Essays in Political Economy and Mechanism Design

Vadim Iaralov

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of Economics
Adviser: Roland J. M. Bénabou

September 2013

Abstract

This thesis studies extended models of choice in political economy and mechanism design. In some situations, economic agents' decision problem does not fit within the traditional "economic man" framework of expected-utility maximization.

The first chapter looks at citizens' indirect political opposition to a dictator when the political institutions do not allow open elections and the dictator uses physical force to punish protesters. The first finding states that as the dictator's power is weakening over time, citizens' anticipation of his eventual downfall makes the uncertain time-frame of the revolution happen sooner. Secondly, the dictator is most oppressive when his power is moderate. Thus, the policing is non-monotone in the state: little in good times, progressively more during hardship right up to the tipping point when it is completely withdrawn. The government puts up an intense short-term fight to stay in power, even though the times are changing and its authoritarian grip is loosening.

The second chapter also looks at political choice – this time it is a democracy with loss-averse voters and "career-concerned" politicians. This rational-expectations model confirms the empirical finding that the voters prefer incumbents during good times and take a chance on challengers when experiencing a bad shock. The model is also consistent with the second empirical finding that while incumbent's average disaster relief increases in the magnitude of an unrelated crisis, their average probability of winning decreases. The politician's decision involves a tradeoff between personal rent and increasing the probability of being elected by choosing a higher signal. Therefore, when the voters suffer a loss from a natural disaster, the incumbent cuts his rents and provides more public goods as the electoral race tightens. By combining "career-concerned" incumbents with behavioral voters, the same model can explain both facts, whereas individually these parts are not enough.

The third chapter looks at social choice from a mechanism designer's point of view, where some of the constituents make mistakes under the exclusive information setting. This theoretical chapter derives novel necessary and sufficient conditions for full implementation (matching desirable outcomes to equilibria), even when the faulty players lie about their private information.

Acknowledgements

First, I would like to thank my advisor Roland Bénabou, Stephen Morris, and the participants of seminars here at Princeton for providing feedback and guidance. I would also like to thank the Economics department for providing financial support during my studies as well as the SSHRC Doctoral Fellowship Award #752-07-1421. Finally, I would like to thank my wife, Mary, without whom this dissertation would not possible.

To my wife, Mary.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis studies extended models of choice in political economy and mechanism design. In some situations, economic agents' decision problem does not fit within the traditional "economic man" framework of expected-utility maximization.

The first chapter considers a coordination game of small and short-lived players that interact with a large, long-lived player that prefers one of their actions. This is a dynamic model where citizens choose to supply labor and the government chooses its policing level. In the stationary case with extreme police (un)productivity, the government's action produces a unique outcome. For intermediate productivity, there are multiple equilibria. If policing productivity is expected to decline in the future (doesn't matter how slowly), then most – but not all – of the indeterminacy is resolved. Forward-looking unraveling argument has the government give up for most of the intermediate region just as for the low levels (in the future) but right before it does, it polices harsher than ever before or after. Thus, the policing is non-monotone in the state: little in good times, progressively more during hardship right up to the tipping point when it is completely withdrawn. This mirrors how the government may put up an intense short-term fight to stay in power, even if it's doomed in the long-run as the times are changing and its authoritarian grip is loosening.

The second chapter allows for elections but voters are no longer expected-utility maximizers as above. The political science literature has identified a salient phenomenon known as *incumbency advantage*, where politicians in office stand a higher chance of being reelected than challengers vying for the same seat. Secondly, more recent research has described the opposite circumstance of *incumbency disadvantage* when challengers do better in bad times after an exogenous shock to the economy, which is unrelated to the government's actions. The third related stylized fact is that while incumbent's average disaster relief increases in the magnitude of the (exogenous) crisis, their average probability of winning decreases. In other words, an average incumbent wins more often when he is lucky to avoid an unrelated crisis and loses more often when he is not lucky, while providing the expected disaster relief for that particular crisis.

This paper develops a model to explain all of these facts by linking politicians with "career concerns" and forward-looking voters with reference-dependent, loss-averse utility. The personal equilibrium of Koszegi and Rabin (2007) is applied to voters who rationally expect their own reference point formed by their future rational voting decision. With an S-shaped value function, they are risk-seeking in the losses region and risk-averse in the gains region. If the incumbent represents the continuation of the status quo and the challenger is a risky gamble, then the incumbent should tend to get more support, except during bad times (with risk-seeking to attempt recouping losses). As incumbent's probability of losing rises, he increases spending by matching the marginal benefit of winning more by appearing more talented, against the desire for personal rents.

The third and final chapter studies a theoretical problem of a mechanism designer who wants to create a mechanism with only desirable outcomes of its equlibria, while having an equilibrium for each desirable outcome. Whether this is possible when some players may be irrational (faulty) and possess private information depends on

the properties of the social choice set: some sets are fully implementable and some are not. Specifically, this model looks at implementation under incomplete information in general environments of Jackson (1991) but with a robust notion of k-fault tolerant equilibrium of Eliaz (2002). The environment may be non-economic and allows for exclusive information and up to k players could be making mistakes. Assuming closure on the socially desirable set, a new condition, called, k-Incentive Compatibility is found to be both necessary and sufficient for partial implementation. When the desirable set also satisfies k-Monotonicity-no-veto, which is a combination of k-no-veto hypothesis and k-Bayesian Monotonicity, then the desired set can be fully implemented.

Chapter 2

Protest Dynamics in a Police State

There is no denying that protests and revolutions are an important force that shapes society. The collapse of the Berlin Wall in 1989 and the more recent Arab Spring of 2011 were both sudden and unexpected, an upheaval following a reasonably stable period in their respective societies. Why is revolution spontaneous and surprising and can the authoritarian government do something about it? Supposing the citizens take into account the likely future repercussions for their protests, how does the timing of the government response affect the evolution of protests over time?

Protesting against a totalitarian government has a distinct chronological friction. While the government may fall in the future, it is in power today and may punish its opposition with violence. This anticipation of future actions taken by the citizens and the strategic government affects equilibrium play. Focusing on a dynamic story instead of an informational one allows comparing the short-run and the long-run predictions of a possible revolution. In the short-run there is greater uncertainty about whether a government with moderate police productivity can successfully deter its opposition from starting a revolution. With a long-run view of a gradual decline, the government is much more likely to fail at an earlier time. There comes a point when rational anticipation of its fall eliminates any optimistic beliefs about the gov-

ernment surviving. Moreover, the police state finds it more costly to put down brazen opposition.

The model also predicts the government to match its policing levels to the regime's political outlook. There is little policing in good times, more during hardship at the tipping point before it is completely withdrawn. This reflects that the government may put up resistance to stay in power, even if it will inevitably collapse in the long-run because of a structural decline. The government polices to delay the inevitable and buy itself some time in power, while it still can afford the necessary expense.

This paper considers a (full-information) coordination game of small and short-lived players that interact with a large, long-lived player that prefers one of their actions. The application at hand has citizens with symmetric preferences for coordinating on two outcomes of work and protest, while incurring a (fixed) personal cost when working and a punishment when protesting. This is a dynamic model where citizens choose to supply labor – acquiesce to the oppressive regime or protest (rebel) against it, based on the state of the economy (high or low current labor force). The government is a strategic agent who strictly prefers coordination on the "work" outcome. When the government has the means of sufficiently productive police, its costly action can force coordination on its preferred outcome of work.

Here the focus is on police productivity as the key state variable affecting the evolution of protests. It measures how effective the government is at converting its budget into punishment, a disutility of protest. First, police productivity is taken as a constant parameter and changing it affects the set of equilibria. This gives a short-run analysis of potential protests and revolution. Later, police productivity is going to be described by a deterministic (downward) trend. This gives a long-run analysis of how anticipation of the eventual fall of the government brings about a certain revolution. This revolution will happen at an earlier time than is likely in the

short-run model without aligned expectation of its fall. Still, the exact date of the revolution is unpredictable and may vary for different equilibrium paths.

In the stationary case with high police productivity, the government's costly action can force coordination on its preferred outcome of work, retaining power. For the stationary case with low police productivity, the government cannot afford to cover individual citizen's cost of work and there are always protests and never policing, so the government loses control to the rebel opposition.

For moderate productivity, there are multiple equilibria types, which stems from citizen's self-reinforcing behavior when a sizable fraction moves simultaneously. Citizens may coordinate on different policing thresholds because their individual deviations don't have the full force as they are small and their individual labor choice doesn't change the total. If everyone else is expected to work, a citizen would require a small policing presence, which the government can afford under moderate productivity and, thus, polices as required. The citizens will keep on working for two reasons: they like to work when others do and, furthermore, more importantly, anticipate policing to continue in the future because a small police force will also be affordable later.

However, if everyone else is expected to protest, the same citizen would require a large policing presence to stay at work, which the government cannot afford under moderate productivity. The citizens will keep protesting because they would rather protest when others do but also because they are unlikely to face policing in the future because it would have to be similarly large and likely unaffordable. At the same time, moderate police productivity implies the government optimally enforces in equilibria with low citizen's policing thresholds and, thus, the citizens work. However, for other equilibria the government faces high policing thresholds that it cannot afford. Therefore, the government gives up and citizens protest in some cases, which makes these high thresholds rational. Unlike the static models with multiplicity, slightly

more can be said here. For the case of upper-moderate productivity, the government will always police in at least the "high" state when the old are already working but there is indeterminacy when the old are protesting. The young citizens can at least coordinate on working with today's old, which means less policing is required. The opposite is true for the lower-moderate productivity: there are always protests in the "low" state when the old are protesting as minimum policing requirement is high relative to the police productivity.

If police productivity is expected to decline in the future (doesn't matter how slowly), then most – but not all – of the indeterminacy is resolved. Forward-looking unraveling argument has the government give up in both states for most of moderate productivity levels just as it does for the low levels (in the future). Interestingly, just before the government gives up, it polices harsher than before or after. Thus, the policing is non-monotone in the productivity: little in the early stable period, progressively more during hardship right up to the tipping point when it is completely withdrawn.

Long before the revolution, the citizens had expected a stable period of autocratic rule with no chance for revolution. The threat of punishment was credible because the government's police was very effective under the assumption of a downward trend. Secondly, it didn't need to police a lot in a given period because every potential rebel had realized they would be punished for two periods and, worse yet, they would be rebelling alone. On the last period of the government's rule, everyone knows that there will be no policing next period. Therefore, the police essentially has to exert two periods worth of punishment plus offset tomorrow's utility of coordinating with tomorrow's (protesting) young. Policing anything less and the revolution would have happened right there and then, contradicting the hypothesis of it being the last period of the government's power. This rise in policing is intimately linked with the

unraveling argument because this required increase is impossible when productivity is moderately low and multiplicity is then resolved to always revolt.

This mirrors how the government may put up an intense short-term fight to stay in power, even if it is doomed in the long-run as the times are changing and its authoritarian grip is loosening. Lenin was intially arrested in Imperial Russia in 1895 and sent into exile for spreading revolutionary literature. Then the Revolution of 1905 was put down by the military using artillery against the textile district in Moscow, killing over a thousand rebel workers. By February 1917 the Tsar could no longer suppress workers' strikes with military force as the soldiers sympathized with the protesters and mutiny occurred.[1]

While the transition is inevitable, the indeterminate length of the transition in the short-run may be instant or maybe prolonged. There is multiplicity in the short-run transition paths taken – even the top revolutionaries are often surprised how fast or slow the revolution actually happens.

The classic papers on the theory of protest highlighted that there may be multiple equilibria and the actual timing of a protest or revolution is unexpected, even by the opposition. Kuran (1991) documents everyone's surprise at the Berlin Wall falling when it did. The early models tended to be static such as Kuran (1989), which focused on supporters of the opposition falsifying their preferences until it was clear that they were going to win. This could be thought of as the citizens' preference to coordinate to be on the winning side. The equilibrium outcomes were fragile to small changes in distribution of private preferences. While it talks about revolution being the "inevitable outcome of a long period of gestation," it misses out how this anticipation affects the revolution process itself. Secondly, it doesn't let the autocratic government, an interested party to be sure, to act strategically in its own self-interest.

[1]See Service (2009) and Pipes (1996) for detailed historical accounts of the Russian revolutions.

These models ignore the effects of potential government interference with protests and the anticipated revolution.

Yin (1998) looks at how equilibria in a threshold model of turnout with heterogeneous agents vary across different families of threshold distributions. A "threshold" here is simply the minimal fraction of the population who must protest before a given agent chooses to protest. The allowed government policies are comparative statics on the parameters that describe a given distribution within its family. For example, a government that is more popular reduces discontent and increases the average threshold of protest. Alternatively, a government that alienates itself from social forces reduces integration and increases dispersion of the distribution. Here the option of physical deterrence is framed in terms of reducing government's popularity (increasing discontent), while intimidating protestors and may backfire when used against the wrong kind of challenger.

Lohmann (1994) looks at informational frictions involved in protesting as a costly signaling of private experiences between differently informed agents about the regime policies. It notes that the actual turnout relative to the expected turnout provides information about regime's vulnerability, though the government is simply a passive participant. Here, individuals who take a political action at a private cost are publicly observed and influence followers' subsequent moves. Similarly, Acemoglu and Jackson (2011) looks at how "leadership" by publicly-observable prominent agents can create coordination on a unique outcome in an overlapping generation repeated game, though with focus on social norms rather than political economy of protests. Just like in the present paper, the current young's single action will coordinate with today's old and tomorrow's young. However, one difference is their paper has a representative agent whose action is guaranteed to move the state, which is important for incentives of public leaders anticipating tomorrow's young action to align to their own benefit. In contrast, the current paper focuses on small citizens that take the

sequence of states as given in any Markov Perfect Equilibrium, which generates additional within-period multiplicity as agents can find it optimal to demand various policing levels inside an interval as long as everyone else in the current period does and deviates otherwise.

Another informational model by Edmond (2011) allows for a strategic government to manipulate quality and quantity of information through propaganda. On one hand, the innovation of centralized mass-media like newspapers and television makes it easier for the government to stay in power, but on the other hand, the more decentralized social networks make it more difficult to prevent protest through a relative increase in informational reliability. This model emphasizes informational rather than time frictions as it studies propaganda and signal filtering rather than relationship between anticipation and dynamic evolution of play.

Like the present paper, Cho and Matsui (2005) also studies a repeated game of asymmetric moves but focuses on the private sector (a single representative agent) that coordinates with the government on inflation-setting and its expectation. The private sector isn't coordinating with itself, though - only with the government's last action and tomorrow's action. The idea to use a time-varying fundamental to reduce equilibrium multiplicity was used by Burdzy, Frankel, and Pauzner (2001). The anticipation of future play with locked-in actions had them focus on a risk-dominant outcome. The present paper introduces variation in small player payoffs through equilibrium actions of a large strategic player, rather than exogenous shocks. Even if the players' own costs of work are fixed, they may still anticipate their endogenous cost of protest to vary in the future because the large player's incentives change. This makes revolution happen sooner without completely pinning down its timing, which would go against observers' surprise at the collapse of the Berlin Wall as was extensively documented by Kuran (1991).

The rest of the paper is structured as follows. Section 2.1 presents a simple static model which will be the stage game in the subsequent dynamic framework. Section 2.2 repeats the stage game in a dynamic model with a stationary police productivity. Section 2.3 introduces a downward trend in police productivity and Section 2.4 concludes. Finally, Appendix B contains some of the proofs from the main text.

2.1 Simplified Model (one-shot)

Consider a static, one-period model that will highlight some of the flavor of later results in a simpler setting. We will find that the government's policing is non-monotone in citizen's cost of work for different equilibria. One limitation of the static model is that it doesn't capture the spontaneity and turbulence of revolution. There are no interactions via expectations for adjacent states - in the static model these belong to different equilibria. On the other hand, in the dynamic model with a trend, knowing that the government will eventually fall can coordinate expectations against it much sooner. Knowing that, the government may have to increase policing before revolution to keep agitated citizens working. Such increase would push the timing of the revolution closer to the present because with declining productivity, the government wouldn't be able to afford it in the future when it may have survived with optimistic citizens.

While the static model doesn't capture the dynamic interactions, the setup and the solution of the stage game is illustrative of the steps taken to solve the repeated game.

The government observes the fundamental state $\theta \in [0, \infty)$, which is publicly known, and represents citizens' cost of work[2]. Next, the government commits to a policing level $p(\theta) \in [0, \infty)$.

[2]In the later, more general model this will be denoted as fixed parameter B.

After observing that the government has already committed to some policing level p and the fundamental is θ, measure 1 of citizens pick an action $a \in \{0, 1\}$ where $a = 0$ represents "protest" or "joining the opposition" and $a = 1$ represents "work."

Focusing on the symmetric pure strategies, aggregate choice $a \in \{0, 1\}$ can be thought of as the labor force. For simplicity of exposition we will focus on a subset of symmetric, pure-strategy Subgame Perfect Equilibria where citizen's strategy is a cutoff (c_θ).

$$\hat{a}_\theta(p) = \begin{cases} 1 & \text{if } p \geqslant c_\theta, \\ 0 & \text{if } p < c_\theta, \end{cases} \tag{2.1.1}$$

Citizens prefer work relative to protest more when policing rises as they prefer to avoid pain. They also work more when labor force rises because they prefer to conform or because the cost of repression is higher for smaller crowd of remaining protestors. Let the relative preference for work over leisure, given labor force aggregate L and policing p, be denoted as

$$\Delta u(L, p; \theta) = u(1, L, p; \theta) - u(0, L, p; \theta) = \alpha L + p - \Theta \tag{2.1.2}$$

The parameter α is the measure of social cohesion (strategic complementarity), how strong the preference for conformity is and θ is a cost of working (preference for leisure).

Government's payoff increases in the labor force (less unrest, more taxes - not modeled) and decreases in the police force (police and justice department budgets are costly).

$$g(L, p) = L - \frac{1}{\gamma} p : \{0, 1\} \times [0, \infty) \to \mathbb{R}, \tag{2.1.3}$$

where $\frac{1}{\gamma}$ is the marginal cost of policing and γ is a measure of policing productivity which is high when policing cost is low.

A pair of strategies $(c_\theta, p^*(\theta))$ form a Subgame-Perfect Equilibrium when they have no profitable deviations in every state. While the government faces state θ, the citizens face state (θ, p).

An individual citizen recognizes the equilibrium labor-force in state (θ, p) to be $L = \mathbb{1}_{p \geqslant c_\theta}$. They find it optimal to work if and only if $\Delta u(L, p; \theta) \geqslant 0$ which happens if and only if

$$p \geqslant \theta - \alpha L = \theta - \alpha \mathbb{1}_{p \geqslant c_\theta} \tag{2.1.4}$$

It can't be the case that the citizen finds it optimal to protest for any policing level (even out-of-equilibrium) above the cutoff strategy, $p \geqslant c_\theta$ as that would violate Subgame-Perfection. Equation (2.1.4) becomes a restriction on the equilibrium cutoff strategy:

$$c_\theta \geqslant \theta - \alpha \tag{2.1.5}$$

Similar considerations give another restriction to prevent citizen from deviating to work when everyone protests in some state below the cutoff with $p < c_\theta$:

$$c_\theta \leqslant \theta \tag{2.1.6}$$

Combining equations (2.1.5) and (2.1.6), c_θ satisfies *collectively-sustained best-response* (BR) if and only if

$$c_\theta \in [\theta - \alpha, \theta] \tag{2.1.7}$$

As explained above, $c_\theta > \theta$ violates equation (2.1.6) because if policing p satisfies $c_\theta > p > \theta$, then each citizen finds it optimal to work and deviates from equilibrium-prescribed protest. Similarly, $c_\theta < \theta - \alpha$ violates equation (2.1.5) because if policing

p satisfies $c_\theta < p < \theta - \alpha$, then each citizen finds it optimal to protest and deviates from equilibrium-prescribed work.

The government takes citizen's cutoff strategy c_θ as given. Thus, government's optimal choice maximizes:

$$g(L, p) = L - \frac{1}{\gamma}p \tag{2.1.8}$$

The optimal policing level turns out to be either zero or equal to the citizen's cutoff c_θ.

$$p(\theta) = \arg\max_p \{\mathbb{1}_{p \geqslant c_\theta} - \frac{1}{\gamma}p\} \in \{0, c_\theta\} \tag{2.1.9}$$

Observe that $p(\theta) = c_\theta$ if and only if $1 - \frac{c_\theta}{\gamma} \geqslant 0$ if and only if $c_\theta \leqslant \gamma$ and, otherwise, $p(\theta) = 0$ if and only if $c_\theta > \gamma$.

Thus,

$$p(\theta) = \begin{cases} c_\theta & \text{if } c_\theta \leqslant \gamma, \\ 0 & \text{if } c_\theta > \gamma, \end{cases} \tag{2.1.10}$$

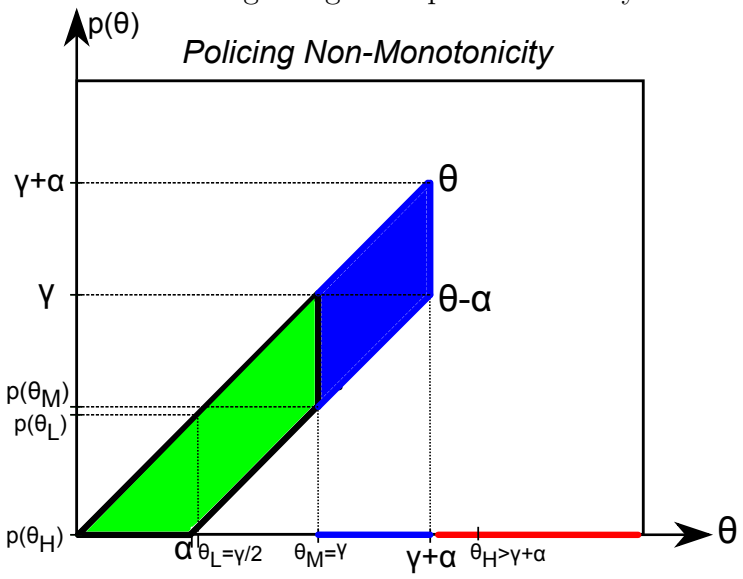
Assumption 2.1. *: $\gamma > 2\alpha$, so social cohesion isn't too great.*

The purpose of this assumption is to ensure $p(\gamma) > p(\alpha)$ for the next proposition. It also ensures $p(\theta) \equiv 0$ is not an equilibrium outcome for all $\theta \geqslant 0$. In particular, the proof of the following proposition will establish that in every SPE, there has to be a positive police level at $\theta = \frac{\gamma}{2}$:

$$p^*(\gamma/2) = c_\theta \geqslant \theta - \alpha = \frac{\gamma}{2} - \alpha > 0. \tag{2.1.11}$$

Proposition 2.2 (Non-Monotonicity)**.** *There exist costs of work $\theta_L \leqslant \theta_M \leqslant \theta_H$: for any equilibrium selection picking arbitrary SPE $(p^*(\theta), c^*_\theta)$ for each state, the policing*

Figure 2.1: One-shot game gives a preview of a dynamic result



in these states is non-monotone and satisfies

$$p^*(\theta_H) < p^*(\theta_L) < p^*(\theta_M)$$

Proof. See Appendix. □

For every cost of work θ, a Subgame-Perfect Equilibrium $(p^*(\theta), c^*_\theta)$ has *Never Revolt* at $\hat{\theta}$ when the equilibrium labor supply is 1 (full employment, no protests) because $p^*(\hat{\theta}) = c_{\hat{\theta}}$. Likewise, a Subgame-Perfect Equilibrium has *Never Revolt* at $\hat{\theta}$ when the equilibrium labor supply is 0 (no employment, everyone protests) because $p^*(\hat{\theta}) = 0 < c_{\hat{\theta}}$.

Proposition 2.3. *1. If $0 \leqslant \theta \leqslant \gamma$, then all equilibria have Never Revolt (NR) at θ and policing satisfies*

$$\theta - \alpha \leqslant p^*(\theta) = c^*_\theta \leqslant \theta$$

2. If $\theta > \gamma + \alpha$, then all equilibria have Always Revolt (AR) at θ and policing satisfies

$$p^*(\theta) = 0 < \theta - \alpha \leqslant c^*_\theta \leqslant \theta$$

3. If $\gamma < \theta \leqslant \gamma + \alpha$, then both NR and AR are attained in different equilibria at θ.

Proof. See Appendix. □

The main idea of Proposition 2.2 will be recreated for the dynamic case in Theorem 2.31. They both say that when the government police is productive relative to the cost of work it enforces, then there is a moderate amount of policing. As the cost of work rises in Proposition 2.2, while keeping police productivity fixed (alternatively: as the government police gets less productive in Theorem 2.31, while cost of work is fixed), policing first increases and at some point when policing is so unproductive it's useless, then policing stops completely (abruptly, rather than smoothly). While the basic results are similar, the mechanism is different. In the one-shot model, cost of work increases exogenously and continuously, so slightly more policing today keeps the citizens indifferent between work and protest at their threshold cutoff. On the other hand, in the dynamic case what's changing is that on the period before the revolution begins (i) the tomorrow's old protest which discourages work today, (ii) continuation utility of receiving policing tomorrow becomes zero since the government gives up, also reducing payoff to work. These two factors cause a discontinuous drop in relative utility of work, so today's policing needs to be higher by a "jump" to compensate.

The same two factors also have a qualitative effect on the equilibrium set. Taking the dynamic model with the stationary states as a baseline and then adding cascading endogenous anticipation resolves multiplicity for some states adjacent to the dominance region. For example, in the stationary region with low-moderate police productivity it is possible to sustain multiple equilibria (at least Traditional Play and Always Revolt) that rely on self-fulfilling beliefs about the future coordination (Proposition 2.18). Once we introduce anticipation of eventual and certain (no matter how far in the future) deterioration of police productivity, today's equilibrium path

gets uniquely resolved into Always Revolt by contagion of dominance (Proposition 2.29).

2.2 Fundamentals of the Repeated Game

This section is going to make the first step towards a dynamic version of the one-shot model - the citizens will live for two periods, not only playing a coordination game with today's young but also with today's old (yesterday's young) and tomorrow's young (when they're themselves old). Secondly, the government's policing problem has a substitution trade-off between tomorrow's policing and today's policing, though Markov Perfection will be used to pin down tomorrow's equilibrium choice, so time inconsistency problem doesn't arise directly[3].

The government is a long-lived player with $0 < \delta < 1$ discount and citizens are short-lived, assumed to live for two periods with $0 < \beta < 1$ discount. There is a measure $\frac{1}{2}$ of citizens that are born every period and commit to an action $a \in \{0, 1\}$ for both periods, where $a = 0$ represents "protest" or "joining the opposition" and $a = 1$ represents "work." Citizens are "Young" when they are born and decide their action and are "Old" when they are stuck playing what they chose last period.

At the beginning of period t, the government observes the average action of the Old a^o before picking a policing level $p_t \in [0, \infty)$. Focusing on the symmetric pure strategies, $a^o \in \{0, 1\}$. Then the Young are born and they observe both (a^o, p_t) before picking work or protest $a \in \{0, 1\}$. The labor force is the total amount of work done

[3]Time-consistency is achieved through matching particular cutoffs. When the government is less patient than the citizens, then among all NR equilibria with full employment, the most preferred equilibrium has the government commit to "maximum" (in a certain sense) policing tomorrow and every other period after by having maximum (pessimistic) citizen's cutoff, as if government was giving up its bargaining power

by the Old and the Young combined,

$$L_t = (1/2)a^o + (1/2)a^y. \tag{2.2.1}$$

The labor force is restricted to $\{0, \frac{1}{2}, 1\}$ for symmetric equilibria in pure strategies.

Government's Markov strategy in state a, the Old's average work level, is the policing level:

$$p(a) : \{0, 1\} \to [0, \infty) \tag{2.2.2}$$

(Young) citizen's Markov strategy in state (a, p), which is the Old's work level and government policing level, is the choice between protest and work:

$$\hat{a}(a, p) : \{0, 1\} \times [0, \infty) \to \{0, 1\} \tag{2.2.3}$$

For simplicity of exposition we will focus on a subset of symmetric, pure-strategy Markov Perfect Equilibria where citizen's strategy is a cutoff (c_0, c_1).[4] Citizen works when policing in state $a \in \{0, 1\}$ exceeds c_a and protests otherwise.

$$\hat{a}(a, p) = \begin{cases} 1 & \text{if } p \geqslant c_a, \\ 0 & \text{if } p < c_a, \end{cases} \tag{2.2.4}$$

At the end of period t, payoffs are realized based on (L, p), which is current labor and policing. Citizens prefer work relative to protest more when policing rises as they prefer to avoid pain. They also work more when labor force rises because they prefer to conform or because the cost of repression is higher for smaller crowd of remaining

[4]For each cutoff $c-$strategy, there is a family $\mathcal{S}(c)$ of strategies that are the same for $p \in [0, c_a]$ but possibly equal to 0 on an open set $p \in (c_a, c_a + \epsilon)$ for some ϵ and equal to 1 for greater p. The behavior above c_a relies on out-of-equilibrium calculation but may be consistent because of coordination. Using $\tilde{c} \in \mathcal{S}(c)$ doesn't change the results because of Lemma 2.10.

protestors. Government's payoff increases in the labor force (less unrest, more taxes - not modeled) and decreases in the police force (police and justice department budgets are costly).

For simplicity of exposition, government's one-period payoff is assumed to be linear in labor and policing:

$$g(L,p) = L - \frac{1}{2\gamma}p : \{0, \frac{1}{2}, 1\} \times [0,\infty) \to \mathbb{R}, \tag{2.2.5}$$

where $\frac{1}{2\gamma}$ is marginal cost of policing and γ is a measure of policing productivity which is high when policing cost is low.

Government is a long-lived agent and receives normalized discounted total payoff of

$$(1-\delta)\sum_{t=0}^{\infty}\delta^t(L_t - \frac{1}{2\gamma}p_t) \tag{2.2.6}$$

Citizen's one-period payoff for choosing a, when total labor force is L and policing is p, is denoted by

$$u(a,L,p) : \{0,1\} \times \{0, \frac{1}{2}, 1\} \times [0,\infty) \to \mathbb{R}. \tag{2.2.7}$$

Citizen born at t receives total utility for playing a_t as their current one-period payoff plus discounted tomorrow's payoff for playing a_t as well:

$$u(a_t, L_t, p_t) + \beta u(a_t, L_{t+1}, p_{t+1}) \tag{2.2.8}$$

Next, we will impose a linearity assumption on citizen's payoffs as follows. Let the relative preference for work over leisure within one period, given (L,p), be denoted

as

$$\Delta u(L,p) = u(1,L,p) - u(0,L,p) = \alpha L + p - B \tag{2.2.9}$$

where α, B are parameters. α is the measure of social cohesion (strategic complementarity), how strong the preference for conformity is and B is a cost of working (preference for leisure).

Assumption 2.4. $0 < \alpha < B$. *The cost of work exceeds gains from complete coordination on work without policing (unpopular dictator).*[5]

We can make the following simple observations by using linear functional forms for flow preferences of citizens and the government.

Observation 2.5. $\Delta u(L,p) = \alpha L + p - B$. *It is monotonically increasing in L (preference for conformity).*

Observation 2.6. $\Delta u(L,p) = \alpha L + p - B$ *is monotonically increasing in p (preference for avoiding pain).*

Observation 2.7. $g(L,p) = L - \frac{1}{2\gamma}p$ *is monotonically increasing in L (government is more popular, country more productive).*

Observation 2.8. $g(L,p =)L - \frac{1}{2\gamma}p$ *is monotonically decreasing in p (costly budgets).*

2.2.1 Markov Perfect Equilibrium conditions

The solution concept used is Markov Perfect Equilibrium in pure strategies. A pair of Markov strategies (a^*, p^*) are MPE when they withstand one-shot deviation in every state. Optimality on off-equilibrium path will be relevant for citizen's cutoff choice. Citizens consider facing arbitrary policing levels to which the government has

[5]This assumption will be discussed in greater detail in Section 2.2.2.

previously committed to on its turn, not simply specific equilibrium quantity (p_0^*, p_1^*) and ensuring the citizen does indeed protest in all states below the cutoff and works for all states above the cutoff.

The government moves first and its policing function, $p(a)$, depends only on the observed old's action, a. $p^*(a) : \{0,1\} \to [0,\infty)$. The young citizen moves after observing government's choice as well as the old's action and its cutoff strategy is $a^*(a,p) = \mathbb{1}_{\{p \geqslant c_a^*\}} : \{0,1\} \times [0,\infty) \to \{0,1\}$ Government receives utility from p^* at state a, taking the citizen cutoff strategy (c_0, c_1) as given, as follows

$$G(a|p^*) = (1-\delta)\left(\frac{a}{2} + \frac{\mathbb{1}_{\{p_a^* \geqslant c_a\}}}{2} - \frac{1}{2\gamma}p_a^*\right) + \delta G(\mathbb{1}_{\{p^* \geqslant c_a\}}|p^*) \tag{2.2.10}$$

Denote government utility from **one-shot deviation** to $\tilde{p} \in [0,\infty)$ and later going back to p^* as

$$\tilde{G}(a|p^*) = (1-\delta)\left(\frac{a}{2} + \frac{\mathbb{1}_{\{\tilde{p} \geqslant c_a\}}}{2} - \frac{1}{2\gamma}\tilde{p}\right) + \delta G(\mathbb{1}_{\{\tilde{p} \geqslant c_a\}}|p^*) \tag{2.2.11}$$

Taking (a^*) as given, government's choice p_a^* is optimal for every $a \in \{0,1\}$:$\forall \tilde{p} \in [0,\infty)$:

$$(1-\delta)g\left(L^*, p_a^*\right) + \delta G\left(\mathbb{1}_{\{p_a^* \geqslant c_a\}}|p^*\right) \geqslant (1-\delta)g\left(\tilde{L}, \tilde{p}\right) + \delta G\left(\mathbb{1}_{\{\tilde{p} \geqslant c_a\}}|p^*\right) \tag{2.2.12}$$

Expanding the payoff functions and simplifying, the government does not benefit in any state $a \in \{0,1\}$ from a one-shot deviation today to $\tilde{p}$ from p^* :

$$\begin{aligned}(1-\delta)\left(\frac{\mathbb{1}_{\{p_a^* \geqslant c_a\}}}{2} - \frac{p_a^*}{2\gamma}\right) + \delta G\left(\mathbb{1}_{\{p_a^* \geqslant c_a\}}|p^*\right) \\ \geqslant (1-\delta)\left(\frac{\mathbb{1}_{\{\tilde{p} \geqslant c_a\}}}{2} - \frac{\tilde{p}}{2\gamma}\right) + \delta G\left(\mathbb{1}_{\{\tilde{p} \geqslant c_a\}}|p^*\right)\end{aligned} \tag{2.2.13}$$

Citizens are "small" players, who individually cannot move tomorrow's state, which is the next period's Old labor contribution. They are followers in a Stackelberg repeated subgame where the government leads with some p_a. Therefore, there is an aggregate best-response strategy that is played, so that citizens don't have incentive to unilaterally deviate from it.

Suppose that in state (a, p) a young citizen will face p policing today and p' policing tomorrow as well as L labor force today and L' labor force tomorrow. The difference in a young citizen's total utility from choosing to work instead of protest today is the following:

$$\begin{aligned}\Delta U &= (u(1, L, p) + \beta u(1, L', p')) - (u(0, L, p) + \beta u(0, L', p')) \\ &= \Delta u(L, p) + \beta \Delta u(L', p') \end{aligned} \tag{2.2.14}$$

Note that the young citizen's unilateral choice doesn't affect the state, the transition path of the labor force or the policing levels in either period. The citizen works when

$$\Delta u(L, p) + \beta \Delta u(L', p') \geqslant 0 \tag{2.2.15}$$

and the citizen protests when

$$\Delta u(L, p) + \beta \Delta u(L', p') < 0. \tag{2.2.16}$$

Next we will state conditions on other citizen's c^*-strategy cutoff for MPE. In each state (a, p) and taking government strategy $\hat{p}$ as given, each citizen playing $\mathbb{1}_{\{\hat{p}_a \geqslant c_a^*\}}$ needs to be a best-response to other citizens playing the same c^*-strategy both periods and government playing $\hat{p}$ next period.[6]

[6] At the time of Young citizen's move, the observed p today is already fixed and need not derive from $\hat{p}$ as MPE conditions require optimality in all states.

Suppose we are in state (a, p) with other citizens following c^*-strategy prescribing work ($c_a^* \leqslant p$) and government following $\hat{p}$-strategy. An individual citizen also *prefers to work* over protest if:

$$\Delta U = \Delta u\left(\frac{a}{2} + \frac{1}{2}, p\right) + \beta \Delta u\left(\frac{1}{2} + \frac{1}{2}\mathbb{1}_{\{\hat{p}_1 \geqslant c_1^*\}}, \hat{p}_1\right) \geqslant 0 \tag{2.2.17}$$

Expanding the payoff functions, noting today's young and tomorrow's old work and simplifying, we get:

$$\forall a \in \{0,1\}, \forall p \geqslant c_a^* : p \geqslant (1+\beta)B - \alpha\left(\frac{1+\beta}{2} + \frac{a}{2} + \frac{\beta \mathbb{1}_{\{\hat{p}_1 \geqslant c_1^*\}}}{2}\right) - \beta \hat{p}_1 \tag{2.2.18}$$

The above condition on work, $\hat{a} = 1$, to be a collectively sustained citizen best-response holds for all $p \geqslant c_a^*$ if and only if

$$\forall a \in \{0,1\}, c_a^* \geqslant (1+\beta)B - \alpha\left(\frac{1+\beta}{2} + \frac{a}{2} + \frac{\beta \mathbb{1}_{\{\hat{p}_1 \geqslant c_1^*\}}}{2}\right) - \beta \hat{p}_1 \tag{2.2.19}$$

To describe (2.2.19) condition, define the following auxiliary function:

$$\underline{p}(p_1, a, a_1) \equiv (1+\beta)B - \alpha\left(\frac{1+\beta}{2} + \frac{a + a_1\beta}{2}\right) - \beta p_1 \tag{2.2.20}$$

In each state $a \in \{0,1\}$, if others use c_a^* and government uses p_a^*, it is optimal to work $\forall p \geqslant c_a^*$:

$$\boxed{\forall a \in \{0,1\}, c_a^* \geqslant \underline{p}(\hat{p}_1, a, \mathbb{1}_{\{\hat{p}_1 \geqslant c_1^*\}})} \tag{2.2.21}$$

Suppose in state (a, p) with other citizens following c^*-strategy prescribing protest ($c_a^* > p$) and government follows $\hat{p}$-strategy. An individual citizen also *prefers to*

protest over work if:

$$\forall a \in \{0,1\}, \forall p < c_a^*, \Delta U = \Delta u\left(\frac{a}{2}, p\right) + \beta \Delta u\left(\frac{1}{2}\mathbb{1}_{\{\hat{p}_0 \geqslant c_0^*\}}, \hat{p}_0\right) < 0 \tag{2.2.22}$$

Simplifying, noting today's young protest and so tomorrow's old protest, we get:

$$\forall a \in \{0,1\}, \forall p < c_a^* : \alpha\left(\frac{a}{2} + \frac{\beta \mathbb{1}_{\{\hat{p}_0 \geqslant c_0^*\}}}{2}\right) + p + \beta \hat{p}_0 < (1+\beta)B \tag{2.2.23}$$

The above condition on protest being a collectively sustained best-response holds if and only if

$$\forall a \in \{0,1\}, c_a^* \leqslant (1+\beta)B - \beta \hat{p}_0 - \alpha\left(\frac{a}{2} + \frac{\beta \mathbb{1}_{\{\hat{p}_0 \geqslant c_0^*\}}}{2}\right) \tag{2.2.24}$$

To describe (2.2.24) condition, define the following auxiliary function:

$$\overline{p}(p_0, a, a_0) \equiv (1+\beta)B - \beta p_0 - \alpha\left(\frac{a + a_0\beta}{2}\right) \tag{2.2.25}$$

In each state $a \in \{0,1\}$, if others use c_a^* and government uses p_a^*, it is optimal to protest $\forall p < c_a^*$:

$$\boxed{\forall a \in \{0,1\}, c_a^* \leqslant \overline{p}(\hat{p}_0, a, \mathbb{1}_{\{\hat{p}_0 \geqslant c_0^*\}})} \tag{2.2.26}$$

Compare the α coefficient on the RHS of (2.2.19) when young citizens coordinate on work and RHS of (2.2.24) when young citizens coordinate on protest. In the former case, citizen derives coordination utility of $\frac{1}{2}$ from working alongside with 1/2 population of the young today and $\frac{\beta}{2}$ from working with 1/2 population of the old tomorrow. In the later case, today's young and tomorrow's old protest instead because $p < c_a^*$.

2.2.2 Characterizing Government's Best Response

Assumption 2.9. *Police productivity γ is constant over time.*

For the purposes of generating benchmark equilibrium sets in Section 2.2, Assumption 2.9 fixes γ within the scope of each game. In each case, each equilibrium set is parametrized by γ. In contrast, Section 2.3 will relax this assumption and let γ vary over time, focusing on decreasing police productivity.[7] One important implication is that under constant γ, time t is not a payoff-relevant state variable for the purposes of MPE in Section 2.2. However, observing a publicly known and anticipated sequence γ_t makes time t a payoff-relevant state variable in Section 2.3.

When the old play a, citizen strategy $a^*(a,p) = \mathbb{1}_{\{p \geqslant c_a^*\}}$ assigns work or leisure for each p. Markov perfection requires optimality for all p, even those not reached in equilibrium. This consistency requires that no one has the incentive to deviate against the prescribed action, $a^*(a,p)$. For very large values of p, Δu is large and eventually individual incentives to work override preference for cohesion. Therefore, for every equilibrium, $a^*(a,p) = 1$ for p sufficiently large. In this case, today's young prefer to work even if there is no policing tomorrow. This means it is always feasible for the government to enforce work by policing high enough, though not necessarily always optimal.

The following argument establishes this upper-dominance region using Eq. (2.2.26) by putting an upper bound on citizen's cutoff used in any MPE. When $p > (1+\beta)B$ citizen's best-response is always work because policing today is high enough to cover cost of work for both periods even if everyone else protests and any additional coordination is a bonus. Recall that

$$\overline{p}(p_0, a, a_0) = (1+\beta)B - \beta p_0 - \alpha\left(\frac{a + a_0\beta}{2}\right) \tag{2.2.27}$$

[7] It is known that policing will be less effective in the future, perhaps because of military and law-enforcement beginning to sympathize with the opposition.

and note it is monotonically decreasing in all arguments because some policing today can be substituted by policing tomorrow or coordination with other citizens.

Fix any MPE $\{(p_0^*, p_1^*), (c_0^*, c_1^*)\}$ and apply (2.2.26):

$$\forall a \in \{0,1\}, c_a^* \leqslant \overline{p}(p_0^*, a, \mathbb{1}_{\{p_0^* \geqslant c_0^*\}}) \leqslant \overline{p}(p_0^*, 0, 0) = (1+\beta)B - \beta p_0^* \leqslant (1+\beta)B \tag{2.2.28}$$

The second inequality from monotonicity. Similarly, $a^*(a,p) = 0$ for sufficiently small p if government's strategy were to prescribe small policing in the good state tomorrow that is reached when today's young pick work.

We now establish there is a similar lower-dominance region where very low policing today makes protest dominant. This means government will face definite protests if it never polices. If tomorrow's policing is not too great, so that today's policing choice is meaningful $p_1^* < \frac{1+\beta}{\beta}(B-\alpha)$, then $\forall a \in \{0,1\} : c_a^* > 0$. [8]. To see this, recall that $\underline{p}(p_1, a, a_1) = (1+\beta)B - \beta p_1 - \alpha\left(\frac{1+\beta}{2} + \frac{a+a_1\beta}{2}\right)$ and note it is monotonically decreasing in all arguments.

Fix any MPE $\{(p_0^*, p_1^*), (c_0^*, c_1^*)\}$ that satisfies $p_1^* < \frac{1+\beta}{\beta}(B-\alpha)$ and apply (2.2.21):

$$\forall a \in \{0,1\}, c_a^* \geqslant \underline{p}(p_1^*, a, \mathbb{1}_{\{p_1^* \geqslant c_a^*\}}) \geqslant \underline{p}(p_1^*, 1, 1) = (1+\beta)(B-\alpha) - \beta p_1^* > 0 \tag{2.2.29}$$

The second inequality from monotonicity. Then $c_a^* > 0$ and citizen's best-response is always protest when facing $p \in [0, c_a^*)$ because the level of policing today and tomorrow are not enough to cover the cost of work for both periods even if everyone else works. Recall that Assumption 2.4 stated $0 < \alpha < B$ and note that it is sufficient to generate this lower-dominance region. Its economic interpretation is that citizens work only when government sufficiently polices enough and, in particular, never work

[8] Here $p_1^* \geqslant 0$ is well defined because $B > \alpha$ by Assumption 2.4

when government never polices. This baseline Assumption 2.4 supposes some level of policing is necessary for this authoritarian government to remain in power.

The next Lemma is going to remove strictly dominated levels of policing from government decision in (2.2.13). It will reduce the choice set of policing from $[0, \infty)$ down to two relevant actions: no policing, which is the least policing to incentivize protest, or $p_a^* = c_a^*$, which is the least policing to incentivize work. This result is straightforward but it will be used repeatedly in further derivations and relies on citizens playing a cutoff strategy. Recall the previous discussion about dominance regions – the government may choose to operate at a boundary but not inside those regions to save on policing costs and getting the same outcome. The economic significance of this Lemma is that government is a large player and a Stackelberg leader that can feasibly force any play as a best-response for the citizens that are small followers. The outcome that the government actually selects for the young citizens to choose depends on γ, how effective the government police is.

Lemma 2.10. *Suppose citizens follow the cutoff strategy $c^* = (c_0^*, c_1^*)$. The government's optimal action in state a is then $p_a^* \in \{0, c_a^*\}$.*

Proof. Recall government's optimal decision problem given by Eq. (2.2.13). Consider any one-shot deviation $\tilde{p}$ in state a from government's p^*-strategy in Eq. (2.2.13). $\tilde{p} \neq c_a^*$ and $\tilde{p} \neq 0$ then $\tilde{p}$ is a never-best response to citizen's policing threshold c_a^* because $\tilde{p}$ is always strictly dominated by one of $\{0, c_a^*\}$. In the government's decision problem, note that excessive policing $\tilde{p} > c_a^*$ is strictly dominated by $\tilde{p} = c_a^*$, which gives the same outcome of work because $\mathbb{1}_{\{\tilde{p} \geqslant c_a^*\}} = 1 = \mathbb{1}_{\{c_a^* \geqslant c_a^*\}}$. But $\tilde{p}$ costs more than c_a^*, so government gets a smaller present payoff:

$$(1-\delta)\left(\frac{a+1}{2} - \frac{1}{2\gamma}\tilde{p}\right) < (1-\delta)\left(\frac{a+1}{2} - \frac{1}{2\gamma}c_a^*\right) \tag{2.2.30}$$

Of course, the continuation values are the same, $G(1|p^*)$, because tomorrow's state $a = 1$ in both cases.

$$(1-\delta)\left(\frac{1}{2}-\frac{1}{2\gamma}c_a^*\right)+\delta G\left(1|p^*\right) > (1-\delta)\left(\frac{1}{2}-\frac{1}{2\gamma}\tilde{p}\right)+\delta G\left(1|p^*\right) \tag{2.2.31}$$

Similar argument eliminates any positive policing that is strictly below the cutoff, $0 < p(a) < c_a^*$, because such policing is strictly dominated by 0 where tomorrow's state is 0 in both cases but is cheaper to achieve with no policing than below-threshold positive policing:

$$(1-\delta)\left(0\right)+\delta G\left(0|p^*\right) > (1-\delta)\left(-\frac{1}{2\gamma}\tilde{p})\right)+\delta G\left(0|p^*\right) \tag{2.2.32}$$

□

This means, the best-response for the government at state a is always in $\{0, c_a^*\}$, which simplifies government's decision problem into binary choice.

Next, we consider government's best-response to some given (c_0, c_1) citizen strategy, resilient to one-shot deviation from Eq. (2.2.13). There are four possible government best-responses (two possible actions in each of two states).

Categorize government's continuation strategy by the unique labor paths it induces:

- "Always Revolt" (AR) if $p_0^* = p_1^* = 0$. Young and old both protest.
- History-dependent "Traditional Play" (TR) if $p_0^* = 0$, $p_1^* = c_1^*$. Young and old play the same.
- "Counter-Culture" (CC) if $p_0^* = c_0^*$, $p_1^* = 0$. Young play the oppoite of old.

- "Never Revolt" (NR) if $p_0^* = c_0^*$, $p_1^* = c_1^*$. Young and old both work.

Every optimal government strategy belongs to exactly one of these types by Lemma 2.10.

Then the proposed strategy p^* survives one-shot deviation precisely when deviating to $\tilde{p}(a) \in \{0, c_a^*\} \backslash p^*(a)$ is not profitable for each $a \in \{0, 1\}$. Thus we only need to consider a single deviation in each state.

Combining (2.2.13) with Lemma 2.10, we can derive conditions on γ, which is the police productivity. Each of the four government strategies has to withstand one-shot deviation in the current period when that initial strategy is used for continuation play. Table 2.1 lists payoff to playing AR and NR strategies, payoffs for corresponding one-shot deviations and required condition so that deviation is not profitable.

Table 2.1: Always Revolt(AR) and Never Revolt(NR) as Best-Responses

	Government's Continuation Strategy	
Govt. Current Play	AR: $p^* = (0, 0)$	NR: $p^* = (c_0^*, c_1^*)$
Play $p^*(0)$	$G(0) = 0$	$G(0) = \frac{1}{2\gamma}\left[\gamma(1+\delta) - (1-\delta)c_0^* - \delta c_1^*\right]$
Play $p^*(1)$	$G(1) = \frac{(1-\delta)}{2}$	$G(1) = 1 - \frac{1}{2\gamma}c_1^*$
Dev. to $\tilde{p}_0 \neq p^*(0)$	$\tilde{G}(0) = (1-\delta)\frac{1}{2\gamma}\left[\gamma(1+\delta) - c_0^*\right]$	$\tilde{G}(0) = \delta G(0)$
Dev. is unprofitable if	$\tilde{G}(0) \leqslant G(0) \iff c_0^* \geqslant (1+\delta)\gamma$	$\tilde{G}(0) \leqslant G(0) \iff c_0^* \leqslant \gamma\frac{(1+\delta)}{1-\delta} - \frac{\delta}{1-\delta}c_1^*$
Dev. to $\tilde{p}_1 \neq p^*(1)$	$\tilde{G}(1) = (1-\delta)\frac{1}{2\gamma}\left[\gamma(1+\delta) - c_1^*\right] + G(1)$	$\tilde{G}(1) = \frac{1-\delta}{2} + \delta G(0)$
Dev. is unprofitable if	$\tilde{G}(1) \leqslant G(1) \iff c_1^* \geqslant (1+\delta)\gamma$	$\tilde{G}(1) \leqslant G(1) \iff c_1^* \leqslant \gamma + \frac{\delta}{(1+\delta)}c_0^*$

Consider payoffs in each state to government picking *"Always Revolt"* strategy with $p_0^* = 0$ and $p_1^* = 0$.

$$G(0|p^*) = 0, G(1|p^*) = \frac{1-\delta}{2}$$

In state $a = 0$, the relevant deviation to consider by Lemma 2.10 is $\tilde{p}_0 = c_0^*$ that forces today's young to work and then reverting to $p^* = (0, 0)$ next period, so that all future young protest as before. The corresponding payoff is:

$$\begin{aligned}\tilde{G}(0|p^*) =&(1-\delta)\left(\frac{1}{2}-\frac{1}{2\gamma}c_0^*\right)+\delta\frac{(1-\delta)}{2}\\ &=(1-\delta)\left[\frac{(1+\delta)}{2}-\frac{1}{2\gamma}c_0^*\right]=(1-\delta)\frac{1}{2\gamma}\left[\gamma(1+\delta)-c_0^*\right]\end{aligned} \tag{2.2.33}$$

The initial strategy induces protest in every period. Under the deviation, only the first two periods differ in outcomes: after observing today's old protesting, the young work today and tomorrow's old work but tomorrow's young protest just like under "AR." The deviation is not profitable, if policing today with cost $\frac{1}{2\gamma}c_0^*$ that exceeds the benefit from today's young working for two periods $\frac{(1+\delta)}{2}$.

$$0 = G(0|p^*) \geqslant \tilde{G}(0|p^*) \iff c_0^* \geqslant \gamma(1+\delta) \tag{2.2.34}$$

In state $a = 1$, the relevant deviation to consider by Lemma 2.10 is $\tilde{p}_1 = c_1^*$ that forces today's young to work and then reverting to $p^* = (0, 0)$ next period, so that all future young protest as before. The corresponding payoff is:

$$\begin{aligned}\tilde{G}(1|p^*) =&(1-\delta)\left(\frac{1+1}{2}-\frac{1}{2\gamma}c_1^*\right)+\delta\frac{(1-\delta)}{2}\\ &=(1-\delta)\frac{1}{2\gamma}\left[\gamma(1+\delta)-c_1^*\right]+\frac{1-\delta}{2}\end{aligned} \tag{2.2.35}$$

Under the deviation, only the first two periods differ in outcomes: after observing today's old working, the young work today and tomorrow's old work but tomorrow's young protest just like under "AR." The deviation is not profitable, if policing today

with cost $\frac{1}{2\gamma}c_1^*$ that exceeds the benefit from today's young working for two periods $\frac{(1+\delta)}{2}$.

$$\begin{aligned} G(1|p^*) &\geqslant \tilde{G}(1|p^*) \\ &\iff \frac{1-\delta}{2} \geqslant (1-\delta)\frac{1}{2\gamma}\left[\gamma(1+\delta) - c_1^*\right] + \frac{1-\delta}{2} \\ &\iff c_1^* \geqslant \gamma(1+\delta) \end{aligned} \tag{2.2.36}$$

Consider payoffs in each state to government picking the *"Never Revolt"* strategy $p_0^* = c_0^*, p_1^* = c_1^*$.

When the old work, the government's payoff along the equilibrium path, after inducing today's young to work, is

$$G(1|p^*) = (1-\delta)(1 - \frac{1}{2\gamma}c_1^*) + \delta G(1|p^*) = 1 - \frac{1}{2\gamma}c_1^*, \tag{2.2.37}$$

When the old protest, the government's payoff along the equilibrium path, after inducing today's young to work, is

$$\begin{aligned} G(0|p^*) &= (1-\delta)\left(\frac{1}{2} - \frac{1}{2\gamma}c_0^*\right) + \delta G(1|p^*) \\ &= \frac{1+\delta}{2} - (1-\delta)\frac{1}{2\gamma}c_0^* - \delta\frac{1}{2\gamma}c_1^* \\ &= \frac{1}{2\gamma}\left(\gamma(1+\delta) - (1-\delta)p_0^* - \delta c_1^*\right) \end{aligned} \tag{2.2.38}$$

In state $a = 0$, the relevant deviation to consider by Lemma 2.10 is $\tilde{p}(0) = 0$ that allows today's young to protest and then reverting to $p^* = (c_0^*, c_1^*)$ next period, so that all future young work as before. The corresponding payoff is smaller if:

$$\tilde{G}(0|p^*) = 0 + \delta G(0|p^*), G(0|p^*) \geqslant \tilde{G}(0|p^*) \iff G(0|p^*) \geqslant 0$$
$$\iff \gamma(1+\delta) - (1-\delta)p_0^* - \delta c_1^* \geqslant 0$$
$$\iff \frac{(1+\delta)}{2} \geqslant (1-\delta)\frac{1}{2\gamma}p_0^* + \frac{1}{2\gamma}\delta c_1^* \tag{2.2.39}$$

Here the requirement is that policing required at $a = 0$ is low enough:

$$c_0^* \leqslant \gamma\frac{(1+\delta)}{1-\delta} - \frac{\delta}{1-\delta}c_1^* \tag{2.2.40}$$

Under the prescribed strategy, it is cheaper to pay p_0^* today and p_1^* tomorrow for total cost of $\frac{1}{2\gamma}(p_0^* + \delta p_1^*)$ and get benefit $\frac{1+\delta}{2}$ of today's young working than pay p_0^* tomorrow for total cost of $\delta\frac{1}{2\gamma}p_0^*$ and get no benefit. The net cost of adhering to the strategy is $\frac{1}{2\gamma}(1-\delta)p_0^* + \delta p_1^*$ is smaller than the net gain $\frac{(1+\delta)}{2}$. Deviation for state $a = 1$ can be evaluated similarly.

Table 2.2 is analogous for CC and TR strategies.

Table 2.2: Counter Culture(CC) and Traditional Play(TR) as Best-Responses

	Government's Continuation Strategy	
Government's Current Play	CC: $p^* = (c_0^*, 0)$	TR: $p^* = (0, c_1^*)$
Play $p^*(0)$	$G(0) = \frac{\frac{1}{2\gamma}}{1+\delta}\left[\gamma(1+\delta) - c_0^*\right]$	$G(0) = 0$
Play $p^*(1)$	$G(1) = \frac{\frac{1}{2\gamma}}{1+\delta}\left[(1+\delta)\gamma - \delta c_0^*\right]$	$G(1) = 1 - \frac{1}{2\gamma}c_1^*$
Dev. to $\tilde{p}_0 \neq p^*(0)$	$\tilde{G}(0) = \delta G(0)$	$\tilde{G}(0) = \frac{1}{2\gamma}\left[\gamma(1+\delta) - (1-\delta)c_0^* - \delta c_1^*\right]$
Dev. is unprofitable if	$\tilde{G}(0) \leqslant G(0) \iff c_0^* \leqslant (1+\delta)\gamma$	$\tilde{G}(0) \leqslant G(0) \iff c_0^* \geqslant \gamma\frac{(1+\delta)}{1-\delta} - \frac{\delta}{1-\delta}c_1^*$
Dev. to $\tilde{p}_1 \neq p^*(1)$	$\tilde{G}(1) = (1-\delta)(1 - \frac{1}{2\gamma}c_1^*) + \delta G(1)$	$\tilde{G}(1) = \frac{(1-\delta)}{2}$
Dev. is unprofitable if	$\tilde{G}(1) \leqslant G(1) \iff c_1^* \geqslant \gamma + \frac{\delta}{(1+\delta)}c_0^*$	$\tilde{G}(1) \leqslant G(1) \iff c_1^* \leqslant \gamma(1+\delta)$

The following Proposition summarizes the results in the tables above.

Proposition 2.11. *Suppose citizens are playing a cutoff strategy $c^* = (c_0^*, c_1^*)$. The government's best-response p^* is then unique a.e. (except when equality holds)*

1. *If $c_0^* \geqslant (1+\delta)\gamma$ and $c_1^* \geqslant (1+\delta)\gamma$ then the Best-Response is AR: $p^* = (0, 0)$*

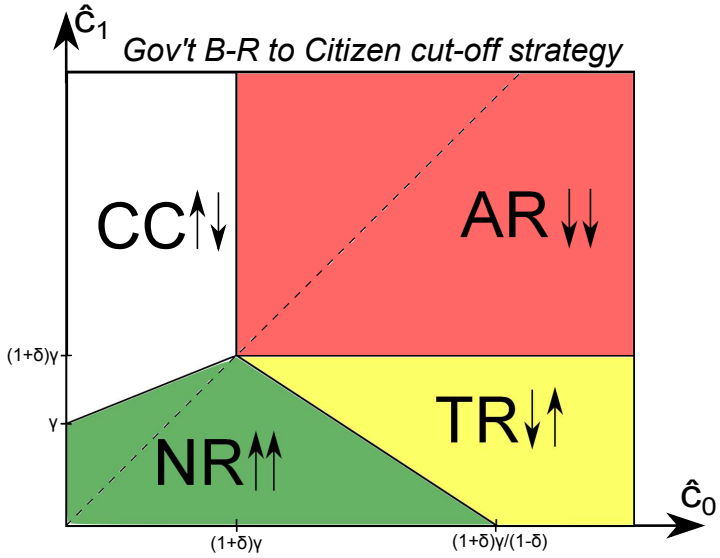
2. *If* $c_0^* \leqslant \gamma\frac{(1+\delta)}{1-\delta} - \frac{\delta}{1-\delta}c_1^*$ *and* $c_1^* \leqslant \gamma + \frac{\delta}{(1+\delta)}c_0^*$ *then the Best-Response is NR:* $p^* = (c_0^*, c_1^*)$

3. *If* $c_0^* \leqslant (1+\delta)\gamma$ *and* $c_1^* \geqslant \gamma + \frac{\delta}{(1+\delta)}c_0^*$ *then the Best-Response is CC:* $p^* = (c_0^*, 0)$

4. *If* $c_0^* \geqslant \gamma\frac{(1+\delta)}{1-\delta} - \frac{\delta}{1-\delta}c_1^*$ *and* $c_1^* \leqslant \gamma(1+\delta)$, *then the Best-Response is TR:* $p^* = (0, c_1^*)$

This can be visualized in the $c_0^* - c_1^*$ space of citizen cutoffs. When citizens have somewhat small demands in each state, then government's best response is to provide matching police force, enforcing work in Never Revolt (NR) region. Likewise, large demands in both states imply government will not be able to afford required policing and it will give up in both states, leading to protests in Always Revolt (AR) region. When citizens have one cutoff much larger than the other, the government provides policing in the state with the lower cutoff and gives up otherwise.

Figure 2.2: Government has unique (a.e.) best-response to citizen strategy



Each category of the best-responses is determined by a pair of inequalities from Proposition 2.12, one for each state. For example, AR best-responses require given thresholds to satisfy $G^0(1) \geqslant G^{c_1^*}(1)$ if and only if $\gamma \leqslant \frac{c_1^*}{1+\delta}$ in state $a = 1$. The second

inqeuality they need to satisfy is $0 \geqslant G^{c_0^*}(0)$ if and only if $\gamma \leqslant \frac{c_0^*}{1+\delta}$ in state $a = 1$, which is a square region in the upper right of $c_0^* - c_1^*$ plane.

2.2.3 Characterizing Citizen's Best-Response Cutoff

Eq. (2.2.21) requires that facing high policing p in state a, $p \geqslant c_a^*$, young citizens find it individually optimal to work, assuming other young conform to the equilibrium strategy cutoff $c^* = (c_0^*, c_1^*)$ and work. Secondly, Eq. (2.2.26) requires that facing low policing p in state a, $p < c_a^*$, young citizens find it individually optimal to protest, assuming other young conform to the equilibrium strategy cutoff $c^* = (c_0^*, c_1^*)$ and protest. To be a best-response, a citizen's strategy thus needs to satisfy both conditions (2.2.21) and (2.2.26).

$$\boxed{BR(\hat{p}_0, \hat{p}_1) = \{(c_0, c_1) : \text{ (2.2.21) and (2.2.26) are satisfied}\}} \tag{2.2.41}$$

$(c_0, c_1) \in BR(\hat{p}_0, \hat{p}_1)$ means that, in all states, (c_0, c_1) is a best-response for individual citizen to government playing $\hat{p}-$strategy and other citizens playing the same cutoff strategy (c_0, c_1). BR is thus the set of all cutoffs that are best-responses to the given government strategy and itself.

Proceed in three steps (see Appendix for details). First, Lemmas A.1 and A.2 characterize sets of cutoffs (c_0^*, c_1^*) that satisfy each of the conditions – Equations (2.2.21) and (2.2.26) respectively. These Lemmas A.1 and A.2 allow for government to play arbitrary p^*. Secondly, by Lemma 2.10, government's best-response is either no policing or matching citizens' cutoff. Using that, Lemma A.3 shows the results from Lemmas 2.2 and 2.3 that are simplified for the special case when Government plays zero policing or matches minimum policing. Thirdly, in Proposition 2.12, the

BR set, which is contained in the $c_0 - c_1$ space, is characterized by an intersection of two intervals for each of the four relevant[9] government plays.

Observe that by monotonicity of $\underline{p}$ in its third argument, a_1, which is the action of the young tomorrow after observing tomorrow's old working:

$$\underline{p}(p_1^*, 1, 1) < \underline{p}(p_1^*, 1, 0). \tag{2.2.42}$$

proposition combines previous two Lemmas to show that when government plays one of $\{NR, AR, TR, CC\}$ citizens' $BR(p_0^*, p_1^*)$ is an interval where the lower bound comes from Lemma 3.2.1 (no policing at $a = 1$, $p_1^* = 0$) or Lemma 3.4.1 (matching policing at $a = 1$, $p_1^* = c_1^*$) and the upper bound comes from Lemma 3.3.1 (no policing at $a = 0$, $p_0^* = 0$ or Lemma 3.4.1 (matching policing at $a = 0$, $p_0^* = c_0^*$). In all cases, the best-response in state a to one of these four government candidate strategies is proven to be an interval and can be succinctly expressed as

$$\forall a \in \{0, 1\}, c_a^* \in [\underline{p}(p_1^*, a, \mathbb{1}_{\{p_1^* \geqslant c_1^*\}}), \overline{p}(p_0^*, a, \mathbb{1}_{\{p_0^* \geqslant c_0^*\}})] \tag{2.2.43}$$

Proposition 2.12. *Suppose (p_0^*, p_1^*) is a given government strategy and (c_0^*, c_1^*) is a citizen cutoff strategy.*

1. *Never Revolt (NR): If $(p_0^* = c_0^*, p_1^* = c_1^*)$ then (c_0^*, c_1^*) satisfies $BR(p_0^*, p_1^*)$ if and only if $\underline{p}(p_1^*, a, 1) \leqslant c_a^* \leqslant \overline{p}(p_0^*, a, 1)$, for $a = 0, 1$.*

2. *Always Revolt (AR): If $(p_0^* = 0, p_1^* = 0)$ then c_0^* then (c_0^*, c_1^*) satisfies $BR(p_0^*, p_1^*)$ if and only if $\underline{p}(p_1^*, a, 0) \leqslant c_a^* \leqslant \overline{p}(p_0^*, a, 0)$, for $a = 0, 1$.*

3. *Traditional Play (TR): If $(p_0^* = 0, p_1^* = c_1^*)$ then (c_0^*, c_1^*) satisfies $BR(p_0^*, p_1^*)$ if and only if $\underline{p}(p_1^*, a, 1) \leqslant c_a^* \leqslant \overline{p}(p_0^*, a, 0)$, for $a = 0, 1$.*

[9] In the sense that these plays are the only government best-responses to an arbitrary citizen strategy.

4. *Counter-culture (CC): If* $(p_0^* = c_0^*, p_1^* = 0)$ *then* (c_0^*, c_1^*) *satisfies* $BR(p_0^*, p_1^*)$ *if and only if* $\underline{p}(p_1^*, a, 0) \leqslant c_a^* \leqslant \overline{p}(p_0^*, a, 1)$, *for* $a = 0, 1$.

Proof. See Appendix. □

Observe that Counter-Culture equilibrium has outcomes that cycle employment and unemployment: today's young play the opposite of what today's old play. All young are rebellious and counter-culture is the norm. This is not a very reasonable description of an oppressive state to have power forever fluctuate between government and opposition between even and odd periods. This kind of play requires (i) very high cutoff c_1^* relative to c_0^* to incentivize the government and (ii) citizens to be very impatient and disregard future gains of permanent work to get one-time leisure payoff today.

We can expanded the second condition of the $BR(p_0^*, p_1^*)$ on c_0^* when the government playing Counter-Culture (CC), that $c_0^* \leqslant \overline{p}(p_0^*, 0, 1)$ as:

$$c_0^* \in \left[(1+\beta)\left(B - \frac{\alpha}{2}\right), B - \frac{\beta}{2(1+\beta)}\alpha\right] \tag{2.2.44}$$

A little bit of citizen's patience destroys this play.[10] As social cohesion gets weaker relative to costs of working ($\frac{B}{\alpha} \to \infty$), the patience requirement to destroy this equilibrium weakens arbitrarily. In other words, fix arbitrarily patience of citizens β to be arbitrarily small, and then if the coordination motive is sufficiently weak, this play is never a Best-Response by citizens.

Define the minimum patience we need to exceed as:

$$\underline{\beta} = \frac{-1 + \sqrt{1 + \frac{4}{2\frac{B}{\alpha} - 1}}}{2} \in (0, 1) \tag{2.2.45}$$

[10] Citizen's patience does not make other three plays AR, NR or TR disappear.

Corollary 2.13. *In Proposition 2.12.4 citizens' Best-Response correspondence to Government playing Counter-culture is* $\varnothing$ *if and only if* $\beta > \underline{\beta}$

Proof. See Appendix. □

We can now look at the intersection of $BR(p_0^*, p_1^*)$ and one-shot deviation conditions. First, consider unproductive police force with low $(1+\delta)\gamma$. Then the government gives up in every equilibrium (AR). Since there is no anticipated policing in the future, today's thresholds are very high, approximately $(1+\beta)B$, which means today's police needs to incentivize two periods of work by itself.

Proposition 2.14. *Let* $\mathcal{E}^{AR}(\gamma)$ *be the set of equilibria where government plays* $p^* = (0,0)$ *and therefore citizens always revolt ("AR") along the equilibrium path.*

1. $\mathcal{E}^{AR}(\gamma) \neq \varnothing$ *if and only if*

$$\gamma \leqslant \overline{\gamma}^{AR} \equiv (1+\beta)B/(1+\delta) - \alpha/(2+2\delta).$$

2. *For all such* $\gamma \leqslant \overline{\gamma}^{AR}$, *citizens using the highest cutoff in* $BR(p^* = (0,0))$ *constitutes an equilibirum,*

$$\{p^* = (0,0), (c_0^*, c_1^*) = ((1+\beta)B, (1+\beta)B - \alpha/2)\} \in \mathcal{E}^{AR}(\gamma),$$

which is thus robust to changes in $\gamma \in [0, \overline{\gamma}^{AR}]$.

Proof. See Appendix. □

As $(1+\delta)\gamma$ reaches intermediate values it can support a TR equilibrium where the government gives up in the low state and polices in the high state a fixed amount that varies among different allowed TR equilibria. These equilibria support a wide range of policing in the high state because its $BR(p_0^*, p_1^*)$ conditions are not tight: the same

top-end ($\overline{p}(p_0^*, a, 0) = (1+\beta)B - \alpha(\frac{a}{2})$) as for AR as they have same continuation for downward deviations with no policing in the low state.

It is consistent for citizens to be "pessimistic" about tomorrow under moderate policing – the rest coordinate to protest, entering a low state with no policing. Today pessimistic citizens would require high policing in the high state to offset tomorrow's low payoff when stuck working. On the other hand, it is also consistent to be "optimistic" about others' strategy with bottom-end threshold as in NR ($\underline{p}(p_1^*, a, 1) = (1+\beta)B - c_1^* - \alpha(\frac{1+2\beta+a}{2})$ – when minimal level of policing is involved, everyone works and expects the same payoff tomorrow, making less policing required today. The government's behavior is consistent with history-dependence when citizens are more pessimistic in the low state than in the high state ($c_1^* \leqslant (1+\delta)\gamma) \leqslant c_0^*$ from Proposition 2.11.4).

Proposition 2.15. *Let $\mathcal{E}^{TR}(\gamma)$ be the set of equilibria where government plays $p^* = (0, c_1^*)$ and therefore citizens follow traditional play ("TR") along the equilibrium path, consequently each young copies the previous generation's choice.*

1. *$\mathcal{E}^{TR}(\gamma) \neq \varnothing$ if and only if*

$$\gamma \in [\underline{\gamma}^{TR}, \overline{\gamma}^{TR}] \equiv [(B-\alpha)/(1+\delta), (1+\beta)B/(1+\delta) - \delta\alpha/(2+2\delta)].$$

2. *There is no TR equilibrium that is robust to changes in γ over the whole range $[\underline{\gamma}^{TR}, \overline{\gamma}^{TR}]$.*

$$\forall \gamma \in [\underline{\gamma}^{TR}, \overline{\gamma}^{TR}], \exists \gamma' \in [\underline{\gamma}^{TR}, \overline{\gamma}^{TR}] : \mathcal{E}^{TR}(\gamma) \cap \mathcal{E}^{TR}(\gamma') = \varnothing.$$

Proof. See Appendix. □

The following proposition 2.17 is going to need the government to be somewhat patient.

Assumption 2.16. $\delta \geqslant \underline{\delta} = \frac{\beta}{1+\beta}$.

This Assumption 2.16 states that the government is sufficiently patient: its discount factor is not much smaller than β or it is greater.

Note that as $\beta \to 0$ the assumption 2.16 weakens to $0 < \delta < 1$ and $\delta \geqslant 0.5$ is sufficient for any β. This is not a necessary condition but it's a sufficient condition used in the proof construction. For case of $\delta < \underline{\delta}$, the γ boundaries in the equilibrium set of Theorem 2.18 will be slightly different.

Proposition 2.17. *Assume $\underline{\delta} \leqslant \delta < 1$.*[11] *Let $\mathcal{E}^{NR}(\gamma)$ be the set of equilibria where government plays $p^* = (c_0^*, c_1^*)$ and thus citizens never revolt ("NR") along the equilibrium path.*

1. *$\mathcal{E}^{NR}(\gamma) \neq \varnothing$ if and only if $\gamma \geqslant \underline{\gamma}^{NR} \equiv B/(1+\delta) - \alpha/2$*
2. *For all such $\gamma \geqslant \underline{\gamma}^{NR}$, citizens using the smallest cutoff in $BR(p^* = (c_0^*, c_1^*))$ constitutes an equilibirum:*

$$\{p^* = (c_0^*, c_1^*) = (B - \alpha/2, B - \alpha)\} \in \mathcal{E}^{NR}(\gamma)$$

which is thus robust to changes in $\gamma \geqslant \underline{\gamma}^{NR}$]. Furthermore, it is the only equilibrium of $\mathcal{E}^{NR}$ with that robustness property.

Proof. See Appendix. □

Fixing all parameters $(\frac{1}{2\gamma}, \alpha, B)$, we can characterize the set of different Markov Perfect equilibria in threshold strategies. First we will look at best-response citizen's threshold strategies that satisfy Equations [2]-[3] for each state (for each of three category's of government's best-response). Recall that by Corollary 2.1, we eliminate counter-culture ("CC") equilibria for sufficient citizen patience $\beta > \underline{\beta} = \frac{-1+\sqrt{1+\frac{4}{2\frac{B}{\alpha}-1}}}{2} \in (0,1)$.

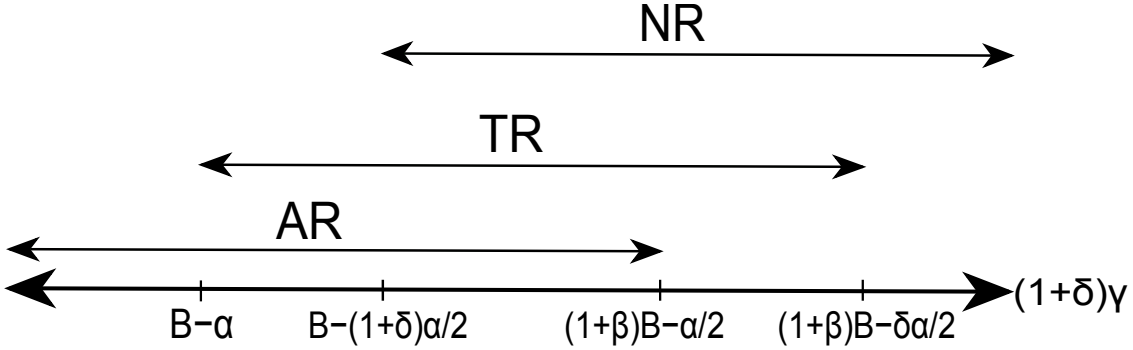
[11] For smaller δ, ranges on γ will be slightly different and different construction should be used.

Theorem 2.18. *Assume $\beta > \underline{\beta}$ and $\delta \geqslant \frac{\beta}{1+\beta}$. If $(1+\delta)\gamma \in$*

1. *$(0, B-\alpha)$ then only Always Revolt (AR) equilibria exist.*
2. *$[B-\alpha, B-(1+\delta)\alpha/2)$ then only AR and Traditional Play (TR) classes of equilibria exist.*
3. *$[B-(1+\delta)\alpha/2, (1+\beta)B-\alpha/2]$ then AR, TR and No Revolt (NR) classes of equilibria exist.*
4. *$((1+\beta)B-\alpha/2, (1+\beta)B-\delta\alpha/2]$ then only TR and NR classes of equilibria exist.*
5. *$((1+\beta)B-\delta\alpha/2, \infty)$ then only NR class of equilibria exists.*

Proof. Simply intersect the productivity regions from propositions 2.14, 2.15 and 2.17 as they precisely identify where specific class of equilibria is located. □

Figure 2.3: The full set of MPE in cutoff strategies for discount factors high enough



Letting γ vary, we can see how equilibrium sets vary over 5 regions: only "AR", "AR" and "TR", "AR" and "TR" and "NR", "TR" and "NR", only "NR".

When government's police productivity is very low, $(1+\delta)\gamma < B-\alpha$, then every equilibrium is in Always Revolt class: government never polices, citizens never work and labor force falls to 0. If government's productivity is moderately-low,

$$(1+\delta)\gamma \in [B-\alpha, B-(1+\delta)\frac{\alpha}{2}), \tag{2.2.46}$$

then AR equilibria as well as history-dependent Traditional Play equilibria exist. If the old worked, then the government polices and the young work, and thus the labor force remains at 1. But if the old revolted, then the government gives up and the young revolt, and thus labor force remains at 0. Here, the labor force in the low state is always 0 and labor force in the high state depends on equilibrium.

If government's productivity is moderate,

$$(1+\delta)\gamma \in [B-(1+\delta)\frac{\alpha}{2}, (1+\beta)B-\frac{\alpha}{2}], \tag{2.2.47}$$

then AR and TR equilibria still exist and also No Revolt (NR) equilibria are allowed: government always polices and citizens always work and labor force rises to 1. Here labor force in both states is indeterminate and depends on equilibrium. If government's productivity is moderately-high,

$$(1+\delta)\gamma \in ((1+\beta)B-\frac{\alpha}{2}, (1+\beta)B-\delta\frac{\alpha}{2}], \tag{2.2.48}$$

then AR equilibria disappear and only TR and NR equilibria remain. Here labor force in the low state is indeterminate and labor force in the high state rises to 1.

Finally, if government's productivity is very high,

$$(1+\delta)\gamma > (1+\beta)B-\delta\frac{\alpha}{2}, \tag{2.2.49}$$

then NR is the only equilibrium class that is allowed. Here labor force in both states rises to 1.

Among all equilibria, the highest cutoffs are seen for the greatest AR equilibrium [12]:

$$\{p^* = (0,0), c^* = \Big((1+\beta)B, (1+\beta)B - \frac{\alpha}{2}\Big)\} \tag{2.2.50}$$

Likewise, the lowest cutoffs are seen for the smallest NR equilibrium[13]:

$$\{p^* = c^* = \Big(B - \frac{\alpha}{2}, B - \alpha\Big)\} \tag{2.2.51}$$

The following corollary observes that one of these two extreme extreme equilibria always exists. Naturally, low γ – unproductive police can always sustain the highest AR, high γ – productive police can always sustain the lowest NR. The cutoffs in these equilibria independent of γ. Intermediate γ can sustain both as well transient TR equilibria that are not robust and depend on γ.

Recall 2.2.45 that

$$\underline{\beta} = \frac{-1 + \sqrt{1 + \frac{4}{2\frac{B}{\alpha} - 1}}}{2} \in (0,1) \tag{2.2.52}$$

Corollary 2.19. *Assume $\beta > \underline{\beta}$. Then for every γ, at least one of the following is an equilibrium: {greatest cutoff AR, smallest cutoff NR}.*

Proof. By Theorem 2.18, there is either NR or AR equilibrium (or both). By Propositions 2.3 and 2.5, those are the respective robust equilibria. □

2.3 Police Productivity γ with a Downward Trend

In the previous analysis, $\gamma_t = \gamma$ police productivity was constant over time. Instead, consider a downward trend in $\gamma_t > 0$ over time. This can be interpreted as police

[12] Without policing tomorrow, thresholds today are especially high

[13] Policing infinitely into the future keeps thresholds today especially low

becoming less effective per dollar invested over time. The law enforcement and the military are defecting to the opposition, making marginal policing more expensive.

Assumption 2.20. *Public knowledge about* $\{\gamma_t\}_{-\infty}^{\infty}$ *sequence.*

Assumption 2.21. *Downward trend* $\gamma_t \geqslant \gamma_{t+1}$.

Assumption 2.22. *Initial (high-productivity region* $\exists L : (1+\delta)\gamma_L > (1+\beta)B$

Assumption 2.23. $\frac{\alpha}{B} < \min\{\frac{1-\beta^2}{1+\frac{1}{2}\beta}, \frac{2\beta(1-\beta^2)}{1+\beta-\beta^3}\} \in (0, 0.55)$

Assumption 2.24. *Eventual low-productivity region*

$$\exists N : (1+\delta)\gamma_N < (1-\beta^2)B - \alpha\left(1+\frac{\beta}{2}\right)$$

Assumptions 2.22 and 2.24 will establish extreme dominance regions to "Never Revolt" initially and "Always Revolt" eventually, respectively.

Next, Assumption 2.23 is a technical requirement to prove sufficiency for some of the results (it is not a necessary restriction). $\frac{\alpha}{B} < \frac{1-\beta^2}{1+\frac{1}{2}\beta}$ ensures $(1-\beta^2)B-\alpha(1+\frac{\beta}{2}) > 0$, and thus Assumption 2.24 is well-defined. This stems from considering the lowest possible best-response level of policing required and government not being able to afford that level. The second part of Assumption 2.23, $\frac{\alpha}{B} < \frac{2\beta(1-\beta^2)}{1+\beta-\beta^3}$ will be used to prove the non-monotonicity result in Theorem 2.31.

The government is a long-lived player with $0 < \delta < 1$ discount and citizens are short-lived, asummed to live for two periods with $0 < \beta < 1$ discount. There is a measure $\frac{1}{2}$ of citizens that are born every period and commit to an action $a \in \{0, 1\}$ for both periods, where $a = 0$ represents "protest" or "joining the opposition" and $a = 1$ represents "work." Citizens are "Young" when they are born and decide their action and are "Old" when they are stuck playing what they chose last period.

At the beginning of period t, the government observes the average action of the Old a^o before picking a policing level $p_t \in [0, \infty)$. Focusing on the symmetric pure

strategies, $a^o \in \{0, 1\}$. Then the Young are born and they observe both (a^o, p_t) before picking work or protest $a \in \{0, 1\}$. The labor force is the total amount of work done by the Old and the Young combined,

$$L_t = (1/2)a^o + (1/2)a^y. \tag{2.3.1}$$

The labor force is restricted to $\{0, \frac{1}{2}, 1\}$ for symmetric equilibria in pure strategies.

Government's Markov strategy in state (a, t), the Old's average work level and time t, is the policing level:

$$p(a, t) : \{0, 1\} \times \mathbb{N} \to [0, \infty) \tag{2.3.2}$$

(Young) citizen's Markov strategy in state (a, p, t), which is the Old's work level and government policing level, is the choice between protest and work:

$$\hat{a}(a, p, t) : \{0, 1\} \times [0, \infty) \times \mathbb{N} \to \{0, 1\} \tag{2.3.3}$$

We are still going to use Markov Perfect Equilibrium in pure strategies as the main solution concept, with the added generation of using time t as one of the states. A pair of Markov strategies (a^*, p^*) are MPE when they withstand one-shot deviation in every state.

Again, restricting attention to citizens playing a threshold strategy this can be rewritten as $\{(c_0^{*,t}, c_1^{*,t})\}_t$:

$$a^*(a, p, t) = \begin{cases} 1 & \text{if } p \geqslant c_a^{*,t}, \\ 0 & \text{if } p < c_a^{*,t}, \end{cases} \tag{2.3.4}$$

Government utility from p^*, taking $\{(c^{t_0}, c^{t_1})\}_t$ as given, at state a and at time t is

$$G^t(a|p^*) = (1-\delta)\left(\frac{a}{2} + \frac{\mathbb{1}_{\{p_a^{*,t} \geqslant c_a^t\}}}{2} - \frac{1}{2\gamma}p_a^{*,t}\right) + \delta G^{t+1}(\mathbb{1}_{\{p_a^{*,t} \geqslant c_a^t\}})|p^*) \tag{2.3.5}$$

Denote government utility from **one-shot deviation** to $\tilde{p} \in [0, \infty)$ and later going back to p^* as

$$\tilde{G}^t(a|p^*) = (1-\delta)\left(\frac{a}{2} + \frac{\mathbb{1}_{\{\tilde{p} \geqslant c_a^t\}}}{2} - \frac{1}{2\gamma}\tilde{p}\right) + \delta G^{t+1}(\mathbb{1}_{\{\tilde{p} \geqslant c_a^t\}}|p^*) \tag{2.3.6}$$

Taking $\{(c_0^t, c_1^t)\}_t$ as given, government's choice $\{(p_0^{*,t}, p_1^{*,t})\}_t$ is optimal for every $a \in \{0,1\}$, for every $t \in \mathbb{N}$, for every deviation $\tilde{p} \in [0, \infty)$:

$$\begin{aligned}(1-\delta)\left(\frac{\mathbb{1}_{\{p_a^{*,t} \geqslant c_a^t\}}}{2} - \frac{1}{2\gamma}p_a^{*,t}\right) + \delta G\left(\mathbb{1}_{\{p_a^{*,t} \geqslant c_a^t\}}, t+1\right) \\ \geqslant (1-\delta)\left(\frac{\mathbb{1}_{\{\tilde{p} \geqslant c_a^t\}}}{2} - \frac{1}{2\gamma}\tilde{p}\right) + \delta G\left(\mathbb{1}_{\{\tilde{p} \geqslant c_a^t\}}, t+1\right)\end{aligned} \tag{2.3.7}$$

Corollary 2.25. *Suppose citizens follow c^*-strategy, then government's best-response from Eq.* (2.3.7) *in state (a,t) is $p_a^{*,t} \in \{0, c_a^{*,t}\}$.*

Proof. Note that proof of Lemma 2.10 applies here because continuation value $G(1, t+1)$ is the same for $\tilde{p}$ and $c_a^{*,t}$ whenever $\tilde{p} \geqslant c_a^{*,t}$ and $G^{t+1}(0)$ for 0 and $\tilde{p} < c_a^{*,t}$, so pick policing that gives lowest cost today and the same continuation. □

Suppose in state (a, p, t) with other citizens following prescribed c^*-strategy with $p \geqslant c_a^{*,t}$ and government follows $\hat{p}$-strategy, the young citizen also prefers to work over protest if:

$$p \geqslant (1+\beta)B - \alpha\left(\frac{1+\beta}{2} + \frac{a}{2} + \frac{\beta\mathbb{1}_{\{\hat{p}_1^{t+1} \geqslant c_1^{*,t+1}\}}}{2}\right) - \beta\hat{p}_1^{t+1} \tag{2.3.8}$$

The above condition on work, $\hat{a} = 1$, to be a collectively sustained citizen best-response holds for all $p \geqslant c_a^{*,t}$ if and only if

$$\forall a \in \{0,1\}, c_a^{*,t} \geqslant (1+\beta)B - \alpha\left(\frac{1+\beta}{2} + \frac{a}{2} + \frac{\beta \mathbb{1}_{\{\hat{p}_1^{t+1} \geqslant c_1^{*,t+1}\}}}{2}\right) - \beta \hat{p}_1^{t+1} \tag{2.3.9}$$

Suppose in state (a,p,t) with other citizens following prescribed c^*-strategy with $p < c_a^{*,t}$ and government follows $\hat{p}$-strategy, citizen also prefers to protest over work if:

$$p < (1+\beta)B - \alpha\left(\frac{a}{2} + \frac{\beta \mathbb{1}_{\{\hat{p}_0^{t+1} \geqslant c_0^{*,t+1}\}}}{2}\right) - \beta \hat{p}_0^{t+1} \tag{2.3.10}$$

The above condition on protest being a collectively sustained best-response holds for all $p < c_a^{*,t}$ if and only if

$$\forall a \in \{0,1\}, c_a^{*} \leqslant (1+\beta)B - \alpha\left(\frac{a}{2} + \frac{\beta \mathbb{1}_{\{\hat{p}_0^{t+1} \geqslant c_0^{*,t+1}\}}}{2}\right) - \beta \hat{p}_0^{t+1} \tag{2.3.11}$$

The difference between LHS of (2.3.8) and (2.3.10) is that in the former case citizen derives coordination utility of $\frac{1}{2}$ from working alongside with 1/2 population of the young today and $\frac{\beta}{2}$ from working with 1/2 population of the old tomorrow. In the later case, today's young and tomorrow's old protest instead because $p < c_a^{*,t}$.

Exactly like in the static case, to describe (2.3.9) condition, define the following auxiliary function, $\underline{p}(p_1, a, a_1)$. In this notation, a is the old's labor choice from last period and (a_1, p_1) are tomorrow's work and policing choices after observing work today.

$$\underline{p}(p_1, a, a_1) \equiv (1+\beta)B - \beta p_1 - \alpha\left(\frac{1+\beta}{2} + \frac{a + a_1\beta}{2}\right). \tag{2.3.12}$$

To describe (2.3.11) condition, define the following auxiliary function:

$$\overline{p}(p_0, a, a_0) \equiv (1+\beta)B - \beta p_0 - \alpha\left(\frac{a + a_0\beta}{2}\right). \tag{2.3.13}$$

In this notation, a is the old's labor choice from last period and (a_0, p_0) are tomorrow's work and policing choices after observing protest today.

Given tomorrow's play

$$\{\hat{p}_0^{t+1}, \hat{p}_1^{t+1}, \hat{c}_0^{t+1}, \hat{c}_1^{t+1}\}, \tag{2.3.14}$$

define *a static, collectively-sustained citizen best-response at t* as an element of the following two-dimensional region bounded by four inequalities:

$$BR^t(\hat{p}_0^{t+1}, \hat{p}_1^{t+1}, \hat{c}_0^{t+1}, \hat{c}_1^{t+1}) = \tag{2.3.15}$$

$$\{(c_0^t, c_1^t) : \underline{p}(\hat{p}_1^{t+1}, a, \mathbb{1}_{\{\hat{p}_1^{t+1} \geqslant \hat{c}_1^{t+1}\}}) \leqslant c_a^t \leqslant \overline{p}(\hat{p}_0^{t+1}, a, \mathbb{1}_{\{\hat{p}_0^{t+1} \geqslant \hat{c}_0^{t+1}\}}) \text{ for a=0,1.}\} \tag{2.3.16}$$

By construction, $x \in BR^t$ is equivalent to x satisfies (2.3.9) and (2.3.11).

We can now describe the circumstances when an individual citizen has no profitable unilateral deviations in all periods. Given a sequence of policing $(\hat{p}_0^t, \hat{p}_1^t)$, a sequence of cutoffs $(\hat{c}_0^t, \hat{c}_1^t)$ is said to be *a dynamic, collectively sustained best-response* if $(c_0^t, c_1^t) \in BR^t(\hat{p}_0^{t+1}, \hat{p}_1^{t+1}, c_0^{t+1}, c_1^{t+1})$ for all periods t. The distinction is that the static requirement only makes sure today's cutoffs work given tomorrow's policing and tomorrow's cutoffs. The dynamic requirement also needs to make sure that yesterday's play is consistent with today's cutoff. It is possible that some cutoff $(\hat{c}_0^T, \hat{c}_1^T)$ satisfies BR^t but there is no sequence $\{c_0, c_1\}_t$ satisfying BR^t for all t, coinciding at $T : c_0^T = \hat{c}_0^t, \hat{c}_1^T = c_1^T$.

By construction, (2.3.8) part of BR is violated if $c_a^{*,t} < \underline{p}(\hat{p}_1^{t+1}, \mathbb{1}_{\{\hat{p}_1^{t+1} \geqslant c_1^{*,t+1}\}})$ and (2.3.10) part of BR is violated if $c_a^{*,t} > \overline{p}(\hat{p}_0^{t+1}, a, \mathbb{1}_{\{\hat{p}_0^{t+1} \geqslant c_0^{*,t+1}\}})$, hence this is a necessary condition.

The following lemma is similar to Proposition 2.12 but now $c_a^{*,t} \neq c_a^{*,t+1}$ (citizen thresholds may or may not be equal across time), which allows more varied labor behavior tomorrow to be consistent with some best-response threshold today for any $\beta \in (0,1)$. For example, Counter-culture tomorrow requires $c_0^{*,t+1} \leqslant \frac{\alpha}{2\beta}$, which is small for large β, while every $c_0^{*,t}$ satisfying BR at t is always large because $\hat{p}_1^{t+1} = 0$. This is why in the static case requiring thresholds to be constant over time lead to breaking BR at today when CC is played next period. Thus the dynamic thresholds allow for new kind of equilibrium action CC, with revolt in a good state and working in a bad state, that was eliminated under "reasonably high" patience with stationary strategies.

Define the length of side of the BR set along a as:

$$\boxed{\Delta^t = \overline{p}(p_0^{t+1}, a, a_0) - \underline{p}(p_1^{t+1}, a, a_1) = \beta(p_1^{t+1} - p_0^{t+1}) + \alpha\left(\frac{(1+\beta)}{2} + \frac{\beta(a_1 - a_0)}{2}\right)} \tag{2.3.17}$$

The second effect of letting thresholds vary over time is that the shape of $\{(c_0^{*,t}, c_1^{*,t})$: satisfying $BR^t(p_0^{*,t+1}, p_1^{*,t+1})$ conditions at every $t\}$ is a square. Previously the restriction of time-stationary thresholds would lead to downward sloping boundary for small c_0^*. With exogenously fixed $(c_0^{*,t+1}, c_1^{*,t+1})$ the shape is a rectangle and Δ^t, the length of the side, is independent of a, so it is a square with lower-left corner at

$$(\underline{p}(p_1^{t+1}, 0, \mathbb{1}_{\{p_1^{*,t+1} \geqslant c_1^{*,t+1}\}}), \underline{p}(p_0^{*,t+1}, 1, \mathbb{1}_{\{p_1^{*,t+1} \geqslant c_1^{*,t+1}\}})) \tag{2.3.18}$$

and upper-right corner at

$$\overline{p}(p_0^{*,t+1}, 0, \mathbb{1}_{\{p_0^{*,t+1} \geqslant c_0^{*,t+1}\}}), \overline{p}(p_0^{*}, 1, \mathbb{1}_{\{p_0^{*,t+1} \geqslant c_0^{*,t+1}\}})). \tag{2.3.19}$$

Lemma 2.26. *Suppose at time $t+1$ the government plays Never Revolt (NR), Always Revolt (AR), Traditional Play (TR) or Counter-culture (CC) that pins tomorrow young's action at (a_0, a_1)[14]. Then $(c_0^{*,t}, c_1^{*,t})$ satisfies $BR^t(p_0^{*,t+1}, p_1^{*,t+1})$ if and only if*

$$\underline{p}(p_1^{*,t+1}, a, a_1) \leqslant c_a^{*,t} \leqslant \overline{p}(p_0^{*,t+1}, a, a_0)$$

Outline. The result is similar to Proposition (2) but the proof is much simpler because next period's cutoffs depend on $t+1$ and are uncoupled from today's cutoffs. This follows straight from the definition of 2.3.17. □

The above Lemma means you only need to know tomorrow's policing and tomorrow old's profile of equilibrium actions in each state $a \in \{0, 1\}$ to verify $(c_0^{*,t}, c_1^{*,t}) \in BR^t(p_0^{*,t+1}, p_1^{*,t+1})$ – in other words, today's cutoffs are independent of each other as long as the corresponding equilibrium policing levels don't violate feasibility at $t-1$, in other words that $BR^{t-1}(p_0^{*,t}, p_1^{*,t}) \neq \varnothing$ (see next Lemma).

The following lemma gives a necessary and sufficient condition on play at $t+1$ to ensure BR set at t is non-empty or equivalently $\Delta^t \geqslant 0$. It cannot be the case that $c_1^{*,t+1} \ll c_0^{*,t+1}$ as that moves lower boundary of BR at t above upper boundary of BR at t, making BR empty.

Lemma 2.27. *$\exists\{(c_0^{*,t}, c_1^{*,t})\}$ satisfying $BR^t(p_0^{*,t+1}, p_1^{*,t+1})$ if and only if ($\Delta^t \geqslant 0$)*

$$p_1^{*,t+1} - p_0^{*,t+1} \geqslant -\frac{\alpha}{2\beta}\left(1 + \beta + \beta(\mathbb{1}_{\{p_1^{*,t+1} \geqslant c_1^{*,t+1}\}} - \mathbb{1}_{\{p_0^{*,t+1} \geqslant c_0^{*,t+1}\}})\right)$$

[14]The subscript i indicates that the current young's action is i. Therefore, tomorrow's young will observe i before making their choice, $a_i = \mathbb{1}_{\{p_i^{*,t+1} \geqslant c_i^{*,t+1}\}}$.

Proof. See Appendix. □

The previous Lemma highlights the desired restriction on next the period's policing levels. It cannot be too high in the bad state relative to good state, or BR conditions fail for the previous period. This is similar to how we eliminated Counter-Culture (CC) equilibria in the stationary case.

The following proposition defines a special "FP" fixed-point set of BR thresholds for Never Revolt (NR) play. If a sequence of thresholds satisfy BR and some point later in the sequence is in "FP," then every point before that is also in FP. But even if none of the points are in "FP", as the sequence gets longer, the first point $(c_0^{*,t=1}, c_1^{*,t=1})$ is arbitrarily close to FP's boundary. In other words, whenever government polices within this boundary (a moderate amount), then it policed inside the boundary every period before that. And every sequence of positive policing supporting NR play, always starts arbitrarily close to the boundary if the sequence is long enough.

The following expression $\overline{P}_a$ is the maximum possible positive policing in any MPE in state a, when today's young are pessimistic[15], the following period will have no policing and tomorrow's young will protest:

$$\boxed{\overline{P}_a = \overline{p}(p^{t+1} = 0, a, a_0 = 0) = (1+\beta)B - \alpha\frac{a}{2}} \tag{2.3.20}$$

Likewise, the following expression $\underline{P}_a$ is the minimum possible positive policing in any MPE in state a, when today's young are optimistic[16] the following period will

[15]Protest is dominant above any lower policing level, by expecting full mutual protest in such a case.

[16]Work is dominant below any higher policing level, by expecting full mutual protest in such a case.

have maximum policing $\overline{P}_a$ and tomorrow's young will work:

$$\boxed{\underline{P}_a = \underline{p}(p^{t+1} = \overline{P}_a, a, a_0 = 1) = (1-\beta^2)B - \alpha\left(\frac{1+a+\beta}{2}\right)} \tag{2.3.21}$$

Proposition 2.28. *Suppose Never Revolt (NR) is played for $t \leqslant L$. The following BR set, denoted as "FP", is the smallest fixed-point set.*

1. *If $\beta \in (0, \frac{1}{2})$: "FP"=*

$$\left[B - \frac{\alpha(1+\beta-\beta^2)}{2(1-\beta^2)}, B + \frac{\alpha\beta^2}{2(1-\beta^2)}\right] \times \left[B - \frac{\alpha(2+\beta-2\beta^2)}{2(1-\beta^2)}, B - \frac{\alpha(1-2\beta^2)}{2(1-\beta^2)}\right]$$

2. *If $\beta \in (\frac{1}{2}, 1)$: "FP"=*

$$\left[B - \frac{\alpha(1+\beta-\beta^2)}{2(1-\beta^2)}, B + \frac{\alpha\beta^2}{2(1-\beta^2)}\right] \times \left[B - \frac{\alpha(2+\beta-2\beta^2)}{2(1-\beta^2)}, B - \frac{\alpha(1-2\beta^2)}{2(1-\beta^2)}\right]$$
$$\cap\{(p_0, p_1) : \textit{Lemma (2.27) holds.}\}$$

Proof. (OUTLINE) 1. BR set at $t-1 \subset$ BR set at t. (i) $\underline{p}(p_1^{*,t-1}, a, 1) \geqslant \underline{p}(p_1^{*,t}, a, 1)$ and (ii) $\overline{p}(p_0^{*,t-1}, a, 1) \leqslant \underline{p}(p_0^{*,t}, a, 1)$ This means once you start in that square region, then you stay there.

The second restriction for $\beta \in (0.5, 1)$ ensures $p_0 \ll p_1$ doesn't happen.

To show that this is the unique FP: any BR set at t that contains the BR set at $t-1$, also contains "FP." As $L \to \infty$, the candidate set shrinks to FP arbitrarily close (pick any BR point outside of FP at $t = 2$, it is no longer contained in BR set at $t = 1$ for L large enough).

To show that it has no "slack," pick top left corner. BR set at $t-1$ starting there is "FP." The size of "FP" square $\Delta = \frac{\alpha}{2(1-\beta)}$

Observation: Δ^{t-1}, size of the BR set the previous period increases in Δ_{p^t} - the difference in policing levels at t. This difference in policing levels is maximized at the top left corner of any BR set. Starting with a larger candidate for "FP", shrinks the BR set strictly the period before: if $\Delta^t > \frac{\alpha}{2(1-\beta)} \implies \Delta^{t-1} < \Delta^t$. Similarly, starting with a smaller candidate for "FP", grows the BR set strictly the period before: if $\Delta^t < \frac{\alpha}{2(1-\beta)} \implies \Delta^{t-1} > \Delta^t$. This shows "FP" is the unique such set.

□

The interpretation of the Fixed-Point set is it's precisely the set of policing levels that are feasible under Never Revolt equilibrium with infinite horizon (e.g. if γ always remained in the upper-dominance region).

Proposition 2.29 (Opposition eventually takes over)**.** *1. (Lower Dominance Region) Consider subgame starting at* (N, a).[17] *Unique labor outcome is "Always Revolt" for every Markov Perfect Equilibrium for this subgame :* $p_a^{*,t} = 0 < c_a^{*,t}$.

2. (Contagion) Extend the subgame above to a supergame $(t = K \leqslant N, a)$ *such that* $(1+\delta)\gamma_K < (1+\beta)B - \alpha\left(1+\frac{\beta}{2}\right)$ *. for every Markov Perfect Equilibrium for this subgame :* $\forall t \geqslant N : p_a^{*,t} = 0 < c_a^{*,t}$.

Proof. See Appendix. □

Similar to the stationary dynamic model, there are four possible equilibrium labor outcomes in a given period. The above proposition established a contagion argument where forward-looking (pessimistic) expectations lead to a unique labor outcome of Always Revolt. Previously, it was only the case that AR was unique for $(1+\delta)\gamma < B - \alpha$ and allowed for multiplicity from TR and NR above that.

Now, the NR region has been extended to

$$(1+\delta)\gamma < (1+\beta)B - \alpha\left(1+\frac{\beta}{2}\right) \tag{2.3.22}$$

[17] From Asmp 4, $(1+\delta)\gamma_N < (1-\beta^2)B - \alpha\left(1+\frac{\beta}{2}\right) = \underline{P}_1$

Looking at it another way, the stationary model had multiplicity on a region with width increasing in βB, the cost of work, plus a factor proportional to α, because with stationary policing strategy tomorrow's policing was correlated with today's policing. In the dynamic model with a downward trend, the multiplicity region has width proportional to α only, coordination among the citizens only. This is a qualitative sense in which the indeterminacy has been reduced.

The next theorem focuses on describing a general pattern present in any equilibrium with a downward productivity trend. The first stage is "Tyranny" where the government enforces work in every period for a while.

It is followed by a turbulent stage called "Revolution" and for a given productivity trend, different equilibria allow for different paths taken on this region. If citizens coordinate on instant revolution, they could have it on the first day of the period or at the other extreme, instead, they could coordinate to do it on the last period with some specified play in the interim. This is consistent with the literature where timing of the revolution in the short-run is somewhat indeterminate, even when it is certain to happen in the long-run.

The final stage is called "Opposition in Power" and it corresponds to the government completely giving up forever, never policing and citizens always revolting.

Theorem 2.30. *Consider γ_t satisfying Asmp 1-5. Let $L = \max_l\{(1+\delta)\gamma_l > (1+\beta)B\}$, the last period in the upper dominance region. Fix any MPE equilibrium.*

1. *(TYRANNY) The first $L > 0$ periods have Never Revolt (NR) outcome.*

2. *(REVOLUTION) Next $k \geqslant 0$ periods play one of {Never Revolt (NR), Always Revolt (AR), Traditional play (TR), Counter-culture (CC) }*

3. *(OPPOSITION IN POWER): The infinite tail starting from $K = L + k > 0$, has Always Revolt (AR) outcome.*

Proof. 1. At $t = L$, the worst case scenario is the government faces infinite tail of Always Revolt starting next period at $t = L + 1$. From the first column of Table 1, playing $\tilde{p} = c_a^{*,L}$ gives strictly greater payoff than playing $p = 0 \iff c_a^{*,L} < (1+\delta)\gamma_L$. This is true because $c_a^{*,L} \leqslant (1 + \beta)B < (1 + \delta)\gamma_L$

3. This follows from Proposition 2.29. Let

$$K = \min_k \left\{ (1 + \delta)\gamma_k > (1 + \beta)B - \alpha \left(1 + \frac{\beta}{2}\right) \right\}$$

2. This follows directly from Corollary 4.1 and $k = K - L$ depends on K, when the contagion kicks in. □

The final result establishes the non-monotonicity of policing result for the dynamic model with a trend. While the previous Theorem 2.30 focused on how anticipation of future play affected the labor path, Theorem 2.31 looks at how every equilibrium labor path affects government's equilibrium policing response. An important contribution of this paper is to highlight the following pattern.

On the period before the revolution begins: (i) the tomorrow's old protest, which discourages work today, (ii) continuation utility of receiving policing tomorrow becomes zero since the government gives up, also reducing payoff to work. These two factors cause a discontinuous drop in relative utility of work, so today's policing needs to be higher by a discrete increase to compensate. Less policing is needed to incentivize work where for many future periods there is going to be a guarantee of tomorrow's old working, plus a positive amount of policing tomorrow.

This means when the government had held the power firmly a long time ago, early in the "Tyranny" region, it policed a moderate amount. On the last period of "Tyranny" before "Revolution," it must police a lot more, else protests would have started even earlier. The government subdues mild opposition with mild polic-

ing, deters strong opposition with heavy policing and gives up against unstoppable opposition at the start of the final region where "Opposition is in Power."

Theorem 2.31 (Non-monotonicity of policing in γ)**.** *Consider γ_t satisfying Asmp 1-5 and for every MPE, there exist time periods M and $T, (M < T)$, that exhibit non-monotonicity of observed policing with respect to γ.*

$$0 = p_1^{*,T} < p_1^{*,M} < p_1^{*,T-1}$$

Proof. See Appendix. □

2.4 Conclusion

This paper extends the literature on protests and revolutions to include government as a strategic agent. Choosing to highlight time frictions, rather than informational frictions, the model makes a connection between anticipation of inevitable revolution and non-monotonic government's policing.

The dynamic model allows for forward-looking expectations to reduce equilibrium indeterminacy by using a downward trend of police productivity. On one hand, it fully characterizes the different kinds of Markov equilibrium labor paths that arise in a repeated game. On the other hand, it makes a prediction that the government will police mildly when its power is secured, give up when the opposition is too strong and fiercely fight back when its government's rule is about to collapse, even if the collapse is inevitable. This roughly matches the Imperial Russian government's response to the Socialist opposition. First it used spies, imprisonment and exile. Then when its rule was in peril, it crushed striking civilians with artillery bombardment. Next time, however, its policing effectiveness has declined and it could not oppose the revolutionaries.

The model with stationary police productivity gives a short-run analysis of potential protests. Later, police productivity is described by a deterministic (downward) trend. This gives a long-run analysis of how anticipation of eventual fall of government brings about certain revolution. This revolution will happen at an earlier time than is likely in the short-run model without aligned expectation of its fall. Still, the exact date of the revolution is unpredictable and may vary across equilibria, which is consistent with the widespread surprise of the Berlin Wall collapsing.

In the stationary case with high police productivity, the government's costly action can force coordination on its preferred outcome of work, retaining power. For the stationary low police productivity, the government cannot afford to cover individual citizen's fixed cost of work and there are always protests and never policing. Therefore, the government loses control to the rebel opposition. In the intermediate regions, there are also possible outcomes of Traditional Play, where rebels' children protest and workers' children work. However, much of this is resolved by a downward trend as an Always Revolt outcome by contagion.

An extension for this line of research would be to model the government-opposition game as "matching pennies" where the citizens want to allocate themselves across time or place (a clandestine meeting), preferring to be together but to avoid the government. Meanwhile, the government uses a finite policing budget to allocate its police to minimize the oppositions' gatherings as much as it could.

Chapter 3

Reference-Dependent Attitudes to Risk, Incumbency Advantage and Response to Crisis

The political science literature has identified a salient phenomenon known as *incumbency advantage*, where politicians in office stand a higher chance of being reelected than challengers vying for the same seat. Conventional explanations include deterrence of challengers, lower quality of challengers than incumbents (incl. incumbents having political experience), use of office for reelection (e.g. franking letters, greater access to media, pork-barrel spending).

Secondly, more recent research has described the opposite circumstance of *incumbency disadvantage* when challengers do better in bad times after an exogenous shock to the economy, which is unrelated to the government's actions.

A non-conventional explanation using prospect theory may be able to explain both incumbency advantage during good times and incumbency disadvantage during a time of crisis, when the loss-gains function is strictly concave with risk-seeking in the losses and risk-aversion in the gains.

The third stylized fact is that while incumbent's average disaster relief increases in the magnitude of the (exogenous) crisis, their average probability of winning decreases. In other words, an average incumbent wins more often when he is lucky to avoid an unrelated crisis and loses more often when he is not lucky, while providing the expected disaster relief for that particular crisis. Thus, these election patterns cannot be explained by backward-looking voters who punish only lazy politicians for underperformance.

Over 90% of House Representatives seeking reelection were successful since World War II (Levitt and Wolfram, 1997). Gelman and King (1990) developed the first consistent and unbiased measure in terms of vote percentage margins, constructing a proper time-series for congressional elections. They found the average incumbency advantage to be 2% between 1900 and 1950, then steadily rising to about 10% in 1990. In fact, Levitt and Wolfram (1997) found that deterrence trumps officeholder benefits, including free mail, media access, fund-raising advantages to explain the rise of incumbency advantage post-World War II.

Ansolabehere and Snyder Jr (2002) showed that incumbency advantage exists at a similar level (co-movements over time) at state and federal legislatures, gubernatorial and other state executives. Therefore, these phenomena do not depend on specific from features of the legislature like redistricting (gerrymandering), diffusion of responsibility or pork-barrel politics because similar incumbency advantage exists in elections for offices without these specific benefits. The proposed decline in challenger quality was also rejected as an explanation for the rise in incumbency advantage. There has to be a more general explanation that applies both to legislatures and the executive.

More recently, Wolfers (2007) analyzed comprehensive empirical evidence that voters reelect incumbents during good times (high oil prices for oil-producing states, national boom for pro-cyclical states) and elect challengers during bad times. One

natural question is whether incumbency advantage disappears in bad times, or just that it's really big in good times and smaller in bad times. Wolfers (2007) finds that, on average, the incumbent governor is reelected 56.7% of the time. Tables 3.1 and 3.2 summarize (somewhat imprecise) estimates of how a governor's chances at his reelection are affected by a sudden rise or fall in the price of oil that can be inferred from the 1950-1988 subset of his data[1]. In some cases, the magnitudes suggested are more than enough to cause an incumbency disadvantage. For example, when the price of oil increases by one standard deviation, an incumbent governor's probability for reelection in a rust-belt manufacturing state decreases by about 7.1%. Likewise, when the price of oil decreases by one standard deviation, an incumbent governor's probability for reelection in an oil-producing state decreases by about 22%. Wolfers (2007) concludes that it is most probable that voters do not efficiently process information and compare the state economy to national reference point, as well as making an attribution error.

Table 3.1: Effect of Oil Price Increase on Incumbent's Probability of reelection

	Oil Price[a] Shock 1σ above mean	Largest(+) Oil Shock
Oil producing state[b]	[0.11, 0.38][c]	[0.25, 0.90][c]
Rust-Belt state[d]	[−0.032, −0.11][c]	[−0.077, −0.27][c]

[a] ΔLog Real Oil, annual averages
[b] Using largest positive state-specific coefficient (0.23), belonging to Alaska, Wyoming or Texas
[c] 68% confidence interval because the coefficient was significant at 10% but not 5%.
[d] Using largest negative state-specific coefficient (-0.07), belonging Michigan or Indiana

Rational models of asymmetric information do not generate these results of incumbency advantage alternating with incumbency disadvantege by pure luck. In rational models, either elections solve a moral-hazard problem of shirking incumbents or a bad

[1] Baseline model from Column 1 of Table 5.C: Incumbent elected$_{s,t}$ = λ National employment gap$_t$ + $\delta(\beta_s$*Oil Shock$_t$) + α(State employment gap$_{s,t}$ −Δ National employment gap$_t$ − β_s *Oil Shock$_{s,t}$) + $\varepsilon_{s,t}$

outcome reveals the incumbent is lower quality than his expected challenger. A model with loss-averse voters is necessary as the recorded external shocks were specifically noted to be outside of the incumbent's influence.

Table 3.2: Effect of Oil Price Decrease on Incumbent's Probability of reelection

	Oil Price[a] Shock 1σ. below mean	Largest$(-)$ Oil Shock
Oil producing state[b]	$[-0.10, -0.34]$[c]	$[-0.31, -1.11]$[c]
Rust-Belt state[d]	$[0.029, 0.10]$[c]	$[0.096, 0.34]$[c]

[a] ΔLog Real Oil, annual averages
[b] Using largest positive state-specific coefficient (0.23), belonging to Alaska, Wyoming or Texas
[c] 68% confidence interval because the coefficient was significant at 10% but not 5%.
[d] Using largest negative state-specific coefficient (-0.07), belonging Michigan or Indiana

In a similar vein, Achen and Bartels (2004) find voters favoring challengers after Acts of God like droughts, floods, and shark attacks, concluding this is an unresolved puzzle for rational choice theory of voting. In a follow-up to test this "blind retrospect" theory, Cole, Healy, and Werker (2008) find that after natural disasters in India, vigorous and responsive administrations that offer relief fare better than unresponsive ones in the next election but worse than expected if no disaster had taken place. Tables 3.3, 3.4 summarize their linear and non-linear models relating disaster severity (flood/drought), amount of relief assistance and change in the incumbent's vote share. During severe weather, an incumbent administration that provides an average amount of disaster relief loses 4.04% votes, essentially being punished for bad luck. During extreme weather, average response is 3.45 times larger and it is not surprising that doing nothing is worse during extreme weather (-10%) than during severe weather (-4.6%). Also, citizens value relief twice as much during the extreme weather (coefficient of 0.0672) as during severe weather (coefficient of 0.032). But, paradoxically, the incumbent party is punished even more severely on average (loses 6.18% votes). While they do not quantify the level of of incumbency (dis)advantage,

they do note that rainfall ("luck") is relevant for deciding the winner because a quarter of elections in their sample have margin smaller than 5.26%.

Table 3.3: Effect of Variable Rain on Disaster Relief and Votes for Incumbent Party

	Expenditure[a] per σ of Rain away from Optimal Rain[b]	Δ Vote(%)
No Response	0	-3.77%
Ave. Response	0.178	-3.25%
2x ave. Response	0.356	-2.73%

[a]A governmental response is an increase in log(disaster spending)

[b]Optimal rain is computed to be about 1 standard deviation (σ) above average, where farming output is at maximum

It thus appears that there is a robust puzzle that incumbency advantage vanishes during a time of crisis, even if the crisis has nothing to do with the incumbent. Secondly, disaster relief increases in the magnitude of the crisis, while the incumbent is disadvantaged and only extreme responses merit reelection. Conventional explanations either should continue to hold (e.g. if pork-barrel spending was the cause) or are inadequate (e.g. principal-agent arguments about effort or quality with rational voters). This paper takes the stance that when the loss-averse voters experience a sudden cut to their consumption, the required minimum ability cutoff for the incumbent rises. When, in equilibrium, the average disaster relief reveals the incumbent to have an average ability, he is then disadvantaged at the election, relative to a risky prospect of the unknown challenger. It takes an extreme performance from the incumbent to reveal extremely high ability to pass the muster.

This paper develops a model to explain all of these facts by linking politicians with career concerns and forward-looking voters with reference-dependent utility. (Holmström, 1999) developed a framework of moral hazard (with symmetric information) where the agent was considered for a promotion based on their unknown ability and

Table 3.4: Non-linear Effect of Rain on Disaster Relief and Votes for Incumbent Party

	Severe Weather[a]		Extreme Weather[b]	
	Exp.[c]	Δ Vote(%)	Exp.	Δ Vote(%)
No Response	0	−4.57%	0	−9.99%
Ave. Response	0.1641	−4.04%	0.5663	−6.18%
2.62x ave. response	0.4306	−3.19%	1.486	0

[a]Severe Weather accounts for rainfall difference from optimal amount in 80-90 percentiles
[b]Extreme Weather accounts for rainfall difference from optimal amount in 90-100 percentiles
[c]A governmental response is an increase in log(disaster spending)

took a hidden action, such as effort, that potentially clouded inference about their innate ability. While in some equilibria it may be the case that the agent's action ended up revealing his ability after-the-fact, his initial incentives involved considerations to improve the signal by exerting some additional effort.

Whether voters are assumed to be loss-averse or rational expected-utility maximizers, a politician will want to provide a non-trivial amount of the public good before the election. The incumbent will take a less-than-maximal personal rent, because raising the rents further would make them look bad as if they had low ability to create the public good. Thus, the politician does not want to reduce his ex-ante probability of winning by shirking. With rational voters, politicians only care about affecting signals that relate to their ability and ignore, for example, exogenous shocks to voters' future income, because their probability of winning only depends on the manifestation of their own ability.

Secondly, even if the politicians had no hidden action to take (such as fixed or zero personal rents) and the incumbent's type was public information, that simpler model would still generate incumbency advantage in good times and incumbency disadvantage in bad times. The uncertainty about the challenger's ability, relative to the better-known incumbent, plus prospect theory gives the result.

Combining both aspects in one model leads to "fickle" voters that take into account irrelevant signals as long as their consumption relative to the reference point

is affected. When reference-dependence is combined with career concerns, the third stylized fact is also matched. Thus, politicians signal higher ability when they vary their hidden action relative to these irrelevant signals since these now enter into their probability of winning, even though in equilibrium their type is revealed anyway.

In this case, the incumbent provides greater disaster relief when the crisis worsens as the voters are risk-seeking and lean more and more towards the challengers. Incumbents increase public goods g_1^* in the first period precisely when their probability of losing increases. Their decision involves trade-off between personal rent and increasing probability of being elected by choosing a higher signal. In equilibrium, all incumbents' signals are perfectly correlated with their ability, so their type becomes known. However, a simpler model without career concerns that has known incumbent's type does not have the tension between choosing signal to affect margin of the probability of winning, thus more public goods wont necessarily be provided for rising moderate shock s. On the other hand, when the crisis does not occur, the incumbent produces less of the public good and takes more personal rents by enjoying his incumbency advantage.

A possible explanation for incumbency advantage via prospect theory was first proposed by Quattrone and Tversky (1988), described in terms of classroom questionnaires and psychology intuition. The questions included policy proposals by candidates, framed in terms of losses and gains for separate groups of responders. Instead of using expected-utility, consumers were supposed to be loss-averse with respect to a particular, given reference point. With an S-shaped value function, they are risk-seeking in the losses region and risk-averse in the gains region. If the incumbent represents the continuation of the status quo and the challenger is a risky gamble, then the incumbent should tend to get more support, except during bad times (with risk-seeking to attempt recouping losses). However, this modelling device of loss-averse voters has not been theoretically developed into an equilibrium model beyond

validation in simple classroom experiments. Prior to the present paper, this direction has not been pursued in the literature on incumbency advantage. Since Quattrone and Tversky (1988) hah not constructed an equilibrium model, they made no predictions about the government's disaster-relief response to the presence of these loss-averse voters.

While explaining the vulnerability of incumbency advantage to economic and outside shocks, a limitation of this approach is that it may be challenging to explain the positive trend in the magnitude of the incumbency advantage.

Loss aversion has been used to offer an alternative explanation for why the Presidents party tends to lose seats in midterm congressional elections (Patty, 2006). Outside of political economy, Fershtman (1996) used prospect theory in IO to demonstrate that variation in the reference points affects market decisions of incumbent and challenger firms, changing the market equilibrium in a dynamic oligopoly game.

The proposed theoretical model needs to go beyond assuming the challenger is riskier than the incumbent. The model requires assumptions about the political competition and information structure to endogenously generate the structure of pay-offs where reference-dependence will be relevant. For example, Patty (2006) emphasizes non-representative turn-out differences during midterm Congressional election, stemming from the particular framing of seats in terms of losses and gains.

3.1 Career concerns and loss-averse voters

The model has two separate, but related, parts: (I) loss-averse voters have reference-dependent utility that generates incumbency advantage in good times and disadvantage in bad times, (II) "Career-concerned" politicians vary in ability and they want to look good before the election by providing a non-trivial amount of the public good by not shirking. Having (II) in addition to (I) lets politicians respond to signals irrel-

evant for determining their ability that affect voters' consumption. As the incumbent becomes disadvantaged with risk-seeking voters having their consumption in a loss-region, politicians increase expenditure on the public goods, so they don't look bad before the election.

A starting point is to use endogenous formation of the reference point as in Koszegi and Rabin (2007) and apply it to a model of post-election politics. That framework uses a mix of expected utility component and a gain-loss function of prospect theory, where the innovation is that reference-point is formed with forward-looking rational expectations. Taking equilibrium behavior of voters and the incumbent as given (to be described later), voters form a reference point for lottery of the consumption that they rationally expect to receive after the election as a function of the incumbent's state[2]. This reference point will be used when the voters come to the polls and weigh different outcomes resulting from their choice.

The baseline structure of the model is similar to the two-period "career concerns" agency model (Persson and Tabellini, 2002). The incumbent wants to impress voters with his quality in the first period, so they reelect him again. This signalling is expensive and requires a cut to politicians' personal rents. The present paper adds the concept of a reference-dependent personal equilibrium (Koszegi and Rabin, 2007) to the voter preferences.

First, I characterize the representative voter's personal equilibrium and incumbent's choice of provision of public goods and extraction of political rents. The reference-point consumption lottery describes citizens' expected consumption next period before the election takes place. It is the weighted sum of getting the challenger-induced consumption with probaiblity q and incumbent-induced consumption with probaiblity $1-q$, where $q, 0 < q < 1$, is the ex-ante rational expectation of incumbent losing. Here q represents voters' expectation of who is going win the election. In

[2]Voters anticipate to correctly derive it in equilibrium from the government's public good provision in the initial stage, so that the incumbent's type is known when the vote takes place.

equilibrium, this expectation is rational and voters do pick the challenger q^* fraction of the time, given the equilibrium reference-point consumption lottery with q^* weight.

Therefore, by endogenizing the rational reference point as in Koszegi and Rabin (2007), the same number q^* parametrizes both the reference-point consumption lottery and incumbent's probability of losing. We can now describe incumbency advantage and disadvantage in terms of the same q^*.

q^* is said to characterize *incumbent's advantage* for $q^* < \frac{1}{2}$ and *incumbency disadvantage* for $q^* > \frac{1}{2}$. The main finding of Theorem 3.1 is that during good times, every personal equilibrium has $q^* < \frac{1}{2}$ and it captures incumbency advantage comparable in magnitude to Gelman and King (1990)'s $2 - 10\%$ estimates. This is intuitively driven by risk-aversion to picking challenger's consumption lottery.

Instead, suppose the reference point was feasible in the sense it was a linear combination of consumptions induced by the challenger and the incumbent, weighted by q, but not necessarily rational. Then Theorem 3.1 also shows there is a unique cutoff strategy parametrized by η_q, the incumbent's probability of losing. During good times, for for all q, $0 < \eta_q, < \frac{1}{2}$, so that there is incumbency advantage even for irrational feasible reference points that do not form a personal equilibrium.

Suppose a voter finds it optimal to reelect incumbent in two-thirds of possible states (of incumbent's performance), given a reference-point compound lottery of receiving challenger's consumption lottery q percent of the time and receiving incumbent's consumption lottery $1 - q$ percent of the time. This would be a personal equilibrium if only if $q = \frac{1}{3}$.

Either way, it turns out the "career-concerned" incumbent provides more public goods when his probability of losing increases[3].

Second, while fixing voter's ERPCL, I look at how agents react to the surprise negative shock to the voter's income in the second period that is realized just before

[3]In the particular linear parametrization of this paper, the amount of the public good is directly proportional to the probability of losing

the election. This shock is exogenous to the economy and specifically to incumbent's actions (e.g. an earthquake, flu epidemic, global economic slowdown). For tractability, the first pass of this analysis assumes that voters assigned a zero probability to this rare shock when they formed their equilibrium reference-point lottery before the election. The main finding here is for small negative shocks to future income, incumbency advantage diminishes even without strict convexity of the gain-loss function μ. With strict convexity in the loss region, there is incumbency disadvantage for moderate negative shocks.

It is for the analysis of incumbency disadvantage and comparative statics where endogenizing the reference point is especially useful. The surprise shock model takes an equilibrium reference-point lottery from the no-shock environment and generates a voting decision rule $Q(q;s)$ that varies with the size of the shock s. When $q < Q(q;s) < \frac{1}{2}$, the incumbency advantage is said to decrease from the shock, for example when the gain-loss function μ is piece-wise linear. Similarly, when $q < \frac{1}{2} < Q(q;s)$, it describes incumbency disadvantage, such as when μ is strictly convex. In both cases the incumbent provides more public goods because $Q(q;s) > q$ as the probability of losing increases.

Finally, the surprise model is extended to a model with fully rational expectations where the negative income shock s happens with probability $p, 0 < p < 1$, and income y remains unchanged with $1 - p$ probability. In the limit as $p \to 0$, the rational-expectations model captures the surprise model as a special case.

3.1.1 Politicians

The government taxes voters in both periods at a constant rate τ and provides public goods $g_t, t \in \{1, 2\}$. At the end of every period, voters vote in favor of the incumbent or the single challenger. Politicians differ in their ability η to convert private goods into public goods.

We will assume that this ability η is uniformly distributed on $[0, 1]$ for all politicians. That is, ex-ante the incumbents and challengers are equally skilled and the deck is not stacked to generate incumbency (dis)advantage in an ad hoc manner by drawing challengers from a different pool. The only distinction is that the equilibrium play will fully reveal incumbent's type but off-equilibrium incumbents can consider altering their hidden action to signal a different type.

When converting one unit of private good, η units of public good are produced. Once a new politician is picked for office, his competence η remains fixed throughout his career (no learning-by-doing) but no one knows the value of η during his first term (not even himself as the job is new to him). The politician will learn his competence after one term and voters will be able to indirectly, though accurately, infer it in equilibrium just before casting their vote. If a challenger politician is elected, η^c is drawn from the same distribution as incumbent's η^i.

Politicians are purely opportunistic - they only care about extracting personal rent r_t out of taxes and exogenous ego rents of being reelected, R (which may capture continuation value of future rents).

Hence, the politician's value function is

$$v_I = r_1 + \beta p_I(r_2 + R) \tag{3.1.1}$$

where p_I is the probability of being elected, which is driven by voter's decisions; r_t are the rents extracted in period t; $0 < \beta < 1$ is the discount factor; R is the value of being reelected to office.

Voters do not observe rents extracted or η^i but they judge competence based on how many public goods were provided. Incumbent's trade-off to extract everything is to appear incompetent and lose election and forgo future rents. At the beginning of every period $t \in \{1, 2\}$, the incumbent balances the budget between personal rents

and public goods. Since in the last period ($t = 2$) there is no reelection incentive, $r_2^* = \overline{r} < y\tau$, the maximum allowed by feasibility.

The budget constraint for the incumbent in period t is:

$$\eta^i \tau y = \eta^i r_t + g_t \Longrightarrow g_t = \eta^i(\tau y - r_t), \tag{3.1.2}$$

where the action to take is $r_t \in [0, \overline{r}]$ and g_t is residually determined.

The information is symmetric for tractability as the principal-agent story is not appropriate for studying exogenous shocks like an earthquake. At the same time, the ex-ante outcome of elections is non-deterministic – the challenger has a viable chance to win and neither politician is endowed with a competence advantage; there is no deterrence effect here either. Since the incumbent cannot condition his personal rent strategy r_1 on own not-yet-realized type, all incumbents extract the same equilibrium amount of rents r_1^*, so that the amount of public goods provided g_1 becomes a random variable, perfectly correlated with η^i.

3.1.2 Voters

In this model voters only care about their own consumption and do not have an ideological party bias. Their consumption utility in period t is linear in consumption, c, which is composed of disposable income $y(1 - \tau)$ plus the public good:

$$m(c) = c = y(1 - \tau) + \alpha g_t, \tag{3.1.3}$$

where $y > 0$ is the fixed income for both periods, $0 < \tau < 1$ is the fixed taxed rate, $\alpha > 1$ is the preference for public goods and g_t is the amount of the public goods provided.

Koszegi and Rabin (2007) bridged the gap between expected utility and classical prospect theory that only looks at the gains losses by considering both. Voter's utility

in the period after the election is:

$$U(F|G) = \int \Big(\int u(c|r) dG(r) \Big) dF(c) = \int \Big(\int m(c) + \mu(m(c) - m(r)) dG(r) \Big) dF(c), \tag{3.1.4}$$

where F is the consumption lottery, G is any reference (consumption) lottery and $\mu(.)$ is the gain-loss function.

We will assume the following parametric form for the gain-loss function that allows for linearity or strict concavity. While piece-wise linearity has been used in previous research for tractability, risk-seeking in the losses is required to generate strict incumbency disadvantage, rather than simply decreasing the advantage.

$$\mu(x) = \begin{cases} \gamma x^{\frac{1}{k}} & \text{if } x > 0, \\ -\gamma\lambda(-x)^{\frac{1}{k}} & \text{if } x \leqslant 0. \end{cases} \tag{3.1.5}$$

where $\gamma > 0$ is a scale parameter for gains and losses relative to the base utility for consumption ($\gamma = 0$ is expected utility); $\lambda > 1$ scales the degree of loss-aversion relative to gains; $k \in \mathbb{N}$ is a curvature parameter: $k = 1$ corresponds to non-strict concavity and convexity of the standard piece-wise linear gain-loss function and $k > 1$ allows for risk-seeking in the losses.

With only two political candidates, there is no problem of strategic voting. Unlike principal-agent models, where voters are sometimes asked to commit to retrospective punishment strategies, voters here are forward-looking when they select the candidate to maximize their next period's utility, conditional on the reference point.

The voter observes g_1 before voting but not η^i. If the voter picks the incumbent, then he will receive the degenerate consumption lottery:

$$c^i = y(1-\tau) + \alpha g_2^i = y(1-\tau) + \alpha\eta^i(\tau y - \overline{r}) \tag{3.1.6}$$

which is increasing in η^i, which will be inferred accurately in equilibrium from g_1, so c^i is not a random variable.

If the voter picks the challenger, then he will receive non-degenerate consumption lottery:

$$c^c = y(1-\tau) + \alpha g_2^c = y(1-\tau) + \alpha\eta^c(\tau y - \overline{r}) \tag{3.1.7}$$

Because η^c is a random variable that has not been realized, c^c is also a random variable, which inherits the uniform distribution.

Denoting $\theta, \theta \in [0,1]$, to be the random realization of the challenger's ability, an arbitrary reference point is characterized by a fixed $q \in [0,1]$

$$\begin{aligned} c_q &= qc^\theta + (1-q)c^i = q\left(y(1-\tau) + \theta G\right) + (1-q)\left(y(1-\tau) + \tilde{\eta}G\right) \\ &= y(1-\tau) + q\theta G + (1-q)\tilde{\eta}G \end{aligned} \tag{3.1.8}$$

where $G \equiv \alpha(\tau y - \overline{r})$, the public-goods production technology.

It is a lottery where the challenger's lottery is drawn with probability q. As before, $\tilde{\eta}$ is the imputed incumbent's ability from his equilibrium play. As it will be known at the moment of the election, when considering next period's gains and losses, the voter will treat $\tilde{\eta}$ as a degenerate (constant) lottery that puts mass 1 on $\tilde{\eta}$.

Given the reference lottery c_q, the utility of voting for the challenger is evaluated as:

$$U(c^c|c_q) = \int_0^1 \int_0^1 \Big(c^c + \mu(c^c - c_q)\Big) d\theta d\eta^c \tag{3.1.9}$$

Likewise, given the reference lottery c_q, the utility of voting for the incumbent is evaluated as:

$$U(c^i|c_q) = \int_0^1 \Big(c^i + \mu(c^i - c_q)\Big)d\theta \tag{3.1.10}$$

Finally, the voter picks the incumbent when $U(c^i|c_q) \geqslant U(c^c|c_q)$.

3.2 Equilibrium

3.2.1 Incumbent's choice of rents today, r_1

Let $r_1 \in [0, \tau y]$ be the incumbent's choice of rents. Suppose in equilibrium the voters observe g_1, infer $\tilde{\eta} = \frac{g_1}{\tau y - \tilde{r}_1}$, where $\tilde{r}_1$ is incumbent's equilibrium rent (known to the voters) and are following some cutoff rule[4].

$$\tilde{p_I} = \begin{cases} 1 & \text{if } \tilde{\eta} \geqslant Q(\lambda, \gamma; \alpha, \tau, k, \overline{r}) \equiv q, \\ 0 & \text{if } \tilde{\eta} < Q(\lambda, \gamma; \alpha, \tau, k, \overline{r}). \end{cases} \tag{3.2.1}$$

where $Q(\lambda, \gamma; \alpha, \tau, k, \overline{r}) \equiv q \in [0, 1]$

Note that today's public goods don't enter into voter's utility function for tomorrow's gains and losses, so the incumbent's choice of r_1 does not affect q. Here there is a one-way channel from the citizen decision rule to the public goods provision, through choice of rent depending on q. This is enough to generate the key stylized facts. The government's second period's decision is fixed by extracting the maximal feasible personal rent, $r_2 = \overline{r}$, and using the rest of the taxes to make the public good. In a dynamic model with an interior decision for next period's rent r_2, there would also be a reverse feedback channel, where the amount of public good could affect tomorrow's gains and losses. This richer model could allow for more patterns

[4]To be shown optimal in Sec. 3.1

of disaster-relief spending.

$$p_I = Pr(p_I = 1) = Pr(\tilde{\eta} \geqslant q) = Pr\Big(\frac{g_1}{\tau y - \tilde{r}_1} \geqslant q\Big)$$
$$= Pr\Big(\frac{\eta^i(\tau y - r_1)}{\tau y - \tilde{r}_1} \geqslant q\Big) \tag{3.2.2}$$

Since η^i not known to the incumbent at the time of choosing the rent and has uniform distribution on $[0, 1]$,

$$Pr(\eta^i \leqslant x) = x \Longrightarrow Pr(\eta^i \geqslant x) = 1 - x. \tag{3.2.3}$$

Thus, the probability of incumbent being reelected can be rephrased in terms of the uniform distribution over the signal of the public good his choice of r_1 generates.

$$p_I = Pr\Big(\eta^i \geqslant \frac{(\tau y - \tilde{r}_1)}{\tau y - r_1} q\Big)$$
$$= 1 - \frac{(\tau y - \tilde{r}_1)}{\tau y - r_1} q \tag{3.2.4}$$

The decision problem, given $r_2 = \overline{r}$ becomes:

$$\max_{r_1} v_I = r_1 + p_I \beta(R + \overline{r}) =$$
$$r_1 + \Big(1 - \frac{(\tau y - \tilde{r}_1)}{\tau y - r_1} q\Big)\beta(R + \overline{r}) \tag{3.2.5}$$

FOC:

$$1 + \beta(R + \overline{r})\Big(- \frac{(\tau y - \tilde{r}_1)}{(\tau y - r_1)^2} q\Big) = 0 \tag{3.2.6}$$

Since in equilibrium voters must have correct anticipation, $r_1 = \tilde{r_1}$. Then FOC becomes:

$$1 + \beta(R + \overline{r})\Big(- \frac{1}{(\tau y - r_1)} q\Big) = 0 \Longrightarrow \tau y - r_1 = \beta(R + \overline{r}) q \tag{3.2.7}$$

Thus,

$$r_1^* = \tau y - \beta(R + \overline{r})q. \tag{3.2.8}$$

The equilibrium probability of winning comes from the voter's cutoff: $p_I^* = 1 - q$. Also,

$$\boxed{g_1^* = \eta^i \beta(R + \overline{r})q} \tag{3.2.9}$$

The amount of public goods provided increases (linearly) in the ex-ante probability of losing, q.[5]

$$\frac{dg_1^*}{dq} = \eta^i \beta(R + \overline{r}) > 0 \tag{3.2.10}$$

(almost surely)[6]. A decrease of incumbency advantage is good for the voters' welfare as they get more public goods in the first period (though irrelevant for their voting decision).

The equilibrium ex-ante value accruing to the incumbent is:

$$\boxed{V_I^* = \tau y - \beta(R + \overline{r})q + (1 - q)\beta(R + \overline{r}) = \tau y + (1 - 2q)\beta(R + \overline{r})} \tag{3.2.11}$$

Naturally, a decrease of incumbency advantage is bad for the incumbent:

$$\frac{dV_I^*}{dq} = -2\beta(R + \overline{r}) < 0, \tag{3.2.12}$$

3.2.2 Voters' personal equilibrium

We are now ready to define the solution concept for the voters' problem. Consider an exogenous reference point c_q putting a weight $q \in [0, 1]$ on electing the challenger and a weight $1 - q \in [0, 1]$ on picking the incumbent. After the incumbent picks his

[5]Incumbent that never expects to lose (if $q = 0$) provides no public goods in the first period. Note, however, that the value of q depends on a forward-looking consideration of $U(c^i|c_q)$ vs. $U(c^c|c_q)$, so if the pool of challengers' competence was not $[0, 1]$ but some inferior set, $q^* = 0$ may conceiveably turn out to be a unique subgame-perfect refinement where voters would suffer from inability to punish restrospectively.

[6]$\Pr(\eta^i = 0) = 0$

private rent r_1^*, the voter observes incumbent's realized ability $\tilde{\eta}$ and elects either the incumbent or the challenger. The strategy $\tilde{p_I}(\tilde{\eta}; q) : [0,1] \to \{0,1\}$ is optimal if it maximizes the next period's the reference-dependent utility, given c_q.

Thus, voters optimal choice is a pair $\big(q, \tilde{p_I}(\tilde{\eta}; q)\big)$. For each decision rule, we can compute incumbent's probability of losing and it will generally vary with q. When this losing probability is different from q, it is inconsistent with rational expectations. We can think of endogenizing the reference point as a process of equilibrium selection to be consistent with rational expectations. This equilibrium is then analogous to "unacclimating personal equilibrium" (UPE) from Koszegi and Rabin (2007). If UPE is unique, then it is also optimal and becomes their "preferred personal equilibrium" (CPE).

Suppose that voters are forward-looking and sometime before they go to the polls, they form rational-expectations about their decision process. The endogenous reference point q^* will also equal to the incumbent's probability of losing in a personal equilibrium.

1. Given any reference point q, a rational decision rule of voting for the incumbent $\tilde{p_I}(\tilde{\eta}; q)$ equals 1 when $U(c^i|c_q) \geqslant U(c^c|c_q)$ and 0 when $U(c^i|c_q) < U(c^c|c_q)$.

2. Given any rational decision rule $\tilde{p_I}(\tilde{\eta}; q_0)$, the ex-ante rational expectation of the challenger winning $E[1 - \tilde{p_I}(\tilde{\eta}; q_0)] = 1 - Pr\Big(U(c^i|c_{q_0}) \geqslant U(c^c|c_{q_0})\Big) = q$.

Personal equilibrium for voters is q^* iff $q^* = 1 - Pr\Big(U(c^i|c_{q^*}) \geqslant U(c^c|c_{q^*})\Big)$.

The following Theorem 3.1 shows that for *any* reference point c_q putting a weight $q \in [0,1]$ on electing the challenger and a weight $1-q \in [0,1]$ on picking the incumbent, there is a unique decision rule that picks the incumbent with probability over $\frac{1}{2}$, that is $q < \frac{1}{2}$. This is true for all feasible reference points, not just rational ones. In other words, there is incumbency advantage for every feasible reference point c_q with

$q \in [0,1]$. Of course, all personal equilibria [7]. also experience incumbency advantage because they stem form a subset of feasible reference points.

Theorem 3.1 (Incumbency Advantage)**.** *Suppose the loss-aversion coefficient is* $\lambda > 1$*, the gain-loss weight is* $\gamma > 0$ *and the convexity parameter is* $k \in \mathbb{N}$*. For any fixed reference point* $q \in [0,1]$*, there is a unique cutoff decision rule* $Q(q) = \eta_q^* : [0,1] \to (0,1/2)$ *such that the optimal voting rule is a cutoff for incumbent's realized ability:*

$$\tilde{p_I}(\tilde{\eta}; q) = \begin{cases} 1 & \text{if } \tilde{\eta} \geqslant \eta_q^*, \\ 0 & \text{if } \tilde{\eta} < \eta_q^*. \end{cases}$$

Proof Outline. The detailed verifications are in the Appendix B.2.

1. Let $f(\tilde{\eta}) = U(c^i|c_q), g(\tilde{\eta}) = U(c^c|c_q)$ be the utilities of picking the incumbent and the challenger, given c_q reference point.

2. f is trivially *strictly increasing* (take derivative). Intuition: choosing the incumbent for sure is better when he is more able. The expected utility component increases and the gain-loss component of incumbent vs. the reference lottery of the incumbent plus the challenger improves.

3. g is *weakly decreasing* in $\tilde{\eta}$, for all $q \in [0,1]$ and $k \in \mathbb{N}$ (take derivative). The expected utility component from the challenger is unchanged and the gain-loss component worsens (in FOSD sense). g is strictly decreasing for $q < 1$ and constant in $\tilde{\eta}$ when $q = 1$ (the incumbent never wins in the reference point).

[7] The existence of personal equilibria be shown in Proposition 3.3.

4. $(f-g)(1/2) > 0$: picking an average incumbent is strictly better than an unknown challenger. It can be shown that if $q = 0$ then

$$(f-g)\left(\frac{1}{2}\right) = (\lambda-1)\frac{k\gamma}{(k+1)2^{\frac{k+1}{k}}} > 0.$$
$$(f-g)\left(\frac{1}{2}\right) = (\lambda-1)\frac{M}{2^{\frac{2+k}{k}}}\left(k(1+q)^{\frac{2k+1}{k}} - k(1-q)^{\frac{2k+1}{k}} - 2(2k+1)q^{\frac{1+k}{k}}\right)$$
$$\text{if } q \in (0,1],$$

where $M \equiv \frac{\gamma k G^{1/k}}{q(1+k)(2k+1)} > 0$ and the second term in the product is positive by Lemma B.7, using binomial series around $q = 0$.

5. $(f-g)(0) < 0$: picking worst incumbent is strictly worse than an unknown challenger.

6. *Intermediate value theorem* on [0,1], $(f-g)$ strictly increasing, and thus has a unique root on (0,1). Denote this root as η_q^*. This is the unique value solves $U(c^i|c_q;\eta_q^*) = U(c^c|c_q;\eta_q^*)$. It is the interior threshold when the citizen is indifferent between a random challenger and an incumbent of known ability η_q^* while the reference point is fixed at q.

□

If loss-aversion is removed because losses and gains are treated equally ($\lambda = 0$) or the gain-loss function has no weight ($\gamma = 0$), then consumer maximizes expected utility and he is completely indifferent between the incumbent and the challenger.

Corollary 3.2 (Parity under expected utility)**.** *If $\lambda = 1$ or $\gamma = 0$, then the unique decision rule for any $q \in [0,1]$: is $\eta_q = \frac{1}{2}$*

Proof. Same as above but now $(f-g)(1/2) = 0$ in Step 4. □

The following proposition shows that the optimal map $Q(q)$ from Theorem 3.1 has a fixed point, proving existence of a personal equilibrium.

Proposition 3.3 (Existence of Personal Equilibrium)**.** *Let $Q : [0,1] \to (0,1/2)$ be the unique cutoff decision rule from Theorem 3.1. Then there exists a personal equilibrium (fixed point) such that $q = Q(q)$. Furthermore, every personal equilibrium $q^* \in \left(0, \frac{1}{2}\right)$.*

Proof Outline.

- Theorem 3.1 showed that there is a unique exogenous equilibrium (unique decision rule η_q) for each q.
- The expression $U(c^i|c_q) - U(c^c|c_q)$ is continuous in q because it's a difference of integrals of gain-loss functions $\mu(x)$, which were continuous in q.
- Thus, $(f-g)(\eta) = U(c^i|c_q)(\eta) - U(c^c|c_q)(\eta)$ is continuous in q on $[0,1/2]$. It is a known result that unique real root of an algebraic expression that is continuous in parameters, is also continuous in q as per Henriksen and Isbell (1953).
- So $Q(q) : [0,1/2] \to [0,1/2]$ is continuous, and has at least one fixed point on [0,1/2] by Brouwer's fixed point theorem.
- Theorem 3.1 showed that $q = 0$ and $q = 1/2$ are not fixed points as there is a strict preference for one of the candidates when $\tilde{\eta} = q$ (optimal decision rule maps to $(0,1/2)$).

It remains to show that any fixed-point $q^* = Q(q^*)$ does, in fact, satisfy the second condition of the personal equilibrium. That is, ex-ante probability of the incumbent's loss equals to q^*. At period 0, g_1 is not yet observed, so $\tilde{\eta}$ cannot be inferred and is a random variable with uniform distribution along the equilibrium path:

$$\begin{aligned} \tilde{\eta} &= \frac{g_1}{\tau y - \tilde{r}_1} = \frac{\eta^i(\tau y - \tilde{r}_1)}{\tau y - \tilde{r}_1} \\ &= \eta^i \sim U[0,1]. \end{aligned} \tag{3.2.13}$$

Recall that the optimal decision rules p_I are step functions with η_q^* cutoff for incumbent's ability and he loses for $\tilde{\eta}$ realizations below the corresponding cutoff. In any

equilibrium, q with decision rule p_I the incumbent loses with probability equal to the corresponding cutoff rule $Q(q) = \eta_q^*$:

$$E[1 - p_I] = 1 - Pr[\tilde{\eta} \geqslant \eta_q^*] = Pr[\eta^i \leqslant \eta_q^*] = \eta_q^*. \tag{3.2.14}$$

But when q is a fixed point, then indeed $q = E[1 - p_I] = \eta_q^*$. □

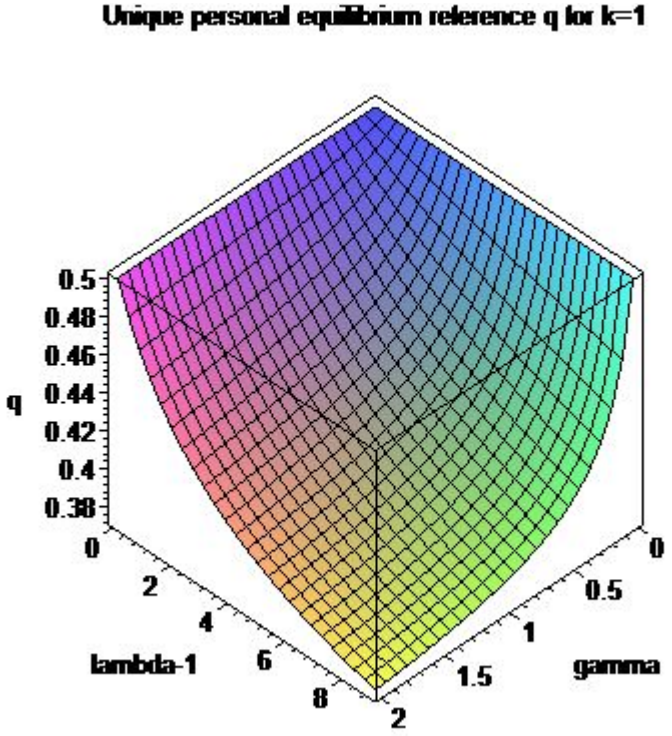
In a special case where $k = 1$ (piece-wise linear μ), setting the reference point q weight on the challenger lottery equal to the probability of the challenger being chosen, given that reference point $0 \leqslant q = \eta_q^* \leqslant 1$, generates the following fourth-degree polynomial in q:

$$3(\lambda - 1)\gamma q^4 - 12q^3\gamma(\lambda - 1) + 13\gamma(\lambda - 1)q^2 + (-3\gamma(\lambda - 1) + 6\gamma + 6)q - 3 - 3\gamma = 0 \tag{3.2.15}$$

By using the exact solution to a quartic equation, it can be shown to have a unique root on $(0, 1/2)$ for $\gamma > 0, \lambda > 1$, which is graphed in 3.1. Thus, the personal equilibrium $q^* = \eta_q^*(k = 0)$ is independent of $(\alpha, \tau, y, \overline{r})$ but generally when allowing for concavity of the gain-loss function, $k_0 \geqslant 2$, the resulting algebraic expression in q will depend on all of these parameters and the relevant algebraic equation has fractional powers.

As expected, along the boundaries of Figure 3.1 with $[\gamma = 0, \lambda]$ and $[\gamma, \lambda = 1], q = \eta_q^* = \frac{1}{2}$. There is no incumbency advantage when loss-aversion stops to matter ($\gamma = 0$) or pain from losses equals pleasure from gains ($\lambda - 1 = 0$), expected value of the challenger equals average incumbent. Here it is important to note that utility $m(c)$ is linear in both private and public consumption outcomes. These values of incumbency advantage are reasonably comparable to estimates in (Gelman and King, 1990) of 2-10%.

Figure 3.1: Unique personal equilibrium reference q for linear gain-loss ($k = 1$)



Although for $k > 1$, there is no closed-form expression for solutions for the fixed point $q = \eta_q^*$, the decision rule iteration $q^{n+1} = \eta_{q_n}^*$ converges rapidly $\forall q \in [0, 1]$, to the same fixed point with more than 1 decimal digit per iteration, even when initial conditions are chosen as far as possible from the center. The numerical evidence suggests uniqueness of personal equilibrium is generic because for a various picks of $(\alpha, \tau, y, \overline{r})$, the decision rule η_q^* is increasing and concave on a compact set, which are conditions for a fixed-point theorem in (Kennan, 2001). While the slope of the decision rule may be steep near $q = 0$ for some parameters λ, k it gets much flatter as it approaches $q = 1/2$.

3.3 Responses to surprise crisis

Suppose just prior to election there is a surprise event that affects the voters, a negative shock (e.g. earthquake or flu) to future income that they did not anticipate when

playing their reference-point equilibrium q^* from Section 3.2. Before we formulate a rational expectations model where the reference-point takes into account the true probability p of the crisis, we will consider a simpler limiting case of a surprise shock. This is equivalent to taking $p \to 0$ in the rational-expectations model. Nonetheless, the surprise case is a generalization of the model with no shock in the sense that taking $s \to 0$ returns the baseline "no shocks" model from the previous section. For now, consider the surprise model as a model of convenience to leverage the existence of personal equilibrium result (Prop. 3.3) in a model without shocks and then using this reference point to verify the second and the third of the stylized facts: (i) there is incumbency disadvantage for moderate shocks s for $k > 1$ and (ii) the incumbent's response to crisis is to increase public goods provision[8].

First, voters anticipate a reference consumption c_{q*} lottery expecting no shock. Then they observe a negative shock to future income, $y_2 = y - s$, before they go to the polls. Thus, they modify their voting decision, so that, in general, $q^* = \eta^*_{q*}(y) \neq \eta^*_{q*}(y - s)$ and the later is the decision rule they use.

If the incumbent has already committed to the rents and public goods choice by expecting $\eta^*_{q*}(y)$, then the agents can still infer the same $\tilde{\eta}$ correctly. Furthermore, incumbency disadvantage arises but there is no response to unanticipated crisis.

If the incumbent has enough time to change his choice of r_1 (η^i not yet realized), then by previous argument from Sec. 3.2, his rent $r_1^* = \tau y - \beta(R + \overline{r})q$ is decreasing linearly in probability of losing, $q(s) = 1 - E[\tilde{p_I(s)}]$. Similarly, his public good provision $g_1^* = \eta^i \beta(R + \overline{r})q$ is increasing in his probability of losing $q(s)$. Thus, the model predicts that the government will "respond" to an unexpected crisis to the extent that it loses its incumbency advantage. But this response is not rewarded

[8]Recall that voters' are forward-looking and next period's public goods are a corner solution of maximal rent, so the incumbent's response to crisis is via career-concern approach as a trade-off between signals is linear in losing probability q. This raises citizen's utility today but after conditioning on politicians' type, does not enter into their choice of tomorrow's politician.

directly[9], rather deviation to not responding would signal low competence and very low chance of reelection. This is broadly consistent with the finding by (Cole, Healy, and Werker, 2008) that Indian administrations that responded to natural disasters (as if $\tilde{\eta}$ is high) did better than those who were less vigorous (as if $\tilde{\eta}$ is low) but not as well as administrations that did not have a disaster on their term (as if $q(s)$ is increasing in s at $s = 0$).

As s increases, the loss region begins to grow and the gain region shrinks. There are two effects: a direct loss of disposable income to buy private goods, $-s(1-\tau)$, and a decrease in tax revenues available for producing public goods (magnitude of loss depends on politician's realized competence).

When picking the incumbent, the only source of uncertainty in $U(c_s^i|c_q; s)$ stems from the unknown challenger in the reference point. Suppose the voter observes $\tilde{\eta}$ and $s > 0$. For small values of $\tilde{\eta}$, the voter is entirely in the loss region because even if the challenger draw is 0, the flat amount $-s(1-\tau)$ dominates the realized gain from $\tilde{\eta}$. For large values of $\tilde{\eta}$, the voter is in the gains against low reference (challenger) draws θ and in the losses against high θ. Thus, Eq. B.1.52[10] derives $f(\tilde{\eta}) = U(c_s^i|c_q; s)$ to be a piecewise-continuous, increasing function in two segments in terms of $\tilde{\eta}$. As the gain-loss function $\mu(c_s^i - c_q)$ is integrated over $\theta \in [0, 1]$, there is a kink (λ valued) in at most one point, which moves towards $\theta = 0$ as $\tilde{\eta}$ decreases to a low cutoff.

When picking the challenger, there is an additional source of uncertainty in $U(c^c|c_q; s)$ from comparing the actual (not-yet-realized) challenger draw against all values of the reference point lottery. The gain-loss function $\mu(c_s^c - c_q)$ is integrated over $\theta \in [0, 1] \times \eta^c \in [0, 1]$, with a kink ($\lambda$ valued) along a straight line in the $[0, 1]^2$ region that moves towards a corner as $\tilde{\eta}$ increases. The shape of the gain-loss region changes twice as the kink fold moves through the $[0, 1]^2$ space – the Eq. B.1.36 derives

[9] In fact, losing incumbency advantage (increase in q) means the incumbent government is worse-off.

[10] The equation is actually from the rational expectations model but setting $p = 0$ does not change the result.

$g(\tilde{\eta}) = U(c_s^c|c_q; s)$ to be a piecewise-continuous, decreasing function in three segments in terms of $\tilde{\eta}$.

Therefore, both $f(\tilde{\eta}) = U(c^c|c_q; s)(\tilde{\eta})$ and $g(\tilde{\eta}) = U(c_s^i|c_q; s)(\tilde{\eta})$ have analytical expressions, each is monotonic (See Appendix B for details) and have a unique intersection. The optimal cutoff $Q(q; s) = \eta_q(s)$ has to be determined numerically[11]

First, for fixed $(\alpha, y, \tau, \overline{r}, k, \lambda)$, we get personal equilibrium q^* in the no-shock model using $q^{n+1} = \eta_{q_n}^*$ iteration. Convergence is fast, about 1 decimal place per iteration. The fixed point is independent of s and arises as an endogenous reference point.

Second, for a fixed a grid of s values, for each fixed s find $\tilde{\eta}(s)$ that makes the voter indifferent between candidates, satisfying the piecewise-continuous, algebraic equation $U(c_s^i|c_{q^*}; s)(\tilde{\eta}) = U(c^c|c_{q^*}; s)(\tilde{\eta})$ as explained above. This describes how the optimal decision rule $Q(q^*; s)$ varies with s such that $q^* = Q(q^*; 0)$.

At $s = 0$, we have $Q(q^*; 0) = q^* < 1/2$ by Theorem 3.1 as before, but for $s > 0$ incumbency advantage starts to diminish $q^* < Q(q^*; s)$. If also $k > 1$, strict convexity gives incumbency disadvantage in Figures 3.2 and 3.3: for s large enough, $Q(q^*; s) > \frac{1}{2} > q^*$.

While this model violates rational expectations for large p, probability of crisis, the model is a good approximation to rational expectations for small p, when inrepreted correctly. Although we noted that in the no-shock model, whenever the reference lottery anticipates elections in a matter inconsistent with the actual decision, given that lottery, we said that $q \neq \eta_q^*$ showed the model violated rational expectations. However, given $s > 0, Q(q^*; s) \nrightarrow q^* asp \rightarrow 0$.

In the next Sec. 3.4, the rational expectations equilibrium is going to be a *pair* of conditional decisions $\{q_{ns}(p), Q_s(p)\}$ – how often the voters pick the challenger when observing no shock and shock, respectively. In the limit as $p \rightarrow 0$, we can interpret

[11]The thresholds for $\tilde{\eta}$, where the functional form of each utility function changes, are non-linearly depend on s, see Figures B.4 and B.2.

this $Q(q^*; s) = \lim_{p\to 0} Q_s(p)$ as the probability of picking the challenger when shock of s magnitude *is observed* along this sequence of rational-expectation models, with the observed shock becoming less and less likely. The corresponding $q^* = \lim_{p\to 0} q_{ns}(p)$ that was derived as the equilibrium of the $s = 0$ model would match the probability of the picking the challenger when the shock *is not observed* along this sequence of the rational-expectations models, with the unobserved shock becoming less and less likely. Therefore, there is no reason to expect these limits to be the same as they are reached along entirely different sequences.

Figure 3.2 shows, for various curvatures of the gain-loss function, the effect of increasing negative shocks to future income(%) on incumbency advantage. The values of the parameters used are as follows: $(\lambda, \gamma, \alpha, y, \tau, \overline{r}) = (3, 1, \frac{5}{2}, 1, \frac{2}{5}, \frac{1}{10})$. For $k = 1$ incumbency advantage diminishes as the magnitude of catastrophe increases (no strict convexity in the loss region). For $2 \leqslant k \leqslant 5$ (higher curves respectively) risk-seeing in the losses allows for stronger effect as was predicted and even incumbency disadvantage (peaking at $s = 0.2y$. For large $s, q(s)$ is decreasing towards neutrality because risk-seeking is predominant for "small" losses and not for "large" losses.

Figure 3.3 is the effect of highr loss-aversion as a sort of limiting guideline. The values of the parameters used here are as follows: $(\lambda, \gamma, \alpha, y, \tau, \overline{r}) = (100, 10, \frac{5}{2}, 1, \frac{2}{5}, \frac{1}{10})$. Note that $\gamma_2 = 10 >> \alpha = \frac{5}{2} > \gamma_1 = 1, \lambda_2 = 100 >> 3 = \lambda_1$. The effects of losses and gains has been exaggerated through increase in γ and losses have higher weight than gains (large λ). For $s \to 0, k = 1$, strong risk-aversion gives large incumbency advantage as $q = 0.352$ On the other end of the spectrum, $s = 0.2, k = 5$ gives $q = 0.646$ - a comparable incumbency disadvantage.

Figure 3.2: Incumbency (dis)advantage as shock increases for moderate loss-aversion, higher convexity upwards.

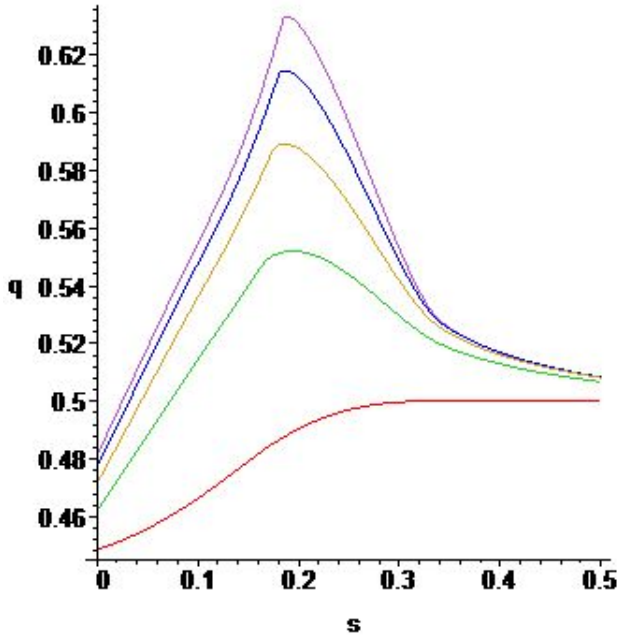
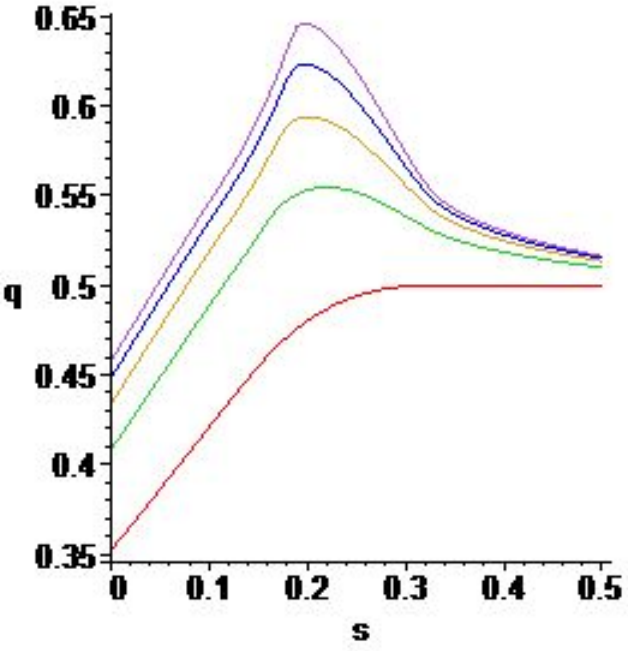


Figure 3.3: Incumbency (dis)advantage as the shock increases for high loss-aversion, higher convexity upwards



3.4 Rational Expectations of Crisis

Suppose the voter observes whether a negative shock to next period's income of s magnitude happens or doesn't happen before he votes. The shock is a Bernoulli event with probability p, uncorrelated[12] with politician's ability draw. We will also assume the challenger's unobserved ability θ is fixed in secret before the crisis is (un)observed. The voter's decision function now depends on $(\tilde{\eta}|s)$. Tomorrow's income is either y or $y - s$.

When the voter conceives of his reference point for voting, he takes into account what information he will know at the time[13]:

1. With probability pQ_s, there is a crisis and the challenger is elected with unknown ability θ, which will be reveiled only in period 2,

2. With probability $p(1 - Q_s)$, there is a crisis and the incumbent is elected with inferred ability $\tilde{\eta}$,

3. With probability $(1 - p)q$, there is no crisis and the challenger is elected with unknown ability θ,

4. With probability $(1-p)(1-Q_s)$, there is no crisis and the incumbent is elected with inferred ability $\tilde{\eta}$.

The reference point is a stochastic lottery to the extent it maintains the residual uncertainty about θ that will remain unresolved in the voting booth when evaluating next period's consumption level.

[12]The shock is an earthquake, change of world price of oil etc, while the ability is personal competence of converting private goods into public goods.

[13]This way the distinction between the incumbent and the challenger remains meaningful. It would be pointless to ignore which information will be reveiled in the interim: the incumbent's ability is also unknown when making his reference point and a key tension would be lost. The incumbent's decision doesn't influence the reference-point formation in the model for tractability, and this is why the incumbent moves after the reference point is already formed.

Let reference point be

$$c_p = p(Q_s c_s^\theta + (1 - Q_s)c_s^i) + (1 - p)(qc^\theta + (1 - q)c^i) \tag{3.4.1}$$

where with probability p the voter will learn there is s shock next period before voting and with probability $1 - p$ he learns there will be no shock next period. Q_s is the cutoff for incumbent, when the shock happens and q the cutoff when no shock happens. Here ability θ emphasizes a hypothetical draw of the challenger in the reference point, giving the corresponding stochastic consumption $c_j^\theta, j \in \{ns, s\}$. For more details calculations of the reference consumption in the rational expectations model, see Sec. B.1. In contrast, when the challenger is elected the actual realization of his ability is η^c, with a corresponding consumption $c_j^c, j \in \{ns, s\}$. It is critical to note that just like in the previous section, η^c is different and independent of θ and the gain-losses are calculated by a double integral over $[0, 1]^2$ in Lemma B.1.

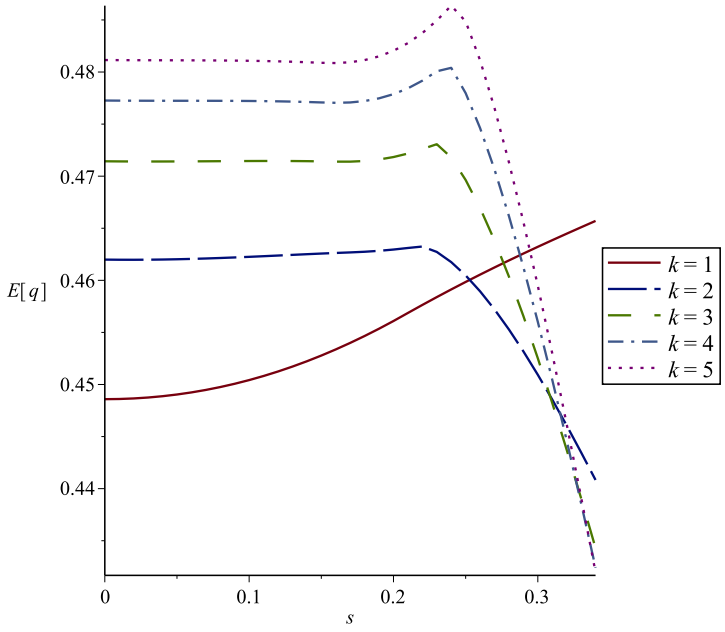
There are two ways to treat the politician's decision. (i) He observes the shock before choosing the rent, which means he knows he'll be losing election with either q(no shock) or Q_s(shock) probability. This is equivalent to the original decision problem, conditional on the state: public goods vary proportionally to the corresponding probability of losing. This assumption that the politician can respond to the shock before choosing his rents in period 1 will be used throughout the rest of the paper.

(ii) He picks rent before he sees the shock. By similar derivation of first-order conditions, get

$$g_1^* = \tau y - r_1^* = \beta(R + \overline{r})(pQ_s + (1 - p)q) \tag{3.4.2}$$

In this case, the amount of public goods is proportional to the ex-ante expectation of losing, $E[q] = pQ_s + (1-p)q$. Figure 3.4 shows that $E[q]$ turns out to be relatively

Figure 3.4: Ex-ante expected probability of picking the challenger, crisis probability $p = \frac{1}{3}$



flat for concave gain-loss function μ because initially for small losses[14], q rises with s at a similar rate as Q_s falls with s and the two effects balance out. In this case, the incumbent maintains an ex-ante advantage, which is reasonable since most incumbents do get reelected. However, in this case the politician cannot respond to the crisis because he commits to the same decision in both states. The central empirical analysis, however, looks at variation of disaster relief with respect to the shock, so it cannot be that the incumbent does not respond.

This model (ii) with $k > 1$ does not match the stylized fact of increased government spending and when $k = 1$, it does not match incumbency disadvantage (merely has decreased advantage) in s. For these reasons, we will assume the incumbent observes s before choosing rent and public goods at least some of the time as per option (i) above. As s increases, we find that q falls and Q_s rises, so there will be fewer public goods when a larger shock is averted and more public goods when a larger shock

[14]The expected loss is increasing in s for $k = 1$ with piecewise-linear gain-loss. Incumbency advantage remains, though diminished.

does happen. Voters' behavioral utility leads to variation in incumbent's reelection chances as S grows and this leads to greater variation in the disaster relief across states.

The *rational-expectations equilibrium* is two pairs $\{\big(q, \eta(q, Q_s)\big), \big(Q_s, \eta^s(q, Q_s)\big)\}$ and satisfies the following 2 conditions:

1. Given any reference point (q, Q_s), in state S the rational decision rule picks the incumbent $\tilde{p_I}^s(\tilde{\eta}; q, Q_s)$ that equals 1 when $U(c_s^i|c_p) \geqslant U(c_s^c|c_p)$ and 0 when $U(c_s^i|c_p) < U(c_s^c|c_p)$. In state "not S" the rational decision rule picks the incumbent $\tilde{p_I}(\tilde{\eta}; q, Q_s)$ that equals 1 when $U(c^i|c_p) \geqslant U(c^c|c_p)$ and 0 when $U(c^i|c_p) < U(c^c|c_p)$.

2. Fix refence point c_p arising from (q^0, Q_s^0). Given any pair of rational decision rules derived from c_p, the ex-ante rational expectation of challenger losing in each state is correct:

$$q^0 = E[1 - \tilde{p_I}(\tilde{\eta}; c_p)] = 1 - Pr\Big(U(c^i|c_p) \geqslant U(c^c|c_p)\Big) \tag{3.4.3}$$

$$Q_s^0 = E[1 - \tilde{p_I}^s(\tilde{\eta}; c_p)] = 1 - Pr\Big(U(c_s^i|c_p) \geqslant U(c_s^c|c_p)\Big) \tag{3.4.4}$$

Moreover, condition (1) above implies a generalization of Theorem 3.1 to show the existence of a unique pair of decision rules for every reference point c_p that is parametrized by (q, Q_s) per Eq. 3.4.1. Each of the unique decision rules is conditional on the state and maximizes the reference-dependent utility, given the reference point c_p.

Theorem 3.4 (Unique decision pair for each refence point c_p)**.** *For any reference point (q, Q_s), there is a unique pair of cutoff decision rules (η_{ns}^*, η_s^*), that depend on (q, Q_s), so that in each state $j \in \{s, ns\}, \eta_j^* : [0,1] \times [0,1] \to (0,1)$ defines the following*

optimal voting rule as:

$$\tilde{p_I}^j(\tilde{\eta}; q, Q_s) = \begin{cases} 1 & \text{if } \tilde{\eta} \geqslant \eta_j^*, \\ 0 & \text{if } \tilde{\eta} < \eta_j^*. \end{cases}$$

Proof outline. 1. Fix state $j \in \{s, ns\}$ and define $f_j(\tilde{\eta}) = U(c_j^i | c_p)$. Use the explicit formulation in Appendix 1 to show that f is strictly increasing in $\tilde{\eta}$. Apply the proof of Lemma B.6 piecewise for each of the two segments in terms of $\tilde{\eta}$ with cutoff $H_j^i(s)$, that only depends on parameters other than $\tilde{\eta}$. Intuitively, the direct utility of picking an incumbent rises faster in his ability than a reference point that has a $p_j(1-q_j)$ weight on $\tilde{\eta}$.

2. Define $g_j(\tilde{\eta}) = U(c_j^c | c_p)$. Use the explicit formulation in Appendix 1 to show that g is weakly decreasing in $\tilde{\eta}$. Apply the proof of Lemma B.5 piecewise for each of the three segments in terms of $\tilde{\eta}$ with cutoffs $\underline{H}_j^i(s)$ and $\overline{H}_j^i(s)$. These cutoffs only depend on parameters other than $\tilde{\eta}$. Intuitively, the utility of picking a challenger falls in the incumbent's ability as the reference point rises with $p_j(1-q_j)$ weight on $\tilde{\eta}$.

3. $(f_j - g_j)(1) > 0$ and $(f_j - g_j)(0) < 0$. Each state fixes the after-tax income for consumption of the private good, but consumer's choice affects the variation of the public goods next period (pinned-down by politician's ability). Selecting the best incumbent of highest ability 1 is strictly preferred to pulling an unknown challenger and the worst incumbent is strictly worse than a random challenger draw, for any reference point c_p.

4. Then $f_j - g_j$ has a unique root on $(0, 1)$ by IVT. Denote this root as η_j^*. This is the unique value solves $U(c_j^i | c_p; \eta_j^*) = U(c_j^c | c_p; \eta_j^*)$. It is the interior threshold in state $j \in \{s, ns\}$ when the citizen is indifferent between a random challenger and an incumbent of known ability η_j^*, while the reference point is fixed at (q, Q_s). □

Let

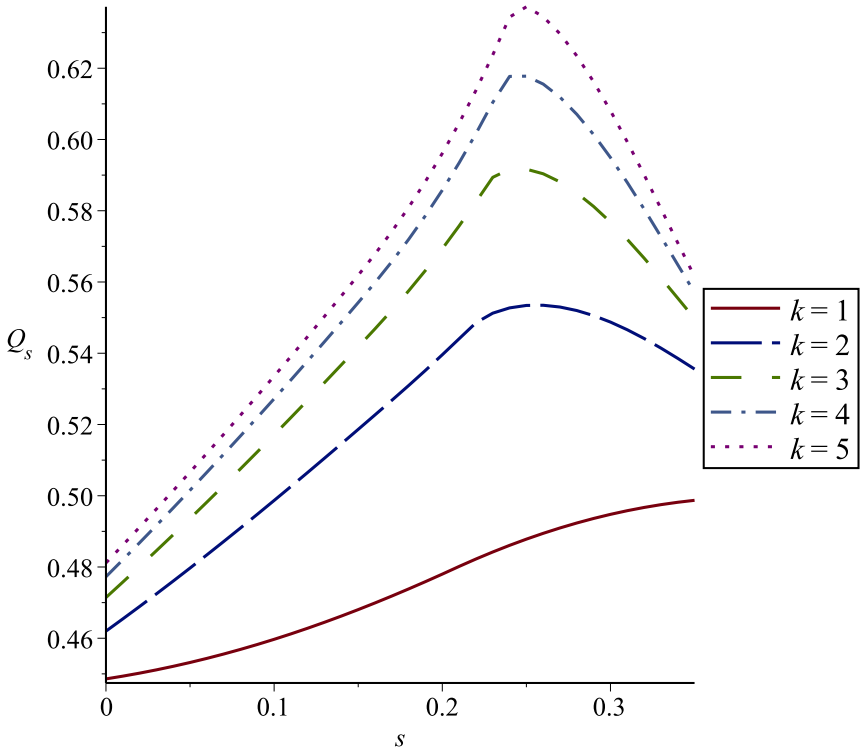
$$F(q, Q_s) = (\eta^*_{ns}, \eta^*_s) : [0,1]^2 \to [0,1]^2 \tag{3.4.5}$$

Proposition 3.5 (Existence of personal equilibrium with shock)**.** *There exists a personal equilibrium (fixed point) such that* $F(q, Q_s) = (q, Q)$.

Proof. As before by (Henriksen and Isbell, 1953), roots of algebraic expressions with continuous coefficients in (q, Q_0) remain continuous in these parameters. Thus, F is continuous. By Brouwer's Fixed Point theorem, it has a fixed point on $[0,1]^2$. □

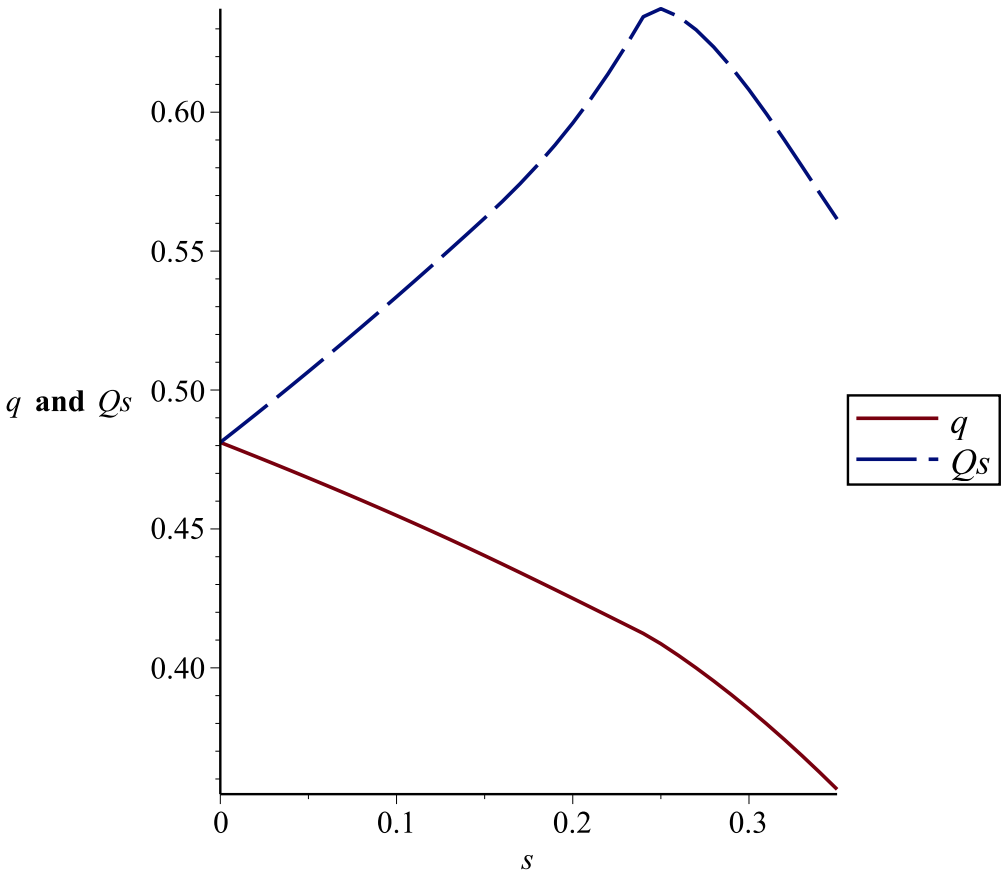
The Figure 3.5 showing comparative statics of Q_s with respect to the size of shock and convexity for $p = \frac{1}{3}$. It is qualitatively similar to when $p = 0$ in Figure 3.2 but shifted to the right.

Figure 3.5: Probability of picking challenger during crisis, Q_s as convexity increases, when $p = \frac{1}{3}$



In Figure 3.6, we can see that conditional on the shock, there is increase in incumbency disadvantage but conditional on no shock, there is increase in incumbency advantage. The first result shows that this finding in the ad hoc model was robust. The second result is new information that wasn't available in the ad hoc model. The interpretation is as follows: the voter is pleasantly surprised to find himself in the gains relative to the reference point that was putting positive weight on the shock. In the gains, gain-loss function generates risk-aversion, so the incumbent gets reelected more often. Thus, the presence of the shock event makes incumbency advantage stronger, so the conditional change from q to Q_s is greater under rational expectations of shock than in the "surprise" model.

Figure 3.6: Conditional probabilities of picking the challenger $(q, Q_s), p = \frac{1}{3}, k = 5$



In the previous section we've used an ad hoc model of "surprise" shock. It was unreasonable in the sense that voters were sophisticated enough to calculate q^* as a rational expectations equilibrium in the no-shock model but were not sophisticated enough to account for the true probability p of the shock. In particular, the incumbency disadvantage was not rationally anticipated, ever.

The purpose of that model was of convenience – it used existence results for the simpler no-shock model and that simpler model only needed to be solved once for the fixed point. Then taking that fixed point, the decision rule as a function of s was a sequence of roots of piecewise-continuous expressions of utility differences in $\tilde{\eta}$. Qualitatively, the ad hoc model was successful in that it verified stylized facts from the empirical literature about incumbency advantage switching to disadvantage during crisis and about the incumbent increasing public goods expenditure with shock, while losing more often.

The full-blown rational expectations model is a robustness check. Numerical calculations show that the stylized results still hold even for arbitrary $0 < p < 1$, though the calculations become more intensive. After fixing the other parameters, except for the shock magnitude s, the RE model is solved for a fixed-point pair (q, Q_s) for each s on a grid as follows. Taking an arbitrary pair (q^0, Q_s^0) and the corresponding c_p^0 with those weights, we solve for the optimal decision rule in each state, given c_p^0. Then the new pair (q^1, Q_s^1) corresponds to c_p^1 and the process is re-iterate to get the new decision pair etc, until convergence is attained to arbitrary precision. The convergence is still fast, on the order of 1 decimal place per iteration for any starting condition. This iterative process is then repeated for each value of s on a grid, to get a grid of fixed points $\{s, (q, Q_s)\}$. In contrast, in the ad hoc model there was only one calculation per s because the refence point q^* was iterated only for $s = 0$ and then reused.

The following Proposion 3.6 is an analytic result that leverages continuity and monotonicity of the algebraic functional forms for utility functions from Sec.B to get continuity of the resulting fixed points. This Proposition highlights the general connections between the fixed-point maps of the rational expectations and the ad hoc models.

Proposition 3.6 (Surprise model as a limit under RE)**.** *Take the personal equilibrium q_0 from the no-shock model and the corresponding decision rule $Q(q_0; S)$ as a function of shock S in the "surprise model," satisfying $q_0 = Q(q_0; 0)$. Now fix p and pick a personal equilibrium pair (q, Q_s) for each S in the rational-expectations model with the corresponding decision-pair map $F(x, y; S, p) = (F^{ns}(x, y; S, p), F^s(x, y; S, p))$, whose components depend on S and p, satisfying $F(q, Q_s; S, p) = (q, Q_s)$.*

Then

$$F^{ns}(a, a; 0, p) = F^s(a, a; 0, p) = Q(a; 0) = a \qquad \textit{equality along } S = 0 \tag{3.4.6}$$

$$F^{ns}(b, B; S, 0) = Q(b; 0) = b \qquad \textit{equality along } p = 0 \tag{3.4.7}$$

$$F^s(d, D; S, 0) = Q(d; S) \qquad \textit{equality along } p = 0 \tag{3.4.8}$$

Proof. By previous Theorems 3.1 and 3.4, all decision rules are continuous in parameters (because the utilities are and gain-loss functions are, from which the decision rules are derived as roots of algebraic expressions). Thus, the decision rule in the limit model equals to the limit of the decision rules.

Recall that,

$$c_p = p(Q_s c_s^\theta + (1 - Q_s)c_s^i) + (1 - p)(qc^\theta + (1 - q)c^i) \tag{3.4.9}$$

$$c_q = qc^\theta + (1 - q)c^i \tag{3.4.10}$$

1. When $s = 0$, we have

$$c_p = c^{\theta}\big(pQ_s + (1-p)q\big) + c^{i}\big(p(1-Q_s) + (1-p)(1-q)\big)$$

If also $Q_s = q = a$, then $c_p = c_q$. The rational expectations reference point of 0 shock with probability p is equal to the no-shock reference point by construction, whenever $q = Q_S$, which are both equal to a here. Thus, the RE decision in the shock state with $S = 0$, given the same reference point, coincides with the ad hoc decision when $S = 0$: $F^s(a, a; 0, p) = Q(a; 0)$. The decision rule in the model without shocks is the same as the decision rule in the RE model of shock $s \to 0$ along the $q = Q_s$ path. When $s = 0$, moreover, the utilities of the RE model are equal in each state for arbitrary reference-point pair: $U(c^c|c_p; S = 0, p) = U(c^{cs}|c_p; S = 0, p)$ and $U(c^i|c_p; S = 0, p) = U(c^{is}|c_p; S = 0, p)$ Thus, the corresponding decisions are equal in each state: $F^{ns}(q^1, Q^1; 0, p) = F^s(q^1, Q^1; 0, p)$ and take $q^1 = Q^1 = a$.

2. When $p = 0$, we have an identity for the reference points in the RE and the ad hoc models, independent of Q_s,

$$c_p = bc^{\theta} + (1-b)c^{i} = c_b$$

Conditional on observing no shock, when shocks occurs with $p = 0$ the decision is the same as being always surprised by observing a shock of $S = 0$, when not anticipating any shock, $F^{ns}(b, B; S, 0) = Q(b; 0)$ because utilities coincide as $U(c^c|c_p; p = 0) = U(c^c_s|c_b; S = 0)$ and $U(c^i|c_p; p = 0) = U(c^i_s|c_b; S = 0)$.

3. When $p = 0$, then $c_d = c_p(d, Q_s; p = 0)$ for any Q_s as above. The utilities in the "surprise" model are equal to the rational expectations utilities when the shock is observed under reference point c_d : $U(c^c_s|c_p; S, p = 0) = U(c^c_s|c_d; S)$ and $U(c^i_s|c_p; S, p = 0) = U(c^i_s|c_d; S)$. Equivalently, the reference point is c_p restricted to

$p = 0$. Surprise of shock S produces the same decision rule as under RE of probability $p = 0$ shock, conditional on the shock being observed: $F^s(d, D; S, 0) = Q(d; S)$. □

As a special case of Proposion 3.6, the following Corollary 3.7 shows that the personal equilibrium in the model with no shocks from 3.2.2 is the same as the rational expectations equilibrium in the model with vanishing shocks s. Secondly, the ad hoc model with shock s is the rational expectations equilibrium with shock s with vanishingly probability p.

Corollary 3.7. *(q_0, q_0) is a RE personal equilibrium when $S = 0$ or the limit of models as $s \to 0$, for any $p \in [0, 1]$. $(q_0, Q(q_0; S))$ is a RE personal equilibrium when $S \geqslant 0$ for $p = 0$ or the limit of models as $p \to 0$.*

Proof. Take $a = b = c = q_0$ and $B = C = Q(q_0; S)$ in 3.6 and it follows immediately that

$$\big(F^{ns}(q_0, q_0; 0, p), F^s(q_0, q_0; 0, p)\big) = (q_0, q_0) \tag{3.4.11}$$

$$\big(F^{ns}(q_0, Q(q_0; S); S, 0), F^s(q_0, Q(q_0; S); S, 0)\big) = \big(q_0, Q(q_0; S)\big) \tag{3.4.12}$$

Also note that as $S \to 0$, the second line becomes a special case of the first line with $p = 0$. The baseline no-shock model's rational equilibrium can be approached along the ad hoc models (take $p \to 0$ first) with $S \to 0$ or along the RE models as $S \to 0$ for arbitrary p. □

Observe that we emphasized how $q_0 \neq Q(q_0, 0)$ violated rational expectations in a model with no shocks but $q^* \neq Q(q^*; S)$ satisfying $q^* = Q(q^*, 0)$, and $p \approx 0$, does not violate rational expectations in the ad hoc model. This is because the ad hoc model has a one-dimensional decision rule – its reference point is a personal equilibrium when $S = 0$ always and its decision rule $Q(q_0; S)$ is conditioned on always observing $S > 0$. The shock drives a wedge between the initial expectation of incumbent's loss

(q less than $\frac{1}{2}$) and the actual ($Q(q;S) > \frac{1}{2}$ for moderate s). But taking the personal equilibrium from the rational expectations model, the fixed-point is now with respect to a different map F with two components that evaluates this wedge correctly and parametrized by $p = 0$: $F(q_0, Q(q_0;S);S,0) = (q_0, Q(q_0;S))$.

3.5 Conclusion

This paper takes a new look at the studied problem of incumbency advantage using behavioral loss-averse, reference-dependent preferences to explain recent behavioral-like empirical findings. Personal equilibrium of Koszegi and Rabin (2007) is applied to voters who rationally expect their own reference point formed by their future rational voting decision. This gives analytical precision to the suggestive intuition found in Quattrone and Tversky (1988). The explanation does not depend on specific features of legislature politics and can be applied to explain the pervasive incumbency advantage found in any kind of elections (state and federal legislatures, gubernatorial and state executives) as Ansolabehere and Snyder Jr (2002) has found. Wolfers (2007) and Achen and Bartels (2004) both find that challengers do better in bad times after an exogenous shock (economic and natural shocks, respectively), unrelated to the government's actions, and incumbents during normal/good times is consistent with $q(s)$ increasing in s for small s. These authors suspect some sort of behavioral mechanism but do not construct one. Reference-dependent risk preferences offer just such a mechanism.

A model that only has politicians with career concerns and rational voters will have the incumbent exert some effort in the first period as well but because his probability of winning wont be related to irrelevant signals (exogenous shock to income), he won't alter the public goods spending as the shock varies. In fact, $\gamma = 0$ gives $q = Q_s = \frac{1}{2}$

for all parameter values – judging politicians solely on their innate talent makes incumbents and challengers interchangeable ex-ante.

This paper finds that the analytical incumbency advantage requires only a feasible reference point, which need not be endogenous, and a weakly concave gain-loss function. To generate incumbency disadvantage (numerically) and the corresponding disaster relief through the "career-concerns" channel, the endogeneous reference point is used as a selection device to make the model's prediction tighter for the purposes of the comparative statics as s increases. Incumbency disadvantage is first illustrated in a simpler ad hoc model with where the voters are surprise by anticipated crisis that always happens. Then, that model is nested within a consistent rational expectations model where crisis happens with known probability p, which enters (and complicates) the reference point of a voter. Finally, a limiting continuty argument connects all three models together.

Cole, Healy, and Werker (2008)'s finding about electoral performance of Indian administrations during terms with and without earthquakes is consistent with the story of politicians showing their competence η^i (or lack-of-thereof) through government responses (public good provision, $g_1(q)$, yet doing worse-off ex-post: both value function of the incumbent $v_I(q)$ and probability of winning p_I decrease when incumbency advantage falls at the same time as rents $r_1(q)$ fall and public outlays increase $g_1^*(q)$.

Another avenue for loss aversion would be politicians' preferences. A challenger may be unwilling to fight as hard to gain the position in question, compared to incumbent determined to avoid losing. This could explain the powerful deterrence phenomenon where a fraction of incumbent house races goes uncontested (Diermeier, Keane, and Merlo, 2005; Stone, Maisel, and Maestas, 2004). If this was one important reason for deterrence of entry and if deterrence was causing the upward trend for incumbency advantage, it is unclear why effects of loss aversion would increase over

time. Reference-dependent preferences may give reasonable explanation for cross-sectional data but less plausible to explain temporal drift found by Gelman and King (1990).

Chapter 4

Fault-Tolerant Bayesian Implementation in General Environments

Classical implementation theory uses Nash equilibrium as its solution concept, which assumes that each agent is fully rational and can choose his most preferred strategy. In this case, player i expects all other players will follow their equilibrium strategy. If some other players deviated by a mistake, then i may want to deviate as well.

The concept of fault-tolerant implementation was introduced by Eliaz (2002) for environments of complete information. The idea is to consider a stronger equilibrium notion than Nash equilibrium, so that players have no incentive to deviate from the equilibrium strategy for any private belief about the identity and behavior of up to k faulty players. Perhaps, a minority of players are making mistakes, malicious, behavioral or did not understand instructions. The 0-fault-tolerant equilibrium reduces to Nash equilibrium and $(N-1)$-fault-tolerant equilibrium is weak dominance.

A stronger statement of full implementation[1] for social choice correspondences includes a new requirement that all k-deviations from k-FTNE equilibria are desirable. This is because even when all rational players adhere to their equilibrium strategies, the outcome of the mechanism may change because of deviations by k faulty players. In some sense, these k-deviations near the equilibrium come bundled with the equilibrium. Since we want each equilibrium to be desirable, it is reasonable that the bundle of k-deviations is also desirable.

Eliaz (2002) showed that k-monotonicity[2] of the social-choice correspondence and standard no-veto power[3] are sufficient for full implementation. Furthermore, weak k-monotonicity is necessary for full implementation.

The current paper extends fault-tolerant implementation from Eliaz (2002)'s complete information setting to an incomplete information setting, similar to Jackson (1991). Here, agents have private information about the state of the world and exclusive[4] information is allowed. Exclusive information is relevant for the study of auctions, public goods provision, delegation games, partnership arrangements, etc.

There is no objective measure of faultiness that is revealed to the social planner or any of the players. It is a key assumption that players are allowed to have arbitrary specific beliefs[5] about who (if any) is faulty and what state-dependent strategy they play. These beliefs are degenerate lotteries as there is no uncertainty about faultiness for any fixed belief. A different analysis may allow for subjective and arbitrary non-degenerate lotteries over faultiness of others.

[1]Every equilibrium of a mechanism is desirable and every desirable outcome can be supported by an equilibrium.

[2]Reduces to Maskin monotonicity for $k = 0$

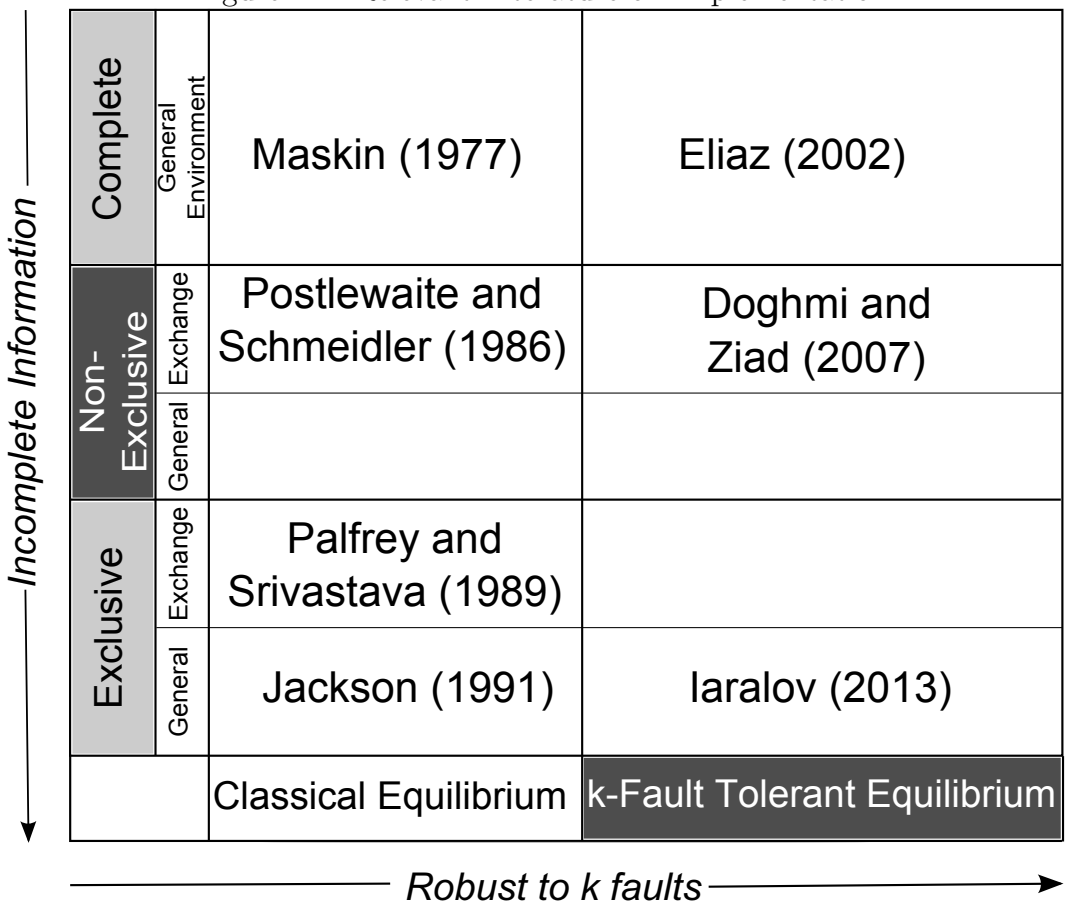
[3]No single player can be a dictator in the senes that if everyone but that player find an outcome most preferred, it must be desirable in the social choice set.

[4]Non-exclusive information means any $N-1$, or possibly any $N-k-1$, players collectively have complete information.

[5]But even under the most favorable beliefs, there is no profitable deviation for any non-faulty player as long as upper bound of k is not violated.

Similar to Jackson (1991), the environment in the current paper is a general one. It is not necessarily economic[6], best and worst allocations across all states sometimes may not exist and full support is relaxed.

Figure 4.1: Relevant Literature on Implementation



There are different ways we can model faulty players under incomplete information and it ties-in to the choice of the equilibrium and implementation concepts. For example, allowing only pure or also mixed strategies, assuming players have some specific or arbitrary beliefs about the faultiness of others, requiring (or not) that k-deviations from the equilibrium need to be desirable. Consider combining exclusive information setting with faulty players in the way that reproduces the approach from

[6]An economic environment has at least two players that aren't satiated at any social choice function and each prefers some other social choice function. An exchange economy with $N \geqslant 3$ players is an economic environment because there are always 2 players who don't get the full endowment. For an example of a non-economic environment consider an economy where in some state everyone is indifferent between every feasible social choice function.

Eliaz (2002) with a single state when each player has exactly one type and reproduces the approach from Jackson (1991) when no faulty players are allowed ($k = 0$). Even this straightforward extension develops some novel difficulties.

Non-exclusive information settings, such as complete information structure, are not sensitive to deceptions by a minority because the correct state can be correctly inferred if the majority is truthful. This covers both a rational agent considering a unilateral deception and a group of faulty players deceiving by a mistake. The exclusive information setting means that individual players' reports of their state are unverifiable and the mechanism takes them at face value. The incentive compatibility condition keeps rational agents honest when everyone else is truthful. The added complication is in considering deceptions by a minority of faulty players because they are not constrained by incentives and their reports are unverifiable. Therefore, the mechanism cannot decipher the types of faulty players directly in an exclusive information setting.

The equilibrium requirement has $N-k$ players continuing to play their prescribed strategies, even when k players deviate. To get partial implementation, the incentive compatibility condition is strengthened to be insensitive to deceptions that are in the k-ball around the truth. That is, as long as the other $N-k-1$ players are truthful, the individual player i prefers to tell the truth. This sustains any desirable outcomes as an equilibrium.

The second complication arises when going in the other direction to get full implementation. The mechanism and a monotonicity condition work together to ensure deceptions of the majority[7] that lead to undesirable outcomes are broken up by a "whistleblowing" minority. The mechanism designer is concerned that a malicious minority may deviate to break a desirable equilibrium when no deception is taking place. The key ingredient of the mechanism is to block any minority objection that

[7]Incentive compatibility doesn't apply when other players are not truthful.

is preferred for some type of some player in the minority when no deception is taking place and the majority proposes some x that is desirable. Similarly, the key ingredient of the monotonicity condition is, for each non-desirable majority deception, generating a state s and an objection y less preferable than x for "whistleblowers" when the majority is truthful, for all types of the objecting players, while preferring the objection in that specific state only if the majority is deceptive.

This second complication is that a minority of $k+1$ players under exclusive information setting have different information sets. In Jackson (1991), the objecting minority consists of exactly one player and his information set is trivially constant across the minority (equal to itself). In Eliaz (2002), the objecting minority consists of $k+1$ players but their information set is the same singleton because information is complete. If the information set is common[8], then the faulty players' types are constant along the rational player's information set. Thus, the faulty players have a constant message under arbitrary deceptions along the rational player's information set. Since the planner doesn't block the faulty players' message at s (by construction), he doesn't block their message at all states in rational player j's information set. This was the case in the previous literature – in Doghmi and Ziad (2007) faulty deceptions did not matter because the planner knew the true state in every $k-$deviation of every equilibrium and the special structure of the environment allowed punishments to stay within a single rule of the mechanism.

On the other hand, when the faulty players' types may vary across the rational player j's information set, the mechanism will allow the "whistleblowing" coalition to replace x with y on a neighborhood of a true state s and potentially block it on a disjoint subset of the rational player j's information set at s^j. This block is to prevent potential "conflict of interest" by one of the coalition members, had they been rational and had a type that strictly preferred y to x when the majority was, in fact, truthful.

[8]As Observation 4.10 in Section 4.2 notes, it is enough that the rational player's information set is weakly coarser than the common intersection of the coalition's information sets.

The condition k-monotonicity-no-veto (k-MNV) has to track these subsets when it is describing the outcomes of the mechanism for j's deviation and non-deviation along his information set at s^j, so that j is ready to form a $k+1$ minority coalition to break the equilibria with non-desirable outcome when a majority coalition asks for x with a majority deception.

This paper defines a central condition called k-incentive compatibility (k-IC), a direct generalization from Jackson (1991)'s incentive compatibility. k-IC is both necessary and sufficient for partial implementation[9] in k-FTBE of social choice sets that satisfy closure. The usual incentive compatibility requires that whenever everyone is telling the truth, the agent also prefers to tell the truth about his type, for any type realization. This condition is important because under exclusive information the planner has to rely on each agent to report their private state to determine the appropriate allocation. On the other hand, under non-exclusive information (including symmetric complete information), the planner can rely on a consensus of reports, so he doesn't have to rely on individual person's report. Secondly, the planner can enforce consensus by punishing obvious deviators who report information that is inconsistent with the consensus. The complication of k faulty players under exclusive information allows for "hidden" lies. Even when the correct social choice function is picked, a given player is allowed to have arbitrary beliefs about k faulty players lying about their type. When the social choice set satisfies k-IC, as long as the other $N-k-1$ players tell the truth, this given player finds it optimal to also tell the truth.

When the social choice set also satisfies k-monotonicity-no-veto[10], full implementation in k-FTBE is achieved. The downside is that k-MNV is a fairly complicated statement that generalizes Jackson (1991)'s MNV to k faulty players. Originally, MNV was a combination of no-veto power and Bayesian monotonicity where in some

[9]For any desirable social choice function, there is an equilibrium whose outcome agrees with the function.

[10]k-MNV combines k-no-veto-hypothesis and k-Bayesian monotonicity

states no-veto-hypothesis is satisfied[11] and in other states, Bayesian monotonicity generates a profitable deviation to kill equilibria with non-desirable outcomes. The states when NVH holds are precisely the states when the integer game is played and outcomes in those states are indifferent (else someone could deviate their integer strategy to win a better outcome for those states). Under symmetric information in Eliaz (2002), the information set is a singleton and either the integer game is played in equilibrium or not. In contrast, in a Bayesian equilibrium with incomplete information, on a subset of the information set, some states lead to the integer game and for other subsets the integer game is not played. This means the constructed deviation is more involved under incomplete information as it has to identify what outcomes result from different rules of the mechanism under different subsets of a given information set.

In Jackson (1991), monotonicity-no-veto gives the second condition of full implementation: for each equilibrium, there is a social choice function that matches that equilibrium's outcome. Here, k-MNV goes further and considers the whole ball of k-deviation around each equilibrium and finds a desirable function for each element in the ball.[12]

Weak k-Bayesian monotonicity is necessary for full implementation. When players try to play a deception that leads to a non-desirable outcome by manipulating their reports of the state, full implementation cannot admit such an equilibrium and so it generates a group of $k+1$ "whistleblowers." They offer a minority objection as a deviation that is less preferable under the truth for each player, relative to some

[11] When $N-1$ players reach a consensus and the N-th player does something else, he cannot trigger the integer game alone. However, any of the $N-1$ players, when deviating, may trigger the integer game and win any prize. Since they don't find it profitable to break equilibrium in those situations, they must prefer the equilibrium outcome to anything else they could've asked for. Perhaps, there are many ties in that state. No-veto hypothesis maintains the desirability of the equilibrium because $N-1$ players find it most preferred.

[12] Any equilibrium with a non-desirable k-deviation is eliminated by construction in k-MNV.

other desirable outcome[13]. Moreover, this objection is strictly better under the proposed deception for one of the "whistleblowers," who finds it a profitable deviation under the belief that the other k "whistleblowers" are faulty and play the specified minority objection. Doghmi and Ziad (2007) have a corresponding result for an exchange economy, restricting attention to compatible deceptions[14]. In general economies, the statement is a bit stronger by not requiring compatible deceptions. Using the special structure of an exchange economy, their mechanism punishes everyone for non-compatible deceptions by giving everyone the 0 allocation, which is the worst outcome for all states. Since more general environments need not have such an outcome, non-compatible deceptions are also allowed.

In the previous work of Doghmi and Ziad (2007), they extended k-fault implementation to incomplete information in a more specific environment: an exchange economy with non-exclusive information as in Palfrey and Srivastava (1989). Doghmi and Ziad (2007) use *k-non-exclusive information*, which becomes more restrictive as k increases because the smaller remaining group knows everything. While allowing robust implementation despite the presence of faulty players, the informational requirement becomes more restrictive and takes a step back towards symmetric information. More recently, Doghmi and Ziad (2009) re-examines k-FTNE implementation of social-choice correspondences in exchange economies with complete information. In some specific environments (e.g. single-peaked preferences with unanimity), their strict k-monotonicity is a sufficient condition for full implementation because the setting provides a weak form of NVP "for free." In other words, this line of work on robust implementation is trying to tighten the gap between sufficient and necessary conditions in Eliaz (2002) by weakening or dropping the no-veto-power. This setting

[13]It is a *weak* k-BM because this preferred outcome can be different for each member of the group.

[14]Compatible refers to deceptions satisfying $s \in T$ implies $\alpha(s) \in T$, where T is the set of states occurring with non-zero probability.

also does not allow for exclusive information or environments more general than an exchange economy.

Doghmi and Ziad (2007) find that when a social choice set satisfies k-Bayesian monotonicity and assuming non-exclusive information (roughly, every group of at least k players collectively pins down the state of nature to a singleton), full implementation is achieved. The corresponding necessary condition is weak k-Bayesian monotonicity. The authors observe that Bayesian monotonicity neither implies nor is implied by weak k-Bayesian monotonicity under incomplete information, similar to the result in Eliaz (2002) that weak k-monotonicity is distinct from Maskin monotonicity.

An exchange economy is a special case of an economic environment, has a worst and a best element independent of the state[15], which allows to restrict attention only to compatible deceptions. The social planner observing non-compatible deception knows that someone has deviated, not necessarily whom. Mechanisms in exchange economies following Palfrey and Srivastava (1989) give collective punishment of 0 to everyone in such cases to rule them out. They also assume full support and that there is an objective conditional distribution of faultiness over players that is independent of state realization. The special environment structure implies a special result is used to prove sufficiency of k-Bayesian monotonicity and k-non-exclusive information to achieve full implementation. The result states that in all states, the equilibrium outcomes come from the first rule of their mechanism: at least $N - k$ non-faulty players agree on the state, agree on the allocation rule and don't initiate the integer game.

An exchange economy is unlike a more general environment that allows outcomes of equilibria to satisfy NVH for the subset of the information set where the integer game is played. In contrast, there are equilibria in general environments that trigger multiple rules of the mechanism across an information set of some fixed player's

[15]In any state, receiving the full endowment E is ideal. In any state, receiving 0 endowment is the worst outcome.

type because players with exclusive information find it harder to coordinate: on a subset of his information set others coordinate on the same desirable x, on a disjoint subset others coordinate on some other $\overline{x}$, on another disjoint subset coordination fails and the integer game rule is triggered but they cannot improve on the equilibrium outcome.

4.1 Environment

Notation and definitions are borrowed or adapted from Jackson (1991), Eliaz (2002), Saglam (2007), and Doghmi and Ziad (2007).

An environment is a collection of $\{N, S, \mathcal{A}, \{q^i(s)\}, \{U^i\}\}$: N is the finite number of players[16]; S is the set of states describing agents' information and preference: $s = (s^1, \cdots, s^N) \in S$, where $|S^i|$ is finite; $\mathcal{A}$ is the set of feasible allocations or consequences, fixed across states; q^i is the prior of agent i on S; $U^i : \mathcal{A} \times S \to \mathbb{R}_+$ is the state-dependent utility function. The set of all social choice functions is $X = \{x | x : S \to \mathcal{A}\}$.

Agents agree on which states occur with positive probability: whenever $q^i(s) > 0$ then $q^j(s) > 0$ for all $j \neq i$. The common support of their priors is denoted by $T = \{s \in S | q^i(s) > 0\}$, which is equal for all i. The priors q^i define partitions of T, Π^i, with elements referenced by π^i. For a given information signal (type) $s^i \in S^i$, let

$$\pi^i(s^i) = \{t \in S | t^i = s^i \text{ and } q^i(t) > 0\} \in \Pi^i$$

be the plausible true states given agent i's information. Without loss of generality, each type has a non-empty information set: $\forall i \in N, \forall s^i \in S^i, \pi^i(s^i) \neq \varnothing$. Agents'

[16]With slight abuse of notation, N will subsequently refer to both the set of all players and the cardinality of that set.

preferences (complete and transitive) over social choice functions have a conditional expected utility representation. For fixed $x, y \in X, s^i \in S^i$,

$$x\mathcal{R}^i(s^i)y \iff \sum_{s\in\pi^i(s^i)} q^i(s)U^i[x(s),s] \geqslant \sum_{s\in\pi^i(s^i)} q^i(s)U^i[y(s),s]$$

A social choice set $F \subset X$ is a collection of "desirable" social choice functions.

Out of N players, at most k players are faulty who do not act according to incentives. Non-faulty players know they are not faulty but do not know who (if any) faulty players are, except that there are between 0 and k (inclusive) faulty players.

A *mechanism* consists of an action space $\mathcal{M} = \mathcal{M}^1 \times \cdots \times \mathcal{M}^N$ and a function $g : \mathcal{M} \to \mathcal{A}$. A (pure) *strategy* for i is a map from information sets to messages, $\sigma^i : S^i \to \mathcal{M}^i$. Denote vector of strategies $\sigma = (\sigma^1, \cdots, \sigma^N); (\sigma^{-i}, \tau^i) = (\sigma^1, \cdots, \sigma^{i-1}, \tau^i, \sigma^{i+1}, \cdots, \sigma^N)$. The outcome social choice function of σ is $x(s) = g(\sigma(s))$.

4.2 Definitions

By relaxing the full-support assumption, the objects of interest are the outcomes of social choice functions along the (plausible) states of non-zero measure.

Definition 4.1. *The social choice functions x and y are equivalent if $\forall s \in T, x(s) = y(s)$. The social choice sets F and $\hat{F}$ are equivalent if $\forall x \in F, \exists \hat{x} \in \hat{F}$ which is equivalent to x, and $\forall \hat{x} \in \hat{F}, \exists x \in F$ which is equivalent to $\hat{x}$.*■

The following notation of splicing will be useful to give meaning to one social choice function z being preferred to $\tilde{z}$ over a subset of states C, e.g. $z\mathcal{R}^j(s^j)\tilde{z}/_C z$.

Definition 4.2. *Let $x/_C z$ be a splicing of two social choice functions x and z along a set $C \subset S$. The social choice function $x/_C z$ is defined as $\forall s \in C, [x/_C z](s) = x(s)$ and $[x/_C z](s) = z(s)$ otherwise.*■

When the social planner faces incomplete and exclusive information, he relies on citizen's incentives to report their type as he has no other way of learning the aggregate state, which is simply a collection of realized types of players. When citizens try to cheat by asking for a desirable social choice function $x \in F$ but manipulate their reporting to get wrong allocations in various states, we say they are playing a deception $x \circ \alpha$. When this deception leads to a non-desirable outcome, a form of Bayesian monotonicity will be later used to eliminate the underlying equilibrium with deception α.

Definition 4.3. *A deception for $i \in N$ is a mapping $\alpha^i : S^i \to S^i$. Let $A^i = \{\alpha^i | \alpha^i : S^i \to S^i\}$ be the set of all deceptions of player i, $A = A^1 \times A^2 \cdots \times A^N$ and A_{-i} are defined accordingly. Let $\alpha = (\alpha^1, \cdots, \alpha^N) \in A$ and $\alpha(s) = [\alpha^1(s^1), \cdots, \alpha^N(s^N)]$. The notation $x \circ \alpha$ represents the social-choice function, which results in $x[\alpha(s)]$ for each $s \in S$.*■

Jackson (1991) points out that *closure* is a basic requirement for implementation of social choice sets. Suppose that the common knowledge concatenation Π has two elements. Pick any two equilibrium strategies of a specific mechanism. Construct a third strategy as follows: players use the first strategy on the first element of Π and the second strategy on the second element of Π, must also be an equilibrium of the mechanism (otherwise, any deviation against the third option would eliminate one of the two original equilibria, based on which element of Π the deviation was in). By full implementation, the outcome of the third equilibrium is desirable.

Definition 4.4 (**Closure**). *Let B and D be any disjoint sets of states such that $B \cup D = T$ and $\forall \pi \in \Pi$, either $\pi \subset B$ or $\pi \subset D$. A social choice set F satisfies closure (C) if $\forall x, y \in F, \exists z \in F$ s.t. $\forall s \in B, z(s) = x(s)$ and $\forall s \in D, z(s) = y(s)$.*■

The stronger notion of full implementation requires that any outcome in a k-ball of an equilibrium to be desirable. Likewise, beliefs about arbitrary "hidden lies" can

be translated into a k-ball of deceptions around the non-faulty deception α (or the identity map, id, when planning to tell the truth).

Definition 4.5 (k**-neighborhood**). *Given vector profile $v^* \in V$ and an upper bound for faulty players k, $B(v^*, k; V) = \{v \in V : |\{i \in N : v_i^* \neq v_i\}| \leqslant k\}$ is a k-neighborhood of v^* (with respect to V), is the set of profiles that are different from v^* by not more than k entries.*■

The complication of k faulty players under exclusive information allows for "hidden" lies. Even when the correct social choice function is picked, a given player is allowed to have arbitrary beliefs about k faulty players lying about their type. When the social choice set satisfies k-IC, whenever the other $N-k-1$ players tell the truth, this given player finds it optimal to tell the truth also.

Definition 4.6 (k**-IC**). *Given $i \in N$, define $id : S \to S$ to be the identity map and $\beta^{-i} \in B(id, k; A_{-i})$ is a profile of deceptions for players $N\backslash\{i\}$ in the k-neighborhood of the truth. A social choice set F satisfies k-incentive compatibility (k-IC) if*

$$\forall x \in F, \forall i \in N, \forall \beta^{-i} \in B(id, k; A_{-i}), \forall t^i, s^i \in S^i, x \circ (\beta^{-i}, id) \mathcal{R}^i(s^i) x \circ (\beta^{-i}, t^i).$$■

When players try to play a deception that leads to a non-desirable outcome by manipulating their reports of the state, full implementation cannot admit such an equilibrium and so it generates a group of $k+1$ "whistleblowers." They offer a minority objection as a deviation that is less preferable under the truth for each player, relative to some other desirable outcome. It is a *weak* k-BM because this preferred outcome can be different for each member of the group. Moreover, this objection is strictly better under the proposed deception for one of the "whistleblowers," j, who finds it a profitable deviation under the belief that the other k "whistleblowers" are faulty and play the specified minority objection.

Definition 4.7 (**weak k-Bayesian monotonicity**). *Given deception α and $x \in F$, a social choice set F is weak k-Bayesian monotonic (wk-BM) if whenever there is no social choice function in F which is equivalent to $x \circ \alpha$,*

1. *Each "whistleblower" i in M offers a minority objection y that is less preferable than some other desirable outcome x^i whenever the rest are telling the truth.*

$$\exists M \subset N : |M| \geqslant k+1 : \exists s \in S, \exists y \in X, \forall i \in M,$$
$$Q_i \equiv M \backslash \{i\}, \exists x^i \in F, \forall \beta^{Q_i} \in B(id, k; A_{Q_i}), \forall t^i \in S^i :$$
$$x^i \circ (\beta^{Q_i}, id^{N \backslash Q_i}) \mathcal{R}^i(t^i) y \circ (\beta^{Q_i}, id^{N \backslash M}, \alpha^i(s^i))$$

2. *There is a rational objector j in M who strictly prefers to deviate together with k faulty players, when the majority is using deception α.*

$$\exists j \in M, \exists \beta_0^{-j} \in B(\alpha, k; A_{-j}) : y \circ (\beta_0^{-j}, \alpha^j) \mathcal{P}^j(s^j) x^j \circ (\beta_0^{-j}, \alpha^j). \blacksquare$$

Condition k-no-veto-hypothesis is a generalization of no-veto-hypothesis from Jackson (1991), where instead of one player not being a dictator over the rest, here there is a group of $k+1$ players that are not dictators over the majority. For z choice function to satisfy NVH for a set of (plausible) states and a deception α means that no player in the majority can improve on z by only altering his integer-game play. It is not a profitable deviation to pick any replacement $\tilde{z}$ (under the α deception) from winning the integer game on a subset C of D and getting z as before when the integer game is not played (outside C). The meaning of the state D is that includes all states where at least one non-faulty person, i, is not playing the first "consensus" rule of the mechanism. This allows j players of the majority to reach the integer game under the corresponding belief of k faulty players playing along with i. Here it could be that i cannot reach the integer game along the equilibrium since he is

already outside the consensus in D and his deviation doesn't increase the size of the minority (when $N-k-1$ players are in consensus, i's beliefs about the faulty players are not sufficient to reach the integer game because the size of the minority is less than $k+2$). Thus, i doesn't necessarily prefer z more than all other social choice functions $\tilde{z}$ on states where the integer game is reached.

Definition 4.8 (k-**NVH**). *Given deception α, a set of faulty players $P \subset N, |P| \leqslant k$ and a subset of plausible states $D \subset T$, a social choice function $z \in X$ satisfies the k-no-veto hypothesis (k-NVH) for α and D, if $\forall s \in D, \exists i$ (non-faulty) $\in N \backslash P : \forall j \in N \backslash (P \cup \{i\})$, a majority of players $\geqslant n-k-1, \forall \tilde{z} \in X, \exists C \subset D : C \ni s$ and $z \mathcal{R}^j(s^j) \tilde{z} \circ \alpha/_C z$.*■

The following are three differences between MNV and k-MNV conditions. First, k-MNV uses k-NVH, whereas Eliaz (2002) used no-veto power that did not depend on k. In a world of symmetric information, $N-1$ players preferred z to anything arising from the integer game because every player, except for the one i who broke the consensus, could reach (and win) the integer game. That argument showed that $N-1$ players are indifferent between all outcomes when the integer game is reachable (so z must be desirable when $N-1$ find it optimal). Even the faulty players leading to z, which is a k-deviation of a given equilibrium, could reproduce z in Eliaz (2002) when they are playing their equilibrium strategy because the state was a singleton. But in a world of exclusive information, the faulty players leading to z, when they individually turn out to be rational may not (necessarily) be able to trigger the integer game on a neighborhood of every point in D and still get z outside of the neighborhood because D was defined relative to the initial faulty players P.

Monotonicity-no-veto of Jackson (1991) and k-MNV[17] highlight a signal s^j when the player j has a profitable deviation as a "whistleblower" on his information set

[17]These conditions are used, in part, to establish Bayesian full implementation and full implementation in k-FTBE, respectively.

$\pi^j(s^j)$[18]. The construction states that the proposed deviator j strictly prefers the outcome of his deviation $\overline{z}$, spelled-out for different subsets of his information set leading to different rules of the mechanism, compared to the outcome z of non-deviating. Thus, MNV only needs to construct the deviation $\overline{z}$ because the given non-deviation z is given "for free" by the hypothesis. When the non-faulty deviator j makes his deviation, he requires particular beliefs about k faulty players also deviating as "whistleblowers." The second difference between MNV and k-MNV is the construction of the z' k-deviation in the later definition that results from player j deciding to abandon his coalition of k faulty players and not deviation from his prescribed strategy. This means they are deviating regardless of his action and the status-quo of j sticking to equilibrium entails the same k-deviation by the faulty players as along j's personal deviation, so k-MNV needs to construct not only the outcome of j's deviation $\overline{z}$ (under "whistleblower" beliefs about k faulty players) but also the outcome z' when j sticks to the equilibrium under beliefs that the k faulty players are forming a "whistleblower" coalition with j.

MNV states player j prefers not to deviate when others play the truth instead of the deception α and j's utility is computed over j's information set, given all possible types t^i. He does not want to falsely announce that his type is $\alpha^i(s^i)$ when the majority is truthful. Here s^i is the "whistleblowing" type when he does prefer to deviate against majority that pays deception α. However, in k-MNV we have a coalition of $k+1$ "whistleblowers," any one of whom could potentially be non-faulty and strategic as far as the social planner is concerned. Thus, none of them may prefer to deviate along *their* information sets, which is why $R^i_{\alpha,x,y}$ states[19] are imported into the definition from the proof.

[18]He deviates against the alternative of an undesirable equilibrium z and undesirable k-deviation z' in the two cases, respectively.

[19]$R^i_{\alpha,x,y}$ are types s^i of player i such that when the majority is truthful and k faulty players play an arbitrary deception, all types t^i of player i prefer announcing x to announcing $y \circ \alpha^i(s^i)$ on the information set $\pi^i(t^i)$.

The primary function of the mechanism (defined in section 4) is to allow players pick desirable outcomes and to incentivize a "whistleblower" coalition of $k+1$ players to break undesirable deceptions. Furthermore, the designer cannot distinguish the rational pivotal player from those k faulty players, so he has to check incentives of every one of these $k+1$ players individually as if any particular player could be the rational one. The mechanism designer is cautious in allowing a coalition to break the underlying equilibrium (with undesirable k-deviations) only if there was no circumstance when some player type would strictly benefit to create a malicious coalition to break an equilibrium leading to a desirable x with no deception taking place. This is a generalization of the requirement from Jackson (1991) that all types of the single "whistleblower" do not benefit from breaking up desirable x when no deception takes place.

k-MNV makes sure that the particular coalition containing the rational "whistleblower" does not include faulty players with preferences that would get the whole coalition blocked by the mechanism rule 2b in every state of the information set of the pivotal rational "whistleblower." This is a simple requirement when $k = 0$ or the information set is a singleton (complete information setting) because the coalition shares the same information set.

The preferences of faulty players[20] need to be considered on π^j because their information isn't held constant by rational player j's information set $\pi^j(s^j)$, which is why $R^i_{\alpha,x,y}$ sets are introduced in addition to B^i_x sets[21] t as per Jackson (1991)in k-MNV.

$$R^i_{\alpha,x,y} = \{s^i \in S^i : \forall \beta^{-i} \in B(id, k; A_{-i}), \forall t^i \in S^i, x \circ (\beta^{-i}, id)\mathcal{R}^i(t^i)y \circ (\beta^{-i}, \alpha^i(s^i))\}$$

[20]The desirable $x \in F$ needs to be preferred to y by all possible types of members of the "whistleblowing" coalition as long as the majority that is asking for x is truthful. Otherwise, y will be blocked by the mechanism.

[21]B^i_x is the set of types of player i when he asks for x, while playing deception α^i.

Definition 4.9 (**k-monotonicity-no-veto**). *Given deception α, a set of faulty players $P \subset N : |P| \leqslant k$ and $\forall x \in F, \forall y \in X, \forall i \in N$, given sets of types $B_x^i, B_{x,y}^i, R_{\alpha,x,y}^i \subset S^i$. Let $B_x^M = \{t \in S \,|\, \forall l \in M, t^l \in B_x^l\}$, a set of states when M players find their type in B_x^i. Suppose $\exists z \in X : \forall x \in F, \forall R \subset N : P \subset R, |R| = k, z(t) = x \circ \alpha(t)$ when $t \in B_x^{N \setminus R}$. That is, z is such that $x \circ \alpha$ is played whenever at least $N - k$ non-faulty players have signal in B_x^i. Also, suppose z satisfies (k-NVH) for α, P and for $D \equiv T \setminus \left(\cup_{x \in F} \cup_{P \subset R, |R| = k} B_x^{N \setminus R} \right)$, the set of states when at least one of the non-faulty player's type is not in B_x^i. Then F satisfies k-monotonicity-no-veto (k-MNV) if, whenever there is no social choice function in F which is equivalent to z, $\exists Q \subset N : P \subset Q, |Q| = k, \exists j \in N \setminus Q$, denote $M = Q \cup \{j\}$ ($k+1$ whistleblowers), $\exists x \in F, \exists s \in B_x^{N \setminus Q}, \exists y \in X$ (reward), $\tilde{z} \in X$ (integer-game prize), $\overline{z} \in X$ (deviation outcome), $z' \in X$ (no-deviation outcome). $\exists \beta^{-j} \in B(\alpha, k; A_{-j})$:*[22]

If j deviates on $\pi^j(s^j)$, the mechanism produces:

$$\overline{z}(t) = \begin{cases} y \circ (\beta^{-j}, \alpha^j)(t) & \text{when } t^i \in B_x^i, \forall i \in N \setminus Q, \text{ and } t^q \in R_{\beta,x,y}^q, \forall q \in Q, \\ x \circ (\beta^{-j}, \alpha^j)(t) & \text{when } t^i \in B_x^i, \forall i \in N \setminus Q, \text{ and } t^q \notin R_{\beta,x,y}^q, \exists q \in Q, \\ \overline{x} \circ (\beta^{-j}, \alpha^j)(t) & \text{when } t^i \in B_{\overline{x}}^i, (\overline{x} \neq x), \forall i \in N \setminus M, \text{ and } t^j \in B_x^j, \\ \tilde{z} \circ (\beta^{-j}, \alpha^j)(t) & \text{otherwise.} \end{cases}$$

[22] Rational players continue to play α deception but faulty players Q have arbitrary deceptions: $\forall i \in N \setminus M, \beta^i = \alpha^i$ and $\forall q \in Q, \beta^q$ is an arbitrary deception.

Otherwise, if j does not deviate on $\pi^j(s^j)$, the mechanism produces:

$$z'(t) = \begin{cases} x \circ (\beta^{-j}, \alpha^j)(t) & \text{when } t^i \in B^i_x, \forall i \in N \backslash Q, \\ y \circ (\beta^{-j}, \alpha^j)(t) & \text{when } t^n \in B^n_{x,y} \cap R^n_{\alpha,x,y}, \exists n \in N \backslash M, \\ & \text{and } t^i \in B^i_x, \forall i \in N \backslash \{Q \cup n\}, \text{ and } t^q \in R^q_{\beta,x,y}, \forall q \in Q, \\ x \circ (\beta^{-j}, \alpha^j)(t) & \text{when } t^n \notin B^n_{x,y} \cap R^n_{\alpha,x,y}, \exists n \in N \backslash M, \\ & \text{or } t^q \notin R^q_{\beta,x,y}, \exists q \in Q, \text{ and } t^i \in B^i_x, \forall i \in N \backslash \{Q \cup n\}, \\ \overline{x} \circ (\beta^{-j}, \alpha^j)(t) & \text{when } t^i \in B^i_{\overline{x}}, (\overline{x} \neq x), \forall i \in N \backslash M, \text{ and } t^j \in B^j_x, \\ \tilde{z} \circ (\beta^{-j}, \alpha^j)(t) & \text{otherwise.} \end{cases}$$

satisfying

1. *$s^j \in R^j_{\alpha,x,y}$ and $\forall q \in Q, s^q \in R^q_{\beta,x,y}$, i.e. $x \circ (\cdot, id))$ is preferred to $y \circ (\cdot, deception(s^q))$ for all types of player q.*

2. *$\overline{z}\mathcal{P}^j(s^j)z'$, i.e. j has profitable deviation at s^j.* ■

Consider a special information structure when the information set of the rational player is (weakly) coarser than the common intersection of the information sets of the coalition. The following observation holds for the setting in Jackson (1991) because $k = 0$ and for the setting in Eliaz (2002) because the state space $S = \Pi$ is a singleton. In general, it does not hold and the faulty players' types may vary across rational player j's information set. Therefore, this is a novel complication in k-fault-tolerant implementation with exclusive information.

When this observation is true, k-MNV can be slightly simplified because in construction of $\overline{z}$ the mechanism never blocks the coalition by rule 2b on a suspicion that one of the faulty coalition members is a rational player trying to falsify a majority deception. In that case, $t^q \in R^q_{\beta,x,y}$ always holds.

Observation 4.10. *Given a coalition $M = \{j\} \cup Q$, a set of states $R^i_{\gamma,x,y}$ and state $s \in S$ satisfying $\forall i \in M : s^i \in R^i_{\gamma,x,y}$, if*

$$\pi^j(s^j) \subset \cap_{i \in M} \pi^i(s^i), \textit{then}$$
$$\forall t \in \pi^j(s^j), \forall i \in M : t^i = s^i \in R^i_{\gamma,x,y}.$$

Proof. In k-MNV notation the deception was $(\beta^{-j}, \alpha^j) = (\alpha^{N \setminus Q}, \beta^Q) \equiv \gamma$. Consider an arbitrary set t in the rational player j's information set $\pi^j(s^j)$. It is contained in the common intersection of the information sets of the coalition by hypothesis: $\forall t \in \pi^j(s^j), \forall i \in M : t \in \pi^i(s^i)$. Then player i's type $t^i = s^i$. □

4.3 Implementation

Every player i of arbitrary type s^i considers every deviation by at most k other (faulty) players σ^M. The resulting outcome, on his information set $\pi^i(s^i)$, from playing σ^{*i} is preferred to outcome of him deviating to any $\tilde{\sigma}^i$, assuming the faulty players play the same way σ^M.

Definition 4.11 (k**-FTBE**)**.** *A profile of strategies $\sigma^* = (\sigma^{*1}, \cdots, \sigma^{*N}) \in \Sigma$ is a k-fault-tolerant Bayesian equilibrium (k-FTBE), if $\forall i \in N, \forall s^i \in S^i, \forall \tilde{\sigma}^i \in \Sigma^i, \forall M \subset N :$ $|M| \leqslant k, \forall \sigma^M : S^M \to \mathcal{M}^M$, have $g(\sigma^{*i}, \sigma^{*N \setminus M \cup \{i\}}, \sigma^M) \mathcal{R}^i(s^i) g(\tilde{\sigma}^i, \sigma^{*N \setminus M \cup \{i\}}, \sigma^M)$.*

Let $\mathcal{B}^k(\mathcal{M}, g)$ be the set of all k-fault-tolerant Bayesian Nash equilibria in the game $(\mathcal{M}, g)$.

Definition 4.12 (k**-FTBNE implementation**)**.** *Given environment*

$$\{N, S, A, \{q^i(s)\}, \{U^i\}\},$$

a social choice set F is (fully) implementable if a mechanism $(\mathcal{M}, g)$ (fully) implements it:

1. $\forall x^* \in F, \exists \sigma^* \in \mathcal{B}^k(\mathcal{M}, g) : \forall s \in T, g[\sigma^*(s)] = x^*(s)$.

2. $\forall \sigma^* \in \mathcal{B}^k(\mathcal{M}, g), \exists x^* \in F : \forall s \in T, g[\sigma^*(s)] = x^*(s)$.

3. $\forall \sigma^* \in \mathcal{B}^k(\mathcal{M}, g), \forall \tilde{\sigma} \in B(\sigma^*, k), \exists \tilde{x} \in F : \forall s \in T, g[\tilde{\sigma}(s)] = \tilde{x}(s)$.

In other words, the mechanism $(\mathcal{M}, g)$ implements F if (1,2) the set of equilibrium outcomes of the mechanism is equivalent to F and (3) if every strategy from k-neighborhood of every equilibrium leads to an outcome that is equivalent to one of the (desirable) outcomes in F.

The first requirement of k-fault-tolerant Bayesian Nash equilibrium implementation is also called partial implementation and it is tied to k-IC. It is equilibrium for everyone to tell the truth and ask for x^* because any personal deviation in reporting from truth-telling would violate k-IC when other players tell the truth. Deviating along other dimensions of the message won't matter because minorities of $k+1$ players by construction of the mechanism are ignored if they stand to gain from their message along truthtelling by others. If they don't stand to gain, the deviation would not be profitable.

The second requirement, in addition to partial implementation, gives classical full implementation. It is a special case of the third requirement where $k = 0$, no faulty players deviated from the equilibrium outcome.

The third requirement comes from Eliaz (2002) and does not appear in Jackson (1991). k-MNV is going to be sufficient to kill equilibria with undesirable k-deviations, though not necessary. Starting from a given social choice set that is fault-tolerant implementable, the necessary condition is the weak k-Bayesian monotonicity.

4.4 Mechanism

The mechanism is adapted from Jackson (1991) but minorities are now $k+1$ and a simpler integer game is played in the last step, instead of the more complicated vector-matching game[23].

Denote player i's message as $\mathcal{M}^i = S^i \times F \times \{\varnothing \cup \mathbb{N}\} \times X \times \{\varnothing \cup X\}$. In this notation, $\mathcal{M} = \mathcal{M}^1 \times \cdots \times \mathcal{M}^N$ is an action space.

$\mathcal{M}$ is partitioned into sets:

$$d_0 = \{m \in \mathcal{M} | \exists M_0 \subset N : |M_0| \geqslant n-k, \exists x \in F \text{ s.t. } \forall j \in M_0, m^j = (\cdot, x, \varnothing, \cdot, \varnothing)\}$$

$$d_1 = \{m \in \mathcal{M} \backslash \{d_0\} | \exists M_1 \subset N : |M_1| = n-k-1, \exists x \in F \text{ s.t. } \forall j \in M_1,$$

$$m^j = (\cdot, x, \varnothing, \cdot, \varnothing), \text{ and } \forall i \in N \backslash M_1, m^i = (\cdot, x, \cdot, \cdot, y)\}$$

$$d_2 = \{m \in \mathcal{M} \backslash \{d_0 \cup d_1\} | \exists M_2 \subset N : |M_2| = n-k-1, \exists x \in F \text{ s.t. } \forall j \in M_2,$$

$$m^j = (\cdot, x, \varnothing, \cdot, \varnothing)\}$$

$$d_3 = \{m \in \mathcal{M} \backslash \{d_0 \cup d_1 \cup d_2\}\}$$

The consequence function, $g : \mathcal{M} \to \mathcal{A}$, defined as:

Rule 1: If $m \in d_0, g(m) = x(m_1)$

At least $N-k$ players agree on $x \in F$ and do not play the integer game.

Rule 2a: If $m \in d_1$ and

$$\forall i \in N \backslash M_1, \forall \beta^{-i} \in B(id, k; A_{-i}), \forall t^i \in S^i :$$

$$x \circ (\beta^{-i}, id) \mathcal{R}^i(t^i) y \circ (\beta^{-i}, m_1^i), g(m) = y(m_1)$$

[23]Vector-matching has the aesthetic advantage of keeping the message space of finite dimension when there are finite number of states.

Exactly $N-k-1$ players as per rule 1 and the remaining $k+1$ players protest x with suggested y choice function that is (weakly) inferior for all of them, in every state - protest approved.

Rule 2b: If $m \in d_1$ and

$$\exists i \in N \backslash M_1, \exists \beta^{-i} \in B(id, k; A_{-i}), \exists t^i \in S^i :$$

$$y \circ (\beta^{-i}, m_1^i)\mathcal{P}^i(t^i)x \circ (\beta^{-i}, id), g(m) = x(m_1)$$

Take the same groups as in rule 3a but one of the protestors strictly benefits from the objection in some state and the protest is denied.

Rule 3: If $m \in d_2, g(m) = x(m_1)$

As in rule 2, except there is no group of $k+1$ protestors with a consistent message that contains a replacement choice function.

Rule 4: If $m \in d_3, g(m) = m_4^{i^*}(m_1)$

i^* is the proposer with the largest integer (if no integers were submitted, take $i^* = 1$) gets to pick an arbitrary choice rule in X (fourth field is mandatory in the message).

4.5 New Results

Theorem 4.13 (**Necessity**). *If a social choice set F is implementable in k-FTBE, then there exists equivalent $\hat{F}$ that satisfies (k-IC) and (wk-BM).*

Proof. See Appendix. □

Theorem 4.14 (**Sufficiency**). *If $|N| \geqslant 3, k < \frac{|N|}{2} - 1$. A social choice set F that satisfies (C), (k-IC) and (k-MNV), is implementable (in k-FTBE).*

1. If F satisfies (k-IC), then: $\forall x^* \in F, \exists \sigma^* \in \mathcal{B}^k(\mathcal{M}, g) : \forall s \in T, g[\sigma^*(s)] = x^*(s)$.

Proof. See Appendix. □

2. $\forall \sigma^* \in \mathcal{B}^k(\mathcal{M}, g), \exists x^* \in F : \forall s \in T, g[\sigma^*(s)] = x^*(s)$.

Proof. Follows from part (3) and $P = \varnothing$. □

3. $\forall \sigma^* \in \mathcal{B}^k(\mathcal{M}, g), \forall \tilde{\sigma} \in B(\sigma^*, k), \exists \tilde{x} \in F : \forall s \in T, z(s) \equiv g[\tilde{\sigma}(s)] = \tilde{x}(s)$.

Proof outline (details are in the Appendix): Want to show that outcome of every k-deviation of FTBE is still desirable. Here Eliaz (2002) used no-veto power that was independent of k (near unanimity for $N-1$ players). In the complete information case, even the faulty players ($i \in P$) that lead to a deviation ($\tilde{\sigma}$) find the outcome (z) weakly best if they were non-faulty in states when i is pivotal: it was possible to reconstruct the outcome (z) when i is non-faulty playing equilibrium (σ^{*i}) by having some other group of faulty players deviate, then the weak superiority of outcome follows from fault-tolerant equilibrium (i pivotal, so can deviate from z but chooses not to).

But in a world of exclusive information, the faulty players leading to z, when they individually turn out to be rational may not be able to trigger the integer game on a neighborhood of every point in D and still get z outside of the neighborhood because D was defined relative to the initial P faulty players.

The idea here is to show z satisfies k-NVH for non-faulty players where players find themselves pivotal with the integer game. Then if z has no equivalent $\tilde{x} \in F$, k-MVN creates a profitable deviation $\overline{z}$ for j under appropriate belief about faulty players where not deviating for j leads to z'. Here Jackson (1991) has to construct only the outcome of deviation ($\overline{z}$) because z is given. Also, player j's interim preference assumption ($x\mathcal{R}^j(t^j)y_{\alpha(s^j)}\forall t^j \in S^j$) covers all $s \in S$ containing s^i. However, now i requires k faulty players $Q = M\backslash\{i\}$ to support his deviation, so the status quo is not z but z' (only Q deviate from σ^*) and so z' needs to be constructed from scratch just like $\overline{z}$, making the definition of k-MNV even longer. The preferences of faulty players

(x vs. y) need to be considered because their information isn't held constant by s^i, which is why $R^i_{\alpha,x,y}$ are introduced in addition to B^i_x in k-MNV.

$$R^i_{\alpha,x,y} = \{s^i \in S^i : \forall \beta^{-i} \in B(id, k; A_{-i}), \forall t^i \in S^i, x \circ (\beta^{-i}, id)\mathcal{R}^i(t^i)y \circ (\beta^{-i}, \alpha^i(s^i))\}.$$

4.6 Conclusion

This paper studies implementation under incomplete information in general environments of Jackson (1991) but with a robust notion of k-fault-tolerant equilibrium of Eliaz (2002). The environment may be non-economic and allows for exclusive information and up to k players could be making mistakes. Assuming closure on the socially desirable set, a new condition, called, k-incentive compatibility is found to be both necessary and sufficient for partial implementation. When the desirable set also satisfies k-monotonicity-no-veto, which is a combination of k-no-veto hypothesis and k-Bayesian monotonicity, then the desired set can be fully implemented.

Several novel challenges arise when faulty players have exclusive information. While the mechanism can safely disregard a minority of k or fewer players asking for the wrong social choice function, their stated information is non-verifiable. Thus, the mechanism cannot filter out any minority deceptions and the social planner understands that these minority deceptions are going to persist in k deviations from desirable equilibria and the corresponding outcomes need to be desirable to achieve full implementation in k-FTBE. The k-IC is strengthened that as long as the other $N-k-1$ players tell the truth, this given player finds it optimal to also tell the truth, while a minority of k players are engaged in arbitrary deceptions (the same deceptions whether the given player deviates or not). In contrast, Jackson (1991) fully reveals the true state of the world for every outcome of Bayesian equilibrium because $k = 0$

and, similarly, Doghmi and Ziad (2007) and Eliaz (2002) observe the true state in every k-deviation from $k-$FTBE because their information is non-exclusive.

Second, in general, the coalition of $k+1$ "whistleblowers" with exclusive information does not share the same information set. Thus, the faulty players generally do not have a constant message along the rational player's information set when he is expected to lead a "whistleblowing" coalition. When the planner does not block the faulty players' messages at s (by construction), he cannot block their messages across states in rational player j's information set $\pi(s^j)$. The planner cannot identify which of the coalition members is rational, and thus he has to give each of the players benefit of the doubt that they may be a rational player and also make sure they do not have a type who would want to lead a false coalition to break a truthful majority playing desirable x.

In Eliaz (2002), a particular state of the world only triggers some rules of the mechanism because information sets are singletons. In some states the integer game is reached, in some it is not. In Doghmi and Ziad (2007) only the first rule of the mechanism is reached in all states of the world because of the specific interplay between exchange economy setting and non-exlusive information. In Jackson (1991), potentially all rules of the mechanism can trigger along a rational player's information set because others' play varies along his information set. However, his own preferences are constant along his information set, and thus the "conflict of interest" condition from the mechanism[24] that needs to hold at a specific state s, will also extend over the whole information set at s^j and does not need to be re-verified in the statement of the Monotonicity-no-Veto that spells out the result of his deviation to z' on different subsets of his information set. In contrast, because this paper uses faulty players with private information in the "whistleblowing" coalition, they may fail the "conflict of interest" condition on a subset of the rational player's information set. Therefore,

[24]That the rational player has no type to pretend to be a whistleblower when the rest is truthful.

whether the whistleblowers are successful in changing the outcome from a majority deception on x to y on some set depends not only on the majority's play but also on the faulty players' preferences, which need to be reverified when the outcome of the rational player's deviation is recorded in $k - MNV$.

Thirdly, the no-veto power condition has been extended to account for k faulty players which was not necessary in Eliaz (2002). Doghmi and Ziad (2007) do not use no-veto power because their mechanism never reaches integer games in any k-deviation of any equilibrium because of the special structure of their model.

The price of the generality in this paper is that it is difficult to contruct examples describing which of the standard sets and correspondences are still implementable and which are not.

Appendix A

Protest Dynamics Details

A.1 Proofs of Results

Proof of ***Proposition 2.2****.* Fix any SPE $(p^*(\theta), c^*_\theta)$. In the following arguments, we will frequently utilize assumption 1 that $\gamma > 2\alpha$. Pick

$$\theta_L = \frac{\gamma}{2}(< \gamma), \theta_M = \gamma, \theta_H > \gamma + \alpha.$$

Then by BR equation (2.1.7), $c_{\theta_L} \in [\theta_L - \alpha, \theta]$ and thus $(0 <)\frac{\gamma - 2\alpha}{2} \leqslant c_{\theta_L} \leqslant \frac{\gamma}{2}$. From the optimal choice of $p^*(\theta)$ in equation (2.1.10), because the citizen cutoff satisfies $c_{\theta_L} \leqslant \gamma$, we get $0 < p^*(\theta_L) = c_{\theta_L} \leqslant \gamma/2 < \gamma - \alpha$.

Secondly, by BR equation (2.1.7), $c_{\theta_M} \leqslant \gamma = \theta_M$. From the optimal choice of $p^*(\theta)$ in equation (2.1.10), because the citizen cutoff satisfies $c_{\theta_H} \leqslant \gamma$, we get $p^*(\theta_M) = c_{\theta_M} \geqslant \gamma - \alpha > p(\theta_L)$.

Thirdly, by BR equation (2.1.7), $c_{\theta_H} \in [\theta_H - \alpha, \theta]$ and thus $(\gamma <)\theta_H - \alpha \leqslant c_{\theta_L} \leqslant \theta_H$. From the optimal choice of $p^*(\theta)$ in equation (2.1.10), because the citizen cutoff satisfies $c_{\theta_H} > \gamma$, we get $p^*(\theta_H) = 0$.

Combining,

$$0 = p^*(\theta_H) < p^*(\theta_L) \leqslant \frac{\gamma + \alpha}{2} < \gamma - \alpha \leqslant p^*(\theta_M)$$

and

$$\theta_L = \frac{\gamma + \alpha}{2} < \theta_M = \gamma < \gamma + \alpha\theta_H.$$

□

Proof of ***Proposition 2.3***. **Statement 1**: Suppose $0 \leqslant \theta \leqslant \gamma$. Then by BR equation (2.1.7), $c^*_\theta \in [\theta - \alpha, \theta]$ and thus $c^*_\theta \leqslant \gamma$. From the optimal choice of $p^*(\theta)$ in equation (2.1.10), because the citizen cutoff satisfies $c_\theta < \gamma$, we get $p^*(\theta) = c_\theta$. Therefore, the equilibrium labor force is 1 and every equilibrium has Never Revolt (NR) at θ that satisfies the hypothesis.

Statement 2: Suppose $\theta > \gamma + \alpha$. Then by BR equation (2.1.7), $c^*_\theta \in [\theta - \alpha, \theta]$ and thus $c^*_\theta \geqslant \theta - \alpha > \gamma$. From the optimal choice of $p^*(\theta)$ in equation (2.1.10), because the citizen cutoff satisfies $c_\theta > \gamma$, we get $p^*(\theta) = 0 < \gamma < c^*_\theta$. Therefore, the equilibrium labor force is 0 and every equilibrium has Always Revolt (AR) at θ that satisfies the hypothesis.

Statement 3: Suppose $\gamma < \theta \leqslant \gamma + \alpha$. Case (i): Take an equilibrium where citizen cutoff satisfies

$$\theta - \alpha \leqslant c_\theta \leqslant \gamma < \theta,$$

then $p^*(\theta) = c_\theta$, the equilibrium labor force is 1 and NR is attained at θ. Otherwise, case (ii): Take an equilibrium where citizen cutoff satisfies

$$\theta - \alpha \leqslant \gamma < c_\theta \leqslant \theta,$$

then $p^*(\theta) = 0 < \gamma < c_\theta$, the equilibrium labor force is 0 and AR is attained at θ. □

The following lemma describes the set of (c_0^*, c_1^*) that is a "half" BR to an arbitrary government strategy p^* in the sense that (2.2.21) holds. Specifically, it starts with Eq. (2.2.21) that $c_a^* \geqslant \underline{p}(p_1^*, a, a_1)$ and evaluates the term $a_1 = \mathbb{1}_{\{p_1^* \geqslant c_1^*\}}$ for arbitrary fixed p_1^*, whether tomorrow's Young are expected to protest or work after observing that today's Young (tomorrow's Old) worked, given some government strategy p^* (not necessarily optimal one for this analysis).

Lemma A.1. *Suppose (p_0^*, p_1^*) is a given government strategy and (c_0^*, c_1^*) is a citizen cutoff strategy.*

1. *If $0 \leqslant p_1^* < B - \alpha$, then c_1^* satisfies (2.2.21) if and only if $c_1^* \geqslant \underline{p}(p_1^*, 1, 0)$*
2. *If $B - \alpha \leqslant p_1^* < B - \frac{2+\beta}{2(1+\beta)}\alpha$, then c_1^* satisfies (2.2.21) if and only if $c_1^* \in [\underline{p}(p_1^*, 1, 1), p_1^*] \cup [\underline{p}(p_1^*, 1, 0), \infty)$*
3. *If $p_1^* \geqslant B - \frac{2+\beta}{2(1+\beta)}\alpha$, then c_1^* satisfies (2.2.21) if and only if $c_1^* \geqslant \underline{p}(p_1^*, 1, 1)$*
4. *c_0^* satisfies (2.2.21) if and only if $c_0^* \geqslant \underline{p}(p_1^*, 0, \mathbb{1}_{\{p_1^* \geqslant c_1^*\}})$*

*Proof of **Lemma A.1**.* Simple arithmetic shows that

$$p_1^* < B - \alpha \iff p_1^* < \underline{p}(p_1^*, 1, 1) = (1+\beta)B - \beta p_1^* - (1+\beta)\alpha. \tag{A.1.1}$$

Similarly,

$$p_1^* \geqslant B - \frac{2+\beta}{2(1+\beta)}\alpha \iff p_1^* \geqslant \underline{p}(p_1^*, 1, 0) = (1+\beta)B - \beta p_1^* - \left(\frac{1+\beta+1}{2}\right)\alpha \tag{A.1.2}$$

Statement 1: In this region, $p_1^* < B - \alpha$ which by (A.1.1) is equivalent to

$$p_1^* < \underline{p}(p_1^*, 1, 1) < \underline{p}(p_1^*, 1, 0), \tag{A.1.3}$$

where the second inequality follows by monotonicity (2.2.42).

"$\Longrightarrow$:" Suppose

$$c_1^* \geqslant \underline{p}(p_1^*, 1, \mathbb{1}_{\{p_1^* \geqslant c_1^*\}}).$$

Then by monotonicity (2.2.42),

$$c_1^* \geqslant \underline{p}(p_1^*, 1, 1) > p_1^*,$$

where the second inequality follows from (A.1.3). Thus, $\mathbb{1}_{\{p_1^* \geqslant c_1^*\}} = 0$.

"$\Longleftarrow$:" Suppose $c_1^* \geqslant \underline{p}(p_1^*, 1, 0)$. Then by monotonicity (2.2.42),

$$c_1^* > \underline{p}(p_1^*, 1, 1) > p_1^*,$$

where the second inequality follows from (A.1.3). Thus, $\mathbb{1}_{\{p_1^* \geqslant c_1^*\}} = 0$.

Statement 2: In this region,

$$B - \alpha \leqslant p_1^* < B - \frac{2+\beta}{2(1+\beta)}\alpha,$$

which by (A.1.1) and (A.1.2) is equivalent to

$$\underline{p}(p_1^*, 1, 1) \leqslant p_1^* < \underline{p}(p_1^*, 1, 0).$$

"$\Longrightarrow$:" Case (i): $c_1^* < \underline{p}(p_1^*, 1, 1)$, then $a_1 = 1$. Take $p : c_1^* < p < \underline{p}(p_1^*, 1, a_1 = 1)$, so (2.2.21) fails by construction of $\underline{p}(p_1^*, 1, a_1 = 1)$ (protest dominant, work prescribed).

Case (ii): otherwise, have

$$c_1^* : p_1^* < c_1^* < \underline{p}(p_1^*, 1, 0),$$

then $a_1 = 0$. Take $p : c_1^* < p < \underline{p}(p_1^*, 1, a_1 = 0)$, so (2.2.21) fails by construction of $\underline{p}(p_1^*, 1, a_1 = 0)$ (protest dominant, work prescribed).

"$\Longleftarrow$:" Case (i):

$$c_1^* \in [\underline{p}(p_1^*, 1, 1), p_1^*],$$

then $a_1 = 1$. For all $p \geqslant c_1^*, p \geqslant \underline{p}(p_1^*, 1, a_1 = 1)$, so (2.2.21) holds by construction of $\underline{p}(p_1^*, 1, a_1 = 1)$.

Case (ii): otherwise,

$$c_1^* \in [\underline{p}(p_1^*, 1, 0), \infty),$$

so $c_1^* > p_1^*$, and then $a_1 = 0$. For all $p \geqslant c_1^*, p \geqslant \underline{p}(p_1^*, 1, a_1 = 1) > \underline{p}(p_1^*, 1, a_1 = 0)$, so (2.2.21) holds by construction of $\underline{p}(p_1^*, 1, a_1 = 0)$.

Statement 3: In this region,

$$p_1^* \geqslant B - \frac{2 + \beta}{2(1 + \beta)},$$

which means

$$p_1^* \geqslant \underline{p}(p_1^*, 1, 0) > \underline{p}(p_1^*, 1, 1)$$

by initial observation.

"$\Longrightarrow$:" Suppose $c_1^* < \underline{p}(p_1^*, 1, 1)$, so $c_1^* < p_1^*$, and then $a_1 = 1$. Like for Part 2, take $p : c_1^* < p < \underline{p}(p_1^*, 1, a_1 = 1)$, so (2.2.21) fails by construction of $\underline{p}(p_1^*, 1, a_1 = 1)$ (protest dominant, work prescribed).

"$\Longleftarrow$:" Suppose $c_1^* \geqslant \underline{p}(p_1^*, 1, 1)$. Case (i): $c_1^* \in [\underline{p}(p_1^*, 1, 1), p_1^*]$, then $a_1 = 1$. For all $p \geqslant c_1^*, p \geqslant \underline{p}(p_1^*, 1, 1a_1 =)$, so (2.2.21) holds by construction of $\underline{p}(p_1^*, 1, a_1 = 1)$.

Case (ii): otherwise, $c_1^* \in (p_1^*, \infty)$, so $c_1^* > p_1^*$, and then $a_1 = 0$. For all $p \geqslant c_1^*, p > p_1^* \geqslant \underline{p}(p_1^*, 1, a_1 = 0)$, so (2.2.21) holds by construction of $\underline{p}(p_1^*, 1, a_1 = 0)$.

Statement 4: Let $a_1 = \mathbb{1}_{\{p_1^* \geqslant c_1^*\}}$. "$\Longrightarrow$:" $c_0^* < \underline{p}(p_1^*, 0, a_1)$. Take $p : c_0^* < p < \underline{p}(p_1^*, 0, a_1)$, so (2.2.21)fails by construction of $\underline{p}(p_1^*, 0, a_1)$ (protest dominant, work prescribed).

"$\Longleftarrow$:" $c_0^* \geqslant \underline{p}(p_1^*, 0, a_1)$. For all $p \geqslant c_0^*, p \geqslant \underline{p}(p_1^*, 0, a_1)$, so (2.2.21) holds by construction of $\underline{p}(p_1^*, 0, a_1)$.

□

Observe that:

$$\overline{p}(p_0^*, 1, 1) + \frac{\alpha}{2} = \overline{p}(p_0^*, 1, 0), \overline{p}(p_0^*, 1, 1) < \overline{p}(p_0^*, 1, 0). \tag{A.1.4}$$

The following lemma describes the set of (c_0^*, c_1^*) that is a "half" BR to an arbitrary government strategy p^* in the sense that (2.2.26) holds. Specifically, it starts with Eq. (2.2.26) that $c_a^* \leqslant \underline{p}(p_0^*, a, a_0)$ and evaluates the term $a_0 = \mathbb{1}_{\{p_0^* \geqslant c_0^*\}}$ for arbitrary fixed p_0^*, whether tomorrow's Young are expected to protest or work after observing that today's Young (tomorrow's Old) protested, given some government strategy p^* (not necessarily optimal one for this analysis).

Lemma A.2. *Suppose (p_0^*, p_1^*) is a given government strategy and (c_0^*, c_1^*) is a citizen cutoff strategy.*

1. *If $0 \leqslant p_0^* \leqslant B - \frac{\beta}{2(1+\beta)}\alpha$, then c_0^* satisfies (2.2.26) if and only if $c_0^* \leqslant \overline{p}(p_0^*, 0, 0)$*
2. *If $B - \frac{\beta}{2(1+\beta)} < p_0^* < B$, then c_0^* satisfies (2.2.26) if and only if*

$$c_0^* \in [0, \overline{p}(p_0^*, 0, 1)] \cup (p_0^*, \overline{p}(p_0^*, 0, 0)]$$

3. *If $p_0^* \geqslant B$, then c_0^* satisfies (2.2.26) if and only if $c_0^* \leqslant \overline{p}(p_0^*, 0, 1)$*
4. *c_1^* satisfies (2.2.26) if and only if $c_1^* \leqslant \overline{p}(p_0^*, 1, \mathbb{1}_{\{p_0^* \geqslant c_0^*\}})$*

*Proof of **Lemma A.2**.* Simple arithmetic shows that

$$p_0^* \leqslant B - \frac{\beta}{2(1+\beta)}\alpha \iff p_0^* \leqslant \overline{p}(p_0^*, 0, 1) = (1+\beta)B - \beta p_0^* - \frac{\beta}{2}\alpha \tag{A.1.5}$$

Similarly,

$$p_1^* \geqslant B \iff p_0^* \geqslant \overline{p}(p_0^*, 0, 0) = (1+\beta)B - \beta p_1^* \tag{A.1.6}$$

Statement 1: In this region,

$$0 \leqslant p_0^* \leqslant B - \frac{\beta}{2(1+\beta)}\alpha$$

. Therefore,

$$p_0^* \leqslant \overline{p}(p_0^*, 0, 1) < \overline{p}(p_0^*, 0, 0)$$

by (A.1.5) and (A.1.6).

"$\Longrightarrow$:" Suppose $c_0^* > \underline{p}(p_1^*, 0, 0)$, so $c_0^* > p_0^*$, and then $a_0 = 0$. Take

$$p : c_0^* > p > \overline{p}(p_0^*, 0, a_0 = 0),$$

so (2.2.26) fails by construction of $\overline{p}(p_0^*, 0, a_0 = 0)$. At this p young citizen has dominant strategy to work ($p > \overline{p}(p_0^*, 0, 0)$ designates the dominance region as $a_0 = 0$ for $c_0^* > p_0^*$) but the prescribed strategy is protest ($p < c_1^*$).

"$\Longleftarrow$:" $c_0^* \leqslant \underline{p}(p_1^*, 0, 0)$, Case (i): $c_0^* \leqslant p_0^*$, ($a_0 = 1$). For all $p : p < c_0^* \leqslant p_0^* \leqslant \overline{p}(p_0^*, 0, a_0 = 1)$, so $p < \overline{p}(p_0^*, 0, a_0 = 1)$ and (2.2.26) holds by construction of $\overline{p}(p_0^*, 0, a_0 = 1)$.

Case (ii): otherwise, $c_0^* : p_0^* < c_0^*$ (so $a_0 = 0$) then for all $p : p < c_0^* \leqslant \overline{p}(p_0^*, 0, a_0 = 0)$, so $p < \overline{p}(p_0^*, 0, a_0 = 0)$ and (2.2.26) holds by construction of $\overline{p}(p_0^*, 0, a_0 = 0)$.

Statement 2: In this region,

$$B - \frac{\beta}{2(1+\beta)}\alpha < p_0^* < B,$$

and thus

$$\overline{p}(p_0^*, 0, 1) < p_0^* < \overline{p}(p_0^*, 0, 0).$$

"$\Longrightarrow$:" Case (i): $c_0^* > \overline{p}(p_0^*, 0, 0)$, then $c_0^* > p_0^*$, so $a_0 = 0$. Take $p : c_0^* > p > \overline{p}(p_0^*, 0, a_0 = 0)$, so (2.2.26) fails by construction of $\overline{p}(p_0^*, 0, a_0 = 0)$ (work dominant, protest prescribed).

Case (ii): otherwise, have $c_0^* \in (\overline{p}(p_0^*, 0, 1), p_0^*]$, then $a_0 = 1$. Take

$$p : c_0^* > p > \overline{p}(p_0^*, 0, a_0 = 1),$$

so (2.2.26) fails by construction of $\overline{p}(p_0^*, 0, a_0 = 1)$ (work dominant, protest prescribed).

"$\Longleftarrow$:" Case (i): $c_0^* \in [0, \overline{p}(p_0^*, 0, 1)]$, then

$$c_0^* \leqslant \overline{p}(p_0^*, 0, 1) < p_0^*,$$

and so $a_0 = 1$. For all $p : p < c_0^* \leqslant \overline{p}(p_0^*, 0, a_0 = 1)$, so (2.2.26) holds by construction of $\overline{p}(p_0^*, 0, 1)$.

Case (ii): otherwise, $c_0^* \in (p_0^*, \overline{p}(p_0^*, 0, 0)]$, so $c_0^* > p_0^*$, and then $a_0 = 0$. For all $p : p < c_0^* \leqslant \overline{p}(p_0^*, 0, a_0 = 0)$, so (2.2.26) holds by construction of $\overline{p}(p_0^*, 0, a_0 = 0)$.

Statement 3: In this region, $p_0^* \geqslant B$ which means

$$p_0^* \geqslant \overline{p}(p_0^*, 0, 0) > \overline{p}(p_0^*, 0, 1)$$

by initial observation.

"$\Longrightarrow$:" Suppose $c_0^* > \overline{p}(p_0^*, 0, 1)$, Case (i):

$$c_0^* \in (\overline{p}(p_0^*, 0, 1), p_0^*],$$

then $a_0 = 1$. Take $p : c_0^* > p > \overline{p}(p_0^*, 0, a_0 = 1)$, so (2.2.26) fails by construction of $\overline{p}(p_0^*, 0, a_0 = 1)$ (work dominant, protest prescribed).

Case (ii): otherwise, $c_0^* \in (p_0^*, \infty)$, so

$$c_0^* > p_0^* \geqslant \overline{p}(p_0^*, 0, 0),$$

and then $a_0 = 0$ and $c_0^* > \overline{p}(p_0^*, 0, 0)$. Take

$$p : c_0^* > p > \overline{p}(p_0^*, 0, a_0 = 0),$$

so (2.2.26) fails by construction of $\overline{p}(p_0^*, 0, a_0 = 0)$ (work dominant, protest prescribed)

"$\Longleftarrow$:" Suppose $c_0^* \leqslant \overline{p}(p_0^*, 0, 1)$, so $c_0^* < p_1^*$, and then $a_0 = 1$. For all $p : p < c_0^* \leqslant \overline{p}(p_0^*, 0, a_0 = 1)$, so (2.2.26) holds by construction of $\overline{p}(p_0^*, 0, a_0 = 1)$.

Statement 4: Let $a_0 = \mathbb{1}_{\{p_0^* \geqslant c_0^*\}}$. "$\Longrightarrow$:" $c_1^* > \overline{p}(p_0^*, 1, a_0)$. Take $p : c_1^* > p > \overline{p}(p_0^*, 1, a_0)$, so (2.2.26) fails by construction of $\overline{p}(p_0^*, 1, a_0)$ (work dominant, protest prescribed).

"$\Longleftarrow$:" $c_1^* \leqslant \overline{p}(p_0^*, 1, a_0)$. For all p:

$$p < c_1^* \leqslant \overline{p}(p_0^*, 1, a_0 = 1),$$

so (2.2.26) holds by construction of $\overline{p}(p_0^*, 1, a_0 = 1)$. □

Given $(p_0^*, p_1^*), (c_0^*, c_1^*)$ cutoff is a best-response to itself and p^*-strategy if and only if c^*-strategy satisfies 2.2.21 and 2.2.26. Lemmas (A.1) and (A.2) considered arbitrary government strategies but from Lemma 1, we only need to focus on four classes of government strategies that partition government's best-responses: NR, AR, TR and CC. In all of these cases even when $p^* > 0$, the citizen's "half" best-response will be an interval. Lemma A.3 is simply a special case of Lemmas (A.1) and (A.2).

Lemma A.3. *Suppose* (p_0^*, p_1^*) *is a given government strategy and* (c_0^*, c_1^*) *is a citizen cutoff strategy.*

1. *If* $c_1^* = p_1^*$, *then* c_1^* *satisfies* (2.2.21) *if and only if* $p_1^* \in [\underline{p}(p_1^*, 1, 1), \infty) = [B - \alpha, \infty)$

2. *If* $c_0^* = p_0^*$, *then* c_0^* *satisfies* (2.2.26) *if and only if* $p_0^* \in [0, \overline{p}(p_0^*, 0, 1)] = [0, B - \frac{\beta}{2(1+\beta)}\alpha]$

Proof of ***Lemma A.3***. **Statement 1**: Since $c_1^* = p_1^* \equiv p_1$, so $a_1 = 1$. Combine part 1 of A.1 and part 3 of A.1. First, suppose $p_1^* \in [0, B - \alpha)$ satisfying necessary condition on p_1^* in Lemma A.1.1. When this happens, by Lemma A.1.1, c_1^* satisfies (2.2.21) if and only if

$$\begin{aligned} c_1^*(= p_1^*) &\geqslant \underline{p}(p_1^*, 1, 0) \\ &= (1+\beta)B - \beta p_1^* - \left(\frac{2+\beta}{2}\right)\alpha \\ &= B - \frac{2+\beta}{2(1+\beta)}\alpha. \end{aligned} \tag{A.1.7}$$

This leads to a contradiciton that $p_1^* < B - \alpha$ and

$$p_1^* \geqslant B - \frac{2+\beta}{2(1+\beta)}\alpha > B - \alpha,$$

hence for $p_1^* \in [0, B - \alpha)$, (2.2.21) doesn't hold.

Secondly, suppose

$$p_1^* \in [B - \alpha, B - \frac{2+\beta}{2(1+\beta)}\alpha)$$

satisfying necessary condition on p_1^* in Lemma A.1.2. When this happens, by Lemma A.1.2, c_1^* satisfies (2.2.21) if and only if

$$c_1^*(= p_1^*) \in [\underline{p}(p_1^*, 1, 1), p_1^*] \cup [\underline{p}(p_1^*, 1, 0, \infty).$$

Note that

$$p_1^* \in [\underline{p}(p_1^*, 1, 1), p_1^*]$$

if and only if $p_1^* \geqslant B - \alpha$. Then (2.2.21) holds

$$\iff p_1^* \in [B - \alpha, B - \frac{2+\beta}{2(1+\beta)}\alpha)$$

as every point for the initial hypothesis satisfied (2.2.21) in this case.

Finally, suppose

$$p_1^* \geqslant B - \frac{2+\beta}{2(1+\beta)}\alpha)$$

satisfying necessary condition on p_1^* in Lemma A.1.3. When this happens, by Lemma A.1.2, c_1^* satisfies (2.2.21) $\iff c_1^*(= p_1^*) \geqslant \underline{p}(p_1^*, 1, 1) = B - \alpha$. Then (2.2.21) holds if and only if

$$p_1^* \geqslant B - \frac{2+\beta}{2(1+\beta)}\alpha),$$

where again the initial hypothesis is the tighter condition.

Combining these three cases: (2.2.21) holds if and only if

$$p_1^* \in [B - \alpha, B - \frac{2+\beta}{2(1+\beta)}\alpha) \cup [B - \frac{2+\beta}{2(1+\beta)}\alpha, \infty) = [B - \alpha, \infty).$$

So $c_1^* \in [\underline{p}(p_1^*, 1, 0), \infty)$ if and only if (2.2.21) holds.

Statement 2: The second proof is analogous. Since $c_0^* = p_0^* \equiv p_0$, so $a_0 = 1$. Combine results from parts Lemma A.2.1-A.2.3. First, consider

$$p_0^* \in [0, B - \frac{\beta}{2(1+\beta)}\alpha] = [0, \overline{p}(p_0^*, 0, 1)]$$

satisfying part 2.3.1 and have (2.2.26) if and only if $c_0^*(= p_0^*) \in [0, \overline{p}(p_0^*, 0, 0)] = [0, B]$, and (2.2.26) holds for every $p_0^* \in [0, B - \frac{\beta}{2(1+\beta)}\alpha]$ as the hypothesis was tighter.

Secondly, if

$$p_0^* \in (B - \frac{\beta}{2(1+\beta)}\alpha, B) = (\overline{p}(p_0^*, 0, 1), \overline{p}(p_0^*, 0, 0))$$

satisfying 2.3.2, have (2.2.26) if and only if

$$c_0^*(=p_0^*) \in [0, \overline{p}(p_0^*, 0, 1)] \cup (p_0^*, \overline{p}(p_0^*, 0, 0)] = [0, B - \frac{\beta}{2(1+\beta)}\alpha] \cup (p_0^*, B]$$

Since by hypothesis, $p_0^* > B - \frac{\beta}{2(1+\beta)}\alpha$, it doesn't belong to

$$[0, B - \frac{\beta}{2(1+\beta)}\alpha] \cup (p_0^*, B]$$

and (2.2.26) fails for every such p_0^*.

Finally,if $p_0^* \geqslant B$ satisfying Lemma A.2.3, have (2.2.26) $\iff$ $c_0^*(=p_0^*) \in [0, \overline{p}(p_0^*, 0, 1)] = [0, B - \frac{\beta}{2(1+\beta)}\alpha]$. Since

$$p_0^* \geqslant B \notin [0, B - \frac{\beta}{2(1+\beta)}\alpha],$$

(2.2.26) fails for every such p_0^*.

Combining the three cases,

$$p_0^* \in [0, \overline{p}(p_0^*, 0, 1)] = [0, B - \frac{\beta}{2(1+\beta)}\alpha]$$

if and only if (2.2.26) holds.

□

Proof of ***Proposition 2.12****.* **Statement 1**: In the NR case policing strategy equals the cutoff strategy and citizens always work. (c_0^*, c_1^*) satisfies $BR(p_0^*, p_1^*)$ when (2.2.21) and (2.2.26) both hold if and only if

$$p_1 \in [\underline{p}(p_1^*, 1, 1), \overline{p}(p_0^*, 1, 1)]$$

by combining Lemma A.3.1 with Lemma A.2.4, and

$$p_0 \in [\underline{p}(p_1^*, 0, 1), \overline{p}(p_0^*, 0, 1)]$$

by comining Lemma A.1.4 with Lemma A.3.2.

Statement 2: In the AR case policing is always zero and citizens never work. (c_0^*, c_1^*) satisfies $BR(p_0^*, p_1^*)$ when (2.2.21) and (2.2.26) both hold if and only if

$$p_1 \in [\underline{p}(p_1^*, 1, 0), \overline{p}(p_0^*, 1, 0)]$$

by combining Lemma A.1.1 and Lemma A.2.4, and

$$p_0 \in [\underline{p}(p_1^*, 0, 0), \overline{p}(p_0^*, 0, 0)]$$

by combining Lemma A.1.4 with Lemma A.3.1.

Statement 3: In the TR case labor is history-dependent as government only polices if the old were working (if it policed last period). (c_0^*, c_1^*) satisfies $BR(p_0^*, p_1^*)$ when (2.2.21) and (2.2.26) both hold if and only if

$$p_1 \in [\underline{p}(p_1^*, 1, 1), \overline{p}(p_0^*, 1, 0)]$$

by combining Lemma 3.4.1 and Lemma 3.3.4, and

$$p_0 \in [\underline{p}(p_1^*, 0, 1), \overline{p}(p_0^*, 0, 0)]$$

by combining Lemma A.1.4 with Lemma A.3.1.

Statement 4: Here $(p_0^* = c_0^*, p_1^* = 0)$ and $a_0 = 1, a_1 = 0$. (c_0^*, c_1^*) satisfies $BR(p_0^*, p_1^*)$ when

$$p_1 \in [\underline{p}(p_1^*, 1, 0), \overline{p}(p_0^*, 1, 1)]$$

by combining Lemma 3.2.1 and Lemma 3.3.4. Secondly, from Lemmas 3.2.4 and 3.4.2,

$$c_0^* \in \left[\underline{p}(p_1^*, 0, 0), \overline{p}(p_0^*, 0, 1)\right] = \left[(1+\beta)B - \left(\frac{1+\beta}{2}\right)\alpha, (1+\beta)B - \beta c_0^* - \frac{\beta}{2}\alpha\right] = \left[(1+\beta)\left(B - \frac{\alpha}{2}\right), B - \frac{\beta}{2(1+\beta)}\alpha\right]$$

□

Proof of ***Proposition 2.13****.* When

$$\beta > \underline{\beta} = \frac{-1 + \sqrt{1 + \frac{4}{2\frac{B}{\alpha} - 1}}}{2},$$

citizens are patient and

$$\beta(1+\beta)B > \frac{\alpha}{2}\left((1+\beta)^2 - \beta\right),$$

rearranging to get

$$\beta B > \frac{\alpha}{2}\left((1+\beta) - \frac{\beta}{1+\beta}\right)$$

implies

$$B + \beta B - (1+\beta)\frac{\alpha}{2} > B - \frac{\beta}{2(1+\beta)}\alpha$$

Thus,

$$(1+\beta)\left(B - \frac{\alpha}{2}\right) > B - \frac{\beta}{2(1+\beta)}\alpha,$$

giving a contradiction. This means, arbitrary (c_0^*, c_1^*) are not in $BR(p_0^*, p_1^*) = \varnothing$ because (2.2.21) and (2.2.26) contradict each other in this case. □

Proof of ***Proposition 2.14***. "$\Longrightarrow$:" Will show that if $(1+\delta)\gamma > (1+\beta)B - \frac{\alpha}{2}$, then there is no AR equilibrium. Suppose there is: from one-shot deviation results in Proposition 2.11,

$$c_1^* \geqslant (1+\delta)\gamma > (1+\beta)B - \frac{\alpha}{2},$$

so $c_1^* > (1+\beta)B - \frac{\alpha}{2}$. From $BR(p_0^*, p_1^*)$ Proposition 3.2, $c_1^* \leqslant (1+\beta)B - \frac{\alpha}{2}]$, a contradiction.

"$\Longleftarrow$:" If $(1+\delta)\gamma \leqslant (1+\beta)B - \frac{\alpha}{2}$, will verify

$$\{(c_0^*, c_1^*) = \left((1+\beta)B, (1+\beta)B - \frac{\alpha}{2}\right)$$

is an equilibrium. No profitable deviation of Proposition 2.11 holds for each $a \in \{0,1\}$, have

$$c_a^* \geqslant (1+\beta)B - \frac{\alpha}{2} \geqslant (1+\delta)\gamma.$$

$BR(p_0^*, p_1^*)$ Proposition 2.12 also holds:

$$c_0^* = (1+\beta)B \in [(1+\beta)\left(B - \frac{\alpha}{2}\right), (1+\beta)B]$$

holds and

$$c_1^* = (1+\beta)B - \frac{\alpha}{2} \in \left[(1+\beta)B - \left(1 + \frac{\beta}{2}\right)\alpha, (1+\beta)B - \frac{\alpha}{2}\right]$$

also holds. □

Proof of ***Proposition 2.15***. **Statement 1**. "$\Longrightarrow$:" Case (i): Will show that if

$$(1+\delta)\gamma > (1+\beta)B - \delta\frac{\alpha}{2},$$

then there is no TR equilibrium. Suppose there is. From $BR(p_0^*, p_1^*)$ Proposition 3.3: $c_1^* \leqslant (1+\beta)B - \frac{\alpha}{2}$ and $c_0^* \leqslant (1+\beta)B$. Also, from government's choice in Proposition 2.11.4,

$$c_1^* \geqslant \frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}c_0^* > \frac{(1+\beta)B}{\delta} - \frac{\alpha}{2} - \left(\frac{1}{\delta} - 1\right)(1+\beta)B = (1+\beta)B - \frac{\alpha}{2}, \quad \text{(A.1.8)}$$

Combining we get that $c_1^* > (1+\beta)B - \frac{\alpha}{2}$, which is a contradiction.

Case (ii): Will show that if $(1+\delta)\gamma < B - \alpha$, then there is no TR equilibrium. Suppose there is:from government's choice in Proposition 2.11.4, $c_1^* \leqslant (1+\delta)\gamma < B - \alpha$, so $c_1^* < B - \alpha$. From $BR(p_0^*, p_1^*)$ Proposition 2.12.3, $c_1^* \geqslant (1+\beta)B - \beta c_1^* - (1+\beta)\alpha$. Grouping and simplifying, this is equivalent to $c_1^* \geqslant B - \alpha$, which is a contradiction.

"$\Longleftarrow$:" If $(1+\delta)\gamma \in \left[B - \alpha, (1+\beta)B - \delta\frac{\alpha}{2}\right]$, then the set of TR equilibria is non-empty. Consider a monotonic sequence of proposed equilibria parametrized by value of police productivity $(1+\delta)\gamma$ as follows:

$$\left\{p^* = (0, c_1^*), c_0^* = (1+\beta)B, c_1^* = \max\left\{\frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}(1+\beta)B, B - \alpha\right\}\right\} \quad \text{(A.1.9)}$$

Government's choice in Proposition 2.10.4 for $a = 0$ requires that

$$c_1^* \geqslant \frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}c_0^*.$$

This is satisfied by construction of c^* in (A.1.9). Next, we're going to expand the maximum operator in $c_1^* = \max\{\frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}(1+\beta)B, B-\alpha\}$ as the police-productivity parameter varies over the region of interest $(1+\delta)\gamma \in [B - \alpha, (1+\beta)B - \delta\frac{\alpha}{2}]$

First, note that

$$c_1^* = B - \alpha \iff (1+\delta)\gamma = \delta(B-\alpha) + (1-\delta)(1+\beta)B \equiv \overline{\gamma}.$$

Since c_1^* is weakly increasing in $(1+\delta)\gamma$, for all $(1+\delta)\gamma < \overline{\gamma}$, we also have $c_1^* = B - \alpha$. On the other hand, for all $(1+\delta)\gamma \geqslant \overline{\gamma}$, we have

$$c_1^* = \frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}(1+\beta)B.$$

Next we will show that

$$\overline{\gamma} \in (B - \alpha, (1+\beta)B - \delta\frac{\alpha}{2}),$$

so that the constructed c_1^* is a step function at $\overline{\gamma}$. First check that $\overline{\gamma}$ is smaller than $(1+\beta)B - \delta\frac{\alpha}{2})$.

$$\overline{\gamma} = \delta(B-\alpha) + (1-\delta)(1+\beta)B < \frac{B-\alpha}{2} + \frac{(1+\beta)B}{2} = \frac{3}{2}B - \frac{\alpha}{2} < (1+\beta)B - \delta\frac{\alpha}{2} \tag{A.1.10}$$

Secondly, check that $\overline{\gamma}$ is greater than $B - \alpha$.

$$\overline{\gamma} = \delta(B-\alpha) + (1-\delta)(1+\beta)B > \delta(B-\alpha) + (1-\delta)(B-\alpha) = B - \alpha \tag{A.1.11}$$

Now we can write c_1^* as a step function over the region of interest:

$$c_1^* = \begin{cases} B - \alpha & \text{if } (1+\delta)\gamma \in [B-\alpha, \overline{\gamma}], \\ \frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}(1+\beta)B & \text{if } (1+\delta)\gamma \in (\overline{\gamma}, (1+\beta)B - \delta\frac{\alpha}{2}]. \end{cases} \tag{A.1.12}$$

Case (i): Suppose

$$(1+\delta)\gamma \in [B-\alpha, \delta(B-\alpha)+(1-\delta)(1+\beta)B].$$

Here $c_1^* = B - \alpha \leqslant (1+\delta)\gamma$, thus Proposition 2.10.4 for $a = 1$ holds.

Case (ii): Suppose

$$(1+\delta)\gamma \in (\delta(B-\alpha)+(1-\delta)(1+\beta)B, (1+\beta)B - \delta\frac{\alpha}{2}].$$

Here

$$(1+\delta)\gamma < (1+\beta)B - \delta\frac{\alpha}{2} < (1+\beta)B.$$

therefore

$$c_1^* = \frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}(1+\beta)B \tag{A.1.13}$$

$$< \frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}(1+\delta)\gamma \tag{A.1.14}$$

$$= (1+\delta)\gamma\frac{1-1+\delta}{\delta} = (1+\delta)\gamma \tag{A.1.15}$$

Once again, $c_1^* \leqslant (1+\delta)\gamma$ satisfies 2.10.4 $a = 1$.

$BR(p_0^*, p_1^*)$ at $a = 0$, requires that

$$c_0^* \in [\underline{p}(p_1^*, 0, 1), \overline{p}(p_0^*, 0, 0)]$$

It should be clear the following is true:

$$(1+\beta)B = c_0^* \in [(1+\beta)B - \beta p_1^* - (\frac{1+2\beta}{2}\alpha, (1+\beta)B].$$

It remains to check $BR(p_0^*, p_1^*)$ at $a = 1$, which requires that:

$$c_1^* \in [(1+\beta)B - \beta c_1^* - (1+\beta)\alpha, (1+\beta)B - \frac{\alpha}{2}].$$

The lower bound holds if and only if

$$c_1^* \geqslant (1+\beta)B - \beta c_1^* - (1+\beta)\alpha \iff c_1^* \geqslant B - \alpha,$$

which is true by construction of the second argument of the maximum in

$$p_1 = \max(\frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}(1+\beta)B, B - \alpha)$$

Checking upper-bound: Case (i): if $c_1^* = B - \alpha < (1+\beta)B - \frac{\alpha}{2}$ holds.

Case (ii): otherwise,

$$c_1^* = \frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta}(1+\beta)B \tag{A.1.16}$$

$$\leqslant \frac{(1+\beta)B - \delta\frac{\alpha}{2}}{\delta} - \frac{1-\delta}{\delta}(1+\beta)B = (1+\beta)B - \frac{\alpha}{2} \tag{A.1.17}$$

Simplifying, we confirm that $c_1^* \leqslant (1+\beta)B - \frac{\alpha}{2}$ holds as well.

Statement 2. There is no robust equilibrium that holds over the whole range where the TR set is non-empty.

$$\varnothing = \cap_{\gamma \in [\underline{\gamma}(TR), \overline{\gamma}(TR)]} \mathcal{E}^{TR}(\gamma).$$

Suppose there was, then no one-shot deviation 2.10.4 for $a = 1$ at the lowest productivity it gives $c_1^* \leqslant (1+\delta)\gamma = B - \alpha$. From $BR(p_0^*, p_1^*)$ $c_0^* \leqslant (1+\beta)B$ and at the

highest productivity Proposition 2.4 for $a = 0$ gives

$$c_1^* \geqslant \frac{(1+\delta)\gamma}{\delta} - \frac{1-\delta}{\delta} c_0^* \tag{A.1.18}$$

$$\geqslant \frac{(1+\beta)B - \delta\frac{\alpha}{2}}{\delta} - \frac{1-\delta}{\delta}(1+\beta)B \tag{A.1.19}$$

$$= (1+\beta)B - \frac{\alpha}{2} > B - \alpha \Longrightarrow c_1^* > B - \alpha, \tag{A.1.20}$$

reaching a contradiction. □

Proof of ***Proposition 2.17****.* **Statement 1**: "$\Longrightarrow$:" Will show that if $(1+\delta)\gamma < B - (1+\delta)\frac{\alpha}{2}$, then there is no NR equilibrium. Suppose there was, then from $BR(p_0^*, p_1^*)$ Proposition 3.1 ($c_a^* \in [\underline{p}(p_1^*, a, 1), \overline{p}(p_0^*, a, 1)]$) at $a = 1$, get $c_1^* \geqslant B - \alpha$. From $BR(p_0^*, p_1^*)$ at $a = 0$, we have $c_0^* \geqslant (1+\beta)B - \beta c_1^* - \frac{1+2\beta}{2}\alpha$. From Proposition 2.11.2 for $a = 0$ get $c_0^* \leqslant \frac{(1+\delta)\gamma}{1-\delta} - \frac{\delta}{1-\delta} c_1^*$. Combining,

$$\frac{(1+\delta)\gamma}{1-\delta} - \frac{\delta}{1-\delta} c_1^* \geqslant c_0^* \geqslant (1+\beta)B - \beta c_1^* - \frac{1+2\beta}{2}\alpha \tag{A.1.21}$$

Rearranging and grouping terms gives:

$$(\delta - \beta(1-\delta))c_1^* < (\delta - \beta(1-\delta))B - (\delta - \beta(1-\delta))\alpha$$

When government patience δ is not high enough,

$$\delta = \frac{\beta}{1+\beta} \iff (\delta - \beta(1-\delta)) = 0$$

This gives $0 < 0$ contradiction. Otherwise, for

$$\frac{\beta}{1+\beta} < \delta < 1 \iff \delta - \beta(1-\delta) > 0$$

and dividing both sides of inequality by it, $c_1^* < B - \alpha$, a contradiction to $BR(p_0^*, p_1^*)$ for $a = 1$.

"$\Longleftarrow$:" If the following inequality is satisfied,

$$(1+\delta)\gamma \geqslant B - (1+\delta)\frac{\alpha}{2},$$

will verify $\{p^* = (c_0^*, c_1^*) = (B - \frac{\alpha}{2}, B - \alpha)$ is an equilibrium. 2.11.2 $(a = 0)$ requires

$$c_0^* \leqslant \frac{(1+\delta)\gamma}{1-\delta} - \frac{\delta}{1-\delta}c_1^*.$$

Substituting $c_1^* = B - \alpha$, and restriction on productivty, $RHS \geqslant \frac{B-(1+\delta)\frac{\alpha}{2}}{1-\delta} - \frac{\delta}{1-\delta}(B - \alpha) = B - \frac{\alpha}{2} = c_0^*$ holds. 2.11.2 $(a = 1)$ requires

$$c_1^*(1+\delta) \leqslant \gamma(1+\delta) + \delta c_0^*.$$

Expanding similarly,

$$RHS \geqslant B - \frac{(1-\delta)\alpha}{2} + \delta(B - \frac{\alpha}{2}) = (1+\delta)B - \frac{\alpha}{2} >$$
$$(1+\delta)B - 2(1+\delta)\frac{\alpha}{2} = (1+\delta)(B-\alpha) = (1+\delta)c_1^*.$$

Next we will check that $(c_0^*, c_1^*) \in BR(p_0^*, p_1^*)$. Proposition 2.12 requires

$$c_a^* \in [\underline{p}(p_1^*, a, 1), \overline{p}(p_0^*, a, 1)]$$

Evaluating the expression for $a = 0$ and $a = 1$,

$$B - \frac{\alpha}{2} = c_0^* \in [B - \frac{\alpha}{2}, B]$$

holds and

$$B - \alpha = c_1^* \in [B - \alpha, B - \frac{\alpha}{2}]$$

holds as well.

Statement 2: We have

$$\frac{\beta}{1+\beta} < \delta < 1 \iff \delta - \beta(1-\delta) > 0$$

and

$$(1+\delta)\gamma = B - (1+\delta)\frac{\alpha}{2}$$

Will show that every NR equilibrium is the same and equal to

$$\{p^* = (c_0^*, c_1^*) = (B - \frac{\alpha}{2}, B - \alpha\}$$

. From similar calculations as above for (A.1.21),

$$(\delta - \beta(1-\delta))c_1^* \leqslant (\delta - \beta(1-\delta))B - (\delta - \beta(1-\delta))\alpha \iff c_1^* \leqslant B - \alpha$$

and from $BR(p_0^*, p_1^*)$ Proposition 2.12 at $a = 1, c_1^* \geqslant B - \alpha$ as above. Thus, $p_1^* = c_1^* = B - \alpha$.

Substitute $c_1^* = B - \alpha$ and $(1+\delta)\gamma = B - (1+\delta)\frac{\alpha}{2}$ into (A.1.21):

$$\frac{B - (1+\delta)\frac{\alpha}{2}}{1-\delta} - \frac{\delta}{1-\delta}(B - \alpha) \geqslant c_0^* \geqslant (1+\beta)B - \beta c_1^* - \frac{1+2\beta}{2}\alpha$$

$$B\frac{1-\delta}{1-\delta} - \frac{\alpha}{2}\left(\frac{1+\delta-2\delta}{1-\delta}\right) = B - \frac{\alpha}{2} \geqslant c_0^* \geqslant B - \frac{\alpha}{2}.$$

Thus, $p_0^* = c_0^* = B - \frac{\alpha}{2}$. □

Proof of ***Proposition 2.27****.* From definition of Δ^t we get $\Delta^t \geqslant 0$ if and only if

$$\underline{p}(p_1^{t+1}, a, a_1) \leqslant \overline{p}(p_0^{t+1}, a, a_0)$$

for each $a \in \{0, 1\}$. Suppose $\Delta^t \geqslant 0$, then take $c_a^{*,t} = \underline{p}(p_1^{t+1}, a, a_1)$, which will satisfy $BR^t(p_0^{*,t+1}, p_1^{*,t+1})$.

Suppose $\exists\{(c_0^{*,t}, c_1^{*,t})\}$ satisfying $BR^t(p_0^{*,t+1}, p_1^{*,t+1})$, and thus

$$\underline{p}(p_1^{t+1}, a, a_1) \leqslant c_a^{*,t} \leqslant \overline{p}(p_0^{t+1}, a, a_0).$$

This implies $\Delta^t =\geqslant 0$.

Finally,

$$0 \leqslant \Delta^t = \beta(p_1^{t+1} - p_0^{t+1}) + \alpha\left(\frac{(1+\beta)}{2} + \frac{\beta(a_1 - a_0)}{2}\right)$$

if and only if

$$p_1^{*,t+1} - p_0^{*,t+1} \geqslant -\frac{\alpha}{2\beta}\left(1 + \beta + \beta(\mathbb{1}_{\{p_1^{*,t+1} \geqslant c_1^{*,t+1}\}} - \mathbb{1}_{\{p_0^{*,t+1} \geqslant c_0^{*,t+1}\}})\right)$$

□

Proof of ***Proposition 2.29****.* **Statement 1**. First, observe Government's continuation utility in every equilibrium p^* is bounded by 1 in every state when it receives full employment (maximum benefit) and pays for no policing (minimum cost): $\forall a \in \{0, 1\} : G(a|p^*) \leqslant 1$. Since the payoffs are bounded, the following least-upper bound on government payoffs exists and it's taken across all future periods and all Markov Perfect Equilibria: $\overline{G}_a \equiv \sup_{p^*, t \geqslant N} G^t(a|p^*) \leqslant 1$.

(i) Suppose the state is $a = 1$. From definition of $\overline{G}_1$ as supremum we can identify an equilibrium with a state in this region where the government receives a similar payoff:

$$\forall \epsilon > 0, \exists p^*, \exists T \geqslant N : \overline{G}_1 < G^T(1|p^*) + \epsilon \tag{A.1.22}$$

We are going to show that the dominant action at $a = 1$ is no policing and corresponding protest. First, take

$$2\epsilon = (1-\delta)\left(-\frac{\delta}{2} - \left(\frac{1}{2} - \frac{1}{2\gamma}\underline{P}_1\right)\right)$$

Re-arranging Assumption 4 implies ϵ is positive.

$$(1-\delta)\left(\frac{1}{2} - \frac{1}{2\gamma}\underline{P}_1\right) = -2\epsilon - \frac{\delta(1-\delta)}{2}$$

Adding $\frac{1-\delta}{2} + \epsilon$ to both sides gives:

$$(1-\delta)\left(1 - \frac{1}{2\gamma}\underline{P}_1\right) + \epsilon = -\epsilon - \frac{\delta(1-\delta)}{2} + \frac{1-\delta}{2} = -\epsilon + \frac{(1-\delta)^2}{2} \tag{A.1.23}$$

Suppose instead the government's optimal action is policing $p_1^{*,T} > 0$ to induce work $a^*(1, p^{*,T}) = 1$.

$$\overline{G}_1 < G^T(1|p^*) + \epsilon \leqslant (1-\delta)\left(1 - \frac{1}{2\gamma}\underline{P}_1\right) + \epsilon + \delta\overline{G}_1 \tag{A.1.24}$$

$$= \frac{(1-\delta)^2}{2} - \epsilon + \delta\overline{G}_1 \leqslant \frac{(1-\delta)^2}{2} + \delta\overline{G}_1 \tag{A.1.25}$$

The first inequality comes from A.1.22 and the second inequality comes from A.1.23.

Simplifying, $\overline{G}_1 < \frac{1-\delta}{2}$, which is a contradiction because government can always guarantee itself a payoff of $\frac{1-\delta}{2}$ in state $a = 1$ by playing no policing forever. Thus,

$p^{*,T} = 0$ and

$$\overline{G}_1 \leqslant \frac{1-\delta}{2} + \delta\overline{G}_0$$

(ii) Suppose the state is $a = 0$. Now by a similar argument, using the same ϵ, there exists time S and equilibrium p' satisfying $\overline{G}_0 < G^S(0|q') + \epsilon$. Suppose to the contrary that there is positive policing, $p_0'^S > 0$,

$$\overline{G}_0 < (1-\delta)\left(\frac{1}{2} - \gamma\underline{P}_0\right) + \epsilon + \delta\overline{G}_1 < (1-\delta)\left(-\frac{\delta}{2}\right) + \delta\overline{G}_1$$

Combining with $\overline{G}_1 \leqslant \frac{1-\delta}{2} + \delta\overline{G}_0$, we get:

$$\overline{G}_0 < (1-\delta)\left(\frac{-\delta}{2}\right) + \delta\left(\frac{1-\delta}{2}\right) + \delta^2 G_0 = \delta^2 G_0$$

This implies $\overline{G}_0 < 0$, which is a contradiction because a payoff of zero is attainable by playing no policing forever. The only other option is that $p_0'^S = 0$, and thus protest is optimal in that state. In this case, $\overline{G}_0 \leqslant \delta\overline{G}_0$ if and only if $\overline{G} = 0$.

Finally, $\overline{G}_1 \geqslant \frac{1}{\delta}\overline{G}_0 + \frac{1-\delta}{2}$, giving

$$\frac{1-\delta}{2} \leqslant \overline{G}_1 \leqslant \frac{1-\delta}{2}$$

Under Assumption 4, the payoffs of $\overline{G}_0 = 0$ and $\overline{G}_1 = \frac{1-\delta}{2}$ are attained in every period of all MPE by never policing forever. It follows that $p_0^* = p_1^* = 0$ for all $t \geqslant N$.

Statement 2. Continuation game has AR in every period from part 1. From the first column of Table 1, playing $\tilde{p} = c_a^{*,K}$ gives strictly smaller payoff than playing $p = 0$ if and only if $c_a^{*,K} > (1+\delta)\gamma_K$. This is true because $c_a^{*,K} \geqslant (1+\beta)B - \alpha\left(1 + \frac{\beta}{2}\right) > (1+\delta)\gamma_K$ □

Proof of ***Theorem 2.31****.* $L = \max_l\{(1+\delta)\gamma_l > (1+\beta)B\}$, the last period in the upper dominance region.

Using Prop. 2.28, going long enough back into the past, gives policing in state $a = 1$ that is close to the boundary of the fixed-point set, "FP". In the worst case, p_1 is slightly larger than its maximal boundary.

Formally, $\forall \epsilon > 0, \exists M < L$:

$$p_1^{*,M} < B - \frac{\alpha(1 - 2\beta^2)}{2(1 - \beta^2)} + \epsilon.$$

[1]

Let

$$2\epsilon \equiv (1 + \beta)B - \alpha\left(1 + \frac{\beta}{2}\right) - \left(B + \frac{\alpha\beta^2}{2(1 - \beta^2)}\right)$$

By Asmp.4 (second argument), $\epsilon > 0$, hence $p_1^{*,t=M} < p_1^{*,T-1}$.

When government gives up power, the previous period has a greater policing level. $\exists T : p_1^{*,T} = 0$ and

$$p_1^{*,T-1} \geqslant (1 + \beta)B - \alpha\left(1 + \frac{\beta}{2}\right).$$

By Asmp 5, $p_1^{*,T-1} > p_1^{*,T} = 0$. □

[1] This is the historical policing awhile before crisis is small.

Appendix B

Incumbency Advantage Details

B.1 Deriving Utilities

B.1.1 General case: rational expectations (q,Qs)

Characterizing Challenger's utility, given no shock

Let reference point be

$$c_p = p(Q_s c_s^{\theta} + (1 - Q_s)c_s^{i}) + (1 - p)(qc^{\theta} + (1 - q)c^{i})$$

where with probability p the voter will learn there is s shock next period before voting and with probability $1 - p$ he learns there will be no shock next period. Q_s is the cutoff for incumbent, when shock happens and q the cutoff when no shock happens. Here the reference point c_p depends on θ ability of a reference draw for the challenger, distinct from η^c which would be an actual draw (different and independent of θ).

For brevity of notation define, $G = \alpha(\tau y - \overline{r})$ is the coefficient of the ability in making the public good.

The corresponding consumptions (next period) are:

1. $c_s^{\theta} = (y - s)(1 - \tau) + \theta G$. Reference consumption under challenger type $\theta : s \geqslant 0$

2. $c^{\theta} = y(1-\tau) + \theta G$. Reference consumption under challenger type $\theta : s = 0$

3. $c_s^i = (y-s)(1-\tau) + \tilde{\eta} G$. Consumption under incumbent type $\tilde{\eta}$: under s

4. $c^i = y(1-\tau) + \tilde{\eta} G$. Consumption under incumbent type $\tilde{\eta} : s = 0$

5. $c_s^c = (y-s)(1-\tau) + \eta^c G$. Consumption under challenger type $\eta^c : s \geqslant 0$

6. $c^c = y(1-\tau) + \eta^c G$. Consumption under challenger type $\eta^c : s = 0$

Lemma B.1 (Expanding $\int_0^1 \int_0^1 \mu(.) d\eta^c d\theta$)**.** *Given differentiable $x(\theta, \eta^c)$ with constant $\frac{\partial x}{\partial \theta} < 0$, and constant $\frac{\partial x}{\partial \eta^c} > 0$, and $\frac{d\eta^c}{d\theta} \in (0,1)$ let*

$$h_\theta = \sup\{\{0\} \cup \{\eta^c \in [0,1] : x(\theta, \eta^c) \leqslant 0\}\}.$$

Then

1. *$x(\theta, \eta^c) \leqslant 0$ a.e. for $\eta^c \in [0, h_\theta]$ (losses) and $x(\theta, \eta^c) \geqslant 0$ a.e. for $\eta^c \in [h_\theta, 1]$. (gains)*

2.

$$\begin{aligned}\int_0^1 \mu\left(x(\theta,\eta^c)\right) d\eta^c = A\Big[&\lambda\left(|x(\theta,h_\theta)|^K - \lambda|x(\theta,0)|^K\right) \\ &+ \left(|x(\theta,1)|^K - |x(\theta,h_\theta)|^K\right)\Big] = \\ &\begin{cases} A\left[x(\theta,1)^K - x(\theta,0)^K\right] & \text{if } h_\theta = 0, \\ A\left[x(\theta,1)^K - \lambda\left(-x(\theta,0)\right)^K\right] & \text{if } h_\theta \in (0,1), \\ A\lambda\left[(-x(\theta,1))^K - (-x(\theta,0))^K\right] & \text{if } h_\theta = 1. \end{cases}\end{aligned}$$

where $A \equiv \frac{\gamma k}{(k+1)\left(\frac{\partial x}{\partial \eta^c}\right)}$ and $K \equiv \frac{k+1}{k}$

3. Let

$$\underline{\Theta} = \sup\{\{0\} \cup \{\theta \in [0,1] : h_\theta = 0\}\}$$
$$\overline{\Theta} = \inf\{\{1\} \cup \{\theta \in [0,1] : h_\theta = 1\}\}.$$

Then $h_\theta = 0$ *a.e. for* $\theta \in [0, \underline{\Theta}], h_\theta \in (0,1)$ *a.e. for* $\theta \in [\underline{\Theta}, \overline{\Theta}]$ *and* $h_\theta = 1$ *a.e. for* $\theta \in [\overline{\Theta}, 1]$.

4.

$$\frac{1}{B}\int_0^1 \int_0^1 \mu\left(x(\theta, \eta^c)\right) d\eta^c d\theta =$$
$$\begin{cases} x(1,0)^L - x(0,0)^L - x(1,1)^L + x(0,1)^L & \text{if } \underline{\Theta} = \overline{\Theta} = 1, \\ -\lambda(-x(1,0))^L - x(0,0)^L - x(1,1)^L + x(0,1)^L & \text{if } \underline{\Theta} \in [0,1), \overline{\Theta} = 1, \\ -\lambda(-x(1,0))^L + \lambda\left(-x(0,0)\right)^L - x(1,1)^L + x(0,1)^L & \text{if } \underline{\Theta} = 0, \overline{\Theta} = 1, \\ -\lambda(-x(1,0))^L + \lambda\left(-x(0,0)\right)^L + \lambda\left(-x(1,1)\right)^L + x(0,1)^L & \\ & \text{if } \underline{\Theta} = 0, \overline{\Theta} \in (0,1], \\ -\lambda(-x(1,0))^L + \lambda\left(-x(0,0)\right)^L + \lambda\left(-x(1,1)\right)^L - \lambda\left(-x(\overline{\Theta},1)\right)^L & \\ & \text{if } \underline{\Theta} = \overline{\Theta} = 0. \end{cases}$$

Where $B \equiv \frac{\gamma k^2}{(k+1)(k+2)\left(-\frac{\partial x}{\partial \theta}\frac{\partial x}{\partial \eta^c}\right)}, L \equiv \frac{2k+1}{k}$.

Proof. 1. If $x(\theta, 0) \geqslant 0$, then $x(\theta, \eta^c) > 0$ for all $\eta^c \in (0,1]$ because $\frac{\partial x}{\partial \eta^c} > 0$. Thus, $h_\theta = 0$ by construction and $x(\theta, \eta^c) \geqslant 0$ for all $\eta^c \in [h_\theta, 1] = [0,1]$. Since $[0, h_\theta] = [0,0]$ is measure $0, x(\theta, \eta^c) \leqslant 0$ a.e. on $[0, h_\theta]$.

If $x(\theta, 1) \leqslant 0$, then $x(\theta, \eta^c) < 0$ for all $\eta^c \in [0,1)$ because $\frac{\partial x}{\partial \eta^c} > 0$. Thus, $h_\theta = 1$ by construction and $x(\theta, \eta^c) \leqslant 0$ for all $\eta^c \in [0, h_\theta] = [0,1]$. Since $[h_\theta, 1] = [1,1]$ is measure $0, x(\theta, \eta^c) \geqslant 0$ a.e. on $[h_\theta, 1]$.

Otherwise, $x(\theta,0) < 0$ and $x(\theta,1) > 0$. Since $x(\theta,.)$ is continuous (it's differentiable), then by IVT $0 < h_\theta < 1$ satisfies $x(\theta,h_\theta) = 0$. Since $\frac{\partial x}{\partial \eta^c} > 0$, we get $x(\theta,\eta^c) > 0$ for all $\eta^c \in (h_\theta, 1]$ and $x(\theta,\eta^c) < 0$ for all $\eta^c \in [0,h_\theta)$.

2. Using the above results, we can evaluate the inner gain-loss integral of $\mu(x)$ on $\eta^c \in [0,1]$:

$$\int_0^1 \mu\left(x(\theta,\eta^c)\right) d\eta^c = \gamma\left(\int_{h_\theta}^1 |x|^{1/k} d\eta^c - \lambda \int_0^{h_\theta} |x|^{1/k} d\eta^c\right)$$
$$= \begin{cases} -\gamma\lambda \int_0^1 (-x)^{1/k} d\eta^c & \text{if } h_\theta = 1, \\ \gamma \int_{h_\theta}^1 x^{1/k} d\eta^c - \gamma\lambda \int_0^{h_\theta} (-x)^{1/k} d\eta^c & \text{if } h_\theta \in (0,1), \\ \gamma \int_0^1 x^{1/k} d\eta^c & \text{if } h_\theta = 0. \end{cases}$$

Because $\frac{\partial x}{\partial \eta^c}$ is a constant, we can sum the gains for cases when $h_\theta \in [0,1)$:

$$\int_{h_\theta}^1 x^{1/k} \left(\frac{\frac{\partial x}{\partial \eta^c}}{\frac{\partial x}{\partial \eta^c}}\right) d\eta^c = \frac{1}{\frac{\partial x}{\partial \eta^c}} \int_{h_\theta}^1 x^{1/k} \frac{\partial x}{\partial \eta^c} d\eta^c =$$
$$\frac{k}{(k+1)\frac{\partial x}{\partial \eta^c}} \left[x(\theta,1)^{(1+k)/k} - x(\theta,h_\theta)^{(1+k)/k}\right]$$

Similarly, sum the losses for cases when $h_\theta \in (0,1]$:

$$\int_0^{h_\theta} (-x)^{1/k} \left(\frac{-\frac{\partial x}{\partial \eta^c}}{-\frac{\partial x}{\partial \eta^c}}\right) d\eta^c = \frac{k}{(k+1)\left(-\frac{\partial x}{\partial \eta^c}\right)} \left[(-x(\theta,h_\theta))^{(1+k)/k} - (-x(\theta,0))^{(1+k)/k}\right]$$
$$= \frac{k}{(k+1)\frac{\partial x}{\partial \eta^c}} \left[(-x(\theta,0))^{(1+k)/k} - (-x(\theta,h_\theta))^{(1+k)/k}\right]$$

Combining, this evaluates the integral as stated.

3. Observe that h_θ is weakly increasing in θ because $\frac{\partial x}{\partial \theta} < 0$. It is strictly increasing along $x(\theta,h_\theta) = 0$ with slope $\frac{d\eta^c}{d\theta} \in (0,1)$ by assumption.

Case I: $h_1 = h_0 = 0 \Longrightarrow h_\theta = 0$ for all $\theta \in [0,1]$, and $\underline{\Theta} = \overline{\Theta} = 1$. Then $h_\theta = 0$ on $[0, \underline{\Theta}] = [0,1]$, while $[\underline{\Theta}, \overline{\Theta}] = [1,1]$ and $[\overline{\Theta}, 1] = [1,1]$ are measure 0.

Case II: $h_1 \in (0,1)$, while $h_0 = 0$. Then $0 \leqslant \underline{\Theta} < 1$ and $\overline{\Theta} = 1$. This means $h_\theta = 0$ on $[0, \underline{\Theta}]$, while $h_\theta \in (0,1)$ on $(\underline{\Theta}, \overline{\Theta}] = (\underline{\Theta}, 1]$ and $[\overline{\Theta}, 1] = [1,1]$ is measure 0.

Case III: $h_1 \in (0,1)$, while $h_0 \in (0,1) \Longrightarrow h_\theta \in (0,1)$ for all $\theta \in [0,1]$. Then $\underline{\Theta} = 0$ and $\overline{\Theta} = 1$. This means $h_\theta \in (0,1)$ on $[\underline{\Theta}, \overline{\Theta}] = [0,1]$, while $[0, \underline{\Theta}] = [0,0]$ and $[\overline{\Theta}, 1] = [1,1]$ are measure 0.

Case IV: $h_1 = 1$, while $h_0 \in (0,1)$. Then $\underline{\Theta} = 0$ and $0 < \overline{\Theta} \leqslant 1$. This means $h_\theta \in (0,1)$ on $[\underline{\Theta}, \overline{\Theta}) = [0, \overline{\Theta})$, while $h_\theta = 1$ on $[\overline{\Theta}, 1]$ and $[0, \underline{\Theta}] = [0,0]$ is measure 0.

Case V: $h_1 = h_0 = 1 \Longrightarrow h_\theta = 1$ for all $\theta \in [0,1]$, and $\underline{\Theta} = \overline{\Theta} = 0$. Then $h_\theta = 1$ on $[\overline{\Theta}, 1] = [0,1]$, while $[0, \underline{\Theta}] = [0,0]$ and $[\underline{\Theta}, \overline{\Theta}] = [0,0]$ are measure 0.

4. Using the values of the inner integral from part (2), $I(\theta, h_\theta)$, can evaluate the outer integral as follows:

$$I = \int_0^1 I(\theta, h_\theta) d\theta = \int_0^{\underline{\Theta}} I(\theta, h_\theta) d\theta + \int_{\underline{\Theta}}^{\overline{\Theta}} I(\theta, h_\theta) d\theta + \int_{\overline{\Theta}}^1 I(\theta, h_\theta) d\theta \tag{B.1.1}$$

$h_\theta = 0$ a.e. for all $\theta \in [0, \underline{\Theta}]$:

$$\int_0^{\underline{\Theta}} I(\theta, h_\theta) d\theta = \int_0^{\underline{\Theta}} A \left[x(\theta,1)^K - x(\theta,0)^K \right] d\theta$$

$$= \frac{Ak}{(k+2)\left(-\frac{\partial x}{\partial \theta}\right)} \left(\left[x(\underline{\Theta},0)^L - x(0,0)^L \right] - \left[x(\underline{\Theta},1)^L - x(0,1)^L \right] \right) \tag{B.1.2}$$

$$= \frac{Ak}{(k+2)\left(-\frac{\partial x}{\partial \theta}\right)} \left(\left[|x(\underline{\Theta},0)|^L - |x(0,0)|^L \right] - \left[|x(\underline{\Theta},1)|^L - |x(0,1)|^L \right] \right) \tag{B.1.3}$$

because $\frac{\partial x}{\partial \theta}$ was assumed to be constant and $L \equiv \frac{2k+1}{k}$. Observe that Eq. (B.1.2) expression is valid for $\underline{\Theta} > 0$ and equal to 0 for $\underline{\Theta} = 0$. In the later case, x under the fractional power may be negative, so to keep the same expression valid for both cases (for all $\underline{\Theta}$), we can impose absolute values under the power.

Similarly, $h_\theta \in (0,1)$ a.e. for all $\theta \in [\underline{\Theta}, \overline{\Theta}]$:

$$\int_{\underline{\Theta}}^{\overline{\Theta}} I(\theta, h_\theta) d\theta = \int_{\underline{\Theta}}^{\overline{\Theta}} A\left[x(\theta,1)^K - \lambda\left(-x(\theta,0)\right)^K\right] d\theta$$
$$= \frac{Ak}{(k+2)\left(-\frac{\partial x}{\partial \theta}\right)}\left(\lambda\left[\left(|-x(\underline{\Theta},0)|\right)^L - \left(|-x(\overline{\Theta},0)|\right)^L\right] - \left[|x(\overline{\Theta},1)|^L - |x(\underline{\Theta},1)|^L\right]\right) \tag{B.1.4}$$

Similarly, $h_\theta = 1$ a.e. for all $\theta \in [\overline{\Theta}, 1]$:

$$\int_{\overline{\Theta}}^{1} I(\theta, h_\theta) d\theta = \int_{\overline{\Theta}}^{1} A\lambda\left[\left(|-x(\theta,1)|\right)^K - \left(|-x(\theta,0)|\right)^K\right] d\theta$$
$$= \frac{Ak\lambda}{(k+2)\left(-\frac{\partial x}{\partial \theta}\right)}\left(\left(|-x(1,1)|\right)^L - \left(|-x(\overline{\Theta},1)|\right)^L - \left(|-x(1,0)|\right)^L + \left(|-x(\overline{\Theta},0)|\right)^L\right) \tag{B.1.5}$$

Let $B \equiv \frac{Ak}{k+2} = \frac{\gamma k^2}{(k+1)(k+2)\left(-\frac{\partial x}{\partial \theta}\frac{\partial x}{\partial \eta^c}\right)}$. I in Eq. (B.1.1) by adding equations (B.1.2)-(B.1.5) and simplifying 4 extra terms.

$$\frac{1}{B}\int_0^1\int_0^1 \mu\left(x(\theta,\eta^c)\right) d\eta^c d\theta = |x(\underline{\Theta},0)|^L - |x(0,0)|^L + |x(0,1)|^L + \lambda\left(|-x(\underline{\Theta},0)|\right)^L$$
$$-|x(\overline{\Theta},1)|^L + \lambda\left(|-x(1,1)|\right)^L - \lambda\left(|-x(\overline{\Theta},1)|\right)^L - \lambda(|-x(1,0)|)^L \tag{B.1.6}$$

Eq. (B.1.6) is valid for all $\underline{\Theta} \in [0,1]$ and for all $\overline{\Theta} \in [0,1]$. The sum of 8 terms always reduces to the values of the function x at the four corners of the unit square when $\underline{\Theta}, \overline{\Theta}$ are known. The following two facts are used: corner solutions (such as $\underline{\Theta} = 0$) are substituted directly and interior solutions (such $\overline{\Theta} \in (0,1]$) satisfy $x(\cdot,\cdot) = 0$. The five cases from part (3) are:

$$\frac{1}{B}\int_0^1\int_0^1 \mu\left(x(\theta,\eta^c)\right)d\eta^c d\theta =$$

$$\begin{cases} x(1,0)^L - x(0,0)^L - x(1,1)^L + x(0,1)^L & \text{if } \underline{\Theta} = \overline{\Theta} = 1, \\ -\lambda(-x(1,0))^L - x(0,0)^L - x(1,1)^L + x(0,1)^L & \text{if } \underline{\Theta} \in [0,1), \overline{\Theta} = 1, \\ -\lambda(-x(1,0))^L + \lambda\left(-x(0,0)\right)^L - x(1,1)^L + x(0,1)^L & \text{if } \underline{\Theta} = 0, \overline{\Theta} = 1, \\ -\lambda(-x(1,0))^L + \lambda\left(-x(0,0)\right)^L + \lambda\left(-x(1,1)\right)^L + x(0,1)^L & \text{if } \underline{\Theta} = 0, \overline{\Theta} \in (0,1], \\ -\lambda(-x(1,0))^L + \lambda\left(-x(0,0)\right)^L + \lambda\left(-x(1,1)\right)^L - \lambda\left(-x(\overline{\Theta},1)\right)^L & \\ & \text{if } \underline{\Theta} = \overline{\Theta} = 0. \end{cases}$$

In the first case, x is positive at four corners of the unit square. In the second case, $x(1,0) < 0$ is the only negative corner. In the third case, $x(0,0) < 0$ and $x(1,0) < 0$ (the slope $0 < \frac{d\eta^c}{d\theta} < 1$ along $x = 0$ by assumption, and thus $x(1,0) > 0$ here). In the fourth case, $x(1,1) < 0$ also. In the fifth case, all corners are negative. The corresponding element is scaled by $-\lambda$ for each transition. □

The next corollary highlights the importance of the assumptions that $\frac{dx}{d\theta} \neq 0$ and $\frac{dx}{d\eta^c} \neq 0$.

Corollary B.2. *1. Given differentiable $x(\theta)$ with constant $\frac{dx}{d\theta} = \frac{dx}{d\eta^c} = 0$,*

$$\int_0^1\int_0^1 \mu\left(x(\theta,\eta^c)\right)d\eta^c d\theta = \mu\left(x\right) = \begin{cases} \gamma x^{\frac{1}{k}} & \text{if } x > 0, \\ -\gamma\lambda(-x)^{\frac{1}{k}} & \text{if } x \leqslant 0. \end{cases}$$

2. Constant $\frac{dx}{d\theta} = 0$ *and constant* $\frac{dx}{d\eta^c} > 0$.

$$\int_0^1 \int_0^1 \mu\left(x(\theta, \eta^c)\right) d\eta^c d\theta = \begin{cases} A\left[x(1)^K - x(0)^K\right] & \text{if } h = 0, \\ A\left[x(1)^K - \lambda\left(-x(0)\right)^K\right] & \text{if } h \in (0,1), \\ A\lambda\left[(-x(1))^K - (-x(0))^K\right] & \text{if } h = 1. \end{cases}$$

where $A \equiv \frac{\gamma k}{(k+1)\left(\frac{\partial x}{\partial \eta^c}\right)}$

First, consider the no-shock state. Let $x_p^c(\theta, \eta^c; s, \tilde{\eta}) = c^c - c_p$ be the gain-loss input of the μ. The challenge is to consider the set of points where the argument switches signs because of the λ kink in μ. Observe that c^c is independent of s, while c_p depends on s. This means as s goes up, reference point looks worse and worse relative to the actual consumption and so the gains grow at any point in the parameter space.

Here we verify that $\frac{\partial x_p^c}{\partial \theta}$ satisfies B.1 condition (constant and negative).

$$\frac{\partial x_p^c}{\partial \theta} = 0 - \frac{\partial c_p}{\partial \theta} = -pQ_s \frac{\partial c_s^\theta}{\partial \theta} - (1-p)q\frac{dc^\theta}{d\theta} = -G\left(pQ_s + (1-p)q\right) < 0 \qquad \text{(B.1.7)}$$

Similarly, verify that $\frac{\partial x_p^c}{\partial \eta^c}$ satisfies B.1 condition (constant and positive).

$$\frac{\partial x_p^c}{\partial \eta^c} = G > 0 \qquad \text{(B.1.8)}$$

Thus, $\frac{-\frac{\partial x_p^c}{\partial \theta}}{\frac{\partial x_p^c}{\partial \eta^c}} = pQ_s + (1-p)q \in (0,1)$ also holds as long as both q, Q_s are not equal to 0 or both equal to 1. That is, both the challenger and the incumbent have an interior probability of being chosen in each state.

Furthermore, we calculate $\frac{\partial x_p^c}{\partial \tilde{\eta}}$:

$$\frac{\partial x_p^c}{\partial \tilde{\eta}} = 0 - \frac{\partial c_p}{\partial \tilde{\eta}} = G\left[-p(1-Q_s) - (1-p)(1-q)\right] < 0 \tag{B.1.9}$$

If q and Q_s are not both identically 0, then reference point is not identical to picking incumbent for sure: $\frac{\partial x_p^c}{\partial \theta} < 0$. If q, Q_s are not both equal to 1, then $\frac{\partial x_p^c}{\partial \tilde{\eta}} > 0$ Otherwise, when $q = Q_s = 0$, we have $\frac{\partial x_p^c}{\partial \theta} = 0$ and when $q = Q_s = 1$, then $\frac{\partial x_p^c}{\partial \tilde{\eta}} = 0$ (the reference point doesn't depend on incumbent when challenger is always picked).

Likewise, we calculate $\frac{\partial x_p^c}{\partial s}$ for $p > 0$:

$$\frac{\partial x_p^c}{\partial s} = 0 - \frac{\partial c_p}{\partial s} = -pQ_s\frac{\partial c_s^\theta}{\partial s} - p(1-Q_s)\frac{\partial c_s^i}{\partial s} = p(1-\tau) > 0 \tag{B.1.10}$$

Let

$$h_\theta^c(s, \tilde{\eta}) = \sup\{\{0\} \cup \{\eta^c \in [0,1] : x_p^c(\theta, \eta^c; s, \tilde{\eta}) \leqslant 0\}\} \tag{B.1.11}$$

Given $(s, \tilde{\eta})$, h_θ^c describes the largest η^c until x_p^c hits gains or it's equal to 0 if gains happen for all η^c.

Since

$$\begin{aligned} &x_p^c(0,1;s,1) > x_p^c(1,1;s,1) = \\ &= y(1-\tau) + G - p((y-s)(1-\tau) + G) - (1-p)(y(1-\tau) + G) \\ &= ps(1-\tau) > 0, \end{aligned} \tag{B.1.12}$$

then $\forall \tilde{\eta} \in [0,1], x_p^c(1,1;s,\tilde{\eta}) > 0$ and $x_p^c(1,1;s,\tilde{\eta}) > 0$ because $\frac{\partial x_p^c}{\partial \tilde{\eta}} > 0$.

In words, we know that given a reference draw of the best challenger mixed with the best incumbent and picking the best (actual) challenger, there will be maximal G of public goods in both reference point and in the actual draw. The challenger is better because no shock has happened but the shock has $p > 0$ weight in the reference

point. This means any incumbent's ability less than 1 will lead to even larger gains. If the reference challenger was worse, $\theta = 0$, the gains grow. By continuity of payoffs in η^c and strict inequality (for $p > 0, s > 0$), we can choose $\eta^c < 1$ so that loss is still strictly positive.

This shows:

$$\forall \theta \in [0,1], \forall s > 0, \forall \tilde{\eta} \in [0,1], h_\theta^c(s, \tilde{\eta}) < 1. \tag{B.1.13}$$

The cutoff boundary between gains and losses is strictly to below the top of the unit square where $\eta^c = 1$. As s goes up $\left(\frac{\partial x_p^c}{\partial s} > 0\right)$, h_θ^c moves further down.

As in part (3) of Lemma B.1, the following cutoffs describe either the location of the interior kink in h_θ^c or its corners:

$$\underline{\Theta}^c(s, \tilde{\eta}) = \sup\{\{0\} \cup \{\theta \in [0,1] : h_\theta^c(s, \tilde{\eta}) = 0\}\} \tag{B.1.14}$$

$$\overline{\Theta}^c(s, \tilde{\eta}) = \inf\{\{1\} \cup \{\theta \in [0,1] : h_\theta^c(s, \tilde{\eta}) = 1\}\}. \tag{B.1.15}$$

The restriction that x^c imposes on $h_\theta^c \in [0,1)$ corresponds to cases I-III of Lemma B.1. In particular,

$$\overline{\Theta}^c(s, \tilde{\eta}) = 1. \tag{B.1.16}$$

Meanwhile, there is no restriction on $\underline{\Theta}^c(s, \tilde{\eta}) \in [0,1]$. This reduces to a one-dimensional problem similar to the incumbent's case without shock, except in that case Θ^i is restricted to away from a corner, so the single remaining corner gave a single condition $H^i(s)$ on $\tilde{\eta}$ when that corner was attained. In contrast, the challenger's problem allows for both $\underline{\Theta}^c(s, \tilde{\eta}) = 0$ and $\underline{\Theta}^c(s, \tilde{\eta}) = 1$, which defines two separate boundary cutoffs for $\tilde{\eta}$ as follows:

$$\underline{H}^c(s) = \sup\{\{0\} \cup \{\tilde{\eta} \in [0,1] : \underline{\Theta}^c(s,\tilde{\eta}) = 1\}\} \tag{B.1.17}$$

$$\overline{H}^c(s) = \inf\{\{1\} \cup \{\tilde{\eta} \in [0,1] : \underline{\Theta}^c(s,\tilde{\eta}) = 0\}\} \tag{B.1.18}$$

As $\tilde{\eta}$ grows, x_p^c decreases because $\frac{\partial x_p^c}{\partial \tilde{\eta}} < 0$. Then $\underline{H}^c(s)$ is the cutoff for $\tilde{\eta}$, below which for all $\tilde{\eta} \in [0, \underline{H}^c(s))$, get $\underline{\Theta}^c(s,\tilde{\eta}) = 1$, which says that $x_p^c = 0$ boundary intersects the right edge of the unit square and through its bottom-left corner for $\tilde{\eta} = \underline{H}^c(s)$. Moreover, $h_\theta^c(s,\tilde{\eta}) = 0$ for all $\theta \in [0,1]$. Hence, for all $\eta^c \in [0,1], x_p^c(\theta,\eta^c;s,\tilde{\eta}) \geqslant 0$. In other words, there are gains for all η^c, for all θ (everywhere on the unit square), whenever $\tilde{\eta}$ is below the $\underline{H}^c(s)$ cutoff. This corresponds to case I boundary from part (3) of Lemma B.1.

Similarly, $\overline{H}^c(s)$ is the cutoff for $\tilde{\eta}$, above which for all $\tilde{\eta} \in (\overline{H}^c(s), 1], \underline{\Theta}^c(s,\tilde{\eta}) = 0$, which says that $x_p^c = 0$ boundary intersects the left edge of the unit square and through its bottom-left corner for $\tilde{\eta} = \underline{H}^c(s)$. Moreover, $h_\theta^c(s,\tilde{\eta}) > 0$ for all $\theta \in [0,1]$.[1] Hence, for all $\eta^c \in [h_\theta^c(s,\tilde{\eta}), 1]$, there are gains $x_p^c(\theta,\eta^c;s,\tilde{\eta}) \geqslant 0$. In contrast, for all $\eta^c \in [0, h_\theta^c(s,\tilde{\eta})]$, there are losses $x_p^c(\theta,\eta^c;s,\tilde{\eta}) \leqslant 0$, whenever $\tilde{\eta}$ is above the $\overline{H}^c(s)$ cutoff. Thus, the top two corners of the unit square involve gains and the bottom two corners involve losses. This corresponds to case III boundary from part (3) of Lemma B.1.

Otherwise, $\tilde{\eta} \in [\underline{H}^c(s), \overline{H}^c(s)]$. This corresponds to a loss at bottom-right corner $x_c^p(1,0;s,\tilde{\eta}) < 0$, and a gain for the other three corners. This is case II boundary from part (3) of Lemma B.1.

[1] When $\theta \in (0,1]$, $h_\theta^c(s,\tilde{\eta}) > 0$ because $\underline{\Theta}^c(s,\tilde{\eta}) = 0$ directly. It's also true for $\theta = 0$ by continuity of $x_p^c = 0$ boundary in $\tilde{\eta}$ (x_p^c is decreasing) and in η^c (x_p^c is increasing). Note that $h_0^c(s,\overline{H}^c(s)) = 0$ means $x_p^c(0, h_0^c(s,\overline{H}^c(s)); s, \overline{H}^c(s)) = 0$ implies $h_0^c(s,\tilde{\eta}) > 0$ for all $\tilde{\eta} > \overline{H}^c(s)$.

Let

$$\underline{S}^c = \sup\{s : \underline{H}^c(s) < 1\} \tag{B.1.19}$$

$$\overline{S}^c = \sup\{s : \overline{H}^c(s) < 1\} \tag{B.1.20}$$

which are positive because $\underline{H}^c(0) = \overline{H}^c(0) = 0$. Note that $\underline{H}^c(s) \leqslant overline H^c(s)$ for all s, so it's enough to show that $\underline{S}^c$ is finite as that implies $1 = \underline{H}^c\left(\underline{S}^c\right) \leqslant \overline{H}^c\left(\underline{S}^c\right) \leqslant 1$, and $\underline{S}^c$ is an upper bound for $\overline{S}^c$. $\underline{S}^c$ is finite because $\frac{\partial x_p^c}{\partial s} = p(1-\tau) > 0$. Thus, we get an upper bound[2] for $\underline{S}^c \leqslant \tilde{S}$: for $\tilde{S}$ large enough, $x_p^c(1, 0; \tilde{S}, 1) \geqslant 0$. Then

$$h_1^c(\tilde{S}, 1) = 0 \Longrightarrow \underline{\Theta}(\tilde{S}, 1) = 1 \Longrightarrow \underline{H}^c(\tilde{S}) = 1.$$

This $\underline{S}^c$ is the cutoff when $\underline{H}^c(s) = 1$ for all $s > \underline{S}^c$. Similarly, $\overline{S}^c$ is the cutoff when $\overline{H}^c(s) = 1$ for all $s > \overline{S}^c$.

Figure B.1 shows the $\underline{H}^c(s)$ and $\overline{H}^c(s)$ boundaries in the $s \times \tilde{\eta}$ region of the parameter space for a low probability of shock with $p = \frac{1}{100}$. The illustrated boundary is similar to the "surprise" shock special case. In general, for fixed p, $\underline{H}^c(s)$ and $\overline{H}^c(s)$ split the $s \times \tilde{\eta}$ space into three sections.

The upper-left section, satisfying $s < \overline{S}^c$ with $\overline{H}^c(s) > 0$, is a set of points $(s, \tilde{\eta})$ that satisfy $\overline{H}^c(s) < \tilde{\eta} \leqslant 1$. For these points, $x_p^c(\cdot, \cdot; s, \tilde{\eta}) = 0$ forms a case III boundary in the $\theta \times \eta^c$ space from part (3) of Lemma B.1 with losses at the bottom two corners of the unit square and gains at the top two. Since $\underline{\Theta}^c(s, \tilde{\eta}) = 0$ and for all $\theta \in [0, 1]$, we have $h_\theta^c(s, \tilde{\eta}) \in (0, 1)$ meaning there are strict gains on $(h_\theta^c(s, \tilde{\eta}), 1]$ and strict losses on $[0, h_\theta^c(s, \tilde{\eta}))$.

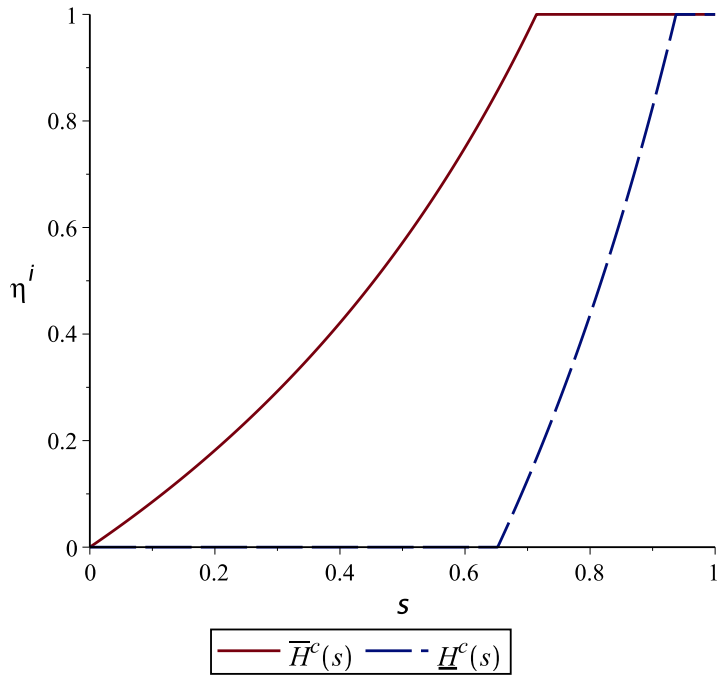
The middle section, is a set of points $(s, \tilde{\eta})$ that satisfy $\underline{H}^c(s) < \tilde{\eta} < \overline{H}^c(s)$. For these points, $x_p^c(\cdot, \cdot; s, \tilde{\eta}) = 0$ forms a case II boundary in the $\theta \times \eta^c$ space from part

[2]This upper bound $\tilde{S}$ requires that for $s > \tilde{S}$, there are gains everywhere on the unit square, in particular at the bottom right corner for the worst challenger against the best incumbent.

(3) of Lemma B.1 with losses at the bottom right corner of the unit square and gains at the other four corners. Since $\underline{\Theta}^c(s,\tilde{\eta}) \in (0,1)$, we locate the kink in the $x_p^c = 0$ boundary to get $h_\theta^c(s,\tilde{\eta}) = 0$ for all $\theta \in [0, \underline{\Theta}^c(s,\tilde{\eta})]$ and $h_\theta^c(s,\tilde{\eta}) \in (0,1)$ for all $\theta \in (\underline{\Theta}^c(s,\tilde{\eta}), 1]$ Meaning there are strict gains on $\eta^c \in (h_\theta^c, 1]$ for all $\theta \in [0,1]$ and strict losses on $\eta^c \in [0, h_\theta^c)$ for all $\theta \in (\underline{\Theta}^c(s,\tilde{\eta}), 1]$, whenever $\tilde{\eta} \in (\underline{H}^c(s), \overline{H}^c(s))$.

The lower-right section, satisfying $\underline{H}^c(s) < 1$, is a set of points $(s,\tilde{\eta})$ that satisfy $0 \leqslant \tilde{\eta} < \underline{H}^c(s)$. For these points, $x_p^c(\cdot,\cdot;s,\tilde{\eta}) = 0$ forms a case I boundary in the $\theta \times \eta^c$ space from part (3) of Lemma B.1 with gains everywhere on the unit square. Since $\underline{\Theta}^c(s,\tilde{\eta}) = 1$ and for all $\theta \in [0,1]$, we have $h_\theta(s,\tilde{\eta}) = 0$ meaning there are strict gains on $(\theta,\eta^c) \in [0,1] \times [0,1]$ for $\tilde{\eta} < \underline{H}^c(s)$.

Figure B.1: Boundaries for $\tilde{\eta}$ when $x_p^c(\theta = 0, \eta^c = 0; s, \tilde{\eta}) = 0, x_p^c(1, 0; s, \tilde{\eta}) = 0, p = \frac{1}{2}$



Picking challenger's lottery subject to reference compound lottery c_p and shock $s \geqslant 0$.

$$U(c^c|c_p; s, \tilde{\eta}) = \Big((1-\tau)(y-s) + \frac{1}{2}G\Big) + \frac{\gamma k^2}{(k+1)(2k+1)\left(-\frac{\partial x_p^c(\theta,\eta^c)}{\partial \theta}\right)\left(\frac{\partial x_p^c(\theta,\eta^c)}{\partial \eta^c}\right)} I_p^c(s) \tag{B.1.21}$$

where the integral $I_p^c(s)$ on the right-hand side depends on the four values of the $x_p^c(\cdot,\cdot)$ at corners of the $(\theta, \eta^c) \in \{0,1\} \times \{0,1\}]$ and the functional form switches between gains and losses at the two boundaries of the parameters: $\overline{H}^c(s), \underline{H}^c(s)$ where $\mu(x_p^c)$ vanishes. Combining the boundary analysis in the Figure B.1 with B.1 get:

$$I_p^c(s, \tilde{\eta}) = \begin{cases} -\lambda(-x_p^c(1,0))^L + \lambda(-x_p^c(0,0))^L - x_p^c(1,1)^L & +x_p^c(0,1)^L \\ & \text{if } \overline{H}^c(s) \leqslant \tilde{\eta} \leqslant 1 \\ -\lambda(-x_p^c(1,0))^L - x_p^c(0,0)^L - x_p^c(1,1)^L & +x_p^c(0,1)^L \\ & \text{if } \underline{H}^c(s) \leqslant \tilde{\eta} \leqslant \overline{H}^c(s) \\ x_p^c(1,0)^L - x_p^c(0,0)^L - x_p^c(1,1)^L & +x_p^c(0,1)^L \\ & \text{if } 0 \leqslant \tilde{\eta} \leqslant \underline{H}^c(s) \end{cases}$$

where $L \equiv \frac{2k+1}{k}$.

Characterising Challenger's Utility, given shock

Second, consider the shock state.

Let $x_p^{cs}(\theta; s, \tilde{\eta}) = c_s^c - c_p$. Here we verify that $\frac{\partial x_p^{cs}}{\partial \theta}$ satisfies B.1 condition (constant and negative).

$$\frac{\partial x_p^{cs}}{\partial \theta} = 0 - \frac{\partial c_p}{\partial \theta} = -pQ_s \frac{\partial c_s^{\theta}}{\partial \theta} - (1-p)q\frac{dc^{\theta}}{d\theta} = -G\left(pQ_s + (1-p)q\right) < 0 \quad \text{(B.1.22)}$$

Similarly, verify that $\frac{\partial x_p^{cs}}{\partial \eta^c}$ satisfies B.1 condition (constant and positive).

$$\frac{\partial x_p^{cs}}{\partial \eta^c} = G > 0 \quad \text{(B.1.23)}$$

Thus, $\frac{-\frac{\partial x_p^{cs}}{\partial \theta}}{\frac{\partial x_p^{cs}}{\partial \eta^c}} = pQ_s + (1-p)q \in (0,1)$ also holds as long as both q, Q_s are not equal to 0 or both equal to 1. That is, both the challenger and the incumbent have an interior probability of being chosen in each state.

Furthermore, we calculate $\frac{\partial x_p^{cs}}{\partial \tilde{\eta}}$:

$$\frac{\partial x_p^{cs}}{\partial \tilde{\eta}} = 0 - \frac{\partial c_p}{\partial \tilde{\eta}} = G\left[-p(1-Q_s) - (1-p)(1-q)\right] < 0 \quad \text{(B.1.24)}$$

If q and Q_s are not both identically 0, then reference point is not identical to picking incumbent for sure: $\frac{\partial x_p^{cs}}{\partial \theta} < 0$. If q, Q_s are not both equal to 1, then $\frac{\partial x_p^{cs}}{\partial \tilde{\eta}} > 0$ Otherwise, when $q = Q_s = 0$, we have $\frac{\partial x_p^{cs}}{\partial \theta} = 0$ and when $q = Q_s = 1$, then $\frac{\partial x_p^{cs}}{\partial \tilde{\eta}} = 0$ (the reference point doesn't depend on incumbent when challenger is always picked).

Likewise, we calculate $\frac{\partial x_p^{cs}}{\partial s}$ for $p > 0$:

$$\frac{\partial x_p^{cs}}{\partial s} = -(1-\tau) - \frac{\partial c_p}{\partial s} = -pQ_s \frac{\partial c_s^{\theta}}{\partial s} - p(1-Q_s)\frac{\partial c_s^{i}}{\partial s} = -(1-p)(1-\tau) < 0 \quad \text{(B.1.25)}$$

Let

$$h_{\theta}^{cs}(s, \tilde{\eta}) = \inf\{\{1\} \cup \{\eta^c \in [0,1] : x_p^{cs}(\theta, \eta^c; s, \tilde{\eta}) \geqslant 0\}\} \quad \text{(B.1.26)}$$

Given $(s, \tilde{\eta})$, h_{θ}^{cs} describes the smallest η^c until x_p^{cs} hits losses or it's equal to 1 if losses happen for all η^c.

Since

$$\begin{aligned} & x_p^{cs}(1,0;s,0) < x_p^{cs}(0,0;s,0) = \\ & = (y-s)(1-\tau) - p((y-s)(1-\tau)) - (1-p)(y(1-\tau)) \\ & = -s(1-\tau) + ps(1-\tau) = -(1-p)(1-\tau) < 0, \end{aligned} \tag{B.1.27}$$

then $\forall \tilde{\eta} \in [0,1], x_p^{cs}(1,0;s,\tilde{\eta}) < 0$ and $x_p^{cs}(0,0;s,\tilde{\eta}) < 0$ because $\frac{\partial x_p^{cs}}{\partial \tilde{\eta}} < 0$.

In words, we know that given a reference draw of the worst challenger mixed with the the worst incumbent and picking the worst (actual) challenger, there will be no public goods in both reference point and in the actual draw. The challenger is a loss because the shock has happened but this shock has a smaller weight $p < 1$ in the reference point and the after-tax income is higher. This means any positive incumbent's ability will lead to even larger losses. If the reference challenger was better, $\theta = 1$, the losses grow. By continuity of payoffs in η^c and strict inequality (for $p > 0, s > 0$), we can choose a slightly better challenger with $\eta^c > 0$ so that there is still a strict loss.

This shows:

$$\forall \theta \in [0,1], \forall s > 0, \forall \tilde{\eta} \in [0,1], h_\theta^{cs}(s,\tilde{\eta}) > 0. \tag{B.1.28}$$

The cutoff boundary between gains and losses is strictly to above the bottom of the unit square where $\eta^c = 0$. As s goes up $\left(\frac{\partial x_p^{cs}}{\partial s} < 0\right)$, h_θ^{cs} moves further up.

As in part (3) of Lemma B.1, the following cutoffs describe either the location of the interior kink in h_θ^{cs} or its corners:

$$\underline{\Theta}^{cs}(s,\tilde{\eta}) = \sup\{\{0\} \cup \{\theta \in [0,1] : h_\theta^{cs}(s,\tilde{\eta}) = 0\}\} \tag{B.1.29}$$

$$\overline{\Theta}^{cs}(s,\tilde{\eta}) = \inf\{\{1\} \cup \{\theta \in [0,1] : h_\theta^{c}(s,\tilde{\eta}) = 1\}\}. \tag{B.1.30}$$

The restriction that x^{cs} imposes on $h_\theta^{cs} \in (0,1]$ corresponds to cases III-V of Lemma B.1. In particular,

$$\underline{\Theta}^{cs}(s,\tilde{\eta}) = 0. \tag{B.1.31}$$

Meanwhile, there is no restriction on $\overline{\Theta}^{cs}(s,\tilde{\eta}) \in [0,1]$. This reduces to a one-dimensional problem similar to the incumbent's case with shock, except in that case Θ^{is} is restricted to away from a corner, so the single remaining corner gave a single condition $H^{is}(s)$ on $\tilde{\eta}$ when that corner was attained. In contrast, the challenger's problem allows for both $\overline{\Theta}^{cs}(s,\tilde{\eta}) = 0$ and $\overline{\Theta}^{cs}(s,\tilde{\eta}) = 1$, as well as intermediate values. The challenger's problem with shock is a mirror of the challenger's problem without the shock in the sense that it covers boundary cases III-V, rather than I-III. With shock, the $x_p^{cs} = 0$ boundary is uniformly bounded away from $\eta^c = 0$, the bottom of the unit square, while without the shock this boundary $x_p^c = 0$ is uniformly bounded away from $\eta^c = 1$, the top of the unit square. We define these two separate boundary cutoffs for $\tilde{\eta}$ as follows:

$$\underline{H}^{cs}(s) = \sup\{\{0\} \cup \{\tilde{\eta} \in [0,1] : \overline{\Theta}^{cs}(s,\tilde{\eta}) = 1\}\} \tag{B.1.32}$$

$$\overline{H}^{cs}(s) = \inf\{\{1\} \cup \{\tilde{\eta} \in [0,1] : \overline{\Theta}^{cs}(s,\tilde{\eta}) = 0\}\} \tag{B.1.33}$$

As $\tilde{\eta}$ grows, x_p^{cs} decreases (losses grow) because $\frac{\partial x_p^{cs}}{\partial \tilde{\eta}} < 0$. Then, $\underline{H}^{cs}(s)$ is the cutoff for $\tilde{\eta}$, below which for all $\tilde{\eta} \in [0, \underline{H}^{cs}(s))$, get $\overline{\Theta}^{cs}(s,\tilde{\eta}) = 1$, which says that $x_p^{cs} = 0$ boundary intersects the right edge of the unit square and through its top-right

corner for $\tilde{\eta} = \underline{H}^{cs}(s)$. Moreover, $h_\theta^{cs}(s,\tilde{\eta}) < 1$ for all $\theta \in [0,1]$.[3] Hence, for all $\eta^c \in [h_\theta^{cs}(s,\tilde{\eta}), 1]$, there are gains $x_p^{cs}(\theta, \eta^c; s, \tilde{\eta}) \geqslant 0$. In contrast, for all $\eta^c \in [0, h_\theta^{cs}(s,\tilde{\eta})]$, there are losses $x_p^{cs}(\theta, \eta^c; s, \tilde{\eta}) \leqslant 0$, whenever $\tilde{\eta}$ is below the $\underline{H}^{cs}(s)$ cutoff. Thus, the top two corners of the unit square involve gains and the bottom two corners involve losses. This corresponds to case III boundary from part (3) of Lemma B.1.

Similarly, $\overline{H}^{cs}(s)$ is the cutoff for $\tilde{\eta}$, above which for all $\tilde{\eta} \in [0, \underline{H}^{cs}(s))$, get $\overline{\Theta}^{cs}(s,\tilde{\eta}) = 0$ which says that $x_p^{cs} = 0$ boundary intersects the left edge of the unit square and through its top-left corner for $\tilde{\eta} = \overline{H}^{cs}(s)$. Moreover, $h_\theta^{cs}(s,\tilde{\eta}) = 1$ for all $\theta \in [0,1]$. Hence, for all $\eta^c \in [0,1], x_p^{cs}(\theta, \eta^c; s, \tilde{\eta}) \leqslant 0$. In other words, there are losses for all η^c, for all θ (everywhere on the unit square), whenever $\tilde{\eta}$ is above the $\overline{H}^{cs}(s)$ cutoff. This corresponds to case V boundary from part (3) of Lemma B.1.

Otherwise, $\tilde{\eta} \in [\underline{H}^{cs}(s), \overline{H}^{cs}(s)]$. This corresponds to a gain at top-left corner $x_c^p(0,1;s,\tilde{\eta}) > 0$, and a loss for the other three corners. This is case IV boundary from part (3) of Lemma B.1.

Let

$$\underline{S}^{cs} = \sup\{s : \underline{H}^{cs}(s) > 0\} \tag{B.1.34}$$

$$\overline{S}^{cs} = \sup\{s : \overline{H}^{cs}(s) > 0\} \tag{B.1.35}$$

which are positive because $\underline{H}^{cs}(0) = \overline{H}^{cs}(0) = 1$. Note that $\underline{H}^{cs}(s) \leqslant overline H^{cs}(s)$ for all s, so it's enough to show that $\overline{S}^{cs}$ is finite as that implies $0 \leqslant \underline{H}^{cs}(\underline{S}^{cs}) \leqslant \overline{H}^{cs}(\overline{S}^{cs}) = 0$, and $\overline{S}^{cs}$ is an upper bound for $\underline{S}^{cs}$. $\overline{S}^{cs}$ is finite because $\frac{\partial x_p^{cs}}{\partial s} = -(1-p)(1-\tau) < 0$. Thus, we get an upper bound[4] for $\overline{S}^{cs} \leqslant \tilde{S}$ as follows: for $\tilde{S}$ large

[3] When $\theta \in [0,1)$, $h_\theta^{cs}(s,\tilde{\eta}) < 1$ because $\overline{\Theta}^{cs}(s,\tilde{\eta}) = 1$ directly. It's also true for $\theta = 1$ by continuity of $x_p^{cs} = 0$ boundary in $\tilde{\eta}$ (x_p^{cs} is decreasing) and in η^c (x_p^{cs} is increasing). Note that $h_0^{cs}(s, \underline{H}^{cs}(s)) = 1$ means $x_p^{cs}(0, h_0^{cs}(s, \underline{H}^{cs}(s)); s, \underline{H}^{cs}(s)) = 0$ implies $h_1^{cs}(s,\tilde{\eta}) < 1$ for all $\tilde{\eta} < \underline{H}^{cs}(s)$.

[4] This upper bound $\tilde{S}$ requires that for $s > \tilde{S}$, there are losses everywhere on the unit square, in particular at the top-left corner for the best (actual) challenger against the worst incumbent mixed with the worst reference challenger.

enough, $x_p^{cs}(0,1;\tilde{S},0) \leqslant 0$. Then

$$h_0^{cs}(\tilde{S},0) = 1 \Longrightarrow \overline{\Theta}(\tilde{S},0) = 0 \Longrightarrow \overline{H}^{cs}(\tilde{S}) = 0.$$

This $\overline{S}^{cs}$ is the cutoff when $\overline{H}^{cs}(s) = 1$ for all $s > \underline{S}^{cs}$. Similarly, $\underline{S}^{cs}$ is the cutoff when $\underline{H}^{cs}(s) = 1$ for all $s > \underline{S}^{cs}$.

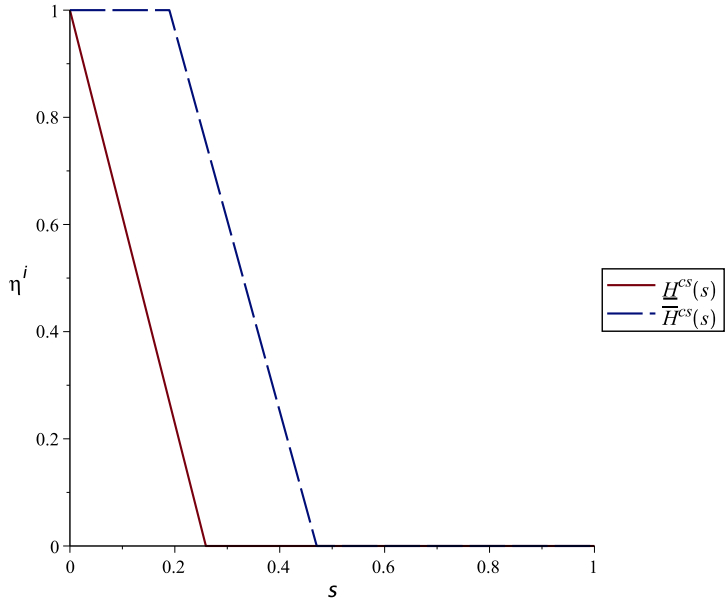
Figure B.2 shows the $\underline{H}^{cs}(s)$ and $\overline{H}^{cs}(s)$ boundaries in the $s \times \tilde{\eta}$ region of the parameter space for a low probability of shock with $p = \frac{1}{100}$. The illustrated boundary is similar to the "surprise" shock special case. In general, for fixed p, $\underline{H}^{cs}(s)$ and $\overline{H}^{cs}(s)$ split the $s \times \tilde{\eta}$ space into three sections.

The lower-left section, satisfying $s < \overline{S}^{cs}$ with $\underline{H}^{cs}(s) > 0$, is a set of points $(s,\tilde{\eta})$ that satisfy $0 \leqslant \tilde{\eta} < \underline{H}^{cs}(s)$. For these points, $x_p^{cs}(\cdot,\cdot;s,\tilde{\eta}) = 0$ forms a case III boundary in the $\theta \times \eta^c$ space from part (3) of Lemma B.1 with losses at the bottom two corners of the unit square and gains at the top two. Since $\overline{\Theta}^{cs}(s,\tilde{\eta}) = 1$ and for all $\theta \in [0,1]$, we have $h_\theta^{cs}(s,\tilde{\eta}) \in (0,1)$ meaning there are strict gains on $(h_\theta^{cs}(s,\tilde{\eta}),1]$ and strict losses on $[0,h_\theta^{cs}(s,\tilde{\eta}))$.

The middle section, is a set of points $(s,\tilde{\eta})$ that satisfy $\underline{H}^{cs}(s) < \tilde{\eta} < \overline{H}^{cs}(s)$. For these points, $x_p^{cs}(\cdot,\cdot;s,\tilde{\eta}) = 0$ forms a case IV boundary in the $\theta \times \eta^c$ space from part (3) of Lemma B.1 with gains at the top-left corner of the unit square and losses at the other four corners. Since $\overline{\Theta}^{cs}(s,\tilde{\eta}) \in (0,1)$, we locate the kink in the $x_p^{cs} = 0$ boundary to get $h_\theta^{cs}(s,\tilde{\eta}) = 1$ for all $\theta \in [\overline{\Theta}^{cs}(s,\tilde{\eta}),1]$ and $h_\theta^{cs}(s,\tilde{\eta}) \in (0,1)$ for all $\theta \in [0,\overline{\Theta}^{cs}(s,\tilde{\eta}))$. This means there are strict losses on $\eta^c \in [0,h_\theta^{cs})$ for all $\theta \in [0,1]$ and strict gains on $\eta^c \in (h_\theta^{cs},1]$ for all $\theta \in [0,\overline{\Theta}^{cs}(s,\tilde{\eta}))$, whenever $\tilde{\eta} \in (\underline{H}^{cs}(s),\overline{H}^{cs}(s))$. The upper-right section, is a set of points $(s,\tilde{\eta})$ that satisfy $\overline{H}^{cs}(s) < \tilde{\eta} \leqslant 1$. For these points, $x_p^{cs}(\cdot,\cdot;s,\tilde{\eta}) = 0$ forms a case V boundary in the $\theta \times \eta^c$ space from part (3) of Lemma B.1 with losses everywhere on the unit square. Since $\overline{\Theta}^{cs}(s,\tilde{\eta}) = 0$ and for all

$\theta \in [0,1]$, we have $h_\theta(s,\tilde{\eta}) = 1$ meaning there are strict losses on $(\theta, \eta^c) \in [0,1] \times [0,1]$ for $\tilde{\eta} > \overline{H}^{cs}(s)$.

Figure B.2: Boundaries for $\tilde{\eta}$ when $x_p^{cs}(\theta = 0, \eta^c = 1; s, \tilde{\eta}) = 0, x_p^{cs}(1,1; s, \tilde{\eta}) = 0, p = \frac{1}{100}$



By Lemma B.1 and Figure B.2, we can evaluate the utility function that corresponds to picking the challenger when the shock happens, subject to reference compound lottery c_p and shock $s \geqslant 0$. Recall that $x_p^{is}(\theta; s, \tilde{\eta}) = c_s^i - c_p$.

$$U(c_s^c | c_p; s) = (1-\tau)y + \frac{1}{2}G + \frac{\gamma k^2}{(k+1)(2k+1)\left(-\frac{\partial x_p^{cs}(\theta,\eta^c)}{\partial \theta}\right)\left(\frac{\partial x_p^{cs}(\theta,\eta^c)}{\partial \eta^c}\right)} I_p^{cs}(s) \tag{B.1.36}$$

where the integral $I_p^{cs}(s)$ on the right-hand side depends on the four values of the $x_p^{cs}(\cdot,\cdot)$ at corners of the $(\theta, \eta^c) \in \{0,1\} \times \{0,1\}]$ and the functional form switches between gains and losses at the two boundaries of the parameters: $\overline{H}^{cs}(s), \underline{H}^{cs}(s)$ where $\mu(x_p^{cs})$ vanishes.

$$I_p^{cs}(s,\tilde{\eta}) = \begin{cases} x^{cs}(0,1)^L - x^{cs}(1,1)^L + \lambda(-x^{cs}(0,0))^L - \lambda(-x^{cs}(1,0))^L \\ \qquad\qquad \text{if } 0 \leqslant \tilde{\eta} \leqslant \underline{H}^{cs}(s). \\ x^{cs}(0,1)^L + \lambda(-x^{cs}(1,1))^L + \lambda(-x^{cs}(0,0))^L - \lambda(-x^{cs}(1,0))^L \\ \qquad\qquad \text{if } \underline{H}^{cs}(s) \leqslant \tilde{\eta} \leqslant \overline{H}^{cs}(s) \\ -\lambda(-x^{cs}(0,1))^L + \lambda(-x^{cs}(1,1))^L + \lambda(-x^{cs}(0,0))^L - \lambda(-x^{cs}(1,0))^L \\ \qquad\qquad \text{if } \overline{H}^{cs}(s) \leqslant \tilde{\eta} \leqslant 1 \end{cases}$$

where $L \equiv \frac{2k+1}{k}$.

Characterising Incumbent's Utility, given no shock

Let reference point be

$$c_p = p(Q_s c_s^{\theta} + (1 - Q_s)c_s^i) + (1-p)(qc^{\theta} + (1-q)c^i)$$

where with probability p the voter will learn there is s shock next period before voting and with probability $1-p$ he learns there will be no shock next period. Q_s is the cutoff for incumbent, when shock happens and q the cutoff when no shock happens. Here θ ability emphasizes a hypothetical draw of the challenger in the reference point, distinct from c_s^c which would be an actual draw (different and independent of θ).

For brevity of notation define, $G = \alpha(\tau y - \overline{r})$ is the coefficient of the ability in making the public good.

The corresponding consumptions are:

1. $c_s^{\theta} = (y-s)(1-\tau) + \theta G$. Reference consumption under challenger type $\theta : s \geqslant 0$
2. $c^{\theta} = y(1-\tau) + \theta G$. Reference consumption under challenger type $\theta : s = 0$

3. $c_s^i = (y-s)(1-\tau) + \tilde{\eta}G$. Consumption under incumbent type $\tilde{\eta} : s \geqslant 0$

4. $c^i = y(1-\tau) + \tilde{\eta}G$. Consumption under incumbent type $\tilde{\eta} : s = 0$

5. $c_s^c = (y-s)(1-\tau) + \eta^c G$. Consumption under challenger type $\eta^c : s \geqslant 0$

6. $c^c = y(1-\tau) + \eta^c G$. Consumption under challenger type $\eta^c : s = 0$

Recall that

$$\mu(x) = \begin{cases} \gamma x^{\frac{1}{k}} & \text{if } x > 0, \\ -\gamma\lambda(-x)^{\frac{1}{k}} & \text{if } x \leqslant 0. \end{cases}$$

Lemma B.3 (Expanding $\int_0^1 \mu(.)d\theta$)**.** *Given differentiable $x(\theta)$ with constant $\frac{dx}{d\theta} < 0$, let*

$$\Theta = \sup\{\{0\} \cup \{\theta \in [0,1] : x(\theta) \geqslant 0\}\}.$$

Then

1. $x(\theta) \geqslant 0$ a.e. for $\theta \in [0, \Theta]$ (gains) and $x(\theta) \leqslant 0$ a.e. for $\theta \in [\Theta, 1]$. (losses) 2.

$$\int_0^1 \mu\left(x(\theta)\right) d\theta = C\left[-\left(|x(\Theta)|^K - |x(0)|^K\right) - \lambda\left(|x(1)|^K - |x(\Theta)|^K\right)\right]) =$$

$$\begin{cases} C\lambda\left[(-x(0))^K - (-x(1))^K\right] & \text{if } \Theta = 0, \\ C\left[x(0)^K - \lambda\left(-x(1)\right)^K\right] & \text{if } \Theta \in (0,1), \\ C\left[x(0)^K - x(1)^K\right] & \text{if } \Theta = 1. \end{cases}$$

where $C \equiv \frac{\gamma k}{(k+1)\left(-\frac{dx}{d\theta}\right)}$ and $K \equiv \frac{k+1}{k}$

3. Let $b = |\frac{dx}{d\theta}|$, then $x(\theta) = a - b\theta$ and $\Theta = \begin{cases} 0 & \text{if } a \leqslant 0, \\ \frac{a}{b} & \text{if } 0 < a < b, \\ 1 & \text{if } a \geqslant b, \end{cases}$

Proof. 1. If $x(0) \leqslant 0$, then $x(\theta) < 0$ for all $\theta \in (0, 1]$ because $\frac{dx}{d\theta} < 0$. Thus, $\Theta = 0$ by construction and $x(\theta) \leqslant 0$ for all $\theta \in [\Theta, 1]$. Since $[0, \Theta] = [0, 0]$ is measure $0, x(\theta) \geqslant 0$ a.e. on $[0, \Theta]$.

If $x(1) \geqslant 0$, then $x(\theta) > 0$ for all $\theta \in [0, 1)$ because $\frac{dx}{d\theta} < 0$. Thus, $\Theta = 1$ by construction and $x(\theta) \geqslant 0$ for all $\theta \in [0, \Theta]$. Since $[\Theta, 1] = [1, 1]$ is measure $0, x(\theta) \leqslant 0$ a.e. on $[\Theta, 1]$.

Otherwise, $x(0) > 0$ and $x(1) < 0$. Since x is continuous (it's differentiable), then by IVT $0 < \Theta < 1$ satisfies $x(\Theta) = 0$. Since $\frac{dx}{d\theta} < 0$, we get $x(\theta) > 0$ for all $\theta \in [0, \Theta)$ and $x(\theta) < 0$ for all $\theta \in (\Theta, 1]$.

2. Using the above results, we can evaluate the gain-loss integral of $\mu(x)$ on $\theta \in [0, 1]$:

$$\int_0^1 \mu\left(x(\theta)\right) d\theta = \gamma \left(\int_0^{\Theta} |x|^{1/k} d\theta - \lambda \int_{\Theta}^1 |x|^{1/k} d\theta \right)$$

$$= \begin{cases} -\gamma\lambda \int_0^1 (-x)^{1/k} d\theta & \text{if } \Theta = 0, \\ \gamma \int_0^{\Theta} x^{1/k} d\theta - \gamma\lambda \int_{\Theta}^1 (-x)^{1/k} d\theta & \text{if } \Theta \in (0, 1), \\ \gamma \int_0^1 x^{1/k} d\theta & \text{if } \Theta = 1. \end{cases}$$

Because $\frac{dx}{d\theta}$ is a constant, we can sum the gains:

$$\int_0^{\Theta} x^{1/k} \frac{\frac{dx}{d\theta}}{\frac{dx}{d\theta}} d\theta = \frac{1}{\frac{dx}{d\theta}} \int_0^{\Theta} x^{1/k} \frac{dx}{d\theta} d\theta = \frac{-k}{(k+1)\left(-\frac{dx}{d\theta}\right)} \left[x(\Theta)^{(1+k)/k} - x(0)^{(1+k)/k} \right]$$

Similarly, sum the losses:

$$\int_{\Theta}^1 (-x)^{1/k} \left(\frac{-\frac{dx}{d\theta}}{-\frac{dx}{d\theta}} \right) d\theta = \frac{k}{(k+1)\left(-\frac{dx}{d\theta}\right)} \left[(-x(1))^{(1+k)/k} - (-x(\Theta))^{(1+k)/k} \right]$$

Combining, this evaluates the integral as stated.

3. Let $a = x(0)$. Then

$$x(\theta) = \int \frac{dx}{d\theta} d\theta = \frac{dx}{d\theta}\theta + a$$

□

The Lemma B.3 shows that the value of the integral depends only on the value of the function at the end-points, where the functional form changes based on whether Θ (the split between the gains and losses) is at a corner $\Theta \in \{0, 1\}$ or interior $\Theta \in (0, 1)$. It does not depend on the specific value of Θ explicitly. However, because x was assumed to be a linear function of θ, the situations when Θ is higher are precisely when $x(0)$ is higher.

The next corollary highlights the importance of the assumption that $\frac{dx}{d\theta} \neq 0$.

Corollary B.4. *Given differentiable $x(\theta)$ with constant $\frac{dx}{d\theta} = 0$,*

$$\int_0^1 \mu\left(x(\theta)\right) d\theta = \mu\left(x\right) = \begin{cases} \gamma x^{\frac{1}{k}} & \text{if } x > 0, \\ -\gamma\lambda(-x)^{\frac{1}{k}} & \text{if } x \leqslant 0. \end{cases}$$

First, consider the no-shock state. Let $x_p^i(\theta; s, \tilde{\eta}) = c^i - c_p$ be the gain-loss input of the μ. Here we calculate $\frac{\partial x_p^i}{\partial \theta}$:

$$\frac{\partial x_p^i}{\partial \theta} = 0 - \frac{\partial c_p}{\partial \theta} = -pQ_s \frac{\partial c_s^\theta}{\partial \theta} - (1-p)q\frac{dc^\theta}{d\theta} = -G\left(pQ_s + (1-p)q\right) < 0 \qquad \text{(B.1.37)}$$

Furthermore, we calculate $\frac{\partial x_p^i}{\partial \tilde{\eta}}$:

$$\frac{\partial x_p^i}{\partial \tilde{\eta}} = \frac{\partial c^i}{\partial \tilde{\eta}} - \frac{\partial c_p}{\partial \tilde{\eta}} = G\left[1 - p(1-Q_s) - (1-p)(1-q)\right] > 0 \qquad \text{(B.1.38)}$$

If q and Q_s are not both identically 0, then reference point is not identical to picking incumbent for sure: $\frac{\partial x_p^i}{\partial \theta} < 0$ and $\frac{\partial x_p^i}{\partial \tilde{\eta}} > 0$ Otherwise, when $q = Q_s = 0$, we have $\frac{\partial x_p^i}{\partial \tilde{\eta}} = \frac{\partial x_p^i}{\partial \theta} = 0$.

Likewise, we calculate $\frac{\partial x_p^i}{\partial s}$ for $p > 0$:

$$\frac{\partial x_p^i}{\partial s} = 0 - \frac{\partial c_p}{\partial s} = -pQ_s \frac{\partial c_s^\theta}{\partial s} - p(1 - Q_s)\frac{\partial c_s^i}{\partial s} = p(1 - \tau) > 0 \tag{B.1.39}$$

The challenge is to consider the set of points where the argument switches signs because of the λ kink in μ. Observe that c^i is independent of s, while c_p depends on s. This means as s goes up, reference point looks worse and worse relative to the actual consumption and so the gains grow at any point in the parameter space.

Let

$$\Theta^i(s, \tilde{\eta}) = \sup\{\{0\} \cup \{\theta \in [0, 1] : x_p^i(\theta; s, \tilde{\eta}) \geqslant 0\}\} \tag{B.1.40}$$

Given $(s, \tilde{\eta})$, Θ^i describes the largest θ until x_p^i hits losses or it's equal to 1 if gains happen for all θ.

Since

$$x_p^i(0; s, 0) = y(1 - \tau) - p((y - s)(1 - \tau)) - (1 - p)(y(1 - \tau)) = ps(1 - \tau) > 0,$$

then $\forall \tilde{\eta}, x_p^i(0; s, \tilde{\eta}) > 0$ because $\frac{\partial x_p^i}{\partial \tilde{\eta}} > 0$.

In words, we know that given a reference draw of the worst challenger and picking the worst incumbent, there will be no public goods in both reference point and in the actual draw. The incumbent is still better because no shock has happened but the shock has a positive weight in the reference point and the after-tax income is higher. This means any positive incumbent's ability will lead to even larger gains. By continuity of payoffs in θ and strict inequality, we can choose $\theta > 0$ so that gain is still strictly positive.

This shows:

$$\forall s > 0, \forall \tilde{\eta} \in [0,1] : \Theta^i(s,\tilde{\eta}) > 0. \tag{B.1.41}$$

The cutoff boundary between gains and losses is strictly to the right of $\theta = 0$. As s goes up $\left(\frac{\partial x_p^i}{\partial s} > 0\right)$, Θ^i moves further right.

$$H^i(s) = \inf\{\{1\} \cup \{\tilde{\eta} \in [0,1] : \Theta^i(s,\tilde{\eta}) = 1\}\} \tag{B.1.42}$$

This is the cutoff for $\tilde{\eta}$, above which gains are achieved for any $0 \leqslant \theta \leqslant 1$ because $\frac{\partial x_p^i}{\partial \tilde{\eta}} > 0$.

Let

$$S^i = \sup\{s : H^i(s) > 0\} \tag{B.1.43}$$

which is positive because $H^i(0) = 1$. S^i is finite because $\frac{\partial x_p^i}{\partial s} = p(1-\tau) > 0$. Thus, we get an upper bound: for $\overline{S}$ large enough, $x_p^i(1;\overline{S},0) \geqslant 0$. Then

$$\Theta^i(\overline{S},0) = 1 \Longrightarrow H^i(\overline{S}) = 0.$$

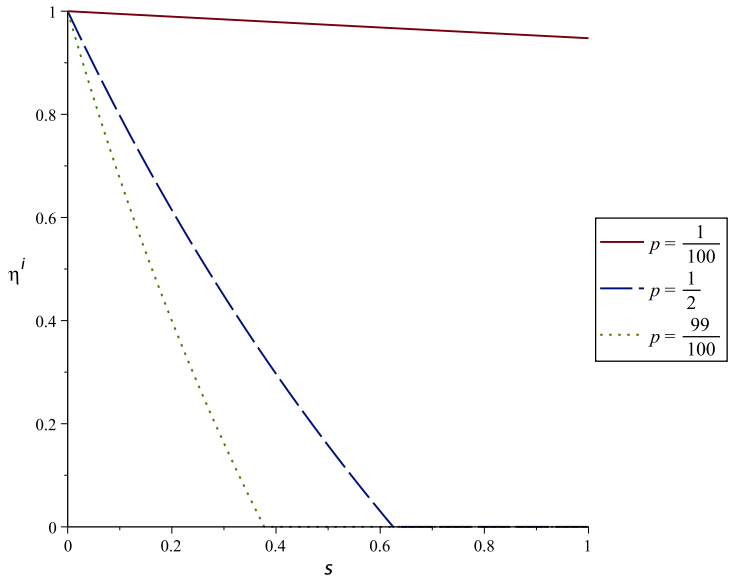
This S^i is the cutoff when $H^i(s) = 0$ for all $s > S^i$.

Figure B.3 shows the $H^i(s)$ boundary in the $s \times \tilde{\eta}$ region of the parameter space for three values of $p \in \{\frac{1}{100}, \frac{1}{2}, \frac{99}{100}\}$. For fixed $p, H^i(s)$ splits $s \times \tilde{\eta}$ space into two sections. The lower-left section, satisfying $s < S^i$ and $H^i(s) > 0$, is a set of points $(s,\tilde{\eta})$ that satisfy $0 \leqslant \tilde{\eta} < H^i(s)$. For these points, $\Theta^i(s,\tilde{\eta}) < 1$, meaning there are strict gains on $[0, \Theta^i(s,\tilde{\eta}))$[5] and strict losses on $(\Theta^i(s,\tilde{\eta}), 1]$.

The upper-right section, satisfying $H^i(s) < 1$, is a set of points $(s,\tilde{\eta})$ that satisfy $H^i(s) < \tilde{\eta} \leqslant 1$. For these points, $\Theta^i(s,\tilde{\eta}) = 1$, and $x_p^i(1;s,\tilde{\eta}) > 0$, meaning there are strict gains on $[0,1]$.

[5]Recall that $\Theta^i(s,\tilde{\eta}) > 0$ by Eq. (B.1.41).

Figure B.3: Boundary $H^i(s)$ for $\tilde{\eta}$ when $x_p^i(\theta = 1; s, \tilde{\eta}) = 0$ for $p \in \{\frac{1}{100}, \frac{1}{2}, \frac{99}{100}\}$



Along the boundary, $\Theta(s, H^i(s)) = 1$ and $x_p^i(1; s, H^i(s)) = 0$, meaning there are strict gains on $\theta \in [0, 1)$.

Recall that $x_p^i(\theta; s, \tilde{\eta}) = c^i - c_p$. We can evaluate the utility function that corresponds to picking incumbent's degenerate lottery when no shock happens, subject to reference compound lottery c_p and shock $s \geqslant 0$.

$$U(c^i|c_p; s, \tilde{\eta}) = \Big((1-\tau)y + \tilde{\eta}G\Big) + \gamma \frac{k}{(1+k)(-\frac{\partial x_p^i}{\partial \theta})} I_p^i(s, \tilde{\eta}) \tag{B.1.44}$$

By Lemma B.3 and Figure B.3, $0 \leqslant \tilde{\eta} \leqslant H^i(s) \implies \Theta^i \in (0, 1)$. Similarly, $H^i(s) \leqslant \tilde{\eta} \leqslant 1 \implies \Theta^i = 1$. Thus,

$$I_p^i(s, \tilde{\eta}) = \begin{cases} x_p^i(0; s, \tilde{\eta})^{\frac{1+k}{k}} - \lambda(-x_p^i(1; s, \tilde{\eta}))^{\frac{1+k}{k}} & \text{if } 0 \leqslant \tilde{\eta} \leqslant H^i(s) \\ x_p^i(0; s, \tilde{\eta})^{\frac{1+k}{k}} - x_p^i(1; s, \tilde{\eta})^{\frac{1+k}{k}} & H^i(s) \leqslant \tilde{\eta} \leqslant 1 \end{cases}$$

Characterizing Incumbent's Utility, given shock

Second, consider the shock state.

Let $x_p^{is}(\theta; s, \tilde{\eta}) = c^{is} - c_p$ be the gain-loss input of the μ. Here we calculate $\frac{\partial x_p^{is}}{\partial \theta}$:

$$\frac{\partial x_p^{is}}{\partial \theta} = 0 - \frac{\partial c_p}{\partial \theta} = -G\left(pQ_s + (1-p)q\right) = \frac{\partial x_p^{i}}{\partial \theta} < 0 \tag{B.1.45}$$

Furthermore, we calculate $\frac{\partial x_p^{is}}{\partial \tilde{\eta}}$:

$$\frac{\partial x_p^{is}}{\partial \tilde{\eta}} = \frac{\partial c^{is}}{\partial \tilde{\eta}} - \frac{\partial c_p}{\partial \tilde{\eta}} = G\left[1 - p(1-Q_s) - (1-p)(1-q)\right] = \frac{\partial x_p^{i}}{\partial \tilde{\eta}} > 0 \tag{B.1.46}$$

If q and Q_s are not both identically 0, then reference point is not identical to picking incumbent for sure: $\frac{\partial x_p^{is}}{\partial \theta} < 0$ and $\frac{\partial x_p^{is}}{\partial \tilde{\eta}} > 0$ Otherwise, when $q = Q_s = 0$, we have $\frac{\partial x_p^{is}}{\partial \tilde{\eta}} = \frac{\partial x_p^{is}}{\partial \theta} = 0$.

While the derivatives with respect to $\theta, \tilde{\eta}$ were equal for x_p^i and x_p^{is}, consider $\frac{\partial x_p^{is}}{\partial s}$ for $p < 1$:

$$\frac{\partial x_p^{is}}{\partial s} = -(1-\tau) - \frac{\partial c_p}{\partial s} = -(1-\tau) + p(1-\tau) = -(1-p)(1-\tau) < 0. \tag{B.1.47}$$

Let

$$\Theta^{is}(s, \tilde{\eta}) = \inf\{\{1\} \cup \{\theta \in [0,1] : x_p^{is}(\theta; s, \tilde{\eta}) \leqslant 0\}\} \tag{B.1.48}$$

Given $(s, \tilde{\eta})$, Θ^{is} describes the smallest θ until x_p^{is} hits gains or it's equal to 1 if losses happen for all θ.

Since

$$\begin{aligned} x_p^{is}(1; s, 1) &= (y-s)(1-\tau) + G - p((y-s)(1-\tau) + G) - (1-p)(y(1-\tau) + G) \\ &= -s(1-\tau) + ps(1-\tau) = -(1-p)s(1-\tau) < 0, \end{aligned}$$

then $\forall \tilde{\eta} \in [0,1], x_p^{is}(1; s, \tilde{\eta}) < 0$ because $\frac{\partial x_p^{is}}{\partial \tilde{\eta}} > 0$.

In words, we know that given a reference draw of the best challenger and picking the best incumbent, there will be maximal G of public goods in both reference point and in the actual draw. The incumbent is worse because the shock has happened but the shock has $p < 1$ weight in the reference point. This means any incumbent's ability less than 1 will lead to even larger losses. By continuity of payoffs in θ and strict inequality, we can choose $\theta < 1$ so that loss is still strictly positive.

This shows:

$$\forall s > 0, \forall \tilde{\eta} \in [0,1] : \Theta^{is}(s, \tilde{\eta}) < 1. \tag{B.1.49}$$

The cutoff boundary between gains and losses is strictly to the left of $\theta = 1$. As s goes up $\left(\frac{\partial x_p^{is}}{\partial s} < 0\right)$, Θ^{is} moves further left.

Let

$$H^{is}(s) = \sup\{\{0\} \cup \{\tilde{\eta} \in [0,1] : \Theta^{is}(s, \tilde{\eta}) = 0\}\} \tag{B.1.50}$$

This is the cutoff for $\tilde{\eta}$, below which losses are achieved for any $0 \leqslant \theta \leqslant 1$ because $\frac{\partial x_p^{is}}{\partial \tilde{\eta}} > 0$.

Let

$$S^{is} = \sup\{s : H^{is}(s) < 1\} \tag{B.1.51}$$

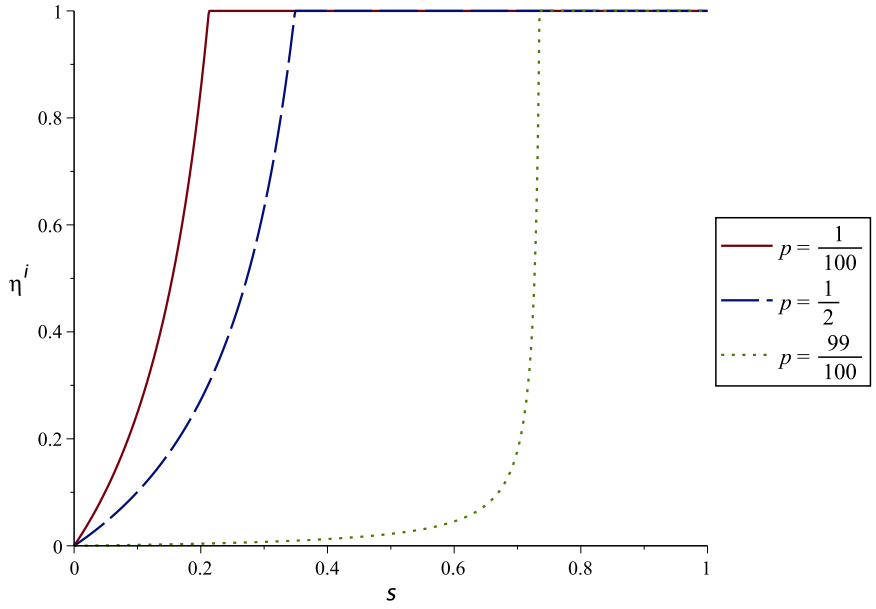
which is positive because $H^{is}(0) = 0$. S^{is} is finite because $\frac{\partial x_p^{is}}{\partial s} = -(1-p)(1-\tau) < 0$. Thus, we get an upper bound: for $\overline{S}$ large enough, $x_p^{is}(0; \overline{S}, 0) \leqslant 0$. Then

$$\Theta^{is}(\overline{S}, 1) = 0 \Longrightarrow H^{is}(\overline{S}) = 1.$$

This S^{is} is the cutoff when $H^{is}(s) = 1$ for $s > S^i$.

Figure B.4 shows the $H^{is}(s)$ boundary in the $s \times \tilde{\eta}$ region of the parameter space for three values of $p \in \{\frac{1}{100}, \frac{1}{2}, \frac{99}{100}\}$. For fixed $p, H^{is}(s)$ splits $s \times \tilde{\eta}$ space into two sections. The upper-left section, satisfying $s < S^{is}$ and $H^{is}(s) < 1$, is a set of points

Figure B.4: Boundary $H^{is}(s)$ for $\tilde{\eta}$ when $x_p^{is}(\theta = 0; s, \tilde{\eta}) = 0$ for $p \in \{\frac{1}{100}, \frac{1}{2}, \frac{99}{100}\}$



$(s, \tilde{\eta})$ that satisfy $H^{is}(s) < \tilde{\eta} \leqslant 1$. For these points, $\Theta^{is}(s, \tilde{\eta}) > 0$, meaning there are strict gains on $[0, \Theta^{is}(s, \tilde{\eta}))$[6] and strict losses on $(\Theta^{is}(s, \tilde{\eta}), 1]$.

The lower-right section, satisfying $H^{is}(s) > 0$, is a set of points $(s, \tilde{\eta})$ that satisfy $0 \leqslant \tilde{\eta} < H^{is}(s)$. For these points, $\Theta^{is}(s, \tilde{\eta}) = 0$, and $x_p^{is}(0; s, \tilde{\eta}) < 0$, meaning there are strict losses on $[0, 1]$.

Along the boundary, $\Theta(s, H^{is}(s)) = 0$ and $x_p^{is}(0; s, H^{is}(s)) = 0$, meaning there are strict losses on $\theta \in (0, 1]$.

Recall that $x_p^{is}(\theta; s, \tilde{\eta}) = c_s^i - c_p$. We can evaluate the utility function that corresponds to picking incumbent's degenerate lottery when shock happens, subject to reference compound lottery c_p and shock $s \geqslant 0$:

$$U(c_s^i | c_p; s, \tilde{\eta}) = \Big((1-\tau)(y-s) + \tilde{\eta}G\Big) + \gamma \frac{k}{(1+k)(-\frac{\partial x_p^{is}}{\partial \theta})} I_p^{is}(s, \tilde{\eta}) \tag{B.1.52}$$

[6]Recall that $\Theta^{is}(s, \tilde{\eta}) < 1$ by Eq. (B.1.49).

By Lemma B.3 and Figure B.4, $0 \leqslant \tilde{\eta} \leqslant H^{is}(s) \Longrightarrow \Theta^{is} = 0$. Similarly, $H^{is}(s) \leqslant \tilde{\eta} \leqslant 1 \Longrightarrow \Theta^{is} = (0,1)$. Thus,

$$I_p^{is}(s,\tilde{\eta}) = \begin{cases} x_p^{is}(0;s,\tilde{\eta})^{\frac{1+k}{k}} - \lambda(-x_p^{is}(1;s,\tilde{\eta}))^{\frac{1+k}{k}} & \text{if } H^{is}(s) \leqslant \tilde{\eta} \leqslant 1 \\ \lambda(-x_p^{is}(0;s,\tilde{\eta}))^{\frac{1+k}{k}} - \lambda(-x_p^{i}(1;s,\tilde{\eta}))^{\frac{1+k}{k}} & 0 \leqslant \tilde{\eta} \leqslant H^{is}(s) \end{cases}$$

B.1.2 Rational Expectation q with no shock

Let reference point be

$$c_q = qc^{\theta} + (1-q)c^{i}$$

where with probability 1 the voters learns there will be no shock next period. q is the cutoff for incumbent, when no shock happens. Here θ ability emphasizes a hypothetical draw of the challenger in the reference point, distinct from c^c which would be an actual draw (different and independent of θ).

For brevity of notation define, $G = \alpha(\tau y - \overline{r})$ is the coefficient of the ability in making the public good.

The corresponding consumptions are:

1. $c^{\theta} = y(1-\tau) + \theta G$. Reference consumption under challenger type $\theta : s = 0$
2. $c^{i} = y(1-\tau) + \tilde{\eta} G$. Consumption under incumbent type $\tilde{\eta} : s = 0$
3. $c^{c} = y(1-\tau) + \eta^c G$. Consumption under challenger type $\eta^c : s = 0$

Characterizing Challenger's Utility under q

Recall that

$$x_q^c(\theta, \eta^c; \tilde{\eta}) = c^c - c_q = \big(\eta^c - q\theta - (1-q)\tilde{\eta}\big)G \tag{B.1.53}$$

$$
\begin{aligned}
h_\theta^c(0,\tilde{\eta}) &= \sup\{\{0\} \cup \{\eta^c \in [0,1] : x_q^c(\theta,\eta^c;\tilde{\eta}) \leqslant 0\}\} \\
&= \sup\{\{0\} \cup \{\eta^c \in [0,1] : (\eta^c - q\theta - (1-q)\tilde{\eta})G \leqslant 0\}\} \\
&= \sup\{\{0\} \cup \{\eta^c \in [0,1] : \eta^c \leqslant q\theta + (1-q)\tilde{\eta}\}\} \\
&= q\theta + (1-q)\tilde{\eta} \qquad \text{(B.1.54)}
\end{aligned}
$$

$$
\begin{aligned}
\underline{\Theta}^c(0,\tilde{\eta}) &= \sup\{\{0\} \cup \{\theta \in [0,1] : h_\theta^c(0,\tilde{\eta}) = 0\}\} \\
&= \sup\{\{0\} \cup \{\theta \in [0,1] : q\theta + (1-q)\tilde{\eta} = 0\}\} = 0 \qquad \text{(B.1.55)}
\end{aligned}
$$

because $\forall\theta > 0, q\theta + (1-q)\tilde{\eta} > 0$ for all $\tilde{\eta} \in [0,1]$.

$$
\begin{aligned}
\underline{H}^c(0) &= \sup\{\{0\} \cup \{\tilde{\eta} \in [0,1] : \underline{\Theta}^c(0,\tilde{\eta}) = 1\}\} \\
&= \sup\{\{0\} \cup \{\tilde{\eta} \in [0,1] : 0 = 1\}\} = \sup\{\{0\} \cup \varnothing\} = 0 \qquad \text{(B.1.56)}
\end{aligned}
$$

$$
\begin{aligned}
\overline{H}^c(0) &= \inf\{\{1\} \cup \{\tilde{\eta} \in [0,1] : \underline{\Theta}^c(0,\tilde{\eta}) = 0\}\} \\
&= \inf\{\{1\} \cup \{\tilde{\eta} \in [0,1] : 0 = 0\}\} = \inf\{\{1\} \cup [0,1]\} = 0 \qquad \text{(B.1.57)}
\end{aligned}
$$

Thus, $\underline{H}^c(0) = \overline{H}^c(0) = 0$, which is consitent with Figure B.2.

For $q \in (0,1]$,

$$
\frac{\partial x_q^c(\theta,\eta^c;\tilde{\eta})}{\partial\theta} = -qG < 0 \qquad \frac{\partial x_q^c(\theta,\eta^c;\tilde{\eta})}{\partial\eta^c} = G > 0 \qquad \text{(B.1.58)}
$$

$$x_q^c(0,1) = (1-(1-q)\tilde{\eta})G \quad \text{(B.1.59)} \qquad x_q^c(1,1) = (1-q)(1-\tilde{\eta})G \quad \text{(B.1.61)}$$

$$-x_q^c(0,0) = (1-q)\tilde{\eta}G \quad \text{(B.1.60)} \qquad -x_q^c(1,0) = (q+(1-q)\tilde{\eta})G \quad \text{(B.1.62)}$$

$$U(c^c|c_q;0,\tilde{\eta}) = (1-\tau)y + \frac{1}{2}G + \frac{\gamma k^2\big(-\lambda(-x_q^c(1,0))^L + \lambda(-x_q^c(0,0))^L - x_q^c(1,1)^L + x_q^c(0,1)^L\big)}{(k+1)(2k+1)G^2q} \tag{B.1.63}$$

where $L = \frac{2k+1}{k}$.

Otherwise, if $q = 0$ then $\frac{\partial x_q^c(\theta;\tilde{\eta})}{\partial\theta} = 0$ and Corollary B.2 applies where $h \in (0,1)$:

$$\begin{aligned} U(c^c|c_0;0,\tilde{\eta}) &= (1-\tau)y + \frac{1}{2}G + \frac{\gamma k}{(k+1)\left(\frac{\partial x}{\partial \eta^c}\right)}\left[x(1)^{\frac{k+1}{k}} - \lambda\left(-x(0)\right)^{\frac{k+1}{k}}\right] \\ &= (1-\tau)y + \frac{1}{2}G + \frac{\gamma k\big(x(1)^{\frac{k+1}{k}} - \lambda\left(-x(0)\right)^{\frac{k+1}{k}}\big)}{(k+1)G} \end{aligned} \tag{B.1.64}$$

where $x(\eta^c) = (\eta^c - \tilde{\eta})G$

Lemma B.5. *Utility of challenger is decreasing in $\tilde{\eta}$ for all reference points $q \in [0,1]$*

Proof. Note that for all $q \in [0,1]$, the derivative of challenger's gain-loss w.r.t. to incumbent's ability, $\tilde{\eta}$, is independent of challenger's ability, η^c:

$$\frac{\partial x_q^c}{\partial \tilde{\eta}} = -(1-q)G \leqslant 0 \tag{B.1.65}$$

If $q = 0$, differentiating $U(c^c|c_0;0,\tilde{\eta})$ w.r.t. to $\tilde{\eta}$ gives:

$$\begin{aligned} \frac{\partial U(c^c|c_0;0,\tilde{\eta})}{\partial\tilde{\eta}} &= 0 + \frac{-(1-q)G\gamma\left(x(1)^{1/k} + \lambda(-x(0))^{1/k}\right)}{G} \\ &= -\big(\left[(1-\tilde{\eta})G\right]^{1/k} + \lambda\left[\tilde{\eta}G\right]^{1/k}\big) < 0 \end{aligned} \tag{B.1.66}$$

If $q \in (0,1]$, differentiating $U(c^c|c_q;0,\tilde{\eta})$ w.r.t. to $\tilde{\eta}$ gives:

$$\frac{\partial U(c^c|c_q;0,\tilde{\eta})}{\partial \tilde{\eta}} = \frac{-(1-q)\gamma k\big(\lambda(-x_q^c(1,0))^K - \lambda(-x_q^c(0,0))^K - x_q^c(1,1)^K + x_q^c(0,1)^K\big)}{(k+1)Gq} \tag{B.1.67}$$

Note that $-x_q^c(1,0) > -x_q^c(0,0)$ and $x_q^c(0,1) > x_q^c(1,1)$ because x_q^c is decreasing in its first argument, θ. Thus,

$$\frac{\partial U(c^c|c_q;0,\tilde{\eta})}{\partial \tilde{\eta}} = -(1-q)D \tag{B.1.68}$$

where $D > 0$.

Thus, utility of challenger strictly decreases in incumbent's ability, $\tilde{\eta}$, for $q \in [0,1)$ and weakly decreases for $q = 0$. □

Characterizing Incumbent's Utility under q

Recall that

$$\begin{aligned} x_q^i(\theta;\tilde{\eta}) &= c^i - c_q = \big(\tilde{\eta} - q\theta - (1-q)\tilde{\eta}\big)G \\ &= q(\tilde{\eta} - \theta)G \end{aligned} \tag{B.1.69}$$

$$\begin{aligned} \Theta^i(0,\tilde{\eta}) &= \sup\{\{0\} \cup \{\theta \in [0,1] : x_q^i(\theta;\tilde{\eta}) \geqslant 0\}\} \\ &= \sup\{\{0\} \cup \{\theta \in [0,1] : q(\tilde{\eta} - \theta)G \geqslant 0\}\} \\ &= \sup\{\{0\} \cup \{\theta \in [0,1] : \theta \leqslant \tilde{\eta}\}\} \\ &= \tilde{\eta} \end{aligned} \tag{B.1.70}$$

$$H^i(0) = \inf\{\{1\} \cup \{\tilde{\eta} \in [0,1] : \Theta^i(0,\tilde{\eta}) = 1\}\}$$
$$= \inf\{\{1\} \cup \{\tilde{\eta} \in [0,1] : \tilde{\eta} = 1\}\} = \inf\{\{1\} \cup \{1\}\} = 1 \tag{B.1.71}$$

Thus, $H^i(0) = 1$, which is consistent with Figure B.3. For $q \in (0,1]$,

$$\frac{\partial x_q^i(\theta;\tilde{\eta})}{\partial \theta} = -qG < 0 \tag{B.1.72}$$

$$x_q^i(0;\tilde{\eta}) = \tilde{\eta} q G \tag{B.1.73}$$

$$-x_q^i(1;\tilde{\eta}) = (1-\tilde{\eta}) q G \tag{B.1.74}$$

By Lemma B.3 and Figure B.3, $0 \leqslant \tilde{\eta} \leqslant H^i(0) \Longrightarrow \Theta^i \in (0,1)$.

Similarly, $H^i(0) \leqslant \tilde{\eta} \leqslant 1 \Longrightarrow \Theta^i = 1$.

$$U(c^i|c_q;\tilde{\eta}) = \Big((1-\tau)y + \tilde{\eta}G\Big) + \frac{\gamma k\big(x_q^i(0;0,\tilde{\eta})^{\frac{1+k}{k}} - \lambda(-x_q^i(1;0,\tilde{\eta}))^{\frac{1+k}{k}}\big)}{(1+k)qG} \tag{B.1.75}$$

Otherwise, if $q = 0$ then $\frac{\partial x_q^i(\theta;\tilde{\eta})}{\partial \theta} = 0$ and Corollary B.4 applies:

$$U(c^i|c_0;\tilde{\eta}) = (1-\tau)y + \tilde{\eta}G + \mu\left(0(\tilde{\eta}-\theta)G\right)$$
$$= (1-\tau)y + \tilde{\eta}G \tag{B.1.76}$$

Lemma B.6. *Utility of incumbent is strictly increasing in $\tilde{\eta}$ for all reference points $q \in [0,1]$*

Proof. Note that for all $q \in [0,1]$, the derivative of incumbent's gain-loss w.r.t. to incumbent's ability, $\tilde{\eta}$, is independent of his ability:

$$\frac{\partial x_q^i}{\partial \tilde{\eta}} = qG \geqslant 0 \tag{B.1.77}$$

If $q = 0$, differentiating $U(c^i|c_0; 0, \tilde{\eta})$ w.r.t. to $\tilde{\eta}$ gives:

$$\frac{\partial U(c^i|c_0; 0, \tilde{\eta})}{\partial \tilde{\eta}} = G > 0 \tag{B.1.78}$$

If $q \in (0, 1]$, differentiating $U(c^i|c_q; 0, \tilde{\eta})$ w.r.t. to $\tilde{\eta}$ gives:

$$\frac{\partial U(c^i|c_q; 0, \tilde{\eta})}{\partial \tilde{\eta}} = G + \gamma\big(x_q^i(0; 0, \tilde{\eta})^{\frac{1}{k}} + \lambda(-x_q^i(1; 0, \tilde{\eta}))^{\frac{1}{k}}\big) > 0 \tag{B.1.79}$$

Thus, utility of incumbent strictly decreases in his ability, $\tilde{\eta}$, for all $q \in [0, 1]$. □

B.2 Proofs of Results

The following Lemma will be used in a key step in the following proposition to show that incumbent of ability 1/2 is strictly preferred to an unknown challenger.

Lemma B.7. *For all $q \in (0, 1]$, for all $k \in \mathbb{N}$,*

$$S \equiv k(1 + q)^{\frac{2k+1}{k}} - k(1 - q)^{\frac{2k+1}{k}} > 2(2k + 1)q^{\frac{1+k}{k}}$$

Proof. Note that the power series expansion of $(1 + y)^{\frac{2k+1}{k}}$ converges aboslutely for $|y| < 1$ for any exponent and also for $|y| = 1$ because $\frac{2k+1}{k} > 0$.

Using binomial-series expansion around $q = 0$,

$$\begin{aligned} S &= k \sum_{m=0}^{\infty} \binom{2+1/k}{m} q^m - k \sum_{m=0}^{\infty} \binom{2+1/k}{m} (-q)^m \\ &= k \sum_{m=0}^{\infty} \binom{2+1/k}{m} \left(q^m - (-1)^m q^m\right) \\ &= 2k \sum_{n=0}^{\infty} \binom{2+1/k}{2n+1} q^{2n+1} \\ &= 2k \frac{(2+1/k)}{1!} q^1 + 2k \frac{(2+1/k)\,(1+1/k)\,(1/k)}{3!} q^3 + O(q^5) \end{aligned} \tag{B.2.1}$$

because if m is odd, then $q^m - (-1)^m q^m = 2q^m$. If m is even, then $q^m - (-1)^m q^m = 0$.

Note that the first term equals to $2(2k+1)q$ and dominates $2(2k+1)q^{\frac{1+k}{k}}$ because it has smaller exponent and $q \leqslant 1$.

$$S - 2(2k+1)q^{\frac{1+k}{k}} > 2k \frac{(2+1/k)\,(1+1/k)\,(1/k)}{3!} q^3 + 2k \sum_{n=2}^{\infty} \binom{2+1/k}{2n+1} q^{2n+1} \tag{B.2.2}$$

Next we will show that the remainder term is strictly positive for all $q \in (0, 1]$ and all $k \in \mathbb{N}$ and this will complete the proof.

Suppose $n \geqslant 2$ and let $t_n = 2k\binom{2+1/k}{2n+1} q^{2n+1}$ be the n-th term of the binomial expansion and let $r_n = \frac{t_{n+1}}{t_n}$. We will use the ratio test to bound the remainder by two geometric series.

$$\begin{aligned} r_n &= \frac{\binom{2+1/k}{2n+3}}{\binom{2+1/k}{2n+1}} q^2 = \frac{\left[\frac{(2+1/k)(1+1/k)(1/k)(1-1/k)(2-1/k)(3-1/k)\cdots(2(n-1)-1/k)(2n-1-1/k)(2n-1/k)}{(2n+3)!}\right]}{\left[\frac{(2+1/k)(1+1/k)(1/k)(1-1/k)(2-1/k)(3-1/k)\cdots(2(n-1)-1/k)}{(2n+1)!}\right]} q^2 \\ &= \frac{(2n-1-1/k)(2n-1/k)}{(2n+2)(2n+3)} = \frac{4n^2 - 4\frac{n}{k} - 2n + \frac{1}{k^2} + \frac{1}{k}}{(2n+2)(2n+3)} q^2 \end{aligned} \tag{B.2.3}$$

Next to show that r_n is strictly increasing for $n \geqslant 2$, differentiate it with respect to n and then observing that the numerator is increasing in n, bound the derivative, for $n \geqslant 2$, from below, by substituting $n = 2$ into the numerator:

$$\begin{aligned}\frac{dr_n}{dn} &= \frac{(1+3k)(8n^2k + (8k-4)n - 2k - 5)}{2k^2(n+1)^2(2n+3)^2}q^2 \\ &\underbrace{\geqslant}_{n \geqslant 2} \frac{(1+3k)(32k + (8k-4)2 - 2k - 5)}{2k^2(n+1)^2(2n+3)^2}q^2 = \frac{(1+3k)(46k-13)}{2k^2(n+1)^2(2n+3)^2}q^2 > 0\end{aligned} \tag{B.2.4}$$

Since r_n is strictly increasing and $r_n \to q^2$, we will bound the remainder by two geometric series. Since all terms of the remainder are positive, a geometric series starting with the same first term and a ratio that is smaller than all r_n will correspond to a sum of smaller terms and form a lower bound. Vice-versa with ratio exceeding all r_n to form an upper bound.

$$\forall n \geqslant 2,\ r_n \text{ increasing: } r_2 \leqslant r_n < q^2 \leqslant 1 \tag{B.2.5}$$

$$t_2 = \binom{2+1/k}{5} q^5 = \frac{4k^4 - 5k^2 + 1}{60k^4} q^5 \tag{B.2.6}$$

$$r_2 = \frac{(3-1/k)(4-1/k)}{42} q^2 \tag{B.2.7}$$

$$\frac{t_2}{1-r_2} < 2k \sum_{n=2}^{\infty} \binom{2+1/k}{2n+1} q^{2n+1} < \frac{t_2}{1-q^2} \tag{B.2.8}$$

Thus,

$$S - 2(2k+1)q^{\frac{1+k}{k}} > 2k \frac{(2+1/k)\,(1+1/k)\,(1/k)}{3!} q^3 + \frac{t_2}{1-r_2} > 0$$

□

Theorem 1: Incumbency Advantage. First will show tha every rational decision rule $\tilde{p_I}(\tilde{\eta}; q)$ is a step function with some cutoff $\eta_q^* = Q(q; \lambda, \gamma; \alpha, \tau, k, \overline{r})$.

This is equivalent to showing: for all $\tilde{\eta} \in [0, Q)$, the challenger is strictly preferred to the incumbent, $U(c^i|c_q) < U(c^c|c_q)$, while for all $\tilde{\eta} \in (Q, 1]$, the incumbent is strictly preferred to the challenger, $U(c^i|c_q) > U(c^c|c_q)$, with indifference at $\tilde{\eta} = Q$.

Define incumbent's utility as a function f of parameter $\tilde{\eta}$, challenger's utility as a function g of parameter $\tilde{\eta}$ and the difference utility of picking incumbent as d :

$$f(\tilde{\eta}) = U(c^i|c_q; \tilde{\eta}) \tag{B.2.9}$$

$$g(\tilde{\eta}) = U(c^c|c_q; \tilde{\eta}) \tag{B.2.10}$$

$$d(\tilde{\eta}) \equiv f(\tilde{\eta}) - g(\tilde{\eta}) \tag{B.2.11}$$

For all $q \in [0, 1]$, we found f is strictly increasing by Lemma B.6 and g is weakly decreasing by Lemma B.5. Thus, d is strictly increasing, so it has at most one root on $[0, 1]$.

Next, we will show d has exactly one root on $(0, 1/2)$ by establishing that $d(0) < 0 < d(1/2)$ for all $q \in [0, 1]$.

Case I. For $q = 0$, because $\lambda > 1$,

$$d(\tilde{\eta}) = G\left(\tilde{\eta} - \frac{1}{2}\right) - \frac{\gamma k G^{1/k}\left((1-\tilde{\eta})^{\frac{k+1}{k}} - \lambda\left(\tilde{\eta}\right)^{\frac{k+1}{k}}\right)}{(k+1)}$$

$$d(0) = -\frac{G}{2} - \frac{\gamma k G^{1/k}}{k+1} < 0 \tag{B.2.12}$$

$$d\left(\frac{1}{2}\right) = 0 + (\lambda - 1)\frac{k\gamma}{(k+1)2^{\frac{k+1}{k}}} > 0 \tag{B.2.13}$$

Note that if $\lambda = 1$ (no loss-aversion), then the unique root is $\tilde{\eta} = \frac{1}{2}$ because $d\left(\frac{1}{2}\right) = 0$.

Case II. For $q \in (0, 1]$:

For any $0 < q \leqslant 1$, for $L = \frac{2k+1}{k}$:

$$d(\tilde{\eta}) = G\left(\tilde{\eta} - \frac{1}{2}\right) + \frac{\gamma k\big(x_q^i(0;\tilde{\eta})^{\frac{1+k}{k}} - \lambda(-x_q^i(1;\tilde{\eta}))^{\frac{1+k}{k}}\big)}{(1+k)qG} - \frac{\gamma k^2\big(-\lambda(-x_q^c(1,0))^L + \lambda(-x_q^c(0,0))^L - x_q^c(1,1)^L + x_q^c(0,1)^L\big)}{(k+1)(2k+1)G^2q} \tag{B.2.14}$$

Evaluating $d(0)$:

$$x_q^i(0;0) = 0 \tag{B.2.15}$$

$$x_q^c(0,1;0) = G \tag{B.2.16}$$

$$-x_q^c(0,0;0) = 0 \tag{B.2.17}$$

$$-x_q^i(1;0) = qG \tag{B.2.18}$$

$$x_q^c(1,1;0) = (1-q)G \tag{B.2.19}$$

$$-x_q^c(1,0;0) = qG \tag{B.2.20}$$

Substituting,

$$\begin{aligned} d(0) &= -\frac{G}{2} - \lambda\gamma\frac{k(qG)^{1/k}}{1+k} \\ &\quad - \frac{\gamma k^2 G^{1/k}\big(-\lambda(q)^{\frac{2k+1}{k}} + \lambda(0)^{\frac{2k+1}{k}} - (1-q)^{\frac{2k+1}{k}} + (1)^{\frac{2k+1}{k}}\big)}{(k+1)(2k+1)q} \\ &= -\frac{G}{2} - \lambda\gamma\frac{k(qG)^{1/k}}{1+k} + \frac{\gamma k^2 G^{1/k}\big(\lambda(q)^{\frac{2k+1}{k}} + (1-q)^{\frac{2k+1}{k}} - 1\big)}{(k+1)(2k+1)q} \end{aligned} \tag{B.2.21}$$

Factoring, $d\,(0)$ get:

$$d(0) = -\frac{G}{2} - M\big[\lambda(2k+1)q^{1+1/k} - \lambda k q^{2+1/k} + 1 - (1-q)^{2+1/k}\big] < -\frac{G}{2} < 0 \tag{B.2.22}$$

where $M \equiv \frac{\gamma k G^{1/k}}{q(1+k)(2k+1)} > 0$

because for all natural numbers k and for all $q \in (0,1]$, raising q to a larger power makes it (weakly) smaller. Therefore,

$$(2k+1)q^{1+1/k} > kq^{2+1/k} \text{ and } (1-q)^{2+1/k} < 1. \tag{B.2.23}$$

Evaluating $d\left(\frac{1}{2}\right)$:

$$x_q^i\left(0;\frac{1}{2}\right)=\frac{qG}{2} \quad \text{(B.2.24)} \qquad -x_q^i\left(1;\frac{1}{2}\right)=\frac{qG}{2} \quad \text{(B.2.27)}$$

$$x_q^c\left(0,1;\frac{1}{2}\right)=\frac{(1+q)G}{2} \quad \text{(B.2.25)} \qquad x_q^c\left(1,1;\frac{1}{2}\right)=\frac{(1-q)G}{2} \quad \text{(B.2.28)}$$

$$-x_q^c\left(0,0;\frac{1}{2}\right)=\frac{(1-q)G}{2} \quad \text{(B.2.26)} \qquad -x_q^c\left(1,0;\frac{1}{2}\right)=\frac{(1+q)G}{2} \quad \text{(B.2.29)}$$

Substituting,

$$\begin{aligned} d\left(\frac{1}{2}\right) &= 0-(\lambda-1)\gamma\frac{k(qG)^{1/k}}{(1+k)2^{\frac{1+k}{k}}} \\ &\quad - \frac{\gamma k^2 G^{1/k}\big(-\lambda(\frac{1+q}{2})^{\frac{2k+1}{k}}+\lambda(\frac{1-q}{2})^{\frac{2k+1}{k}}-(\frac{1-q}{2})^{\frac{2k+1}{k}}+(\frac{1+q}{2})^{\frac{2k+1}{k}}\big)}{(k+1)(2k+1)q} \\ &= -(\lambda-1)\gamma\frac{k(qG)^{1/k}}{(1+k)2^{\frac{1+k}{k}}}+(\lambda-1)\frac{\gamma k^2 G^{1/k}\big((1+q)^{\frac{2k+1}{k}}-(1-q)^{\frac{2k+1}{k}}\big)}{(k+1)(2k+1)q2^{\frac{2k+1}{k}}} \end{aligned} \quad \text{(B.2.30)}$$

Factoring, $d\left(\frac{1}{2}\right)$ get:

$$d\left(\frac{1}{2}\right)=(\lambda-1)\frac{M}{2^{\frac{2+k}{k}}}\big(k(1+q)^{\frac{2k+1}{k}}-k(1-q)^{\frac{2k+1}{k}}-2(2k+1)q^{\frac{1+k}{k}}\big) \quad \text{(B.2.31)}$$

where $M\equiv\frac{\gamma k G^{1/k}}{q(1+k)(2k+1)}>0$

By Lemma B.7, $d\left(\frac{1}{2}\right)>0$.

By intermediate value theorem and d continuous, $\exists\eta_q^*\in(0,\frac{1}{2}) : d(\eta^*)=0$ Thus, $U(c^i|c_q;\eta_q^*)=U(c^c|c_q;\eta_q^*)$ (unique η_q^* by strict monotonicity of d in $\tilde{\eta}$ for fixed q). $\square$

Appendix C

FTBE Details

C.1 Proofs of Results

Theorem (**Necessity**). *If a social choice set F is implementable in k-FTBE, then there exists equivalent $\hat{F}$ that satisfies (k-IC) and (wk-BM).*

Proof. 1. Let $(\mathcal{M}, g)$ implement F and define $\hat{F} = \{x | \exists \sigma \in \mathcal{B}^k(\mathcal{M}, g) : \forall s \in S, x(s) = g[\sigma(s)]\}$. By construction, F and $\hat{F}$ are equivalent.

$\forall x \in \hat{F}, \exists \sigma \in \mathcal{B}^k(\mathcal{M}, g) : \forall s \in S, x(s) = g[\sigma(s)]$. Fix arbitrary $i \in N, t^i \in S^i$. Define $\tilde{\sigma}^i$ as follows: $\forall s^i \in S^i, \tilde{\sigma}^i(s^i) = \sigma^i(t^i)$. Furthermore, consider arbitrary deceptions of up to k other players: $\forall \beta^{-i} \in B(id, k; A_{-i})$, let $\beta = (\beta^{-i}, id)$ and $Q = \{q \in N \backslash \{i\} | \beta^q \neq id\}$. By construction, $|Q| \leqslant k$. Define $\forall j \in N \backslash \{i\}, \overline{\sigma}^j$ as follows: $\forall s^j \in S^j, \overline{\sigma}^j(s^j) = \sigma^j(\beta^j(s^j))$. Thus, $\forall j \in N \backslash Q, \overline{\sigma}^j(s^j) = \sigma^j(s^j)$. Combining these we get the outcome as follows: $g[\overline{\sigma}(s)] = g[(\sigma(s)] \circ \beta = x \circ (\beta^{-i}, id)$, where $\overline{\sigma}$ is in k-neighborhood of σ with player i playing according to σ^i. Now define $\tilde{\sigma}^{-i} = \overline{\sigma}^{-i}$, so $\tilde{\sigma}$ furthermore has player i make a deviation from σ^i: $g[\tilde{\sigma}(s)] = x \circ (\beta^{-i}, t^i)$.

Since σ is FTBE: $g[\overline{\sigma}(s)] \mathcal{R}^i(s^i) g[\tilde{\sigma}(s)]$, that is $x \circ (\beta^{-i}, id) \mathcal{R}^i(s^i) x \circ (\beta^{-i}, t^i)$ as required, so $\hat{F}$ satisfies (k-IC).

2. Now will show that $\hat{F}$ satisfies (wk-BM).

$\forall x \in \hat{F}, \exists \sigma \in \mathcal{B}^k(\mathcal{M}, g) : \forall s \in S, x(s) = g[\sigma(s)]$. By hypothesis of (wk-BM), suppose for deception $\alpha, \nexists z \in F : \forall s \in T, z(s) = x \circ \alpha(s)$. Then $\sigma \circ \alpha \notin \mathcal{B}^k(\mathcal{M}, g)$ because F is implementable. Thus, $\exists s \in T, \exists Q \subset N : |Q| = \tilde{k} \leqslant k, \exists j \in N \backslash Q, \exists (\sigma'^j, \sigma'^Q \circ \beta^Q)$:

$$g(\sigma'^j, \sigma^{*-M} \circ \alpha^{-M}, \sigma'^Q \circ \beta^Q) \mathcal{P}^j(s^j) g(\sigma^j \circ \alpha^j, \sigma^{*M} \circ \alpha^{-M}, \sigma'^Q \circ \beta^Q) \tag{C.1.1}$$

where $M \equiv \{Q \cup j\}$ and $\forall t^j \in S^j, \sigma'^j(t^j) = m'^j \in \mathcal{M}_j$ (j's constant message). Thus, $\forall \beta^j \in A_j, \sigma'^j \circ \beta^j = \sigma'^j$.

If $\tilde{k} < k$, expand the set of faulty players to k as follows: $Q \subset N \backslash \{j\} : Q \subset R, |R| = k$. The strategy and deception for the expanded group R are:

$$\sigma'^R = ((\sigma'^i)_{i \in Q}, (\sigma^{*i})_{i \in R \backslash Q}),$$
$$\beta^R = ((\beta^i)_{i \in Q}, (\alpha^i)_{i \in R \backslash Q}),$$

where β^i is an arbitrary deception and the "whistleblowing" coalition becomes $M \equiv R \cup \{j\}$. Thus, without loss of generality we can take $|Q| = k$ in the following arguments.

Denote $Q_i \equiv M \backslash \{i\}$. Since σ^* is k-FTBE, $\forall i \in Q, \forall t^i \in S^i$: define $x^i = g[(\sigma^i, \sigma'^{Q_i}, \sigma^{*N \backslash M})]$ and $y = g[(\sigma'^M, \sigma^{*N \backslash M})]$.

Next, $\forall \beta^{Q_i} \in B(id, k; A_{Q_i})$:

$x^i \circ (\beta^{Q_i}, id^{N \backslash Q_i}) = g[(\sigma^{*i}, \sigma'^{Q_i}, \sigma^{*N \backslash M}) \circ (id, \beta^{Q_i}, id^{N \backslash M})] = g(\sigma^{*i}, \sigma'^{Q_i} \circ \beta^{Q_i}, \sigma^{*N \backslash M})$

is a k-fault deviation from σ^* equilibrium, so by implementation of $\hat{F}$, $x^i \in \hat{F}$ and $x^i \circ (\beta^{Q_i}, id^{N \backslash Q_i}) \in \hat{F}$. Secondly, it is not profitable for i to deviate from σ^{*i} to σ'^i given Q_i are k faulty players: $y \circ (\beta^{Q_i}, id^{N \backslash M}, \alpha^i(s^i)) = g[(\sigma'^{Q_i}, \sigma^{*N \backslash M}, \sigma'^i) \circ (\beta^{Q_i}, id^{N \backslash M}, \alpha^i(s^i))] = g(\sigma'^{Q_i} \circ \beta^{Q_i}, \sigma^{*N \backslash M}, \sigma'^i)$.

Meaning, $\forall t_i \in S^i, x^i \circ (\beta^{Q_i}, id^{N \backslash Q_i}) \mathcal{R}^i(t^i) y \circ (\beta^{Q_i}, id^{N \backslash M}, \alpha^i(s^i))$, satisfying part (1) of (wk-BM) for $\hat{F}$.

$x^i \circ (id, \beta^{M \backslash \{i\}}, id^{N \backslash M}) = g[(\sigma^{*i}, \sigma'^{M \backslash \{i\}}, \sigma^{*N \backslash M}) \circ (id, \beta^{M \backslash \{i\}}, id^{N \backslash M})]$

Since $g(\sigma'^j, \sigma^{*N\setminus M} \circ \alpha^{N\setminus M}, \sigma'^{Q_j} \circ \beta^{Q_j}) = g(\sigma'^j \circ \alpha^j, \sigma^{*N\setminus M} \circ \alpha^{N\setminus M}, \sigma'^{Q_j} \circ \beta^{Q_j})$, equation (4.1) can be rewritten as:

$g[(\sigma'^j, \sigma^{*N\setminus M}, \sigma'^{Q_j}) \circ (\alpha^j, \alpha^{N\setminus M}, \beta^{Q_j})]\mathcal{P}^j(s^j)g[(\sigma^j, \sigma^{*N\setminus M}, \sigma'^{Q_j}) \circ (\alpha^j, \alpha^{N\setminus M}, \beta^{Q_j})]$

Letting $\beta_0^{-j} = (\alpha^{N\setminus M}, \beta^{Q_j}) \in B(\alpha, k; A_{-j})$ gives $y \circ (\beta_0^{-j}, \alpha^j)\mathcal{P}^j(s^j)x^j \circ (\beta_0^{-j}, \alpha^j)$, which satisfies part (2) of (wk-BM) for $\hat{F}$.

□

Theorem (**Sufficiency**). *If* $|N| \geqslant 3, k \leqslant \frac{|N|}{2} - 1$. *A social choice set* F *satisfies (C), (k-IC) and (k-MNV), is implementable (in k-FTBE).*

1. If F satisfies (k-IC), then: $\forall x^* \in F, \exists \sigma^* \in \mathcal{B}^k(\mathcal{M}, g) : \forall s \in T, g[\sigma^*(s)] = x^*(s)$.

Proof. (similar to Lemma 1 in Jackson (1991)) Pick any $x \in F$. We will construct k-FTBE equilibrium using the following strategy:

$$\forall i \in N, \forall s^i \in S^i, \sigma^i(s^i) \equiv (s^i, x, \varnothing, \cdot, \varnothing).$$

Consider player i's deviation $\tilde{m}^i$ in state s^i, having freedom over beliefs for faulty deviations by arbitrary k other players. Observe that d_3 and the corresponding rule 4 (the integer game) are unreachable by construction of the strategy σ as there are always at least $N - k - 1$ players agreeing on $(\cdot, x, \varnothing, \cdot, \varnothing)$. A belief about k faulty players defines deception for the other $N - 1$ players:

$$\forall M \subset N, |M| \leqslant k, \beta^{-i} \in B(id, k; A_{-i}).$$

$\forall \tilde{s}^i \in S$, if $\tilde{m}^i = (\tilde{s}^i, x, \cdot, \cdot, \varnothing)$ or $\tilde{m}^i = (\tilde{s}^i, \overline{x}, \cdot, \cdot, \cdot)$, then

$$[\tilde{m}^i, \sigma^{*N\setminus M\cup\{i\}}, \sigma^M] \in d_0 \cup d_1 \cup d_2$$

with the outcome $x \circ (\beta^{-i}, \tilde{s}^i)$, which is weakly worse than the expected equilibrium outcome of $x \circ (\beta^{-i}, id)$ by (k-IC).

Second, if $\tilde{m}^i = (\tilde{s}^i, x, \cdot, \cdot, y)$ and $\exists j \in M : |M| = k, \sigma^j \neq (\cdot, x, \cdot, \cdot, y)$ (where M is the set of k faulty players, according to i's belief), then $[\tilde{m}^i, \sigma^{*N\setminus M\cup\{i\}}, \sigma^M] \in d_2$ and the outcome is still $x \circ (\beta^{-i}, \tilde{s}^i)$. The same logic applies if $|M| < k$.

Third, if $\tilde{m}^i = (\tilde{s}^i, x, \cdot, \cdot, y)$ and $\forall j \in M : |M| = k, \sigma^j = (\cdot, x, \cdot, \cdot, y)$, then this message belongs to rule 1 of the mechanism: $[\tilde{m}^i, \sigma^{*N\setminus M\cup\{i\}}, \sigma^M] \in d_1$.

If $\forall j \in M \cup \{i\}, \forall \beta^{-j} \in B(id, k; A_{-j}), \forall t^j \in S^j : x \circ (\beta^{-j}, id)\mathcal{R}^j(t^j)y \circ (\beta^{-j}, m_1^j)$, then the outcome is $y \circ (\beta^{-i}, \tilde{s}^i)$, which is not improving for i by the mechanism construction.

Fourth, messages as before but now if $\exists j \in M \cup \{i\}, \exists \beta^{-j} \in B(id, k; A_{-j}), \exists t^j \in S^j : y \circ (\beta^{-j}, m_1^j)\mathcal{P}^j(t^j)x \circ (\beta^{-j}, id)$, then the outcome is $x \circ (\beta^{-i}, \tilde{s}^i)$, which is non-improving by (k-IC).

□

2. $\forall \sigma^* \in \mathcal{B}^k(\mathcal{M}, g), \exists x^* \in F : \forall s \in T, g[\sigma^*(s)] = x^*(s)$.

Proof. Follows from part (3) and $P = \varnothing$. □

3. $\forall \sigma^* \in \mathcal{B}^k(\mathcal{M}, g), \forall \tilde{\sigma} \in B(\sigma^*, k), \exists \tilde{x} \in F : \forall s \in T, z(s) \equiv g[\tilde{\sigma}(s)] = \tilde{x}(s)$.

Proof. Fix $\sigma^* \in \mathcal{B}^k(\mathcal{M}, g)$ with corresponding k-deviation $z = g(\sigma^{*-P}, \hat{\sigma}^P), |P| \leqslant k$ (we will show z is desirable) and let $\alpha(s) = m_1(s). \forall i \in N, \forall x, y \in F$, let

$$B_x^i = \{s^i \in S^i : \sigma^{*i}(s^i) = (\alpha^i(s^i), x, \varnothing, \cdot, \varnothing)\}. \tag{C.1.2}$$

$$B_{x,y}^i = \{s^i \in S^i : \sigma^{*i}(s^i) = (\alpha^i(s^i), x, \varnothing, \cdot, y)\}, \tag{C.1.3}$$

$$R_{\alpha,x,y}^i = \{s^i \in S^i : \forall \beta^{-i} \in B(id, k; A_{-i}), \forall t^i \in S^i, x \circ (\beta^{-i}, id)\mathcal{R}^i(t^i)y \circ (\beta^{-i}, \alpha^i(s^i))\}. \tag{C.1.4}$$

Part I: First, we will show that k-NVH holds for z and α on

$$D = T\backslash(\cup_{x\in F} \cup_{P\subset R,|R|=k} B_x^{N\backslash R}). \tag{C.1.5}$$

Suppose not. Since k-NVH requires a certain statement to hold for all but one rational player among $N\backslash P$, the negated statement holds for at least two of these rational players.

$$\exists s \in D : \exists i_1, i_2 \in N\backslash P, z^{i_1}, z^{i_2} \in X : \forall C \subset D(s \in C) : \forall n \in \{1, 2\}, \tag{C.1.6}$$

$$z^{i_n} \circ \alpha/_C z\mathcal{P}^{i_n}(s^{i_n})z. \tag{C.1.7}$$

Enlarge the set P to contain exactly k elements without drawing on the two labelled rational players. Call this enlarged set $Q \subset N\backslash\{i_1, i_2\} : P \subset Q, |Q| = k$.[1]

Since

$$s \in T\backslash \left(\cup_{x\in F} \cup_{P\subset R,|R|=k} B_x^{N\backslash R}\right), \tag{C.1.8}$$

then $\forall x \in F, s \notin B_x^{N\backslash R}$. That is, $\exists l \in N\backslash Q : s^l \in S^l\backslash B_x^l$. Let $i \in \{i_1, i_2\}\backslash l$.[2] Now will construct profitable deviation for i, who is a rational player and not contained in P. Therefore, this deviation would be relative to i's equilibrium strategy σ_i^* under a particular belief about Q. Denote $M = \{Q \cup i\}$ as the minority consensus.

Since z is the outcome of $(\mathcal{M}, g)$ under $\alpha, \exists \hat{z} \in X : z = \hat{z} \circ \alpha$. Let

$$\forall q \in P, \tilde{\sigma}^q = (\hat{\sigma}_1^q, \hat{\sigma}_2^q, \hat{z}, N^*, \hat{\sigma}_5^q) \tag{C.1.9}$$

[1]Negation of k-NVH gave us 2 rational players and the superset of faulty players Q with k elements does not contain those rational players.

[2]l is the rational player who didn't play B_x^l at s^l

where

$$N^* > \max_{k \in N, t^k \in S^k} \{\sigma_4^{*k}(t^k), \hat{\sigma}_4^k(t^k)\}. \tag{C.1.10}$$

Also define

$$\forall q \in Q \backslash P, \tilde{\sigma}^q = (\sigma_1^{*q}, \sigma_2^{*q}, \hat{z}, N^*, \sigma_5^{*q}). \tag{C.1.11}$$

Combining these we get the no-deviation as follows:

$$g(\sigma^{*-Q}, \tilde{\sigma}^Q) = \begin{cases} z & \text{if } (\sigma^{*-Q}, \tilde{\sigma}^Q) \notin d_3, \\ \hat{z} \circ \alpha & \text{if } (\sigma^{*-Q}, \tilde{\sigma}^Q) \in d_3. \end{cases} = z \tag{C.1.12}$$

Let $\tilde{\sigma}^i = (\sigma_1^{*i}, \sigma_2^{*i}, z^i, N^* + 1, \sigma_5^{*i})$.

The deviation outcome is:

$$g(\sigma^{*-M}, \tilde{\sigma}^i, \tilde{\sigma}^Q) = \begin{cases} z & \text{if } (\sigma^{*-M}, \tilde{\sigma}^i, \tilde{\sigma}^Q) \notin d_3, \\ z^i \circ \alpha & \text{if } (\sigma^{*-M}, \tilde{\sigma}^i, \tilde{\sigma}^Q) \in d_3. \end{cases} \tag{C.1.13}$$

By construction, $(\sigma^{*,-M}, \tilde{\sigma}^i, \tilde{\sigma}^Q)(s) \in d_3$ (Rule 4 of the mechanism, the integer game). Let

$$C^i = \{s \in \pi^i(s^i) : (\sigma^{*-M}, \tilde{\sigma}^i, \tilde{\sigma}^Q)(s) \in d_3\}, \tag{C.1.14}$$

thus $s^i \in C^i$. If $t \in \pi^i(s^i) \backslash C^i$ (in other words, t triggers one of Rules 1-3), the outcome is unaffected (remains z). Otherwise, if $t \in C^i$ then $g(\sigma^{*-M}, \tilde{\sigma}^i, \tilde{\sigma}^Q) = z^i \circ \alpha$. This deviation is profitable for i at s^i from negation of (k-NVH):

$$z^i \circ \alpha /_{C^i} z \mathcal{P}^i(s^i) z \iff$$

$$g(\sigma^{*-M}, \tilde{\sigma}^i, \tilde{\sigma}^Q)\mathcal{P}^i(s^i)g(\sigma^{*-Q}, \tilde{\sigma}^Q),$$

a contradiction.

Part II: Second, now can apply (k-MVN) hypothesis to construct: $\exists Q \subset N : P \subset Q$, and $|Q| = k$. The set of k players Q contains the original faulty players P (k or fewer) that generated the k-deviation z from the original hypothesis. Next, we are given a rational player j who will form a "whistleblowing" coalition with players Q : $\exists j \in N \backslash Q$. There is a state s in which all other rational players are asking for $x \circ \alpha$ in the original equilibrium σ^* but the rational player j has a profitable deviation by asking for y at s^j. When player j deviates in a coalition with Q, the resulting outcome is $\overline{z}$ but when he does not deviate under the same beliefs about players Q protesting, the outcome of this k-deviation from σ^* is z'. Whenever the integer game is reached, the coalition wins $\tilde{z}$ outcome.

$$\exists x \in F, \exists s \in B_x^{N\backslash Q}, \exists y, \tilde{z}, \overline{z}, z' \in X : \overline{z}\mathcal{P}^j(s^j)z' \tag{C.1.15}$$

As before, denote $M = \{Q \cup j\}$ as the minority consensus. We also have the corresponding deception from k-MNV: $\exists \beta^{-j} \in B(\alpha, k; A_{-j})$.

Consider the following deviation from $\sigma^* : \forall i \in M, \tilde{\sigma}^i = (\alpha^i, x, N^*, \tilde{z}, y)$. k-MNV will ensure that if $z \notin F$, then player j at information s^j, under the belief of Q playing $\tilde{\sigma}^Q = (\beta^i, x, N^*, \tilde{z}, y,$ will deviate from σ^* to $\tilde{\sigma}^j$, which is a different strategy from that prescribed by equilibrium σ^* because $\sigma^{*j}(s^j) = (\alpha^j, x, \varnothing, \cdot, \varnothing)$.

The lengthy construction for $z', \overline{z}$ in k-MNV ensures that these outcomes are derived from the mechanism, finding

$$z' = g(\sigma^{*j}, \sigma^{*-M}, \tilde{\sigma}^R),$$

and

$$\overline{z} = g(\tilde{\sigma}^j, \sigma^{*-M}, \tilde{\sigma}^R)$$

with the later strictly preferred, violating σ^* as k-FTBE.

□

C.2 Related Definitions

Definition C.1. *An environment is said to be economic (E) if*

$$\forall z \in X, \forall s \in S, \exists i, j \in N(i \neq j), \exists x, y \in X$$

such that x and y are constant, $x/_C z P^i(s^i) z$ and

$$y/_C z P^i(s^i) z, \forall C \ni s(C \subset S).$$

An environment is called noneconomic if it is not economic.

Definition C.2 (**Incentive compatibility**)**.** *Given $x \in X, i \in N$ and $t^i \in S^i$, define $\forall s \in S, x_{t^i}$ by $x_{t^i}(s) = x(s^{-i}, t^i)$. A social choice F satisfies incentive compatibility (IC) if $\forall x \in F, \forall i \in N, \forall t^i, s^i \in S^i, x\mathcal{R}^i(s^i)x_{t^i}$.*

Definition C.3 (**Bayesian monotonicity**)**.** *Given deception α and $x \in F$, a social choice set F is Bayesian monotonic if whenever there is no social choice function in F which is equivalent to $x \circ \alpha, \exists i \in N, \exists s^i \in S^i, \exists y \in X : \forall t^i \in S^i, x\mathcal{R}^i(t^i)y_{\alpha^i(s^i)}$ and $y \circ \alpha \mathcal{P}^i(s^i) x \circ \alpha$.*

Definition C.4 (**k-Bayesian monotonicity**)**.** *Given deception α and $x \in F$, a social choice set F is k-Bayesian monotonic if whenever there is no social choice function in F which is equivalent to $x \circ \alpha$,*

1. $\exists M \subset N : |M| \geqslant k+1, \exists s \in S, \exists y \in X : \forall i \in M, \forall \beta^{-i} \in B(id, k; A_{-i}), \forall t^i \in S^i, x \circ (\beta^{-i}, id)\mathcal{R}^i(t^i)y \circ (\beta^{-i}, \alpha^i(s^i))$

2. $\exists j \in M, \exists \beta_0^{-j} \in B(\alpha, k; A_{-j}) : y \circ (\beta_0^{-j}, \alpha^j)\mathcal{P}^j(s^j)x \circ (\beta_0^{-j}, \alpha^j)$.

Definition C.5 (**No-veto hypothesis**). *Given deception α and a subset of plausible states $D \subset T$, a social choice function $z \in X$ satisfies the no-veto hypothesis (NVH) for α and D, if $\forall s \in D, \exists i \in N : \forall j \in N\backslash\{i\}, \forall \tilde{z} \in X, \exists C \subset D : s \in C$ and $z\mathcal{R}^j(s^j)\tilde{z} \circ \alpha/_C z$.*

Definition C.6 (**Monotonicity-no-veto**). *Given deception α, and $\forall x \in F, \forall i \in N$, given a set $B_x^i \subset S^i$. Let $B_x = B_x^1 \times \cdots \times B_x^N$. Suppose $\exists z \in X : \forall x \in F, \forall s \in B_x, z(s) = x \circ \alpha(s)$. Also, suppose z satisfies (NVH) for α and for $D \equiv T\backslash(\cup_{x \in F} B_x)$. Then F satisfies monotonicity-no-veto (MNV) if, whenever there is no social choice function in F which is equivalent to z, $\exists i \in N, \exists x \in F, \exists s \in B_x, \exists y, \tilde{z}, \overline{z} \in X$:*

1. $\forall t \in B_x, \overline{z}(t) = y \circ \alpha(t)$,

2. $\forall \overline{x} \in F\backslash\{x\}, \forall j \in N\backslash\{i\} : t^j \in B_{\overline{x}}^j, \overline{z}(t) = z(t)$

3. $\overline{z}(t) = \tilde{z} \circ \alpha(t)$ *otherwise.*

satisfying $\forall t^i \in S^i, x\mathcal{R}^i(t^i)y \circ (id^-i, \alpha^i(s^i))$, and $\overline{z}\mathcal{P}^i(s^i)z$.

Bibliography

Acemoglu, D., and M. Jackson (2011): "History, Expectations, and Leadership in the Evolution of Social Norms," Discussion paper, National Bureau of Economic Research.

Achen, C., and L. Bartels (2004): "Blind Retrospection: Electoral Responses to Drought, Flu, and Shark Attacks," *Estudio/Working Paper 2004/199*.

Ansolabehere, S., and J. Snyder Jr (2002): "The Incumbency Advantage in US Elections: An Analysis of State and Federal Offices, 1942-2000," *Election Law Journal*, 1(3), 315–338.

Burdzy, K., D. Frankel, and A. Pauzner (2001): "Fast equilibrium selection by rational players living in a changing world," *Econometrica*, 69(1), 163–189.

Cho, I., and A. Matsui (2005): "Time consistency in alternating-move policy games," *Japanese Economic Review*, 56(3), 273–294.

Cole, S., A. Healy, and E. Werker (2008): "Do Voters Appreciate Responsive Governments? Evidence from Indian Disaster Relief," *Harvard Business School Finance Working Paper No. 09-050*.

Diermeier, D., M. Keane, and A. Merlo (2005): "A Political Economy Model of Congressional Careers," *The American economic review*, 95(1), 347–373.

Doghmi, A., and A. Ziad (2007): "Fault tolerant Bayesian implementation in exchange economies," Discussion paper, Mimeo.

——— (2009): "Faulty Nash Implementation in Exchange Economies with Single-peaked Preferences," *Jena Economic Research Papers*, 2009, 073.

Edmond, C. (2011): "Information manipulation, coordination, and regime change," Discussion paper, National Bureau of Economic Research.

Eliaz, K. (2002): "Fault tolerant implementation," *The Review of Economic Studies*, 69(3), 589–610.

Fershtman, C. (1996): "On the Value of Incumbency: Managerial Reference Point and Loss Aversion," *Journal of Economic Psychology*, 17(2), 245–257.

Gelman, A., and G. King (1990): "Estimating Incumbency Advantage Without Bias," *American Journal of Political Science*, 34(4), 1142–1164.

Henriksen, M., and J. R. Isbell (1953): "On the continuity of the real roots of an algebraic equation," *Proceedings of the American Mathematical Society*, 4(3), 431–434.

Holmström, B. (1999): "Managerial incentive problems: A dynamic perspective," *The Review of Economic Studies*, 66(1), 169–182.

Jackson, M. (1991): "Bayesian implementation," *Econometrica*, 59(2), 461–477.

Kennan, J. (2001): "Uniqueness of Positive Fixed Points for Increasing Concave Functions on n: An elementary result," *Review of Economic Dynamics*, 4, 893–899.

Koszegi, B., and M. Rabin (2007): "Reference-Dependent Risk Attitudes," *American Economic Review*, 97(4), 1047–1073.

Kuran, T. (1989): "Sparks and prairie fires: A theory of unanticipated political revolution," *Public Choice*, 61(1), 41–74.

——— (1991): "Now out of never: The element of surprise in the East European revolution of 1989," *World politics*, 44(01), 7–48.

Levitt, S., and C. Wolfram (1997): "Decomposing the Sources of Incumbency Advantage in the US House," *Legislative Studies Quarterly*, 22, 45–60.

Lohmann, S. (1994): "The dynamics of informational cascades," *World politics*, 47(1), 42–101.

Palfrey, T., and S. Srivastava (1989): "Implementation with incomplete information in exchange economies," *Econometrica*, 57(1), 115–134.

Patty, J. (2006): "Loss Aversion, Presidential Responsibility, and Midterm Congressional Elections," *Electoral Studies*, 25(2), 227–247.

Persson, T., and G. Tabellini (2002): *Political Economics: Explaining Economic Policy*. MIT press.

Pipes, R. (1996): *A concise history of the Russian revolution*. Vintage.

Quattrone, G., and A. Tversky (1988): "Contrasting Rational and Psychological Analyses of Political Choice," *American Political Science Review*, 82(3), 719–736.

Saglam, I. (2007): "A Unified Theory of Implementation," *Economics Bulletin*, 4(19), 1–10.

Service, R. (2009): *The Russian Revolution 1900-1917*. New York: Palgrave Macmillan.

Stone, W., L. Maisel, and C. Maestas (2004): "Quality Counts: Extending the Strategic Politician Model of Incumbent Deterrence," *American Journal of Political Science*, 48(3), 479–495.

Wolfers, J. (2007): "Are Voters Rational? Evidence From Gubernatorial Elections," *Revise and resubmit, Review of Economics and Statistics*.

Yin, C. (1998): "Equilibria of collective action in different distributions of protest thresholds," *Public Choice*, 97(4), 535–567.