Unbiased Information Packets

Daniel Halpern [*] **Ariel Procaccia** [*]

Abstract

Forums that allow average citizens to deliberate over policy issues are gaining prominence around the world. It is widely understood that a prerequisite to successful deliberation is that participants have access to information that is "balanced"; but what does that mean, precisely? In this paper we propose to measure how (un)balanced a set of information sources (which we call an *information packet*) is by considering the bias of linear regression on these data. Our main theoretical result formalizes and quantifies the idea that judicious selection of information sources can significantly reduce bias. Our experiments reinforce this message in a variety of conditions.

1. Introduction

In recent decades, deliberation has emerged as one of the most promising directions for democratic reform. Fishkin (2018) argues that, in an age of misinformation and polarization, the "will of the people" can only be uncovered by allowing ordinary citizens to have a thoughtful and informed discussion of policy questions.

This idea — coupled with the randomized selection of representatives (also known as *sortition*) — underlies the worldwide surge in *citizens' assemblies*, which consist of people from all walks of life and aim to inform policy makers. Prominent examples of national citizens' assemblies include We the Citizens in Ireland (which famously led to the legalization of abortions following a referendum), Citizens' Convention for Climate in France (which was initiated by President Emmanuel Macron) and Climate Assembly UK (which was sponsored by the House of Commons).

A prerequisite for the success of citizens' assemblies is that participants have access to a high-quality information about the question at hand. For example, the organizers of Climate Assembly UK[1] place a significant emphasis on this aspect: expert leads were responsible for ensuring that "the information provided to Climate Assembly UK is balanced, accurate and comprehensive," and an academic panel provided another layer of oversight to make absolutely certain that information was, once again, "balanced, accurate and comprehensive."

Of these requirements, the first — that the information be "balanced," or *unbiased* — is the one that is most open to interpretation in our view. To be clear, this requirement pertains not to a single *information source* (such as a news article or scientific paper), but to a compilation of different information sources, which we refer to as an *information packet*. We ask:

> *What is a principled way of measuring the bias of an information packet? And to what degree does careful selection of information sources reduce bias?*

To our knowledge we are the first to address these questions from a mathematical viewpoint.

1.1. Overview of Our Model and Results

We are interested in deliberation over a binary issue, which we think of as enacting or rejecting a policy proposal for concreteness. For example, in the context of climate change, the proposal might be to reduce greenhouse gas emissions nationally to a certain level by a certain year. Citizens' assemblies sometimes consider multiple issues, in which case our framework would apply to each issue separately.

The information packet we wish to construct consists of m information sources selected from a larger set of information sources. The choice of m is exogenous and conceptually aims for comprehensiveness subject to the limited attention span of panelists.

Information sources are represented as points $(x, y) \in \mathbb{R}^2$. The first coordinate, x, is the *strength of evidence*, where positive numbers correspond to evidence in favor of enacting the proposal (the larger the stronger) and negative numbers correspond to evidence in favor of rejecting the proposal (the smaller the stronger).

In practice evidence would typically be multi-dimensional; in the context of climate change these dimensions might

[*]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. Correspondence to: Daniel Halpern <dhalpern@g.harvard.edu>.

[1]`https://www.climateassembly.uk`

be indicators like air temperature, sea levels and many others. For the purposes of our framework, then, we assume that multi-dimensional evidence has been aggregated into a single number in way that is consistent across information sources. This does *not* mean that panelists are not exposed to nuanced, multi-dimensional evidence — the aggregation is only used to *select* information sources. See Section 5, though, for a discussion of how to extend our model to directly work with multi-dimensional evidence.

At first glance, it seems tempting to define the bias of an information packet based on strength of evidence alone: the average strength of evidence should be close to 0, meaning that evidence in favor of enacting the proposal is balanced with evidence in favor of rejecting it. However, we do not believe that this is a good definition. Going back to our example of climate change, reliable information sources would typically provide evidence in favor of enacting a measured policy proposal; it seems undesirable to artificially balance positive and negative evidence even when one side is sparse.

For this reason, each information source is also associated with a real number y representing *strength of support* (where positive and negative numbers are again interpreted as support for enacting or rejecting the proposal, respectively). For example, an editorial that explicitly advocates a particular policy would be interpreted as a strong level of support, regardless of the evidence it presents.

The opinion expressed in an information source, then, is manifested through the relation between strength of evidence and and strength of support. To determine the relation between strength of evidence and strength of support with respect to an information packet (subset of information sources) we apply ordinary least squares (OLS) regression to the data, which yields a linear function.

The linearity assumption is ostensibly restrictive, but there are two reasons why we believe it is not so. First, recall that multi-dimensional evidence is aggregated into a single number, representing overall strength of evidence, through an arbitrarily rich function, so we are implicitly considering the composition of the linear function and the rich function.[2] Second, we are not claiming that the linear function reflects the real world in terms of, say, the conclusions people would draw from the data; it is merely a tool that allows us to reason about bias in a rigorous way. That said, extending our model beyond linear functions would be valuable; we discuss this open problem in Section 5.

Now, the linear function $f(x) = \beta_1 \cdot x + \beta_0$ describes a relation between strength of evidence and strength of support that is obviously biased if it maps very weak evidence (x close to 0) to significant support for enacting or rejecting the proposal ($|y|$ far from 0).[3] Conversely, we think of f as unbiased if β_0 is close to 0. It is remarkably convenient that β_0, the intercept of the linear function f, is also widely known as its *bias*. To summarize, the *associated bias* of an information packet is measured through the (absolute value of the) bias of the linear function obtained by applying OLS regression to it — this gives an answer to our first question.

To address the second question ("to what degree does careful selection of information sources reduce bias"), we consider an *unbiased* ground-truth linear function $f^\star(x) = \beta_1^\star \cdot x$ that relates strength of evidence to the appropriate strength of support. We draw mk information sources[4] by, for each, independently drawing $x \sim \mathcal{D}$, where $\mathcal{D}$ is an arbitrary distribution over $\mathbb{R}$, and then drawing $y \sim \mathcal{N}(f^\star(x), \sigma^2)$. The symmetric noise over strength of support reflects the bias of individual information sources.

Roughly speaking, our main theoretical result is that, for any $m \geq 2$ and with high probability as k grows, there is a subset of m information sources whose associated bias is smaller by a factor of $\sqrt{k}$ than that of the entire set of mk information sources. We conclude that, in our model, it is possible to extract a relatively unbiased information packet from a large and relatively biased set of information sources.

In our experiments, we verify that our theoretical results continue to hold even in more relaxed settings, reinforcing their robustness. In particular, we analyze settings with practical values of k and non-normal noise distributions. Further, we compare several different algorithms for choosing an unbiased subset. Our results show that even with a minimal amount of computation, it is possible to do much better than the naive algorithm used in our theoretical analysis.

1.2. Related Work

Existing work that studies sortition from a statistical perspective focuses on the process of selecting participants for an assembly (Benadè et al., 2019; Flanigan et al., 2020).

In the social sciences there is a rich literature, dating back at least seven decades (White, 1950), on *media bias*. In contrast to our work, the focus is on the bias of individual news stories rather than that of an entire information packet.

[2] The same viewpoint is seen in many areas of machine learning. For example, in the literature on inverse reinforcement learning the reward of executing an action is often assumed to be linear in the features of the state-action pair (Abbeel & Ng, 2004), but these features can be obtained through a rich extractor.

[3] A linear function with negative slope may also seem obviously biased, but if such a function arises from OLS then either the information sources are extremely unreliable or the representation is defective; we do not explicitly handle this unlikely scenario.

[4] The assumption that the number of information sources is divisible by m is made purely for ease of exposition.

Moreover, to our knowledge computational work in this area aims to alleviate media bias without precisely defining it (Park et al., 2009). Groseclose & Milyo (2005) do measure media bias, but they adopt a purely empirical methodology.

We would be remiss if we did not mention the rich literature on fairness in ML, as "fairness" typically refers to "unbiasedness" (Dwork et al., 2012; Hardt et al., 2016; Joseph et al., 2016; Pleiss et al., 2017). However, the bias in this literature is with respect to the outcomes or labels assigned to individuals or groups, whereas the bias we examine is manifested in the relation between evidence and support.

On a technical level, our results are related to properties of the *folded normal distribution* (Leone et al., 1961; Tsagris et al., 2014), as we analyze the absolute value of normally distributed random variables.

2. Terminology

Let us present our terminology and notation more formally. We refer the reader to Section 1 for a detailed justification of our modeling choices.

Let bias be a function mapping (multi)sets of points (corresponding to information sources) in $\mathbb{R}^2$ to $\mathbb{R}$ called the *associated bias* function: Given n points from $\mathbb{R}^2$, $p_1 = (x_1, y_1), \ldots, p_n = (x_n, y_n)$, $\mathsf{bias}(\{p_1, \ldots, p_n\})$ is the bias term of the ordinary least squares (OLS) linear regression function on $p_1, \ldots, p_n$. For brevity of notation, we will often simply write $\mathsf{bias}(p_1, \ldots, p_n)$. Note that if the regression happens to be degenerate (which occurs when $x_1 = \cdots = x_n$), we assume that the slope is 0, in which case $\mathsf{bias}((x, y_1), \ldots, (x, y_n))$ will be the mean of the y values.

Fix some positive integers k and m. We are given km samples from $\mathbb{R}^2$, $p_1 = (x_1, y_1), \ldots, p_{km} = (x_{km}, y_{km})$. Our goal is to pick a subset of size m, $\{p_{i_1}, \ldots, p_{i_m}\}$, so as to minimize the absolute value of the associated bias, $|\mathsf{bias}(p_{i_1}, \ldots, p_{i_m})|$.

We assume the points are randomly sampled and thus we treat them as random variables $P_1 = (X_1, Y_1), \ldots, P_{km} = (X_{km}, Y_{km})$. Further, we assume they are sampled from an *unbiased linear model*. In our theoretical results (Section 3), we assume each point $P_i = (X_i, Y_i)$ is sampled i.i.d. in the following way: first, X_i is sampled from some fixed distribution $\mathcal{D}$ over the reals, then, Y_i is sampled from $\mathcal{N}(\beta_1^\star \cdot X_i, \sigma^2)$, where $\beta_1^\star$ and σ are fixed parameters. We relax the normal distribution assumption in our experiments (Section 4).

We are interested in a good *selection procedure*, a function that maps km points to m points that (hopefully) have low associated bias. Denote the bias-minimizing selection procedure by F^{opt}; it satisfies

$$F^{opt}(p_1, \ldots, p_{km}) \in \operatorname{argmin}_{S \subset \{p_1, \ldots, p_{km}\}, |S|=m} |\mathsf{bias}(S)|.$$

We also study the selection procedure F^{min}, which works as follows. First, F^{min} partitions the points into k sets of size m, $S_1 = \{p_1, \ldots, p_m\}, S_2 = \{p_{m+1}, \ldots, p_{2m}\}, \ldots, S_k = \{p_{(k-1)m} + 1, \ldots p_{km}\}$, and then it chooses the subset S_i with smallest absolute value of associated bias, $|\mathsf{bias}(S_i)|$. When dealing with independent random variables, since each $\mathsf{bias}(S_i)$ has the same distribution for all i, we are able to discuss the distribution of $\mathsf{bias}(S_i)$ in general.

3. Theoretical Analysis

The main goal of our theoretical analysis is to establish an upper bound of $o(1)$ on the ratio

$$\frac{|\mathsf{bias}(F^{opt}(P_1, \ldots, P_{km}))|}{|\mathsf{bias}(P_1, \ldots, P_{km})|}$$

between the associated bias of the optimal selection procedure (which selects m information sources) and that of the full set of km information sources. Such a bound would formalize the idea that as k (the ratio between the overall number of information sources and the desired size of the information packet) grows large, judicious selection of information sources will significantly decrease bias, even when compared to a highly unbiased, yet infeasible, alternative.

A major obstacle, however, is that F^{opt} is very difficult to reason about directly. In fact, we have the following theorem, whose proof is relegated to Appendix A.1.

Theorem 1. *It is NP-hard to approximate F^{opt} to any factor.*

Roughly speaking, the theorem implies that F^{opt} does not have a simple representation, and the same is true even for worst-case approximations thereof.

Our strategy, therefore, is to use F^{min} — which is amenable to theoretical analysis — as a proxy for F^{opt}. In particular, the bias of F^{min} is obviously an upper bound on the bias of F^{opt}, so any upper bound on the bias of F^{min} also holds for that of F^{opt}. In our experiments in Section 4 we directly measure the bias of F^{opt}.

3.1. Upper Bound

Our main result is the following theorem.

Theorem 2. *For all distributions $\mathcal{D}$, model parameters $\beta_1^\star, \sigma^2 > 0$, and $m \geq 2$, it holds that*

$$\lim_{k \to \infty} \Pr\left[\frac{|\mathsf{bias}(F^{min}(P_1, \ldots, P_{km}))|}{|\mathsf{bias}(P_1, \ldots, P_{km})|} \leq g(k)\right] = 1$$

for all $g(k) \in \omega\left(1/\sqrt{k}\right)$.

We remark that another natural benchmark to compare against is a random subset of m information sources. Against this benchmark, the bound of $1/\sqrt{k}$ improves to $1/k$. However, it is intuitive that a careful selection of m information sources outperforms a random selection. By contrast, since in the case of an unbiased ground truth the associated bias of mk information sources converges to 0 faster than that of m information sources, *a priori* it is unclear that the type of asymptotic improvement claimed by Theorem 2 would actually hold.

We now turn to the theorem's proof. We start with three lemmas about normally distributed random variables. The first is proved in Appendix A.2. The third is a special case of a folklore result that is mentioned, e.g., by Amemiya (1985); we prove it in Appendix A.3 for completeness. Slightly abusing notation, define min-abs to be a function that given multiple arguments returns the one that is smallest in absolute value (breaking ties arbitrarily): $\mathsf{min\text{-}abs}(x_1, \ldots, x_n) \in \operatorname{argmin}_{x_i} |x_i|$.

Lemma 1. *Let $f : \mathbb{N} \to (0, 1]$. If $X_1, \ldots, X_n$ are i.i.d. random variables drawn from a standard normal distribution, $\mathcal{N}(0, 1)$, then,*

$$\begin{aligned}&(1 - 2 \cdot \phi(0) \cdot f(n))^n \\ &\quad \le \Pr\left[|\mathsf{min\text{-}abs}(X_1, \ldots, X_n)| > f(n)|\right] \\ &\quad \le (1 - 2 \cdot \phi(1) \cdot f(n))^n.\end{aligned}$$

Lemma 2. *Let $p \in (0, 1]$ and let $f(n) : \mathbb{N} \to \mathbb{R}_{\ge 0}$ such that $f(n) \in \omega(1/n)$. If $X_1, \ldots, X_{\lceil pn \rceil}$ are independent, normally-distributed random variables with mean 0 and (not necessarily equal) variance at most 1, $\lim_{n\to\infty} \Pr\left[|\mathsf{min\text{-}abs}(X_1, \ldots, X_{\lceil pn \rceil})| > f(n)\right] = 0$.*

Proof. Fix f and p. Let $g(\ell) = \min_{n:\lceil pn \rceil = \ell} f(n)$. Note that $\lceil p \lfloor \ell/p \rfloor \rceil = \ell$, so g is well defined and $g(\ell) \ge f(\lfloor \ell/p \rfloor)$. We will show that as ℓ grows large, $\Pr\left[|\mathsf{min\text{-}abs}(X_1, \ldots, X_\ell)| > g(\ell)\right]$ approaches 0. As $g(\lceil pn \rceil) \le f(n)$ for all n, this will imply the desired result.

We would now like to invoke Lemma 1 on g and on $X_1, \ldots, X_\ell$; however, there are two conditions on g and each X_i that are not currently met, that g has range $(0, 1]$ and each X_i has variance equal to 1. We will show that without loss of generality, we may assume both.

First, we claim it is sufficient to prove the lemma for f with range $(0, 1]$, as this will imply it is true for all f. Indeed, replacing a value of 1 with a value higher than 1 can only decrease $\Pr\left[|\mathsf{min\text{-}abs}(X_1, \ldots, X_{\lceil pn \rceil})| > f(n)\right]$ (which we are trying to upper bound). Additionally, as $f(n) \in \omega(1/n)$, there is some N such that for all $n > N$, $f(n) > 0$. Hence, replacing the output of f where $f(n) = 0$ with something arbitrary in $(0, 1]$ will not affect $\Pr\left[|\mathsf{min\text{-}abs}(X_1, \ldots, X_{\lceil pn \rceil})| > f(n)\right]$ as n grows large. As $g(\ell) = f(n)$ for all ℓ and some n, under this assumption, g will also have range $(0, 1]$.

Additionally, we claim it is sufficient to prove the statement when each X_i has variance equal to 1. Indeed, as each X_i is normally distributed, decreasing the variance can only decrease $\Pr\left[|\mathsf{min\text{-}abs}(X_1, \ldots, X_{\lceil pn \rceil})| > f(n)\right]$ and thus would not affect the limit approaching 0.

By Lemma 1, we now have that

$$\Pr\left[|\mathsf{min\text{-}abs}(X_1, \ldots, X_\ell)| > g(\ell)\right] \le (1 - 2 \cdot \phi(1) \cdot g(\ell))^\ell.$$

Next, we will show that $g(\ell) \in \omega(1/\ell)$. Fix an arbitrary $c > 0$. As $f(n) \in \omega(1/n)$, there is some N such that for all $n > N$, $f(n) > \frac{c}{pn}$. Let $L \in \mathbb{N}$ be such that $\lfloor L/p \rfloor \ge N$ ($L = \lceil pN \rceil$ will do). Then, for all $\ell > L$,

$$g(\ell) \ge f(\lfloor \ell/p \rfloor) > \frac{c}{p \cdot \lfloor \ell/p \rfloor} \ge \frac{c}{\ell}$$

as needed.

As $g(\ell) \in \omega(1/\ell)$, for all c and sufficiently large ℓ,

$$(1 - 2 \cdot \phi(1) \cdot g(\ell))^\ell < (1 - c/\ell)^\ell.$$

However, $(1 - c/\ell)^\ell < e^{-c}$ for all ℓ. Since c was arbitrary, as ℓ grows large, $\Pr\left[|\mathsf{min\text{-}abs}(X_1, \ldots, X_\ell)| > g(\ell)\right]$ approaches 0. □

Let us define $\mathsf{mean}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$ to be the mean of $x_1, \ldots, x_n$. Let

$$\mathrm{SV}(x_1, \ldots, x_n) = \frac{\sum_{i=1}^n (x_i - \mathsf{mean}(x_1, \ldots, x_n))^2}{n}$$

be the (adjusted) sample variance.[5] Then we have the following.

Lemma 3. *Conditioned on the x values of a linear model with parameters β_0, β_1, and σ having values $x_1, \ldots, x_n$, the following hold:*

- *If $\mathrm{SV}(x_1, \ldots, x_n) > 0$ ($x_1, \ldots, x_n$ are not all identical), the associated bias follows a normal distribution with mean β_0 and variance $\frac{\sigma^2}{n}\left(1 + \frac{\mathsf{mean}(x_1,\ldots,x_n)^2}{\mathrm{SV}(x_1,\ldots,x_n)}\right)$.*
- *If $\mathrm{SV}(x_1, \ldots, x_n) = 0$, so $x = x_1 = \ldots = x_n$ for some $x \in \mathbb{R}$, then the associated bias follows a normal distribution with mean $\beta_1 \cdot x + \beta_0$ and variance $\frac{\sigma^2}{n}$.*

We are now ready to prove the theorem.

[5] It is a slight variant of the standard definition, with n instead of $n - 1$ in the denominator.

Proof of Theorem 2. Fix $\mathcal{D}$, $\beta_1^\star$, σ, m, and g. There are two cases depending on whether $\mathcal{D}$ has zero variance or not. First we will handle the non-zero case.

The case of non-zero variance. Here, we will assume $\mathrm{Var}(\mathcal{D}) > 0$.

Fix $\varepsilon > 0$. Our goal is to show that there is a constant $K \in \mathbb{N}$ such that for all $k > K$,

$$\Pr\left[\frac{|\mathsf{bias}(F^{min}(P_1, \ldots, P_{km}))|}{|\mathsf{bias}(P_1, \ldots, P_{km})|} \leq g(k)\right] < \varepsilon. \tag{1}$$

To do this, we will find a sequence of thresholds $T_k \in \mathbb{R}$ for each $k \in \mathbb{N}$ such that for sufficiently large k, both

$$\Pr\left[|\mathsf{bias}(P_1, \ldots, P_{km})| < T_k\right] < \frac{\varepsilon}{2}. \tag{2}$$

and

$$\Pr\left[|\mathsf{bias}(F^{min}(P_1, \ldots, P_{km}))| > T_k \cdot g(k)\right] < \frac{\varepsilon}{2} \tag{3}$$

If neither of the events of Equation (2) nor Equation (3) occur, then

$$\frac{|\mathsf{bias}(F^{min}(P_1, \ldots, P_{km}))|}{|\mathsf{bias}(P_1, \ldots, P_{km})|} \leq g(k)$$

holds. Hence, if Equation (2) and Equation (3) are true, then Equation (1) will hold as well by a union bound.

We will now define T_k. Let $s > 0$ be a number of standard deviations such that the probability of a normally-distributed random variable falling within this number of standard deviations of the mean is at most $\varepsilon/2$; more formally, let s be such that $\Phi(s) - \Phi(-s) = \int_{-s}^{s} \phi(x)\,dx < \varepsilon/2$. Note that $\varepsilon/2$ is in fact an upper bound of a normal random variable falling within a range of $2s$ standard deviations even if the interval is not centered around the mean, that is, $\Pr[X \in [y, y + 2s \cdot \sigma']] < \varepsilon/2$ for all $y \in \mathbb{R}$ where $X \sim \mathcal{N}(\mu, (\sigma')^2)$ for any parameters μ, σ'. Let

$$T_k = \sqrt{\frac{\sigma^2}{mk}} \cdot s.$$

Note that as we are treating σ^2, m, and s as constants depending only on the instance, so $T_k \in \Theta(1/\sqrt{k})$.

Next, we will prove Equation (2) holds for *all* values of k (not just those that are sufficiently large). Fix some arbitrary x values $x_1, \ldots, x_{km}$. We will show that, regardless of what these values are,

$$\Pr\left[|\mathsf{bias}(P_1, \ldots, P_{km})| < T_k \mid X_1 = x_1, \ldots, X_{km} = x_{km}\right] < \frac{\varepsilon}{2} \tag{4}$$

holds. As this is true for all values of $x_1, \ldots, x_{km}$, by the law of total probability, Equation (2) follows.

To establish Equation (4), we will first show that regardless of the x values, the distribution of $\mathsf{bias}(P_1, \ldots, P_{km})$ conditioned on these values is normal with variance at least $\frac{\sigma^2}{mk}$. Indeed, if all x values are identical, this follows directly from Lemma 3. On the other hand, suppose $x_1, \ldots, x_{km}$ are not all identical. Then, by Lemma 3, $\mathsf{bias}(P_1, \ldots, P_{km})$ follows a normal distribution with variance $\frac{\sigma^2}{km}(1 + \frac{\mathsf{mean}(x_1, \ldots, x_n)^2}{\mathrm{SV}(x_1, \ldots, x_n)})$. However, as $\frac{\mathsf{mean}(x_1, \ldots, x_n)^2}{\mathrm{SV}(x_1, \ldots, x_n)}$ is always nonnegative, the variance is at least $\frac{\sigma^2}{km}$ as needed. Therefore, in all cases, $T_k = \sqrt{\frac{\sigma^2}{mk}} \cdot s$ is at most s standard deviations, so the probability of falling in any range of this size (even if it is not centered at the mean) is at most $\varepsilon/2$, as needed.

It remains show that there exists a constant $K \in \mathbb{N}$ such that for all $k > K$, Equation (3) holds. Let $R_1 = \{X_1, \ldots, X_m\}, \ldots, R_k = \{X_{(k-1)m+1}, \ldots X_{km}\}$ be the partition of x values that F^{min} uses (recall that $S_1, \ldots, S_k$ were the sets of points). Let v be a function that takes a vector $(x_1, \ldots, x_m)$ and returns

$$1 + \frac{\mathsf{mean}(x_1, \ldots, x_n)^2}{\mathrm{SV}(x_1, \ldots, x_n)}$$

(or ∞ if $\mathrm{SV}(x_1, \ldots, x_n) = 0$). Note that if $v(R_i)$ is "not too large," this will let us bound the variance of the associated bias using Lemma 3. We will begin by showing there are "enough" sets R_i that have $v(R_i)$ "not too large." In particular, there exist constants $p \in (0, 1)$, $C \in \mathbb{R}$, and $K_1 \in \mathbb{N}$ such that for all $k > K_1$,

$$\Pr\left[\sum_{i=1}^{k} \mathbb{I}[v(R_i) < C] < p \cdot k\right] < \frac{\varepsilon}{4}. \tag{5}$$

To show this, we simply need to know that $v(R_i) < C$ for some constant C occurs with strictly positive probability. Then, since each R_i is independent, if we take $p < \frac{1}{2}\Pr[v(R_i) < C]$, as k grows large, the probability that fewer than a p fraction of $R_1, \ldots, R_k$ will have $v(R_i) < C$ will approach 0. Therefore, for sufficiently large k, the probability this occurs will be less than $\varepsilon/4$.

It remains to be shown that such a C exists. The only slight complication is that if it happens that $\mathrm{SV}(R_i) = 0$, $v(R_i)$ will be infinite. However, this only occurs if every element of R_i is the same. As $\mathrm{Var}(\mathcal{D}) > 0$, and the elements of R_i are sampled i.i.d. from $\mathcal{D}$, they must not be all identical with some probability $p_1 > 0$. Conditioned on $\mathrm{Var}(R_i) \neq 0$, the distribution of $v(R_i)$ only takes on real values, so there is some constant probability p_2 such that it is less than some constant C. Therefore, with probability at least $p_1 \cdot p_2 > 0$, $v(R_i) < C$ as needed.

Next, we will show that if the condition of Equation (5) does not hold, then we can bound the bias. In particular, there is

some constant $K_2 \in \mathbb{N}$ such that for all $k > K_2$,

$$\Pr\Bigg[|\textsf{min-abs}(\textsf{bias}(R_1), \ldots, \textsf{bias}(R_k))| > g(k) \cdot T_k \\ \Bigg| \sum_{i=1}^{k} \mathbb{I}[v(R_i) < C] \geq p \cdot k\Bigg] < \frac{\varepsilon}{4}. \tag{6}$$

Suppose $\sum_{i=1}^{k} \mathbb{I}[v(R_i) < C] \geq p \cdot k$. Without loss of generality, suppose that $v(R_i) < C$ for $1 \leq i \leq \lceil pk \rceil$. Note that

$$|\textsf{min-abs}(\textsf{bias}(R_1), \ldots, \textsf{bias}(R_n))| \\ \leq |\textsf{min-abs}(\textsf{bias}(R_1), \ldots, \textsf{bias}(R_{\lceil pk \rceil}))|.$$

Thus, it is sufficient to show that,

$$\Pr\left[|\textsf{min-abs}(\textsf{bias}(R_1), \ldots, \textsf{bias}(R_{\lceil pk \rceil}))| > g(k) \cdot T_k\right] \\ < \frac{\varepsilon}{4}. \tag{7}$$

Next, note that as $v(R_i) < C$, $\textsf{bias}(R_i)$ follows a normal distribution with mean 0 and variance at most $\frac{\sigma^2}{m}C$. Importantly, $\frac{\sigma^2}{m}C$ does not depend on k, thus, for the purposes of this proof, it can be treated as a constant. Therefore, $\frac{\textsf{bias}(R_i)}{\sqrt{\frac{\sigma^2}{m}C}}$ follows a normal distribution with mean 0 and variance at most 1.

In addition,

$$\left|\textsf{min-abs}\left(\frac{x_1}{c}, \ldots, \frac{x_n}{c}\right)\right| = \frac{|\textsf{min-abs}(x_1, \ldots, x_n)|}{c}$$

for all constants $c > 0$. We conclude that Equation (7) is equivalent to

$$\Pr\left[\left|\textsf{min-abs}\left(\frac{\textsf{bias}(R_1)}{\sqrt{\frac{\sigma^2}{m}C}}, \ldots, \frac{\textsf{bias}(R_{\lceil pk \rceil})}{\sqrt{\frac{\sigma^2}{m}C}}\right)\right| > \frac{g(k) \cdot T_k}{\sqrt{\frac{\sigma^2}{m}C}}\right] \\ < \frac{\varepsilon}{4} \tag{8}$$

Finally, note that $g(k) \in \omega(1/\sqrt{k})$, $T_k \in \Omega(\sqrt{k})$, and $\sqrt{\frac{\sigma^2}{m}C}$ can be treated as a constant. Hence, the right-hand side of the inequality inside the probability in Equation (8) is $\omega(1/k)$. By Lemma 2, the left-hand side of Equation (8) approaches 0. Thus for sufficiently large k, the probability is at most $\varepsilon/4$ as needed. Finally, combining Equations (5) and (6) by way of a union bound yields Equation (3) as desired.

The case of zero variance. If $\mathcal{D}$ has variance equal to zero, it must be the case that there is some $r \in \mathbb{R}$ such that $\Pr[X = r] = 1$. This means that points $P = (X, Y)$ are in fact sampled as (r, Y) where Y is normally distributed with mean $r \cdot \beta_1^\star$ and variance σ^2. Call this mean $\mu = r \cdot \beta_1^\star$. As all the points have the same x value, by Lemma 3, $\textsf{bias}(P_1, \ldots, P_{km})$ is normally distributed with mean μ and variance $\frac{\sigma^2}{mk}$.

Here, there are two more cases we must consider, one where $\mu \neq 0$ and one where $\mu = 0$.

Suppose $\mu \neq 0$, and without loss of generality suppose $\mu > 0$. Let p be the probability that $\textsf{bias}(S_i)$, a sample of m points, (which will simply be the average of $Y_1, \ldots, Y_m$, normal with mean μ and variance σ^2/m) falls in $(-2, 0)$. Note that

$$\Pr\left[\textsf{bias}(S_i) \in \left(-\frac{1}{n}, \frac{1}{n}\right)\right] \geq \frac{p}{n} \tag{9}$$

for all $n \geq 1$. This is because the probability of $(-2/n, 0)$ is at least p/n as the density in $(-2/n, 0)$ is strictly higher than the rest of $(-2, -2/n)$, and then the density of $(0, 1/n)$ must be higher than $(-2/n, -1/n)$.

Fix $\varepsilon > 0$ and let N be such that $(1/e)^N < \varepsilon/2$. Note that for sufficiently large k,

$$\Pr\left[\textsf{bias}(P_1, \ldots, P_{km}) < \frac{\mu}{2}\right] < \frac{\varepsilon}{2}.$$

Additionally, for sufficiently large k, $\mu/2 \cdot g(k) \geq N/(kp)$ as N, p and $\mu/2$ are constants, and $g(k) \in \omega(1/\sqrt{k})$. Therefore, all that needs to be shown is that for large enough k,

$$\Pr\left[\textsf{bias}(F^{min}(P_1, \ldots, P_{km})) \in \left(-\frac{N}{kp}, \frac{N}{kp}\right)\right] \geq 1 - \frac{\varepsilon}{2}, \tag{10}$$

as a union bound will yield the desired result.

To that end, consider k large enough that $N/(kp) \leq 1$. By Equation (9), it holds that

$$\Pr\left[\textsf{bias}(S_i) \in \left(\frac{-N}{kp}, \frac{N}{kp}\right)\right] \geq \frac{p}{\frac{kp}{N}} = \frac{N}{k}.$$

Therefore, the left-hand side of Equation (10) is at least $1 - (1 - N/k)^k$. We then have,

$$\left(1 - \frac{N}{k}\right)^k = \left(1 - \frac{N}{k}\right)^{\frac{k}{N} \cdot N} = \left(1 - \frac{1}{\frac{k}{N}}\right)^{\frac{k}{N} \cdot N} \\ = \left(\left(1 - \frac{1}{\frac{k}{N}}\right)^{\frac{k}{N}}\right)^N \leq \left(\frac{1}{e}\right)^N \leq \frac{\varepsilon}{2},$$

as needed.

Next, suppose that $\mu = 0$. Recall that $\textsf{bias}(P_1, \ldots, P_{km}) \sim \mathcal{N}(0, \frac{\sigma^2}{mk})$. Additionally, each $\textsf{bias}(S_i) \sim \mathcal{N}(0, \frac{\sigma^2}{m})$. The

remainder of this proof follows a similar structure to the case of non-zero variance in that we can prove Equation (2) and Equation (3) hold for sufficiently large k, except the arguments are much simpler. Indeed, Equation (2) is immediate from the choice of T_k. For Equation (3), this can be shown more directly as

$$\mathsf{bias}(F^{min}(P_1,\dots,P_{km})) = \mathsf{min\text{-}abs}(X_1,\dots,X_k)\cdot\sqrt{\frac{\sigma^2}{m}}$$

where each $X_i \sim \mathcal{N}(0,1)$. Hence, Equation (3) follows as a consequence of Lemma 2. □

3.2. Lower Bound

A careful reading of the proof of Theorem 2 suggests that the analysis is asymptotically tight. Our final theoretical result proves that this is indeed the case.

Theorem 3. *For all distributions $\mathcal{D}$ such that $\mathrm{Var}(\mathcal{D}) \neq 0$, model parameters $\beta_1^\star$, $\sigma^2 > 0$, and $m \geq 2$, the probability*

$$\Pr\left[\frac{|\mathsf{bias}(F^{min}(P_1,\dots,P_{km}))|}{|\mathsf{bias}(P_1,\dots,P_{km})|} \leq g(k)\right]$$

does not approach 1 as k grows for all $g(k) \in O(1/\sqrt{k})$.

The proof relies on the following lemma, whose proof can be found in Appendix A.4.

Lemma 4. *For all distributions $\mathcal{D}$ such that $\mathrm{Var}(\mathcal{D}) \neq 0$, there is some constant $C \in \mathbb{R}$ such that if $X_1,\dots,X_n$ are independent samples drawn from $\mathcal{D}$, then*

$$\lim_{n\to\infty} \Pr\left[\frac{\mathsf{mean}(X_1,\dots,X_n)^2}{\mathrm{SV}(X_1,\dots,X_n)} < C\right] = 1.$$

Proof of Theorem 3. Since $g(k) \in O(1/\sqrt{k})$, there is some $c > 0$, such that sufficiently large k, $g(k) < c/\sqrt{k}$. We will show that as k grows large,

$$\Pr\left[\frac{|\mathsf{bias}(F^{min}(P_1,\dots,P_{km}))|}{|\mathsf{bias}(P_1,\dots,P_{km})|} > \frac{c}{\sqrt{k}}\right]$$

is lower bounded by some positive constant.

By Lemma 4, there is some constant $C-1$ such that for sufficiently large k,

$$\Pr\left[1 + \frac{\mathsf{mean}(X_1,\dots,X_{km})^2}{\mathrm{SV}(X_1,\dots,X_{km})} < C\right] > \frac{1}{2}. \tag{11}$$

From now on we condition on this event. By Lemma 3, $\mathsf{bias}(P_1,\dots,P_{km})$ follows a normal distribution with mean 0 and variance at most $\frac{\sigma^2}{mk}\cdot C$. Let $C' = \sqrt{\frac{\sigma^2}{m}\cdot C}$. We have that $\frac{C'}{\sqrt{k}}$ is an upper bound on the standard deviation, and further, C' does not depend on k, so, for the purposes of this proof, it can be treated as a constant.

We will make use of the following normal distribution tail bound (Feller, 1968):

$$1 - \Phi(x) < x^{-1}\phi(x)$$

for all $x > 0$. In our setting, this implies that

$$\begin{aligned}\Pr\left[|\mathsf{bias}(P_1,\dots,P_{km})| < \frac{c\cdot x\cdot C'}{\sqrt{k}}\right] &\geq \Phi(cx) - \Phi(-cx)\\ &= \Phi(cx) - (1-\Phi(cx)) = 2\Phi(cx) - 1\\ &> 2\left(1 - \frac{1}{cx}\cdot\frac{e^{-cx^2}}{\sqrt{2\pi}}\right) - 1 = 1 - \frac{2}{cx}\cdot\frac{e^{-cx^2}}{\sqrt{2\pi}}\end{aligned}$$

for all $x > 0$. On the other hand, by Lemma 1,

$$\begin{aligned}\Pr\left[|\mathsf{bias}(F^{min}(P_1,\dots,P_{km}))| > \frac{x\cdot C'}{k}\right] &\geq \left(1 - \frac{2\phi(0)\cdot x}{k}\right)^k\\ &\geq e^{-2\phi(0)\cdot x} - \varepsilon\end{aligned}$$

for any $\varepsilon > 0$ as long as k is sufficiently large. However, for sufficiently large x,

$$\frac{2}{cx}\cdot\frac{e^{-cx^2}}{\sqrt{2\pi}} + \left(1 - e^{-2\phi(0)\cdot x}\right) < 1,$$

so for small enough ε, the two events will occur with probability at least $\varepsilon' > 0$, conditioned on the event of Equation (11). By the law of total probability we conclude that for large enough k,

$$\Pr\left[\frac{|\mathsf{bias}(F^{min}(P_1,\dots,P_{km}))|}{|\mathsf{bias}(P_1,\dots,P_{km})|} \leq g(k)\right] \leq 1 - \frac{\varepsilon'}{2},$$

as needed. □

4. Experiments

The purpose of this section is twofold. First, we verify that our theoretical results empirically generalize when some of our assumptions are relaxed, in particular for practical values of k and for noise distributions that are not necessarily normal. Second, we compare different selection procedures for a variety of parameterized instances. All experiments were run on synthetically-generated data.

The experiments were conducted as follows. All x values were drawn from $\mathcal{U}(-1,1)$. We tested different values of $m \in \{3,10,20\}$, values of $k \in \{5,20,50,100\}$ and values of $\beta_1^\star \in \{0,1,5\}$. In addition, we tested noise distributions, $\mathcal{N}(0,0.5^2)$, $\mathcal{U}(-2,2)$, and $\beta(2,2)$. Note that as $\beta(2,2)$ has support $[0,1]$, we mirror it by flipping the sign with probability $1/2$. Each point is the mean squared bias of a specific selection procedure based on 3000 samples.

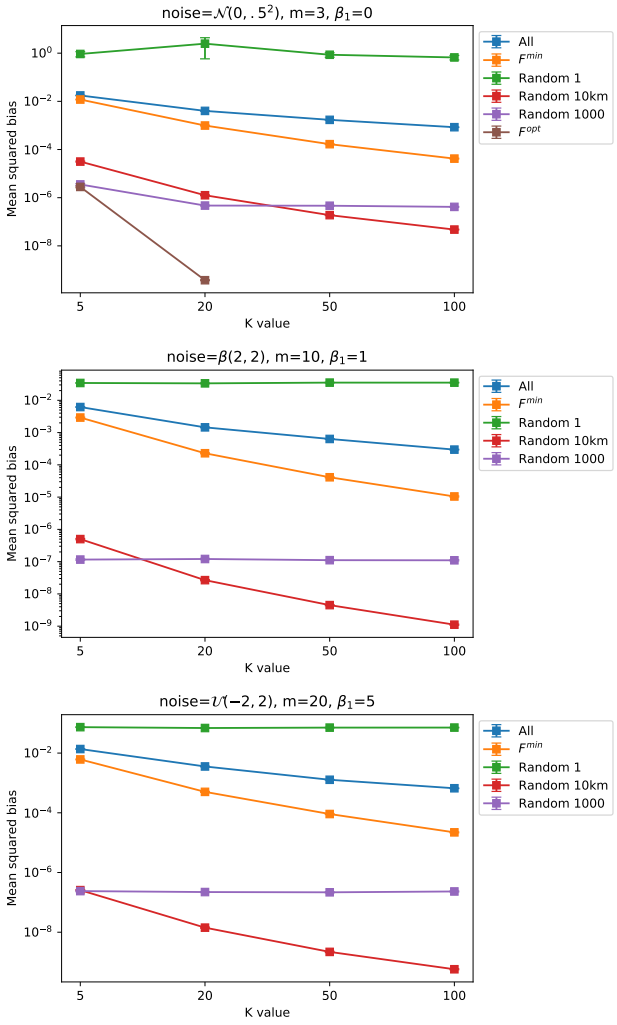


Figure 1: Representative plots. Error bars (which are so small they are invisible in most cases) show the standard error of the mean. Note that the y-axis is in log scale.

We compared the following selection procedures: F^{opt} and F^{min} were defined in Section 2; *All* uses all mk points; Random 1, 1000, and $10mk$ each take 1, 1000 and $10mk$ samples of size m, respectively, and then choose the best of these samples. Due to computational constraints, F^{opt} could only be run on instances with $m = 3$ and $k \leq 20$.

Representative plots are shown in Figure 1; other plots are relegated to Appendix B. Note that the different settings of parameters did very little to affect the relative performance of the selection procedures (except differing values of m for Random $10km$ and Random 1000 for obvious reasons).

We draw several conclusions from these results. First, F^{min} provides significant improvements over All and (even more so) Random 1 across the board, thereby showing that the qualitative message of Theorem 2 indeed holds true even for small values of k and different noise distributions. Second, randomly choosing the best of many m-subsets is a significant improvement over F^{min}, which suggests that very simple algorithms can lead to low-bias selections. Third, in cases where it can be computed, F^{opt} does much better than other selection procedures, which highlights the question of designing more sophisticated algorithms in our model.[6]

5. Discussion

We wrap up by discussing the limitations of our model and possible extensions thereof.

First, in formulating our model we assumed that for each information source, strength of evidence is represented as a single number. As mentioned in Section 1, in practice evidence would typically be multi-dimensional and an arbitrarily rich function can be used to arrive at an aggregate number. Alternatively, it is possible to directly work with multi-dimensional evidence vectors. However, this poses additional technical challenges for proving theoretical results. In particular, when adding more dimensions, the formula for the variance of the bias is no longer as simple to work with as that of Lemma 3.

A related issue is the means by which an actual information source is mapped to its numerical representation. This step could be performed by the same experts that today compile information packets; in fact, it would require a lower level of expertise. We acknowledge, however, that measures must be taken to prevent bias from creeping in through this step of the process.

Another modeling choice — which we justified in Section 1 but may still prove controversial — is our use of *linear* regression to measure bias. The challenges with going beyond linear functions are not only technical but also conceptual: it is not immediately clear how to interpret and measure bias. Note that a homogeneous linear function satisfies $f(x) = -f(-x)$ for all $x \in \mathbb{R}$; one idea is to measure bias more generally through the relation between $f(x)$ and $-f(-x)$, suitably aggregated across different values of x.

To conclude, there are admittedly gaps between our model and reality (like all models) but we believe it is useful in two ways. First, our work serves to show that the problem of creating unbiased information packets can be approached from a mathematical perspective. Second, we believe that our qualitative message — algorithms for selection of information sources greatly reduce bias — is robust to the details of the model. Taken together, these contributions reveal a new way in which computer scientists and statisticians can support democratic innovation.

[6] That said, sophisticated algorithms would typically be specific to the details of the model. We view the qualitative message derived from our model as more important.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pp. 1–8, 2004.

Amemiya, T. *Advanced Econometrics*. Harvard University Press, 1985.

Benadè, G., Gölz, P., and Procaccia, A. D. No stratification without representation. In *Proceedings of the 20th ACM Conference on Economics and Computation (EC)*, pp. 281–314, 2019.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 214–226, 2012.

Feller, W. *Introduction to Probability Theory and its Applications*, volume 1, pp. 254. Wiley, 3rd edition, 1968.

Fishkin, J. *Democracy When the People Are Thinking: Revitalizing Our Politics Through Public Deliberation*. Oxford University Press, 2018.

Flanigan, B., Gölz, P., Gupta, A., and Procaccia, A. D. Neutralizing self-selection bias in sampling for sortition. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020. Forthcoming.

Groseclose, T. and Milyo, J. A measure of media bias. *Quarterly Journal of Economics*, 120:1191–1237, 2005.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3315–3323, 2016.

Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 325–333, 2016.

Kleinberg, J. M. and Tardos, E. *Algorithm Design*. Addison-Wesley, 2005.

Leone, F., Nelson, L., and Nottingham, R. The folded normal distribution. *Technometrics*, 3(4):543–550, 1961.

Park, S., Kang, S., Chung, S., and Song, J. NewsCube: Delivering multiple aspects of news to mitigate media bias. In *Proceedings of the 27th ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 443–453, 2009.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., and Weinberger, K. Q. On fairness and calibration. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5680–5689, 2017.

Tsagris, M., Beneki, C., and Hassani, H. On the folded normal distribution. *Mathematics*, 2(1):12–28, 2014.

White, D. M. The "gate keeper": A case study in the selection of news. *Journalism & Mass Communication Quarterly*, 27(4):383–390, 1950.

Appendix

A. Missing Proofs

A.1. Proof of Theorem 1

To more formally define the computational problem let us extend F^{opt} to accept a number of points m to be selected. Specifically, F^{opt} works as follows: Given a list of n points $(x_1, y_1), \ldots, (x_n, y_n)$ and a number $m \leq n$ it returns m points that have optimal bias.

Our hardness proof starts with the classic NP-hard decision problem SUBSET SUM (Kleinberg & Tardos, 2005). The problem is as follows: *Given natural numbers $w_1, \ldots, w_n$, and a target number W, is there a subset of $\{w_1, \ldots, w_n\}$ that adds up precisely to W?* We present a folklore reduction to a slight variant that we call SUBSET SUM TO ZERO (SSZ) : *Given integers $z_1, \ldots, z_n$, is there a nonempty subset of $\{z_1, \ldots, z_n\}$ that adds up precisely to 0?*

Given an instance of SUBSET SUM $w_1, \ldots, w_n$ and W, construct an instance of SSZ by first removing all elements w_i such that $w_i = 0$ then adding a new number $-W$. Clearly this can be done in polynomial time; we will show that this reduction is correct. Indeed if there was a subset $S \subseteq \{w_1, \ldots, w_n\}$ that summed to W, S without any 0 valued elements would still add to W, and along with the new element $-W$, would then add to 0 (and be nonempty). On the other hand, if there is a nonempty subset $S \subseteq \{w_i | w_i \neq 0\} \cup \{-W\}$ that summed to 0, note that S must include $-W$ as all other elements are positive, hence $S \setminus \{-W\}$ would be a subset of $\{w_1, \ldots, w_n\}$ that adds to W, as needed.

Suppose f is an approximation to F^{opt}, that is, given n points $(x_1, y_1), \ldots, (x_n, y_n)$ and a number $m \leq n$, f chooses a subset of the points such that

$$\mathsf{bias}(f((x_1, y_1), \ldots, (x_n, y_n), m)) \leq C(n) \cdot \mathsf{bias}(F^{opt}(((x_1, y_1), \ldots, (x_n, y_n), m))$$

for some function $C(n)$. Importantly,

$$\mathsf{bias}(F^{opt}(((x_1, y_1), \ldots, (x_n, y_n), m)) = 0 \iff \mathsf{bias}(f((x_1, y_1), \ldots, (x_n, y_n), m)) = 0.$$

We will next show that if f can be computed in polynomial time, then SSZ can be solved in polynomial time, proving the NP-hardness of computing f. Indeed, given an SSZ instance $z_1, \ldots, z_n$, all we need to do is compute $\mathsf{bias}(f((0, z_1), \ldots, (0, z_n), m)$ for all m such that $1 \leq m \leq n$. If any of these are 0, then output YES, otherwise, output NO. Since f can be run in polynomial time, this algorithm can as well. The key idea for correctness is that $\mathsf{bias}((0, y_1), \ldots, (0, y_\ell)) = \mathsf{mean}(y_1, \ldots, y_\ell)$. In particular, note that $\mathsf{mean}(y_1, \ldots, y_\ell) = 0$ if and only if $\sum_{i=1}^{\ell} y_i = 0$. Hence, if there is some nonempty zero-sum subset $S = \{z_{i_1}, \ldots, z_{i_m}\}$, then $\mathsf{bias}(f((0, z_1), \ldots, (0, z_n), |S|)) = 0$. However, if there is no nonempty zero-sum subset then $\mathsf{bias}(f((0, z_1), \ldots, (0, z_n), m)) \neq 0$ for any m, as needed. □

A.2. Proof of Lemma 1

Fix such a function f. We have that,

$$\begin{aligned} &\Pr\left[|\mathsf{min\text{-}abs}(X_1, \ldots, X_n)| > f(n)\right] \\ &\quad = \Pr\left[\forall i \in [n], |X_i| > f(n)\right] \\ &\quad = \Pr\left[|X_i| > f(n)\right]^n \\ &\quad = (1 - \Pr\left[|X_i| \leq f(n)\right])^n, \end{aligned} \tag{12}$$

where the second equality holds because the X_is are independent (and follow the same distribution).

Let us now consider $\Pr\left[|X_i| \leq f(n)\right]$. As $X_i \sim \mathcal{N}(0, 1)$, we have that

$$\Pr\left[|X_i| \leq f(n)\right] = \int_{-f(n)}^{f(n)} \phi(x)\, dx. \tag{13}$$

Since $f(n) \leq 1$ for all n, for $x \in [-f(n), f(n)]$, $\phi(1) \leq \phi(x) \leq \phi(0)$, hence

$$2 \cdot \phi(1) \cdot f(n) \leq \int_{-f(n)}^{f(n)} \phi(x)\, dx \leq 2 \cdot \phi(0) \cdot f(n). \tag{14}$$

Combining Equations (12), (13), and (14) yields the desired result. □

A.3. Proof of Lemma 3

Fix β_0, β_1, σ, and $x_1, \ldots, x_n$. We have that each $Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

First let us suppose that $\mathrm{SV}(x_1, \ldots, x_n) > 0$. In this case, the x_is are not all identical, so the associated linear regression is not degenerate. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS estimators of β_0 and β_1 respectively and let $\beta = \begin{pmatrix}\beta_1 \\ \beta_0\end{pmatrix}$ and $\hat{\beta} = \begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_0\end{pmatrix}$. Let $\mathbf{X} = \begin{pmatrix}x_1 & 1 \\ \vdots & \vdots \\ x_n & 1\end{pmatrix}$, $\mathbf{Y} = \begin{pmatrix}Y_1 \\ \vdots \\ Y_n\end{pmatrix}$, and $\varepsilon = \begin{pmatrix}\varepsilon_1 \\ \vdots \\ \varepsilon_n\end{pmatrix}$. Note as each ε_i is sampled independently from $\mathcal{N}(0,1)$, ε follows a multivariate normal distribution with mean $\vec{0}$ and variance $\sigma^2\mathbf{I}$. Then we have that

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{X}\beta + \varepsilon\right)$$

Note that this shows that $\hat{\beta}$ is a linear function of a (multivariate) normal distribution, and thus is is also a (multivariate) normal distribution. To completely characterize this distribution, it is sufficient to find it's expectation and covariance matrix. We have that,

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}\right] &= \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{X}\beta + \varepsilon\right)\right] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{X}\beta + \mathbb{E}\left[\varepsilon\right]\right) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{X}\beta\right) = \beta\end{aligned}$$

where the second transition holds because $\mathbf{X}$ and β are being treated as constants. Next, we have that

$$\begin{aligned}\mathrm{Var}\left[\hat{\beta}\right] &= \mathrm{Var}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{X}\beta + \varepsilon\right)\right] \\ &= ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\,\mathrm{Var}\left[\mathbf{X}\beta + \varepsilon\right]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\ &= ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\sigma^2\mathbf{I})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\end{aligned}$$

where the second and third transitions holds because $\mathbf{X}$ and β are being treated as a constant.

Finally, let us consider the marginal distribution of $\hat{\beta}_0$. Note that it is normally distributed with mean β_0. What remains to be shown is that its variance is $\left(1 + \frac{\mathsf{mean}(x_1,\ldots,x_n)}{\mathrm{SV}(x_1,\ldots,x_n)}\right)$. By the definition of $\mathbf{X}$, we have that

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix}\sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n\end{pmatrix}$$

This implies that the bottom right entry of $(\mathbf{X}^T\mathbf{X})^{-1}$ (the one relevant for the marginal we are interested in) is equal to

$$\begin{aligned}\frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} &= \frac{n\frac{\sum_{i=1}^n x_i^2}{n}}{n^2\frac{\sum_{i=1}^n x_i^2}{n} - n^2\mathsf{mean}(x_1,\ldots,x_n)^2} \\ &= \frac{1}{n}\left(1 + \frac{\mathsf{mean}(x_1,\ldots,x_n)^2}{\frac{\sum_{i=1}^n x_i^2}{n} - \mathsf{mean}(x_1,\ldots,x_n)^2}\right) \\ &= \frac{1}{n}\left(1 + \frac{\mathsf{mean}(x_1,\ldots,x_n)^2}{\mathrm{SV}(x_1,\ldots,x_n)}\right)\end{aligned}$$

where the final transition holds due to $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ for the discrete distribution that uniformly picks each x_i with probability $1/n$.

Finally, let us suppose that $\mathrm{SV}(x_1, \ldots, x_n) = 0$, in which case $x_1 = \cdots = x_n = x$ for some $x \in \mathbb{R}$. Then, we have that each point $P_i = (x, Y_i)$ where $Y_i \sim \mathcal{N}(\beta_1 \cdot x + \beta_0, \sigma^2)$. By our definition for bias in this degenerate case, it will be precisely $\mathsf{mean}(Y_1, \ldots, Y_n) \sim \mathcal{N}(\beta_1 \cdot x + \beta_0, \frac{\sigma^2}{n})$, as needed. □

A.4. Proof of Lemma 4

Fix the distribution $\mathcal{D}$. First we will show that as n grows large, if $X_1, \ldots, X_n \sim \mathcal{D}$, then $\Pr[\mathrm{SV}(X_1, \ldots, X_n) \geq \ell]$ approaches 1 for some constant $\ell > 0$. To do this, we will show that there are values a, b with $a < b$ such that $\Pr[X_i < a] > 0$ and $\Pr[X_i > b] > 0$. Indeed, if cdf is the CDF of $\mathcal{D}$, there must be some point a such that $0 < \mathsf{cdf}(a) < 1$, as if this was not the case, $\mathcal{D}$ would have 0 variance. Further, as CDFs are right continuous, there is some point $b > a$ such that $\mathsf{cdf}(b) < 1$. With high probability, at least $\mathsf{cdf}(a)/2$ of the samples will be at most a and at least $(1 - \mathsf{cdf}(b))/2$ of the samples will be at least b. If both of these events occur, regardless of the sample mean, the variance is lower bounded by

$$\ell = \min\left(\frac{\mathsf{cdf}(a)}{2} \cdot \left(\frac{b-a}{2}\right)^2, \frac{1-\mathsf{cdf}(b)}{2} \cdot \left(\frac{b-a}{2}\right)^2\right) > 0,$$

as needed.

Next, note that there is some value C' such that $\Pr[X_i \in (-C', C')] > 1/2$. This means that as n grows large, with high probability, at least $1/2$ of the samples will fall in $(-C', C')$.

Suppose both of these events occur in some arbitrary realization $x_1, \ldots, x_n$, that is the $\mathrm{SV}(x_1, \ldots, x_n) \geq \ell$ and at least $n/2$ of $x_1, \ldots, x_n$ fall in $(-C', C')$. Let $\mu = \mathsf{mean}(x_1, \ldots, x_n)$ be the sample mean. If $|\mu| \leq 2C'$, then

$$\frac{\mathsf{mean}(x_1, \ldots, x_n)^2}{\mathrm{SV}(x_1, \ldots, x_n)} \leq \frac{4C'^2}{\ell}.$$

On the other hand, if $|\mu| > 2C'$, then note that

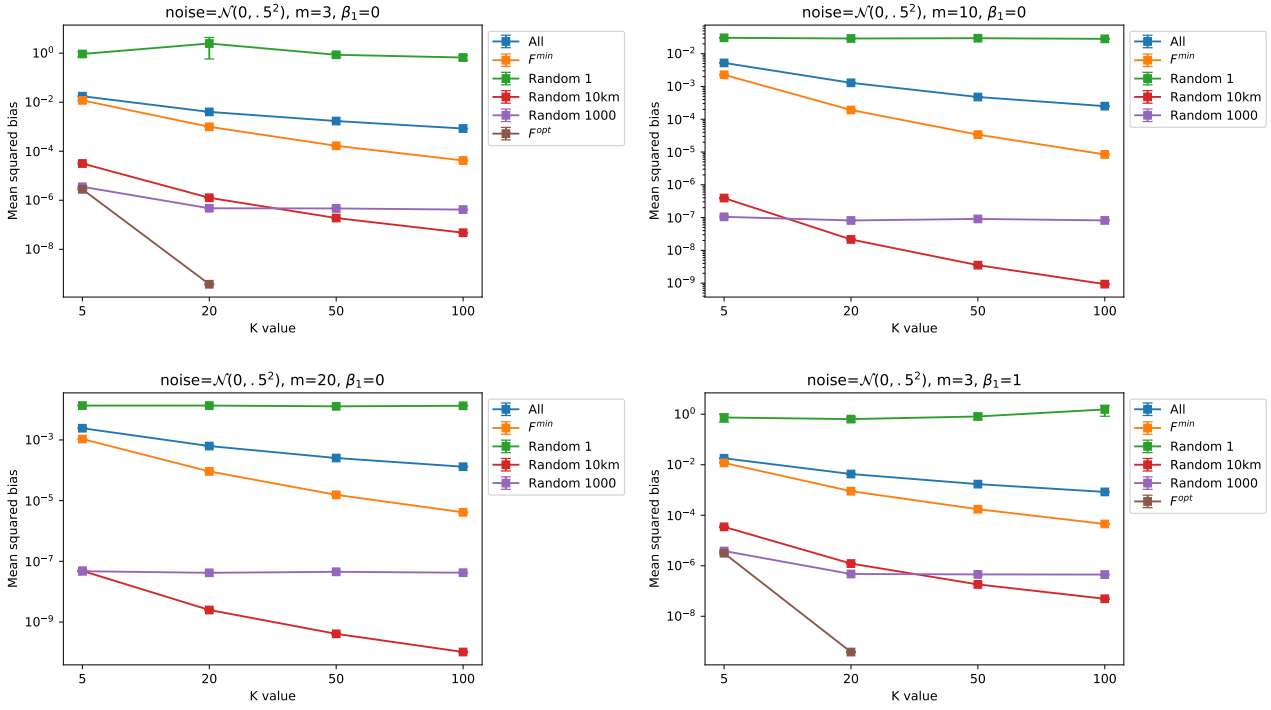
$$\mathrm{SV}(x_1, \ldots, x_n) \geq \frac{1}{2} \cdot (|\mu| - C')^2 \geq \frac{\mu^2}{8}.$$

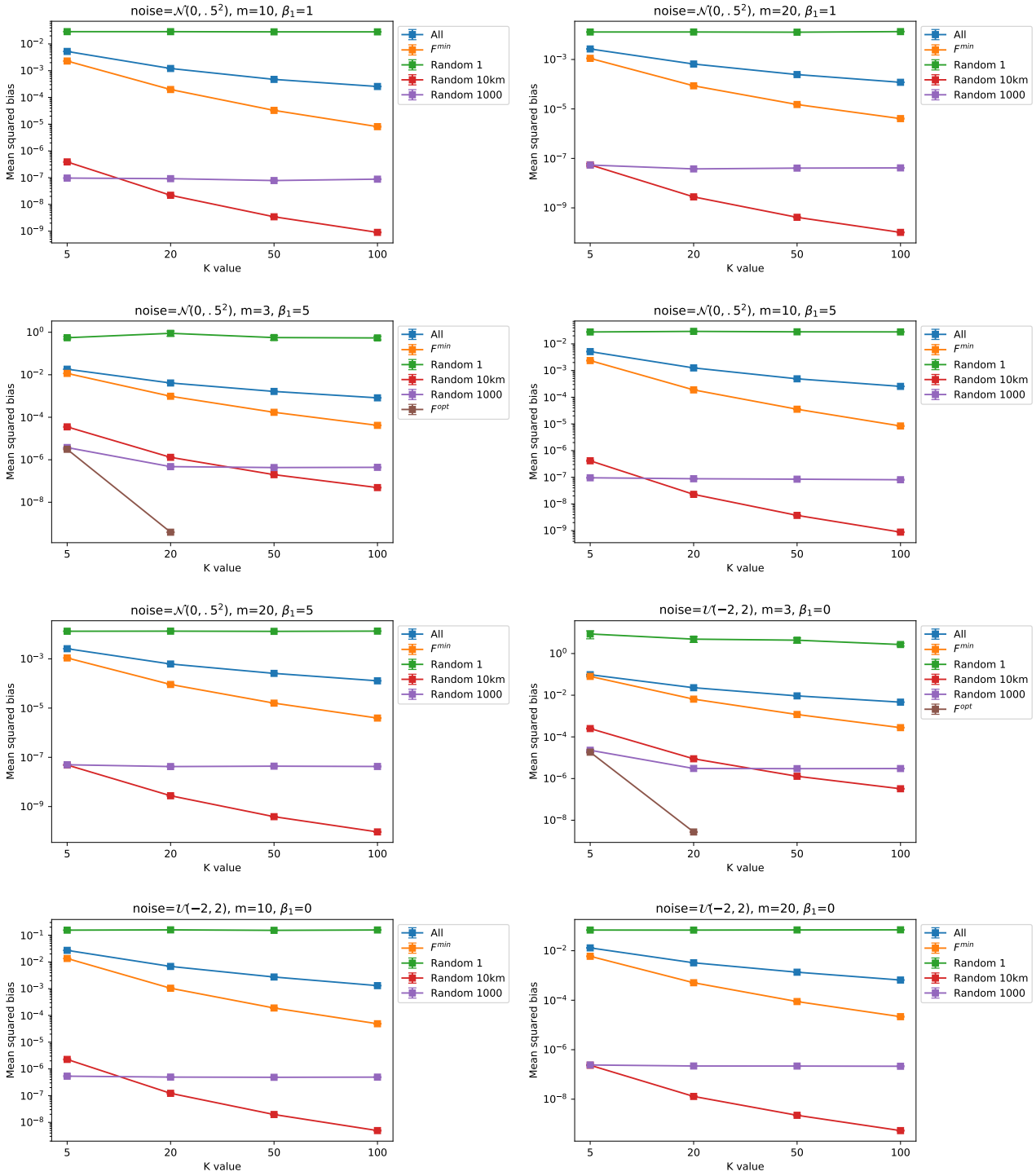
Hence,

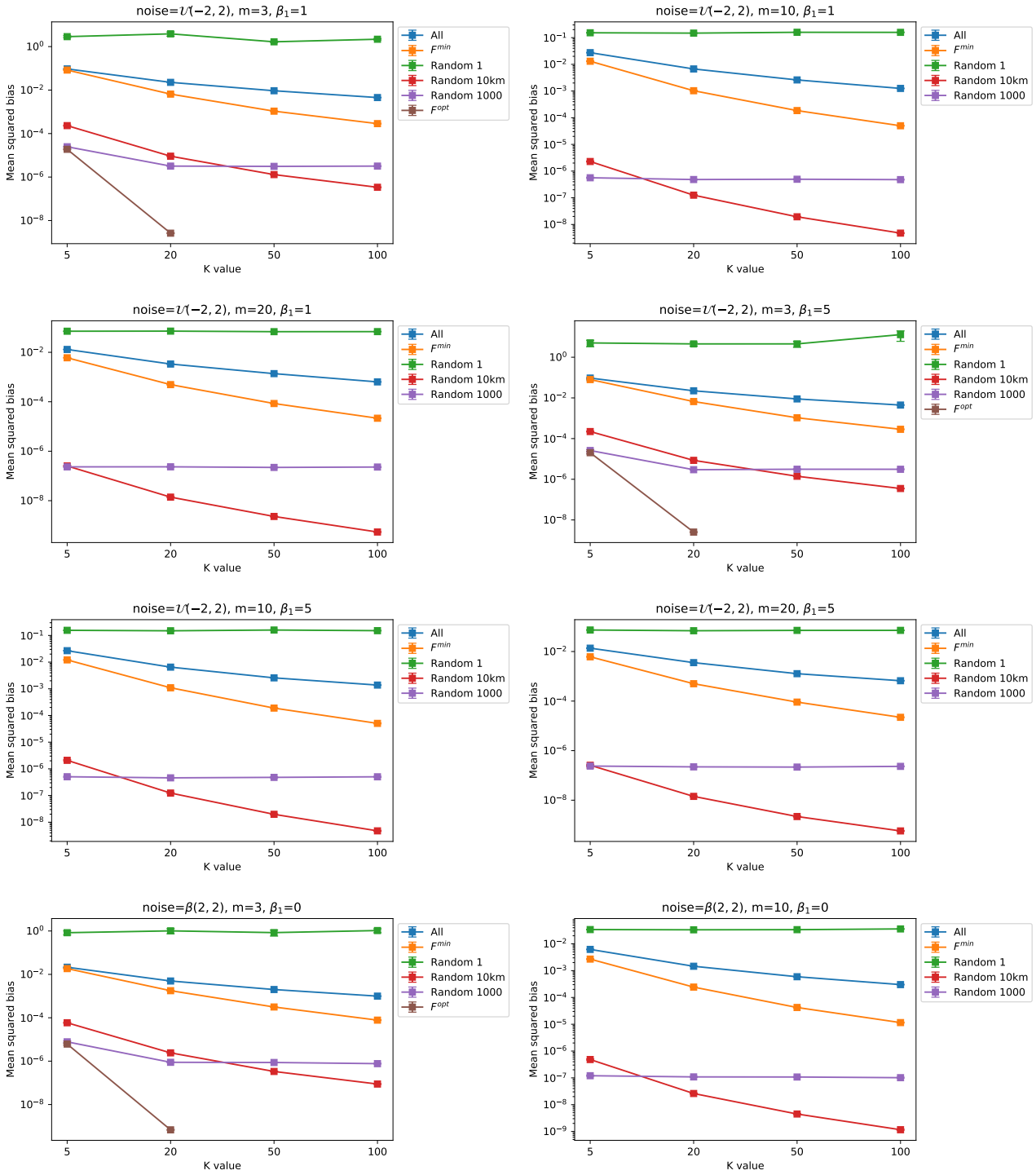
$$\frac{\mathsf{mean}(x_1, \ldots, x_n)^2}{\mathrm{SV}(x_1, \ldots, x_n)} \leq \frac{\mu^2}{\frac{\mu^2}{8}} \leq \frac{1}{8}.$$

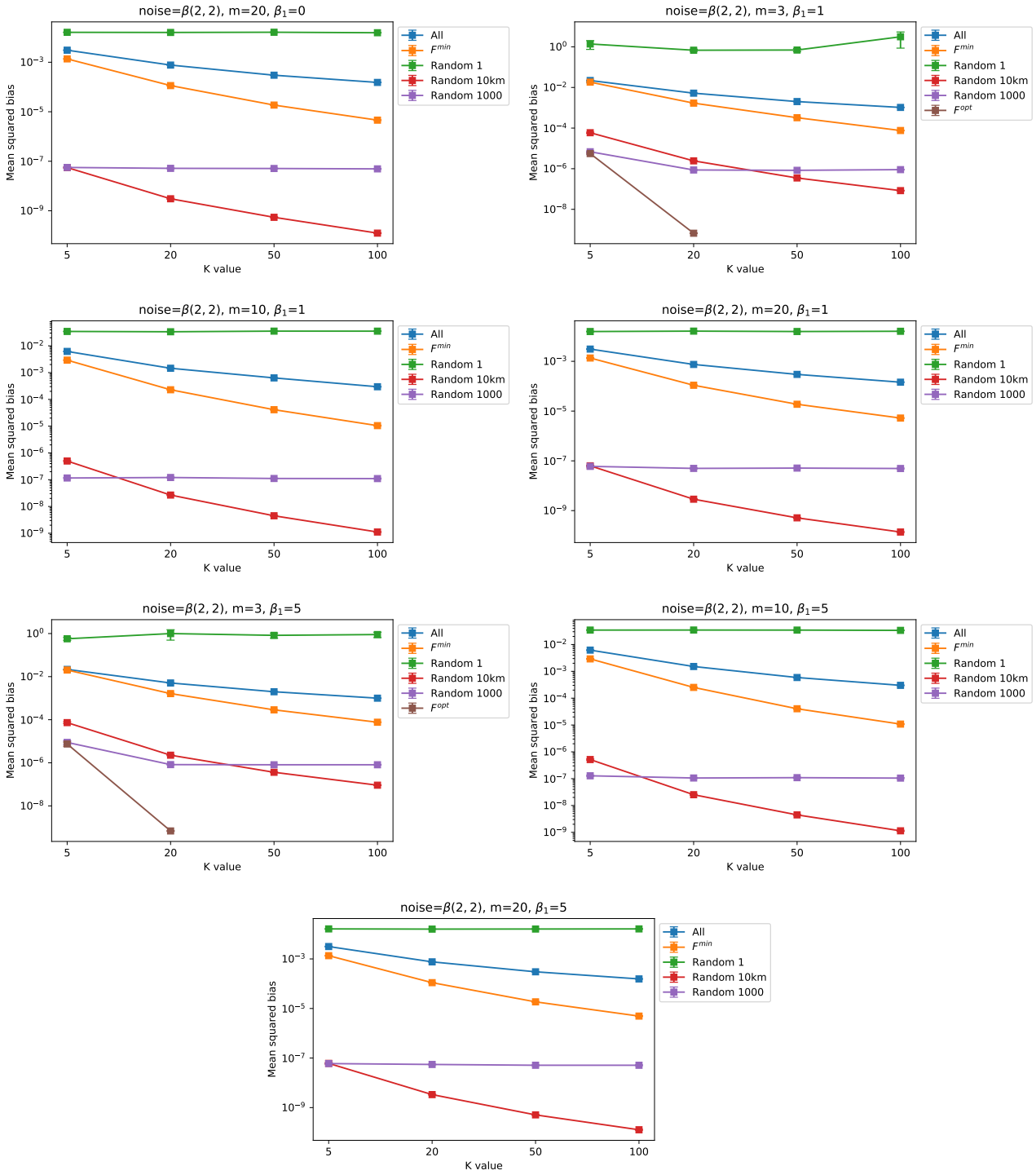
$\frac{\mu^2}{\mathrm{SV}(x_1,\ldots,x_n)} = \frac{1}{8}$. □

B. Additional Plots




noise=$\mathcal{N}(0, .5^2)$, m=10, β_1=1
noise=$\mathcal{N}(0, .5^2)$, m=20, β_1=1
noise=$\mathcal{N}(0, .5^2)$, m=3, β_1=5
noise=$\mathcal{N}(0, .5^2)$, m=10, β_1=5
noise=$\mathcal{N}(0, .5^2)$, m=20, β_1=5
noise=$\mathcal{U}(-2, 2)$, m=3, β_1=0
noise=$\mathcal{U}(-2, 2)$, m=10, β_1=0
noise=$\mathcal{U}(-2, 2)$, m=20, β_1=0
Mean squared bias
K value
All
F^{min}
Random 1
Random 10km
Random 1000
F^{opt}


noise=$\mathcal{U}(-2,2)$, m=3, β_1=1
noise=$\mathcal{U}(-2,2)$, m=10, β_1=1
noise=$\mathcal{U}(-2,2)$, m=20, β_1=1
noise=$\mathcal{U}(-2,2)$, m=3, β_1=5
noise=$\mathcal{U}(-2,2)$, m=10, β_1=5
noise=$\mathcal{U}(-2,2)$, m=20, β_1=5
noise=$\beta(2,2)$, m=3, β_1=0
noise=$\beta(2,2)$, m=10, β_1=0
Mean squared bias
K value
All
F^{min}
Random 1
Random 10km
Random 1000
F^{opt}


noise=β(2, 2), m=20, β1=0
noise=β(2, 2), m=3, β1=1
noise=β(2, 2), m=10, β1=1
noise=β(2, 2), m=20, β1=1
noise=β(2, 2), m=3, β1=5
noise=β(2, 2), m=10, β1=5
noise=β(2, 2), m=20, β1=5
Mean squared bias
K value
All
F^min
Random 1
Random 10km
Random 1000
F^opt