Approximately Strategy-Proof Voting

Eleanor Birrell*
Cornell University
eleanor@cs.cornell.edu

Rafael Pass†
Cornell University
rafael@cs.cornell.edu

June 3, 2013

Abstract

The classic Gibbard-Satterthwaite Theorem establishes that only dictatorial voting rules are strategy-proof; under any other voting rule, players have an incentive to lie about their true preferences. We consider a new approach for circumventing this result: we consider randomized voting rules that only *approximate* a deterministic voting rule and only are *approximately* strategy-proof. We show that *any* deterministic voting rule can be approximated by an approximately strategy-proof randomized voting rule, and we provide lower bounds on the parameters required by such voting rules.

1 Introduction

The classic Gibbard-Satterthwaite Theorem [5, 13] considers the question of when voters will honestly report their preferences. It shows that if the voting rule has at least three outcomes, then only *dictatorial* functions (i.e., one player determines the output) can be honestly computed by rational agents.

The earliest approach to circumventing this limitation, first suggested by Gibbard [6] and later advocated by Conitzer and Sandholm [2], consists of using randomized approximations as a means to bypass the limitations of deterministic rules. Unfortunately, the potential of this approach is limited by two negative results. Gibbard showed that the only strategy-proof randomized voting rules are *trivial* in that they consist of simple probability distributions over rules that depend on only one voter (*unilateral rules*) and rules that have at most two possible outputs (*duple rules*). In recent work, Procaccia [12] quantified the quality of approximation that can be achieved by such trivial functions (for certain types of voting rules): in particular, he constructed a simple approximation of Plurality (the voting rule that returns the outcome that receives the most first-choice votes). However, his approximation only guarantees that the expected number of votes received by the returned

*This material is based upon work supported under a National Science Foundation Graduate Research Fellowship

†Supported in part by a Alfred P. Sloan Fellowship, Microsoft New Faculty Fellowship, NSF CAREER Award CCF-0746990, AFOSR Award FA9550-08-1-0197, BSF Grant 2006317.

output is 2/3 the number received by the true winner, and he proves that his mechanism is asymptotically optimal.

An alternative approach that consists of restricting the class of preference functions. Notably, Moulin [11] considers voting rules defined over an output set with a natural total ordering with single-peaked preferences, that is utilities that have an optimal outcome (the *peak*) and decrease monotonically with distance from the peak. He constructs rules that are strategy-proof with respect to this (very restricted) class of utility functions.

A final intriguing approach to circumventing this limitation, first suggested by Bartholdi et al. [1], is to construct voting rules that are computationally difficult to manipulate. However, there are strong impossibility results that limit the potential of this approach. In their original work, Bartholdi et al. showed that many standard voting rules can be efficiently manipulated. Moreover, recent work demonstrates that for any voting rule, "bad inputs" (preference profiles that admit a successful non-honest strategy) are not rare [2, 4, 8]; these papers develop lower bounds for the number of preference profiles which admit some kind of manipulation and show that when the number of outputs is small it is easy to find successful manipulations.

In this work, we consider a new approach to circumventing these previous negative results: *approximately* strategy-proof voting rules. This approach is motivated by the observation that previous work assumes that people will deviate from the honest strategy even if the improvement in their utility is extremely small. In practice, people often don't deviate if the gain is very small; a small gain in utility may be offset by a psychological cost associated with lying or by the computational cost of computing an effective deviation [7]. This phenomenon is compounded by the fact that when risk-averse individual voters are uncertain about the preferences of the other voters, they may not pursue a deviation that is only expected to yield a small benefit. In other contexts, these observations have led to the introduction of relaxed solution concepts (e.g., ε-Nash equilibria and ε-dominant strategies). In the context of voting, we thus consider ε-strategy-proof voting rules; that is voting rules for which no deviating strategy can improve a player's expected utility by more than ε.

Unfortunately, a corollary of the Gibbard-Satterthwaite theorem shows that the only deterministic voting rules with three or more outcomes that are ε-strategy-proof are dictatorial rules. We thus consider approximate strategy-proofness in the context of randomized voting rules. As it turns out, the parameter ε quite sharply determines whether or not there exist non-trivial ε-strategy-proof voting rules.

For the remainder of this section, let n be the number of voters and let the number of outcomes be a fixed constant k.

$\varepsilon = \omega(1/n)$: In this regime, we show that there exist natural, non-trivial, ε-strategy-proof voting rules. Moreover, we show that *every* deterministic voting rule f can be approximated by a non-trivial, ε-strategy-proof voting rule g. Towards formalizing this, we require a notion of what it means for a randomized mechanism to approximate a deterministic voting rule. Intuitively, we want to capture the idea that with high probability, the output of g is "close" to that of f. To define closeness, we consider a specialized distance metric inspired by the idea of vote corruption, that is the observation that in real-life elections, votes get miscounted, recounted, lost, etc. We say that the output $y \in g(\vec{x})$ is δ-close to the right answer if there exists some vector $\vec{x}'$ that differs from $\vec{x}$ in only δ positions, such that $f(\vec{x}') = y$; in other

words, when the output is δ-correct, it means that we could have reached this output by flipping only δ votes. Our main theorem can now be informally stated as follows:

Theorem 1.1 (Upper Bound – Informal Statement)**.** *Let $\varepsilon = \omega(1/n)$, $\beta > 0$, and let f be a deterministic voting rule over n players. For sufficiently large n, there exists an ε-strategy-proof randomized voting rule g that is a βn-approximation of f.*

Intuitively speaking, our mechanism guarantees that the outcome returned by g will be the correct one modulo a change in a few votes. When dealing with large election (e.g., national elections) this seems like a reasonable guarantee.

One may ask whether an alternative notion of approximation can be achieved: For instance, could we hope to get a mechanism that yields the *correct* output with high probability? In Section B we consider this alternative definition, and we show that this is impossible unless the underlying voting rule f is dictatorial. We believe this negative result highlights why our definition of approximation is both reasonable and minimal for circumventing the negative results of Gibbard and Satterthwaite.

$\varepsilon = o(1/n^2)$**:** In this setting, we show that the Gibbard's result characterizing strategy-proof randomized voting rules [6] still applies:

Theorem 1.2 (Lower Bound – Informal Statement)**.** *If g is a $o(1/n^2)$-strategy-proof randomized voting rule, then g is trivial (i.e., a distribution over unilateral and duple rules).*

We additionally show that natural voting rules (e.g., Plurality) cannot be well approximated by such mechanisms. This result can be informally stated as follows:

Theorem 1.3 (Plurality – Informal Statement)**.** *Let g be a trivial voting rule over n players. There exists β such that for sufficiently large n, g cannot be a $\beta n^2 2$-approximation of* Plurality*.*

In fact, in Appendix B we extend this result to show that there is no trivial voting rule that approximates Plurality even with high probability. The two theorems combined bound the approximation parameters that can be achieved for natural voting rules like Plurality.

Additional Properties Finally, we observe that there are many properties (other than strategy-proofness) that are desirable in a voting rule. We show that in addition to being ε-strategy-proof, the mechanisms we construct are *collusion-resistant*—a group of t players cannot increase their collective utility by more than a small amount. Moreover, when we consider a *neutral* voting rule—one whose outcome is independent of voter identities—and a constant number of outcomes, our mechanism is computationally efficient. Finally, we show that our definitions are robust: in Appendix B we extend our results to a relaxed notion of approximation that considers functions that are close to correct with high probability.

2 Definitions and Preliminaries

A voting rule f is a mapping from player "votes" to an outcome in the set $[k] = \{1, \ldots, k\}$. Player votes are total preference orderings over the set of outcomes $[k]$; these preference

orderings are represented as a permutation $\sigma_i \in \Sigma_k$ (here Σ_k denotes the set of permutations over $[k]$). We use the notation $\sigma_i(j) > \sigma_i(j')$ when the preference type $\sigma_i \in \Sigma_k$ ranks outcome j higher than outcome j'. For convenience, a preference profile $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n)$ may also be denoted by $(\sigma_i, \vec{\sigma}_{-i})$ for any $i \in [n]$.

Players are motivated by determinism—that is, they prefer that outcomes ranked higher in their preference ordering σ_i are chosen. More formally, each player i has a preference ordering σ_i and a utility function $u_i \in \mathcal{U}_{\sigma_i}$, the class of utility functions u_i that satisfy the following two properties: First, for all $\vec{\sigma}_{-i}$, $u_i(\vec{\sigma}, j) \geq u_i(\vec{\sigma}, j')$ if $\sigma_i(j) > \sigma_i(j')$. Second, for all input-output pairs $(\vec{\sigma}, j)$, $u_i(\vec{\sigma}, j) \in [0, 1]$.

Observe that although it is traditional to talk exclusively about the preference relations σ_i, the underlying utility function is necessary in order to quantify *approximate* strategy-proofness. The restriction to utilities in the range $[0, 1]$ is not strictly necessary (our techniques only rely on the fact that the utilities are bounded), however we believe that utilities in this range lend themselves to a more intuitive interpretation.

2.1 Approximately Strategy-proof Voting

We say that a voting rule is (approximately) strategy-proof if for all players, honestly reporting their preferences (approximately) dominates any other reporting strategy.

Definition 2.1 (ε-strategy-proof)**.** A voting rule g is *ε-strategy-proof* if for all players i, all preference profiles $\vec{\sigma}$, all alternative preferences σ_i', and all utility functions $u_i \in \mathcal{U}_{\sigma_i}$,

$$\mathbb{E}[u_i(\vec{\sigma}, g(\vec{\sigma}))] + \varepsilon \geq \mathbb{E}[u_i(\vec{\sigma}, g(\sigma_i', \vec{\sigma}_{-i})]$$

The notion of strategy-proof voting rules can also be extended to handle collusions between groups of t players.

Definition 2.2. A voting rule $g : (\Sigma_k)^n \to [k]$ is *(t, ε)-strategy-proof* if for all subsets $S \subseteq [n]$ such that $|S| \leq t$, all preference profiles $\vec{\sigma}$ and $\vec{\sigma}'$ such that $\sigma_i = \sigma_i'$ for all $i \notin S$, and all utility profiles $\vec{u} \in \mathcal{U}_{\vec{\sigma}}$,

$$\sum_{i \in S} \mathbb{E}\left[u_i(\vec{\sigma}, g(\vec{\sigma}))\right] + \varepsilon \geq \sum_{i \in S} \mathbb{E}\left[u_i(\vec{\sigma}, g(\sigma')\right].$$

Intuitively, we interpret ε-strategy-proofness to mean that if there is a small cost associated with deviating (e.g., a psychological cost to lying or a computational cost to finding a successful deviation) then the voters will honestly report their preferences.

2.2 Voter Max-Influence

We introduce a new concept inspired by the notion of variable influence [9] which we call *max-influence*; max-influence describes the maximum amount by which a single player can impact the probability that a particular output is returned by a voting rule.

Definition 2.3 (ξ-max-influence)**.** A player $i \in [n]$ has *ξ-max-influence* over a voting rule $g : (\Sigma_k)^n \to [k]$ if there exists a preference profile $\vec{\sigma}$ and an alterative preference σ_i' such that the statistical distance between $g(\vec{\sigma})$ and $g(\sigma_i', \vec{\sigma}_{-i})$ is at least ξ, that is if

$$\frac{1}{2} \sum_{j \in [k]} |\Pr[g(\vec{\sigma}) = j] - \Pr[g(\sigma_i', \vec{\sigma}_{-i}) = j]| > \xi$$

.

We observe that if no player i has ξ-max-influence over a voting rule g then the rule is approximately strategy-proof.

Lemma 2.4. *Let $g : (\Sigma_k)^n \to [k]$ be a voting rule such that no player $i \in [n]$ has ξ-max-influence over g. Then g is 2ξ-strategy-proof.*

Proof. Let $\vec{\sigma}, \vec{\sigma}'$ be preference profiles that differ only on input i (that is $\sigma_j = \sigma_j'$ for all $j \neq i$), and let u_i be voter i's utility function. Let $S_{\vec{\sigma}} = \{j : \Pr[g(\vec{\sigma}) = j] \geq \Pr[g(\vec{\sigma}') = j]\}$ and let $S_{\vec{\sigma}'}$ be its complement (the set of outputs that are more likely to be generated by $\vec{\sigma}'$).

$$\begin{aligned}
E[u_i(\vec{\sigma}, g(\vec{\sigma}'))] - E[u_i(\vec{\sigma}, g(\vec{\sigma}))] &= \sum_j Pr[g(\vec{\sigma}') = j] u_i(\vec{\sigma}, j) - \sum_j Pr[g(\vec{\sigma}) = j] u_i(\vec{\sigma}, j) \\
&= \sum_{j \in S_{\vec{\sigma}'}} (\Pr[g(\vec{\sigma}') = j] - \Pr[g(\vec{\sigma}) = j]) u_i(\vec{\sigma}, j) \\
&\quad + \sum_{j \in S_{\vec{\sigma}}} (\Pr[g(\vec{\sigma}') = j] - \Pr[g(\vec{\sigma}) = j]) u_i(\vec{\sigma}, j) \\
&\leq \sum_{j \in S_{\vec{\sigma}'}} (\Pr[g(\vec{\sigma}') = j] - \Pr[g(\vec{\sigma}) = j]) \cdot 1 + 0 \\
&\leq 2\xi
\end{aligned}$$

Therefore the constructed approximation g is 2ξ-strategy-proof. □

We also present an analogous result for coalitions.

Lemma 2.5. *Let $g : (\Sigma_k)^n \to [k]$ be a voting rule such that no player $i \in [n]$ has ξ-max-influence over g. Then g is $(t, 2t^2\xi)$-strategy-proof.*

Proof. For any subset of t players $S = \{s(1), \ldots, s(t)\} \subseteq [n]$, let $\vec{\sigma}, \vec{\sigma}'$ be two preference profiles that differ only on components $i \in S$, that is for all $i \notin S$, $\sigma_i = \sigma_i'$. Define a sequence of hybrid profiles $\vec{\sigma}^0, \ldots, \vec{\sigma}^t$ such that in profile $\vec{\sigma}^\ell$, players $j \in \{s(1), \ldots, s(\ell)\}$ have preference $\sigma_j^\ell = \sigma_j'$, players $j \in \{s(\ell+1), \ldots, s(t)\}$ have preference $\sigma_j^\ell = \sigma_j$ and all other players $j \notin S$ have preference $\sigma_j^\ell = \sigma_j = \sigma_j'$. Observe that $\vec{\sigma} = \vec{\sigma}^0$ and $\vec{\sigma}' = \vec{\sigma}^t$.

Observe that for all $i \in S$ and all $\ell \in [t]$, $\mathbb{E}[u_i(\vec{\sigma}, g(\vec{\sigma}^\ell))] - \mathbb{E}[u_i(\vec{\sigma}, g(\vec{\sigma}^{\ell-1}))] \leq 2\xi$ (the proof is identical to the proof of 2ξ-strategy-proofness in Lemma 2.4). It follows the difference in the group's collective utility between any two consecutive hybrid profiles is bounded by $\sum_{i \in S} \mathbb{E}[u_i(\vec{\sigma}, g(\vec{\sigma}))] - \sum_{i \in S} \mathbb{E}[u_i(\vec{\sigma}, g(\vec{\sigma}')] \leq 2t\xi$ and therefore that the total difference in collective utility between the honest strategy and the collective deviating strategy is at most

$$\sum_{i \in S} \mathbb{E}[u_i(\vec{\sigma}, g(\vec{\sigma}))] - \sum_{i \in S} \mathbb{E}[u_i(\vec{\sigma}, g(\vec{\sigma}')] \leq 2t^2\xi$$ □

Observe that Gibbard and Satterthwaite's classic result shows that every non-dictorial voting rule with at least three outputs has a player with 1-max-influence. By contrast, we will show that every voting rule can be approximated by a randomized voting rule over which no player has much max-influence.

2.3 Approximations

In order to formally define the notion of a randomized approximation, it is necessary to define a closeness metric d for voting rules. The appropriate choice of metric in such a context is not immediately clear. Since outcomes are assigned an arbitrary number $j \in [k]$ and votes are permutations σ_i over the outcomes (interpreted as a total preference ordering), standard notions of "closeness" are meaningless. While some particular voting rules have natural quality scores that can be used to define an approximation [12], such techniques are not fully generalizable.

We instead introduce a pseudometric d_v inspired by vote corruption, that is the observation that in real-life elections, votes get miscounted, recounted, lost, etc. This pseudometric d_v associates a number ℓ with each pair $(\vec{\sigma}, j) \in (\Sigma_k)^n \times [k]$ defined by

$$\ell(\vec{\sigma}, j) = \min_{\vec{\sigma}' \text{ s.t. } f(\vec{\sigma}')=j} \Delta(\vec{\sigma}, \vec{\sigma}')$$

where $\Delta(\vec{\sigma}, \vec{\sigma}')$ is the number of components which differ between $\vec{\sigma}$ and $\vec{\sigma}'$. We define an induced pseudometric $d_v((\vec{\sigma}, j), (\vec{\sigma}', j')) = |\ell(\vec{\sigma}, j) - \ell(\vec{\sigma}', j')|$. This says that an output is close to the correct answer if there exists an input close to the true input which generates that output.

Using this metric, we consider an approximation g to be a function whose value is *close* to that of f.

Definition 2.6 (δ-approximation)**.** A (randomized) voting rule g is a δ-approximation of a voting rule f if for all inputs $\vec{\sigma}$ and all possible random coins,

$$d_v((\vec{\sigma}, g(\vec{\sigma})), (\vec{\sigma}, f(\vec{\sigma}))) \leq \delta.$$

Remark 2.7. In Section B, we relax our definition of an approximation to consider functions that are close to the correct outcome with high probability and show that our lower bounds extend to the relaxed definition.

2.4 Trivial Voting Rules

There are two classes of simple voting rules, collectively referred to as trivial, that will be used to characterize the set of strategy-proof voting rules under certain conditions. The first is the class of rules that depend on only one player's inputs, and the second is the class that returns at most two outputs.

Definition 2.8 (Gibbard [6])**.** A deterministic voting rule $f : (\Sigma_k)^n \to [k]$ is *unilateral* if there exists a player i such that for all preferences profiles $\vec{\sigma}, \vec{\sigma}' \in (\Sigma_k)^n$ satisfying $\sigma_i = \sigma_i'$, $f(\vec{\sigma}) = f(\vec{\sigma}')$. A voting rule that is unilateral and onto is also called *dictatorial*.

Definition 2.9 (Gibbard [6])**.** A deterministic voting rule $f : (\Sigma_k)^n \to [k]$ is a *duple* if the range is at most 2, that is if $|\{j \in [k] : \exists \vec{\sigma} \in (\Sigma_k)^n \text{ such that } f(\vec{\sigma}) = j\}| \leq 2$.

A voting rule is called *trivial* if it is a probability distribution over unilateral and duple rules, otherwise it is called *non-trivial*.

3 Previous Work

Randomized voting has been recently explored by Procaccia [12] who shows how to define 0-strategy-proof approximations of certain voting rules that are derived from natural quality scores. Our work differs from his approach both in our use of ε-strategy-proof rules and in the way we define and construct approximations. In particular, Procaccia's work relies on a natural quality score to define an approximation, a technique that prevents discussion of certain types of voting rules including multi-layered rules like run-off elections or the electoral college-based system employed in U.S. presidential elections. Furthermore, as Procaccia demonstrates, the quality-score based approach is inherently limited in the quality of approximations that can be achieved; for example, it is impossible to construct a strategy-proof approximation of PLURALITY that will (in expectation) return an outcome with quality score greater $\Omega(1/\sqrt{k})$ times the optimal. Although he does construct an approximation for PLURALITY, if his approximation were employed during an election between four candidates, the expected number of votes received by the candidate it returns would be 2/3 the number of votes received by the true winner.

There are two important negative results in the context of voting. The first, proven independently by Gibbard [5] and Satterthwaite [13], demonstrates that only a trivial collection of voting rules are strategy-proof. Restated with our definitions, they show that only *dictatorial* functions are 0-strategy-proof.

Theorem 3.1 (Gibbard-Satterthwaite)**.** *Let $f : (\Sigma_k)^n \to [k]$ be a deterministic, onto voting rule with $k \geq 3$. Then f is 0-strategy-proof if and only if it is dictatorial.*

This result has been quantitatively extended to give lower bounds on the number of input profiles which admit manipulations [4, 8]. Their work shows that not only are manipulations relatively common, but they can also be found efficiently.

The Gibbard-Satterthwaite theorem has also been extended to characterize the class of strategy-proof randomized voting rules [6]. In terms of our definitions, this extension shows that only trivial voting rules are 0-strategy-proof.

Theorem 3.2 (Gibbard)**.** *Let $g : (\Sigma_k)^n \to [k]$ be a randomized voting rule. Then g is 0-strategy-proof if and only if it is trivial.*

4 Approximate Voting

Leveraging our new notion of approximate strategy-proofness, we now show that *every* voting rule can be approximated by a randomized voting rule. Intuitively, we construct an approximation by adding noise to the original deterministic voting rule. The probability

that these approximations return a particular outcome decreases linearly with the distance between the outcome under consideration and the correct outcome; the slope of this linear function bounds the influence that a voter can have on the resulting approximation. If the slope is sufficiently steep, then we can guarantee that all "bad" outputs (ones that are δ-far from the correct output) are chosen with probability 0.

Our techniques can be seen as a linear analog of the exponential mechanism [10] that has been used to establish differential privacy [3].

Theorem 4.1. *For any deterministic voting rule $f : (\Sigma_k)^n \to [k]$, any $\varepsilon > 0$ and any $\delta \geq k(k+1+\varepsilon)/\varepsilon - 1$, f has a ε-strategy-proof δ-approximation g.*

Proof. We construct an approximation g of the voting rule f as follows: we assign each input-output pair $(\vec{\sigma}, j)$ a quality score $q(\vec{\sigma}, j) = -d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, j))$. Observe that $q(\vec{\sigma}, j)$ decreases linearly with the (minimal) number of votes that must be corrupted before f returns j instead of $f(\vec{\sigma})$. Let $\xi = \varepsilon/k(k+1+\varepsilon)$—this value is chosen to guarantee ε-strategy-proofness. The mechanism g returns the value j with probability proportional to $\max\{1 + \xi q(\vec{\sigma}, j), 0\}$. Note that g never returns an outcome more than $1/\xi$-far from $f(\vec{\sigma})$.

First, we bound the max-influence a voter can have over g. For any $\vec{\sigma}'$ that differs from $\vec{\sigma}$ in only one position and any outcome $j \in [k]$, the difference $|\Pr[g(\vec{\sigma}') = j] - \Pr[g(\vec{\sigma}) = j]|$ is equal to

$$\left| \frac{\max\{1 + \xi q(\vec{\sigma}', j), 0\}}{\sum_{\iota \in [k]} \max\{1 + \xi q(\vec{\sigma}', \iota), 0\}} - \frac{\max\{1 + \xi q(\vec{\sigma}, j), 0\}}{\sum_{\iota \in [k]} \max\{1 + \xi q(\vec{\sigma}, \iota), 0\}} \right|.$$

For the purpose of clarity, let $A = \max\{1 + \xi q(\vec{\sigma}, j), 0\}$ and let $B = \sum_{\iota \in [k]} \max\{1 + \xi q(\vec{\sigma}, \iota), 0\}$. Using this notation we observe that the difference can be bounded above by

$$\left| \frac{A + \xi}{B - k\xi} - \frac{A}{B} \right| = \left| \frac{AB + \xi B - AB + k\xi A}{B^2 - k\xi B} \right| = \left| \frac{\xi - k\xi A/B}{B - k\xi} \right|$$

Where the first expression follows from the definition of g, the second from cross multiplication, and the third from canceling terms.

Since $A = \max\{1 + \xi q(\vec{\sigma}, j), 0\} \leq 1$ (recall that quality scores are negative) and $B = \sum_{\iota \in [k]} \max\{1 + \xi q(\vec{\sigma}, j), 0\} \geq 1$ (since the correct output, which is included in the sum, contributes 1 and there are no negative components), we can maximize this bound by setting $A = B = 1$ giving

$$|\Pr[g(\vec{\sigma}') = j] - \Pr[g(\vec{\sigma}) = j]| \leq \frac{\xi + k\xi}{1 - k\xi} = \varepsilon/k$$

which implies that no player has max-influence greater than $\varepsilon/2$. Therefore by Lemma 2.4 the constructed approximation g is ε-strategy-proof.

Second, we claim that g is a good approximation for f according to the distance metric d_v. Observe that the approximation g returns a outcome with distance greater than $1/\xi$ from the correct answer with probability 0. Since we fixed $\delta \geq 1/\xi$, all "bad" outputs are sufficiently far from the correct answer that they are returned with probability zero, therefore g always returns an answer that is δ-close to the correct outcome. □

Corollary 4.2. *Let $\varepsilon = \omega(1/n)$, $\beta > 0$, and let $f : (\Sigma_k)^n \to [k]$ be a deterministic voting rule. For sufficiently large n, there exists an ε-strategy-proof randomized voting rule g that is a βn-approximation of f.*

Proof. Fix $\beta > 0$ and let $\delta = \beta n$. Let g be defined as in the proof of Theorem 4.1 and recall that (as shown in the previous proof) g is an ε-strategy-proof βn-approximation of f if $\beta n \geq k(k+1+\varepsilon)/\varepsilon - 1$. Since $\varepsilon = \omega(1/n)$, the result follows immediately. □

Example 4.3. For concreteness, consider an election with 100 million voters and three outputs (about the scale of a United States presidential election). Fixing $\varepsilon = .001$, we can construct an approximation g that is guaranteed to *always* return an answer within 12,500 votes of the correct answer, in practice, well within the vote corruption in such an election. Looked at in another way, this says that in any such election in which one outcome wins by at least 12,500 votes, this mechanism will always return the correct answer.

Observe that if $\varepsilon < 1/n$ then all outputs are chosen with positive probability. For such small values of ε, the approximation g is well-defined, but it is a trivial n-approximation of f. In Section 5 we show that this $\omega(1/n)$ restriction is an inherent bound on the achievable approximation parameters and not an artifact of the construction employed in Theorem 4.1. In contrast, when $\varepsilon = \omega(1/n)$, the approximations both offer good guarantees and are non-trivial.

Corollary 4.4. *Let $\varepsilon = \omega(1/n)$ and $k = O(1)$. For sufficiently large numbers of voters n, there exist non-trivial ε-strategy-proof randomized voting rules.*

Proof. Let f be PLURALITY, the voting rule that returns the outcome j that receives the most first-choice votes (using some deterministic tie-breaking rule). By Corollary B.4 we know that there exist arbitrarily good approximations of f for $\varepsilon = \omega(1/n)$. By Theorem 5.4 this implies that for such ε, the mechanism from Theorem 4.1 is non-trivial. □

In addition to being non-trivial and approximately strategy-proof, the randomized voting rule g constructed in the proof of Theorem 4.1 has several other nice properties. First, the voting rule g is collusion-resistant: its guarantees degrade gracefully if the mechanism is extended to protect against collusion by t players.

Corollary 4.5. *For any voting rule $f : (\Sigma_k)^n \to [k]$, any $t < n$, any $\varepsilon > 0$, and any $\delta \geq (k(k_1)t^2 + k\varepsilon)/\varepsilon - 1$, f has a (t, ε)-strategy-proof δ-approximation.*

Proof. The proof is equivalent to that of Theorem 4.1 except that we use a different parameter $\xi = \varepsilon/k(k+1)t^2 + k\varepsilon$. The proof that the randomized voting rule g—constructed as in the proof of Theorem 4.1 (except with the new value of ξ)—is (t, ε)-strategy-proof follows immediately from Lemma 2.5. The proof that g is a δ-approximation of f is identical to the proof in Theorem 4.1. □

Second, we observe that for *neutral* voting rules—those whose outcome is independent of voter identities—our approximations are computationally efficient.

Corollary 4.6. *If $f : (\Sigma_k)^n \to [k]$ is a neutral, efficiently computable voting rule with a constant number of outputs k, then the approximation g defined in the proof of Theorem 4.1 can be computed in polynomial time.*

Proof. To compute the output of the approximation $g(\vec{\sigma})$, where g is defined as in the proof of Theorem 4.1, we begin by computing the quality score $q(\vec{\sigma}, j)$ for each possible outcome $j \in [k]$. Since f is neutral, the correct outcome $f(\vec{\sigma})$ depends only on the vote configuration—that is the unordered set of votes that were cast—and not on the full preference profile $\vec{\sigma}$ per se. We can therefore compute the quality scores of all outputs $j \in [k]$ simply by evaluating the original rule f on each of the possible vote configurations and setting the quality score of an output to be equal to the (negative) minimum distance between the configuration associated with $\vec{\sigma}$ and a configuration that returns the outcome j. There are $k!$ different ways a player could vote and each voting preference is cast by at most n voters, therefore there are $O(n^{k!})$ vote configurations that need to be considered.

Having computed all of the quality scores $q(\vec{\sigma}, j)$, we can define a distribution $D_{\vec{\sigma}}$ such that the probability that $D_{\vec{\sigma}}$ returns an outcome $j \in [k]$ is given by $D_{\vec{\sigma}}(j) = \max\{1 + \xi q(\vec{\sigma}, j), 0\} / \sum_{\iota \in [k]} \max\{1 + \xi q(\vec{\sigma}, \iota), 0\}$. Observe that this distribution is efficiently samplable, therefore the approximation g can be computed efficiently computed. □

5 Optimality of our Approximations

In this section, we develop lower bounds on the approximation parameters that can be achieved for approximately strategy-proof voting rules. In particular, we demonstrate that the only voting rules that are ε-strategy-proof (for small values of ε) are close to trivial, and we show that such voting rules cannot provide good approximations of natural voting rules like Plurality.

We begin by giving an extension of the Gibbard-Satterthwaite Theorem which says that ε-strategy-proof voting rules are close to trivial.

Theorem 5.1. *If f is ε-strategy proof then it can be expressed as a probability distribution over voting rules such that with probability at least $1 - 72n^2k^4\varepsilon$, the selected rule is either unilateral or binary.*

The proof, which closely follows that of Gibbard [6], is given in Appendix A. We also observe briefly that this extended result implies that for moderate values of ε, any *deterministic* ε-strategy-proof voting rule is either deterministic or trivial.

In spite of this limitation, some voting rules still have good approximation for small values of ε.

Example 5.2. Define a class of voting rules Subset-$\delta : (\Sigma_k)^n \to [k]$ by Subset-$\delta(\vec{\sigma}) =$ Plurality$(\sigma_1, \ldots, \sigma_\delta)$. The rule g that returns an output $j \in [k]$ uniformly at random is a 0-strategy-proof δ-approximation of Subset-δ.

Example 5.3. Define a class of voting rules Modulo-$k : (\Sigma_k)^n \to [k]$ that returns the *number* of first-choice votes given to the winner (under the voting rule Plurality) modulo k. Again the rule g that returns an output k uniformly at random is a 0-strategy-proof $\lfloor k/2 \rfloor$-approximation.

However, neither of these voting rules satisfy our intuition concerning what constitutes a "good" voting rule. Subset-δ is not *neutral*—that is the outcome depends on the order of the players—and Modulo-k is not monotonic.

We focus instead on a common, natural voting rule: PLURALITY. We show that under our definition of an approximation, no trivial voting rule is a good approximation of PLURALITY. The proof closely follows that of Procaccia [12].

Theorem 5.4. *If $\varepsilon = o(1/n^2)$ then there exists $\beta > 0$ such that for all n,* PLURALITY *does not have a trivial βn-approximation.*

Proof. Fix a constant value $c > 0$ (e.g., $c = \sqrt{k}$). Observe that any randomized voting rule g can be considered as a probability distribution over three types of voting rules: unilateral rules, duple rules, and other rules. Let g be an ε-strategy proof approximation of PLURALITY. We bound the quality of approximation we can achieve by seperately considering these three types of voting rules.

First consider the case where g selects a duple. Define $q_j = \Pr[j \in \text{Range}(g) | g \text{ selects a duple}]$ and observe that $\sum_{j\in[k]} q_j = 2$ which implies that there exists a set $A' \subseteq [k]$ of k/c alternatives such that $\sum_{j\in A'} q_j \leq 2/c$.

Now consider the case where g selects a unilateral rule. Observe that there exists a set $N' \subseteq [n]$ of n/c players such that a unilateral rule that considers only an agent $i \in N'$ is selected with probability at most $1/c$ (conditioned on g selecting a unilateral rule). Construct a partial preference profile defined for players $i \in [n] \setminus N'$ that satisfies the property that each outcome $j \in [k] \setminus A'$ is ranked first by equally many agents $i \in [n] \setminus N'$ (and no such agent ranks any other outcome first). Since $|A'| \geq k/c$, there exist an output $x^* \in A'$ such that for all preference profiles $\vec{\sigma}$ consistent with the defined preferences σ_i, the probability that $g(\vec{\sigma}) = x^*$ (conditioned on g selecting a unilateral rule that considers only an agent $i \in [n] \setminus N'$) is at most c/k. Complete the preference profile $\vec{\sigma}$ so that it satisfies the property that for all $i \in N'$, x^* is ranked first. Observe that

$$\begin{aligned}
\Pr[g(\vec{\sigma}) = x^* | g \text{ selects unilateral rule}] &= \Pr[g(\vec{\sigma}) = x^* | g \text{ selects unilateral rule in } N'] \\
&\quad \cdot \Pr[g \text{ selects unilat. rule in } N' | g \text{ selects unilat. rule}] \\
&\quad + \Pr[g(\vec{\sigma}) = x^* | g \text{ selects unilateral rule in } [n] \setminus N'] \\
&\quad \cdot \Pr[g \text{ selects unilat. rule in } [n] \setminus N' | g \text{ selects unilat.}] \\
&\leq 1 \cdot 1/c + c/k \cdot 1 \\
&\leq 1/c + c/k
\end{aligned}$$

Finally, consider the case where g selects a voting rule that is neither unilateral nor duple. In this case, the behavior of the selected voting rule (and hense the quality of approximation) is arbitrary. However, since g is ε-strategy proof this case occurs with probability at most $72n^2k^4\varepsilon$.

Since $x^* \in A'$, by combining the two cases, we get that $\Pr[g(\vec{\sigma}) = x^*] \leq \max\{2/c, 1/c + c/k\} + 72n^2k^4\varepsilon$.

Finally, we observe that the number of first-choice votes given to x^* is at least n/c. and the number of first-choice votes given to any output $j \neq x^*$ is at most $(n - n/c)/(k - k/c) = n/k$.

We interpret this as saying that simultaneously, all outcomes $j \neq x^*$ are bad (at least $(n/c - n/k)/2$ corrupted votes are required to effect the output of f) and the probability that g returns the one good output x^* is low. This means that if $\max\{2/c, 1/c + c/k\} + 72n^2k^4\varepsilon < 1$ then f does not have a trivial $(n/c - n/k)/2 - 1$-approximation. □

References

[1] J. Bartholdi, C. Tovey, and M. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6:227–241, 1989.

[2] V. Conitzer and T. Sandholm. Nonexistence of voting rules that are usually hard to manipulate. In *Proc. 21st AAAI Conference*, pages 627–634, 2006.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *3rd TCC*, pages 265–284, 2006.

[4] E. Friedgut, G. Kalai, and N. Nisan. Elections can be manipulated often. In *Proc. 49th FOCS*, pages 243–249, 2009.

[5] A. Gibbard. Manipulation of voting schemes: a general result. *Econometrica*, 41(4):587–601, 1973.

[6] A. Gibbard. Manipulation of schemes that mix voting with chance. *Econometrica*, 45:665–681, 1977.

[7] J. Halpern and R. Pass. Game theory with costly computation: Formulation and application to protocol security. In *ICS*, pages 120–142, 2010.

[8] M. Isaksson, G. Kindler, and E. Mossel. The geometry of manipulation - a quantitative proof of the Gibbard Satterthwaite theorem. In *Proc. 50th FOCS*, 2010.

[9] Jeff Kahn, Gil Kalai, and Nathan Linial. The influence of variables on boolean functions (extended abstract). In *FOCS*, pages 68–80, 1988.

[10] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proc. 48th FOCS*, 2007.

[11] H. Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35(4):437–455, 1980.

[12] A. Procaccia. Can approximation circumvent Gibbard-Satterthwaite? In *Proc. 24th AAAI*, pages 836–841, 2010.

[13] M. Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.

A Proof of Theorem 5.1

We begin by making a general claim about the likelihood of returning outcomes that everyone dislikes.

Lemma A.1. *If $\vec{\sigma}$ and $\vec{\sigma}'$ are two preference profiles in which all players rank x last and f is ε-strategy proof, then $|\Pr[f(\vec{\sigma}') = x] - \Pr[f(\vec{\sigma}) = x]| \leq n\varepsilon$.*

Proof. Assume for contradiction that $\delta = \Pr[f(\vec{\sigma}) = x] - \Pr[f(\vec{\sigma}') = x] > n\varepsilon$. By a hybrid argument, there exist preference profiles $\vec{\sigma}^{(i)}, \vec{\sigma}^{(i)'}$ that differ only in a single position (player i) in which all players rank x last and satisfying $\Pr[f(\vec{\sigma}^{(i)}) = x] - \Pr[f(\vec{\sigma}^{(i)'}) = x] > \varepsilon$.

Define a parameter $0 < \xi < 1$ satisfying $1 - \Pr[f(\vec{\sigma}^{(i)'}) = x] > \xi \cdot (\delta/n - \varepsilon)$. Define a utility u_i consistent with $\vec{\sigma}^{(i)}$ that assigns the first $k-1$ ranked outcomes utilities in the range

$$\frac{1 - \Pr[f(\vec{\sigma}^{(i)'}) = x] - \xi(\frac{\delta}{n} - \varepsilon)}{1 - \Pr[f(\vec{\sigma}^{(i)'}) = x]}$$

and assigns the last-ranked utility (x) value 0. Observe that:

$$\begin{aligned}
\mathbb{E}[u_i(f(\vec{\sigma}^{(i)}))] &\leq (1 - \Pr[f(\vec{\sigma}^{(i)}) = x]) \cdot 1 + \Pr[f(\vec{\sigma}^{(i)}) = x] \cdot 0 \leq 1 - \Pr[f(\vec{\sigma}^{(i)'}) = x] - \delta/n \\
\mathbb{E}[u_i(f(\vec{\sigma}^{(i)'}))] &\geq (1 - \Pr[f(\vec{\sigma}^{(i)'}) = x]) \frac{1 - \Pr[f(\vec{\sigma}^{(i)'}) = x] - \xi \cdot (\delta/n - \varepsilon)}{1 - \Pr[f(\vec{\sigma}^{(i)'}) = x]} + \Pr[f(\vec{\sigma}^{(i)'}) = x] \cdot 0 \\
&\geq 1 - p_{x,\ell} - \delta/n + \varepsilon
\end{aligned}$$

Therefore if player i has utility u_i and other players report preferences $\vec{\sigma}_{-i}^{(\ell+1)} = \vec{\sigma}_{-i}^{(\ell)}$, then player i can improve his expected utility by at least ε by reporting preferences $\vec{\sigma}_i^{(\ell)}$ instead of true preference ordering $\vec{\sigma}_i^{(\ell+1)}$. This breaks the ε-strategy proofness of f and thus yields a contradiction. □

We then prove several technical lemmas establishing a sequence of properties that hold for any ε-strategy-proof voting rule.

Definition A.2. A voting rule f is ε-localized if for all $i \in [n]$ and all preference profiles $\vec{\sigma}, \vec{\sigma}'$ that differ only in position i and both σ_i, σ_i' rank the elements of $X \subseteq [k]$ in the first $|X|$ positions, $|\Pr[f(\vec{\sigma}') \in X] - \Pr[f(\vec{\sigma}) \in X]| \leq \varepsilon$.

Lemma A.3. *If f ε-strategy proof then f is ε-localized.*

Proof. If there exist $i \in [n]$, $X \subseteq [k]$, $\vec{\sigma}, \vec{\sigma}'$ such that $\sigma_j = \sigma_j'$ for all $j \neq i$ and both σ_i, σ_i' rank the elements of X in the first $|X|$ positions, let $p(X, \vec{\sigma}) = \Pr[f(\vec{\sigma}) \in X]$ and let $\delta = p(X, \vec{\sigma}') - p(X, \vec{\sigma})$. Assume for contradiction that $\delta > \varepsilon$.

Define a parameter $0 < \xi < 1$ such that $(1 - \xi)\delta > \xi + \varepsilon$ (observe that such a ξ must exist since $\delta > \varepsilon$). Define a utility function u_i such that the first $|X|$ outcomes (according to preference σ_i) are assigned utilities evenly distributed between 1 and $1 - \xi$ and the remaining $k - |X|$ outcomes (according to σ_i) are assigned values evenly distributed between ξ and 0. Observe that:

$$\begin{aligned}
\mathbb{E}[u_i(f(\vec{\sigma}))] &\leq 1 \cdot p(X, \vec{\sigma}) + \xi \cdot (1 - p(X, \vec{\sigma})) = (1 - \xi)p(X, \vec{\sigma}) + \xi \\
\mathbb{E}[u_i(f(\vec{\sigma}'))] &\geq (1 - \xi)p(X, \vec{\sigma}') + 0 \cdot (1 - p(X, \vec{\sigma})) > (1 - \xi)(p(X, \vec{\sigma}) + \delta)
\end{aligned}$$

Since $(1 - \xi)\delta > \xi + \varepsilon$, if player i has utility u_i and all other players report preference $\vec{\sigma}_{-i} = \vec{\sigma}_{-i}'$ then i can improve his expected utility by at least ε by reporting preferences σ_i' instead of true preferences σ. This breaks the ε-strategy proofness of f and thus yields a contradiction. □

Definition A.4. Let σ_i^{x+1} denote the preference profile derived from σ_i by moving x up one position. We say that f is *ε-pairwise responsive* if for all preference profiles $\vec{\sigma}$, all players i and all outcomes $x, y, z \in [k]$, if σ_i ranks x directly above y and $z \notin \{x, y\}$ then $|p(z, (\vec{\sigma}_{-i}, \sigma_i^{x+1})) - p(z, \vec{\sigma})| \leq \varepsilon$.

Lemma A.5. *If f is ε-strategy proof then f is 2ε-pairwise responsive.*

Proof. If σ_i ranks x directly above y and $z \notin \{x, y\}$, let W_z be the set of outcomes ranked above z. Observe that both W_z and $W_z \cup \{z\}$ head both σ_i and σ_i^{x+1}. Since f is ε-localized,

$$|p(W_z, (\vec{\sigma}_{-i}, \sigma_i^{x+1})) - p(W_z, \vec{\sigma})| \leq \varepsilon \text{ and}$$
$$|p(W_z \cup \{z\}, (\vec{\sigma}_{-i}, \sigma_i^{x+1})) - p(W_z \cup \{z\}, \vec{\sigma})| \leq \varepsilon$$

Therefore by the triangle inequality, $|p(z, (\vec{\sigma}_{-i}, \sigma_i^{x+1})) - p(z, \vec{\sigma})| \leq 2\varepsilon$ □

Definition A.6. Let $\sigma_i|_{\{x,y\}}$ denote the relative ordering of outcomes $x, y \in [k]$ under preference σ_i. Define $e_i^y(f, \vec{\sigma}) = p(y, \vec{\sigma}_{-i}, \sigma_i^{y+1}) - p(y, \vec{\sigma})$. f is *ε-pairwise isolated* if for all players i and all preference profiles $\vec{\sigma}, \vec{\sigma}'$ such that $\sigma_i = \sigma_i'$ ranks x directly above y and for all $j \in [n], \sigma_j|_{\{x,y\}} = \sigma_j'|_{\{x,y\}}$, $|e_i^y(f, \vec{\sigma}') - e_i^y(f, \vec{\sigma})| \leq \varepsilon$.

Lemma A.7. *If f is ε-strategy-proof then f is $4nk^2\varepsilon$-pairwise isolated.*

Proof. Fix an outcome $y \in [k]$. Let $\vec{\sigma}, \vec{\sigma}'$ be preference profiles that agree on $\sigma_i = \sigma_i'$, where σ_i ranks x directly above y, and where $\sigma_j|_{\{x,y\}} = \sigma_j'|_{\{x,y\}}$ for all $j \in [n]$. Then there exists a sequence of hybrid preference profiles $\vec{\sigma} = \vec{\sigma}^{(0)}, \ldots, \vec{\sigma}^{(nk^2/2)} = \vec{\sigma}'$ where any two adjacent hybrid profiles $\vec{\sigma}^{(\ell)}, \vec{\sigma}^{(\ell+1)}$ differ only in that a single voter $j \neq i$ switches the order of two adjacent outcomes $\{w, z\} \neq \{x, y\}$ (note that the pairs are not necessarily disjoint, just distinct).

Without loss of generality, let z denote the outcome ranked directly below w in profile $\sigma_j^{(\ell)}$. We now consider two cases: either (1) $y \notin \{w, z\}$ or (2) $x \notin \{w, z\}$.

In case (1), since since $y \notin \{w, z\}$ and since by Lemma A.5 f is 2ε-pairwise responsive, it immediately holds that

$$|p(y, \vec{\sigma}) - p(y, (\vec{\sigma}_{-j}, \sigma_j^{z+1}))| \leq 2\varepsilon$$
$$|p(y, (\vec{\sigma}_{-i}, \sigma_i^{y+1})) - p(y, (\vec{\sigma}_{-i,j}, \sigma_i^{y+1}, \sigma_j^{z+1}))| \leq 2\varepsilon$$

Therefore the difference in effect

$$|e_i^y(f, \vec{\sigma}^{(\ell)}) - e_i^y(f, \vec{\sigma}^{(\ell+1)})| \leq 4\varepsilon.$$

In case (2), when $x \notin \{w, z\}$, by first switching the positions of x, y under i's preference ordering and then applying the above argument, we bound the difference in the effect of x on the revised profiles, that is $|e_i^x(f, (\vec{\sigma}_{-i}^{(\ell)}, \sigma_i^{y+1})) - e_i^x(f, (\vec{\sigma}_{-i}^{(\ell+1)}, \sigma_i^{y+1}))| \leq 4\varepsilon$. Now we observe that if V is the set of outcomes ranked above x, y in σ_i then both σ_i and σ_i^{y+1} rank the elements of V in the first $|V|$ positions and the elements of $V \cup \{x, y\}$ in the first $|V| + 2$ positions, therefore since f is ε-localized (by Lemma A.3), the probability of returning an element of $\{x, y\}$ differs by at most 2ε. We can therefore bound the desired quantity

$$|e_i^y(f, \vec{\sigma}^{(\ell)}) - e_i^y(f, (\vec{\sigma}^{(\ell+1)}| \leq 8\varepsilon.$$

Combining these two cases, we observe that the difference in effect in any single step is at most 8ε. Therefore the total difference in effect

$$|e_i^y(f,\vec{\sigma}) - e_i^y(f,\vec{\sigma}')| \leq 4nk^2\varepsilon.$$

□

Definition A.8. f is ε-decomposable if for all voters i and all outcomes $x \neq y$, there exist functions γ, δ such that for all preference profiles $\vec{\sigma}$ where σ_i ranks x directly above y, $|e_i^y(f,\vec{\sigma}) - \gamma(\vec{\sigma}|_{\{x,y\}}) - \delta(\sigma_i)| \leq \varepsilon$.

Lemma A.9. *If $f : (\Sigma_k)^n \to [k]$ is ε-strategy proof voting rule where $k \geq 2$ then it is $16nk^3\varepsilon$-decomposable.*

Proof. Fix $i \in [n]$ and $x \neq y \in [k]$. Let $\vec{\sigma}^{\{x,y\}\downarrow}$ be the preference profile constructed from $\vec{\sigma}$ by moving x, y to the last two positions in every component σ_i (but maintaining their relative order).

Let ID be the preference profile in which all voters rank the outcomes in numerical order, that is $ID_j = (1, \ldots, k)$ for all $j \in [n]$.. Let g^* be a function that maps restricted orderings of the form $\vec{\sigma}|_{\{x,y\}}$ to full preference orderings in the following deterministic way: the jth component of the profile $g^*(\vec{\sigma}|_{\{x,y\}})$ is $ID^{\{x,y\}\downarrow}$ if the relative ordering of x, y according to σ_j agrees with their numerical ordering and else the jth component is $ID^{\{x,y\}\downarrow,z+1}$ where $z \in \{x, y\}$ is the element ranked first according to the restricted ordering $\sigma_j|_{\{x,y\}}$. Observe that g^* ranks $\{x, y\}$ are ranked in the last two positions according to all voters j, and that $\vec{\sigma}|_{\{x,y\}} = g^*(\vec{\sigma}|_{\{x,y\}})|_{\{x,y\}}$.

Define component functions γ, δ as follows:

$$\gamma(\vec{\sigma}|_{\{x,y\}}) = e_i^y(f, g^*(\vec{\sigma}|_{\{x,y\}}))$$
$$\delta(\sigma_i) = e_i^y(f,(ID_{-i},\sigma_i)) - \gamma((ID_{-i},\sigma_i)|_{\{x,y\}})$$

We now claim that if $\vec{\sigma}$ is a preference profile where σ_i ranks x directly above y then $\gamma(\vec{\sigma}|_{\{x,y\}}) + \delta(\sigma_i)$ is close to $e_i^y(f,\vec{\sigma})$. Specifically we will show the equivalent statement that $e_i^y(f,(ID_{-i},\sigma_i)) - e_i^y(f,\vec{\sigma})$ is close to $e_i^y(f, g^*((ID_{-i},\sigma_i)|_{\{x,y\}}) - e_i^y(f, g^*(\vec{\sigma}|_{\{x,y\}}))$ (this is obtained from the previous claim by substituting in the definitions of γ and δ and then rearranging the terms). We will prove the rearranged statement by giving an explicit bound on the former quantity.

We begin by observing that $e_i^y(f,(ID_{-i},\sigma_i))$ is within error $4nk^2\varepsilon$ of $e_i^y(f,(ID_{-i}^{\{x,y\}\downarrow},\sigma_i))$ and likewise $e_i^y(f,\vec{\sigma})$ is within $4nk^2\varepsilon$ of $e_i^y(f,\vec{\sigma}_{-i}^{\{x,y\}\downarrow},\sigma_i)$ (since by Lemma A.7 f is approximately pairwise responsive). It thus suffices to bound the quantity that $e_i^y(f,(ID_{-i}^{\{x,y\}\downarrow},\sigma_i)) - e_i^y(f,\vec{\sigma}_{-i}^{\{x,y\}\downarrow},\sigma_i)$ which can be rewritten as

$$\Big[p(y,(ID_{-i}^{\{x,y\}\downarrow},\sigma_i^{y+1})) - p(y,(ID_{-i}^{\{x,y\}\downarrow},\sigma_i))\Big] - \Big[p(y,(\vec{\sigma}_{-i}^{\{x,y\}\downarrow},\sigma_i^{y+1})) - p(y,(\vec{\sigma}_{-i}^{\{x,y\}\downarrow},\sigma_i))\Big]$$

We now construct σ_i^{y+1} from σ_i via the following sequence of steps: (a) Progressively switch y down to the last position, (b) Progressively switch x down to just above y. (c) Switch

the positions of x and y, (d) Progressively switch y back up to its original position, and (e) Progressively switch x up to just below y.

Steps of type (a) and (d), of which there are at most $2k$, consist of switching y with various alternatives $z \notin \{x, y\}$. Since $ID_{-i}^{\{x,y\}\downarrow}$ and $\vec{\sigma}_{-i}^{\{x,y\}\downarrow}$ agree on the relative ordering of y, z and since f is $4nk^2\varepsilon$-pairwise isolated, the difference in the probability of returning y between the two cases is within an error of $4nk^2\varepsilon$. (Observe that this analysis applies to both steps of type (d) and steps of type (a) since the change in probabilities due to moving y down is just the negative change in probabilities due to moving y up.)

Steps of type (b) and (e) consist of switching x with various alternatives $z \notin \{x, y\}$. Since f is 2ε responsive, this changes each case by at most 2ε. Therefore each step of type (b) or (e) (of which there are at most $2k$) introduces an error of at most 4ε.

Finally we observe that in the unique step of type (c) we are moving y from the last position (where it is ranked immediately below x) up one position, therefore the effect of this step is exactly $e_i^y(f, (ID_{-i}^{\{x,y\}\downarrow}, \sigma_i^{\{x,y\}\downarrow})) - e_i^y(f, \vec{\sigma}^{\{x,y\}\downarrow})$ which, since f $4nk^2\varepsilon$-pairwise isolated, is within $4nk^2\varepsilon$ of $e_i^y(f, g^*(\vec{\sigma}^*)_{-i}, \sigma_i^{\{x,y)\downarrow}) - e_i^y(f, g^*(\vec{\sigma}|_{\{x,y\}})_{-i}, \sigma_i^{\{x,y\}\downarrow})$. Since f is ε-localized, this in turn is within 8ε of $e_i^y(f, g^*((\vec{\sigma}_{-i}^*, \sigma_i)|_{\{x,y\}}) - e_i^y(f, g^*(\vec{\sigma}|_{\{x,y\}}))$.

Combining all of these steps, we can successfully bound the desired quantity:

$$\begin{aligned}\left|\left[e_i^y(f, (\vec{\sigma}_{-i}^*, \sigma_i)) - e_i^y(f, \vec{\sigma})\right] - \left[e_i^y(f, g^*((\vec{\sigma}_{-i}^*, \sigma_i)|_{\{x,y\}})) - e_i^y(f, g^*(\vec{\sigma}|_{\{x,y\}}))\right]\right| \\ \leq 4nk^2\varepsilon + 4nk^2\varepsilon + 2k \cdot 4nk^2\varepsilon + 2k \cdot 2\varepsilon + 4nk^2\varepsilon + 8\varepsilon \\ \leq 16nk^3\varepsilon\end{aligned}$$

□

We now use the above Lemmas to prove an extension of Gibbard's theorem characterizing strategy-proof voting rules. The proof employs the concept of a *pseudorule*. A pseudorule is a mapping $r : [k] \times \Sigma_k^n rightarrow \mathbb{R}^{\geq 0}$ that takes an outcome and a preference profile and returns a non-negative value. A pseudorule r induces a voting rule f_r defined by $\Pr[f_r(\vec{\sigma}) = x] \propto r(x, \vec{\sigma})$. All of the above properties can be applies to pseudorules as well as voting rules: we say that a pseudorule r is localized (resp. pairwise responsive, pairwise isolated, decomposable) if the induced voting rule f_r is localized (resp. pairwise responsive, pairwise isolated, decomposable). We observe briefly that using the notation defined above, f is the voting rule induced by p.

We now proceed to prove Theorem 5.1.

Theorem A.10. *If f is ε-strategy proof then it can be expressed as a probability distribution over voting rules such that with probability at least $1 - 72n^2k^4\varepsilon$, the selected rule is either unilateral or binary.*

Proof. We define a collection of pseudorules as follows:

- $p(x, \vec{\sigma}) = \Pr[f(\vec{\sigma}) = x]$
- $p_0(x, \vec{\sigma}) = \Pr[f(\vec{\sigma}^{x\downarrow}) = x]$
- $p_i(x, \vec{\sigma}) = \min_{\vec{\sigma}'}\{p(x, (\sigma'_{-i}, \sigma_i)) - p(x, (\vec{\sigma}'_{-i}, \sigma_i^{x\downarrow}))\}$

- $c(x, \vec{\sigma}) = p(x, \vec{\sigma}) - \sum_{i=0}^{n} p_i(x, \vec{\sigma})$
- $c_{yz}(x, \vec{\sigma}) \begin{cases} c(x, \vec{\sigma}^{\{y,z\}\downarrow}) & \text{if } x \in \{y, z\} \\ 0 & \text{else} \end{cases}$

The proof proceeds via a sequence of four claims:

Claim A.11. Voting rules f_{p_i} are unilateral for all $i \in [n]$.

Proof of Claim A.11 For $i \in [n]$ by construction p_i depends only on σ_i and not on $\vec{\sigma}_{-i}$, therefore the induced voting rules are unilateral.

Claim A.12. Voting rules $f_{c_{yz}}$ are duple for all $y \neq z \in [k]$.

Proof of Claim A.12 This follows immediately by construction. For any value $x \notin \{y, z\}$, $f_{c_{yz}}$ never returns v, so it is duple.

Claim A.13. There exists a unilateral voting rule g such that with probability at least $1 - nk\varepsilon$ over the internal randomness, the voting rule $f_{p_0} = g$

Proof of Claim A.13 Let $\vec{\sigma}^*$ be a fixed preference profile and define $g(\vec{\sigma}) = f_{p_0}(\vec{\sigma}^*)$. Observe that g is constant, that is it ignores its input $\vec{\sigma}$, and therefore it is unilateral.

For all outcomes $x \in [k]$, $|p_0(x, \vec{\sigma}) - \Pr[g(\vec{\sigma}) = x]| \leq n\varepsilon$ (this follows immediately from Lemma A.1, since we are considering the difference in probability between two outcomes in $\vec{\sigma}^{x\downarrow}$ and $\vec{\sigma}^{*x\downarrow}$ in which all voters rank x last. Therefore the pseudorule p_0 is equal to g with probability at least $1 - nk\varepsilon$.

Claim A.14. $|c(x, \vec{\sigma}) - \sum_{yz} c_{yz}(x, \vec{\sigma})| \leq 72n^2k^4\varepsilon$

Proof of Claim A.14 The proof proceeds by induction on the position of x. As a base case, we show that $c(x, \vec{\sigma}^{x\downarrow}) = \sum_{y \neq z} c_{yz}(x, \vec{\sigma}^{x\downarrow})$. We first claim that since everyone ranks x last in $\vec{\sigma}^{x\downarrow}$, $c(x, \vec{\sigma}^{x\downarrow}) = 0$. This follows immediately from the definition of c. We then observe that if $x \notin \{y, z\}$ then $c_{yz}(x, \vec{\sigma}^{x\downarrow}) = 0$ and if $x \in \{y, z\}$ then x is uniformly at the bottom in $\vec{\sigma}^{x\downarrow,\{y,z\}\downarrow}$ so as before $c_{yz}(x, \vec{\sigma}^{x\downarrow}) = c(z, \vec{\sigma}^{x\downarrow,\{y,z\}\downarrow}) = 0$.

Before moving on to the inductive step, we will show two technical claims that will be used to prove the inductive step:

Sub-Claim A.15. $|e_i^y(p_i, \vec{\sigma}) - \min_{\vec{\tau}} e_i^y(p, (\vec{\tau}_{-i}, \sigma_i))| \leq 4nk^3\varepsilon$

Define $\alpha = p_i(y, \vec{sigma})$ and $\beta = p_i(y, (\vec{\sigma}_{-i}, \sigma_i^{y+1}))$. Define $\eta = \min_{\vec{\tau}} e_i^y(f, (\vec{\tau}_{-i}, \sigma_i))$. Observe that we are trying to bound the difference between β and $\alpha + \eta$.

To show one side of the bound, observe that $\beta = \min_{\sigma'}\{p(y, (\sigma'_{-i}, \sigma_i^{y+1})) - p(y, (\vec{\sigma}'_{-i}, \sigma_i^{y\downarrow}))\}$. Fix $\vec{\sigma}'$ to be a preference profile that minimizes that quantity, that is $\beta = p(y, (\sigma'_{-i}, \sigma_i^{y+1})) - p(y, (\vec{\sigma}'_{-i}, \sigma_i^{y\downarrow}))$. Observe further that $\alpha \leq p(y, (\vec{\sigma}'_{-i}, \sigma_i)) - p(y, (\vec{\sigma}'_{-i}, \sigma_i^{y\downarrow}))$ and $\eta \leq p(y, (\vec{\sigma}'_{-i}, \sigma_i^{y+1})) - p(y, (\vec{\sigma}'_{-i}, \sigma_i))$ and therefore it immediately follows that $\beta \geq \alpha + \eta$.

To show the other side of the bound, we define a particular preference profile $\vec{\tau}$ as follows: Let $\vec{\sigma}^*$ be the profile such that $\sigma_i^* = \sigma_i$ and $\alpha = p(y, \vec{\sigma}^*) - p(y, (\vec{\sigma}^*_{-i}, \sigma_i^{y\downarrow}))$. Let $\vec{\sigma}'$ be the preference profile such that $\sigma'_i = \sigma_i$ and $\eta = e_i^y(f, \vec{\sigma}')$. Define $\vec{\tau}$ from $\vec{\sigma}^*$ by moving x

(in the ranking of each voter $j \neq i$) to just above or just below y (where the relative ordering of x, y in τ_j is defined to match the relative ordering of x, y in σ'_j). Observe that $\alpha = p(y, (\vec{\sigma}^*_{-i}, \sigma_i)) - p(y, (\vec{\sigma}^*_{-i}, \sigma_i^{y\downarrow}))$ is equal to the sum of the effects, in the context of σ^*, of i's successively moving y from the bottom to just below x in σ_i. Since y is never switched with x σ^* and τ agree on the relative ordering of the two outcomes being switched by i in each iteration. Therefore since f is $4nk^2\varepsilon$-pairwise isolated, $\alpha \geq p_i(y, \vec{\tau}) - (k-1) \cdot 4nk^2\varepsilon$. Observe further that since σ', τ agree on the relative ordering of y an since f is $4nk^2\varepsilon$-pairwise isolated, $\eta \geq e_i^y(p, t\vec{a}u) - 4nk^2\varepsilon$. Therefore $\beta \leq \alpha + \eta + 4nk^3\varepsilon$. This concludes the proof of Claim A.15.

We now proceed to show that c approximately ignores external comparisons, that is for all $i \in [n]$ and all $x, y \in [k]$ if $\vec{\sigma}, \vec{\sigma}'$ are preference profiles such that σ_i, σ'_i rank x directly above y and $\vec{\sigma}|_{\{x,y\}} = \vec{\sigma}'|_{\{x,y\}}$ then $e_i^y(c, \vec{\sigma}) \approx e_i^y(c, \vec{\sigma}')$.

Sub-Claim A.16. The pseudorule c $72nk^3\varepsilon$-ignores external comparisons.

Let $\vec{\sigma}, \vec{\sigma}'$ be two preference profiles such that σ_i ranks x directly above y and $\vec{\sigma}|_{\{x,y\}} = \vec{\sigma}'_{\{x,y\}}$. Let δ, γ be the functions whose existence is guaranteed by the fact that f is $16nk^3\varepsilon$-decomposable.

Observe that $e_i^y(c, \vec{\sigma}) = e_i^y(p, \vec{\sigma}) - e_i^y(p_i, \vec{\sigma})$ since all other terms cancel. Therefore $|e_i^y(c, \vec{\sigma}) - e_i^y(c, \vec{\sigma}')| = |e_i^y(p, \vec{\sigma}) - e_i^y(p_i, \vec{\sigma}) - e_i^y(p, \vec{\sigma}') + e_i^y(p_i, \vec{\sigma}')|$. Let $\vec{\tau}^* = \text{argmin}_{\vec{\tau} \text{ s.t. } \tau_i = \sigma_i} e_i^y(p, \vec{\tau})$ and let $\vec{\tau}' = \text{argmin}_{\vec{\tau} \text{ s.t. } \tau_i = \sigma'_i} e_i^y(p, \vec{\tau})$. The by Sub-Claim A.15, $|e_i^y(c, \vec{\sigma}) - e_i^y(c, \vec{\sigma}')| \leq |e_i^y(p, \vec{\sigma}) - e_i^y(p, \vec{\tau}^*) - e_i^y(p, \vec{\sigma}') + e_i^y(p, \vec{\tau}')| + 2 \cdot 4nk^3\varepsilon$. Then since f is $16nk^3\varepsilon$-decomposable, $|e_i^y(c, \vec{\sigma}) - e_i^y(c, \vec{\sigma}')| \leq |\gamma(\vec{\sigma}|_{\{x,y\}}) - \gamma(\vec{\tau}^*|_{\{x,y\}}) - \gamma(\vec{\sigma}'|_{\{x,y\}}) + \gamma(\vec{\tau}'|_{\{x,y\}})| + 4 \cdot 16nk^3\varepsilon + 8nk^3\varepsilon$ (observe that since $\sigma_i = \tau_i^*$ and $\sigma'_i = \tau'_i$, the δ terms cancel. Moreover, since $\vec{\sigma}|_{\{x,y\}} = \vec{\sigma}'|_{\{x,y\}}$, $|e_i^y(c, \vec{\sigma}) - e_i^y(c, \vec{\sigma}')| \leq |\gamma(\vec{\tau}'|_{\{x,y\}}) - \gamma(\vec{\tau}^*|_{\{x,y\}})| + 72nk^3\varepsilon$.

Finally, we observe that since $\vec{\tau}^*$ minimizes $\gamma(\vec{\tau}^*|_{\{x,y\}})$ for $\tau_i^* = \sigma_i$ and $\vec{\tau}'$ minimizes $\gamma(\vec{\tau}'|_{\{x,y\}})$ for $\tau'_i = \sigma'_i$ and since $\tau_i^*|_{\{x,y\}} = \tau'_i|_{\{x,y\}}$, the fact that γ depends only on the relative ordering of x, y implies that $\gamma(\vec{\tau}^*|_{\{x,y\}}) = \gamma(\vec{\tau}'|_{\{x,y\}})$. Therefore $|e_i^y(c, \vec{\sigma}) - e_i^y(c, \vec{\sigma}')| \leq 72nk^3\varepsilon$.

We can now proceed with the inductive step of our argument: let $\vec{\sigma}^{(0)}, \ldots, \vec{\sigma}^{(\mu)}$ be a sequence such that $\vec{\sigma}^{(\mu)} = \vec{\sigma}, \vec{\sigma}^{(0)} = \vec{\sigma}^{x\downarrow}$ and $\vec{\sigma}^{(\ell)}, \vec{\sigma}^{(\ell+1)}$ differ only in that a single voter i moves x up one position, that is $\vec{\sigma}^{(\ell+1)} = (\vec{\sigma}^{(\ell)}_{-i}, \sigma_i^{(\ell)x+1})$ for some $i \in [n]$. From our base case, we know that $c(x, \vec{\sigma}^{(0)}) = \sum_{y \neq z} c_{yz}(x, \vec{\sigma}^{(0)})$. We now proceed by induction on ℓ.

Let i_ℓ be the unique voter such that $\sigma_{i_\ell}^{(\ell)} \neq \sigma_{i_\ell}(\ell+1)$ and let w_ℓ be the outcome ranked directly above x in $\sigma_{i_\ell}^{(\ell)}$. Observe that $c_{w_\ell x}(x, \vec{\sigma}^{(\ell+1)}) - c_{w_\ell x}(x, \vec{\sigma}^{(\ell)}) = c(x, \vec{\sigma}^{(\ell+1), \{w_\ell, x\}\downarrow}) - c(x, \vec{\sigma}^{(\ell), \{w_\ell, x\}\downarrow})$ since $x \in \{w_\ell, x\}$. From how we constructed the sequence, this is $e_{i_\ell}^x(c, \vec{\sigma}^{(\ell), \{w_\ell, x\}\downarrow})$. Since c $72nk^3\varepsilon$-ignores external comparisons, this is within error $72nk^3\varepsilon$ of $c_{i_\ell}^x(c, \vec{\sigma}^{(\ell)}) = c(x, \vec{\sigma}^{(\ell+1)}) - c(x, \vec{\sigma}^{(\ell)})$. We now claim that for all other pairs $\{y, z\} \neq \{w_\ell, x\}$, $c_{yz}(x, \vec{\sigma}^{(\ell+1)}) = c_{yz}(x, \vec{\sigma}^{(\ell)}$. This follows immediately from the observation that in this case, $\vec{\sigma}^{(\ell), \{y,z\}\downarrow} = \vec{\sigma}^{(\ell+1), \{y,z\}\downarrow}$. Therefore $e_i^x(c, \vec{\sigma}^{(\ell)})$ is within error $72nk^3\varepsilon$ of $\sum_{yz} c_{yz}(x, \vec{\sigma}^{(\ell)})$.

Therefore by induction, for all preference profiles $\vec{\sigma}$ and all outcomes $x \in [k]$, $|c(x, \vec{\sigma}) - \sum_{yz} c_{yz}(x, \vec{\sigma})| \leq n \cdot (k-1) \cdot 72nk^3\varepsilon$.

Combining Claims A.11, A.12, A.13, and A.12 with the definition of c we can therefore observe that the voting rule f can be expressed as a probability distribution over voting rules where with probability at least $1 - 72n^2k^4\varepsilon$, the chosen component voting rule is either unilateral or duple. □

B Relaxed Approximations

In this section, we consider a relaxed notion of approximation, called a (δ, μ)-*approximation*, under which we require only that the approximation g return an outcome that is close to the correct outcome *with high probability*. We explore how this relaxed definition can be leveraged to establish a trade-off between the proximity of "good answers" to the true outcome and the probability of returning a good answer. In particular, we demonstrate two distinct constructions of such relaxed approximations: the first construction extends the linear approach employed in Theorem 4.1, the second construction uses an exponential approach inspired by differential privacy and provides improved guarantees for certain values of δ. We also develop lower bounds that shown the asymptotic optimality of our constructions.

B.1 Defining (δ, μ)-Approximations

We begin by formalizing this weaker notion of approximation.

Definition B.1 ((δ, μ)-approximation)**.** A (randomized) voting rule g is a (δ, μ)-approximation of a voting rule f if for all inputs $\vec{\sigma}$,

$$\Pr[d_v((\vec{\sigma}, g(\vec{\sigma})), (\vec{\sigma}, f(\vec{\sigma}))) > \delta] \leq \mu.$$

Observe that a $(\delta, 0)$-approximation is equivalent to a δ-approximation as defined in Section 2.

B.2 Constructing Relaxed Approximations

By relaxing our definition, we inherently facilitate the construction of randomized ε-strategy-proof approximations for any deterministic function. In this section, we show two such constructions. In Theorem B.2 we extend the linear construction from Theorem 4.1 to quantify the μ-values received for smaller values of δ. In Theorem B.3 we provide an alternative exponential construction that provides improved guarantees for small values of δ.

Theorem B.2. *For any deterministic voting rule $f : (\Sigma_k)^n \to [k]$ and any $\varepsilon > 0, \delta \geq 0$, f has a ε-strategy-proof (δ, μ)-approximation g, where $\mu = 0$ if $\varepsilon(\delta + 1)/k(k + 1 + \varepsilon) \geq 1$ and $\mu = (k - 1)(1 - (\delta + 1)\varepsilon/k(k + 1 + \varepsilon))/(1 + (k - 1)(1 - (\delta + 1)\varepsilon/k(k + 1 + \varepsilon)))$ else.*

Proof. The construction is identical to that employed in the proof of Theorem 4.1: we assign each input-output pair $(\vec{\sigma}, j)$ a quality score $q(\vec{\sigma}, j) = -d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, j))$. As before, let $\xi = \varepsilon/k(k + 1 + \varepsilon)$—this value is chosen to guarantee ε-strategy-proofness. The mechanism g returns the value j with probability proportional to $\max\{1 + \xi q(\vec{\sigma}, j), 0\}$.

The proof that the resulting mechanism is ε-strategy-proof is identical to that employed in the proof of Theorem 4.1. It is only necessary to demonstrate the quality of approximation achieved, using the relaxed notion of (δ, μ)-approximations.

As before, let $M(\vec{\sigma}, \iota) = |\{j \in [k] : q(\vec{\sigma}, j) = \iota\}|$, that is $M(\vec{\sigma}, \iota)$ is the number of outputs with quality score $q(\vec{\sigma}, j) = \iota$. We again observe that for all profiles $\vec{\sigma}$,

$$\Pr[d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, g(\vec{\sigma}))) > \delta] = \frac{\sum_{\iota=-\delta-1}^{-n} \max\{1 + \xi\iota, 0\} M(\vec{\sigma}, \iota)}{\sum_{\iota=0}^{-n} \max\{1 + \xi\iota, 0\} M(\vec{\sigma}, \iota)}$$

We now wish to bound this probability for smaller values of δ, so we express the denominator as the sum of two terms $\sum_{\iota=-\delta-1}^{-n} \max\{1 + \xi q(\vec{\sigma}, \iota), 0\} M(\vec{\sigma}, \iota) + \sum_{\iota=0}^{-\delta-1} \max\{1 + \xi q(\vec{\sigma}, \iota), 0\} M(\vec{\sigma}, \iota)$ which is minimized (thereby maximizing the total fraction) when the second term (the "good" outputs) consists only of the one correct answer.

$$\Pr[d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, g(\vec{\sigma}))) > \delta] \leq \frac{\sum_{\iota=-\delta-1}^{-n} \max\{1 + \xi\iota, 0\} M(\vec{\sigma}, \iota)}{\sum_{\iota=-\delta-1}^{n} \max\{1 + \xi\iota, 0\} M(\vec{\sigma}, \iota) + 1}$$

Finally, it is clear that the right hand side of the inequality is maximized when the numerator is maximized, and this occurs when there are $k - 1$ outputs with quality $q(\vec{\sigma}, j) = -\delta - 1$. Therefore

$$\forall \vec{\sigma}, \Pr[d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, g(\vec{\sigma}))) > \delta] \leq \frac{(k-1) \max\{1 + \xi(-\delta - 1), 0\}}{1 + (k-1) \max\{1 + \xi(-\delta - 1), 0\}}$$

The result follows. □

Observe that Theorem B.2 is a simple extension of Theorem 4.1 that quantifies the guarantees achieved for smaller values of δ, that is for $\delta < k(k + 1 + \varepsilon)/\varepsilon - 1$. However, for such small values, this linear construction is not necessarily optimal. We instead present an alternative method for constructing (δ, μ)-approximations in which $\mu > 0$ but which offers improved guarantees for small values of δ.

Theorem B.3. *For any deterministic voting rule $f : (\Sigma_k)^n \to [k]$ and any $\varepsilon > 0, \delta \geq 0$, f has a ε-strategy-proof (δ, μ)-approximation g, where $\mu = (k-1)/((\varepsilon + 1)^{(\delta+1)/2} + k - 1)$.*

Proof. We employ the exponential mechanism of Talwar and McSherry [10] to construct an approximation g of the voting rule f as follows: we assign each input-output pair $(\vec{\sigma}, j)$ a quality score $q(\vec{\sigma}, j) = n - d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, j))$. Observe that $q(\vec{\sigma}, j)$ decreases linearly with the (minimal) number of votes that must be corrupted before f returns j instead of $f(\vec{\sigma})$. Let $\xi = (1/2) \ln(\varepsilon + 1)$—this value is chosen to guarantee ε-strategy-proofness as discussed in Lemma **??**. The mechanism g returns the value j with probability proportional to $\exp(\xi q(\vec{\sigma}, j))$.

First, we claim that the resulting mechanism is ε-strategy-proof. From [10] we know that this mechanism gives 2ξ-differential privacy. We reproduce the proof for the sake of completeness: The probability that an output j is chosen is given by $(\exp(\xi q(\vec{\sigma}, j))/(\sum_{\iota \in [k]} \exp(\xi q(\vec{\sigma}, \iota)))$. A single change in the input profile $\vec{\sigma}$ can, by definition, change the quality score by at most 1, giving a factor of at most $\exp(\xi)$ in the numerator and at least $\exp(-\xi)$ in the dominator,

yielding a total change in probability of at most $\exp(2\xi)$. It follows by Lemma **??** that the approximation g is ε-strategy-proof.

Second, we claim that g is a good approximation for f according to the distance metric d_v. We note that Talwar and McSherry present a general accuracy bound for the exponential mechanism [10], however, that bound is too loose for our purposes.[1] Instead, we take advantage of the particular distance pseudometric d_v to develop a tighter bound. Let $M(\vec{\sigma}, \iota) = |\{j \in [k] : q(\vec{\sigma}, j) = \iota\}|$, that is $M(\vec{\sigma}, \iota)$ is the number of outputs with quality score $q(\vec{\sigma}, j) = \iota$. We observe that for all profiles $\vec{\sigma}$,

$$\begin{aligned}\Pr[d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, g(\vec{\sigma}))) > \delta] &= \frac{\sum_{j:d_v((\vec{\sigma},f(\vec{\sigma})),(\vec{\sigma},g(\vec{\sigma})))>\delta} \exp(\xi q(\vec{\sigma}, g(\vec{\sigma})))}{\sum_{j\in[k]} \exp(\xi q(\vec{\sigma}, g(\vec{\sigma})))} \\ &= \frac{\sum_{\iota=0}^{n-\delta-1} \exp(\xi\iota) M(\vec{\sigma}, \iota)}{\sum_{\iota=0}^{n} \exp(\xi\iota) M(\vec{\sigma}, \iota)}\end{aligned}$$

That is, the probability that g returns an output greater than δ from the true output is equal to sum over such outputs j of the probability that j is chosen divided by the equivalent sum over all outputs. Exploiting our metric d_v, this is then re-indexed over the set of possible quality scores.

We express the denominator as the sum of two terms $\sum_{\iota=0}^{n-\delta-1} \exp(\xi\iota) M(\vec{\sigma}, \iota) + \sum_{\iota=n-\delta}^{n} \exp(\xi\iota) M(\vec{\sigma}, \iota)$ which is minimized (thereby maximizing the total fraction) when the second term (the "good" outputs) consists only of the one correct answer.

$$\Pr[d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, g(\vec{\sigma}))) > \delta] \leq \frac{\sum_{\iota=0}^{n-\delta-1} \exp(\xi\iota) M(\vec{\sigma}, \iota)}{\sum_{\iota=0}^{n-\delta-1} \exp(\xi\iota) M(\vec{\sigma}, \iota) + \exp(\xi n)}$$

Finally, it is clear that the right hand side of the inequality is maximized when the numerator is maximized, and this occurs when there are $k-1$ outputs with quality $q(\vec{\sigma}, j) = n - \delta - 1$. Therefore

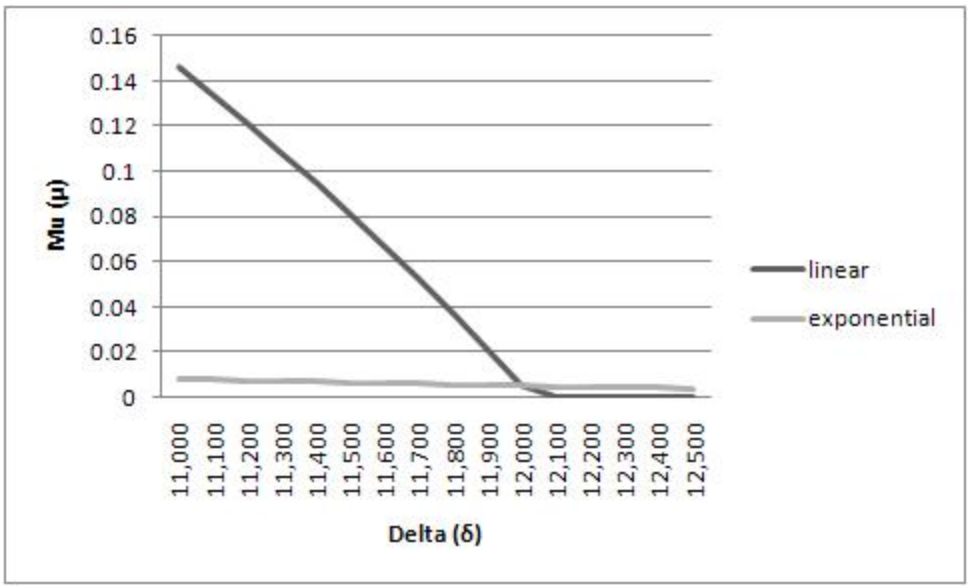
$$\begin{aligned}\forall \vec{\sigma}, \Pr[d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, g(\vec{\sigma}))) > \delta] &\leq \frac{(k-1)\exp(\xi(n-\delta-1))}{\exp(\xi n) + (k-1)\exp(\xi(n-\delta-1))} \\ &= \frac{(k-1)}{(\varepsilon+1)^{(\delta+1)/2} + (k-1)}\end{aligned}$$ □

As before, we can express the quality of the exponential-approximations in terms of their asymptotic behavior.

Corollary B.4. *Let $\varepsilon = \omega(1/n)$, $\beta, \mu > 0$, and let $f : (\Sigma_k)^n \to [k]$ be a deterministic voting rule. For sufficiently large n, there exists an ε-strategy-proof randomized voting rule g that is a $(\beta n, \mu)$-approximation of f.*

Proof. Fix $\beta > 0$ and let $\delta = \beta n$. Let g be defined as in the proof of Theorem 4.1 and recall that (as shown in the previous proof) g is an ε-strategy-proof $(\beta n, \frac{(k-1)}{(\varepsilon+1)^{(\beta n+1)/2}+(k-1)})$-

[1] In particular, for small values of δ (e.g., $\delta < 178$ when $\varepsilon = .05$ and $k = 3$) the original bound makes the trivial statement that $\Pr[d_v((\vec{\sigma}, f(\vec{\sigma})), (\vec{\sigma}, g(\vec{\sigma}))) > \delta] \leq 1$.



approximation of f. Observe that

$$\lim_{n\to\infty} \frac{(k-1)}{(\varepsilon+1)^{(\beta n+1)/2}+(k-1)} = \lim_{n\to\infty} \frac{(k-1)}{((\varepsilon+1)^{1/\varepsilon})^{\varepsilon(\beta n+1)/2}+(k-1)} \tag{1}$$

$$= 0 \tag{2}$$

The first step is simple arithmetic. (2) follows by the following argument, which demonstrates that the denominator goes to infinity as n grows: If $\varepsilon = \Omega(1)$ then by definition $\varepsilon + 1$ is bounded below by some constant, so it suffices that $(\beta n + 1)/2$ goes to infinity, which is obvious. In the case where $\varepsilon = o(1)$, observe that since $\varepsilon = \omega(1/n)$, $\lim_{n\to\infty} \varepsilon(\beta n+1)/2 = \infty$. It therefore suffices to show that $(\varepsilon+1)^{1/\varepsilon}$ is bounded below by some constant. Since $\varepsilon = o(1)$, $\lim_{n\to\infty}(\varepsilon+1)^{1/\varepsilon} = \lim_{\varepsilon\to 0}(\varepsilon+1)^{1/\varepsilon} = e$.

It follows that for any β, μ, for sufficiently large n there exists an ε-strategy-proof $(\beta n, \mu)$-approximation. □

Example B.5. For concreteness, consider the case where you have one hundred players and three outputs. Set $\varepsilon = .005$. The probability that g returns an incorrect answer is at most .67 (that is, any voting scheme $f : (\Sigma_3)^{100} \to [3]$ has a .05-strategy-proof $(0, .67)$-approximation).

Alternatively, consider an election with 100 million voters and three outputs (about the scale of a United States presidential election). Again fixing $\varepsilon = .005$, the probability that g returns an answer further than 5000 votes from the correct answer (in practice, well within the vote corruption in such an election) is at most .00001 (i.e., any voting rule $f : (\Sigma_3)^{100,000,000} \to [3]$ has a .005-strategy-proof $(5000, .00001)$-approximation). Looked at in another way, this says that in any election with 100 million voters and three outputs in which one outcome wins by at least 5000 votes (e.g., a typical national election), this mechanism will return the correct answer with probability at least .99999.

The relative guarantees achieved by these two constructions are shown in the Figure B.2; numbers are calculated with $\varepsilon = .001$, $n = 100,000,000$, and $k = 3$.

B.3 Extended Lower Bounds

In this section, we justify our original definition of approximation by showing two lower bounds that strictly limit the utility of alternative definitions; First, we extend the lower bound from Theorem 5.4 to show that our relaxed definition does not admit asymptotically better approximations. In particular, when $\varepsilon = o(1/n)$, we show that PLURALITY cannot be well-approximated even for $\mu > 0$. We then consider an alternative definition that has been considered previously in the context of 0-strategy-proofness: the requirement that the approximation g be equal to the original function with high probability (in our notation, approximations with $\delta = 0$). We extend the work of Gibbard and Satterthwaite to show that only dictatorial voting rules have ε-strategy-proof approximations with $\delta = 0$.

We first bound the asymptotic behavior of the relaxed approximations introduced in Appendix B.

Theorem B.6. *If $\varepsilon = o(1/n^2)$ then there exists $\beta > 0$ such that for all n,* PLURALITY *does not have a trivial $(\beta n, 1 - 2/\sqrt{k})$-approximation.*

Proof. This extended result follows immediately from the proof of Theorem 5.4 by setting $c = \sqrt{k}$. □