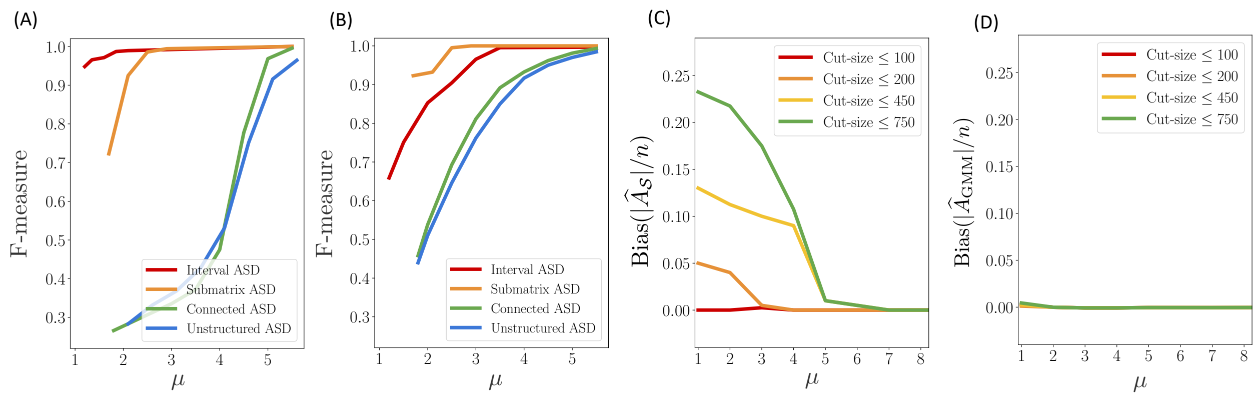

Figure S1: Data $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ as in Figure 2 in main text. (A) F-measure of the MLE $\widehat{A}_{\mathcal{S}}$ versus μ for means $\mu \geq \mu_{\text{detect}}$. Note that the MLEs $\widehat{A}_{\mathcal{S}}$ for the connected family $\mathcal{S} = \mathcal{C}_G$ and unstructured family $\mathcal{S} = \mathcal{P}_n$ have low F-measure for small means μ, consistent with the $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n)$ for these families shown in Figure 2. (B) F-measure of our estimator $\widehat{A}_{\text{GMM}}$ versus μ for means $\mu \geq \mu_{\text{detect}}$. (C) $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n)$ of the MLE versus μ for means $\mu \geq 1$ and for graph cut family $\mathcal{S} = \mathcal{T}_{G,\rho}$ with different bounds ρ on the cut-size and. (D) $\text{Bias}(|\widehat{A}_{\text{GMM}}|/n)$ of our GMM estimator versus μ for means $\mu \geq 1$ and for graph cut family $\mathcal{S} = \mathcal{T}_{G,\rho}$ with different bounds ρ on the cut-size.

A. Calculating μ_{detect}

μ_{detect} is the smallest mean μ such that the GLR test asymptotically solves the ASD Detection Problem with the probability of a type 1 or type 2 error going to 0 as $n \to \infty$ (Sharpnack et al., 2013a). We empirically determine μ_{detect} by finding the smallest mean μ such that the Type I and Type II errors of the GLR test statistic $\widehat{t}_{\mathcal{S}}$ (Equation (2) in the main text) are both less than 0.01.

B. Additional Experiments

B.1. F-measure

Although Conjecture 1 is about the $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n)$ of the MLE $\widehat{A}_{\mathcal{S}}$, we also observe that larger $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n)$ reduces the F-measure between the anomaly A and the MLE $\widehat{A}_{\mathcal{S}}$. Using the data described in Section 3.1 in the main text, we find a noticeable difference in F-measure between anomaly families where $|\breve{\mathcal{S}}(A)|$ is exponential — the connected family $\mathcal{C}_G$ and the unstructured family $\mathcal{P}_n$ — and anomaly families where $|\breve{\mathcal{S}}(A)|$ is sub-exponential — the interval family $\mathcal{I}_n$ and the submatrix family $\mathcal{M}_N$ (Figure S1 A).

In contrast, our GMM-based estimator $\widehat{A}_{\text{GMM}}$ has a much smaller difference in the F-measure for anomaly families where $|\breve{\mathcal{S}}(A)|$ is exponential versus anomaly families where $|\breve{\mathcal{S}}(A)|$ is sub-exponential (Figure S1 B). This result is consistent with the reduced bias of the GMM-based estimator $\widehat{A}_{\text{GMM}}$ (Figure 2C, main text). Interestingly, even for our reduced bias estimator, we still observe a mild difference in F-measure between the families with exponential $|\breve{\mathcal{S}}(A)|$ versus the families with sub-exponential $|\breve{\mathcal{S}}(A)|$.

B.2. Graph Cut Family

We examine the $\text{Bias}(|\widehat{A}_{\mathcal{T}_{G,\rho}}|/n)$ of the size of the MLE $\widehat{A}_{\mathcal{T}_{G,\rho}}$ for the graph cut family $\mathcal{T}_{G,\rho}$, where G is a $\sqrt{n} \times \sqrt{n}$ lattice graph, for different values of the bound ρ on the cut-size. For each value of ρ, we select an anomaly $A \in \mathcal{T}_{G,\rho}$ with size $|A| = 0.05n$ uniformly at random from $\mathcal{T}_{G,\rho}$. (Note that the cut-size of A is not fixed, as we select A uniformly at random from the set $\mathcal{T}_{G,\rho}$ of all subgraphs of G with cut-size less than ρ.) We then draw a sample $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{T}_{G,\rho}}(A, \mu)$ with $n = 900$ observations and compute the MLE $\widehat{A}_{\mathcal{T}_{G,\rho}}$. We repeat for 50 samples to estimate $\text{Bias}(|\widehat{A}_{\mathcal{T}_{G,\rho}}|/n)$.

While the graph cut anomaly family is often studied in the network anomaly literature (Sharpnack et al., 2013b;a; Sharp-

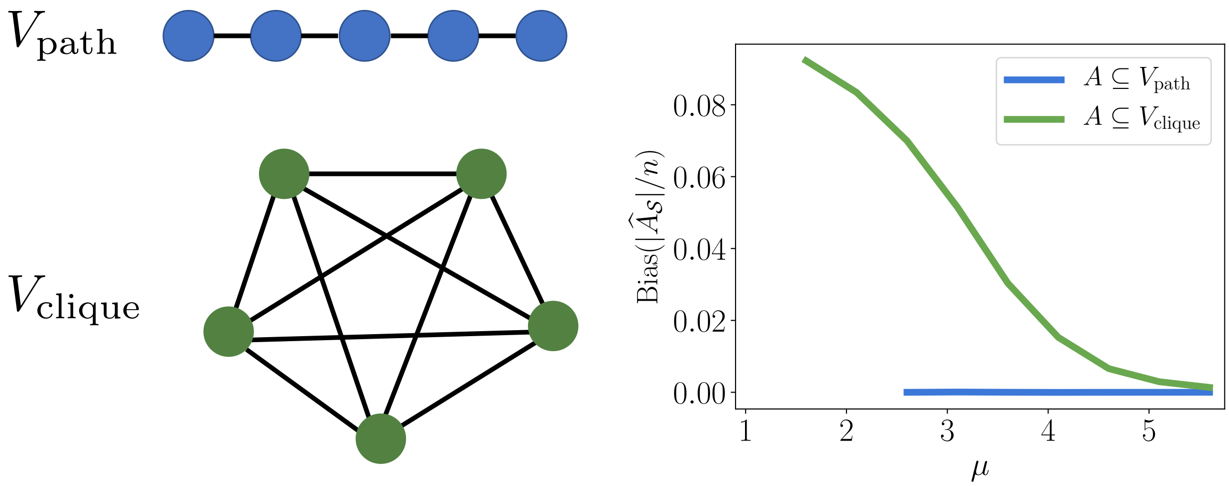

Figure S2: **Left:** Graph $G = (V, E)$ with two disjoint connected components: V_{path}, a path graph, and V_{clique}, a clique graph, with $|V_{\text{path}}| = |V_{\text{clique}}| = \frac{n}{2}$. **Right:** $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n)$ versus mean μ for the connected anomaly family $\mathcal{S} = \mathcal{C}_G$ with $n = |V| = 500$ vertices and an anomaly A with size $|A| = 0.05n$, for means $\mu \geq \mu_{\text{detect}}$. The blue line corresponds to an anomaly $A \subseteq V_{\text{path}}$ and the green line corresponds to an anomaly $A \subseteq V_{\text{clique}}$. This experiment suggests that the $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n)$ is determined by the $|\breve{\mathcal{S}}(A)|$, rather than $|\mathcal{S}|$, consistent with Conjecture 1.

nack, 2013), the cut-size bound ρ is typically left unspecified. When ρ is constant $|\breve{\mathcal{S}}(A)|$ is polynomial in n, but when ρ is close to the number of edges in G then $|\breve{\mathcal{S}}(A)|$ is exponential in n (Nagamochi et al., 1994). So by Conjecture 1 we expect the bias of the MLE $\widehat{A}_{\mathcal{T}_{G,\rho}}$ to depend on ρ. Indeed, we observe that the $\text{Bias}(|\widehat{A}_{\mathcal{T}_{G,\rho}}|/n)$ of the MLE is small when ρ is small and the $\text{Bias}(|\widehat{A}_{\mathcal{T}_{G,\rho}}|/n)$ of the MLE is large when ρ is large (Figure S1 C), which is consistent with Conjecture 1. Our results demonstrate that careful attention to the cut-size bound ρ is required when the MLE $\widehat{A}_{\mathcal{T}_{G,\rho}}$ is used for anomaly estimation.

For the same data, we find that our GMM estimator $\widehat{A}_{\text{GMM}}$ has small bias regardless of the cut-size bound ρ (Figure S1 D). This is consistent with Corollary 1, and demonstrates that our GMM estimator $\widehat{A}_{\text{GMM}}$ is a less biased estimator than the MLE $\widehat{A}_{\mathcal{T}_{G,\rho}}$ regardless of the cut-size bound ρ.

B.3. Dependence of Bias$(|\widehat{A}_{\mathcal{S}}|/n)$ on $|\breve{\mathcal{S}}(A)|$ versus $|\mathcal{S}|$

In this section, we construct an anomaly family $\mathcal{S}$ where $|\mathcal{S}|$ is exponential, but $|\breve{\mathcal{S}}(A)|$ is exponential for some anomalies A and sub-exponential for others. We then use this anomaly family $\mathcal{S}$ to provide evidence that $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n)$ depends on the number $|\breve{\mathcal{S}}(A)|$ of subsets in $\mathcal{S}$ that contain the anomaly A, rather than the size $|\mathcal{S}|$ of the anomaly family.

Let $G = (V, E)$ be a graph whose vertices $V = V_{\text{path}} \cup V_{\text{clique}}$ can be partitioned into two disjoint connected components: V_{path}, a path graph, and V_{clique}, a clique (Figure S2, left). (Note that the path graph V_{path} and the clique V_{clique} are disjoint, unlike the graph from Figure 2.) Both the path graph V_{path} and the clique V_{clique} have size $|V_{\text{path}}| = |V_{\text{path}}| = \frac{n}{2}$, where $n = 900$.

Let $\mathcal{S} = \mathcal{C}_G$ be the connected family for graph G, and let $A \in \mathcal{C}_G$ be a set of size $|A| = 0.05n$. The size $|\mathcal{S}|$ of the anomaly family $\mathcal{S}$ is exponential in n, as $|\mathcal{S}| = O(2^{\frac{n}{2}})$. However, $|\breve{\mathcal{S}}(A)|$ depends on the anomaly A: if the anomaly $A \subseteq V_{\text{path}}$ is in the path graph component, then $|\breve{\mathcal{S}}(A)| = O(n^2)$ is sub-exponential in n. On the other hand, if $A \subseteq V_{\text{clique}}$ is in the clique graph component, then $|\breve{\mathcal{S}}(A)| = O(2^{0.45n})$ is exponential in n.

Empirically, we observe that if $\mu \geq \mu_{\text{detect}}$, then $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n) \approx 0$ if $A \subseteq V_{\text{path}}$ and $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n) > 0$ if $A \subseteq V_{\text{clique}}$ (Figure S2, right). This finding is consistent with Conjecture 1 and demonstrates the dependence of $\text{Bias}(|\widehat{A}_{\mathcal{S}}|/n)$ on $|\breve{\mathcal{S}}(A)|$ rather than $|\mathcal{S}|$.

B.4. Highway Traffic Data with Edge-Dense Family

We compare our estimator $\widehat{A}_{\text{GMM}}$ and the MLE $\widehat{A}_{\mathcal{S}}$ on a real-world highway traffic dataset. This dataset consists of a highway traffic network $G = (V, E)$ in Los Angeles County, CA with $|V| = 1868$ vertices and $|E| = 1993$ edges. The vertices V are sensors that record the speed of cars passing and the edges E connect adjacent sensors. The observations $\mathbf{X} = (X_v)_{v \in V}$ are p-values (where sensors that record higher average speeds have lower p-values) that are transformed to Gaussians using the method in Reyna et al. (2020).

For the edge-dense family $\mathcal{E}_{G,\delta}$ with edge density $\delta = 0.7$, we find that our GMM-based estimator $\widehat{A}_{\text{GMM}}$ is much smaller than the MLE $\widehat{A}_{\mathcal{E}_{G,\delta}}$ ($|\widehat{A}_{\text{GMM}}| = 10$ versus $|\widehat{A}_{\mathcal{E}_{G,\delta}}| = 600$) but with higher average score (4.4 for our estimator versus 0.4 for the MLE). While there is no ground-truth anomaly in this dataset, our results show that our estimator $\widehat{A}_{\text{GMM}}$ yields a smaller anomaly but with higher average values than the MLE $\widehat{A}_{\mathcal{E}_{G,\delta}}$, which also suggests that the MLE $\widehat{A}_{\mathcal{E}_{G,\delta}}$ for the edge-dense family $\mathcal{E}_{G,\delta}$ is biased.

B.5. Additional Details for NYC Breast Cancer

In the NYC breast cancer incidence data (Boscoe et al., 2016), we are given observed disease counts $\mathbf{C} = \{C_1, \dots, C_n\}$ and expected disease counts $\mathbf{B} = \{B_1, \dots, B_n\}$ for each census block $i \in [n]$. As is standard in the disease surveillance and spatial scan statistic literature, (Kulldorff, 1997; Glaz & Naus, 2010; Neill, 2009; 2012), we model the observed counts $\mathbf{C}$ as being distributed as

$$C_i \sim \begin{cases} \text{Pois}(q_{\text{in}} B_i), & \text{if } i \in A, \\ \text{Pois}(B_i), & \text{otherwise,} \end{cases} \tag{1}$$

where $A \in \mathcal{S}$ is the anomaly, $\mathcal{S}$ is the anomaly family, and q_{in} is the *relative risk* of census blocks $i \in A$ in the anomaly A.

The MLE $\widehat{A}_{\mathcal{S}}$ for the anomaly A given the observed counts $\mathbf{C}$ and the expected counts $\mathbf{B}$ — also known as the expectation-based Poisson scan statistic in the spatial scan statistic literature (Neill, 2009) — is given by

$$\widehat{A}_{\mathcal{S}} = \operatorname*{argmax}_{A \in \mathcal{S}} \left[\sum_{i \in A} (B_i) + \left(\sum_{i \in A} C_i \right) \cdot \left(-1 + \log \sum_{i \in A} C_i - \log \sum_{i \in A} B_i \right) \right]. \tag{2}$$

We adapt our estimator to the disease count model in Equation (1) by using the EM algorithm to fit the observed counts $\mathbf{C}$ to the Poisson mixture $C_i \sim \alpha \cdot \text{Pois}(q_{\text{in}} B_i) + (1 - \alpha) \cdot \text{Pois}(B_i)$. The other components of our estimator are unchanged: we use this fit to compute the responsibilities $\widehat{r}_i = P(i \in A \mid C_i, B_i)$ for each census block $i \in [n]$, and then estimate the anomaly using Equation (6) in the main text.

C. Wasserstein Distance between GMM and Unstructured ASD

Let $\mathbf{X} = (X_1, \dots, X_n)$ with $X_i \overset{\text{i.i.d.}}{\sim} \text{GMM}(\mu, \alpha)$ distributed according to the GMM and let $\mathbf{Y} = (Y_1, \dots, Y_n) \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ be distributed according to the unstructured ASD, with $\alpha = |A|/n$. We empirically observe that $d_W\left(\frac{1}{n}\sum_{i=1}^n 1_{X_i}, \frac{1}{n}\sum_{i=1}^n 1_{Y_i}\right) = O(n^{-0.5})$, where d_W is the 1-Wasserstein distance, also known as the earth mover's distance (Figure S3). We note that our empirical observation matches the result that the Wasserstein distance between the normal distribution $N(\mu, \sigma)$ and the empirical distribution of n samples from $N(\mu, \sigma)$ is also $O(n^{-0.5})$ (Rippl et al., 2016; Weed & Bach, 2019).

D. Regularized MLE for Submatrix ASD

For the submatrix family $\mathcal{M}_N$, Liu & Arias-Castro (2019) show that a regularized version of the MLE is asymptotically unbiased. Specifically, for a submatrix $M \in \mathbb{R}^{p \times q}$ of a matrix $N \in \mathbb{R}^{m \times m}$, they define the regularized scan statistic function $\Gamma_{\text{R}}(M) = \Gamma(M) - \sqrt{2 \log\left(m^2 \binom{m}{p}\binom{m}{q}\right)}$ and the regularized MLE $\widehat{A}_{\text{R}} = \operatorname*{argmax}_{M \in \mathcal{M}_N} \Gamma_{\text{R}}(M)$. Liu & Arias-Castro (2019) then show that $\widehat{A}_{\text{R}}$ is asymptotically unbiased.

However, our proof of Theorem 1 shows that the MLE $\widehat{A}_{\mathcal{M}_N}$ for the submatrix ASD, which does not use the above regularization, is also asymptotically unbiased. Thus, the regularization is not required. Empirically, we find that that the

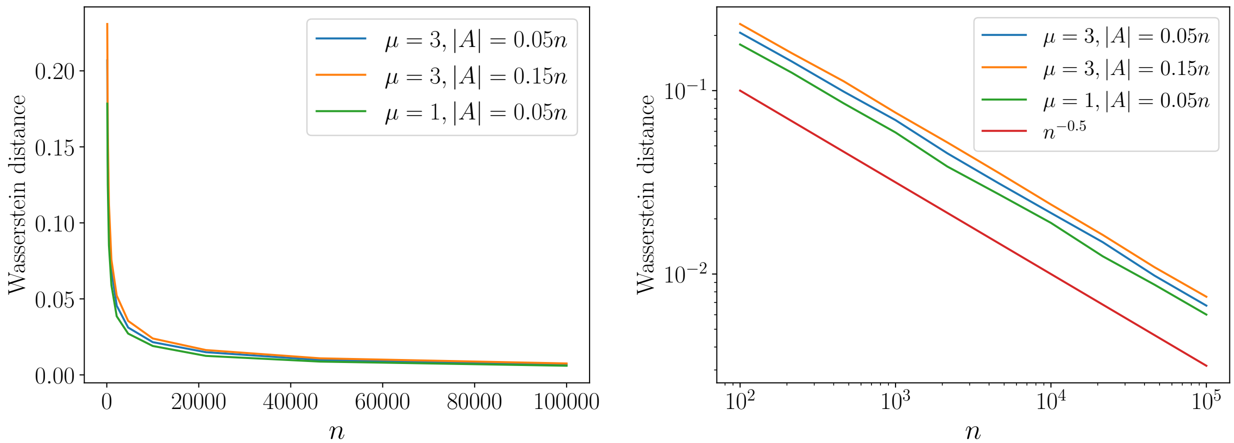

Figure S3: 1-Wasserstein distance between the GMM distribution and the unstructured ASD distribution. $\mathbf{X} = (X_1, \dots, X_n) \overset{\text{i.i.d.}}{\sim} \text{GMM}(\mu, \alpha)$ is distributed according to the GMM and $\mathbf{Y} = (Y_1, \dots, Y_n) \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ is distributed according to the unstructured ASD, with $\alpha = |A|/n$. **Left:** $d_W\left(\frac{1}{n}\sum_{i=1}^n 1_{X_i}, \frac{1}{n}\sum_{i=1}^n 1_{Y_i}\right)$, the 1-Wasserstein distance between the empirical distributions of the GMM and unstructured ASD, versus the number n of observations for various values of μ and $|A|/n$. **Right:** 1-Wasserstein distance on log-log scale. We observe that the 1-Wasserstein distance $d_W\left(\frac{1}{n}\sum_{i=1}^n 1_{X_i}, \frac{1}{n}\sum_{i=1}^n 1_{Y_i}\right)$ is $O(n^{-0.5})$, as each line is parallel to $n^{-0.5}$ in the log-log plot.

MLE $\widehat{A}_{\mathcal{M}_N}$ and the regularized MLE $\widehat{A}_{\text{R}}$ have similar bias and similar F-measure to the anomaly (Figure S4), suggesting that the regularization proposed by (Liu & Arias-Castro, 2019) is not necessary to reduce bias or increase performance in anomaly estimation.

E. Approximating the GMM Estimator for the Submatrix Family and the Connected Family

For the submatrix family $\mathcal{S} = \mathcal{M}_N$ and the connected family $\mathcal{S} = \mathcal{C}_G$, our GMM estimator

$$\widehat{A}_{\text{GMM}} = \underset{\substack{S \in \mathcal{S} \\ \left||S| - \widehat{\alpha}_{\text{GMM}}\right| \leq \sqrt{\frac{\log n}{n}}}}{\operatorname{argmax}} \left(\sum_{i \in S} \widehat{r}_i\right) \tag{3}$$

can be inefficient to compute because of the constraint on the size $|S|$ of the subset S. In our experiments, we relax this constraint by computing the following approximation $\widetilde{A}_{\text{GMM}}$ of our GMM estimator:

$$\widetilde{A}_{\text{GMM}} = \underset{S \in \mathcal{S}}{\operatorname{argmax}} \sum_{i \in S} (\widehat{r}_i - \tau). \tag{4}$$

Here, $\tau > 0$ is a positive number that we use to "shift" the estimated responsibilities $\widehat{r}_i$ to $\widehat{r}_i - \tau$. We select $\tau > 0$ so that the number T of positive "shifted" responsibilities $\widehat{r}_i - \tau$ satisfies $\left|T - \widehat{\alpha}_{\text{GMM}}\right| \leq \sqrt{\frac{\log n}{n}}$. That is, τ is chosen so that $\{i : \widehat{r}_i - \tau > 0\} = T$, where T satisfies $\left|T - \widehat{\alpha}_{\text{GMM}} n\right| \leq \sqrt{\frac{\log n}{n}}$. Because the number of positive shifted responsibilities is T, we expect our approximate estimator $\widetilde{A}_{\text{GMM}}$ to have size $|\widetilde{A}_{\text{GMM}}| \approx T \approx \widehat{\alpha}_{\text{GMM}} n$.

F. Proof of Theorem 1

F.1. Preliminary Lemmas

We first prove the following technical lemmas.

Lemma 1. *Let $\{A_n\}_{n=1,2,\dots}$ and $\{B_n\}_{n=1,2,\dots}$ be two sequences of events in the same probability space. Suppose $\lim_{n\to\infty} P(A_n) = 1$ and $\lim_{n\to\infty} P(B_n) = 1$. Then $\lim_{n\to\infty} P(A_n \cap B_n) = 1$.*

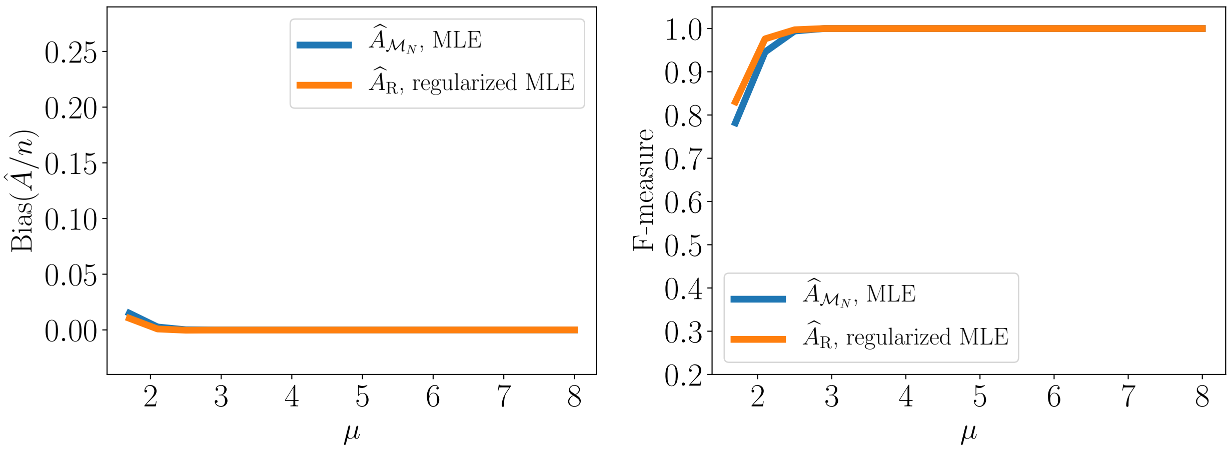

Figure S4: $\mathbf{X} \sim \text{ASD}_{\mathcal{M}_N}(A, \mu)$ is distributed according to the submatrix ASD, where $N \in \mathbb{R}^{30\times 30}$ is a 30×30 matrix. **Left:** $\text{Bias}(|\widehat{A}_{\mathcal{M}_N}|/n)$ and $\text{Bias}(|\widehat{A}_{\mathrm{R}}|/n)$ versus μ for means $\mu \geq \mu_{\text{detect}}$. **Right:** F-measure of $\widehat{A}_{\mathcal{M}_N}$ and $\widehat{A}_{\mathrm{R}}$ versus μ for means $\mu \geq \mu_{\text{detect}}$.

Proof. Let $p_n = P(A_n)$ and $q_n = P(B_n)$. Then

$$P(A_n \cap B_n) = P(A_n) + P(B_n) - P(A_n \cup B_n) = p_n + q_n - P(A_n \cup B_n) \geq p_n + q_n - 1, \tag{5}$$

where in the last inequality we use that $P(A_n \cup B_n) \leq 1$. Thus,

$$\lim_{n\to\infty} P(A_n \cap B_n) \geq \lim_{n\to\infty} (p_n + q_n - 1) = \left(\lim_{n\to\infty} p_n\right) + \left(\lim_{n\to\infty} q_n\right) - 1 = 1.$$

Since $\lim_{n\to\infty} P(A_n \cap B_n) \leq 1$ by definition, it follows that $\lim_{n\to\infty} P(A_n \cap B_n) = 1$ $\square$

Lemma 2. *Let $X_1, X_2, \ldots$ be a sequence of random variables with $X_n < 1$ for all n. If $\lim_{n\to\infty} P(X_n > C) = 0$ for some $C > 0$, then $E[X_n] < 2C$ for sufficiently large n.*

Proof. We have two cases depending on the value of C. First, suppose $C \geq 1$. Then $X_n < 1 < C$ for all n, and it follows that $E[X_n] < C < 2C$.

Next, suppose $C \in (0, 1)$. Let n be sufficiently large so that $P(X_n > C) < \frac{C}{1-C}$. Then

$$E[X_n] \leq C \cdot P(X_n \leq C) + 1 \cdot P(X_n > C) \leq C \cdot \left(1 - \frac{C}{1-C}\right) + \frac{C}{1-C} = 2C.$$ $\square$

Lemma 3. *Let $X_n \sim N(\mu_n, \sigma_n)$, with $\mu_n, \sigma_n \to \infty$ as $n \to \infty$. Then*

$$\lim_{n\to\infty} P\left(\mu_n - \sqrt{2\sigma_n \log n} \leq X_n \leq \mu_n + \sqrt{2\sigma_n \log n}\right) = 1. \tag{6}$$

Proof. We have

$$\begin{aligned} P(X_n > \mu_n + \sqrt{2\sigma_n \log n}) &= P(Z > \sqrt{2 \log n}), \text{ where } Z \sim N(0,1), \\ &\leq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\log n}} \cdot \frac{1}{n} = O\left(\frac{1}{n}\right), \end{aligned} \tag{7}$$

where in the last inequality we use the standard bound $P(Z \geq x) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2}$. By symmetry, we have

$$P(X_n < \mu_n - \sqrt{2\sigma_n \log n}) \leq O\left(\frac{1}{n}\right) \tag{8}$$

Thus,

$$P\left(\mu_n - \sqrt{2\sigma_n \log n} \leq X_n \leq \mu_n + \sqrt{2\sigma_n \log n}\right) > 1 - O\left(\frac{1}{n}\right).$$

Taking the limit as $n \to \infty$ proves the result. □

Lemma 4. *Suppose $X_v \overset{i.i.d.}{\sim} N(0,1)$ for $v = 1, \ldots, n$. Let $\mathcal{S} \subseteq \mathcal{P}_n$ be a family of subsets of $[n]$ with size $|\mathcal{S}| = \Omega(n)$. For any $k \in [n]$ define $\mathcal{S}_k = \{B \in \mathcal{S} : |B| = k\}$ and $Y_k = \max_{B \in \mathcal{S}_k}\left(\sum_{v \in B} X_v\right)$. Then,*

$$\lim_{n\to\infty} P\left(Y_k \leq \sqrt{2n \log |\mathcal{S}|} \;\; \textit{for all} \;\; k = 1, \ldots, \frac{n}{2}\right) = 1 \tag{9}$$

Proof. Let $t = \sqrt{2n \log |\mathcal{S}|}$ and let Φ be the CDF of the standard normal distribution. Fix $k \in \left\{1, \ldots, \frac{n}{2}\right\}$. We have

$$\begin{aligned} P(Y_k > t) &= P\left(\max_{B \in \mathcal{S}_k} \sum_{v \in B} X_v > t\right) \\ &\leq \sum_{B \in \mathcal{S}_k} P\left(\sum_{v \in B} X_v > t\right) \\ &= |\mathcal{S}_k| \cdot (1 - \Phi(t/\sqrt{k})) \\ &\leq |\mathcal{S}| \cdot (1 - \Phi(t/\sqrt{k})). \end{aligned} \tag{10}$$

where the first inequality uses a union bound and the second equality uses that $\sum_{v \in B} X_v \sim N(0,k)$. Plugging in the standard bound $1 - \Phi(x) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2}$ gives us:

$$\begin{aligned} P\left(\max_{B \in \mathcal{S}_k} \sum_{v \in B} X_v > t\right) &\leq |\mathcal{S}| \cdot (1 - \Phi(t/\sqrt{k})) \\ &\leq |\mathcal{S}| \cdot \frac{1}{\sqrt{2\pi}} \frac{\sqrt{k}}{t} e^{-t^2/2k} \\ &= |\mathcal{S}| \cdot \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{k}{2n \cdot \log |\mathcal{S}|}} \cdot e^{-\frac{n \cdot \log |\mathcal{S}|}{k}} \\ &= \left(\sqrt{\frac{k}{4\pi n}}\right) \cdot \frac{1}{\sqrt{\log |\mathcal{S}|}} \cdot |\mathcal{S}|^{1 - \frac{n}{k}} \\ &\leq \left(\sqrt{\frac{1}{4\pi}}\right) \cdot \frac{1}{|\mathcal{S}| \cdot \sqrt{\log |\mathcal{S}|}}, \text{ since } k \leq \frac{n}{2}. \end{aligned} \tag{11}$$

Taking a union bound over all $k = 1, \ldots, \frac{n}{2}$ gives us

$$\begin{aligned} P\left(\max_{B \in \mathcal{S}_k} \sum_{v \in B} X_v > t \;\text{ for any }\; k = 1, \ldots, \frac{n}{2}\right) &\leq \sum_{k=1}^{n/2} P\left(\max_{B \in \mathcal{S}_k} \sum_{v \in B} X_v > t\right) \\ &\leq \frac{n}{2} \cdot \left(\sqrt{\frac{1}{4\pi}}\right) \cdot \frac{1}{|\mathcal{S}| \cdot \sqrt{\log |\mathcal{S}|}}, \text{ by Equation (11)}, \\ &= \left(\sqrt{\frac{1}{16\pi}}\right) \cdot \frac{n}{|\mathcal{S}| \cdot \sqrt{\log |\mathcal{S}|}} \\ &= O\left(\frac{1}{\sqrt{\log n}}\right) \text{ as } |\mathcal{S}| = \Omega(n). \end{aligned} \tag{12}$$

It follows that for sufficiently large n, there exists a constant $C > 0$ such that

$$P\left(\max_{B\in\mathcal{S}_k}\sum_{v\in B} X_v > t \text{ for any } k = 1,\ldots,\frac{n}{2}\right) \le \frac{C}{\sqrt{\log n}}, \tag{13}$$

so that

$$\begin{aligned}\lim_{n\to\infty} P\left(Y_k \le \sqrt{2n\log|\mathcal{S}|} \text{ for all } k = 1,\ldots,\frac{n}{2}\right) &= 1 - \lim_{n\to\infty} P\left(Y_k > \sqrt{2n\log|\mathcal{S}|} \text{ for any } k = 1,\ldots,\frac{n}{2}\right)\\ &\ge 1 - \lim_{n\to\infty}\left(\frac{C}{\sqrt{\log n}}\right), \text{ by Equation (13)},\\ &= 1,\end{aligned} \tag{14}$$

proving the result. $\square$

Lemma 5. *Let* $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{S}}(A,\mu)$ *where* $|A| = \alpha n$ *for* $0 < \alpha < 0.5$. *Then* $\lim_{n\to\infty} P(|\widehat{A}_{\mathcal{S}}| \le 0.5n) = 1$.

Proof. Let $S \in \mathcal{S}$ be a set with size $|S| > 0.5n$. To prove the claim, it suffices to show that

$$\frac{1}{\sqrt{|A|}}\sum_{v\in A} X_v > \frac{1}{\sqrt{|S|}}\sum_{v\in S} X_v. \tag{15}$$

with high probability.

Note that $\frac{1}{\sqrt{|A|}} > \frac{1}{\sqrt{\alpha n}}$ and $\frac{1}{\sqrt{|S|}} \le \frac{1}{\sqrt{0.5n}}$. Thus, to prove (15) it is sufficient to prove that

$$\frac{1}{\sqrt{0.25n}}\sum_{v\in A} X_v > \frac{1}{\sqrt{0.5n}}\sum_{v\in S} X_v \iff \sum_{v\in A} X_v > \left(\sqrt{2\alpha}\right)\sum_{v\in S} X_v. \tag{16}$$

with high probability.

By independence of the X_v, we have that $\sum_{v\in A} X_v \sim N(\mu\alpha n, \alpha n)$, so by Lemma 3 it follows that $\sum_{v\in A} X_v > \mu\alpha n - \sqrt{2\alpha n\log n}$ with high probability. Similarly, we have that $\sum_{v\in S} X_v \sim N(M, \beta n)$ where $M \le \mu\alpha n$ (as there are at most $|A| = \alpha n$ terms in the sum $\sum_{v\in S} X_v$ with mean μ, and the other terms have mean 0). Thus, by Lemma 3, we also have $\sum_{v\in S} X_v < \mu\alpha n + \sqrt{2|S|\log n}$ with high probability. Putting together the lower bound on $\sum_{v\in A} X_v$ and the upper bound on $\sum_{v\in S} X_v$, (16) can be reduced to

$$\mu\alpha n - \sqrt{2\alpha n\log n} > \left(\sqrt{2\alpha}\right)(\mu\alpha n + \sqrt{2|S|\log n}) \iff \left(1-\sqrt{2\alpha}\right)\mu\alpha n > \sqrt{n\log n}\left(\sqrt{\frac{|S|}{\alpha n}} + \sqrt{2\alpha}\right). \tag{17}$$

Because $\alpha < 0.5$, the LHS is $\Theta(n)$. Since the RHS is $o(n)$, then (17) holds with high probability. Thus, (15) also holds with high probability, and the result follows. $\square$

F.2. Main Lemmas

Lemma 6. *Let* $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{S}}(A,\mu)$ *where* $|\mathcal{S}| = \Omega(n)$ *and* $|A| = \alpha n$ *with* $0 < \alpha < 0.5$. *Suppose* $\lim_{n\to\infty} P(A \subseteq \widehat{A}_{\mathcal{S}}) = 1$. *Then for sufficiently large* n, *we have*

$$\mathrm{Bias}\left(\frac{|\widehat{A}_{\mathcal{S}}|}{n}\right) \le 2\alpha\left(\left(\frac{\mu\alpha n + \sqrt{2n\log|\check{\mathcal{S}}(A)|} + \sqrt{2\alpha n\log n}}{\mu\alpha n - \sqrt{2\alpha n\log n}}\right)^2 - 1\right) + o(1). \tag{18}$$

Proof. We will first derive the $o(1)$ term in Equation (18). Let $X_S = \sum_{v\in S} X_v$, and define the following events:

$$D_n = \left[|\widehat{A}_{\mathcal{S}}| \le \frac{n}{2}\right]$$

$$E_n = [A \subseteq \widehat{A}_{\mathcal{S}}],$$
$$F_n = \left[\mu\alpha n - \sqrt{2\alpha n \log n} \leq X_A \leq \mu\alpha n + \sqrt{2\alpha n \log n}\right],$$
$$G_n = \left[\max_{\substack{B \in \breve{\mathcal{S}}(A) \\ |B| \leq \frac{n}{2}}} X_{B \setminus A} \leq \sqrt{2n \log |\breve{\mathcal{S}}(A)|}\right].$$

Let $H_n = D_n \cap E_n \cap F_n \cap G_n$. We claim that $\lim\limits_{n\to\infty} P(H_n) = 1$.

To prove this claim, first note that $\lim\limits_{n\to\infty} P(D_n) = 1$ by Lemma 5 and $\lim\limits_{n\to\infty} P(E_n) = 1$ by assumption. Moreover, because $X_A \sim N(\mu\alpha n, \alpha n)$, it follows from Lemma 3 that $\lim\limits_{n\to\infty} P(F_n) = 1$. Finally, by applying Lemma 4 with the anomaly family $\breve{\mathcal{S}}(A)$, we have that $\lim\limits_{n\to\infty} P(G_n) = 1$. Thus, by a repeated application of Lemma 1, we have $\lim\limits_{n\to\infty} P(H_n) = \lim\limits_{n\to\infty} P(D_n \cap E_n \cap F_n \cap G_n) = 1$.

Now define $p_n = P(H_n)$. Then, we have

$$\begin{aligned}
\text{Bias}\left(\frac{|\widehat{A}_{\mathcal{S}}|}{n}\right) &= p_n \cdot \text{Bias}\left(\frac{|\widehat{A}_{\mathcal{S}}|}{n} \;\middle|\; H_n\right) + (1-p_n) \cdot \text{Bias}\left(\frac{|\widehat{A}_{\mathcal{S}}|}{n} \;\middle|\; H_n^c\right) \\
&\leq \text{Bias}\left(\frac{|\widehat{A}_{\mathcal{S}}|}{n} \;\middle|\; H_n\right) + (1-p_n) \qquad (19) \\
&= \text{Bias}\left(\frac{|\widehat{A}_{\mathcal{S}}|}{n} \;\middle|\; H_n\right) + o(1),
\end{aligned}$$

where in the second line we use that $p_n \leq 1$ and $\text{Bias}\left(\frac{|\widehat{A}_{\mathcal{S}}|}{n} \;\middle|\; H_n^c\right) \leq 1$, and in the third line we use that $\lim\limits_{n\to\infty} P(H_n) = 1$.

To complete the proof, we will bound $\text{Bias}\left(\frac{|\widehat{A}_{\mathcal{S}}|}{n} \;\middle|\; H_n\right)$. Since the bias term conditions on $H_n = D_n \cap E_n \cap F_n \cap G_n$, for the rest of the proof we will assume that the events D_n, E_n, F_n, and G_n hold.

Since E_n holds, we have that

$$\Gamma(A) = \frac{1}{\sqrt{|A|}} \sum_{v \in A} X_v = \frac{1}{\sqrt{\alpha n}} X_A$$
$$\Gamma(\widehat{A}_{\mathcal{S}}) = \frac{1}{\sqrt{|\widehat{A}_{\mathcal{S}}|}} \sum_{v \in \widehat{A}_{\mathcal{S}}} X_v = \frac{1}{\sqrt{|\widehat{A}_{\mathcal{S}}|}} \left(X_A + X_{\widehat{A}_{\mathcal{S}} \setminus A}\right),$$

We will find lower and upper bounds for $X_{\widehat{A}_{\mathcal{S}} \setminus A}$ in terms of $|\widehat{A}_{\mathcal{S}}|$, and use those bounds to derive (18).

We start by finding a lower bound for $X_{\widehat{A}_{\mathcal{S}} \setminus A}$. Since F_n holds, we have:

$$\mu\alpha n - \sqrt{2\alpha n \log n} \leq X_A \leq \mu\alpha n + \sqrt{2\alpha n \log n}. \qquad (20)$$

Combining (20) and the fact that $\Gamma(A) \leq \Gamma(\widehat{A}_{\mathcal{S}})$ yields

$$\frac{1}{\sqrt{\alpha n}}(\mu\alpha n - \sqrt{2\alpha n \log n}) \leq \Gamma(A) \leq \Gamma(\widehat{A}_{\mathcal{S}}) \leq \frac{1}{\sqrt{|\widehat{A}_{\mathcal{S}}|}} \left(\mu\alpha n + \sqrt{2\alpha n \log n} + X_{\widehat{A}_{\mathcal{S}} \setminus A}\right). \qquad (21)$$

By assumption, $\widehat{A}_{\mathcal{S}} \setminus A \neq \varnothing$. Thus, solving for $X_{\widehat{A}_{\mathcal{S}} \setminus A}$ gives us a lower bound on $X_{\widehat{A}_{\mathcal{S}} \setminus A}$:

$$X_{\widehat{A}_{\mathcal{S}} \setminus A} \geq \sqrt{\frac{|\widehat{A}_{\mathcal{S}}|}{\alpha n}} \left(\mu\alpha n - \sqrt{2\alpha n \log n}\right) - \mu\alpha n - \sqrt{2\alpha n \log n}. \qquad (22)$$

Next, we find an upper bound for $X_{\widehat{A}_{\mathcal{S}}\setminus A}$. Since D_n holds, we have $|\widehat{A}_{\mathcal{S}}| \leq \frac{n}{2}$. Since G_n also holds, we have

$$X_{\widehat{A}_{\mathcal{S}}\setminus A} \leq \max_{\substack{B\in\breve{\mathcal{S}}(A)\\ |B|\leq \frac{n}{2}}} X_{B\setminus A} \leq \sqrt{2n\log|\breve{\mathcal{S}}(A)|}. \tag{23}$$

Combining the lower bound from (22) and the upper bound from (23) yields

$$\sqrt{\frac{|\widehat{A}_{\mathcal{S}}|}{\alpha n}}\left(\mu\alpha n - \sqrt{2\alpha n\log n}\right) - \mu\alpha n - \sqrt{2\alpha n\log n} \leq X_{\widehat{A}_{\mathcal{S}}\setminus A} \leq \sqrt{2n\log|\breve{\mathcal{S}}(A)|} \tag{24}$$

Thus, the LHS of (24) is less than the RHS of (24), i.e.

$$\sqrt{\frac{|\widehat{A}_{\mathcal{S}}|}{\alpha n}}\left(\mu\alpha n - \sqrt{2\alpha n\log n}\right) - \mu\alpha n - \sqrt{2\alpha n\log n} \leq \sqrt{2n\log|\breve{\mathcal{S}}(A)|}. \tag{25}$$

Rearranging (25) yields

$$\frac{|\widehat{A}_{\mathcal{S}}|}{n} - \alpha \leq \alpha\left(\left(\frac{\mu\alpha n + \sqrt{2n\log|\breve{\mathcal{S}}(A)|} + \sqrt{2\alpha n\log n}}{\mu\alpha n - \sqrt{2\alpha n\log n}}\right)^2 - 1\right). \tag{26}$$

So by Lemma 2, we have

$$\text{Bias}\left(\frac{|\widehat{A}_{\mathcal{S}}|}{n}\,\middle|\, H_n\right) = E\left[\frac{|\widehat{A}_{\mathcal{S}}|}{n} - \alpha \,\middle|\, H_n\right] \leq 2\alpha\left(\left(\frac{\mu\alpha n + \sqrt{2n\log|\breve{\mathcal{S}}(A)|} + \sqrt{2\alpha n\log n}}{\mu\alpha n - \sqrt{2\alpha n\log n}}\right)^2 - 1\right). \tag{27}$$

The result follows by combining Equations (19) and (27). □

Lemma 7. *Let* $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A,\mu)$ *where* $|\mathcal{S}| = \Omega(n)$ *and* $|A| = \alpha n$ *with* $0 < \alpha < 0.5$. *Assume* $\lim_{n\to\infty} P(A \subseteq \widehat{A}_{\mathcal{S}}) = 1$. *If Bias*$(|\widehat{A}_{\mathcal{S}}|/n) \geq \gamma$, *then*

$$|\breve{\mathcal{S}}(A)| \geq (C_{\mu,\gamma,\alpha})^n \cdot e^{-\Theta(\sqrt{n\log n})} \tag{28}$$

for sufficiently large n, *where* $C_{\mu,\alpha,\gamma} = \exp\left(\frac{1}{2}\mu^2\alpha^2\left(\sqrt{1+\frac{\gamma}{4\alpha}} - 1\right)^2\right)$.

Proof. By Lemma 6, we have

$$\gamma \leq \text{Bias}(|\widehat{A}_{\mathcal{S}}|/n) \leq 2\alpha\left(\left(\frac{\mu\alpha n + \sqrt{2n\log|\breve{\mathcal{S}}(A)|} + \sqrt{2\alpha n\log n}}{\mu\alpha n - \sqrt{2\alpha n\log n}}\right)^2 - 1\right) + o(1). \tag{29}$$

Thus, the LHS of (29) is less than the RHS of (29), i.e.

$$\gamma \leq 2\alpha\left(\left(\frac{\mu\alpha n + \sqrt{2n\log|\breve{\mathcal{S}}(A)|} + \sqrt{2\alpha n\log n}}{\mu\alpha n - \sqrt{2\alpha n\log n}}\right)^2 - 1\right) + o(1) \tag{30}$$

Let n be sufficiently large so that the $o(1)$ term in (30) is less than $\frac{\gamma}{2}$. Then, solving for $|\breve{\mathcal{S}}(A)|$ in (30):

$$\gamma \leq 2\alpha\left(\left(\frac{\mu\alpha n + \sqrt{2n\log|\breve{\mathcal{S}}(A)|} + \sqrt{2\alpha n\log n}}{\mu\alpha n - \sqrt{2\alpha n\log n}}\right)^2 - 1\right) + \frac{\gamma}{2}$$

$$\Rightarrow \sqrt{\frac{\gamma}{4\alpha}+1} \le \frac{\mu\alpha n + \sqrt{2n\log|\breve{\mathcal{S}}(A)|} + \sqrt{2\alpha n \log n}}{\mu\alpha n - \sqrt{2\alpha n\log n}}$$

$$\Rightarrow \mu\alpha n\left(\sqrt{\frac{\gamma}{4\alpha}+1}-1\right) - \Theta(\sqrt{n\log n}) \le \sqrt{2n\log|\breve{\mathcal{S}}(A)|}$$

$$\Rightarrow |\breve{\mathcal{S}}(A)| \ge \left[\exp\left(\frac{1}{2}\mu^2\alpha^2\left(\sqrt{\frac{\gamma}{4\alpha}+1}-1\right)^2\right)\right]^n \cdot e^{-\Theta(\sqrt{n\log n})}$$

completing the proof. □

F.3. Proof of Theorem

Using the above lemmas, we are now ready to prove Theorem 1 from the main text.

Theorem 1. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathrm{ASD}_{\mathcal{S}}(A, \mu)$ *where* $\mathcal{S} = \Omega(n)$ *and* $|A| = \alpha n$ *with* $0 < \alpha < 0.5$. *Suppose* $|\breve{\mathcal{S}}(A)|$ *is sub-exponential in* n *and* $\lim_{n\to\infty} P(A \subseteq \widehat{A}_{\mathcal{S}}) = 1$. *Then* $\lim_{n\to\infty} \mathrm{Bias}(|\widehat{A}_{\mathcal{S}}|/n) = 0$.

Proof. Let $\gamma > 0$ and let $C_{\mu,\alpha,\gamma}$ be as defined in Lemma 7. Note that because $C_{\mu,\alpha,\gamma} > 1$, we have $\frac{2C_{\mu,\alpha,\gamma}}{1+C_{\mu,\alpha,\gamma}} > 1$.

Because $|\breve{\mathcal{S}}(A)|$ is sub-exponential, there exists sufficiently large n so that

$$|\breve{\mathcal{S}}(A)| < \left(\frac{2C_{\mu,\alpha,\gamma}}{1+C_{\mu,\alpha,\gamma}}\right)^n. \tag{31}$$

We also note that because $C_{\mu,\alpha,\gamma} > 1$, then $\frac{2}{1+C_{\mu,\alpha,\gamma}} < 1$. For sufficiently large n, $e^{-\Theta(\sqrt{\log n/n})}$ will get arbitrarily close to 1. Thus, we have

$$e^{-\Theta(\sqrt{\log n/n})} \ge \frac{2}{1+C_{\mu,\alpha,\gamma}}. \tag{32}$$

Combining both (31) and (32) gives us

$$C^n_{\mu,\alpha,\gamma} \cdot e^{-\Theta(\sqrt{n\log n})} = \left(C_{\mu,\alpha,\gamma}\cdot e^{-\Theta(\sqrt{\log n/n})}\right)^n \ge \left(C_{\mu,\alpha,\gamma}\cdot \frac{2}{1+C_{\mu,\alpha,\gamma}}\right)^n > |\breve{\mathcal{S}}(A)|, \tag{33}$$

for sufficiently large n, where the first inequality follows by (32) and the second inequality follows by (31).

Thus, by the contrapositive of Lemma 7, it follows that $\mathrm{Bias}(|\widehat{A}_{\mathcal{S}}|/n) < \gamma$ for sufficiently large n. Taking the limit as $n \to \infty$ yields

$$\lim_{n\to\infty} \mathrm{Bias}(|\widehat{A}_{\mathcal{S}}|/n) \le \gamma. \tag{34}$$

Because (34) holds for all $\gamma > 0$, it follows that $\lim_{n\to\infty} \mathrm{Bias}(|\widehat{A}_{\mathcal{S}}|/n) \le 0$. Furthermore, because $\lim_{n\to\infty} P(A \subseteq \widehat{A}_{\mathcal{S}}) = 1$, we also have $\lim_{n\to\infty} \mathrm{Bias}(|\widehat{A}_{\mathcal{S}}|/n) \ge 0$. Thus, $\lim_{n\to\infty} \mathrm{Bias}(|\widehat{A}_{\mathcal{S}}|/n) = 0$, as desired. □

G. Proof of Theorem 2

In the following proof, we slightly abuse notation and assume that all statements of the form $\lim_{n\to\infty} R_n = Y$, where R_n and Y are random variables, hold almost surely.

Theorem 2. *Let* $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$ *where* $|A| = \alpha n$ *with* $0 < \alpha < 0.5$. *Then* $\lim_{n\to\infty} \mathrm{Bias}(|\widehat{A}_{\mathcal{P}_n}|/n) > 0$.

Proof. From Proposition 1 in the main text, we have

$$\widehat{A}_{\mathcal{P}_n} = \operatorname*{argmax}_{S\subseteq[n]} \frac{1}{\sqrt{|S|}}\sum_{v\in S} X_v. \tag{35}$$

Because the maximum is taken over all subsets of $[n]$, an equivalent formulation of the above is $\widehat{A}_{\mathcal{P}_n} = \{v : X_v > \widehat{T}_{\mathcal{S}}\}$, where

$$\widehat{T}_{\mathcal{S}} = \operatorname*{argmax}_{T\in\mathbb{R}} \left(\frac{1}{\sqrt{\#\{v\in[n] : X_v > T\}}} \sum_{v\in[n]:X_v>T} X_v \right). \tag{36}$$

We start by showing that $\lim_{n\to\infty} \widehat{T}_{\mathcal{S}}$ is finite. To do so, we will find an expression for the RHS as $n\to\infty$. Let $M_T = \{v \in [n] : X_v > T, v\in A\}$ and $N_T = \{v\in[n] : X_v > T, v\notin A\}$. Additionally, let $\nu_{\mu,T}$ be the mean of a $N(\mu,1)$ distribution that is truncated to be above T. Then

$$\sum_{v\in[n]:X_v>T} X_v = \left(\frac{\sum_{v\in M_T} X_v}{|M_T|}\right)\cdot|M_T| + \left(\frac{\sum_{v\in N_T} X_v}{|N_T|}\right)\cdot|N_T| \tag{37}$$

By the strong law of large numbers, $\lim_{n\to\infty} \frac{\sum_{v\in M_T} X_v}{|M_T|} = \nu_{\mu,T}$ and $\lim_{n\to\infty} \frac{\sum_{v\in M_T} X_v}{|M_T|} = \nu_{0,T}$. Similarly, $\lim_{n\to\infty} \frac{|M_T|}{\alpha n\cdot(1-\Phi(T-\mu))} = 1$ and $\lim_{n\to\infty} \frac{|N_T|}{(1-\alpha)n\cdot(1-\Phi(T))} = 1$, where Φ is the CDF of a standard normal. Plugging these limits into (37) gives us

$$\lim_{n\to\infty} \frac{\sum_{v\in[n]:X_v>T} X_v}{\nu_{\mu,T}\cdot(\alpha n(1-\Phi(T-\mu)) + \nu_{0,T}\cdot((1-\alpha)n(1-\Phi(T))} = 1. \tag{38}$$

A similar calculation yields

$$\lim_{n\to\infty} \frac{\sqrt{\#\{v\in[n] : X_v > T\}}}{\sqrt{\alpha n(1-\Phi(T-\mu)) + (1-\alpha)n(1-\Phi(T))}} = 1. \tag{39}$$

Plugging in Equations (38) and (39) into Equation (36) yields

$$\begin{aligned}
\lim_{n\to\infty}\widehat{T}_{\mathcal{S}} &= \lim_{n\to\infty}\left[\operatorname*{argmax}_{T\in\mathbb{R}}\left(\frac{\nu_{\mu,T}\cdot(\alpha n(1-\Phi(T-\mu)) + \nu_{0,T}\cdot((1-\alpha)n(1-\Phi(T))}{\sqrt{\alpha n(1-\Phi(T-\mu)) + (1-\alpha)n(1-\Phi(T))}}\right)\right] \\
&= \lim_{n\to\infty}\left[\operatorname*{argmax}_{T\in\mathbb{R}}\left(\frac{\nu_{\mu,T}\cdot(\alpha(1-\Phi(T-\mu)) + \nu_{0,T}\cdot((1-\alpha)(1-\Phi(T))}{\sqrt{\alpha(1-\Phi(T-\mu)) + (1-\alpha)(1-\Phi(T))}}\cdot\sqrt{n}\right)\right] \\
&= \operatorname*{argmax}_{T\in\mathbb{R}}\left(\frac{\nu_{\mu,T}\cdot(\alpha(1-\Phi(T-\mu)) + \nu_{0,T}\cdot((1-\alpha)(1-\Phi(T))}{\sqrt{\alpha(1-\Phi(T-\mu)) + (1-\alpha)(1-\Phi(T))}}\right).
\end{aligned} \tag{40}$$

Thus $\lim_{n\to\infty}\widehat{T}_{\mathcal{S}}$ is finite.

Next, define $T^* = \lim_{n\to\infty}\widehat{T}_{\mathcal{S}}$. To complete the proof, we use T^* to derive an expression for $\lim_{n\to\infty}\frac{|\widehat{A}_{\mathcal{P}_n}|}{n}$, and then use that expression to bound $\lim_{n\to\infty}\text{Bias}\left(\frac{|\widehat{A}_{\mathcal{P}_n}|}{n}\right)$.

Since the fraction of observations X_i such that $i\in A$ and $X_i > \widehat{T}_{\mathcal{S}}$ is asymptotically $\frac{|A|}{n}\cdot(1-\Phi(\widehat{T}_{\mathcal{S}}-\mu))$, it follows that

$$\lim_{n\to\infty}\frac{|A\cap\widehat{A}_{\mathcal{P}_n}|}{n} = \lim_{n\to\infty}\left(\frac{|A|}{n}\cdot(1-\Phi(\widehat{T}_{\mathcal{S}}-\mu))\right) = \alpha\cdot(1-\Phi(T^*-\mu)). \tag{41}$$

Similarly, the fraction of observations X_i such that $i\notin A$ and $X_i > \widehat{T}_{\mathcal{S}}$ is asymptotically $\left(1-\frac{|A|}{n}\right)\cdot(1-\Phi(\widehat{T}_{\mathcal{S}}))$, so we have

$$\lim_{n\to\infty}\frac{|A^c\cap\widehat{A}_{\mathcal{P}_n}|}{n} = \lim_{n\to\infty}\left(\left(1-\frac{|A|}{n}\right)\cdot(1-\Phi(\widehat{T}_{\mathcal{S}}))\right) = (1-\alpha)\cdot(1-\Phi(T^*)). \tag{42}$$

Combining Equations (41) and (42) gives us

$$\lim_{n\to\infty} \frac{|\widehat{A}_{\mathcal{P}_n}|}{n} = \alpha \cdot (1 - \Phi(T^* - \mu)) + (1-\alpha) \cdot (1 - \Phi(T^*)) \tag{43}$$

Thus, the asymptotic bias of the MLE $\widehat{A}_{\mathcal{S}}$ is

$$\begin{aligned}\lim_{n\to\infty} \text{Bias}(|\widehat{A}_{\mathcal{P}_n}|/n) &= \lim_{n\to\infty} E[|\widehat{A}_{\mathcal{P}_n}|/n] - \alpha \\ &= \alpha \cdot (1 - \Phi(T^* - \mu)) + (1-\alpha) \cdot (1 - \Phi(T^*)) - \alpha.\end{aligned}$$

Since the above expression is always positive for $\alpha < 0.5$, so it follows that $\lim_{n\to\infty} \text{Bias}(|\widehat{A}_{\mathcal{P}_n}|/n) > 0$. □

H. Proof of Theorem 3 and Corollaries 1, 2

To prove Theorem 3, we require the following Lemma.

Lemma 8. *Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Then*

$$P\left(\max_{i\in[n]} X_i \le \mu + 2\sigma\sqrt{\log n}\right) \ge 1 - \frac{1}{n}. \tag{44}$$

Proof. For fixed $i \in [n]$ we have

$$\begin{aligned}P(X_i > \mu + 2\sigma\sqrt{\log n}) &= P(Z > 2\sqrt{\log n}), \text{where } Z \sim N(0,1) \\ &\le \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2\sqrt{\log n}} \cdot \frac{1}{n^2} < \frac{1}{n^2}\end{aligned} \tag{45}$$

where in the second line we use the standard bound $P(Z \ge x) \le \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{x} e^{-x^2/2}$. Thus, $P(X_i \le \mu + 2\sigma\sqrt{\log n}) \ge 1 - \frac{1}{n^2}$. By a union bound, it follows that

$$P(X_i \le \mu + 2\sigma\sqrt{\log n} \quad \text{for all } i \in [n]) \ge 1 - \frac{1}{n}, \tag{46}$$

which implies the desired result. □

Theorem 3. *Let $\mathbf{X} = (X_1, \dots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$, where $|A| = \alpha n$ for $0 < \alpha < 0.5$ and $\mu \ge C\sqrt{\log n}$ for a sufficiently large constant $C > 0$. For sufficiently large n, we have that*

$$|\widehat{\alpha}_{GMM} - \alpha| \le \sqrt{\frac{\log n}{n}} \quad \textit{and} \quad |\widehat{\mu}_{GMM} - \mu| \le 3\sqrt{\frac{\log n}{n}}$$

with probability at least $1 - \frac{1}{n}$.

Proof. For $\widetilde{\alpha} \in (0,1)$ and $\widetilde{\mu} > 0$, let $L_{\widetilde{\alpha},\widetilde{\mu}}(x) = \log\left(\widetilde{\alpha} \cdot \exp\left(-\frac{1}{2}(x-\widetilde{\mu})^2\right) + (1-\widetilde{\alpha}) \cdot \exp\left(-\frac{1}{2}x^2\right)\right)$ be the (scaled) log-likelihood function for the mixture distribution $\widetilde{\alpha} \cdot N(\widetilde{\mu}, 1) + (1-\widetilde{\alpha}) \cdot N(0,1)$, and define $L_{\widetilde{\alpha},\widetilde{\mu}}(\mathbf{X}) = \prod_{i=1}^n L_{\widetilde{\alpha},\widetilde{\mu}}(X_i)$. Then

$$\widehat{\alpha}_{\text{GMM}}, \widehat{\mu}_{\text{GMM}} = \underset{\substack{\widetilde{\alpha}\in(0,1) \\ \widetilde{\mu}\in(0,\infty)}}{\text{argmax}}\, L_{\widetilde{\alpha},\widetilde{\mu}}(\mathbf{X}) \tag{47}$$

To prove the claim, it suffices to show that if $|\widehat{\alpha}_{\text{GMM}} - \widetilde{\alpha}| > \sqrt{\frac{\log n}{n}}$ or $|\widehat{\mu}_{\text{GMM}} - \widetilde{\mu}| > 3\sqrt{\frac{\log n}{n}}$, then $L_{\alpha,\mu}(\mathbf{X}) > L_{\widetilde{\alpha},\widetilde{\mu}}(\mathbf{X})$ with probability at least $1 - \frac{1}{n}$.

We will prove the following equivalent formulation: if κ, τ are real numbers such that $|\kappa| > \sqrt{\frac{\log n}{n}}$ or $|\tau| > 3\sqrt{\frac{\log n}{n}}$, then

$$L_{\alpha,\mu}(\mathbf{X}) > L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) \tag{48}$$

with probability at least $1 - \frac{1}{n}$.

We proceed by a case analysis based on whether κ and τ also satisfy the following additional conditions:

$$\frac{1}{n^2} < \alpha + \kappa < 1 - \frac{1}{n^2}, \tag{49}$$

$$\frac{\mu^2}{\tau^2} > 100. \tag{50}$$

Briefly, the intuition for the above conditions is that if κ and τ satisfy (49) and (50), then we can derive a simplified formula for the likelihood $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})$.

In all cases, we assume that $\mu - 2\sqrt{\log n} \le X_i \le \mu + 2\sqrt{\log n}$ for $i \in A$ and $-2\sqrt{\log n} \le X_i \le 2\sqrt{\log n}$ for $i \notin A$, as these events hold with probability at least $1 - \frac{1}{n}$ by Lemma 8.

Case 1: κ satisfies (49) and τ satisfies (50). The log-likelihood $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})$ can be written as

$$\begin{aligned} L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) &= \log\left(\prod_{i=1}^{n}\left((\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right) + (1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right)\right)\right) \\ &= \sum_{i=1}^{n}\log\left((\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right) + (1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right)\right) \end{aligned} \tag{51}$$

Let $T_1 = (\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right)$ be the first term in the logarithm and let $T_2 = (1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right)$ be the second term. We claim that if $i \in A$, then $T_1 + T_2 = (1+o(n^{-1}))T_1$, while if $i \notin A$ then $T_1 + T_2 = (1+o(n^{-1}))T_2$.

To show that $T_1 + T_2 = (1+o(n^{-1}))T_1$ for $i \in A$, we compute $\frac{T_2}{T_1}$:

$$\begin{aligned} \frac{T_2}{T_1} &= \frac{(1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right)}{(\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right)} \\ &= \left(\frac{1-\alpha-\kappa}{\alpha+\kappa}\right)\cdot\exp\left(-X_i(\mu+\tau)+\frac{1}{2}(\mu+\tau)^2\right) \\ &\le \left(\frac{1-\alpha-\kappa}{\alpha+\kappa}\right)\cdot\exp\left(-(\mu-2\sqrt{\log n})(\mu+\tau)+\frac{1}{2}(\mu+\tau)^2\right) \\ &= \left(\frac{1-\alpha-\kappa}{\alpha+\kappa}\right)\cdot\exp\left(-\frac{1}{2}\mu^2+\frac{1}{2}\tau^2+2\mu\sqrt{\log n}+2\tau\sqrt{\log n}\right) \\ &\le n^2\cdot\exp\left(-\frac{1}{2}\mu^2+\frac{1}{2}\left(\frac{\mu^2}{100}\right)+2\mu\sqrt{\log n}+2\left(\frac{\mu}{10}\right)\sqrt{\log n}\right) \\ &= n^2\cdot\exp\left(-\frac{99}{200}\mu^2+\frac{11}{5}\mu\sqrt{\log n}\right), \end{aligned} \tag{52}$$

where the first inequality uses that $X_i \ge \mu - 2\sqrt{\log n}$ and the second inequality uses that $\tau \le \frac{\mu}{10}$ (which follows from (50)).

Now $-\frac{99}{200}t^2 + \frac{11}{5}t\sqrt{\log n}$ is a concave quadratic with a maximum at $t = \frac{20}{9}\sqrt{\log n}$. Since $\mu \ge C\sqrt{\log n} > \frac{20}{9}\sqrt{\log n}$ (for sufficiently large C), it follows that $-\frac{99}{200}\mu^2 + \frac{11}{5}\mu\sqrt{\log n} \le -\frac{99}{200}(C\sqrt{\log n})^2 + \frac{11}{5}(C\sqrt{\log n})\cdot\sqrt{\log n}$. Plugging this into the above equation yields

$$\begin{aligned} n^2\cdot\exp\left(-\frac{99}{200}\mu^2+\frac{11}{5}\mu\sqrt{\log n}\right) &\le n^2\cdot\exp\left(-\frac{99}{200}(C\sqrt{\log n})^2+\frac{11}{5}(C\sqrt{\log n})\cdot\sqrt{\log n}\right) \\ &= n^2\cdot n^{-\frac{99}{200}C^2+\frac{440}{200}C} \\ &= o(n^{-1}) \text{ for sufficiently large } C. \end{aligned} \tag{53}$$

Thus, $\frac{T_2}{T_1} = o(n^{-1})$, which implies $T_1 + T_2 = (1 + o(n^{-1}))T_1$ for $i \in A$. By a similar derivation, we also have that $T_1 + T_2 = (1 + o(n^{-1}))T_2$ for $i \notin A$.

Using these relationships between T_1 and T_2, we rewrite the log-likelihood in Equation (51) as

$$\begin{aligned}
L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) &= \sum_{i\in A} \log\left((1+o(n^{-1}))\cdot T_1\right) + \sum_{i\notin A} \log\left((1+o(n^{-1}))\cdot T_2\right) \\
&= \sum_{i\in A} \log\left((1+o(n^{-1}))\cdot(\alpha+\kappa)\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right)\right) + \sum_{i\notin A}\log\left((1+o(n^{-1}))\cdot(1-\alpha-\kappa)\exp\left(-\frac{X_i^2}{2}\right)\right) \\
&= \alpha n\cdot\log(\alpha+\kappa) + (1-\alpha)n\cdot\log(1-\alpha-\kappa) - \frac{1}{2}\sum_{i\in A}(X_i-\mu-\tau)^2 - \frac{1}{2}\sum_{i\notin A} X_i^2 + o(1).
\end{aligned} \tag{54}$$

Plugging in $\kappa = \tau = 0$ into (54) yields the following expression for the log-likelihood $L_{\alpha,\mu}(\mathbf{X})$ with the true parameters α, μ:

$$L_{\alpha,\mu}(\mathbf{X}) = \alpha n\cdot\log\alpha + (1-\alpha)n\cdot\log(1-\alpha) - \frac{1}{2}\sum_{i\in A}(X_i-\mu)^2 - \frac{1}{2}\sum_{i\notin A}X_i^2 + o(1). \tag{55}$$

So after equating equations (54) and (55) and simplifying, we have that $L_{\alpha,\mu}(\mathbf{X}) > L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})$ is equivalent to

$$\begin{aligned}
&L_{\alpha,\mu}(\mathbf{X}) > L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) \\
&\quad\Leftrightarrow \alpha\log\left(\frac{\alpha}{\alpha+\kappa}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-\alpha-\kappa}\right) + \frac{\tau\alpha}{2}\left(\tau - 2\frac{\sum_{i\in A}(X_i-\mu)}{\alpha n}\right) + o(n^{-1}) > o(n^{-1}),
\end{aligned}$$

To prove that $L_{\alpha,\mu}(\mathbf{X}) > L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})$ and complete the proof, it suffices to show that

$$\alpha\log\left(\frac{\alpha}{\alpha+\kappa}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-\alpha-\kappa}\right) + \frac{\tau\alpha}{2}\left(\tau - 2\frac{\sum_{i\in A}(X_i-\mu)}{\alpha n}\right) = \Omega(n^{-1}). \tag{56}$$

To bound the above inequality, we first note that the first two terms $\alpha\log\left(\frac{\alpha}{\alpha+\kappa}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-\alpha-\kappa}\right)$ are the KL-divergence between a Bern(α) random variable and a Bern($\alpha+\kappa$) random variable. By Pinsker's inequality, we have

$$\begin{aligned}
\alpha\log\left(\frac{\alpha}{\alpha+\kappa}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-\alpha-\kappa}\right) &= D_{KL}\left(\text{Bern}(\alpha)\ ||\ \text{Bern}(\alpha+\kappa)\right) \\
&\geq 2\left[d_{TV}\left(\text{Bern}(\alpha), \text{Bern}(\alpha+\kappa)\right)\right]^2 \\
&= 2\kappa^2.
\end{aligned} \tag{57}$$

Second, we note that $\frac{\sum_{i\in A}(X_i-\mu)}{\alpha n} \sim N(0, \frac{1}{\alpha n})$, so by Lemma 8, we have that $\frac{\sum_{i\in A}(X_i-\mu)}{\alpha n} < \sqrt{\frac{\log n}{\alpha n}}$ with high probability. Thus,

$$\alpha\log\left(\frac{\alpha}{\alpha+\kappa}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-\alpha-\kappa}\right) + \frac{\tau\alpha}{2}\left(\tau - 2\frac{\sum_{i\in A}(X_i-\mu)}{\alpha n}\right) \geq 2\kappa^2 + \frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right) \tag{58}$$

To prove that the RHS of (58) is $\Omega(n^{-1})$, we use casework depending on whether $|\kappa| > \sqrt{\frac{\log n}{n}}$ or $|\tau| > 3\sqrt{\frac{\log n}{n}}$.

Case 1, Sub-case 1: $|\kappa| > \sqrt{\frac{\log n}{n}}$.

We bound the RHS of (58) as

$$\begin{aligned}
2\kappa^2 + \frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right) &\geq 2\frac{\log n}{n} + \frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right) \\
&\geq 2\frac{\log n}{n} - \frac{\alpha}{2}\cdot\frac{\log n}{n} \\
&= \frac{\log n}{n}\cdot\left(2 - \frac{\alpha}{2}\right) \\
&= \Omega(n^{-1}),
\end{aligned} \tag{59}$$

where the second inequality uses that $\frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right)$ is a quadratic in τ whose minimum is $-\frac{\alpha}{2}\frac{\log n}{n}$.

Case 1, Sub-case 2: $|\tau| > 3\sqrt{\frac{\log n}{n}}$.

Note that the condition on τ implies that either $\tau > 3\sqrt{\frac{\log n}{n}}$ or $\tau < -3\sqrt{\frac{\log n}{n}}$.

We also note that $\frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right)$ is a quadratic in τ that is decreasing for $\tau < \sqrt{\frac{\log n}{n}}$ and is increasing for $\tau > \sqrt{\frac{\log n}{n}}$. Depending on the value of τ, we lower bound $\frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right)$ as follows:

$$\tau > 3\sqrt{\frac{\log n}{n}} \implies \frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right) > \frac{\alpha}{2}\left(3\sqrt{\frac{\log n}{n}}\right)\left(3\sqrt{\frac{\log n}{n}} - 2\sqrt{\frac{\log n}{n}}\right) = \frac{3\alpha}{2}\frac{\log n}{n}$$

$$\tau < -3\sqrt{\frac{\log n}{n}} \implies \frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right) > \frac{\alpha}{2}\left(-3\sqrt{\frac{\log n}{n}}\right)\left(-3\sqrt{\frac{\log n}{n}} - 2\sqrt{\frac{\log n}{n}}\right) = \frac{15\alpha}{2}\frac{\log n}{n}.$$

In either case, we have that $\frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right) > \frac{3\alpha}{2}\frac{\log n}{n}$. Thus the RHS of (58) is

$$2\kappa^2 + \frac{\tau\alpha}{2}\left(\tau - 2\sqrt{\frac{\log n}{n}}\right) \geq \frac{3\alpha}{2}\frac{\log n}{n} = \Omega(n^{-1}), \tag{60}$$

as desired.

Case 2: κ **does not satisfy** (49)**,** τ **satisfies** (50)**.** This means that either $\alpha + \kappa < \frac{1}{n^2}$ or $\alpha + \kappa > 1 - \frac{1}{n^2}$. We will treat each of these sub-cases separately.

Before doing so, we require the following lower bound on $L_{\alpha,\mu}(\mathbf{X})$: for sufficiently large n, $L_{\alpha,\mu}(\mathbf{X}) > -n$. To prove this lower bound, from (55) we have

$$\begin{aligned}
L_{\alpha,\mu}(\mathbf{X}) &= n\cdot\log(1+o(n^{-1})) + \alpha n\cdot\log\alpha + (1-\alpha)n\cdot\log(1-\alpha) - \frac{1}{2}\sum_{i\in A}(X_i-\mu)^2 - \frac{1}{2}\sum_{i\notin A}X_i^2 \\
&= n\left(\log(1+o(n^{-1})) + H(\alpha) - \frac{\alpha}{2}\left(\frac{\sum_{i\in A}(X_i-\mu)^2}{\alpha n}\right) - \frac{1-\alpha}{2}\left(\frac{\sum_{i\notin A}X_i^2}{(1-\alpha)n}\right)\right) \\
&\geq n\left(-\frac{\alpha}{2}\left(\frac{\sum_{i\in A}(X_i-\mu)^2}{\alpha n}\right) - \frac{1-\alpha}{2}\left(\frac{\sum_{i\notin A}X_i^2}{(1-\alpha)n}\right)\right).
\end{aligned} \tag{61}$$

where $H(\alpha)$ is the binary entropy function. By standard tail bounds on χ^2 random variables (see e.g. Lemma 1 of (Laurent & Massart, 2000)), we have that $\frac{\sum_{i\in A}(X_i-\mu)^2}{\alpha n} \leq 2$ and $\frac{\sum_{i\notin A}(X_i-\mu)^2}{(1-\alpha)n} \leq 2$ with probability at least $1 - \frac{1}{n^3}$, for sufficiently large n. Plugging these upper bounds into (61) yields

$$L_{\alpha,\mu}(\mathbf{X}) \geq n\left(-\frac{\alpha}{2}\left(\frac{\sum_{i\in A}(X_i-\mu)^2}{\alpha n}\right)-\frac{1-\alpha}{2}\left(\frac{\sum_{i\notin A}X_i^2}{(1-\alpha)n}\right)\right)\geq n\left(-\alpha-(1-\alpha)\right)=-n. \tag{62}$$

Case 2, Sub-case 1: $\alpha+\kappa<\frac{1}{n^2}$.

Our strategy for this sub-case, as well as the subsequent ones, will be to upper bound $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})$ and show that $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})<-n\leq L_{\alpha,\mu}(\mathbf{X})$.

We have

$$\begin{aligned}L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) &= \sum_{i=1}^{n}\log\left((\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right)+(1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right)\right)\\ &\leq \sum_{i\in A}\log\left((\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right)+(1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right)\right).\end{aligned} \tag{63}$$

We upper bound the first term $(\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right)$ in the logarithm by

$$(\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right)\leq\alpha+\kappa<\frac{1}{n^2}. \tag{64}$$

We upper bound the second term $(1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right)$ in the logarithm as

$$\begin{aligned}(1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right) &\leq \exp\left(-\frac{X_i^2}{2}\right)\\ &\leq \exp\left(-\frac{1}{2}(\mu-2\sqrt{\log n})^2\right)\\ &\leq \exp\left(-\frac{1}{2}(C\sqrt{\log n}-2\sqrt{\log n})^2\right),\\ &= n^{-(C-2)^2/2}\\ &\leq n^{-2}\text{ for sufficiently large }C,\end{aligned} \tag{65}$$

where the first inequality follows from the assumption that, $X_i>\mu-2\sqrt{\log n}$ and the second inequality follows the fact that $\mu\geq C\sqrt{\log n}>2\sqrt{\log n}$ and $-\frac{1}{2}(\mu-2\sqrt{\log n})^2$ is decreasing for $\mu>2\sqrt{\log n}$.

Combining the upper bounds in (64) and (65) gives us the following upper bound on $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})$:

$$L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})\leq\sum_{i\in A}\log(n^{-2}+n^{-2})=\alpha n(\log(2n^{-2}))=-\Theta(n\log n). \tag{66}$$

Thus, for sufficiently large n, we have that $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})\leq\alpha n(\log(2n^{-2}))<-n=L_{\alpha,\mu}(\mathbf{X})$, as desired.

Case 2, Sub-case 2: $\alpha+\kappa>1-\frac{1}{n^2}$.

As in the previous sub-case, we will prove that $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})<-n\leq L_{\alpha,\mu}(\mathbf{X})$.

Since (50) holds, we have that $\frac{\mu^2}{\tau^2} > 100$, or equivalently $\frac{9}{10}\mu < \mu + \tau < \frac{11}{10}\mu$. We upper bound $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})$ as

$$\begin{aligned} L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) &\leq \sum_{i\notin A} \log\left((\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right)+(1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right)\right) \\ &\leq \sum_{i\notin A}\log\left(\exp\left(-\frac{1}{2}(X_i-(\mu+\tau))^2\right)+\frac{1}{n^2}\right) \\ &\leq \sum_{i\notin A}\log\left(\exp\left(-\frac{1}{2}(2\sqrt{\log n}-\frac{9}{10}C\sqrt{\log n})^2\right)+\frac{1}{n^2}\right) \\ &= \sum_{i\notin A}\log\left(n^{-\frac{1}{2}\left(2-\frac{9}{10}C\right)^2}+n^{-2}\right) \\ &\leq \sum_{i\notin A}\log\left(n^{-2}+n^{-2}\right) \text{ for sufficiently large } C \\ &= (1-\alpha)n\log(2n^{-2}), \end{aligned} \tag{67}$$

where the second inequality follows from the assumption that $1-\frac{1}{n^2} < \alpha+\kappa < 1$, and the third inequality follows from the fact that $X_i \leq 2\sqrt{\log n} \leq \frac{9}{10}C\sqrt{\log n} \leq \frac{9}{10}\mu \leq \mu+\tau$ for sufficiently large C.

For sufficiently large n, we have that $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) \leq (1-\alpha)n\log(2n^{-2}) = -\Theta(n\log n) < -n \leq L_{\alpha,\mu}(\mathbf{X})$, as desired.

Case 3: τ does not satisfy (50).

Since τ does not satisfy (50), we have that either $\mu+\tau > \frac{11}{10}\mu$ or $\mu+\tau < \frac{9}{10}\mu$. We treat each of these sub-cases separately. In each sub-case, we use the bound $L_{\alpha,\mu}(\mathbf{X}) > -n$ derived in Case 2.

Case 3, Sub-case 1: $\mu+\tau > \frac{11}{10}\mu$.

As before, we will show that $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) < -n \leq L_{\alpha,\mu}(\mathbf{X})$. We upper bound $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})$ as

$$\begin{aligned} L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) &= \sum_{i=1}^{n}\log\left((\alpha+\kappa)\cdot\exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right)+(1-\alpha-\kappa)\cdot\exp\left(-\frac{X_i^2}{2}\right)\right) \\ &\leq \sum_{i\in A}\log\left(\exp\left(-\frac{1}{2}\big(X_i-(\mu+\tau)\big)^2\right)+\exp\left(-\frac{X_i^2}{2}\right)\right) \\ &\leq \sum_{i\in A}\log\left(\exp\left(-\frac{1}{2}\left(\mu+2\sqrt{\log n}-\frac{11}{10}\mu\right)^2\right)+\exp\left(-\frac{1}{2}(\mu-2\sqrt{\log n})^2\right)\right) \\ &= \sum_{i\in A}\log\left(\exp\left(-\frac{1}{2}\left(\frac{1}{10}\mu-2\sqrt{\log n}\right)^2\right)+\exp\left(-\frac{1}{2}(\mu-2\sqrt{\log n})^2\right)\right) \\ &\leq \sum_{i\in A}\log\left(\exp\left(-\frac{1}{2}\left(\frac{1}{10}C\sqrt{\log n}+2\sqrt{\log n}\right)^2\right)+\exp\left(-\frac{1}{2}(C\sqrt{\log n}-2\sqrt{\log n})^2\right)\right) \\ &= \sum_{i\in A}\log\left(n^{-\frac{1}{2}\left(\frac{C}{10}+2\right)^2}+n^{-\frac{1}{2}(C-2)^2}\right) \\ &\leq \sum_{i\in A}\log\left(n^{-2}+n^{-2}\right) \text{ for sufficiently large } C, \end{aligned} \tag{68}$$

The first inequality uses that $\alpha+\kappa \leq 1$ and $1-\alpha-\kappa \leq 1$. The second inequality uses that $X_i \leq \mu+2\sqrt{\log n} < \frac{11}{10}\mu < \mu+\tau$ (where $\mu+2\sqrt{\log n} < \frac{11}{10}\mu \Leftrightarrow \mu \geq 20\sqrt{\log n}$ holds for sufficiently large C), and $X_i \geq \mu - 2\sqrt{\log n}$. The third inequality uses that $\frac{1}{10}\mu \geq \frac{1}{10}C\sqrt{\log n} > 2\sqrt{\log n}$ and $\mu \geq C\sqrt{\log n} > 2\sqrt{\log n}$ (for sufficiently large C).

Thus, for sufficiently large n we have $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) \leq -\alpha n\log(2n^{-2}) < -n = L_{\alpha,\mu}(\mathbf{X})$, as desired.

Case 3, Sub-case 2: $\mu+\tau < \frac{9}{10}\mu$.

As before, we will show that $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) < -n \leq L_{\alpha,\mu}(\mathbf{X})$. We upper bound $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X})$ as

$$\begin{aligned}
L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) &= \sum_{i=1}^{n} \log\left((\alpha+\kappa)\cdot \exp\left(-\frac{1}{2}(X_i-\mu-\tau)^2\right) + (1-\alpha-\kappa)\cdot \exp\left(-\frac{X_i^2}{2}\right)\right) \\
&\leq \sum_{i\in A} \log\left(\exp\left(-\frac{1}{2}\big(X_i-(\mu+\tau)\big)^2\right) + \exp\left(-\frac{X_i^2}{2}\right)\right) \\
&\leq \sum_{i\in A} \log\left(\exp\left(-\frac{1}{2}\left((\mu-2\sqrt{\log n})-\frac{9}{10}\mu\right)^2\right) + \exp\left(-\frac{1}{2}(\mu-2\sqrt{\log n})^2\right)\right) \\
&= \sum_{i\in A} \log\left(\exp\left(-\frac{1}{2}\left(\frac{1}{10}\mu-2\sqrt{\log n}\right)^2\right) + \exp\left(-\frac{1}{2}(\mu-2\sqrt{\log n})^2\right)\right) \\
&\leq \sum_{i\in A} \log\left(\exp\left(-\frac{1}{2}\left(\frac{1}{10}C\sqrt{\log n}-2\sqrt{\log n}\right)^2\right) + \exp\left(-\frac{1}{2}(C\sqrt{\log n}-2\sqrt{\log n})^2\right)\right) \\
&= \sum_{i\in A} \log\left(n^{-\frac{1}{2}\left(\frac{C}{10}-2\right)^2} + n^{-\frac{1}{2}(C-2)^2}\right) \\
&\leq \sum_{i\in A} \log\left(n^{-2}+n^{-2}\right) \text{ for sufficiently large } C.
\end{aligned} \tag{69}$$

The first inequality uses that $\alpha+\kappa \leq 1$ and $1-\alpha-\kappa \leq 1$. The second inequality uses that $\mu+\tau < \frac{9}{10}\mu < \mu - 2\sqrt{\log n} \leq X_i$ (where the middle inequality $\frac{9}{10}\mu < \mu-2\sqrt{\log n} \Leftrightarrow \mu > 20\sqrt{\log n}$ holds for sufficiently large C) and $X_i > \mu - 2\sqrt{\log n}$. The third inequality uses that $\frac{1}{10}\mu > \frac{1}{10}C\sqrt{\log n} > 2\sqrt{\log n}$ and $\mu > C\sqrt{\log n} > 2\sqrt{\log n}$ for sufficiently large C.

So as before, for sufficiently large n we have $L_{\alpha+\kappa,\mu+\tau}(\mathbf{X}) \leq -\alpha n \log(2n^{-2}) < -n = L_{\alpha,\mu}(\mathbf{X})$, as desired. □

H.1. Proofs of Corollaries

Corollary 1. *Let* $\mathbf{X} = (X_1,\dots,X_n) \sim \mathrm{ASD}_{\mathcal{S}}(A,\mu)$*, where* $|A| = \alpha n$ *for* $0 < \alpha < 0.5$ *and* $\mu \geq C\sqrt{\log n}$ *for a sufficiently large constant* $C > 0$*. Then* $\lim_{n\to\infty} \textit{Bias}(|\widehat{A}_{\mathcal{S}}|/n) = 0$*.*

Proof. Let B_n be the event that $\left|\frac{|\widehat{A}_{\mathcal{S}}|}{n} - |A|\right| \leq \sqrt{\frac{\log n}{n}}$. By Theorem 3, $P(B_n) \geq 1-\frac{1}{n}$. Note that when B_n does not hold, then $\left|\frac{|\widehat{A}_{\mathcal{S}}|}{n} - |A|\right| \leq 1$. So we have

$$\begin{aligned}
|\mathrm{Bias}(|\widehat{A}_{\mathcal{S}}|/n)| &\leq E\left[\left|\frac{|\widehat{A}_{\mathcal{S}}|}{n} - |A|\right| \mid B_n\right] + E\left[\left|\frac{|\widehat{A}_{\mathcal{S}}|}{n} - |A|\right| \mid B_n^c\right] \\
&\leq \left(\sqrt{\frac{\log n}{n}}\right)\cdot P(B_n) + 1\cdot P(B_n^c) \\
&\leq \sqrt{\frac{\log n}{n}} + \frac{1}{n}.
\end{aligned} \tag{70}$$

It follows that $\lim_{n\to\infty} \mathrm{Bias}(|\widehat{A}_{\mathcal{S}}|/n) = \lim_{n\to\infty}\left(\sqrt{\frac{\log n}{n}} + \frac{1}{n}\right) = 0$. □

Corollary 2. *Let* $\mathbf{X} = (X_1,\dots,X_n) \sim ASD_{\mathcal{P}_n}(A,\mu)$*, where* $|A| = \alpha n$ *for* $0 < \alpha < 0.5$ *and* $\mu \geq C\sqrt{\log n}$ *for a sufficiently large constant* $C > 0$*. Then* $\frac{|A \triangle \widehat{A}_{GMM}|}{|A|} \leq \frac{2}{\alpha}\sqrt{\frac{\log n}{n}} = o(1)$ *with probability at least* $1-\frac{1}{n}$*.*

Proof. By Lemma 8, with probability at least $1 - \frac{1}{n}$ we have that $X_i \leq 2\sqrt{\log n}$ for $i \notin A$ and $X_j \geq (C+2)\sqrt{\log n}$ for $j \in A$. Thus $X_i \leq 2\sqrt{\log n} < (C+2)\sqrt{\log n} \leq X_j$ for all $i \notin A$ and $j \in A$, which means A consists of the αn largest observations X_i.

Moreover, because the responsibilities $\widehat{r}_i$ are sorted in the same order as the observations X_i, we have that $\widehat{A}_{\text{GMM}}$ consists of the $|\widehat{A}_{\text{GMM}}|$ largest observations X_i. Thus, by Theorem 3, we have

$$|\widehat{A}_{\text{GMM}} \triangle A| = \left| |\widehat{A}_{\text{GMM}}| - \alpha n \right| \leq \left| |\widehat{A}_{\text{GMM}}| - \widehat{\alpha}_{\text{GMM}} n \right| + |\widehat{\alpha}_{\text{GMM}} - \alpha| \leq 2\sqrt{n \log n}. \tag{71}$$

with probability at least $1 - \frac{1}{n}$, for sufficiently large n. Since $|A| = \alpha n$, it follows that $\frac{|\widehat{A}_{\text{GMM}} \triangle A|}{|A|} \leq \frac{2}{\alpha}\sqrt{\frac{\log n}{n}} = o(1)$ as desired. $\square$

References

Boscoe, F. P., Talbot, T. O., and Kulldorff, M. Public domain small-area cancer incidence data for new york state, 2005-2009. *Geospatial health*, 11(1):304–304, 04 2016. doi: 10.4081/gh.2016.304.

Glaz, J. and Naus, J. I. *Scan Statistics*, pp. 1–7. American Cancer Society, 2010. ISBN 9780471667193.

Kulldorff, M. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, 1997.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000.

Liu, Y. and Arias-Castro, E. A multiscale scan statistic for adaptive submatrix localization. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 44–53, 2019.

Nagamochi, H., Nishimura, K., and Ibaraki, T. Computing all small cuts in undirected networks. In Du, D.-Z. and Zhang, X.-S. (eds.), *Algorithms and Computation*, pp. 190–198, Berlin, Heidelberg, 1994. Springer Berlin Heidelberg. ISBN 978-3-540-48653-4.

Neill, D. B. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics*, 8(1):20, 2009.

Neill, D. B. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

Reyna, M. A., Chitra, U., Elyanow, R., and Raphael, B. J. Netmix: A network-structured mixture model for reduced-bias estimation of altered subnetworks. In *Research in Computational Molecular Biology*, pp. 169–185, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45257-5.

Rippl, T., Munk, A., and Sturm, A. Limit laws of the empirical wasserstein distance. *J. Multivar. Anal.*, 151(C):90–109, October 2016. ISSN 0047-259X.

Sharpnack, J. A path algorithm for localizing anomalous activity in graphs. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 341–344, Dec 2013.

Sharpnack, J., Krishnamurthy, A., and Singh, A. Near-optimal anomaly detection in graphs using lovász extended scan statistic. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 1959–1967, 2013a.

Sharpnack, J., Singh, A., and Rinaldo, A. Changepoint detection over graphs with the spectral scan statistic. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 545–553, 2013b.

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 11 2019.