# Interim Self-Stable Decision Rules

Daeyoung Jeong* Semin Kim†

May 23, 2018

**Abstract**

This study identifies a set of interim self-stable decision rules. In our model, individual voters encounter two separate decisions sequentially: (1) a decision on the change of a voting rule they are going to use later and (2) a decision on the final voting outcome under the voting rule which has been decided from the prior procedure. A given decision rule is self-stable if any other possible rule does not get enough votes to replace the given rule under the given rule itself. We fully characterize the set of interim self-stable decision rules among qualified majority rules. We also characterize the set of interim self-stable constitution among weighted majority rules.

**JEL Classification: C72, D02, D72, D82.**

**Keywords: Weighted majority rules, decision rules, self-stability**

*Economic Research Institute, The Bank of Korea; daeyoung.jeong@gmail.com
†School of Economics, Yonsei University, South Korea; seminkim@yonsei.ac.kr

# Contents

# 1 Introduction

Since different voting rules may result in different voting outcomes, the welfare of an individual voter may depend on a voting rule they are using as well as her preference over possible voting outcomes. A voter with a particular preference over possible voting outcomes may also form preferences over different voting rules. So, when the change on voting rule is possible in a certain way, one individual may want to change the voting rule according to her interest, but another may not. A rule change in a society is, therefore, closely tied up with the individuals' preferences over voting outcomes and voting rules in the society.

This study identifies a set of interim self-stable decision rules. In our model, individual voters encounter two separate decisions sequentially: (1) a decision on the change of a voting rule they are going to use later and (2) a ordinary decision on the final economic outcome under the voting rule which has been decided from the prior procedure. A given decision rule is self-stable if any other possible rule does not get enough votes to replace the given rule under the given rule itself. Unlike the previous studies on self-stability (Barberà and Jackson, 2004; Azrieli and Kim, 2016), which assume that the decision on the change of a voting rule takes place before the individuals' preferences over possible outcomes have been realized, we assume that individuals' preferences have been realized even before making any decision.[1] So, we can demonstrate the direct effect of realized preferences of individuals on the self-stability of decision rules.

On many occasions, voters are aware of their own preferences over economic outcomes even before they decide on a voting rule to use. For example, a legislature, who are aware of the characteristics of an upcoming bill on which they are going to vote, may have a chance to change their decision rule beforehand. In this situation, the voting body may make a decision on the rule change strategically based on their preferences. This strategical component at the "interim" stage makes the analysis more complicated but provide interesting implications.

[1] In this sense, our study is an obvious generalization of Holmström and Myerson (1983), which assumes individuals with realized preferences over possible voting outcomes make decisions on changing the decision rule not under the given decision rule, but only under unanimity.

This study helps us to explore the implications by comparing the set of interim self-stable decision rules with the set of ex-ante self stable decision rules, which has been identified by previous studies. A voting rule may be interim self-stable, but not ex-ante, or possibly vice versa. So, the stability of a particular voting rule may depend on the timing of the rule change; before(ex-ante) or after(ex-post) the preference realization. The comparison, therefore, allows us to examine the impact of the timing of the rule change on the stability of decision rules, and also the feasible set of economic outcomes.

In analysis, by following Barberà and Jackson (2004), we first start with the simplest possible case where the society uses the same voting rule for the rule change at the first stage and the ordinary decision at the second stage. We show that, in the set of the set of qualified majority rules, where all voters have the same voting power, the simple and super majority rule is interim self-stable. Note that this result does not depend on the environmental features of the society, such as the number of voters and the prior distribution of the voters' preferences.

We then move on to more general form of constitutions, which consists of a pair of voting rules, one of which is for the rule change and the other is for the ordinary decision. We show that a constitution is interim self-stable if and only if 1) the voting rule for the ordinary decision is not stronger than the voting rule for the change of the rule, and 2) the combining toughness of those two rules should be higher then a certain level.[2] Intuitively, the first condition prevents the liberal voters from forming a coalition to change the given rule to a weaker voting rule: If there are enough number of liberal voters so that they can change the given rule to a weaker one, they don't have to change it since the given rule is weak enough for them to achieve the liberal outcome. On the other hand, the second condition prevents the conservative voters from forming a coalition to change the given rule to a stronger voting rule: If there are enough number of conservative voters so that they can change the given rule to a stronger rule, they don't have to change the rule since the given rule is strong enough to prevent the others from implementing the liberal outcome.

[2]Similar to the previous result, this one also does not depend on the societal environment.

We generalize our analysis further by considering a set of weighted majority rules where individual voters may have different voting powers. Notwithstanding the asymmetry of voters under a general weighted majority rule makes the analysis considerably complicated, we can still obtain meaningful conditions that characterize the set of self-stable weighted majority rules.

## 1.1 Related Literature

The two papers, Barberà and Jackson (2004) and Holmström and Myerson (1983) motivate this project. Barberà and Jackson (2004) introduce the ex-ante self-stability of voting rules and focus on the qualified majority rules. Unlike them, we define the interim self-stability of voting rules and study not only the qualified majority rules but also general voting rules. The interim self-stability is similar to the durability of decision rules defined by Holmström and Myerson (1983) in that an agent utilizes the preferences information in the interim stage. While they use the unanimous rule to choose between rules, we start with the given rule itself and try to extend the argument with the various rules. It can show the effects of those variations on the set of stable rules.

In our model, agents' preferences over voting rules are endogenously determined from their assessments regarding their preferences over alternatives. Such a model was first suggested in early papers by Rae (1969), Badger (1972), and Curtis (1972). While these papers only consider anonymous voting rules with the same weight to all agents, we study weighted majority rules which allow the heterogenous weights for agents.

The seminal book of Neumann and Morgenstern (1953, Section 5) theoretically investigates weighted majority rules. The main interest of the book is the measures of the voting power of agents under the rule. A common scenario leading to heterogeneous voting weights is that of a representative democracy with heterogenous district sizes. An early paper on this topic is Penrose (1946). Recently, Barberà and Jackson (2006) and Fleurbaey (2008) point out the advantage of weighted majority rules from a utilitarian point of view. Also, Azrieli and Kim (2014) show that, in a standard mechanism design setup, weighted majority rules naturally arise from considerations of efficiency and incen-

tive compatibility. We investigate another property, the stability of weighted majority rules.

The idea that the same voting rule used to choose between alternatives is also used to choose between voting rules can be found in the social choice literature. Koray (2000) introduces the concept of self-selection for social choice functions. See also Barberà and Beviá (2002) and Koray and Slinko (2008).

# 2 Definitions

## 2.1 Environment

A society faces a binary decision whether to implement the Reform ($R$) or to keep the Status-quo ($S$), so the set of alternatives is $A = \{R, S\}$. In the society, there are $n \geq 2$ agents (voters), $N = \{1, 2, ..., n\}$. Each agent can either prefer $R$ or $S$, which indicates the type of the agent, $t_i \in T_i = \{r, s\}$. The probability of agent $i$ being a type $t_i$ is $p_i(t_i)$ and $p_i(t_i = r) + p_i(t_i = s) = 1$. We assume that there is no agent who is indifferent between $R$ and $S$, $p_i(r) \neq p_i(s)$, and that $p_i(t_i) > 0$ for any $t_i \in T_i$. Let $T = T_1 \times \cdots \times T_n$ be the set of type profiles. We assume that types are independent across agents, so we write $P(t) = \prod_{i \in N} p_i(t_i)$ for the probability of a type profile $t \in T$. For the technical convenience, we abuse the notation, $P(t_{-i}) = \frac{p(t)}{p_i(t_i)}$ for the probability of a type profile of other agents excluding agent $i$.

An agent's utility depends on the chosen alternative and on his own type, $u_i : A \times T_i \to \mathbb{R}$. We normalize the utility such that $u(R, r) = a$, $u(R, s) = -1$, and $u(S, r) = u(S, s) = 0$. Thus a society can be characterized by the pair $(p_r, a)$, where $p_r = (p_1(r), ..., p_n(r))$.

## 2.2 Voting Rules and Constitutions

A strategy for an agent $i$ is a function $v_i : T_i \to V_i = \{0, 1\}$ that associates her own type $t_i$ to a voting decision. A voting rule is any mapping $f : V \to [0, 1]$, with the interpretation that, $f(v)$ is the probability that the change or the reform $R$ is chosen when the voting profile of agents is $v = (v_i)_{i \in N} \in V = \prod_{i \in N} V_i$. We mainly focus on weighted majority

rules which can be formally defined as follows.

**Definition 1** (Weighted Majority Rule)**.**

The voting rule $f$ is a *Weighted Majority Rule* if there are non-negative weights $w^f = \left(w_1^f, ..., w_n^f\right)$ and a quota $0 \leq q^f < \sum_{i \in N} w_i^f$ such that

$$f(v) = \begin{cases} 1 & if \quad \sum_{\{i:v_i=1\}} w_i^f > q^f \\ 0 & if \quad \sum_{\{i:v_i=1\}} w_i^f \leq q^f. \end{cases}$$

We denote a weighted majority rule $f$ by $f = \left(w^f, q^f\right)$. We also denote the set of weighted majority rules by $\mathbf{G}$ and the set of weighted majority rules with a certain weights w by $\mathbf{G}(\mathrm{w})$.

We define a constitution as a pair of weighted majority rules.

**Definition 2** (Constitution)**.**

A *Constitution* is a pair of decision rules $(f, F)$ where $F$ is for the rule change and $f$ is for the final outcome $S$ or $R$.

To economize on notation, for $F = (w^F, q^F)$, we drop the superscripts, so write $F = (w, q)$.

## 2.3 Definition of Interim Self-stability

We now define the concept of interim self-stability of a constitution $(f, F)$ with a two-stage voting game, $\Gamma$. Timing of the game $\Gamma$ is as follows. In the first stage, individual voters observe their own type $t_i$. Then under a rule $F = (w, q)$, which is for the decision on the rule change, agents play a simultaneous voting game whether to keep the incumbent rule $f$ or to choose the alternative rule $g$: The rules $f$ and $g$ are for the decision on the final outcome $S$ or $R$. The alternative rule $g$ would be implemented if $\sum w_i > q$, where the sum is taken over all voters who vote for $g$, and $f$ would be maintained otherwise. In the second stage, agents make a decision on $A = \{R, S\}$ by the rule chosen in the first stage, either $f$ or $g$. To restrict our focus on the decision on rule changes, we assume that voters act sincerely in the second stage: $r$-type votes for $R$ and $s$-type for $S$.

Roughly, we would like to say a constitution $(f, F)$ is interim self-stable, if, for any alternative voting rule $g$, there is a Nash equilibrium of a voting game in which the alternative $g$ is defeated by $f$ for any type profile $t \in T$ when the decision is made by the rule $F$. The rest of this section formally defines the interim self-stability.

Let $\sigma_i(t_i)$ be the probability that individual $i$ would vote for $g$ in the first stage when her type is $t_i$.[3] To reject the alternative rule $g$ all the time, the probability that $g$ gets sufficient support should be zero for all $t \in T$. In other words, the alternative $g$ is always rejected if and only if

$$\sum_{\{j:\sigma_j(t_j)>0\}} w_j \leq q, \quad \forall t \in T. \tag{C.1}$$

If Condition (C.1) holds, then the voting strategies in the first stage, $\sigma = (\sigma_1, \ldots, \sigma_n)$, together with honest behavior under $f$ and $g$, form a Nash equilibrium if and only if

$$\sum_{t_{-i}} P\left(t_{-i}\right) \gamma_i(t_{-i}) \left(u_i(f(t), t_i) - u_i(g(t), t_i)\right) \geq 0 \quad \forall i, \ \ \forall t_i \in T_i, \tag{C.2}$$

where

$$\Phi_i = \{H_i \subseteq N \setminus \{i\} |\, q - w_i < \sum_{j \in H_i} w_j \leq q\},$$

and

$$\gamma_i(t_{-i}) = \sum_{H_i \in \Phi_i} \left( \prod_{j \in H_i} \sigma_j(t_j) \right) \left( \prod_{j \in N/(H_i \cup \{i\})} (1 - \sigma_j(t_j)) \right).$$

Voter $i$ is pivotal if the voters in $H_i \in \Phi_i$ vote for the alternative rule $g$ and all others $j \notin H_i$ vote for the given rule $f$.[4] We can interpret that $\gamma_i(t_{-i})$ is the voter $i$'s probability of being pivotal given the strategy profile $\sigma$ and the others' type profile $t_{-i}$. Therefore, Condition (C.2) implies that either voter $i$ is never pivotal, or she is weakly better off

[3] Conceptually, a mixed strategy $\sigma_i$ is a probability distribution over the set of pure strategy $V_i$.

[4] Also denote $\Psi^f$ the set of minimal winning coalitions under a decision rule $f$. Note that, when $f$ is a qualified majority rule, $\Phi_i = \{C \setminus \{i\} : C \in \Psi^f \text{ and } i \in C\}$.

under $f$ than $g$.

Note that, in a simultaneous voting game, there generally exists a trivial Nash equilibrium in which no agent votes for the alternative $g$, unless an agent has the dictatorial power under $F$. In such an equilibrium, where the condition (C.1) and (C.2) are satisfied, any alternative rule $g$ is defeated by the given rule $f$. Therefore, in order to define a reasonable concept of interim self-stability, we need to refine the equilibria of the game $\Gamma$. We require a type of sequential rationality for agents' voting strategy profile $\sigma$ given that they share a consistent 'posterior' belief. To be consistent with Holmström and Myerson (1983), we assume that agents may have some apprehensions for being pivotal coincidentally due to others' mistakes in voting.

We first characterize a posterior distribution given that individual $i$ is pivotal as follows.[5]

$$\mu_i(t_{-i}) = \lim_{k\to\infty} \frac{P\left(t_{-i}\right) \sum_{H_i \in \Phi_i} \rho(H_i; t_{-i}, \sigma^k)}{\sum_{\hat{t}_{-i} \in T_{-i}} P\left(\hat{t}_{-i}\right) \sum_{H_i \in \Phi_i} \rho(H_i; \hat{t}_{-i}, \sigma^k)} \tag{C.3}$$

$$\forall i, \forall t_i \in T_i, \forall t_{-i} \in T_{-i},$$

where

$$\rho(H_i; t_{-i}, \sigma^k) = \left(\prod_{j \in H_i} \sigma_j^k(t_j)\right) \left(\prod_{j \in N \setminus (H_i \cup \{i\})} (1 - \sigma_j^k(t_j))\right)$$

$$\sigma_j^k(t_j) > 0 \quad \forall k, \forall j, \forall t_j \in T_j$$

$$\sigma_j(t_j) = \lim_{k\to\infty} \sigma_j^k(t_j) \quad \forall j, \forall t_j \in T_j$$

So, agent $i$ believe that, when she is coincidentally pivotal, the others' type profile is $t_{-i}$ with the probability $\mu_i(t_{-i})$ given their mistakes $\sigma_j^k$.[6] Given this distribution or belief,

[5] The posterior distribution $\mu_i(t_{-i})$ is not exactly the posterior beliefs in a concept of sequential equilibrium. However, an agent $i$'s decision on the changing rules is only relevant when she is pivotal. Hence, we characterize an agent's posterior distribution given that she is pivotal, and then require a rational behavior at the first stage given the posterior distribution.

[6] Since the limit of denominator of Condition (C.3), which represents $i$'s probability of being pivotal given $\sigma$, could be zero, we characterize the distribution in the style of the trembling hand model.

we require that, for any type $t_i$ of any individual $i$,

$$\begin{aligned} &\text{if } \sigma_i(t_i) = 0, \\ &\text{then } \sum_{t_{-i}} \mu_i(t_{-i}) u_i(f(t), t_i) \geq \sum_{t_{-i}} \mu_i(t_{-i}) u_i(g(t), t_i). \end{aligned} \tag{C.4}$$

Condition (C.4) imposes that, conditional on that the agent $i$ is pivotal, if she is expected to be better off under the alternative rule $g$ than under the rule $f$, then she should vote for $g$ with positive probability.[7]

With the conditions above, we define the concepts of equilibrium rejection and endurance.

**Definition 3** (Equilibrium rejection)**.**
Consider a constitution $(f, F)$. A strategy profile and a belief $(\sigma, \mu)$ consists an equilibrium rejection of $g$ under $F$ if and only if the conditions (C.1) through (C.4) are all satisfied.

**Definition 4** (Endurance)**.**
Consider a constitution $(f, F)$. The voting rule $f$ endures an alternative $g$ under $F$ if and only if there exists some equilibrium rejection of $g$ under $F$.

Now, in a set of voting rules $\hat{G}$, we formally define the interim self-stability of a constitution $(f, F)$ by using the above definitions.

**Definition 5** (Interim Self-stability)**.**
Consider a constitution $(f, F)$. The constitution $(f, F)$ is interim self-stable in $\hat{\mathbf{G}}$ if and only if $f$ endures every alternative rule $g \in \hat{\mathbf{G}}$ under $F$.

# 3 Self-stability of "One Person, One vote"

In this section, as in the previous studies (Holmström and Myerson, 1983; Barberà and Jackson, 2004), we focus on anonymous weighted majority rules, which treats each voter

[7] This condition may prevent an individual who strongly prefers the alternative $g$ from voting against it just because she is never be pivotal given the others' voting strategy.

identically.[8] An anonymous weighted majority rule (or an anonymous voting rule) can be represented by the special type of weighted majority rules where all voters have the same voting power $w = (1, ..., 1)$ and $q \in \{0, 1, ..., n-1\}$. They are classified according to the quota: a simple majority rule ($q = q^s \equiv \frac{n}{2}$ if $n$ is even and $\frac{n-1}{2}$ if $n$ is odd), a sub majority rule ($q < q^s$), and a super majority rule ($q > q^s$). We denote by $\mathbf{G}(\mathrm{I})$ the set of anonymous weighted majority rules where $w = \mathrm{I} \equiv (1, ..., 1)$.

## 3.1 Self-stable Voting Rules

First, we focus on a special type of constitution where the society uses the same voting rule on the decisions of the rule change and the final outcome, so $F = f$. The following proposition characterizes the set of interim self-stable anonymous voting rule. It argues that a voting rule is interim self-stable if and only if it is not a sub majority rule.

**Proposition 1** (Interim Self-stable Voting Rules)**.**
*For any given environment $(p_r, a)$, a voting rule $f \in \mathbf{G}(\mathrm{I})$ is interim self-stable in $\mathbf{G}(\mathrm{I})$ if and only if it is a simple or super majority rule.*

The complete proof of this proposition is in the appendix.[9] The basic intuition is as follows. Suppose the given rule $f$ is a simple or super majority rule. Denote by $N_s(t) = \{i \in N | t_i = s \text{ for a given } t \in T\}$ the set of s-type agents given a type profile $t$. Similarly, define $N_r(t)$.[10] An $s$-type may support a change to a more conservative alternative rule with a higher quota, $g$ with $q^g > q^f$. But if there are enough $s$-type voters to accomplish the change, $|N_s(t)| > q^f$, the final outcome under the given rule $f$ would already be $S$, $|N_r(t)| = n - |N_s(t)| < n - q^f \leq q^f + 1$. So, $s$-types do not have to vote for the stronger alternative. Similarly, an $r$-type may support a change to a less conservative alternative rule with a lower quota, $g$ with $q^g < q^f$. But if there are enough $r$-type voters to make the change, $|N_r(t)| > q^f$, the final outcome under the given rule $f$ would already be $R$. So, $r$-types do not have to vote for the less conservative

[8] See May (1952) for the formal definition of an anonymous voting rule.
[9] This proposition is a corollary of Proposition 2.
[10] By construction, $|N_s(t)| + |N_r(t)| = |N| = n$.

alternative.[11] Therefore, under a simple or super majority rule, when a group of agent can make a change in a certain way, so they form a winning coalition, they don't need to since they can obtain their desirable outcome under the given rule.

On the other hand, if the given rule $f$ is a sub-majority, $s$-type agents in a minimal winning coalition may be worried about the possible reform $R$ supported from another exclusive minimal winning coalition. So, the $s$-types would vote for a change to a more conservative alternative. Here, under a sub-majority rule, even a group of agent can make a change in a certain way, so they form a winning coalition, they may not always get their way especially when they prefer the status-quo $S$. So they would vote for a more conservative rule. Thus, the sub-majority rule $f$ cannot be interim self-stable.

Note that the result in Proposition 1 is environment independent. That is, for any society with the characterization $(p_r, a)$ and the number of voters $n$, a simple and super majority is interim self-stable. In the characterization of Barberà and Jackson (2004) or Azrieli and Kim (2016), on the other hand, the ex-ante self-stability depends on the environment: A rule which is ex-ante self-stable in one environment could not be ex-ante self-stable in another environment. They induce the agents' preferences based on the ex-ante expected utilities calculated with the ex-ante probabilities $p_r$.[12] By construction, therefore, the decision on the rule change would highly depend on the environment. However, in our concept, one's main concern at the interim stage is not the ex-ante probability $p_r$ itself, but the (perceived) posterior belief $\mu_i(t_{-i})$, which depends on the strategy profile of the other agents and the given voting rule.[13] As a result, the agents' voting decisions would naturally be strategically interdependent. This strategic voting behavior allows us to explore the agents' incentive in strategic situations under any majority rules from sub

[11] For the super-majority rules, Barberà and Jackson (2004) provide a similar interpretation to explain why they focus on the ex-ante concept. We further provide an interpretation for the sub-majority rules.

[12] Conceptually, they suppose that an agent believes she is always pivotal in the first stage voting game, and, given each voting rule, calculate the agent's ex-ante expected utility. For a given environment, the induced preferences based on the calculated expected utilities satisfy certain properties such as single-peak, single-crossing, and so on. (See Lemma 1 and 2 in Barberà and Jackson (2004)) Consequentially, agents' induced preferences over voting rules highly depends on the environment of a society.

[13] Since agents make decisions at interim, it is natural that the voting strategies are contingent to the agents' types. An agent's (perceived) pivotal probability would highly depend on the other agents' behavior. Hence, an agent makes a decision focusing mainly on the posterior believes induced by the other agents' voting strategies, rather on the ex-ante probability $p$.

to super majority, and under any general constitution later in this section.

Roughly speaking, Barberà and Jackson (2004) show that a simple majority (or similar) rule is one of the few ex-ante self-stable voting rules.[14] (See Theorem 1 and 2 in Barberà and Jackson (2004).) The set of interim self-stable voting rules specified completely in Proposition 1, which contains super-majority rules, is generically bigger than the set of the ex-ante self-stable voting rules.[15] In reality, we can easily observe super-majority voting rules in use: the unanimity rule under jury conviction systems or super-majority rules under legislatures. Normative reasons aside, one could argue that the rule change schemes in reality fit better with the concept of the interim self-stability, but another may argue that it is not the matter of the timing of the rule change, but the matter of the different environment. The current study provides a theoretical tool to examine how the rule change mechanism affects on the societal values such as the stability of voting rules.

## 3.2 Self-stable Constitutions

Now, we consider a general constitution $(f, F)$ where different voting rules are used for the different decisions. (See Definition 2 for the formal definition.) Proposition 2 presents the complete characterization of the interim self-stable constitution. Remind that $f = (w^f, q^f)$ and $F = (w, q)$.

**Proposition 2** (Interim Self-stable Constitutions)**.**
*A constitution* $(f, F) \in \mathbf{G}(\text{I}) \times \mathbf{G}(\text{I})$ *is interim self-stable in* $\mathbf{G}(\text{I})$ *if and only if*

1. $q \geq q^f$ *and*

2. $q + 1 \geq n - q^f$.

[14] Because the concept highly depends on environments, it is hard to specify the necessity condition of the ex-ante self-stable voting rule.

[15] With this result, one may argue that a society is more stable with the interim rule change since more decision rules can be stably used without any change than with the ex-ante rule change. However, another may say that a society is more conservative with the interim rule change because of the similar reason. So, we cannot argue assertively that which rule change scheme is better. To examine it, we may need to priorly discuss the societal cost of the rule change and/or the direct benefit of the "stability."

The intuition here is similar to that for the previous case with a single voting rule. While there exists only one condition $q \geq \frac{n-1}{2}$ for the previous case, we have two conditions for this case with a constitution; $q \geq q^f$ and $q+1 \geq n-q^f$. The first condition prevents $r$-types from forming a winning coalition for an equilibrium rejection, and the second condition prevents $s$-types from doing that. More specifically, the former says the size of minimal winning coalition of $F$, $q$, is not less than the size of minimal winning coalition of $f$, $q^f$. If there are enough number of $r$-type voters so that they can change the given rule to a weaker one they all prefer, $|N_r(t)| > q$, they don't have to change it since the given rule is weak enough for them to get the Reform in the second stage voting game, $|N_r(t)| > q^f$. The latter condition says the size of minimal winning coalition of $F$, $q+1$, is not less than the size of minimal veto coalition of $f$, $n-q^f$. If there are enough number of $s$-type voters so that they can change the given rule to a stronger rule they all prefer, $|N_s(t)| > q$, they don't have to change the rule since they can veto any attempt to reform under the given rule $f$, $|N_s(t)| > n-q^f$.

Those conditions are consistent with observations from the reality. We usually observe a constitution where $F$ is relatively tougher than $f$. One society may start with a voting rule $f$ which is tougher than $F$, but it may be replaced with an alternative $g$ which is weaker than $F$ under some circumstances. Moreover, we have few real world examples with a long-lived constitution consisting of two weak voting rules, such as a constitution with two sub-majority rules with $q+q^f < n-1$.

The following corollaries demonstrate interesting characteristics and examples of interim self-stable constitutions.

**Corollary 2.1.** *A constitution $(f,F)$ is not interim self-stable if $q < \frac{n-1}{2}$.*

Corollary 2.1 states that, if $F$ is a sub-majority rule, the constitution is not interim self-stable. Thus, to be stably used, a constitution should be consisted with a simple or super majority rule for the decision on the rule change.

**Corollary 2.2.** *Any constitution $(f,F)$ is interim self-stable if $F$ is the unanimity rule.*

Corollary 2.2 states that, if $F$ is a unanimity, a constitution with any $f$ is interim self-stable. Thus, any constitution could be stably used if the rule change requires unanimous

supports.

**Corollary 2.3.** *If a constitution $(f, F)$ with $q$ is interim self-stable, a constitution $(f, F')$ with $q' > q$ is also interim self-stable.*

Corollary 2.3 implies that the stronger the voting rule for the rule change is, the bigger the set of interim self-stable constitutions is.

## 3.3 General Alternatives

In this subsection, we check if the voting rules and/or constitutions characterized in Proposition 1 or 2 are still self-stable against any general weighted majority rules. It is hard to find the real world example where an anonymous voting rule or constitution was replaced by a weighted majority rule with different voting powers. One may argue that it is because of the social norm: The social norm may require that a voting rule or constitution treats all voters equally. Here, in a positive analysis, we examine if certain anonymous rules can endure any weighted majority rules even without such a social norm.

**Proposition 3** (Self-stable Qualified Majority Rules)**.**
*A constitution $(f, F) \in \mathbf{G}(\mathrm{I}) \times \mathbf{G}(\mathrm{I})$ is interim self-stable in $\mathbf{G}$ if*

1. $q \geq q^f$ *and*
2. $q + 1 \geq n - q^f$.

This proposition explains why it is hard for one society to move from an "one-person, one-vote" decision rule to another with asymmetric voting powers: Simple or super majority rules can be stably survived in the long run, even though the society considers any general voting rules as alternatives. The proposition may also imply why it is important to start with the self-stable anonymous rules in the virtue of fairness or equality. If a society starts with a self-stable anonymous decision rule, it may not need any strong normative arguments to keep the fairness or equality in voting powers. The self-stable anonymous decision rule will defend themselves against any "unfair" alternatives.

# 4 Extensions

In this section, we relax the anonymous constraints for the given constitution $(f, F)$ as well as the alternative $g$, so that different agents could have different voting powers under a decision rule. Even though not as common as anonymous voting rules, non-anonymous weighted majority rules where agents have different weights can be found in reality. A stockholder meeting would be one example: stockholders' weights are determined by the amounts of the stocks they possess. Another example is a legislature with a veto player: In a presidential system, the president may have the veto power, so has the power to refuse to approve a bill.

We conduct two separate analyses based on the set of alternatives. Firstly, we examine the interim self-stability in a set of weighted majority rules with fixed weight. So, in this setting, the rules vary only in quota. The shareholders' meeting would have a better fit with this setting: The shareholders' voting powers, in general, have been determined prior to any meeting, and may not changed by the result of it. Second, we further generalize the set of alternatives, and consider the changes in weights. A presidential system may fit better with this situation. Technically, the legislature can pass the bill to amend the constitution from a presidential system with a veto player to a cabinet system without her.

## 4.1 Fixed-weight Constitutions

Here, we restrict our focus on the set of weighted majority rules under which agents' weights are fixed: For any $f$ and $g$, $(w_i^f)_{i=1}^n = (w_i^g)_{i=1}^n = \mathrm{w}$. That is, given "fixed" weights (or voting powers), the rules vary only in quota. Denote $\mathbf{G}(\mathrm{w})$ the set of weighted majority rules with given weights $\mathrm{w} = (w_i)_{i=1}^n$.

The following proposition specifies a sufficient condition of interim self-stable fixed-weight constitutions.

**Proposition 4** (Fixed-weight Constitution: Sufficient condition)**.**
*A constitution* $(f, F) \in \mathbf{G}(\mathrm{w}) \times \mathbf{G}(\mathrm{w})$ *is interim self-stable in* $\mathbf{G}(\mathrm{w})$ *if*

1. $q \geq q^f$ *and*

2. $\forall i$, $\exists \hat{C} \ni i$ *such that* $w(N \setminus \hat{C}) \leq q^f$.

Proposition 4 shares a similar implication with Proposition 2. The first condition prevents $r$-types from forming a winning coalition for an equilibrium rejection, and the second condition prevents $s$-types from doing that. The former says the size of minimal winning coalition of $F$ is not less than the size of minimal winning coalition of $f$, $q^f$. If there are enough number of $r$-type voters so that they can change the given rule to a weaker one they all may prefer, $|N_r(t)| > q$, they don't have to change it since the given rule is weak enough for them to get the Reform, $|N_r(t)| > q^f$. The latter says if any individual under the given rule is a member of a minimal winning coalition of $F$ which is a veto coalition of $f$ at the same time, then the constitution $(f, F)$ is interim self-stable. If there are enough number of $s$-type voters so that they can change the given rule to a stronger rule they all prefer, $w(N_s(t)) > q$, one may believe they don't have to change the rule since they are strong enough $N_s(t) \supseteq \hat{C}$ so that they can veto any attempt to reform under the given rule $f$, $q^f \geq w(N \setminus \hat{C}) \geq w(N \setminus N_s(t))$.

So far, we discuss a sufficient condition of interim self-stable constitutions with fixed weights. The following proposition describes a necessary condition which is also strong enough to cover the cases with the anonymous constraint discussed in Section 3. We denote by $\bar{C}$ the minimal winning coalition which contains agents with highest weights under $F$.

**Proposition 5** (Fixed-weight Constitution: Necessary condition)**.**
*A constitution* $(f, F) \in \mathbf{G}(\mathrm{w}) \times \mathbf{G}(\mathrm{w})$ *is interim self-stable in* $\mathbf{G}(\mathrm{w})$ *only if*

1. $q \geq q^f$ *and*

2. *for some* $i \in \bar{C}$, $\exists \hat{C} \ni i$ *such that* $\hat{C} \cap C^f \neq \emptyset$ *for any* $C^f \in \Psi^f$.

The first condition is the same as the corresponding condition of Proposition 2, and the second condition is the generalized version of the second condition of it.

### 4.2 General Environment

In this subsection, we discuss the self-stability among any possible weighted majority rule. So, now, there is neither the anonymous constraint nor the fixed-weight constraint.[16]

**Proposition 6** (Necessary Condition of Interim Self-stable Constitution)**.**
*A constitution* $(f, F) \in \mathbf{G} \times \mathbf{G}$ *is interim self-stable in* $\mathbf{G}$ *only if*

1. $\exists C^f \in \Psi^f$ *such that* $C^f$ *is not a proper superset of* $C$ *for any* $C \in \Psi^F$ *and*

2. *for some* $i \in \bar{C}$, $\exists \hat{C} \ni i$ *such that* $\hat{C} \cap C^f \neq \emptyset$ *for any* $C^f \in \Psi^f$.

If $r$-type voters cannot be sure of result in the Reform in their beliefs or $s$-type voters cannot be sure of result in the Status-quo, then the constitution is not interim self-stable. So, the conditions in Proposition 6 is a generalized version of the prior propositions.[17]

Now, think about the typical example mentioned above, a presidential system with a veto player. In technical terms, a veto player is an agent who is in any minimal winning coalition. So, if there is a veto agent, without her support, a change cannot be made (or the reform $R$ cannot be achieved). The following lemma shows that a decision rule with a veto agent is interim self-stable.

**Lemma 1** (Sufficient Condition: Veto Agent)**.**
*A weighted majority rule* $f \in \mathbf{G}$ *with a veto agent is interim self-stable in* $\mathbf{G}$.

# 5 Discussion

In this paper, we have identified a set of interim self-stable decision rules. In contrast to the previous studies, which assume that the decision for changing the decision rule takes

[16] Under a given rule, an agent's voting power could be different with another agent's. Moreover, an agent's voting power under a voting rule could be different with the agent's voting power under another voting rule.

[17] Note that this proposition only specifies a necessary condition. Since the set of alternatives is infinite $|G| = \infty$ and there are too many tedious alternatives which have never been considered in reality, it is not easy to pin down a sufficient condition in this setting. For example, for any given constitution $(f, F)$, we can come up with a weird alternative $g$ that gives all powers to one minimal winning coalition of $F$ and assigns zero weights for the others. This alternative may not be interesting to consider, but still in the set of alternative $G$ and makes the given rule $f$ hard to be self-stable. Here, in order to restrict our attention to realistic situations, we examine if a typical example of a general weighted majority rule is interim self-stable.

place before the individuals' preferences over possible outcomes have been realized, we assume that individuals evaluate decision rules after their preferences have been realized. Among anonymous weighted majority rules which are called qualified majority rules, a decision rule is interim self-stable if and only if it is a simple or super majority rule. We also generalize our analysis further by considering a constitution which consists of a pair of voting rules, one of which is for the decision on changing the constitution and the other is for the decision on the final outcome. We show that a constitution is interim self-stable if and only if the voting rule for the final decision is weaker than the voting rule for the change of the rule, and the combining toughness of those two rules should be higher then a certain level. That is, to be stably used, a constitution should consist of the rules for the rule change and for the ordinary decision such that the former is stronger than the latter and the both of them are not too weak.

We also show that the interim self-stable qualified majority constitutions are still interim self-stable even when the society considers more generalized alternatives in a set of weighted majority rules. This result explains why it is hard for one society to move from an "one-person, one-vote" society to another with asymmetric voting powers: Simple or super majority rules can be stably survived in the long run, even though the society considers any general voting rules as alternative. The proposition may also implies why it is important to start with the self-stable qualified majority rules in the virtue of fairness. The society may not need any strong normative arguments to keep the fairness in voting powers. If a society starts with a self-stable constitution consisting of qualified majority rules, the constitution will defend themselves against any "unfair" alternatives.

We characterize a more generalized set of weighted majority rules where individual voters may have different voting powers. We show that, if any individual can be in a minimal winning coalition which veto the ordinary decision together, then the weighted majority rule is interim self stable. We also demonstrate the necessary conditions in general environments which cover the special case with qualified majority rules.

# References


Azrieli, Y. and Kim, S. (2014). Pareto efficiency and weighted majority rules. *International Economic Review*, 55(4):1067–1088.

Azrieli, Y. and Kim, S. (2016). On the self-(in)stability of weighted majority rules. *Games and Economic Behavior*, 100:376 – 389.

Badger, W. (1972). Political individualism, positional preferences, and optimal decision rules. In Niemi, R. G. and Weisberg, H. F., editors, *Probability Models of Collective Decision Making*. Merrill, Columbus, Ohio.

Barberà, S. and Beviá, C. (2002). Self-selection consistent functions. *Journal of Economic Theory*, 105(2):263 – 277.

Barberà, S. and Jackson, M. (2006). On the weights of nations: Assigning voting weights in a heterogeneous union. *Journal of Political Economy*, 114(2):317–339.

Barberà, S. and Jackson, M. O. (2004). Choosing how to choose: Self-stable majority rules and constitutions. *The Quarterly Journal of Economics*, 119(3):1011–1048.

Curtis, R. (1972). Political individualism, positional preferences, and optimal decision rules. In Niemi, R. G. and Weisberg, H. F., editors, *Probability Models of Collective Decision Making*. Merrill, Columbus, Ohio.

Fleurbaey, M. (2008). Weighted majority and democratic theory. *Mimeo*.

Holmström, B. and Myerson, R. B. (1983). Efficient and durable decision rules with incomplete information. *Econometrica*, 51(6):pp. 1799–1819.

Koray, S. (2000). Self-selective social choice functions verify arrow and gibbard- satterthwaite theorems. *Econometrica*, 68(4):981–995.

Koray, S. and Slinko, A. (2008). Self-selective social choice functions. *Social Choice and Welfare*, 31(1):129–149.

May, K. O. (1952). A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica*, 20(4):680–684.

Neumann, J. V. and Morgenstern, O. (1953). *Theory of games and economic behavior,*. Princeton University Press, Princeton.

Penrose, L. S. (1946). The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1):53–57.

Rae, D. W. (1969). Decision-rules and individual values in constitutional choice. *American Political Science Review*, 63:40–56.

# A Appendix

## A.1 Proofs

*Proof of Proposition 1.*

**(Only if part)**

Assume that the current rule $f$ is sub majority rule with the quota $q$ and that it is interim self-stable. Consider the unanimous rule as the alternative rule $g$ with $\frac{n-1}{n} \leq q$. By the assumption, there exists an equilibrium rejection of $g$, $(\sigma, \mu)$. Fix agent $i$ with $\bar{t}_i = s$. Define $\bar{T}_{-i} \equiv \{t_{-i} \in T_{-i} : f(t_{-i}, \bar{t}_i) = R\}$. In the equilibrium rejection $(\sigma, \mu)$, for the agent $i$ the left hand side of Equation (C.4) is

$\sum_{t_{-i}} \mu_i(t_{-i}) u_i(f(t_{-i}, \bar{t}_i), \bar{t}_i) = -\sum_{t_{-i} \in \bar{T}_{-i}} \mu_i(t_{-i})$ and the right hand side of Equation (C.4) is zero.

We claim that $\sum_{t_{-i} \in \bar{T}_{-i}} \mu_i(t_{-i}) > 0$ in any equilibrium rejection. Note that under qualified majority rule, at most $q$ agents vote for $g$ with a positive probability for any $t \in T$ in any equilibrium rejection. In other words, $n - q$ agents never vote for $g$. There are two cases regarding the probability of agent $i$ being pivotal for any equilibrium rejection. First there exists a $t_{-i} \in T_{-i}$ such that $\gamma(t_{-i}) > 0$. It implies that exactly $q$ agents vote for $g$ with a positive probability at $t_{-i}$. Fix these agents and we can find a $\bar{t}_{-i} \in \bar{T}_{-i}$ such that $\gamma(\bar{t}_{-i}) > 0$ since the number of other agents is $n - q - 1 > q$ and they decide $f(\bar{t}) = R$ by themselves. Then by Bayes theorem, $\sum_{t_{-i} \in \bar{T}_{-i}} \mu_i(t_{-i}) > 0$. Second for any $t_{-i} \in T_{-i}$, $\gamma(t_{-i}) = 0$. We can find a $\tilde{t}_{-i} \in T_{-i}$ such that $\mu_i(\tilde{t}_{-i}) > 0$. With the similar trick of the previous case, fix agents in $H_i$ at $\tilde{t}_{-i}$ and we can find a $\bar{t}_{-i} \in \bar{T}_{-i}$ such that $\lim_{k \to \infty} \frac{\left(\prod_{j \in H_i} \sigma_j(\bar{t}_j)\right)\left(\prod_{j \in N/(H_i \cup \{i\})} (1 - \sigma_j(\bar{t}_j))\right)}{\left(\prod_{j \in H_i} \sigma_j(\tilde{t}_j)\right)\left(\prod_{j \in N/(H_i \cup \{i\})} (1 - \sigma_j(\tilde{t}_j))\right)} = \frac{\left(\prod_{j \in H_i} \sigma_j(\tilde{t}_j)\right)}{\left(\prod_{j \in H_i} \sigma_j(\tilde{t}_j)\right)} = 1$. Then, $\mu_i(\bar{t}_{-i}) > 0$ which proves the claim. By the claim, the condition (C.4) implies that $\sigma_i(\bar{t}_i) > 0$. The argument is valid for any agent $i$ with $t_i = s$. However, at the type profile $\bar{t}$ with $|\{i : \bar{t}_i = s\}| > q$, this equilibrium rejection contradicts (C.1).

**(If part)**

We only show that the simple majority rule is interim self-stable because the proof for super majority rules is almost the same. Among alternative rules, we consider the two extreme qualified majority rules, $q = n - 1$ and 0. When the alternative rule is the unanimous rule, i.e., $q = n - 1$, consider the strategy profile and a belief $(\sigma, \mu)$ such that $\sigma_i(t_i) = 0$ for $\forall t_i \in T_i$, $\sigma_i^k(s) = \frac{1}{k}$, and $\sigma_i^k(r) = \frac{1}{k^2}$ for $\forall i \in N$. This trivial strategy profile simply satisfies the conditions (C.1) and (C.2). By (C.3), we can derive $\mu_i(t_{-i}) > 0$ for $t_{-i}$ such that $|\{i : t_i = s\}| = q^s$ and $\mu_i(t_{-i}) = 0$ otherwise. The right hand side of the equation in (C.4) is weakly less than the left for any type and any agent. Thus, the pair $(\sigma, \mu)$ is an equilibrium rejection of $g$. When the alternative rule is the other extreme case of $q = 0$, we can similarly find an equilibrium rejection $(\sigma, \mu)$ such that $\sigma_i(t_i) = 0$ for $\forall t_i \in T_i$, $\sigma_i^k(s) = \frac{1}{k^2}$, and $\sigma_i^k(r) = \frac{1}{k}$ for $\forall i \in N$. For any intermediate qualified majority rule with $0 < q < n - 1$, the same argument is valid, which proves that the simple majority rule is interim self-stable. □

*Proof of Proposition 2.*

**(If part)** First, consider any alternative rule $g$ with $q^g > q^f$. Set $\sigma_i(s) = \sigma_i(r) = 0$ and $\sigma_i^k(s) > \sigma_i^k(r)$. By construction, $f = R$ if $g = R$. So, $\sigma_i(r) = 0$ is always justified. For $t_i = s$, $\mu_i(t_{-i}) > 0$ only when $f(t) = S$.[18] So, $\sigma_i(s) = 0$ is also justified.

Second, consider any alternative rule $g$ with $q^g < q^f$. Set $\sigma_i(s) = \sigma_i(r) = 0$ and $\sigma_i^k(s) < \sigma_i^k(r)$. By construction, $f = R$ only if $g = R$. So, $\sigma_i(s) = 0$ is always justified. For $t_i = r$, $\mu_i(t_{-i}) > 0$ only when $f(t) = R$.[19] So, $\sigma_i(r) = 0$ is also justified.

**(Only if part)** Suppose $q < q^f$. Consider an alternative rule $g$ with $q^g = 0$. By construction, $f = R$ only if $g = R$. For any arbitrary $i$ with $t_i = r$, $g = R$. Moreover, if there exists an equilibrium rejection, $\mu_i(t_{-i}) > 0$ for some $t_{-i}$ with $f(r, t_{-i}) = S$. Hence, from Condition (C.4), $\sigma_i(r) > 0$. We have picked an arbitrary $i$, so there cannot be an equilibrium rejection of $g$.

Suppose now $q + q^f < n - 1$. Consider the unanimous rule $g$. By construction, $f = R$ if $g = R$. For any arbitrary $i$ with $t_i = s$, $g = S$. Moreover, if there exists an equilibrium rejection, $\mu_i(t_{-i}) > 0$ for some $t_{-i}$ with $f(s, t_{-i}) = R$. Hence, from Condition (C.4), $\sigma_i(s) > 0$. We have picked an arbitrary $i$, so there cannot be an equilibrium rejection of $g$. □

**Lemma 2** (Equilibrium property 1)**.**
*Consider a constitution $(f, F)$. If there exists a equilibrium rejection $(\sigma, \mu)$ of an alternative rule $g$, then, for each $i$ such that $\sigma_j(s) + \sigma_j(s) > 0$ for any $j \geq i$, there exists a set of agents $\tilde{H} \in \Phi_i$ and a type profile $\tilde{t}_{-i} \in T_{-i}$ such that*

$$\lim_{k \to \infty} \frac{\rho(\tilde{H}; \tilde{t}_{-i}, \sigma^k)}{\rho(H; t_{-i}, \sigma^k)} > 0 \qquad \forall H \in \Phi_i, \forall t_{-i} \in T_{-i} \tag{A.1}$$

*which satisfies $\sigma_j(s) = \sigma_j(r) = 0$ for all $j \in N \setminus (\tilde{H} \cup \{i\})$.*

***Proof** of Lemma 2.* Without loss of generality, suppose $w_i \geq w_j$ if and only if $i \geq j$ under $F$.

First, consider $i = n$. Suppose not. So for any $\tilde{H}$ and a type profile $\tilde{t}_{-i} \in T_{-i}$ which makes Equation (A.1) goes to zero in the slowest speed, there is some agent $j \in N \setminus (\tilde{H} \cup \{i\})$ with $\sigma_j(s) + \sigma_j(r) > 0$. Define $\hat{N}^+ \equiv \{j \in N \setminus (\tilde{H} \cup \{i\}) | \sigma_j(s) + \sigma_j(r) > 0\}$. Also define $\tilde{N}^+ \equiv \{j \in \tilde{H} | \sigma_j(\tilde{t}_j) > 0\}$. We know $w(\hat{N}^+ \cup \tilde{N}^+ \cup \{i\}) \leq q$ ($\because \sigma$ consists an equilibrium rejection) and $w(\tilde{H} \cup \{i\}) > q$. Now, pick $j \in \tilde{H} \setminus (\hat{N}^+ \cup \tilde{N}^+)$ with the lowest weight, and add her into the set $(\hat{N}^+ \cup \tilde{N}^+ \cup \{i\})$. Repeat it until the set turns to a winning coalition. Denote the winning coalition $WC'$ and also $H' \equiv WC' \setminus \{i\}$. By construction, $H' \in \Phi_i$ ($\because i = n$ has the highest weight). Then we find a contradiction, since $\rho(H'; t_{-i}, \sigma)$ goes to zero in a speed that is slower than $\rho(\tilde{H}; \tilde{t}_{-i}, \sigma)$.

Second, consider $i = n - 1$. A similar argument from above works for any $j < i$. So, we only need to show that there exists some $\tilde{H}$ such that $j = n$ with $\sigma_j(s) + \sigma_j(r) > 0$ is (also) not in $N \setminus (\tilde{H} \cup \{i\})$. Suppose $j \in N \setminus (\tilde{H} \cup \{i\})$ with $\sigma_j(s) + \sigma_j(r) > 0$. Follow the same logic from above to find $H'$. Because $j = n$ is not in $\tilde{H} \setminus (\hat{N}^+ \cup \tilde{N}^+)$, we still have $H' \in \Phi_{i=n-1}$.

[18] The agent $i$ believes she is pivotal only when at least $q + 1$ agents including her have $s$-type.
[19] The agent $i$ believes she is pivotal only when at least $q + 1$ agents including her have $r$-type.

Then, we can show that $\rho(H'; t_{-i}, \sigma)$ goes to zero in a speed that is slower than $\rho(\tilde{H}; \tilde{t}_{-i}, \sigma)$. Contradiction.

Similar arguments apply for any $i \in N$. □

***Proof** of Proposition 3.* **Case 1:** Consider the case where $\forall C^g$, $\exists C^f \subseteq C^g$. Set $\sigma_i(s) = \sigma_i(r) = 0$ and $\sigma_i^k(s) > \sigma_i^k(r)$. By construction, $f = R$ if $g = R$. So, $\sigma_i(r) = 0$ is always justified. For $t_i = s$, $\mu_i(t_{-i}) > 0$ only when $f(t) = S$. ($\because$ The agent $i$ believes she is pivotal only when at least $q+1$ agents including herself are $s$-type. And we know $q^f \geq n - (q+1)$ by construction.) So, $\sigma_i(s) = 0$ is also justified.

**Case 2:** Consider the case where $\exists \bar{C}^g$ such that $\bar{C}^g \not\supseteq C^f \ \forall C^f$.

Define $\bar{C}^f \equiv \arg\max_{C^f} w^g(C^f)$. (More than one?) Then, by construction, for some $C^g$, $\bar{C}^f \supsetneq C^g$, so $\exists i \in \bar{C}^f$ such that $w^g(\bar{C}^f \setminus \{i\}) > q^g$.

**Case 2-1:** Consider the case where $\forall i \in \bar{C}^f$, we have $w^g(\bar{C}^f \setminus \{i\}) > q^g$. Set, for any $i$, $\sigma_i(s) = \sigma_i(r) = 0$, $\sigma_i^k(s) = k^{-\frac{2}{w_i^g}}$ and $\sigma_i^k(r) = k^{-\frac{1}{w_i^g}}$. So, $\sigma_i^k(s) < \sigma_i^k(r)$ for any $i$, and $\sigma_i^k(r) < \sigma_j^k(r)$ for any $i$ and $j$ such that $w_i^g < w_j^g$. For $t_i = s$, $\mu_i(t_{-i}) > 0$ only when $g(t) = R$. ($\because$ For any $i$, $\mu_i(t_{-i}) > 0$ only when $\forall j \in (\bar{C}^f \setminus \{i\})$ has $t_j = r$.) So, $\sigma_i(s) = 0$ is okay. For $t_i = r$, $\mu_i(t_{-i}) > 0$ only when $f(t) = R$. ($\because$ The agent $i$ believes she is pivotal only when at least $q+1$ agents including herself are $r$-type. And we know $q^f \leq q$ by construction.) So, $\sigma_i(r) = 0$ is okay.

**Case 2-2:** Consider the case where for some $i \in \bar{C}^f$, we have $w^g(\bar{C}^f \setminus \{i\}) \leq q^g$. Define $J \equiv \{i \in \bar{C}^f : w^g(\bar{C}^f \setminus \{i\}) \leq q^g\}$. Pick $\bar{C}^F \supseteq \bar{C}^f$. Set, for any $i \in J$, $\sigma_i(s) = \epsilon$ and $\sigma_i(r) = 1$, and for any $i \in \bar{C}^F \setminus \bar{C}^f$, set $\sigma_i(s) = 1$ and $\sigma_i(r) = \epsilon$. For all others, such that $i \in \bar{C}^f \setminus J$ or $i \in N \setminus \bar{C}^F$, set $\sigma_i(s) = \sigma_i(r) = 0$ and $\sigma_i^k(s) < \sigma_i^k(r)$. We show that this construction can form a equilibrium rejection of $g$ with some small enough positive value of $\epsilon$.

Check if $\sigma_i(r) = 0$ is okay. If she believes $f = R$, then it is okay. If she believes $f(t) = S$, she knows $g(t) = S$. (Suppose not, so she believes $f = S$ and $g = R$. If all agents in $J$ are $r$-type, then some $C^f \supset J$ containing her $i$ contains only $r$-type agents. So, to have $f = S$, she knows at least one agents in $J$ should be $s$-type. Then, by construction of $\bar{C}^f \supset J$, more than $g^f + 1$ agents should be $r$-type to make $g = R$. So, $f = R$. Contradiction.) So for any belief, $\sigma_i(r) = 0$ is okay.

Now, check if $\sigma_i(s) = 0$ is okay. Suppose $\epsilon = 0$. Then she always believes $g = R$. Moreover, she believes $f = S$ with positive probability. (Only agents in $J$ and $q^f - |J|$ more agents are guaranteed to be $r$ types, and all others could be $s$-type with positive probability.) Therefore, $\sum_{t_{-i}} \mu_i(t_{-i}) u_i(f(t), t_i) - \sum_{t_{-i}} \mu_i(t_{-i}) u_i(g(t), t_i) > 0$ when $\epsilon = 0$. If we denote by $\kappa_i$ the event

that makes $g = R$ and $f = S$, then $\lim_{\epsilon \to 0} \mu(\kappa_i) = 1$. Write

$$\begin{aligned}
&\sum_{t_{-i}} \mu_i(t_{-i}) u_i(f(t), t_i) - \sum_{t_{-i}} \mu_i(t_{-i}) u_i(g(t), t_i) \\
&= \mu(\kappa_i) \left( u_i(f(t) = S, t_i = s) - u_i(g(t) = R, t_i = s) \right) \\
&\quad + (1 - \mu(\kappa_i)) \left( \sum_{t_{-i}} \mu_i(t_{-i} | \kappa_i^c) u_i(f(t), t_i = s) - \sum_{t_{-i}} \mu_i(t_{-i} | \kappa_i^c) u_i(g(t), t_i = s) \right) \\
&= \mu(\kappa_i) \left( 0 - (-1) \right) \\
&\quad + (1 - \mu(\kappa_i)) \left( \sum_{t_{-i}} \mu_i(t_{-i} | \kappa_i^c) u_i(f(t), t_i = s) - \sum_{t_{-i}} \mu_i(t_{-i} | \kappa_i^c) u_i(g(t), t_i = s) \right) \\
&\geq \mu(\kappa_i) + (1 - \mu(\kappa_i)) \times (-1) = (2\mu(\kappa_i) - 1) .
\end{aligned}$$

Since $\lim_{\epsilon \to 0} \mu(\kappa_i) = 1$, for a small enough $\epsilon$, $(2\mu(\kappa_i) - 1)$ is always positive. Therefore, $\sigma_i(r) = 0$ is okay. □

***Proof** of Proposition 4.*

Find $H$ such that $H \in \Phi_i$ for any $i \in N \setminus H$ and $(N \setminus C^*) \subset H$ for some $j \in N \setminus H$.[20]

Consider first the case with $q^g < q^f$. Set $\sigma_i(t_i) = 0$ for all $i$ and $t_i$ and $\sigma_i^k(r) = k^{-\frac{1}{w_i}}$ and $\sigma_i^k(s) = k^{-\frac{2}{w_i}}$. For any $s$-type agent, $\sigma_i(s) = 0$ is okay, since $g(\cdot) = S$ implies $f(\cdot) = S$. For an $r$-type agent, $\mu(t_{-i}) > 0$ only when $f(t_i = r, t_{-i}) = R$. So $\sigma_i(r) = 0$ is justified.

Consider now the case with $q^g > q^f$. Set $\sigma_i(t_i) = 0$ for all $i$ and $t_i$ and $\sigma_i^k(r) = k^{-\frac{2}{w_i}}$ and $\sigma_i^k(s) = k^{-\frac{1}{w_i}}$. For any $r$-type agent, $\sigma_i(r) = 0$ is okay, since $g(\cdot) = R$ implies $f(\cdot) = R$. For an $s$-type agent $i$, Equation (A.1) with $\tilde{H} = \hat{C} \setminus \{i\}$ where $\hat{C}$ has the highest weights among the minimal coalitions contain the agent $i$ converges to zero in a speed that is slower than that with any other $\tilde{H}$. By construction, $w(N \setminus \hat{C}) \leq q^f$. So, $\mu(t_{-i}) > 0$ only when $f(t_i = r, t_{-i}) = S$. Therefore, $\sigma_i(s) = 0$ is justified. □

***Proof** of Proposition 5.*

We prove by contradiction.

First, suppose $q < q^f$ and there exists some equilibrium rejection $(\sigma, \mu)$ of $g$ with $q^g = 0$. For any $i$, $\mu_i(t_{-i}) > 0$ for some $t_{-i}$ such that $f(t_i = r, t_{-i}) = S$, while $g(t_i = r, t_{-i}) = R$ always. So, $\sigma_i(r)$ should be positive from Condition (C.4). Contradiction.

Second, suppose $\forall i \in \bar{C}$, $\forall C \ni i$, $w(N \setminus C) > q^f$. And suppose there exists some equilibrium rejection $(\sigma, \mu)$ of the unanimity rule $g$. Consider the agent $i = n$. From Lemma 2, we know for Equation A.1, $j \in N \setminus (\tilde{H} \cup \{i\})$ should have $\sigma_j(s) = \sigma_j(r) = 0$. By construction, there exist a minimal winning coalition $C \ni i$ such that $C \subset (\tilde{H} \cup \{i\})$, and any $j \in \tilde{H} \setminus C$ should have $\sigma_j(s) + \sigma_j(r) > 0$. (If not, there should exists some slower $H'$ which does not contain $j$ with $\sigma_j(s) + \sigma_j(r) = 0$ than $\tilde{H}$.) Also, if $\sigma_j(r) = 0$ for some $j \in \tilde{H} \setminus C$, there exists some $\tilde{H}'$

[20] How? Add $i$ with the smallest weight into the set $N \setminus C^*$. Repeat until the set becomes to be a winning coalition of all $i$ not in the set.

and $t'_{-i}$ where $j \in N \setminus \tilde{H}'$ and $t_j = r$, which gives the same convergence speed for $\rho(\tilde{H}'; t'_{-i}, \sigma)$ with $\rho(\tilde{H}; t_{-i}, \sigma)$. We know $C$ has a mutually exclusive $C^f$ and have shown that all $j \in C^f$ have $r$-types with positive probability in the sense of posterior belief $\mu$. Therefore, for $i = n$, $\mu_i(t_{-i}) > 0$ for some $f(t_i = s, t_{-i}) = R$, while $g(t_i = s, t_{-i}) = S$ always. So, $\sigma_i(s)$ should be positive from Condition (C.4).

For some $i \in \bar{C}$ such that $i \neq n$, a similar logic can be applied. So, for all $i \in \bar{C}$, $\sigma_i(s) > 0$. Contradiction. □

***Proof** of Proposition 6.*

We prove by contradiction.

First, suppose $\forall C^f \in \Psi^f$, $\exists C \in \Psi^F$ such that $C \subseteq C^f$. So, $f = R$ implies $F = R$. Also, suppose there exists some equilibrium rejection $(\sigma, \mu)$ of $g$ with $q^g = 0$. For any $i$, $\mu_i(t_{-i}) > 0$ for some $t_{-i}$ such that $f(t_i = r, t_{-i}) = S$, while $g(t_i = r, t_{-i}) = R$ always. So, $\sigma_i(r)$ should be positive from Condition (C.4). Contradiction.

The second part of the proof is the same as that of Proof of "Only if part" in Proposition 5. □

*Proof of Lemma 1.*

Consider a strategy profile and a belief system $(\sigma, \mu)$ such that, for all $j \in N$, $\sigma_{j \neq i}(r) = 1$, $\sigma_{j \neq i}(s) = 0$, $\sigma_i(s) = \sigma_i(r) = 0$ and $\sigma^k_{j \neq i}(s) > \sigma^k_i(s) > \sigma^k_i(r)$.

For the veto agent $i$, if $t_i = s$, $\sigma_i(s) = 0$ is justified since $f(t_i, t_{-i}) = S$ for any $t_{-i}$. If $t_i = r$, $\sigma_i(r) = 0$ is justified since $i$ is pivotal only when the right enough number of other agents with type $r$ vote for $g$, so only when $f(t_i, t_{-i}) = R$.

For all other agents $j \neq i$, $\sigma_j(s) = 0$ is justified since $\gamma_j(t_j) = 0$ for all $t_j$ and $\mu_j(t_j, t_{-j})$ is positive only when $t_i = s$, so $f(t_j, t_{-j}) = S$. Also, $\sigma_j(r) = 1$ obviously satisfies Equation (C.2) since $\gamma_j(t_j) = 0$ for all $t_j$ and does not violate Equation (C.4). □

## A.2 Consistency of Definition

Here, we discuss the consistency of our definition of interim self stability with the ex-ante self-stability à la Azrieli and Kim (2016) and the durability à la Holmström and Myerson (1983).

Consider the "ex-ante environment" studied in Azrieli and Kim (2016), where agents vote on rule change before their types are realized. We rewrite our conditions and definition as follows.

To reject the alternative rule $g$ all the time, the probability that $g$ gets sufficient weighted votes should be zero. In other words, the alternative $g$ is always rejected if and only if

$$\sum_{\{j:\sigma_j > 0\}} w_j \leq q. \tag{A.2}$$

If Equation (A.2) holds, then honest behavior in $f$ and $g$ (we consider incentive compatible $f$ and $g$), together with the voting strategies in the first stage, $\sigma = (\sigma_1, \ldots, \sigma_n)$ form a Nash

equilibrium if and only if

$$\gamma_i \left( u_i(f) - u_i(g) \right) \geq 0 \quad \forall i, \tag{A.3}$$

where

$$u_i(f) = a \sum_{\{t \in T : t_i = r\}} p(t) f(t) - \sum_{\{t \in T : t_i = s\}} p(t) f(t),$$

$$\Phi_i = \{ H_i \subseteq N / \{i\} \,|\, q - w_i < \sum_{j \in H_i} w_j \leq q \},$$

and

$$\gamma_i = \sum_{H_i \in \Phi_i} \left( \prod_{j \in H_i} \sigma_j \right) \left( \prod_{j \in N/(H_i \cup \{i\})} (1 - \sigma_j) \right).$$

We require that, for any individual $i$,

$$\text{if } \; u_i(f) < u_i(g), \text{then } \; \sigma_i = 1. \tag{A.4}$$

This condition imposes that, if the expected utility of individual $i$ in the alternative decision rule $g$ would be higher than in the current rule $f$, then individual $i$ should vote for $g$.[21]

$w(Y) := \sum_{i \in Y} w_i$ denotes the total weight of coalition $Y$.

**Proposition 7.** *For a given weighted majority rule $f$, $w(\{i : u_i(f) < u_i(g)\}) \leq q$ for any alternative rule $g$ if and only if there exists a strategy profile $\sigma$ that satisfies conditions (A.2), (A.3), and (A.4).*

*Proof of Proposition 7.*

**(Only if part)**

Suppose $w(\{i : u_i(f) < u_i(g)\}) \leq q$. Then, set $\sigma_i = 1$ for any $i \in \{i : u_i(f) < u_i(g)\}$ and $\sigma_i = 0$ for any $i \notin \{i : u_i(f) < u_i(g)\}$. The condition (A.2) and (A.4) are satisfied. For an individual $i$ with $\sigma_i = 1$, $\gamma_i = 0$. For an individual $i$ with $\sigma_i = 0$, $(u_i(f) - u_i(g)) \geq 0$ by construction. Therefore, the condition (A.3) is satisfied.

**(If part)**

Suppose not. That is, the conditions (A.2), (A.3), and (A.4) are all satisfied, but

$$w(\{i : u_i(f) < u_i(g)\}) > q.$$

Since we suppose the condition (A.4) is satisfied, $\{j : u_j(f) < u_j(g)\} \subseteq \{j : \sigma_j > 0\}$, which implies $w(\{i : u_i(f) < u_i(g)\}) \leq w(\{j : \sigma_j > 0\})$. Then, the condition (A.2) is violated, since

[21] In the second stage, since $f$ and $g$ are incentive compatible, we simply assume that all individuals report their true types.

$\sum_{\{j:\sigma_j>0\}} w_j \geq w(\{i : u_i(f) < u_i(g)\}) > q$. Contradiction. □

One may wonder why we don't use the simple condition as $w(\{i : u_i(f) < u_i(g)\}) \leq q$ in Azrieli and Kim (2016) to define interim self-stability. To do that, in our setting, we need to add up the weights of agents $i$'s who have

$$\sum_{t_{-i}} \mu_i(t_{-i}) u_i(f(t), t_i) < \sum_{t_{-i}} \mu_i(t_{-i}) u_i(g(t), t_i),$$

which is a part of the condition (C.4). But as in the condition (C.3), the posterior belief $\mu_i$ can only be calculated with a strategy profile for the first stage voting game $\sigma$. That is, to define interim self-stability in a way analogous to Azrieli and Kim (2016), we need a complete characterization of a Nash equilibrium with a sequentially rational strategy profile and a consistent belief system.

## A.3 Extra Lemmas and Propositions

We show that if any individual can be in a minimal winning coalition which veto the ordinary decision together, then the weighted majority rule is interim self stable.

**Lemma 3** (Fixed-weight Environment: Sufficient condition)**.**
*A weighted majority rule $f \in \mathbf{G}(w)$ is interim self-stable in $\mathbf{G}(w)$ if $\forall i$, $\exists C \ni i$ such that $w(N \setminus C) \leq q$.*

We omit the proof of this lemma since it is a corollary of Proposition 4 for a constitution $(f, f)$. Instead, we provide an intuition behind it. Lemma 3 says if any individual under the given rule is a member of a minimal winning coalition which is a veto coalition at the same time, then the rule is interim self-stable. If there are enough number of $r$-type voters so that they together can change the given rule to a weaker rule, it is obvious that individuals does not have to vote for the change since the given rule is already weak enough. If there are enough number of $s$-type voters so that they can change the given rule to a stronger rule they all prefer, one may believe they don't have to change the rule since they are strong enough to veto any attempt to reform under the given rule.

We show that, if a decision rule is an interim self-stable weighted majority rule, then there exists an individual with a relatively high voting power such that any minimal coalition containing the individual is not mutually exclusive with all other minimal coalitions. Note that this condition is identical to the necessary (and sufficient) condition of interim self-stability from the previous sections.

**Proposition 8** (Necessary Condition in the General Environment)**.**
*A weighted majority rule $f \in \mathbf{G}$ is interim self stable in $\mathbf{G}$ only if for some $i \in \bar{C}$, $\exists \hat{C} \ni i$ such that $\hat{C} \cap C^f \neq \emptyset$ for any $C^f \in \Psi^f$.*

Lemma 4 is an analogy of the sufficient condition of super majority qualified majority rules.

**Lemma 4** (Fixed-weight Environment: Sufficient Condition 2)**.**
*A weighted majority rule $f \in G(w)$ is interim self-stable among $G(w)$ if there exists a minimal winning coalition $C^*$ such that $w(N \setminus C^*) + w_i \leq q$ for some $i \in C^*$.*

*Proof of Lemma 4.*
Let $\Psi^{\hat{f}}$ denote the set of minimal winning coalitions(MWCs) under a decision rule $\hat{f}$. We also define, for a decision rule $\hat{f}$ and a type $t_i$, $T_{t_i}^{\hat{f}} \equiv \left\{t_{-i} : \hat{f}(t_i, t_{-i}) = R\right\}$, and for a set of type profile $\tilde{T} \subseteq T_{-i}$, $\mu_i(\tilde{T}) \equiv \sum_{t_{-i} \in \tilde{T}} \mu_i(t_{-i})$. For the convenience of notation, $w_i \geq w_j$ if and only if $i \geq j$.

Find $H$ such that $H \in \Phi_i$ for any $i \in N \setminus H$ and $(N \setminus C^*) \subset H$.[22] By construction, $w(N \setminus (H \cup \{i\})) \leq q$ for any $i \in N \setminus H$ and $C^* \cap H \neq \emptyset$.

First, consider the case where $q_f > q_g$. So, if $g(t) = S$, then $f(t) = S$. And if $f(t) = R$, then $g(t) = R$. Suppose a strategy profile $(\sigma, \mu)$ such that $\sigma_i(s) = 0$ for any $i$ and $\sigma_j(r) = 0$ for any $j \notin H$ and $\sigma_j(r) = 1$ for $j \in H$, and any arbitrary $\sigma_i^k(t_i)$ which converges to $\sigma_i(t_i)$ for any $i$ and $t_i$. $\sigma_i(s)$ is always justified, since $g(t) = S$ implies $f(t) = S$. For any $i \notin H$, $\gamma_i(t_{-i}) > 0$ only when $t_j = r$ for all $j \in H$. Then, for $t_i = r$, Condition (C.2) is satisfied, and so $\sigma_i(r) = 0$ is justified. For any $i \in H_{\underline{i}}$, $\gamma_i(t_{-i})$ is always zero, $\sigma_j(r) = 1$ it is okay.

Second, consider the case where $q_f < q_g$. So, if $f(t) = S$, then $g(t) = S$. And if $g(t) = R$, then $f(t) = R$. Suppose a strategy profile $(\sigma, \mu)$ such that $\sigma_i(r) = 0$ for any $i$ and $\sigma_j(s) = 0$ for any $j \notin H$ and $\sigma_j(s) = 1$ for $j \in H$, and any arbitrary $\sigma_i^k(t_i)$ which converges to $\sigma_i(t_i)$ for any $i$ and $t_i$. $\sigma_i(r)$ is always justified, since $g(t) = R$ implies $f(t) = R$. For any $i \notin H$, $\gamma_i(t_{-i}) > 0$ only when $f(t_i, t_{-i}) = S$. Then, for $t_i = s$, Condition (C.2) is satisfied, and so $\sigma_i(s) = 0$ is justified. For any $i \in H_{\underline{i}}$, $\gamma_i(t_{-i})$ is always zero, so $\sigma_j(s) = 1$ is okay. □

**Lemma 5.** *If $\exists i^*$ such that, $\forall C^* \ni i^*$, $\exists C \cap C^* = \emptyset$, then $w(N) > 2q$.*

**Lemma 6.** *If, for some $i^*$, $w_{i^*} < w(N) - 2q$, then $\sigma_{i^*}(s) = 1$ in equilibrium.*

**Lemma 7.** *If there exists an equilibrium rejection, then $w(\{i | w_i < w(N) - 2q\}) \leq q$.*

**Lemma 8.** *There exists a minimal winning coalition $C^*$ such that $w(N \setminus C^*) + w_{i^*} \leq q$ for some $i^* \in C^*$ if and only if no pair of minimal winning coalitions is mutually exclusive.*

*Proof.* Consider $A \equiv C^* \setminus \{i^*\}$ and $B \equiv \{i^*\} \cup (N \setminus C^*)$. We know $w(A) \leq q$ and $w(B) \leq q$. Then we cannot find any partition of $N$ with two winning coalitions. □

**Lemma 9.** *There exists an agent $i$ such that for any $C^* \ni i$ $w(N \setminus C^*) \leq q$ if and only if no pair of minimal winning coalitions is mutually exclusive.*

*Proof.* The "If" part is straightforward.

(Only if) Suppose not. So, there exists a pair of minimal winning coalitions $C$ and $C'$ such that $C \cap C' = \emptyset$. By construction $i$ is neither in $C$ nor in $C'$. We know $w(N) \geq w(C) + w(C') +$

[22] How? Add $i$ with the smallest weight into the set $N \setminus C^*$. Repeat until the set becomes to be a winning coalition of all $i$ not in the set.

$w_i > q + q + w_i$. For any $C^* \ni i$, we have $w(N \setminus C^*) \leq q$ and $q < w(C^*) \leq q + w_i$. So, $w(N) = w(N \setminus C^*) + w(C^*) \leq q + q + w_i$. Contradiction. □

**Lemma 10.**

*There is a minimal winning coalition $\bar{C} \in \Psi^f$ where for any $i \in \bar{C}$, any minimal winning coalition $C_i \ni i$ has a mutually exclusive minimal winning coalition $C' \in \Psi^f$ such that $C_i \cap C' = \emptyset$ if and only if any minimal winning coalition in $\Psi^f$ has a mutually exclusive minimal winning coalition in $\Psi^f$.*

*Proof.*

**(Only if)**

Assume a minimal winning coalition $\bar{C} \in \Psi^f$ where for any $i \in \bar{C}$, any minimal winning coalition $C_i \ni i$ has a mutually exclusive minimal winning coalition $C' \in \Psi^f$ such that $C_i \cap C' = \emptyset$. If there exists a minimal winning coalition $\tilde{C}$ which is not mutually exclusive with any other minimal winning coalition, then $\tilde{C}$ should not contain any $i \in \bar{C}$. So, $\tilde{C} \cap \bar{C} = \emptyset$. Contradiction.

**(If)**

It is obvious. □

**Lemma 11.**

*If any minimal winning coalition has a mutually exclusive minimal winning coalition, there exists a pair of mutually exclusive minimal winning coalitions $C$ and $C'$ such that $w_i \geq w_{i'}$ for any $i \in C$ and $i' \in C'$.*

*Proof.* There always exists a minimal winning coalition $\bar{C} \in \Psi^f$ such that for all $i \in \bar{C}$, $w_i \geq w_j$ for any $j \in N \setminus \bar{C}$. If the minimal winning coalition $\bar{C}$ has a mutually exclusive minimal winning coalition $C'$, then $w_i \geq w_{i'}$ for any $i \in \bar{C}$ and $i' \in C'$. □

**Lemma 12** (Necessary Condition 1: Single Agent Minimal Winning Coalition)**.**

*If, under $f$, there exists an agent $i$ who consists a minimal winning coalition by itself $C = \{i\}$ and a mutually exclusive minimal winning coalition $\tilde{C}$ such that $\tilde{C} \cap C = \emptyset$, $f$ is not interim self stable.*

*Proof of Lemma 12.*

By construction, the agent $i$ is always pivotal, $\gamma(t_{-i}) = 1$ for all $t_{-i}$.

Consider an alternative rule $g$ such that $w_i^g = w_i^f$ and $w_j^g = 0$ for all $j \neq i$.

Then, for $t_i = s$, Equation (C.2) is violated since

$$\sum_{t_{-i}} P(t_{-i})\, u_i(f(t), t_i) = \sum_{t_{-i} \in T^f_{t_i = s}} P(t_{-i})\, u_i(f(t), t_i) = -\sum_{t_{-i} \in T^f_{t_i = s}} P(t_{-i}) < 0.$$

So, there is no equilibrium rejection of $g$. □

**Lemma 13** (Necessary Condition 2: Small Quota)**.**
*If there exists a minimal winning coalition $C \in \Psi^f$ such that for any $i \in C$ and for any minimal winning coalition $C_i \ni i$, $w(N \setminus C_i) > 2q$, $f$ is not interim self-stable.*

*Proof of Lemma 13.*
Let an alternative rule $g$ be the unanimous rule. By construction, for any agent $i \in C$, $T^f_{t_i=s} \neq \emptyset$ and $T^g_{t_i=s} = \emptyset$.

We prove by contradiction. Let's suppose there exists an equilibrium of $g$, $(\sigma, \mu)$.

For $t_i = s$, suppose $\sigma_i(s) \neq 1$. From Equation (C.4), $-\mu_i(T^f_s) \geq -\mu_i(T^g_s)$. We know $\mu_i(T^g_r) = 0$, we should have $\mu_i(T^f_s) = 0$.

So, if a type profile $\tilde{t}_{-i} \in T_{-i}$ and some set of agents $\tilde{H} \in \Phi_i$,

$$\lim_{k\to\infty} \left( \prod_{j\in\tilde{H}} \sigma^k_j(\tilde{t}_j) \right) \left( \prod_{j\in N\setminus(\tilde{H}\cup\{i\})} (1 - \sigma^k_j(\tilde{t}_j)) \right)$$

goes to zero in a speed that is no faster than for any other $H \in \Phi_i$ and type profile $t_{-i} \in T_{-i}$, $\tilde{t}_{-i}$ should not be in $T^f_{t_i=s}$. It means that, for a minimal winning coalition $C_i \ni i$ which is a subset of $\tilde{H}$, any $j \in \left(\tilde{H} \setminus C_i\right)$ have either $\sigma_j(s) > 0$ or $\sigma_j(r) > 0$. Also, there should not be any minimal winning coalition with all $r$ types in $\tilde{t}_{-i}$. It means that $w(\left\{ j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = r \right\}) \leq q$. Then, we should have enough number of agents $j \in N \setminus (\tilde{H} \cup \{i\})$ such that $\tilde{t}_j = s$ and

$$w(\left\{ j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = s \right\}) \geq w(N \setminus (\tilde{H} \cup \{i\})) - q.$$

It implies that for such $j$ with $\tilde{t}_j = s$ we should have $\sigma_j(r) = 1$. But, since $w(N \setminus C_i) > 2q$, $w(N \setminus C_i) = w(\tilde{H} \setminus C_i) + w(N \setminus (\tilde{H} \cup \{i\}))$ and

$$\begin{aligned} &w(N \setminus (\tilde{H} \cup \{i\})) \\ &\quad = w(\left\{ j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = s \right\}) + w(\left\{ j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = r \right\}), \end{aligned}$$

we have

$$w(\left\{ j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = s \right\}) + w(\tilde{H} \setminus C_i) > q.$$

The above result violates Equation (C.1). So, $\sigma$ cannot be an equilibrium rejection. Hence, $\sigma_i(s)$ should be 1.

However, this is true for any $i \in C$. Contradiction. □

**Lemma 14.**
*A weighted majority rule $f \in G(w)$ is interim self-stable among $G(W)$ if, for any individual $i$, there exists a minimal winning coalition including $i$ which is not mutually exclusive with any other minimal winning coalition.*

*Proof.*

Denote $\hat{C}_i$ a minimal winning coalition which is not mutually exclusive with any other minimal winning coalition. Since we are focusing on the case where $w_f = w_g$ for any $g$, it is either $q_f < q_g$ or $q_f > q_g$.

First, consider the case where $q_f < q_g$. So, if $f(t) = S$, then $g(t) = S$. And if $f(t) = R$, then $g(t)$ could be either $S$ or $R$. Suppose a strategy profile $(\sigma, \mu)$ such that $\sigma_i(t_i) = 0$ for all $i$ and $t_i \in T_i$, $\sigma_i^k(s) = k^{-\frac{1}{w_i}}$ and $\sigma_i^k(r) = k^{-\frac{2}{w_i}}$. Pick an agent $i$. Suppose for a minimal winning coalition $C_m$ and a type profile $t_{-i}$, the convergence speed of

$$\lim_{k\to\infty} \left( \prod_{j \in C_m \setminus \{i\}} \sigma_j^k(t_j) \right) \left( \prod_{j \in N \setminus C_m} (1 - \sigma_j^k(t_j)) \right)$$

is slower than for any other minimal winning coalition. By construction, $W(C_m) \geq W(\hat{C}_i)$ and $t_j = s$ for $j \in C_m \setminus \{i\}$. For such $t_{-i}$, $f(t) = S$ if $t_i = s$, and $f(t)$ could be either $S$ or $R$ if $t_i = r$, because $C_m$ has no mutually exclusive minimal winning coalition. Thus the right hand side of Equation (C.4) is weakly less than the left for any type and any agent. This is true for any $i$. The strategy profile $\sigma$ and the derived belief system $\mu$ is an equilibrium rejection of $g$.

Second, consider the case where $q_f > q_g$. So, if $f(t) = S$, then $g(t)$ could be either $S$ or $R$. And if $f(t) = R$, then $g(t) = R$. Suppose a strategy profile $(\sigma, \mu)$ such that $\sigma_i(t_i) = 0$ for all $i$ and $t_i \in T_i$, $\sigma_i^k(s) = k^{-\frac{2}{w_i}}$ and $\sigma_i^k(r) = k^{-\frac{1}{w_i}}$. Pick an agent $i$. Suppose for a minimal winning coalition $C_m$ and a type profile $t_{-i}$, the convergence speed of

$$\lim_{k\to\infty} \left( \prod_{j \in C_m \setminus \{i\}} \sigma_j^k(t_j) \right) \left( \prod_{j \in N \setminus C_m} (1 - \sigma_j^k(t_j)) \right)$$

is slower than for any other minimal winning coalition and type profile. By construction, $W(C_m) \geq W(\hat{C}_i)$ and $t_j = r$ for $j \in C_m \setminus \{i\}$. For such $t_{-i}$, $f(t)$ could be either $S$ or $R$ if $t_i = s$, and $f(t) = R$ if $t_i = r$, because $C_m$ has no mutually exclusive minimal winning coalition. Thus the right hand side of Equation (C.4) is weakly less than the left for any type and any agent. This is true for any $i$. The strategy profile $\sigma$ and the derived belief system $\mu$ is an equilibrium rejection of $g$. □

### A.3.1 Super Majority Rules

**Lemma 15** (Weak altenative)**.**

*Consider a qualified majority rule $f$. If $g$ has a $C^g$ where $n^g \notin C^g$ and $w^f(C^g) \leq q^f$, then there exist an equilibrium rejection of $g$.*

*Proof of Lemma 15.* Set $\sigma_i(t_i) = 0$, $\sigma_i^k(r) = k^{-\frac{1}{w_i^g}}$ and $\sigma_i^k(s) = k^{-\frac{2}{w_i^g}}$. Condition (C.1) and (C.2) are trivially satisfied. Since $\sigma_i^k(r) > \sigma_i^k(s)$ for any $k$ and $i$, and $\sigma_i^k(r) > \sigma_j^k(r)$ for any $i > j$, we have $f(t_{-i}, t_i = r) = R$ and $g(t_{-i}, t_i) = R$ for any $i$ and $t_{-i}$ with $\mu_i(t_{-i}) > 0$. Therefore, Condition (C.4) are satisfied for any $i$ and $t_i$. □

**Lemma 16.** *Consider a qualified majority rule $f$ with $q^f \geq \frac{n-1}{2}$. If $g$ has a $C^f$ where $w^g(N \setminus C^f) + w_i^g \leq q^g$ for any $i \in C^f$, then there exists an equilibrium rejection of $g$.*

*Proof.* Consider $C^f$ with highest weights under $g$. Set $\sigma_i(t_i) = 0$, $\sigma_i^k(r) = k^{-\frac{2}{w_i^g}}$ and $\sigma_i^k(s) = k^{-\frac{1}{w_i^g}}$. Since $\sigma_i^k(r) < \sigma_i^k(s)$ for any $k$ and $i$, and $\sigma_i^k(s) > \sigma_j^k(s)$ for any $i > j$, we have $f(t_{-i}, t_i = s) = S$ and $g(t_{-i}, t_i) = S$ for any $i$ and $t_{-i}$ with $\mu_i(t_{-i}) > 0$. Therefore, Condition (C.4) are satisfied for any $i$ and $t_i$. □

**Lemma 17.** *Consider a qualified majority rule $f$ with $q^f \geq \frac{n-1}{2}$. The followings are equivalent.*

1. *A rule $g$ has a $C^f$ where $w^g(N \setminus C^f) + w_i^g \leq q^g$ for any $i \in C^f$.*
2. *$\bar{C}^f$ with highest weights under $g$ satisfies $w^g(N \setminus \bar{C}^f) + w_i^g \leq q^g$ for any $i \in \bar{C}^f$.*
3. *Consider $\bar{C}^f$ with highest weights under $g$. Then, $w^g(N \setminus \bar{C}^f) + w_{n^g}^g \leq q^g$.*