Approximately Strategy-Proof Voting

Eleanor Birrell[*] and **Rafael Pass**[†]
Cornell University
{eleanor,rafael}@cs.cornell.edu

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 asymptotically tight lower bounds on the parameters required by such voting rules.

1 Introduction

The classic Gibbard-Satterthwaite Theorem [Gibbard, 1973; Satterthwaite, 1975] 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 [1977] and later advocated by Conitzer and Sandholm [2006], 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 [2010] 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 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 [1980] 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. [1989], 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 [Conitzer and Sandholm, 2006; Friedgut *et al.*, 2009; Isaksson *et al.*, 2010]; 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 [Halpern and Pass, 2010]. 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

[*]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.

utility by more than ε.

Unfortunately, a corollary of the Gibbard-Satterthwaite theorem shows that the only deterministic voting rules 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 the full version of this paper, 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.

$\varepsilon = o(1/n)$: In this setting, we show that the Gibbard's result characterizing strategy-proof randomized voting rules [Gibbard, 1977] still applies:

Theorem 1.2 (Lower Bound – Informal Statement)**.** *If g is a $o(1/n)$-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 βn-approximation of* PLURALITY.

In fact, in the full version 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. Note that Theorem 1.3 also implies that the approximations constructed in Theorem 1.1 are non-trivial.

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.

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 results 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. More formally,

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 [Kahn *et al.*, 1988] 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 $\vec{\sigma}$, an alterative preference σ_i' and an outcome $j \in [k]$ such that $|\Pr[g(\vec{\sigma}) = j] - \Pr[g(\sigma_i', \vec{\sigma}_{-i}) = j]| > \xi$.

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 $k\xi$-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 $(t, t^2 k\xi)$-strategy-proof.*

The proofs of Lemmas 2.4 and 2.5 are omitted due to space constraints.

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 [Procaccia, 2010], 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 the full version, 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, 1977])**.** 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, 1977])**.** 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 [2010] 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 [1973] and Satterthwaite [1975], 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 [Friedgut *et al.*, 2009; Isaksson *et al.*, 2010]. 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 [Gibbard, 1977]. 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 [McSherry and Talwar, 2007] that has been used to establish differential privacy [Dwork *et al.*, 2006].

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 ε/k. 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 4.2 we know that there exist arbitrarily good approximations of f for $\varepsilon = \omega(1/n)$. By Theorem 5.5 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 only trivial voting rules can be ε-strategy-proof for small values of ε, and we show that such voting rules cannot provide good approximations of natural voting rules like PLURALITY.

We begin by showing that for small values of ε, only trivial voting rules are ε-strategy-proof. The general outline of the proof is as follows: we define a reduction between ε-strategy-proof voting rules and 0-strategy-proof voting rules by including punishments for players who misreport their preferences. Consider the following modified mechanism: with probability $1 - p$ we run the original mechanism and with probability p we choose a single player i and output his first choice (according to his *reported* preferences). This has the effect that for appropriate values of ε, p, no one wants to misreport their first choice. However, they can still successfully deviate by making more subtle changes further down their preference profile. We overcome this by modifying our behavior slightly: when we choose a player i, we return each of his choices with probability inversely proportional to their ranking in his reported preferences. For small values of ε, this style of punishment is sufficient to offset any benefits yielded by a deviating strategy under the original mechanism. The result then follows from Theorem 3.2. More formally:

Theorem 5.1. *Let $\varepsilon < 1/nk^3$. A randomized voting rule $g : (\Sigma_k)^n \to [k]$ is ε-strategy-proof if and only if it is trivial.*

Proof. Let u_i be uniformly distributed for all $i \in [n]$ — that is the utility assigned to the outcome ranked jth is $(k - j)/(k - 1)$. Define a new randomized voting rule g' that for each $i \in [n]$, returns i's jth choice (according to i's reported preference) with probability $k\varepsilon(k - j)$ and otherwise follows the original mechanism g. That is, we use one of the unilateral "punishment" mechanisms with probability $p = n \sum_{j \in [k]} k\varepsilon(k - j) < nk^3\varepsilon$ and we use the original mechanism with probability $1 - p$. Observe that both of these probabilities are in the range [0,1] because $\varepsilon < 1/nk^3$.

Let $\vec{\sigma}, \vec{\sigma}'$ be preference profiles that differ only in the ith component and let $\sigma_{i,j}$ denote the outcome ranked jth by σ_i. We begin by claiming that the expected loss of utility from deviating (due to the unilateral punishment mechanisms) is at least ε. This will offset any utility gained under the original mechanism g.

Claim 5.2. $\sum_j k\varepsilon(k - j)(u_i(\vec{\sigma}, \sigma_{i,j}) - u_i(\vec{\sigma}, \sigma'_{i,j})) \geq \varepsilon$.

We first show how to prove Theorem 5.1 assuming this claim, and then we prove the claim itself.

$$
\begin{aligned}
&E[u_i(\vec{\sigma}, g'(\vec{\sigma}))] - E[u_i(\vec{\sigma}, g'(\vec{\sigma}'))] \\
&\quad = \sum_j k\varepsilon(k - j)(u_i(\vec{\sigma}, \sigma_{i,j}) - u_i(\vec{\sigma}, \sigma'_{i,j})) \\
&\qquad + (1 - p)(E[u_i(\vec{\sigma}, g(\vec{\sigma}))] - E[u_i(\vec{\sigma}, g(\vec{\sigma}'))]) && (1) \\
&\quad \geq \varepsilon + (1 - p)(E[u_i(\vec{\sigma}, g(\vec{\sigma}))] - E[u_i(\vec{\sigma}, g(\vec{\sigma}'))]) && (2) \\
&\quad \geq \varepsilon + (1 - p)(-\varepsilon) && (3) \\
&\quad \geq 0 && (4)
\end{aligned}
$$

(1) follows by the definition of expectation, (2) follows immediately from Claim 5.2, (3) follows from the fact that the original mechanism g is ε-strategy proof, and (4) follows from the

fact that $p \in [0,1]$. Since g' is 0-strategy-proof, by Theorem 3.2 it must be trivial. Since g' follows mechanism g with probability $1-p > 0$, it follows that g is also trivial.

Proof of Claim 5.2: We begin by considering the simplified case where σ_i and σ_i' are identical except that the positions of two outcomes j_1, j_2 are swapped. Assume without loss of generality that $\sigma_i(j_1) > \sigma_i(j_2)$. In this case, most of the k unilateral punishment rules cancel (because the preferences only differ in two positions). Let $r_\sigma(j)$ indicate the rank of j according to preference σ. The expected difference in utility contributed by the punishment mechanism is

$$(k\varepsilon(k - r_{\sigma_i}(j_1))u_i(j_1) + k\varepsilon(k - r_{\sigma_i}(j_2))u_i(j_2)) \\ - \left(k\varepsilon(k - r_{\sigma_i'}(j_1))u_i(j_1) + k\varepsilon(k - r_{\sigma_i'}(j_2))u_i(j_2)\right)$$

Since the difference between σ_i, σ_i' is that the two ranks are switched—$r_{\sigma_i'}(j_1) = r_{\sigma_i}(j_2)$ and $r_{\sigma_i'}(j_2) = r_{\sigma_i}(j_1)$—the expected difference in utility can be written as

$$k\varepsilon\left((k - r_{\sigma_i}(j_1)) - (k - r_{\sigma_i}(j_1))\right)(u_i(j_1) - u_i(j_2))$$

The difference in rank between j_1 and j_2 under preference σ_i is at least 1. Since u_i is uniformly distributed, the difference in utility u_i between j_1, j_2 is at least $1/k$, therefore $\sum_j k\varepsilon(k-j)(u(\vec{\sigma}, \sigma_{i,j}) - u(\vec{\sigma}, \sigma'_{i,j})) \geq k\varepsilon \cdot 1 \cdot (1/k) = \varepsilon$.

We now consider the general case where σ_i and σ_i' differ arbitrarily. Define a sequence $\vec{\sigma}' = \vec{\sigma}^{(1)}, \vec{\sigma}^{(2)}, \ldots, \vec{\sigma}^{(m)} = \vec{\sigma}$ iteratively as follows: given $\vec{\sigma}^{(\iota)}$, let $\sigma_\ell^{(\iota+1)} = \sigma_\ell^{(\iota)}$ for all $\ell \neq i$. Define the preferences $\sigma_i^{(\iota)}$ iteratively using bubble sort [Friend, 1956]: any two adjacent preferences differ only by flipping two adjacent outcomes, and the two adjacent outcomes are only flipped if they are in the wrong relative order (according to the final preference ordering σ_i). The fact that this sequence is well defined (and terminates with $\vec{\sigma}^{(m)} = \vec{\sigma}$) follows immediately from the correctness of bubble sort. The proof that the difference in voter i's expected utility between two adjacent hybrid preferences is at least ε follows by the above argument since only two outcomes are flipped and only if they start in the wrong relative order. Claim 5.2 then follows by a hybrid argument. □

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

Example 5.3. Define a class of voting rules SUBSET-δ : $(\Sigma_k)^n \to [k]$ 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 5.4. 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.

Theorem 5.5. *Let $n \geq k$, let $\beta > 0$. For sufficiently large n,* PLURALITY *does not have a trivial βn-approximation.*

The proof closely follows that of Procaccia [2010] and is omitted due to space constraints. Observe that this result implies that for the voting rule PLURALITY, the approximation constructed in Theorem 4.1 is asymptotically optimal.

References

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

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

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

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

[Friend, 1956] Edward H. Friend. Sorting on electronic computer systems. *J. ACM*, 3:134–168, July 1956.

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

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

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

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

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

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

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

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

[Satterthwaite, 1975] 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.