sigma^2 + 4(b_2 - b_1)\max\{u,m\}t}\right).$$

Theorem 6. *Suppose that $x_i \in [b_1, b_2]$ for all $i \in [N]$. Then, for any $t \geq 0$,*

$$\mathbb{P}\{S_m - \mathbb{E}S_m \geq t\} \leq \exp\left(\frac{-Nt^2}{4mu(b_2 - b_1)^2}\right).$$

Remark. The two bounds above hold under the same assumptions and provide complementary Bernstein-type and Hoeffding-type guarantees. Thus, one can take the minimum of their right-hand sides. The same principle applies to other paired bounds in this paper whenever they concern the same deviation event.

How Serfling-type bounds perform? To illustrate the behavior of Serfling-type bounds, we review related results. [18] showed that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \leq (b_2 - b_1)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}. \tag{2}$$

[33] sharpened the bound (2) by the following inequality: with probability at least $1 - \delta$

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \leq (b_2 - b_1)\sqrt{\frac{(u+1)}{2Nm}\log\frac{1}{\delta}}, \tag{3}$$

which improves the bound (2) by a factor $(u + 1)/N$. As $m \to N$, Serfling's bound improves dramatically, reflecting that without-replacement sampling yields an increasingly accurate estimate of μ. [33] also noted that the factor u/N would be intuitively more natural; the factor u/N is *the finite-sampling correction factor*, as indeed when $m = N$ the mean μ is known exactly, so that the deviation bound is deterministically zero. [6] made this precise: with probability at least $1 - \delta$

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \leq (b_2 - b_1)\sqrt{\frac{\rho_m}{2m}\log\frac{1}{\delta}}, \tag{4}$$

where $\rho_m = \frac{u+1}{N}$ if $m \leq N/2$ and $\rho_m = \frac{u(m+1)}{Nm}$ if $m > N/2$. [6] also established a Bernstein Serfling-type bound, showing that the following holds with probability at least $1 - 2\delta$, if $m \leq N/2$

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \leq \frac{4(b_2 - b_1)}{3m}\log\frac{1}{\delta} + \frac{(b_2 - b_1)}{m}\sqrt{\frac{m(m-1)}{N(u+1)}}\log\frac{1}{\delta} + \sqrt{2\sigma^2\frac{(u+1)}{Nm}\log\frac{1}{\delta}}, \tag{5}$$

and if $m > N/2$

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \leq \frac{4(b_2 - b_1)}{3m}\log\frac{1}{\delta} + \frac{(b_2 - b_1)}{m}\sqrt{\frac{u(u-1)}{N(m+1)}}\log\frac{1}{\delta} + \sqrt{2\sigma^2\frac{u(m+1)}{Nm^2}\log\frac{1}{\delta}}. \tag{6}$$

We now compare with our bounds. Equivalently, Theorem 5 implies that with probability at least $1 - \delta$

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \leq \frac{4(b_2 - b_1)\max\{u, m\}}{Nm}\log\frac{1}{\delta} + \sqrt{\frac{16u\sigma^2}{Nm}\log\frac{1}{\delta}}; \tag{7}$$

and Theorem 6 implies that with probability at least $1 - \delta$

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \leq \sqrt{\frac{4u(b_2 - b_1)^2}{Nm}\log\frac{1}{\delta}}. \tag{8}$$

For the Bernstein-Serfling inequality, our bound (7) gives $\sqrt{(u\sigma^2)/(Nm)}$ in the sub-Gaussian regime ($\sqrt{\log 1/\delta}$), including the *finite-sampling correction factor* and matching the u, m dependence up to constants of (5) and (6); and in the sub-exponential regime ($\log 1/\delta$), the dominant rate in our bound (7) and the bounds (5-6) is $1/m$, again aligning up to constants. A similar conclusion holds for the Hoeffding-Serfling inequality, by comparing the inequality (8) with the bound (4). Overall, under the same sampling setting and boundedness assumption, our bounds match the corresponding Serfling-type rates in [6] up to universal constants. The main contribution is methodological rather than an improvement in concentration rates. Specifically, the bounds in [6] rely on forward/backward Doob martingales and stepwise MGF constructions, whereas our bounds follow directly from the three-step recipe of Theorem 1: identifying g, bounding a, and computing $\mathbb{E}[g(\Pi)]$. This yields a unified derivation without the more elaborate martingale machinery.

4 Empirical Bernstein-Serfling Inequality

This section develops empirical Bernstein inequalities. We first prove Serfling-type deviation bounds for the variance estimator (an essential intermediate step), and then combine them with the Bernstein–Serfling inequality for sums to obtain a data-dependent, variance-sensitive bound.

4.1 Variance Estimate

Let $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^m X_i$ and $\hat{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \hat{\mu}_m)^2$. The unbiased sample variance satisfies

$$\hat{\sigma}_m^2 = \frac{1}{m(m-1)} \sum_{i,j=1}^m \frac{(X_i - X_j)^2}{2} = \frac{1}{m(m-1)} \sum_{i,j=1}^m \frac{(x_{\Pi_i} - x_{\Pi_j})^2}{2}.$$

The functional $\hat{\sigma}_m^2$ is (m, u)-symmetric, so Theorem 1 applies. We first give an upper bound for the population standard deviation, and then provide a reverse-direction control.

Theorem 7. *Suppose that $x_i \in [b_1, b_2]$ for all $i \in [N]$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\sigma \le \sqrt{\frac{N-1}{N}}\hat{\sigma}_m + (b_2 - b_1)\sqrt{\frac{N-1}{N}}\sqrt{\frac{8u}{N(m-1)}\log\frac{1}{\delta}}.$$

Theorem 8. *Suppose that $x_i \in [b_1, b_2]$ for all $i \in [N]$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\hat{\sigma}_m \le \sqrt{\frac{N}{N-1}}\sigma + \sqrt{\frac{2(b_2-b_1)^2 u}{N(m-1)}\log\frac{1}{\delta}}.$$

Discussion. For sampling *with* replacement, [24, Theorem 10] show that, with probability at least $1-\delta$,

$$\sigma \le \hat{\sigma}_m + (b_2 - b_1)\sqrt{\frac{2\log(1/\delta)}{m-1}}. \tag{9}$$

Following the spirit of [24]—namely, avoiding any a priori knowledge of the variance—[6, Lemma 4.1] establish the following inequality with probability at least $1-\delta$

$$\sigma \le \hat{\sigma}_m + (b_2 - b_1)\left(1 + \sqrt{1+\rho_m}\right)\sqrt{\frac{\log(3/\delta)}{2m}}, \tag{10}$$

where $\rho_m = \frac{u+1}{N}$ if $m \le N/2$ and $\rho_m = \frac{u(m+1)}{Nm}$ if $m > N/2$. However, (i) since $(1+\sqrt{1+\rho_m}) \ge 2$, the bound (10) is even weaker than the *with*-replacement bound (9), contradicting the Serfling phenomenon that without-replacement sampling yields sharper concentration [33]; and (ii) since $(1+\sqrt{1+\rho_m})$ is dominated by the constant 1, the bound in (10) is not Serfling-type. Thus [6, Section 5] posed the question: *Can we establish a Serfling-type variance estimate for sampling without replacement?*

Theorem 7 addresses this question. Compared to the *with*-replacement bound (9), Theorem 7 explicitly incorporates the *finite-sampling correction factor* u/N. As $m \to N$, our bound improves dramatically; the term $(b_2-b_1)\sqrt{\frac{N-1}{N}}\sqrt{\frac{8u}{N(m-1)}\log\frac{1}{\delta}}$ vanishes as $u \to 0$, reflecting the Serfling effect. This mirrors how [33] and [6] refine Hoeffding's bound; see the discussion in Section 3.

4.2 Empirical Bernstein Bound

Combining our variance estimate bound with the Bernstein-Serfling inequality for the sum functional and applying a union bound yields an empirical Bernstein-Serfling inequality. In particular, pairing Theorem 7 with Theorem 5 gives the following bound.

Theorem 9. *For any $\delta \in (0,1)$, with probability at least $1-2\delta$*

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \le \frac{4(b_2-b_1)\max\{u,m\}}{Nm}\log\frac{1}{\delta}$$
$$+ \sqrt{\frac{8u}{Nm}\log\frac{1}{\delta}}\left(\sqrt{\frac{N-1}{N}}\hat{\sigma}_m + (b_2-b_1)\sqrt{\frac{N-1}{N}}\sqrt{\frac{8u}{N(m-1)}\log\frac{1}{\delta}}\right).$$

If, instead, we combine Theorem 7 with the bounds (5-6), we obtain the following inequality.

Theorem 10. *For any $\delta \in (0,1)$, the following holds with probability at least $1-3\delta$, if $m \le N/2$*

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \le \frac{4(b_2-b_1)}{3m}\log\frac{1}{\delta} + \frac{(b_2-b_1)}{m}\sqrt{\frac{m(m-1)}{N(u+1)}\log\frac{1}{\delta}}$$
$$+\sqrt{2\frac{(u+1)}{Nm}\log\frac{1}{\delta}}\left(\sqrt{\frac{N-1}{N}}\hat{\sigma}_m + (b_2-b_1)\sqrt{\frac{N-1}{N}}\sqrt{\frac{8u}{N(m-1)}\log\frac{1}{\delta}}\right), \qquad (11)$$

and if $m > N/2$

$$\frac{1}{m}S_m - \frac{1}{m}\mathbb{E}S_m \le \frac{4(b_2-b_1)}{3m}\log\frac{1}{\delta} + \frac{(b_2-b_1)}{m}\sqrt{\frac{u(u-1)}{N(m+1)}\log\frac{1}{\delta}}$$
$$+\sqrt{2\frac{u(m+1)}{Nm^2}\log\frac{1}{\delta}}\left(\sqrt{\frac{N-1}{N}}\hat{\sigma}_m + (b_2-b_1)\sqrt{\frac{N-1}{N}}\sqrt{\frac{8u}{N(m-1)}\log\frac{1}{\delta}}\right). \qquad (12)$$

Because our variance estimate is Serfling–type and sharper, the empirical Bernstein bounds derived from it improve over those in [6].

We next compare with another closely related result. The work [39] also obtained empirical-Bernstein-type bounds for sampling without replacement, but their formulation is structurally different from ours. The closest result is their Corollary 3.2, which gives a fixed-time confidence interval based on a predictable sequence and an associated variance process. Thus, both the center and the radius of their interval are determined by this sequential construction. In contrast, our bounds are centered at the ordinary sample mean and give an explicit closed-form radius, together with a separate variance-estimation inequality. A further difference is that our bounds display the Serfling correction directly. For example, in the exhaustion limit $u = 0$, Theorem 7 reduces to $\sigma \le \sqrt{\frac{N-1}{N}}\,\hat{\sigma}_m = \sigma$, thereby recovering the deterministic endpoint exactly. By comparison, Corollary 3.2 of [39] captures the without-replacement improvement through the sequential scaling term $\sum_{i=1}^{n}\lambda_i\frac{N}{N-i+1}$, rather than through an explicit Serfling-type endpoint correction.

5 U-Statistics

This section studies concentration for U-statistics under sampling without replacement. We consider the following U-statistics of order r with a kernel function $K(\cdot):\mathbb{R}^r \to [b_1,b_2]$ which is symmetric (permutation invariant) in its arguments:

$$U = \frac{1}{\binom{m}{r}}\sum_{1\le i_1<\cdots<i_r\le m} K(X_{i_1},...,X_{i_r}).$$

Define the population mean and variance of the kernel by

$$\mu = \frac{1}{\binom{N}{r}}\sum_{1\le i_1<\cdots<i_r\le N} K(x_{i_1},...,x_{i_r}); \quad \sigma^2 = \frac{1}{\binom{N}{r}}\sum_{1\le i_1<\cdots<i_r\le N}(K(x_{i_1},...,x_{i_r})-\mu)^2.$$

Since $U = \frac{1}{\binom{m}{r}}\sum_{1\le i_1<\cdots<i_r\le m}K(X_{i_1},...,X_{i_r}) = \frac{1}{\binom{m}{r}}\sum_{1\le i_1<\cdots<i_r\le m}K(x_{\Pi_{i_1}},...,x_{\Pi_{i_r}})$ is (m,u)-symmetric, we can apply the general recipe of Section 2.

We first state a Hoeffding-Serfling bound, followed by a Bernstein-Serfling bound.

Theorem 11. *For any $t \ge 0$,*

$$\mathbb{P}\{U-\mu \ge t\} \le \exp\left\{\frac{-Nmt^2}{4ur^2(b_2-b_1)^2}\right\}.$$

Theorem 12. *For any $t \ge 0$,*

$$\mathbb{P}\{U-\mu \ge t\} \le \exp\left\{\frac{-Nmt^2}{8ur^2\sigma^2 + 4(b_2-b_1)(N-2)rt}\right\}.$$

Discussion. In the related work, [1] extend Hoeffding's inequality and Bernstein's inequality for U-statistics to the setting of sampling without replacement. However, their analysis relies on Hoeffding's reduction Lemma and, as a consequence, the resulting bounds mirror those in the with-replacement setting (hence we omit them here). More recently, Theorem 4 of [2] established a Hoeffding-Serfling inequality for U-statistics: For any $t \geq 0$,

$$\mathbb{P}\{U - \mu \geq t\} \leq \exp\left\{\frac{-2(m-r+1)(N-r+1)t^2}{ur^2(1+\frac{1}{m-r+1})(b_2-b_1)^2}\right\}.$$

Although this bound is stated for a possibly nonsymmetric kernel, its formulation is equivalent to our symmetric-kernel representation after symmetrization; see Appendix H for the exact correspondence. Consequently, the relevant assumptions in Theorem 11 and Theorem 4 of [2] are equivalent. Under these equivalent assumptions, by comparison, (i) we give a simple proof of the Hoeffding-Serfling inequality via our three-step guidelines; (ii) our inequality yields a sharper bound in certain parameter regimes, with a different sample-size dependence involving mN instead of $(m-r+1)(N-r+1)$; and (iii) the approach of [2] does not readily extend to a Bernstein-type inequality. A Bernstein-Serfling inequality for U-statistics appears to have been missing from the literature; our framework derives it directly, using the same unified and transparent recipe.

6 Suprema of Empirical Processes

This section studies concentration of the supremum of empirical processes. We introduce some additional notation for this section. Let $\{X_1, ..., X_m\}$ be a sequence of random variables sampled uniformly without replacement from the finite population $\mathcal{X} = \{x_1, ..., x_N\}$. Let $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ satisfy $f(x) \in [-1, 1]$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$, and assume $\mathbb{E}[f(X_1)] = 0$. Define the variance $\sigma^2 = \sup_{f\in\mathcal{F}} \mathbb{V}[f(X_1)] = \sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X_1)]$. The centered supremum under sampling without replacement is $Q_m = \sup_{f\in\mathcal{F}} \sum_{i=1}^m f(X_i)$. Equivalently, if $\Pi \sim \text{Unif}(S_N)$ is a uniform random permutation, then $Q_m = Q_m(\Pi) = \sup_{f\in\mathcal{F}} \sum_{i=1}^m f(x_{\Pi_i})$. Since Q_m is (m, u)-symmetric, we apply the template from Section 2. Define quadratic empirical processes

$$\mathcal{C}_m(f, \Pi) = \mathbb{E}_\Pi\Big[\sup_f \frac{1}{m}\Big(\sum_{i\in I} f(x_{\Pi_i})^2 - \mathbb{E}_\Pi \sum_{i\in I} f(x_{\Pi_i})^2\Big)\Big];$$

and

$$\mathcal{C}_u(f, \Pi) = \mathbb{E}_\Pi\Big[\sup_f \frac{1}{u}\Big(\sum_{j\in J} f(x_{\Pi_j})^2 - \mathbb{E}_\Pi \sum_{j\in J} f(x_{\Pi_j})^2\Big)\Big].$$

With this notation in place, our first concentration inequality is stated below.

Theorem 13. *For any $t \geq 0$*

$$\mathbb{P}\{Q_m - \mathbb{E}Q_m > t\} \leq \exp\left(\frac{-Nt^2}{8mu\sigma^2 + 4mu(C_m(f,\Pi) + C_u(f,\Pi)) + 4\max\{u,m\}t}\right).$$

We then present a concentration inequality for the transductive empirical process, which is proposed in [41, 38] to prove generalization error or excess risk bounds for transductive learning. It is defined as

$$T_{m,u} := \sup_{f\in\mathcal{F}}\left\{\frac{1}{m}\sum_{i=1}^m f(X_i) - \frac{1}{u}\sum_{j=m+1}^N f(X_j)\right\}.$$

Theorem 14. *For any $t \geq 0$*

$$\mathbb{P}\left\{T_{m,u} - \mathbb{E}T_{m,u} > t\right\} \leq \exp\left(\frac{-mut^2}{16N\sigma^2 + 8N(C_m(f,\Pi) + C_u(f,\Pi)) + 4\max\{u,m\}t}\right).$$

In particular, Theorem 13 yields the following inequality, with probability at least $1 - e^{-t}$

$$Q_m - \mathbb{E}Q_m \leq \frac{4\max\{u,m\}}{N}t + \sqrt{\frac{mu}{N}\left(8\sigma^2 t + 4(C_m(f,\Pi) + C_u(f,\Pi))\,t\right.}.$$

This bound exhibits a clear Serfling effect via the *finite-sampling factor correction factor* u/N: as $m \to N$ (i.e., $u \to 0$), the deviation term contracts sharply, showing that without-replacement sampling concentrates more tightly than with replacement—mirroring the way previously discussed results improve the corresponding with-replacement results.

Discussion. Concentration inequalities for the suprema of empirical processes have been widely studied in the literature [41, 12, 37, 28]. We compare our results with the recent work [41], as [41] gives the sharpest bounds to our knowledge. They study concentration for two supremum-type empirical processes under sampling without replacement. Specifically, by aligning with the setting of this section, [41, Theorem 4.1] prove that for any $t \geq 0$ and $\delta \in (0,1)$, with probability at least $1 - e^{-t} - \delta$, the following inequality of transductive empirical process holds

$$T_{m,u} - \mathbb{E}T_{m,u} \lesssim \sqrt{\frac{\sigma^2\, t \log(\frac{1}{\delta})}{\min\{m,u\}}} + \inf_{\alpha>0}\left\{\frac{\mathcal{C}_m(\mathcal{F},\Pi) + \mathcal{C}_u(\mathcal{F},\Pi)}{\alpha} + \frac{\alpha t}{\min\{m,u\}}\right\} + \frac{t}{\min\{m,u\}}.$$

[41, Corollary 4.3] also establish an inequality for the following empirical process $Q_m - \mathbb{E}Q_m$

$$\lesssim \frac{mu}{N}\left(\sqrt{\frac{\sigma^2\, t \log(\frac{1}{\delta})}{\min\{m,u\}}} + \inf_{\alpha>0}\left\{\frac{\mathcal{C}_m(\mathcal{F},\Pi) + \mathcal{C}_u(\mathcal{F},\Pi)}{\alpha} + \frac{\alpha t}{\min\{m,u\}}\right\} + \frac{t}{\min\{m,u\}}\right).$$

By comparison, our Theorems 14 and 13 match the rate of [41, Theorem 4.1; Corollary 4.3] up to absolute constants. The main distinction lies in the proof methodology. Our analysis works directly with the three-step recipe of Theorem 1: identify g, bound a, and compute $\mathbb{E}[g(\Pi)]$, thus offering a streamlined alternative route.

7 Optimality of the Rates

This section demonstrates the optimality of our rates. We examine the sharpness of the rates for the sample mean $\bar{X}_m = S_m/m$ as an example; the same arguments extend verbatim to the other statistics considered.

Let $\Delta = b_2 - b_1$ and $u = N - m$. The Gaussian and Bernstein linear scales are $u\sigma^2/(Nm)$ and $\Delta \max\{u,m\}/(Nm) \asymp \Delta/m$, respectively.

(i) Gaussian scale is tight. For sampling without replacement, $\mathrm{Var}(\bar{X}_m) = u\sigma^2/[m(N-1)]$. Consider an even-sized population with half of its elements equal to b_1 and half equal to b_2. Then $\sigma^2 = \Delta^2/4$, and $\bar{X}_m - \mathbb{E}\bar{X}_m$ is symmetric. Any sub-Gaussian bound with variance proxy $v > 0$ of the form $\mathbb{P}\{\bar{X}_m - \mathbb{E}\bar{X}_m \geq t\} \leq \exp(-t^2/(Cv))$ for all $t \geq 0$ implies, by symmetry and tail integration, $\mathrm{Var}(\bar{X}_m) \leq 2Cv$. Consequently, $v \geq u\Delta^2/[8Cm(N-1)]$, showing that the Gaussian scale is optimal up to universal constants.

(ii) Linear scale is necessary. Consider $m = N-1$ and a population with a fraction $p \in (0, 1/2)$ of its elements equal to b_1 and the remainder equal to b_2, where pN is an integer. Omitting a b_1 observation occurs with probability p and yields $\bar{X}_m - \mathbb{E}\bar{X}_m = (1-p)\Delta/m$, while $\mathrm{Var}(\bar{X}_m) = p(1-p)\Delta^2/m^2$. Suppose a Bernstein-type bound $\mathbb{P}\{\bar{X}_m - \mathbb{E}\bar{X}_m \geq t\} \leq \exp\{-t^2/(C_1 \mathrm{Var}(\bar{X}_m) + C_2 L t)\}$ held uniformly with universal constants $C_1, C_2 > 0$ and $L = o(\Delta/m)$ as $N \to \infty$. Taking $t = (1-p)\Delta/m$ and letting $N \to \infty$ would imply $p \leq \exp(-(1-p)/(C_1 p))$, a contradiction for sufficiently small fixed p. Thus, the linear scale Δ/m cannot be uniformly improved within this Bernstein form.

(iii) Optimality of the two scales. Our Gaussian scale matches the exact variance order, while the Bernstein linear scale $\Delta \max\{u,m\}/(Nm) \asymp \Delta/m$ has the necessary uniform order. In particular, replacing the linear scale by $\Delta u/(Nm)$ would contradict (ii) when $m = N-1$. Together, these establish the sharpness of the two scales for sample-mean concentration up to constants.

8 Conclusion

We developed concentration inequalities for sampling without replacement. Our main result is a general, flexible inequality that unifies and extends prior bounds. As interest grows in analyzing algorithms trained without replacement, our techniques may provide a principled route to tighter and more practical performance bounds.

Limitation. The numerical constants in our concentration bounds may be further improved, and obtaining sharper constants remains an interesting direction for future research.

Acknowledgments and Disclosure of Funding

This work was supported by the National Science Foundation of China (under Key Project No. 92570203). The authors declare no competing interests.

References

[1] J. Ai, O. Kuželka, and Y. Wang. Hoeffding and bernstein inequalities for u-statistics without replacement. *Statistics & Probability Letters*, 187:109528, 2022.

[2] J. Ai, O. Kuželka, and Y. Wang. Hoeffding–serfling inequality for u-statistics without replacement. *Journal of Theoretical Probability*, 36(1):390–408, 2023.

[3] P. Alquier. Transductive and inductive adaptative inference for regression and density estimation. *University Paris*, 6, 2006.

[4] M. Banerjee. Simple random sampling. *Unpublished Manuscript, University of Michigan*, 2012. `https://dept.stat.lsa.umich.edu/~moulib/sampling.pdf`.

[5] R. Bardenet, A. Doucet, and C. Holmes. Towards scaling up markov chain monte carlo: an adaptive subsampling approach. In *International conference on machine learning*, pages 405–413. PMLR, 2014.

[6] R. Bardenet and O.-A. Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 2015.

[7] A. Ben-Hamou, Y. Peres, and J. Salez. Weighted sampling without replacement. *Brazilian Journal of Probability and Statistics*, 32(3):657–669, 2018.

[8] S. G. Bobkov. Concentration of normalized sums and a central limit theorem for noncorrelated random variables. *Annals of probability*, pages 2884–2907, 2004.

[9] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Annals of Probability*, 33:514–560, 2005.

[10] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

[11] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[12] C. Cortes, M. Mohri, D. Pechyony, and A. Rastogi. Stability analysis and learning bounds for transductive regression algorithms. *arXiv preprint arXiv:0904.0814*, 2009.

[13] R. El-Yaniv and D. Pechyony. Stable transductive learning. In *International Conference on Computational Learning Theory*, pages 35–49. Springer, 2006.

[14] R. El-Yaniv and D. Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

[15] P. Esser, L. Chennuru Vankadara, and D. Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. In *Advances in Neural Information Processing Systems*, pages 27043–27056, 2021.

[16] D. Garber, G. Korcia, and K. Levy. Online convex optimization in the random order model. In *International Conference on Machine Learning*, pages 3387–3396. PMLR, 2020.

[17] E. Greene and J. A. Wellner. Exponential bounds for the hypergeometric distribution. *Bernoulli*, 23(3):1911, 2017.

[18] W. Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[19] M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Soc., 2001.

[20] S. Li and Y. Liu. Concentration and moment inequalities for general functions of independent random variables with heavy tails. *Journal of Machine Learning Research*, 25(268):1–33, 2024.

[21] A. Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

[22] A. Maurer. Thermodynamics and concentration. *Bernoulli*, 18(2):434–454, 2012.

[23] A. Maurer. A bernstein-type inequality for functions of bounded interaction. *Bernoulli*, 25(2):1451–1471, 2019.

[24] A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

[25] A. Maurer and M. Pontil. Empirical bounds for functions with weak interactions. In *Conference On Learning Theory*, pages 987–1010. PMLR, 2018.

[26] A. Maurer and M. Pontil. Uniform concentration and symmetrization for weak interactions. In *Conference on Learning Theory*, pages 2372–2387. PMLR, 2019.

[27] K. Oono and T. Suzuki. Optimization and generalization analysis of transduction through gradient boosting and application to multi-scale graph neural networks. *Advances in Neural Information Processing Systems*, 33:18917–18930, 2020.

[28] B. Polaczyk. Concentration bounds for sampling without replacement and hoeffding statistics. *arXiv preprint arXiv:2301.03096*, 2023.

[29] M. Raginsky, I. Sason, et al. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.

[30] I. Safran and O. Shamir. How good is sgd with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.

[31] H. Sambale and A. Sinulis. Modified log-sobolev inequalities and two-level concentration. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, 18(1), 2021.

[32] H. Sambale and A. Sinulis. Concentration inequalities on the multislice and for sampling without replacement. *Journal of Theoretical Probability*, 35(4):2712–2737, 2022.

[33] R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48, 1974.

[34] O. Shamir. Without-replacement sampling for stochastic gradient methods. *Advances in neural information processing systems*, 29, 2016.

[35] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81:73–205, 1995.

[36] I. Tolstikhin. Concentration inequalities for samples without replacement. *Theory of Probability & Its Applications*, 61(3):462–481, 2017.

[37] I. Tolstikhin, G. Blanchard, and M. Kloft. Localized complexities for transductive learning. In *Conference on Learning Theory*, pages 857–884. PMLR, 2014.

[38] I. Tolstikhin, N. Zhivotovskiy, and G. Blanchard. Permutational rademacher complexity: a new complexity measure for transductive learning. In *Algorithmic Learning Theory*, pages 209–223. Springer, 2015.

[39] I. Waudby-Smith and A. Ramdas. Confidence sequences for sampling without replacement. *Advances in Neural Information Processing Systems*, 33:20204–20214, 2020.

[40] I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

[41] Y. Yang. Improved generalization bounds for transductive learning by transductive local complexity and its applications. *arXiv preprint arXiv:2309.16858v4*, 2025.

A Proof for Section 2

A.1 Proof of Theorem 1

We prove Theorem 1 in three coupled steps. First, the modified log-Sobolev inequality and $V_\pi(f) \le g(\pi)$ yield an entropy bound involving the log-MGF of g and a residual entropy term (Step I). Second, the multiplicative self-bounding condition $V_\pi(g) \le a^2 g(\pi)$ provides a log-MGF bound for g in terms of $\mathbb{E}[g(\Pi)]$ (Step II). Third, substituting this bound into the entropy inequality allows us to absorb the residual entropy term, followed by Herbst integration and the exponential Markov inequality to obtain the concentration bound (Step III).

Step I. We first relate the entropy of $e^{\gamma f(\Pi)}$ to the log-MGF of the surrogate g.

Lemma 1. *Let f and $V_\pi(f)$ be as in Theorem 1. Suppose that there exists a (m,u)-symmetric function $g : S_N \to \mathbb{R}_{\ge 0}$ such that, for all $\pi \in S_N$,*

$$V_\pi(f) \le g(\pi).$$

Let $\Pi \sim \mathrm{Unif}(S_N)$. Then for any $\gamma > 0$ and $\theta \ne 0$, we have

$$\mathrm{Ent}(e^{\gamma f(\Pi)}) \le \frac{\theta^{-1}\gamma^2}{N}\left(\mathrm{Ent}(e^{\gamma f(\Pi)}) + \mathbb{E}[e^{\gamma f(\Pi)}]\log\mathbb{E}\left[e^{\theta g(\Pi)}\right]\right),$$

where $\mathrm{Ent}(f) := \mathbb{E}[f\log(f)] - \mathbb{E}[f]\log\mathbb{E}[f]$ for $f > 0$ is the entropy functional.

Proof. The proof builds on the entropy method and the logarithmic Sobolev inequality [19, 11]. By Lemma 3, proved in Appendix A.2, for any (m,u)-symmetric function f, we have the following modified log-Sobolev inequality

$$\begin{aligned}\mathrm{Ent}\left(e^{f(\Pi)}\right) &\le \frac{1}{2N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi) - f(\Pi^{i,j}))(e^{f(\Pi)} - e^{f(\Pi^{i,j})}) \\ &= \frac{1}{N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi) - f(\Pi^{i,j}))(e^{f(\Pi)} - e^{f(\Pi^{i,j})})\mathbb{I}\{f(\Pi) \ge f(\Pi^{i,j})\}, \end{aligned} \tag{13}$$

where as before, $\mathrm{Ent}(e^{f(\Pi)}) = \mathbb{E}[e^{f(\Pi)}\log e^{f(\Pi)}] - \mathbb{E}[e^{f(\Pi)}]\log\mathbb{E}[e^{f(\Pi)}]$, $I = \{1,...,m\}$ and $J = \{m+1,...,N\}$. Since for any $a, b \in \mathbb{R}$ [8],

$$(a-b)(e^a - e^b) \le \frac{e^a + e^b}{2}(a-b)^2,$$

so

$$\begin{aligned}&\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi) - f(\Pi^{i,j}))(e^{f(\Pi)} - e^{f(\Pi^{i,j})})\mathbb{I}\{f(\Pi) \ge f(\Pi^{i,j})\} \\ \le&\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi) - f(\Pi^{i,j}))^2\frac{e^{f(\Pi)} + e^{f(\Pi^{i,j})}}{2}\mathbb{I}\{f(\Pi) \ge f(\Pi^{i,j})\} \\ \le&\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi) - f(\Pi^{i,j}))^2 e^{f(\Pi)}\mathbb{I}\{f(\Pi) \ge f(\Pi^{i,j})\}.\end{aligned}$$

This implies that

$$\mathrm{Ent}\left(e^{f(\Pi)}\right) \le \frac{1}{N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi) - f(\Pi^{i,j}))^2 e^{f(\Pi)}\mathbb{I}\{f(\Pi) \ge f(\Pi^{i,j})\}.$$

Replacing f with γf with $\gamma > 0$ gives

$$\mathrm{Ent}\left(e^{\gamma f(\Pi)}\right) \le \frac{\gamma^2}{N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi) - f(\Pi^{i,j}))^2 e^{\gamma f(\Pi)}\mathbb{I}\{f(\Pi) \ge f(\Pi^{i,j})\}. \tag{14}$$

For any $\theta \neq 0$, we have

$$\begin{aligned}\text{Ent}(e^{\gamma f(\Pi)}) \leq & \frac{\theta^{-1}\gamma^2}{N}\mathbb{E}\theta e^{\gamma f(\Pi)} \sum_{i\in I}\sum_{j\in J}(f(\Pi) - f(\Pi^{i,j}))^2 \mathbb{I}\{f(\Pi) \geq f(\Pi^{i,j})\} \\ \leq & \frac{\theta^{-1}\gamma^2}{N}\mathbb{E}\theta e^{\gamma f(\Pi)} g(\Pi) \\ \leq & \frac{\theta^{-1}\gamma^2}{N}\left(\text{Ent}(e^{\gamma f(\Pi)}) + \mathbb{E}[e^{\gamma f(\Pi)}]\log \mathbb{E}\left[e^{\theta g(\Pi)}\right]\right),\end{aligned}$$

where the second inequality uses the assumption $V_\pi(f) \leq g(\pi)$ for all $\pi \in S_N$ and the last inequality follows from Lemma 5. The proof is complete. □

Step II. Our second inequality develops an upper bound for the logarithm of moment-generating function.

Lemma 2. *Let $g : S_N \to \mathbb{R}_{\geq 0}$ and $V_\pi(g)$ be as in Theorem 1. Suppose that there exists a constant $a \geq 0$ such that for all $\pi \in S_N$*

$$V_\pi(g) \leq a^2 g(\pi).$$

Let $\Pi \sim \text{Unif}(S_N)$. (1) Then for any $\lambda \in (0, N/a^2)$

$$\log \mathbb{E}\left[e^{\lambda(g(\Pi) - \mathbb{E}g(\Pi))}\right] \leq \frac{a^2\lambda^2 \mathbb{E}g(\Pi)}{N - a^2\lambda}.$$

Also for any $t \geq 0$

$$\mathbb{P}\{g(\Pi) - \mathbb{E}g(\Pi) \geq t\} \leq \exp\left(\frac{-Nt^2}{4a^2\mathbb{E}\left[g(\Pi)\right] + 2a^2 t}\right).$$

(2) Then for any $\lambda \in (0, N/a^2)$

$$\log \mathbb{E}\left[e^{\lambda g(\Pi)}\right] \leq \frac{\lambda \mathbb{E}g(\Pi)}{1 - a^2\lambda/N}.$$

Proof. We first prove (1) of Lemma 2. From Eq.(14), we know that for an arbitrary $\lambda > 0$

$$\text{Ent}\left(e^{\lambda g(\Pi)}\right) \leq \frac{\lambda^2}{N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(g(\Pi) - g(\Pi^{i,j}))^2 e^{\lambda g(\Pi)}\mathbb{I}\{g(\Pi) \geq g(\Pi^{i,j})\} \leq \frac{a^2\lambda^2}{N}\mathbb{E}e^{\lambda g(\Pi)} g(\Pi),$$

where the last inequality uses the assumption $V_\pi(g) \leq a^2 g(\pi)$ for all $\pi \in S_N$. By the entropy–log-Laplace identity in Lemma 6, we have

$$\log \mathbb{E}\left[e^{\lambda(g(\Pi) - \mathbb{E}g(\Pi))}\right] = \lambda \int_0^\lambda \frac{\text{Ent}(e^{\gamma g(\Pi)})}{\mathbb{E}[e^{\gamma g(\Pi)}]\gamma^2} d\gamma \leq \frac{a^2}{N}\lambda \int_0^\lambda \frac{\mathbb{E}e^{\gamma g(\Pi)} g(\Pi)}{\mathbb{E}[e^{\gamma g(\Pi)}]} d\gamma = \frac{a^2}{N}\lambda \log \mathbb{E}e^{\lambda g(\Pi)}, \tag{15}$$

where the last identity follows from the fact that $\frac{\mathbb{E}e^{\gamma g} g}{\mathbb{E}[e^{\gamma g}]} = (d/d\gamma)\log \mathbb{E}e^{\gamma g}$. Eq.(15) gives

$$\log \mathbb{E}\left[e^{\lambda(g(\Pi) - \mathbb{E}g(\Pi))}\right] \leq \frac{a^2\lambda}{N}\log \mathbb{E}\left[e^{\lambda(g(\Pi) - \mathbb{E}g(\Pi))}\right] + \frac{a^2\lambda^2}{N}\mathbb{E}g(\Pi),$$

and rearranging this inequality implies that for any $\lambda \in (0, N/a^2)$

$$\log \mathbb{E}\left[e^{\lambda(g(\Pi) - \mathbb{E}g(\Pi))}\right] \leq \frac{a^2\lambda^2 \mathbb{E}g(\Pi)}{N - a^2\lambda}. \tag{16}$$

By Markov's inequality, Eq.(16) implies that for any $t \geq 0$

$$\begin{aligned}\mathbb{P}\{g(\Pi) - \mathbb{E}g(\Pi) \geq t\} \leq & \inf_{0<\lambda<\frac{N}{a^2}} \exp(\log \mathbb{E}[e^{\lambda(g(\Pi) - \mathbb{E}g(\Pi))}] - \lambda t) \\ \leq & \inf_{0<\lambda<\frac{N}{a^2}} \exp\left(\frac{a^2\lambda^2\mathbb{E}g(\Pi)}{N - a^2\lambda} - \lambda t\right) \leq \exp\left(\frac{-Nt^2}{4a^2\mathbb{E}\left[g(\Pi)\right] + 2a^2 t}\right),\end{aligned}$$

where the last inequality uses Lemma 4 (the optimization lemma).

We then prove (2) of Lemma 2. Eq.(15) gives that for any $\lambda \in (0, N/a^2)$

$$\log \mathbb{E}[e^{\lambda g(\Pi)}] \leq \frac{a^2\lambda}{N}\log \mathbb{E}e^{\lambda g(\Pi)} + \lambda \mathbb{E}g(\Pi),$$

and rearranging this inequality establishes the claimed bound. □

Step III. Combining Lemma 1 and Lemma 2, i.e., the entropy's relationship with the logarithm of moment-generating function and the upper bound of the logarithm of moment-generating function, yields a bound on the entropy $\text{Ent}(e^{\gamma f(\Pi)})$. Then, by the entropy's close relationship to the concentration, we prove the inequality of Theorem 1.

Proof. Lemma 1 shows that for any $\gamma > 0$ and $\theta \neq 0$, we have

$$\text{Ent}(e^{\gamma f(\Pi)}) \leq \frac{\theta^{-1}\gamma^2}{N}\left(\text{Ent}(e^{\gamma f(\Pi)}) + \mathbb{E}[e^{\gamma f(\Pi)}]\log\mathbb{E}\left[e^{\theta g(\Pi)}\right]\right),$$

To deal with this inequality, we define $h(\gamma) = \frac{\gamma^2}{N}$. For any $a > 0$ and $\gamma < \frac{N}{a}$ we have

$$a\sqrt{\frac{h(\gamma)}{N}} < 1.$$

Now let $0 < \gamma \leq \lambda < \frac{N}{a}$ and let $\theta = \frac{\gamma}{a}$, so $\theta < \frac{N}{a^2}$ and $\theta > \gamma\sqrt{\frac{h(\gamma)}{N}} = \frac{\gamma^2}{N} = h(\gamma)$. Subtracting $\frac{\theta^{-1}\gamma^2}{N}\text{Ent}(e^{\gamma f})$ from both sides, multiplying by θ, and then using part (2) of Lemma 2 (i.e., $\log\mathbb{E}[e^{\theta g(\Pi)}] \leq \frac{\theta\mathbb{E}g(\Pi)}{1-a^2\theta/N}$), we obtain

$$\text{Ent}(e^{\gamma f(\Pi)})(\theta - \frac{\gamma^2}{N}) \leq \frac{\gamma^2}{N}\mathbb{E}[e^{\gamma f(\Pi)}]\log\mathbb{E}[e^{\theta g(\Pi)}] \leq \frac{\gamma^2}{N}\mathbb{E}[e^{\gamma f(\Pi)}]\frac{\theta}{1-a^2\theta/N}\mathbb{E}[g(\Pi)],$$

which holds since $\theta \in (0, N/a^2)$. Additionally, since $\theta > \frac{\gamma^2}{N}$, we further get

$$\text{Ent}(e^{\gamma f(\Pi)}) \leq \frac{\gamma^2}{N}\mathbb{E}[e^{\gamma f(\Pi)}]\frac{\theta}{(\theta - \frac{\gamma^2}{N})(1-\frac{a^2\theta}{N})}\mathbb{E}[g(\Pi)] = \frac{N\gamma^2}{(N-a\gamma)^2}\mathbb{E}[e^{\gamma f(\Pi)}]\mathbb{E}[g(\Pi)].$$

Together with Lemma 6 gives

$$\begin{aligned}\log\mathbb{E}[e^{\lambda(f(\Pi)-\mathbb{E}f(\Pi))}] &\leq \lambda\mathbb{E}[g(\Pi)]\int_0^\lambda \frac{\frac{N\gamma^2}{(N-a\gamma)^2}}{\gamma^2}d\gamma \\ &= \lambda\mathbb{E}[g(\Pi)]\int_0^\lambda \frac{N}{(N-a\gamma)^2}d\gamma = \lambda\mathbb{E}[g(\Pi)]\frac{\lambda}{N-a\lambda}.\end{aligned}$$

By Markov's inequality, for any $t \geq 0$

$$\begin{aligned}\mathbb{P}\{f(\Pi) - \mathbb{E}f(\Pi) > t\} &\leq \inf_{0<\lambda<\frac{N}{a}} \exp(\log\mathbb{E}[e^{\lambda(f(\Pi)-\mathbb{E}f(\Pi))}] - \lambda t) \\ &\leq \inf_{0<\lambda<\frac{N}{a}} \exp\left(\mathbb{E}[g(\Pi)]\frac{\lambda^2}{N-a\lambda} - \lambda t\right) \\ &\leq \exp\left(\frac{-t^2}{2(\frac{2}{N}\mathbb{E}[g(\Pi)] + \frac{a}{N}t)}\right) \\ &= \exp\left(\frac{-Nt^2}{2(2\mathbb{E}[g(\Pi)] + at)}\right),\end{aligned}$$

where we used Lemma 4 in the last inequality. The proof is complete. □

A.2 Modified Log-Sobolev Inequality on the Slice

The following lemma gives the modified log-Sobolev inequality used in the proof of Theorem 1. We provide a self-contained proof.

Lemma 3 (Modified log-Sobolev inequality)**.** *Let $f : S_N \to \mathbb{R}$ be an (m, u)-symmetric function, where $u = N - m$, and let $\Pi \sim \text{Unif}(S_N)$. Let $I = \{1, \ldots, m\}$ and $J = \{m+1, \ldots, N\}$. Then*

$$\text{Ent}\left(e^{f(\Pi)}\right) \leq \frac{1}{2N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}\left(f(\Pi) - f(\Pi^{i,j})\right)\left(e^{f(\Pi)} - e^{f(\Pi^{i,j})}\right).$$

Proof. To prove the lemma, we first pass from the permutation representation to the fixed-cardinality slice

$$G_{N,m} := \left\{ x \in \{0,1\}^N : \sum_{k=1}^{N} x_k = m \right\}.$$

This representation is natural for sampling without replacement. Indeed, drawing m elements from a population of size N is equivalent to selecting an m-element subset of the population, and each such subset can be represented by its indicator vector in $G_{N,m}$. Thus, the slice is exactly the state space of m-element subsets of the population indices.

The permutation representation used throughout the paper contains the same information, together with an ordering of the selected and unselected elements. For $\pi = (\pi_1, \dots, \pi_N) \in S_N$, define

$$x_k(\pi) := \mathbf{1}\{k \in \{\pi_1, \dots, \pi_m\}\}, \qquad k = 1, \dots, N.$$

Then $x(\pi) \in G_{N,m}$ records which population indices appear in the first m positions of π. If $\Pi \sim \mathrm{Unif}(S_N)$, then $x(\Pi)$ is uniformly distributed on $G_{N,m}$, since every m-element subset has exactly $m!(N-m)!$ permutation preimages.

Moreover, if $f : S_N \to \mathbb{R}$ is (m,u)-symmetric, with $u = N - m$, then f is invariant under reorderings within the first m positions and within the last u positions. Hence $f(\pi)$ depends only on the selected subset encoded by $x(\pi)$, and therefore induces a well-defined function

$$h : G_{N,m} \to \mathbb{R}, \qquad h(x(\pi)) := f(\pi).$$

Under this identification, a cross-block transposition also has a direct interpretation on the slice. If $1 \le i \le m < j \le N$, then π_i is selected and π_j is unselected, and $\pi^{i,j}$ replaces the selected index π_i by the unselected index π_j. In the indicator representation, this is exactly the exchange of one coordinate equal to 1 with one coordinate equal to 0. This correspondence allows us to prove the modified log-Sobolev inequality on the slice and then transfer it back to the permutation representation.

All expectations below are with respect to the uniform measure on $G_{N,m}$. For a positive random variable Z, define

$$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z].$$

For $1 \le a < b \le N$, let x^{ab} exchange coordinates a, b, set $h^{ab}(x) := h(x^{ab})$, and define

$$\mathcal{E}_N(h) := \sum_{a<b} \mathbb{E}\big[(h - h^{ab})(e^h - e^{h^{ab}})\big].$$

If $x_a = x_b$, then $x^{ab} = x$, so the corresponding term vanishes. We prove

$$\mathrm{Ent}(e^h) \le \frac{1}{2N} \mathcal{E}_N(h).$$

For $p + q = 1$ and $r, s > 0$, define

$$\mathrm{Ent}_p(r,s) := pr \log r + qs \log s - (pr + qs) \log(pr + qs).$$

This function is jointly convex. Writing $z = pr + qs$,

$$\partial_{rr} \mathrm{Ent}_p(r,s) = \frac{pqs}{rz}, \qquad \partial_{ss} \mathrm{Ent}_p(r,s) = \frac{pqr}{sz}, \qquad \partial_{rs} \mathrm{Ent}_p(r,s) = -\frac{pq}{z},$$

so its Hessian is positive semidefinite. We claim

$$\mathrm{Ent}_p(r,s) + \mathrm{Ent}_p(s,r) \le pq(r-s)(\log r - \log s).$$

Both sides are unchanged when p, q or r, s are exchanged, so assume $p \le q$ and $r \ge s$. Let $\phi(t) = t \log t$, and for $0 \le t \le p$ define

$$F(t) := \phi(r) + \phi(s) - \phi(tr + (1-t)s) - \phi((1-t)r + ts).$$

Since $p + q = 1$,

$$F(p) = \mathrm{Ent}_p(r,s) + \mathrm{Ent}_p(s,r).$$

Also, $F(0) = 0$ and

$$F'(t) = (r - s) \log \frac{(1-t)r + ts}{tr + (1-t)s}.$$

Set

$$d := \log(r/s), \qquad a := 1 - 2t, \qquad z_1 := tr + (1-t)s, \qquad z_2 := (1-t)r + ts.$$

Since $t \le p \le 1/2$, $a \in [0, 1]$, and

$$\frac{z_2 - z_1}{z_2 + z_1} = a \frac{r - s}{r + s} = a \tanh \frac{d}{2}.$$

Hence

$$\log \frac{z_2}{z_1} = 2 \operatorname{arctanh}\left(a \tanh \frac{d}{2} \right) \le 2a \operatorname{arctanh}\left(\tanh \frac{d}{2} \right) = ad,$$

by convexity of $\operatorname{arctanh}$ and $\operatorname{arctanh}(0) = 0$. Integrating over $[0, p]$ gives

$$F(p) \le (r - s)d \int_0^p (1 - 2t)\, dt = pq(r - s)(\log r - \log s).$$

We now prove the desired inequality by induction on N. The cases $m = 0$ and $m = N$ are immediate since the slice is a singleton. Let $Y = e^h$ and assume the result for slices with $N - 1$ coordinates. For each $c \in [N]$,

$$\operatorname{Ent}(Y) = \mathbb{E}[\operatorname{Ent}(Y \mid x_c)] + \operatorname{Ent}(\mathbb{E}[Y \mid x_c]).$$

Conditionally on $x_c = \varepsilon$, the remaining coordinates are uniform on the $(N - 1, m - \varepsilon)$-slice. Averaging the induction hypothesis over c, each pair $\{a, b\}$ remains for exactly $N - 2$ choices of c, so

$$\frac{1}{N} \sum_{c=1}^{N} \mathbb{E}[\operatorname{Ent}(Y \mid x_c)] \le \frac{N - 2}{2N(N - 1)} \mathcal{E}_N(h).$$

Put

$$p := \frac{m}{N}, \qquad q := \frac{u}{N}, \qquad \alpha_c := \mathbb{E}[Y \mid x_c = 1], \qquad \beta_c := \mathbb{E}[Y \mid x_c = 0].$$

Conditionally on $x_c = 1$, choose B uniformly among the u zero coordinates. Then x^{cB} is uniform conditional on $x_c = 0$: each target has exactly m preimages and

$$\frac{m}{u\binom{N-1}{m-1}} = \frac{1}{\binom{N-1}{m}}.$$

By joint convexity,

$$\operatorname{Ent}_p(\alpha_c, \beta_c) \le \mathbb{E}\left[\operatorname{Ent}_p\big(Y(x), Y(x^{cB})\big) \mid x_c = 1 \right].$$

Moreover,

$$\mathbb{E}[\,\cdot \mid x_c = 1] = \frac{N}{m}\, \mathbb{E}[\mathbf{1}_{\{x_c = 1\}}\, \cdot].$$

Thus summing over c and averaging over B gives the factor $N/(mu)$. For each $a < b$ with $x_a \neq x_b$, the configurations x, x^{ab} form a two-point orbit. Its two orientations give the two terms in the inequality above, while $\mathcal{E}_N(h)$ counts the same symmetric summand twice. Therefore,

$$\sum_{c=1}^{N} \operatorname{Ent}(\mathbb{E}[Y \mid x_c]) \le \frac{N}{mu} \frac{pq}{2} \mathcal{E}_N(h) = \frac{1}{2N} \mathcal{E}_N(h).$$

Dividing by N and combining the two bounds,

$$\operatorname{Ent}(Y) \le \left(\frac{N - 2}{2N(N - 1)} + \frac{1}{2N^2} \right) \mathcal{E}_N(h) \le \frac{1}{2N} \mathcal{E}_N(h).$$

Finally, under the slice–permutation correspondence, $(i, j) \mapsto (\pi_i, \pi_j)$ is a bijection from $I \times J$ to the selected–unselected coordinate pairs, and

$$h(x(\pi^{i,j})) = f(\pi^{i,j}).$$

Hence

$$\operatorname{Ent}\Big(e^{f(\Pi)}\Big) \le \frac{1}{2N}\, \mathbb{E} \sum_{i \in I} \sum_{j \in J} \big(f(\Pi) - f(\Pi^{i,j})\big)\big(e^{f(\Pi)} - e^{f(\Pi^{i,j})}\big),$$

as claimed. □

Remark. Under the slice–permutation correspondence, Bobkov's inequality [8, Theorem 2.1] gives

$$\mathrm{Ent}\Big(e^{f(\Pi)}\Big) \le \frac{1}{N+2}\mathbb{E}\sum_{i\in I}\sum_{j\in J}\big(f(\Pi)-f(\Pi^{i,j})\big)\big(e^{f(\Pi)}-e^{f(\Pi^{i,j})}\big).$$

In contrast, Lemma 3 establishes the sharper coefficient $1/(2N)$. For $N \ge 2$, this improves the constant by a factor of $2N/(N+2)$, which approaches 2 as $N \to \infty$. The improved coefficient follows from the self-contained proof above.

A.3 Auxiliary Lemmas

We introduce some auxiliary lemmas. The first is an optimization lemma.

Lemma 4 (Lemma 12 in [21]). *Let C and b be two positive real numbers, $t > 0$. Then*

$$\inf_{\lambda\in(0,1/b)}\left(-\lambda t+\frac{C\lambda^2}{1-b\lambda}\right) \le \frac{-t^2}{2(2C+bt)}.$$

The next is a decoupling lemma.

Lemma 5 ([22]). *Define the entropy functional* $\mathrm{Ent}(f) := \mathbb{E}[f\log(f)] - \mathbb{E}[f]\log\mathbb{E}[f]$ *(for any positive function f). Then,*

$$\mathbb{E}\left[ge^{\lambda f}\right] \le \mathrm{Ent}\left(e^{\lambda f}\right) + \mathbb{E}\left[e^{\lambda f}\right]\log\mathbb{E}\left[e^{g}\right].$$

The next lemma establishes a connection between the entropy and the log-Laplace transform.

Lemma 6 (Theorem 1 in [22]). *For any $\lambda > 0$*

$$\log\mathbb{E}\left[e^{\lambda(f-\mathbb{E}f)}\right] = \lambda\int_0^\lambda \frac{\frac{\mathbb{E}[\gamma f e^{\gamma f}]}{\mathbb{E}[e^{\gamma f}]} - \log\mathbb{E}[e^{\gamma f}]}{\gamma^2}d\gamma.$$

Lemma 7 (Corollary 9 in [24]). *Suppose $\{x_1,...,x_n\} \subset [0,1]$. Then*

$$\frac{1}{n}\sum_k\left(\frac{1}{n}\sum_j(x_k-x_j)^2\right)^2 \le \frac{1}{2n^2}\sum_{k,j}(x_k-x_j)^2.$$

The next is Hoeffding's reduction lemma, which reduces the analysis of sampling with replacement to the with-replacement case.

Lemma 8 ([18]). *Let $\{U_1,...,I_m\}$ and $\{W_1,...,W_m\}$ be sampled uniformly from a finite set of d-dimensional vectors $\{\mathbf{v}_1,...,\mathbf{v}_N\} \subset \mathbb{R}^d$ with and without replacement, respectively. Then, for any convex function $F : \mathbb{R}^d \to \mathbb{R}$, the following holds*

$$\mathbb{E}\left[F\left(\sum_{k=1}^m W_k\right)\right] \le \mathbb{E}\left[F\left(\sum_{k=1}^m U_k\right)\right].$$

Remark 1. Although Hoeffding's reduction lemma is a simple approach for some basic functionals to prove the concentration inequalities of sampling without replacement, it may fail to yield meaningful bounds for more complex functionals. For example, for average-scale suprema of empirical processes, Theorem 2 and Lemma 3 of [37], which proceed through a with-replacement reduction (i.e., Hoeffding's reduction lemma), yield a bound that is of larger order than the natural fluctuation scale $m^{-1/2}$ once $m \gg N^{2/5}$ and even diverges when $m \gg N^{1/2}$. By contrast, our bounds remain convergent for all sample sizes m. More broadly, our results retain the correct order; for example, when $u = 0$, Theorem 7 reduces to $\sigma \le \sqrt{(N-1)/N}\,\hat{\sigma}_m = \sigma$.

A.4 Details for the Comparison with [31]

We provide the details behind the constant comparison in Section 2.3. We first recall the abstract two-level result of [31], and then specialize it to the sampling-without-replacement setting considered in this paper.

Let P be a probability measure on a state space $\mathcal{X}$, and let Γ be a difference operator acting on functions on $\mathcal{X}$. The convention in [31] is to say that P satisfies a Γ-modified log-Sobolev inequality with constant $\rho > 0$, denoted Γ-mLSI(ρ), if for every function $h : \mathcal{X} \to \mathbb{R}$,

$$\mathrm{Ent}_P(e^h) \le \frac{\rho}{2}\mathbb{E}_P\left[e^h \Gamma(h)^2\right],$$

where

$$\mathrm{Ent}_P(Z) := \mathbb{E}_P[Z \log Z] - \mathbb{E}_P[Z] \log \mathbb{E}_P[Z].$$

A representative two-level bound is Proposition 1.3 of [31]. Assume that P satisfies Γ-mLSI(ρ), and that the difference operator satisfies

$$\Gamma(h^2) \le 2h\Gamma(h)$$

for every positive function h. If there exists a nonnegative auxiliary function q and a constant $b > 0$ such that

$$\Gamma(f) \le q, \qquad \Gamma(q) \le b,$$

then for every $t \ge 0$,

$$P\{f - \mathbb{E}_P f \ge t\} \le \exp\left\{-\frac{1}{4\rho}\min\left(\frac{t^2}{2\mathbb{E}_P q^2}, \frac{t}{b}\right)\right\}.$$

We now specialize this result to sampling without replacement. In this paper, the state space is the permutation space S_N, equipped with the uniform measure $P = \mathrm{Unif}(S_N)$. The elementary move is to swap one coordinate in the sampled block $I = \{1, \dots, m\}$ with one coordinate in the unsampled block $J = \{m+1, \dots, N\}$. For a permutation π, let $\pi^{i,j}$ denote the permutation obtained by swapping the i-th and j-th coordinates. Recall that

$$V_\pi(h) := \sum_{i \in I}\sum_{j \in J}\big(h(\pi) - h(\pi^{i,j})\big)^2 \mathbf{1}\{h(\pi) \ge h(\pi^{i,j})\}.$$

Thus, in the notation of [31], the natural finite difference operator is

$$\Gamma(h)(\pi) := \sqrt{V_\pi(h)}.$$

Equivalently, $\Gamma(h)^2 = V_\pi(h)$.

With this choice of Γ, the modified log-Sobolev inequality used in the proof of Theorem 1 reads

$$\mathrm{Ent}_P(e^h) \le \frac{1}{N}\mathbb{E}_P\left[e^h V_\Pi(h)\right].$$

Since $V_\Pi(h) = \Gamma(h)^2$, this is exactly the Γ-mLSI convention above with

$$\frac{\rho}{2} = \frac{1}{N}, \qquad \text{that is,} \qquad \rho = \frac{2}{N}.$$

This is the meaning of the normalization $\rho = 2/N$ used in Section 2.3.

The chain-rule condition required in Proposition 1.3 is also satisfied by this finite-difference operator. Indeed, for each one-sided swap term, if $h(\pi) \ge h(\pi^{i,j})$ and $h > 0$, then

$$h(\pi)^2 - h(\pi^{i,j})^2 = \big(h(\pi) + h(\pi^{i,j})\big)\big(h(\pi) - h(\pi^{i,j})\big) \le 2h(\pi)\big(h(\pi) - h(\pi^{i,j})\big).$$

After squaring, summing over $i \in I, j \in J$, and taking square roots, this gives

$$\Gamma(h^2)(\pi) \le 2h(\pi)\Gamma(h)(\pi).$$

Therefore, Proposition 1.3 of [31] specializes in the sampling-without-replacement setting to

$$P\{f(\Pi) - \mathbb{E}f(\Pi) \ge t\} \le \exp\left\{-\min\left(\frac{Nt^2}{16\mathbb{E}q(\Pi)^2}, \frac{Nt}{8b}\right)\right\}.$$

The corresponding assumptions are

$$\sqrt{V_\pi(f)} \le q(\pi), \qquad \sqrt{V_\pi(q)} \le b.$$

We now align this bound with Theorem 1. Let g be the surrogate in Theorem 1, so that

$$V_\pi(f) \le g(\pi), \qquad V_\pi(g) \le a^2 g(\pi).$$

Set

$$q := \sqrt{g}.$$

Then

$$\Gamma(f)(\pi) = \sqrt{V_\pi(f)} \le \sqrt{g(\pi)} = q(\pi).$$

It remains to control $\Gamma(q)$. For each swap, write

$$x = g(\pi), \qquad y = g(\pi^{i,j}).$$

On the one-sided event $x \ge y$, if $x > 0$, then

$$(\sqrt{x} - \sqrt{y})^2 = \frac{(x-y)^2}{(\sqrt{x}+\sqrt{y})^2} \le \frac{(x-y)^2}{x}.$$

Applying this inequality to each one-sided swap term gives

$$V_\pi(\sqrt{g}) \le \frac{V_\pi(g)}{g(\pi)} \le a^2.$$

If $g(\pi) = 0$, then $V_\pi(g) \le a^2 g(\pi) = 0$, so the corresponding one-sided terms vanish and the same conclusion holds. Hence

$$\Gamma(q)(\pi) = \sqrt{V_\pi(\sqrt{g})} \le a.$$

Thus the black-box specialization of [31, Proposition 1.3] can be applied with $b = a$. It yields

$$P\{f(\Pi) - \mathbb{E}f(\Pi) \ge t\} \le \exp\left\{-\min\left(\frac{Nt^2}{16\mathbb{E}g(\Pi)}, \frac{Nt}{8a}\right)\right\}.$$

By contrast, Theorem 1 gives the exact Bernstein form

$$P\{f(\Pi) - \mathbb{E}f(\Pi) \ge t\} \le \exp\left\{-\frac{Nt^2}{4\mathbb{E}g(\Pi) + 2at}\right\}.$$

Moreover,

$$\frac{Nt^2}{4\mathbb{E}g(\Pi) + 2at} \ge \min\left(\frac{Nt^2}{8\mathbb{E}g(\Pi)}, \frac{Nt}{4a}\right).$$

Therefore Theorem 1 implies the min-form constants

$$(8\mathbb{E}g(\Pi),\, 4a),$$

whereas the black-box specialization of [31, Proposition 1.3] gives

$$(16\mathbb{E}g(\Pi),\, 8a).$$

Thus, under the same modified log-Sobolev convention and surrogate alignment, the direct multiplicative closure improves the min-form constants of this black-box specialization by a factor of two.

Remark 2 (Beyond the without-replacement setting)**.** The constant comparison above is not specific to the particular sampling-without-replacement application. The same argument applies whenever the difference operator Γ is compatible with the elementary square-root contraction

$$\Gamma(\sqrt{g}) \le a \qquad \text{whenever} \qquad \Gamma(g)^2 \le a^2 g.$$

This condition is satisfied by the usual finite-difference carré-du-champ-type operators built from sums of squared local differences, including product-space replacement differences, permutation or swap differences, and the block-swap operator used in this paper. It is also satisfied, with the standard chain-rule interpretation, by diffusion carré-du-champ operators. Therefore, in these settings, the same direct multiplicative closure can be used to sharpen the min-form constants obtained by the Proposition 1.3 of [31]. We do not pursue these routine extensions here.

B Proofs for Section 3

B.1 Proof of Theorem 5

Proof. We proceed in three steps.

Step 1 (Choose f and construct g with $V_\pi(f) \le g(\pi)$). Let

$$f(\pi) \;=\; S_m(\pi) \;=\; \sum_{i\in I} x_{\pi_i}, \qquad I=\{1,\dots,m\},\; J=\{m+1,\dots,N\}.$$

Then

$$V_\pi(f) = \sum_{i\in I}\sum_{j\in J}\big(x_{\pi_i}-x_{\pi_j}\big)^2\,\mathbb{I}\{x_{\pi_i}\ge x_{\pi_j}\} \;\le\; 2u\sum_{i\in I}(x_{\pi_i}-\mu)^2 \;+\; 2m\sum_{j\in J}(x_{\pi_j}-\mu)^2.$$

Hence we may take the (m,u)-symmetric surrogate

$$g(\pi) \;=\; 2u\sum_{i\in I}(x_{\pi_i}-\mu)^2 \;+\; 2m\sum_{j\in J}(x_{\pi_j}-\mu)^2,$$

for which $V_\pi(f)\le g(\pi)$ holds for all $\pi\in S_N$.

Step 2 (Sensitivity of g: bound $V_\pi(g)\le a^2 g(\pi)$). A direct computation gives

$$\begin{aligned}
V_\pi(g) =& \sum_{i\in I}\sum_{j\in J}(g(\pi)-g(\pi^{i,j}))^2\mathbb{I}\{g(\pi)\ge g(\pi^{i,j})\}\\
\le& 4\sum_{i\in I}\sum_{j\in J}\Big(u\sum_{k\in I}(x_{\pi_k}-\mu)^2+m\sum_{l\in J}(x_{\pi_l}-\mu)^2-u\sum_{k\in I}(x_{\pi_k^{i,j}}-\mu)^2-m\sum_{l\in J}(x_{\pi_l^{i,j}}-\mu)^2\Big)^2\\
=& 4\sum_{i\in I}\sum_{j\in J}\big(u(x_{\pi_i}-\mu)^2-u(x_{\pi_j}-\mu)^2+m(x_{\pi_j}-\mu)^2-m(x_{\pi_i}-\mu)^2\big)^2\\
\le& 4\sum_{i\in I}\sum_{j\in J}\big(u(x_{\pi_i}-\mu)^2+m(x_{\pi_j}-\mu)^2\big)^2\\
\le& 8\Big(\sum_{i\in I}\sum_{j\in J}u^2(x_{\pi_i}-\mu)^4+m^2(x_{\pi_j}-\mu)^4\Big)\\
\le& 8(b_2-b_1)^2\Big(\sum_{i\in I}\sum_{j\in J}u^2(x_{\pi_i}-\mu)^2+m^2(x_{\pi_j}-\mu)^2\Big)\\
=& 8(b_2-b_1)^2\Big(u^3\sum_{i\in I}(x_{\pi_i}-\mu)^2+m^3\sum_{j\in J}(x_{\pi_j}-\mu)^2\Big)\\
\le& 4(b_2-b_1)^2\max\{u^2,m^2\}g(\pi).
\end{aligned}$$

Thus we can take

$$a^2 \;=\; 4(b_2-b_1)^2\;\max\{u^2,m^2\}.$$

Step 3 (Compute $\mathbb{E}[g(\Pi)]$ and plug into Theorem 1). By symmetry of the without-replacement model,

$$\mathbb{E}\left[\sum_{i\in I}(x_{\Pi_i}-\mu)^2\right] \;=\; m\,\sigma^2, \qquad \mathbb{E}\left[\sum_{j\in J}(x_{\Pi_j}-\mu)^2\right] \;=\; u\,\sigma^2,$$

hence

$$\mathbb{E}[g(\Pi)] = 2u\cdot m\sigma^2 + 2m\cdot u\sigma^2 = 4um\sigma^2.$$

Applying Theorem 1 with the above g and a yields, for any $t\ge 0$,

$$\begin{aligned}
\mathbb{P}\{\,S_m-\mathbb{E}S_m\ge t\,\} \;&\le\; \exp\!\left(-\frac{Nt^2}{4\cdot(4um\sigma^2)\;+\;2\cdot a\,t}\right)\\
&= \exp\!\left(-\frac{Nt^2}{16um\sigma^2\;+\;4(b_2-b_1)\max\{u,m\}\,t}\right),
\end{aligned}$$

which is the stated Bernstein-Serfling bound. □

B.2 Proof of Theorem 6

Proof. We proceed in three steps.

Step 1 (Choose f and a *constant* surrogate g). For the same $f(\pi) = S_m(\pi)$, we have the range bound

$$V_\pi(f) = \sum_{i\in I}\sum_{j\in J}\big(x_{\pi_i} - x_{\pi_j}\big)^2\,\mathbb{I}\{x_{\pi_i} \geq x_{\pi_j}\} \;\leq\; mu\,(b_2 - b_1)^2.$$

Set the (m,u)-symmetric surrogate to the constant

$$g(\pi) \;=\; mu\,(b_2 - b_1)^2.$$

Step 2 (Sensitivity of g). Since g is constant, $V_\pi(g) = 0$, hence we may take $a = 0$.

Step 3 (Compute $\mathbb{E}[g(\pi)]$ and apply Theorem 1). Here $\mathbb{E}[g(\Pi)] = mu(b_2 - b_1)^2$. Applying Theorem 1 with $a = 0$ gives, for any $t \geq 0$,

$$\mathbb{P}\{\, S_m - \mathbb{E}S_m \geq t\,\} \;\leq\; \exp\!\left(-\frac{Nt^2}{4\,mu\,(b_2 - b_1)^2}\right).$$

The proof is complete. □

C Proofs for Section 4

C.1 Proof of Theorem 7

We split the proof into three steps.

Step I. The proof of Theorem 7 requires a new inequality.

Lemma 9. *Let $f : S_N \to \mathbb{R}_{\geq 0}$ be a (m,u)-symmetric function, and let $V_\pi(f)$ be as in Theorem 1. Suppose that there exists a constant $a \geq 0$ such that for all $\pi \in S_N$, f satisfies*

$$V_\pi(f) \leq a^2 f(\pi) \quad \textit{and} \quad f(\pi) - f(\pi^{i,j}) \leq 1.$$

Let $\Pi \sim \mathrm{Unif}(S_N)$. Then for any $\lambda > 0$

$$\log \mathbb{E}\left[e^{\lambda(\mathbb{E}f(\Pi) - f(\Pi))}\right] \leq \frac{2a^2\lambda^2}{N}\mathbb{E}f(\Pi).$$

Also for any $t \geq 0$

$$\mathbb{P}\left\{\mathbb{E}[f(\Pi)] - f(\Pi) \geq t\right\} \leq \exp\left(\frac{-Nt^2}{8a^2\mathbb{E}[f(\Pi)]}\right).$$

Proof. From Eq.(13), we know

$$\begin{aligned}
\mathrm{Ent}\left(e^{f(\Pi)}\right) &\leq \frac{1}{2N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi) - f(\Pi^{i,j}))(e^{f(\Pi)} - e^{f(\Pi^{i,j})}) \\
&= \frac{1}{2N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi^{i,j}) - f(\Pi))(e^{f(\Pi^{i,j})} - e^{f(\Pi)}) \\
&= \frac{1}{N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi^{i,j}) - f(\Pi))(e^{f(\Pi^{i,j})} - e^{f(\Pi)})\mathbb{I}\{f(\Pi^{i,j}) \geq f(\Pi)\},
\end{aligned}$$

where as before, $\mathrm{Ent}(f) = \mathbb{E}[f\log(f)] - \mathbb{E}[f]\log\mathbb{E}[f]$, $I = \{1,...,m\}$ and $J = \{m+1,...,N\}$.

Applying this inequality to $-\lambda f$ with $\lambda > 0$ yields

$$\begin{aligned}
&\text{Ent}\left(e^{-\lambda f(\Pi)}\right)\\
\leq&\frac{\lambda}{N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi)-f(\Pi^{i,j}))(e^{\lambda(f(\Pi)-f(\Pi^{i,j}))}-1)e^{-\lambda f(\Pi)}\mathbb{I}\{f(\Pi)\geq f(\Pi^{i,j})\}\\
\leq&\frac{2\lambda^2}{N}\mathbb{E}\sum_{i\in I}\sum_{j\in J}(f(\Pi)-f(\Pi^{i,j}))^2e^{-\lambda f(\Pi)}\mathbb{I}\{f(\Pi)\geq f(\Pi^{i,j})\}\\
\leq&\frac{2a^2\lambda^2}{N}\mathbb{E}e^{-\lambda f(\Pi)}f(\Pi),
\end{aligned}$$

where the second inequality uses the property $e^x - 1 \leq 2x$ for all $x \in [0, 1]$ and the facts $f(\pi) - f(\pi^{i,j}) \leq 1$ and $\mathbb{I}\{f(\pi) \geq f(\pi^{i,j})\} \geq 0$ for all $\pi \in S_N$, and where the third inequality uses the assumption $V_\pi(f) \leq a^2 f(\pi)$ for all $\pi \in S_N$. Since the covariance of $f(\Pi)e^{-\lambda f(\Pi)}$ is non-positive (i.e., $\mathbb{E}f(\Pi)e^{-\lambda f(\Pi)} \leq \mathbb{E}f(\Pi)\mathbb{E}e^{-\lambda f(\Pi)}$), we get

$$\frac{\text{Ent}\left(e^{-\lambda f(\Pi)}\right)}{\mathbb{E}e^{-\lambda f(\Pi)}} \leq \frac{2a^2\lambda^2}{N}\mathbb{E}f(\Pi).$$

Thus, by Lemma 6, we have for any $\lambda > 0$

$$\log\mathbb{E}\left[e^{\lambda(\mathbb{E}[f(\Pi)]-f(\Pi))}\right] = \lambda\int_0^\lambda \frac{\text{Ent}\left(e^{-\gamma f(\Pi)}\right)}{\mathbb{E}e^{-\gamma f(\Pi)}\gamma^2}d\gamma \leq \frac{2a^2\lambda^2}{N}\mathbb{E}f(\Pi).$$

By Markov's inequality, setting $\lambda = \frac{Nt}{4a^2\mathbb{E}f(\Pi)}$, we obtain

$$\begin{aligned}
\mathbb{P}\{\mathbb{E}[f(\Pi)]-f(\Pi)\geq t\} &\leq \inf_{\lambda>0}\exp(\log\mathbb{E}[e^{\lambda(\mathbb{E}[f(\Pi)]-f(\Pi))}]-\lambda t)\\
&\leq \inf_{\lambda>0}\exp\left(\frac{2a^2\lambda^2}{N}\mathbb{E}[f(\Pi)]-\lambda t\right) \leq \exp\left(\frac{-Nt^2}{8a^2\mathbb{E}\left[f(\Pi)\right]}\right).
\end{aligned}$$

The proof is complete. □

Step II. Note that $\hat{\sigma}_m^2 = \frac{1}{m(m-1)}\sum_{i,j=1}^m \frac{(X_i-X_j)^2}{2} = \frac{1}{m(m-1)}\sum_{i,j=1}^m \frac{(x_{\Pi_i}-x_{\Pi_j})^2}{2}$ is (m, u)-symmetric; hence Lemma 9 applies.

Lemma 10. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$*

$$\sqrt{\mathbb{E}\hat{\sigma}_m^2} \leq \sqrt{\hat{\sigma}_m^2} + \sqrt{\frac{8(b_2-b_1)^2u}{N(m-1)}\log\frac{1}{\delta}}.$$

Proof. Let $f(\Pi) = \frac{m}{(b_2-b_1)^2}\hat{\sigma}_m^2$ with $\hat{\sigma}_m^2 = \frac{1}{m(m-1)}\sum_{i,j=1}^m \frac{(x_{\Pi_i}-x_{\Pi_j})^2}{2}$. A direct calculation gives for all $\pi \in S_N$

$$\begin{aligned}
f(\pi)-f(\pi^{i,j}) =&\frac{1}{(b_2-b_1)^2}\frac{1}{m-1}\sum_{k=1}^m((x_{\pi_i}-x_{\pi_k})^2-(x_{\pi_j}-x_{\pi_k})^2)\\
\leq&\frac{1}{(b_2-b_1)^2}\frac{1}{m-1}\sum_{k=1}^m(x_{\pi_i}-x_{\pi_k})^2 \leq 1,
\end{aligned} \tag{17}$$

and then

$$
\begin{aligned}
V_\pi(f) &= \sum_{i\in I}\sum_{j\in J}(f(\pi)-f(\pi^{i,j}))^2\mathbb{I}\{f(\pi)\geq f(\pi^{i,j})\}\\
&\leq \sum_{i\in I}\sum_{j\in J}\left(\frac{1}{(b_2-b_1)^2}\frac{1}{m-1}\sum_{k=1}^{m}(x_{\pi_i}-x_{\pi_k})^2\right)^2\\
&= \sum_{i\in I}\sum_{j\in J}\left(\frac{1}{m-1}\sum_{k=1}^{m}(\frac{1}{b_2-b_1}x_{\pi_i}-\frac{1}{b_2-b_1}x_{\pi_k})^2\right)^2\\
&\leq \frac{um^3}{(m-1)^2}\frac{1}{2m^2}\frac{1}{(b_2-b_1)^2}\sum_{i,k}(x_{\pi_i}-x_{\pi_k})^2\\
&= \frac{mu}{m-1}f(\pi),
\end{aligned}
\tag{18}
$$

where the first inequality uses (17) and the second inequality uses Lemma 7. Thus, plugging $a^2=\frac{mu}{m-1}$ into Lemma 9 gives

$$
\mathbb{P}\{\mathbb{E}[f(\Pi)]-f(\Pi)\geq t\}\leq \exp\left(\frac{-N(m-1)t^2}{8mu\mathbb{E}\left[f(\Pi)\right]}\right),
$$

which means that

$$
\mathbb{P}\left\{\mathbb{E}\frac{m}{(b_2-b_1)^2}\hat{\sigma}_m^2-\frac{m}{(b_2-b_1)^2}\hat{\sigma}_m^2\geq t\right\}\leq \exp\left(\frac{-N(m-1)t^2}{8mu\mathbb{E}\left[\frac{m}{(b_2-b_1)^2}\hat{\sigma}_m^2\right]}\right).
$$

Equivalently, for any $\delta\in(0,1)$

$$
\mathbb{P}\left\{\mathbb{E}\hat{\sigma}_m^2-\hat{\sigma}_m^2\geq\sqrt{\frac{8(b_2-b_1)^2u\mathbb{E}\hat{\sigma}_m^2}{N(m-1)}\log\frac{1}{\delta}}\right\}\leq\delta.
$$

Completing the square on the left hand side, we have the following inequality with probability at least $1-\delta$

$$
\left(\sqrt{\mathbb{E}\hat{\sigma}_m^2}-\sqrt{\frac{2(b_2-b_1)^2u}{N(m-1)}\log\frac{1}{\delta}}\right)^2\leq\hat{\sigma}_m^2+\frac{2(b_2-b_1)^2u}{N(m-1)}\log\frac{1}{\delta}.
$$

Taking the square-root and using $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$, we obtain for any $\delta\in(0,1)$, with probability at least $1-\delta$

$$
\sqrt{\mathbb{E}\hat{\sigma}_m^2}\leq\sqrt{\hat{\sigma}_m^2}+\sqrt{\frac{8(b_2-b_1)^2u}{N(m-1)}\log\frac{1}{\delta}}.
$$

The proof is complete. □

Step III. Using the identity from [4, Page 5]

$$
\mathbb{E}\hat{\sigma}_m^2=\frac{N}{N-1}\sigma^2.
$$

Lemma 10 implies the following inequality with probability at least $1-\delta$

$$
\sqrt{\frac{N}{N-1}}\sigma\leq\hat{\sigma}_m+(b_2-b_1)\sqrt{\frac{8u}{N(m-1)}\log\frac{1}{\delta}},
$$

which completes the proof of Theorem 7.

C.2 Proof of Theorem 8

Proof. Following the proof of Theorem 7, let $f(\Pi) = \frac{m}{(b_2-b_1)^2}\hat{\sigma}_m^2$, we know $a^2 = \frac{mu}{m-1}$ and $\mathbb{E}f(\Pi) = \frac{m}{(b_2-b_1)^2}\mathbb{E}\hat{\sigma}_m^2$. Plugging them into (1) of Lemma 2 gives

$$\mathbb{P}\left\{\frac{m}{(b_2-b_1)^2}\hat{\sigma}_m^2 - \mathbb{E}\frac{m}{(b_2-b_1)^2}\hat{\sigma}_m^2 \geq t\right\} \leq \exp\left(\frac{-N(m-1)t^2}{4mu\mathbb{E}\frac{m}{(b_2-b_1)^2}\hat{\sigma}_m^2 + 2mut}\right).$$

Equivalently, we obtain for any $\delta \in (0,1)$ with probability at least $1-\delta$

$$\hat{\sigma}_m^2 \leq \mathbb{E}\hat{\sigma}_m^2 + \sqrt{\frac{4(b_2-b_1)^2 u\mathbb{E}\hat{\sigma}_m^2}{N(m-1)}\log\frac{1}{\delta}} + \frac{2(b_2-b_1)^2 u}{N(m-1)}\log\frac{1}{\delta}.$$

Completing the square on the right hand side, we have the following inequality with probability at least $1-\delta$

$$\hat{\sigma}_m^2 \leq \left(\sqrt{\mathbb{E}\hat{\sigma}_m^2} + \sqrt{\frac{2(b_2-b_1)^2 u}{N(m-1)}\log\frac{1}{\delta}}\right)^2.$$

Taking the square-root, with probability at least $1-\delta$

$$\sqrt{\hat{\sigma}_m^2} \leq \sqrt{\mathbb{E}\hat{\sigma}_m^2} + \sqrt{\frac{2(b_2-b_1)^2 u}{N(m-1)}\log\frac{1}{\delta}}.$$

Since $\mathbb{E}\hat{\sigma}_m^2 = \frac{N}{N-1}\sigma^2$ [4], we obtain for any $\delta \in (0,1)$, with probability at least $1-\delta$

$$\hat{\sigma}_m \leq \sqrt{\frac{N}{N-1}}\sigma + \sqrt{\frac{2(b_2-b_1)^2 u}{N(m-1)}\log\frac{1}{\delta}}.$$

The proof is complete. □

D Proofs for Section 5

Notation. We introduce some notation to simplify the proof. If B is a set and $r \in \mathbb{N}$, then let $\mathcal{S}_B^r$ denote the set of all those subsets of B which have cardinality r. Also, if $S \subseteq \{1,...,m\}$, we use x_S to denote the vector $(x_{i_1},...,x_{i_{|S|}})$, where $\{i_1,...,i_{|S|}\} = S$ and the i_k are increasingly ordered. For $y, z \in \mathcal{X}$ we use (y, x_S) and (y, z, x_S) to denote respectively, the vectors $(y, x_{i_1},...,x_{i_{|S|}})$ and $(y, z, x_{i_1},...,x_{i_{|S|}})$. With this notation

$$U = \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}_{\{1,...,m\}}^r} K(X_S). \tag{19}$$

With reference to any given $k \in \{1,...,m\}$, using the symmetry of K, we can divide the set x_S into two subsets, those that contain the element x_{π_k} and those that do not, so we have

$$U(\pi) = \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}_{\{\pi_1,...,\pi_m\}\setminus\pi_k}^{r-1}} K(x_{\pi_k}, x_S) + \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}_{\{\pi_1,...,\pi_m\}}^{r}:\pi_k\notin S} K(x_S).$$

In particular, U is (m,u)–symmetric with respect to permutations, so the template of Section 2 applies.

D.1 Proof of Theorem 11

Proof. We proceed in three steps.

Step 1 (Choose f and a *constant* surrogate g). Let $f(\pi) = U(\pi)$. From the above decomposition,

$$\begin{aligned}
&V_\pi(f) = \sum_{i\in I}\sum_{j\in J}(f(\pi) - f(\pi^{i,j}))^2\mathbb{I}\{f(\pi) \geq f(\pi^{i,j})\} \\
\leq &\sum_{i\in I}\sum_{j\in J}\Big(\frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i}, x_S) + \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r}_{\{\pi_1,\ldots,\pi_m\}}:\pi_i\notin S} K(x_S) \\
&- \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_j}, x_S) - \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r}_{\{\pi_1,\ldots,\pi_m\}}:\pi_i\notin S} K(x_S)\Big)^2 \\
= &\sum_{i\in I}\sum_{j\in J}\left(\frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i}, x_S) - \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_j}, x_S)\right)^2 \\
\leq &\sum_{i\in I}\sum_{j\in J}\left(\frac{r}{m}\right)^2(b_2 - b_1)^2 \\
= &mu\left(\frac{r}{m}\right)^2(b_2 - b_1)^2.
\end{aligned}$$

Set the (m, u)-symmetric surrogate to the constant

$$g(\pi) \;=\; mu\left(\frac{r}{m}\right)^2(b_2 - b_1)^2.$$

Step 2 (Sensitivity of g). Since g is constant, $V_\pi(g) = 0$, hence we may take $a = 0$.

Step 3 (Compute $\mathbb{E}[g(\Pi)]$ and apply Theorem 1). Here $\mathbb{E}[g(\Pi)] = mu\left(\frac{r}{m}\right)^2(b_2-b_1)^2$. Applying Theorem 1 with $a = 0$ gives, for any $t \geq 0$,

$$\mathbb{P}\{U - \mu \geq t\} \leq \exp\left(\frac{-Nmt^2}{4ur^2(b_2 - b_1)^2}\right).$$

The proof is complete. □

D.2 Proof of Theorem 12

Proof. We proceed in three steps.

Step 1 (Choose f and construct g with $V_\pi(f) \leq g(\pi)$). Let $f(\pi) = U(\pi)$, we have

$$\begin{aligned}
&V_\pi(f) = \sum_{i\in I}\sum_{j\in J}(f(\pi) - f(\pi^{i,j}))^2\mathbb{I}\{f(\pi) \geq f(\pi^{i,j})\} \\
\leq &\sum_{i\in I}\sum_{j\in J}\Big(\frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i}, x_S) + \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r}_{\{\pi_1,\ldots,\pi_m\}}:\pi_i\notin S} K(x_S) \\
&- \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_j}, x_S) - \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r}_{\{\pi_1,\ldots,\pi_m\}}:\pi_i\notin S} K(x_S)\Big)^2 \\
= &\sum_{i\in I}\sum_{j\in J}\left(\frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i}, x_S) - \frac{1}{\binom{m}{r}}\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_j}, x_S)\right)^2 \\
= &\left(\frac{1}{\binom{m}{r}}\right)^2\sum_{i\in I}\sum_{j\in J}\left(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i}, x_S) - K(x_{\pi_j}, x_S)\right)^2.
\end{aligned}$$

Hence we may take the (m,u)-symmetric surrogate

$$g(\pi) \;=\; \left(\frac{1}{\binom{m}{r}}\right)^2 \sum_{i\in I}\sum_{j\in J}\left(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)\right)^2,$$

for which $V_\pi(f)\le g(\pi)$ holds for all $\pi\in S_N$.

Step 2 (Sensitivity of g: bound $V_\pi(g)\le a^2 g(\pi)$). A direct computation gives

$$\begin{aligned}
V_\pi(g) &= \sum_{k\in I}\sum_{l\in J}(g(\pi)-g(\pi^{k,l}))^2\mathbb{I}\{g(\pi)\ge g(\pi^{k,l})\}\\
&\le \left(\frac{1}{\binom{m}{r}}\right)^4 \sum_{k\in I}\sum_{l\in J}\Big(\sum_{i\in I,i\neq k}\sum_{j\in J,j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)\Big)^2\\
&\quad+\sum_{i=k}\sum_{j\in J,j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)\Big)^2\\
&\quad+\sum_{i\in I,i\neq k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)\Big)^2\\
&\quad+\sum_{i=k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)\Big)^2\\
&\quad-\sum_{i\in I,i\neq k}\sum_{j\in J,j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)\Big)^2\\
&\quad-\sum_{i=k}\sum_{j\in J,j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_l},x_S)-K(x_{\pi_j},x_S)\Big)^2\\
&\quad-\sum_{i\in I,i\neq k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_k},x_S)\Big)^2\\
&\quad-\sum_{i=k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_l},x_S)-K(x_{\pi_k},x_S)\Big)^2\Big)^2\\
&= \left(\frac{1}{\binom{m}{r}}\right)^4 \sum_{k\in I}\sum_{l\in J}\Big(\sum_{i=k}\sum_{j\in J,j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)\Big)^2\\
&\quad+\sum_{i\in I,i\neq k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)\Big)^2\\
&\quad-\sum_{i=k}\sum_{j\in J,j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_l},x_S)-K(x_{\pi_j},x_S)\Big)^2\\
&\quad-\sum_{i\in I,i\neq k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_k},x_S)\Big)^2\Big)^2
\end{aligned}$$

which equals to

$$
\begin{aligned}
&= \left(\frac{1}{\binom{m}{r}}\right)^4 \sum_{k\in I}\sum_{l\in J}\Big(\sum_{i=k}\sum_{j\in J, j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_j},x_S) \\
&\quad + \sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_l},x_S) - K(x_{\pi_j},x_S)\Big) \times \Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_j},x_S) \\
&\quad - \sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_l},x_S) + K(x_{\pi_j},x_S)\Big) + \sum_{i\in I, i\neq k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_j},x_S) \\
&\quad + \sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_k},x_S)\Big) \times \Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_j},x_S) \\
&\quad - \sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) + K(x_{\pi_k},x_S)\Big)\Big)^2 \\
&= \left(\frac{1}{\binom{m}{r}}\right)^4 \sum_{k\in I}\sum_{l\in J}\Big(\sum_{i=k}\sum_{j\in J, j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_j},x_S) \\
&\quad + \sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_l},x_S) - K(x_{\pi_j},x_S)\Big) \times \Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_k},x_S) - K(x_{\pi_l},x_S)\Big) \\
&\quad + \sum_{i\in I, i\neq k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_j},x_S) \\
&\quad + \sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_k},x_S)\Big) \times \Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_k},x_S) - K(x_{\pi_l},x_S)\Big)\Big)^2
\end{aligned}
$$

which also equals to

$$
\begin{aligned}
&= \left(\frac{1}{\binom{m}{r}}\right)^4 \sum_{k\in I}\sum_{l\in J}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_k},x_S) - K(x_{\pi_l},x_S)\Big)^2 \\
&\times \Big(\sum_{i=k}\sum_{j\in J, j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_j},x_S) \\
&+ \sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_l},x_S) - K(x_{\pi_j},x_S)\Big) \\
&+ \sum_{i\in I, i\neq k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_j},x_S) + \sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\ldots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S) - K(x_{\pi_k},x_S)\Big)\Big)^2.
\end{aligned}
$$

Note that

$$\Big(\sum_{i=k}\sum_{j\in J,j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\dots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)+\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\dots,\pi_m\}\setminus\pi_i}} K(x_{\pi_l},x_S)-K(x_{\pi_j},x_S)\Big)$$

$$+\sum_{i\in I,i\neq k}\sum_{j=l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\dots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_j},x_S)+\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\dots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_k},x_S)\Big)\Big)^2$$

$$=\Big(\sum_{j\in J,j\neq l}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\dots,\pi_m\}\setminus\pi_i}} K(x_{\pi_k},x_S)-K(x_{\pi_j},x_S)+\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\dots,\pi_m\}\setminus\pi_i}} K(x_{\pi_l},x_S)-K(x_{\pi_j},x_S)\Big)$$

$$+\sum_{i\in I,i\neq k}\Big(\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\dots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_l},x_S)+\sum_{S\in\mathcal{S}^{r-1}_{\{\pi_1,\dots,\pi_m\}\setminus\pi_i}} K(x_{\pi_i},x_S)-K(x_{\pi_k},x_S)\Big)\Big)^2$$

$$\leq 4(N-2)^2\binom{m-1}{r-1}^2(b_2-b_1)^2.$$

Thus we obtain

$$V_\pi(g)=\sum_{k\in I}\sum_{l\in J}(g(\pi)-g(\pi^{k,l}))^2\mathbb{I}\{g(\pi)\geq g(\pi^{k,l})\}\leq 4\left(\frac{1}{\binom{m}{r}}\right)^2(N-2)^2\binom{m-1}{r-1}^2(b_2-b_1)^2g(\pi).$$

Thus we can take

$$a^2 \;=\; 4\left(\frac{1}{\binom{m}{r}}\right)^2(N-2)^2\binom{m-1}{r-1}^2(b_2-b_1)^2.$$

Step 3 (Compute $\mathbb{E}[g(\Pi)]$ and plug into Theorem 1). By symmetry of the without-replacement model, $\mathbb{E}[g(\Pi)]$ has the following bound

$$\mathbb{E}\left[\left(\frac{1}{\binom{m}{r}}\right)^2\sum_{i\in I}\sum_{j\in J}\left(\sum_{S\in\mathcal{S}^{r-1}_{\{\Pi_1,\dots,\Pi_m\}\setminus\Pi_i}} K(x_{\Pi_i},x_S)-K(x_{\Pi_j},x_S)\right)^2\right]$$

$$=\left(\frac{1}{\binom{m}{r}}\right)^2 mu\mathbb{E}\left[\left(\sum_{S\in\mathcal{S}^{r-1}_{\{\Pi_1,\dots,\Pi_m\}\setminus\Pi_i}}(K(x_{\Pi_i},x_S)-\mu)-(K(x_{\Pi_j},x_S)-\mu)\right)^2\right]$$

$$\leq 2\left(\frac{1}{\binom{m}{r}}\right)^2 mu\binom{m-1}{r-1}^2\mathbb{E}\left[(K(x_{\Pi_i},x_S)-\mu)^2\right]=2\left(\frac{1}{\binom{m}{r}}\right)^2 mu\binom{m-1}{r-1}^2\sigma^2.$$

Applying Theorem 1 with the above g and a yields, for any $t\geq 0$,

$$\mathbb{P}\{U-\mu\geq t\}\leq\exp\left\{\frac{-Nmt^2}{8ur^2\sigma^2+4(b_2-b_1)(N-2)rt}\right\},$$

which is the stated Bernstein–Serfling bound. □

E Proofs for Section 6

We begin by recording an inequality. Consider two functions $f,g:\mathcal{A}\to\mathbb{R}$ defined on an arbitrary set $\mathcal{A}$ and suppose that $\sup_{a\in\mathcal{A}}f(a)=f(\bar{a})$ for some $\bar{a}\in A$. Then

$$\left(\sup_{a\in\mathcal{A}}f(a)-\sup_{a\in\mathcal{A}}g(a)\right)^2\mathbb{I}\left\{\sup_{a\in\mathcal{A}}f(a)\geq\sup_{a\in\mathcal{A}}g(a)\right\}\leq(f(\bar{a})-g(\bar{a}))^2. \tag{20}$$

E.1 Proof of Theorem 13

Proof. We proceed in three steps.

Step 1 (Choose f and construct g with $V_\pi(f) \le g(\pi)$). Suppose, for the purpose of presentation, that the supremum in the definition $Q_m(\pi) = \sup_{f\in\mathcal{F}} \sum_{i=1}^m f(x_{\pi_i})$ is attained at some function $\bar{f} \in \mathcal{F}$. One may wonder whether this supremum is always attained. This point can be made rigorous in a standard way: either by working with ε-maximizers and then letting $\varepsilon \to 0$, or equivalently by passing to the closure of the finite-dimensional evaluation set

$$\mathcal{V} := \{(f(x_1), \dots, f(x_N)) : f \in \mathcal{F}\} \subset [-1,1]^N.$$

Since the population is finite, all quantities used below depend only on these evaluation vectors, and the relevant functionals are continuous. Hence passing to the compact closure $\overline{\mathcal{V}}$ does not change the supremum, and a maximizer exists there. We present the proof using the notation $\bar{f}$ for simplicity; the same inequalities follow by the preceding limiting argument without any additional assumption on $\mathcal{F}$.

With this convention, Eq.(20) implies

$$\begin{aligned}
&\sum_{i\in I}\sum_{j\in J}(Q_m(\pi) - Q_m(\pi^{i,j}))^2 \mathbb{I}\{Q_m(\pi) \ge Q_m(\pi^{i,j})\} \\
\le &\sum_{i\in I}\sum_{j\in J}\left(\sum_{k=1}^m \bar{f}(x_{\pi_k}) - \sum_{k=1}^m \bar{f}(x_{\pi_k^{i,j}})\right)^2 = \sum_{i\in I}\sum_{j\in J}\left(\bar{f}(x_{\pi_i}) - \bar{f}(x_{\pi_j})\right)^2 \\
\le &2\sum_{i\in I}\sum_{j\in J}\left(\bar{f}(x_{\pi_i})\right)^2 + 2\sum_{i\in I}\sum_{j\in J}\left(\bar{f}(x_{\pi_j})\right)^2 \\
\le &2u \sup_{f\in\mathcal{F}} \sum_{i\in I}\left(f(x_{\pi_i})\right)^2 + 2m \sup_{f\in\mathcal{F}} \sum_{j\in J}\left(f(x_{\pi_j})\right)^2.
\end{aligned}$$

Hence we may take the (m,u)-symmetric surrogate

$$g(\pi) \;=\; 2(u \sup_{f\in\mathcal{F}} \sum_{i\in I}\left(f(x_{\pi_i})\right)^2 + m \sup_{f\in\mathcal{F}} \sum_{j\in J}\left(f(x_{\pi_j})\right)^2),$$

for which $V_\pi(f) \le g(\pi)$ holds for all $\pi \in S_N$.

Step 2 (Sensitivity of g: bound $V_\pi(g) \le a^2 g(\pi)$). Suppose that the supremum in $\sup_{f\in\mathcal{F}} \sum_{i\in I}\left(f(x_{\pi_i})\right)^2$ is attained at the function $\bar{f} \in \mathcal{F}$, and that the supremum in $\sup_{f\in\mathcal{F}} \sum_{j\in J}\left(f(x_{\pi_j})\right)^2$ is attained at the function $\hat{f} \in \mathcal{F}$. Again, Eq.(20) implies

$$\begin{aligned}
&\sum_{i\in I}\sum_{j\in J}(g(\pi) - g(\pi^{i,j}))^2 \mathbb{I}\{g(\pi) \ge g(\pi^{i,j})\} \\
\le &4\sum_{i\in I}\sum_{j\in J}\left(u\sum_{k\in I}\left(\bar{f}(x_{\pi_k})\right)^2 + m\sum_{l\in J}\left(\hat{f}(x_{\pi_l})\right)^2 - u\sum_{k\in I}(\bar{f}(x_{\pi_k^{i,j}}))^2 - m\sum_{l\in J}(\hat{f}(x_{\pi_l^{i,j}}))^2\right)^2 \\
= &4\sum_{i\in I}\sum_{j\in J}\left(u(\bar{f}(x_{\pi_i}))^2 - u(\bar{f}(x_{\pi_j}))^2 + m(\hat{f}(x_{\pi_j}))^2 - m(\hat{f}(x_{\pi_i}))^2\right)^2 \\
\le &4\sum_{i\in I}\sum_{j\in J}\left(u(\bar{f}(x_{\pi_i}))^2 + m(\hat{f}(x_{\pi_j}))^2\right)^2 \\
\le &8\left(\sum_{i\in I}\sum_{j\in J} u^2(\bar{f}(x_{\pi_i}))^2 + m^2(\hat{f}(x_{\pi_i}))^2\right) \\
= &8\left(u^3\sum_{i\in I}\bar{f}(x_{\pi_j})^2 + m^3\sum_{j\in J}\hat{f}(x_{\pi_i})^2\right) \\
\le &4\max\{u^2, m^2\} g(\pi),
\end{aligned}$$

where the third inequality uses the fact that $f \in [-1,1]$. Thus we can take

$$a^2 \;=\; 4\max\{u^2, m^2\}.$$

Step 3 (Compute $\mathbb{E}[g(\Pi)]$ and plug into Theorem 1). To prove $\mathbb{E}[g(\Pi)]$, we need an inequality characterizing *the weak and the wimpy variances of sampling without replacement*:

$$\begin{aligned}\mathbb{E}_\Pi\left[\sup_{f\in\mathcal{F}}\sum_{i\in I}f(x_{\Pi_i})^2\right] &= \mathbb{E}_\Pi\Big[\sup_{f\in\mathcal{F}}(\sum_{i\in I}f(x_{\Pi_i})^2-\mathbb{E}_\Pi\sum_{i\in I}f(x_{\Pi_i})^2+\mathbb{E}_\Pi\sum_{i\in I}f(x_{\Pi_i})^2)\Big]\\ &\le\ \mathbb{E}_\Pi\Big[\sup_{f}(\sum_{i\in I}f(x_{\Pi_i})^2-\mathbb{E}_\Pi\sum_{i\in I}f(x_{\Pi_i})^2)\Big]+m\sigma^2, \qquad (21)\end{aligned}$$

by the definition of σ^2: $\sup_{f\in\mathcal{F}}\mathbb{E}_\Pi[\sum_{i\in I}f(x_{\Pi_i})^2]=m\sigma^2$. Now, by symmetry of the without-replacement model,

$$\begin{aligned}&\mathbb{E}\left[u\sup_{f\in\mathcal{F}}\sum_{i\in I}\left(f(x_{\Pi_i})\right)^2+m\sup_{f\in\mathcal{F}}\sum_{j\in J}\left(f(x_{\Pi_j})\right)^2\right]\\ \le& u\left(m\sigma^2+\mathbb{E}_\Pi\Big[\sup_f(\sum_{i=1}^m f(x_{\Pi_i})^2-\mathbb{E}_\Pi\sum_{i=1}^m f(x_{\Pi_i})^2)\Big]\right)\\ &+m\left(u\sigma^2+\mathbb{E}_\Pi\Big[\sup_f(\sum_{j=m+1}^N f(x_{\Pi_j})^2-\mathbb{E}_\Pi\sum_{j=m+1}^N f(x_{\Pi_j})^2)\Big]\right)\\ =&2mu\sigma^2+muC_m(f,\Pi)+muC_u(f,\Pi),\end{aligned}$$

where the inequality uses (21). Applying Theorem 1 with the above g and a yields, for any $t\ge 0$,

$$\mathbb{P}\{Q_m-\mathbb{E}Q_m>t\}\le\exp\left(\frac{-Nt^2}{8mu\sigma^2+4mu(C_m(f,\Pi)+C_u(f,\Pi))+4\max\{u,m\}t}\right).$$

The proof is complete. □

E.2 Proof of Theorem 14

Proof. Let $f(\pi)=\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m f(X_i)-\frac{1}{u}\sum_{j=m+1}^N f(X_j)=\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i\in I}f(x_{\pi_i})-\frac{1}{u}\sum_{j\in J}f(x_{\pi_j})$, which is (m,u) symmetric. Then

$$\begin{aligned}V_\pi(f)&=\sum_{i\in I}\sum_{j\in J}(f(\pi)-f(\pi^{i,j}))^2\mathbb{I}\{f(\pi)\ge f(\pi^{i,j})\}\\ =&\sum_{i\in I}\sum_{j\in J}\left(\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{k\in I}f(x_{\pi_k})-\frac{1}{u}\sum_{l\in J}f(x_{\pi_l})-(\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{k\in I}f(x_{\pi_k^{i,j}})-\frac{1}{u}\sum_{l\in J}f(x_{\pi_l^{i,j}}))\right)^2\\ &\times\mathbb{I}\left\{\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{k\in I}f(x_{\pi_k})-\frac{1}{u}\sum_{l\in J}f(x_{\pi_l})\ge\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{k\in I}f(x_{\pi_k^{i,j}})-\frac{1}{u}\sum_{l\in J}f(x_{\pi_l^{i,j}})\right\}. \qquad (22)\end{aligned}$$

The proof of Theorem 14 follows from the proof of Theorem 13. Suppose that the supremum in the definition $\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i\in I}f(x_{\pi_i})-\frac{1}{u}\sum_{j\in J}f(x_{\pi_j})$ is attained at the function $\bar{f}\in\mathcal{F}$, using Eq.(20), Eq.(22) gives

$$\begin{aligned}&\sum_{i\in I}\sum_{j\in J}(f(\pi)-f(\pi^{i,j}))^2\mathbb{I}\{f(\pi)\ge f(\pi^{i,j})\}\\ \le&\sum_{i\in I}\sum_{j\in J}\left(\frac{1}{m}\bar{f}(x_{\pi_i})-\frac{1}{m}\bar{f}(x_{\pi_j})+\frac{1}{u}\bar{f}(x_{\pi_i})-\frac{1}{u}\bar{f}(x_{\pi_j})\right)^2\\ \le&\frac{2N^2}{m^2u^2}\sum_{i\in I}\sum_{j\in J}\left(\bar{f}(x_{\pi_i})\right)^2+\frac{2N^2}{m^2u^2}\sum_{i\in I}\sum_{j\in J}\left(\bar{f}(x_{\pi_j})\right)^2\\ \le&\frac{2N^2}{m^2u^2}\left(u\sup_{f\in\mathcal{F}}\sum_{i\in I}\left(f(x_{\pi_i})\right)^2+m\sup_{f\in\mathcal{F}}\sum_{j\in J}\left(f(x_{\pi_j})\right)^2\right).\end{aligned}$$

Hence we may take the (m,u)-symmetric surrogate $g(\pi) = \frac{2N^2}{m^2u^2}\Big(u\sup_{f\in\mathcal{F}}\sum_{i\in I}\left(f(x_{\pi_i})\right)^2 + m\sup_{f\in\mathcal{F}}\sum_{j\in J}\left(f(x_{\pi_j})\right)^2\Big)$. Further following the proof of Theorem 13, a direct calculation gives for all $\pi \in S_N$, $V_\pi(g) \le \frac{4N^2}{m^2u^2}\max\{u^2,m^2\}g(\pi)$, so $a^2 = \frac{4N^2}{m^2u^2}\max\{u^2,m^2\}$. Finally, by (21)

$$\begin{aligned}\mathbb{E}[g(\Pi)] =& \mathbb{E}\left[\frac{2N^2}{m^2u^2}\left(u\sup_{f\in\mathcal{F}}\sum_{i\in I}\left(f(x_{\Pi_i})\right)^2 + m\sup_{f\in\mathcal{F}}\sum_{j\in J}\left(f(x_{\Pi_j})\right)^2\right)\right]\\ \le& \frac{2N^2}{m^2u^2}\left(2mu\sigma^2 + muC_m(f,\Pi) + muC_u(f,\Pi)\right).\end{aligned}$$

Plugging these results into Theorem 1, the proof is complete. □

F Equivalence with the condition in [12]

[12] considers a symmetric function $\phi : \mathcal{X}^m \to \mathbb{R}$ satisfying the one-point bounded-difference condition

$$|\phi(x_1,\dots,x_m) - \phi(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_m)| \le c.$$

Here we restrict attention to feasible one-point replacements under sampling without replacement, where the replacement observation is associated with a population index not already included in the sample.

For $\pi = (\pi_1,\dots,\pi_N)$, define

$$f(\pi) = \phi(x_{\pi_1},\dots,x_{\pi_m}).$$

Then f is (m,u)-symmetric: symmetry of ϕ gives invariance in the first block, and the last block does not enter f. For $i \le m < j$, $\pi^{i,j}$ replaces x_{π_i} by x_{π_j}, so the condition above becomes exactly

$$|f(\pi) - f(\pi^{i,j})| \le c.$$

Conversely, an (m,u)-symmetric f induces a well-defined symmetric function on feasible samples of m distinct population indices, and a feasible one-point replacement corresponds to a cross-block transposition. Thus, the two conditions are equivalent on the support induced by sampling without replacement.

G Equivalence with Bobkov's concentration inequality

In this section, we explain the exact correspondence between Theorem 3 and the concentration inequality obtained from Bobkov's logarithmic-Sobolev inequality [8, Theorem 2.1].

Let

$$G_{N,m} := \left\{x \in \{0,1\}^N : \sum_{k=1}^N x_k = m\right\}$$

be the slice of the discrete cube equipped with the uniform measure. For $x \in G_{N,m}$, define

$$I(x) := \{p : x_p = 1\}, \qquad J(x) := \{q : x_q = 0\}.$$

For $p \in I(x)$ and $q \in J(x)$, let $s_{pq}x$ denote the vector obtained by exchanging coordinates p and q. For a function $h : G_{N,m} \to \mathbb{R}$, define

$$|\nabla h(x)|^2 := \sum_{p\in I(x)}\sum_{q\in J(x)}\left(h(x) - h(s_{pq}x)\right)^2.$$

Bobkov's logarithmic-Sobolev inequality [8, Theorem 2.1], together with the standard Herbst argument, implies that if

$$|\nabla h(x)|^2 \le \alpha \qquad \text{for all } x \in G_{N,m},$$

then, for X uniformly distributed on $G_{N,m}$,

$$\mathbb{P}\left\{h(X) - \mathbb{E}h(X) \ge t\right\} \le \exp\left\{-\frac{(N+2)t^2}{4\alpha}\right\}, \qquad t \ge 0.$$

We now identify this formulation with the permutation representation used in Theorem 3. For $\pi = (\pi_1, \dots, \pi_N) \in S_N$, define

$$x_k(\pi) := \mathbf{1}\{k \in \{\pi_1, \dots, \pi_m\}\}, \qquad k = 1, \dots, N.$$

Then $x(\pi) \in G_{N,m}$. Moreover, if $\Pi \sim \mathrm{Unif}(S_N)$, then $x(\Pi)$ is uniformly distributed on $G_{N,m}$. Indeed, for every subset $A \subseteq \{1, \dots, N\}$ with $|A| = m$,

$$\mathbb{P}\{\{\Pi_1, \dots, \Pi_m\} = A\} = \binom{N}{m}^{-1}.$$

Let $f : S_N \to \mathbb{R}$ be an (m, u)-symmetric function, where $u = N - m$. Define $h : G_{N,m} \to \mathbb{R}$ by

$$h(x(\pi)) := f(\pi).$$

This definition is well defined. Indeed, if two permutations $\pi, \pi' \in S_N$ satisfy $x(\pi) = x(\pi')$, then they differ only by permutations within the first m coordinates and within the last u coordinates. By (m, u)-symmetry,

$$f(\pi) = f(\pi').$$

Fix $1 \le i \le m < j \le N$, and set

$$p := \pi_i, \qquad q := \pi_j.$$

Then

$$p \in I(x(\pi)), \qquad q \in J(x(\pi)).$$

Let $\pi^{i,j}$ denote the permutation obtained by exchanging the i-th and j-th coordinates of π. Then

$$x(\pi^{i,j}) = s_{pq} x(\pi),$$

and hence

$$f(\pi) - f(\pi^{i,j}) = h(x(\pi)) - h(s_{pq} x(\pi)).$$

Furthermore, for a fixed permutation π, the correspondence

$$(i, j) \longmapsto (\pi_i, \pi_j)$$

is a bijection between

$$\{(i, j) : 1 \le i \le m < j \le N\}$$

and

$$I(x(\pi)) \times J(x(\pi)).$$

Consequently,

$$\sum_{i=1}^{m} \sum_{j=m+1}^{N} \big(f(\pi) - f(\pi^{i,j})\big)^2 = |\nabla h(x(\pi))|^2.$$

Therefore, the assumption in Theorem 3,

$$\sum_{i=1}^{m} \sum_{j=m+1}^{N} \big(f(\pi) - f(\pi^{i,j})\big)^2 \le \alpha \qquad \text{for all } \pi \in S_N,$$

is exactly Bobkov's discrete-gradient condition under the above identification. Since $h(x(\Pi)) = f(\Pi)$ almost surely, Bobkov's concentration inequality becomes

$$\mathbb{P}\{f(\Pi) - \mathbb{E}f(\Pi) \ge t\} \le \exp\left\{-\frac{(N+2)t^2}{4\alpha}\right\}, \qquad t \ge 0,$$

which is precisely the assertion of Theorem 3.

H Equivalence of the kernel representations

The formulation in [2] allows an arbitrary bounded, possibly nonsymmetric kernel g. Specifically, it defines

$$U_g = \frac{1}{m^{(r)}} \sum_{(i_1,\dots,i_r)\in[m]^r_{\neq}} g(X_{i_1},\dots,X_{i_r}),$$

where $[m]^r_{\neq}$ is the set of all ordered r-tuples of distinct indices and

$$m^{(r)} = m(m-1)\cdots(m-r+1).$$

For a possibly nonsymmetric kernel g, define its symmetrization by

$$\bar{g}(x_1,\dots,x_r) = \frac{1}{r!}\sum_{\sigma\in S_r} g(x_{\sigma(1)},\dots,x_{\sigma(r)}).$$

Then $\bar{g}$ is symmetric. Grouping the ordered r-tuples according to their underlying unordered r-subsets gives

$$\begin{aligned} U_g &= \frac{1}{m^{(r)}} \sum_{1\le i_1<\cdots<i_r\le m} \sum_{\sigma\in S_r} g(X_{i_{\sigma(1)}},\dots,X_{i_{\sigma(r)}}) \\ &= \frac{r!}{m^{(r)}} \sum_{1\le i_1<\cdots<i_r\le m} \bar{g}(X_{i_1},\dots,X_{i_r}) \\ &= \frac{1}{\binom{m}{r}} \sum_{1\le i_1<\cdots<i_r\le m} \bar{g}(X_{i_1},\dots,X_{i_r}), \end{aligned}$$

where we used

$$m^{(r)} = r!\binom{m}{r}.$$

Taking $K = \bar{g}$, the last expression is exactly the symmetric-kernel representation used in Section 5.

Thus, the two kernel representations define exactly the same U-statistic. Moreover, if $b_1 \le g \le b_2$, then also $b_1 \le \bar{g} \le b_2$. The symmetric-kernel representation depends on the first m sampled observations only through the collection of their r-subsets, and hence the resulting U-statistic is (m,u)-symmetric.

Consequently, the relevant assumptions in Theorem 11 and Theorem 4 of [2] are equivalent.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the Conclusion Section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 2 provides the assumptions, and the Appendix provides all the proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper focuses on theory and does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper focuses on theory and does not include experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper focuses on theory and does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper focuses on theory and does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper focuses on theory and does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research cnducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper focuses on theory, and there is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used to assist with writing and editing the manuscript and did not affect the core methodology, scientific rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.