Incumbent performance and electoral control: a comment *

Germán Gieczewski† Christopher Li‡

September 2023

Abstract

We note and correct a flaw in the analysis of Ferejohn (1986)'s seminal model of electoral accountability. In the original solution, it is supposedly optimal for the voter to impose a stationary path of performance targets on officeholders. We show that, in fact, stationary paths are sub-optimal; the voter can do better by choosing a path of performance targets that become increasingly lenient over time, which extracts more effort from the politician earlier on. We explicitly solve for the optimal performance targets for a class of examples and discuss the substantive implications of our exercise.

*We thank the editor and two anonymous reviewers for helpful comments. The paper also benefited from discussions with Avidit Acharya, Matías Iaryczower, Elliot Lipnowski, Navin Kartik, Mattias Polborn, João Ramos, and Kristopher Ramsay. All remaining errors are our own.

†Department of Politics, Princeton University.

‡Department of Economics and Political Science, Vanderbilt University.

John Ferejohn's seminal[1] 1986 paper introduced one of the first game-theoretic models of electoral accountability. It marked the beginning of a research program in formal political theory that is still growing to this day (see Duggan and Martinelli, 2017 for a survey, as well as Meirowitz, 2007; Anesi and Buisseret, 2022; Acharya, Lipnowski and Ramos, 2022 and many others for recent work in the area).

Ferejohn (1986) (Ferejohn henceforth) is notable for treating the accountability of elected officials as a moral hazard problem—that is, how should voters incentivize politicians to exert effort when their observed performance also depends on events outside their control? Ferejohn took seriously the fact that, unlike in typical economic contexts (e.g., Holmström, 1979), the voter can only reward the politician with reelection, not direct payments; and she has no ability to credibly commit (say through a contract) to the conditions for reelecting the incumbent. Instead, the voter-politician interaction follows the logic of a repeated game: the voter reelects the incumbent in each period only if a performance target is met, and she follows through on this threat because she would otherwise lose her credibility in the future.

Ferejohn's analysis yields a strikingly simple solution: the voter's optimal path of performance targets is purportedly shown to be stationary. That is, it is best for the voter to set the same reelection threshold in every period, and the incumbent's resulting reelection probability is the same in every period. In this note, we show that a key step in the original analysis is incorrect, and correcting the mistake qualitatively alters the nature of the solution. Indeed, the true optimal path of performance targets is in general not stationary, and can be quite complex. For a class of examples, we show that the equilibrium reelection threshold *decreases* over time, and the incumbent's probability of reelection *increases* over time. In other words, the voter finds it optimal to become more lenient in later periods as a way to encourage effort in early periods. Now, it is well

[1]The paper has garnered over three thousand citations and is the most cited work by John Ferejohn according to Google Scholar.

known in the dynamic agency and contracting literature that backloading incentives may be optimal (Lazear 1981; Ray 2002; Burdett and Coles 2003; Acemoglu, Golosov and Tsyvinski 2008; Acharya, Lipnowski and Ramos 2022). What we show is that this logic was already present in Ferejohn's simple model of accountability. We discuss the substantive implications of our findings in the conclusion.

Beyond Ferejohn (1986) itself, the most related prior paper is Acharya, Lipnowski and Ramos (2022) (ALR henceforth). ALR study a variation of Ferejohn's model, with simplified productivity shocks and an expanded strategy space for the voter. They fully characterize the voter-optimal equilibrium, which exhibits an extreme form of backloaded incentive provision, with the possibility of politicians achieving full job security ("complete entrenchment") on the path of play. In particular, stationary retention rules are strictly sub-optimal in ALR's setting whenever the moral hazard problem binds. These observations appear to be at odds with Ferejohn's on an intuitive level, though ALR's and Ferejohn's settings differ enough that the two sets of results are not in direct contradiction. Our note revisits Ferejohn's exact model, pinpoints a mathematical error in his analysis, and shows, after correction, that incentive backloading is in fact also a feature of the equilibrium Ferejohn meant to characterize. We comment further on the relation between ALR and our exercise in the conclusion.

1 Preliminaries

In this section we outline the model as presented in Ferejohn (1986) and make some useful observations, adhering to the original notation as much as possible. Propositions from the original paper retain their numbering. New propositions are enumerated by letters. The reader may wish to consult the original paper for a more detailed discussion of the model.

1.1 Model overview

Time is discrete and infinite, and is indexed by $t = 0, 1, \ldots$. There is a long-lived representative voter and an infinite population of homogeneous politicians, one of which is the initial officeholder. In each period, the voter and the current officeholder ("incumbent") engage in the following interaction, described in chronological order:

1. The incumbent privately observes the realization of a state variable[2] $\theta_t \in [0, m]$, drawn iid from a continuously differentiable c.d.f. F.

2. Then, the incumbent chooses an effort level $a_t \in [0, \infty)$.

3. The voter observes the incumbent's performance, $a_t\theta_t$, but not a_t or θ_t separately. Then the voter decides whether to reelect the incumbent or not.

4. If the incumbent is voted out, an identical replacement comes into office and the game proceeds as before.

The voter's flow payoff is $u_t = a_t\theta_t$. The incumbent's flow payoff is $v_t = W - \phi(a_t)$, where $W > 0$ is office rents, and $\phi(a)$ is the incumbent's cost of effort. An officeholder receives zero flow payoff when out of office. The cost function ϕ is assumed to be increasing and convex, with $\phi(0) = 0$. Discounted payoffs take the usual form with a common discount factor $\delta \in (0, 1)$. In particular, voter welfare is defined to be $\sum_{t=0}^{\infty} \delta^t u_t$.

Ferejohn considers the possibility that the officeholder may return to office in the future after being ousted. Formally, the incumbent, if ousted from office, has a probability $\lambda \in [0, 1]$ of returning to power in each future period in which the (new) incumbent fails to be reelected. (It is worth noting that subsequent work mostly focuses on the case of $\lambda = 0$ i.e., the incumbent cannot return to office after being ousted. Some of our results are obtained for this case.)

[2]The state variable represents shocks to the incumbent's productivity.

Ferejohn provides a somewhat loose definition of strategies and equilibrium and does not spell out their relationship to standard solution concepts. To improve transparency, we introduce some notation and terminology of our own. To begin, a generic strategy profile takes the form (a, R), where $a(h)$ is the officeholder's effort at history h and $R(h')$ is the voter's reelection decision at history h', with $R = 1$ being reelect. The structure of the game allows one to define perfect Bayesian equilibria (PBE) or perfect public equilibria (PPE) in the standard way.[3]

Ferejohn imposes a condition on the voter's strategy that effectively selects a class of equilibria from the full set of PBEs (or PPEs). In particular, the voter is only allowed to condition reelection on the incumbent's *current* performance and on calendar time.[4] Formally, Ferejohn assumes that the voter is restricted to using a (retrospective) cutoff rule, denoted by $K_t \in [0, \infty)$ for $t \geq 0$, by which the officeholder at time t is reelected if and only if $a_t\theta_t \geq K_t$. In terms of the notation that we introduced, a strategy profile that satisfies the above restriction can be written as $(a, R(\mathbf{K}))$, where $\mathbf{K} = (K_t)_{t\geq 0}$ is the voter's cutoff rule,[5] and the function $a(t, \theta_t)$ specifies the officeholder's effort in period t and state θ_t. This description presumes that all officeholders use a common strategy, which does not condition on past values of θ_s nor the players' past actions. This is without loss of generality since the voter's own strategy does not discriminate between officeholders or condition on the past.

Following Ferejohn's terminology, we say that a strategy profile $(a, R(\mathbf{K}))$ constitutes an *equilibrium* if the officeholder's effort choice $a(t, \theta_t)$ maximizes her expected utility at all histories given the cutoff rule $\mathbf{K}$. Note that this definition of equilibrium places no

[3] Ferejohn does not explicitly state his solution concept. Various remarks in the paper suggest subgame perfect equilibrium (SPE) as the basis for his notion of equilibrium (see e.g., the last two paragraphs of Section 1 and the remark on p. 15). Technically, SPE does not apply here because the voter does not observe realizations of θ_t, and so there are no proper subgames. However, PBE and PPE both capture the notion of sequential rationality intended by Ferejohn.

[4] Absent this restriction, the voter might in general condition reelection in period t on the entire history of the game up to t.

[5] A cutoff rule $(K_t)_{t\geq 0}$ induces a voter strategy $R(\mathbf{K})$ such that $R(\mathbf{K})(h^t) = \mathbb{1}_{\{a_t\theta_t \geq K_t\}}$ at all period-t histories h^t. Allowing more general off-path continuations does not affect the results.

explicit restrictions on equilibrium cutoff rule **K**. This is not an issue because it turns out that any cutoff rule is sequentially rational and so is compatible with standard solution concepts such as PBE or PPE. The reason is that, if the voter deviated in some period t (say by reelecting when $a_t\theta_t < K_t$ or vice versa), this would have no impact on her flow payoff (as today's performance is sunk); no effect through selection (as officeholders are homogeneous); and no impact on the officeholder(s)' future behavior (as their strategy will not condition on this deviation). Thus, the voter never benefits from a deviation.

Finally, we denote the voter's welfare in equilibrium for a given cutoff rule by $U(\mathbf{K})$, and following Ferejohn's terminology, we say **K** is the *optimal retrospective rule* if it maximizes $U(\mathbf{K})$ compared to any other cutoff rule.[6]

1.2 Voter's problem

We begin by summarizing the initial steps in Ferejohn's original analysis, which are correct and will be useful for later discussion. As noted above, *any* sequence of cutoffs $(K_t)_{t\geq 0}$ is consistent with an equilibrium (not necessarily maximizing voter welfare). Consequently, the *optimal* retrospective rule is what the voter would choose if she could *credibly commit* to a sequence of cutoffs at the start of the game. Thus, the optimal retrospective rule can be viewed as the "full commitment solution," i.e., the rule that the voter would choose if she had the power to credibly and permanently commit to a rule at the start of the game.

Consider the incumbent's response to an arbitrary cutoff rule $(K_t)_{t\geq 0}$. For a cutoff K_t, the incumbent's optimal effort at date t is either $\frac{K_t}{\theta_t}$, which is just enough to secure reelection,

[6] See Ferejohn's definition of the optimal retrospective rule at the end of page 15 of the original paper.

or 0.[7] The incumbent will choose to meet the cutoff if and only if

$$W - \phi\left(\frac{K_t}{\theta_t}\right) + \delta V_{t+1}^I \geq W + \delta V_{t+1}^O, \tag{1}$$

where V_{t+1}^I is the incumbent's continuation value if he remains in office at date $t+1$, and V_{t+1}^O is the continuation value if he is ousted. (This may be positive as the incumbent may return to office in the future i.e., $\lambda > 0$.) Rearranging terms, one obtains the following characterization of the incumbent's optimal effort.

Proposition 1 (Ferejohn (1986))**.** *Given a sequence of cutoffs $(K_t)_{t\geq 0}$, the incumbent's optimal strategy is*

$$a_t = \frac{K_t}{\theta_t} \quad \textit{iff} \quad \theta_t \geq \theta_t^* := \frac{K_t}{\phi^{-1}\left(\delta V_{t+1}^I - \delta V_{t+1}^O\right)}, \tag{2}$$

and $a_t = 0$ otherwise.

Thus, a sequence of cutoffs $(K_t)_{t\geq 0}$ induces a unique sequence of cutoffs in the state variable $(\theta_t^*)_{t\geq 0}$, such that the incumbent exerts effort, and is reelected, if and only if the realization of state is greater than θ_t^*. To make explicit that the cutoff in the state variable is dependent on the cutoff rule, we sometimes write $\theta_t^*(\mathbf{K})$.

Proposition 1 implies that voter welfare can be written as

$$U(\mathbf{K}) = \sum_{t=0}^{\infty} \delta^t K_t \Pr\left[\theta_t \geq \frac{K_t}{\phi^{-1}\left(\delta V_{t+1}^I - \delta V_{t+1}^O\right)}\right] \tag{3}$$

$$= \sum_{t=0}^{\infty} \delta^t K_t (1 - F(\theta_t^*(\mathbf{K}))). \tag{3'}$$

[7] The incumbent has no incentive to over-perform because doing so has no effect on future cutoffs, and therefore his future prospects. This stems from the assumption that the cutoff rule does not condition on past performances; see Footnote 4.

Ferejohn then proceeds to characterize the optimal retrospective rule. It is worth emphasizing that the difference between Ferejohn's characterization of the optimal retrospective rule and ours owes to a subtle error in his solution to the maximization problem, and not to any difference in the setup or the solution concept.

2 The optimal retrospective rule

In this section, we explicate and rectify the error in the original solution. We then trace the impact of our correction on the substantive predictions of the model, and discuss some new empirical implications. Proofs of new results are contained in the Appendix. In addition, the Online Appendix discusses original results on comparative statics, and provides a teaching guide to help walk graduate students through the proof of Proposition C, which is rather involved. This may be useful for instructors of formal modeling classes covering models of accountability.

2.1 Optimality condition

Ferejohn's Proposition 2 states that the optimal retrospective rule satisfies the following equation:

$$K_t = \frac{1 - F(\theta_t^*)}{f(\theta_t^*)} \phi^{-1}(\delta V_{t+1}^I - \delta V_{t+1}^O) \tag{4}$$

for all t. (Our Equation 4 is identical to Ferejohn's equation (4).) The proof in Ferejohn claims that "this [equation (4)] follows directly from the first-order conditions derived from equation (3)" (see p. 16 of Ferejohn; his equation (3) is also the same as our Equation 3).

Essentially, Equation 4 is meant to be the first-order condition, $\frac{\partial U}{\partial K_t} = 0$, slightly rearranged. But it is not. In particular, Equation 4 is obtained under the presumption that V_s^I

and V_s^O in the right-hand side of Equation 3 are independent of K_t for all s.[8] However, V_s^I and V_s^O are in fact *not* independent of K_t for $t > s$: in general, the incumbent's value function in a given period depends on the performance targets in the current period and all future periods. Thus, altering the value of K_t changes $V_0^I, \ldots, V_t^I, V_0^O, \ldots, V_t^O$, and hence $\theta_0^*, \ldots, \theta_{t-1}^*$, in addition to θ_t^*. The reason is simple: if K_t increases, the value of reaching period t in office declines, as it is harder to be reelected at that point. The officeholder's motivation to exert effort then also declines in all periods *before* t, as she has (in expectation) a shorter tenure to look forward to.[9]

To summarize, the original derivation of the first-order condition incorrectly presumes that the choice of cutoff K_t affects voter welfare only through its effect on her flow payoff at date t. The *correct* first-order condition for the optimal retrospective rule can be derived from Equation 3' as follows:

$$\begin{aligned} 0 = \frac{\partial U}{\partial K_t} =& \delta^t(1 - F(\theta_t^*(\mathbf{K}))) + \delta^t K_t(-f(\theta_t^*(\mathbf{K})))\frac{\partial \theta_t^*(\mathbf{K})}{\partial K_t} \\ &+ \sum_{s=0}^{t-1} \delta^s K_s(-f(\theta_s^*(\mathbf{K})))\frac{\partial \theta_s^*(\mathbf{K})}{\partial K_t}. \end{aligned}$$

Again using that $\frac{\partial \theta_t^*(\mathbf{K})}{\partial K_t} = \frac{1}{\phi^{-1}(\delta V_{t+1}^I - \delta V_{t+1}^O)}$ and rearranging (while suppressing the dependence on **K** to simplify notation), we obtain

$$\begin{aligned} K_t =& \frac{1 - F(\theta_t^*)}{f(\theta_t^*)}\phi^{-1}(\delta V_{t+1}^I - \delta V_{t+1}^O) \\ &- \frac{\phi^{-1}(\delta V_{t+1}^I - \delta V_{t+1}^O)}{f(\theta_t^*)} \sum_{s=0}^{t-1} \delta^{s-t} K_s f(\theta_s^*)\frac{\partial \theta_s^*}{\partial K_t}. \end{aligned} \tag{4'}$$

[8] In this case, K_t would appear only in the t-th term of the summation. Differentiating Equation 3' would yield the first-order condition

$$0 = \frac{\partial U}{\partial K_t} = \delta^t \left(1 - F(\theta_t^*) - K_t f(\theta_t^*)\frac{\partial \theta_t^*}{\partial K_t}\right),$$

where $\frac{\partial \theta_t^*}{\partial K_t} = \frac{1}{\phi^{-1}(\delta V_{t+1}^I - \delta V_{t+1}^O)}$ by Equation 2. Rearranging this expression indeed yields Equation 4.

[9] This argument is straightforward if λ is low enough or, in particular, zero. For high values of λ, it is less clear that being fired in period t is harmful because it may make it easier to return to power in period $t + 2$, if the cutoff in period $t + 1$ is very high.

The values of K_t pinned down by Equations 4 and 4' differ by the additional term in the second line of Equation 4'. The two equations are equivalent only when this term is zero. This is trivially the case for $t = 0$, as the last term is an empty summation, but generally false for $t > 0$. Intuitively, the last term of Equation 4' captures the effect of varying K_t on the voter's flow payoff in periods prior to t. In the special case of $\lambda = 0$, or more generally if λ is small, it is easy to show that this last term is negative.

2.2 On the stationarity of equilibrium

An important property of Ferejohn's solution based on the erroneous first-order condition is that the optimal retrospective rule is stationary, which simplifies the calculation of the voter's and officeholder's utilities and the comparative statics. This property also has two empirical implications. First, the officeholder's performance, $a_t\theta_t$, should be constant over time (up until the period she is ousted). Second, the probability of reelection will be constant over time, since the cutoff value of the state variable leading to reelection, θ_t^*, is stationary as well. As shown below, these observations no longer hold in general when one uses the correct FOC given in Equation 4'. The officeholder's behavior and the probability of turnover may instead follow complex dynamics. A complete characterization of equilibrium in the fullest generality is difficult. For tractability, we impose some assumptions on the environment. They are not overly restrictive and are common in the literature.

For the results below, denote the (constant) sequence of cutoffs that solves Equation 4 for all t by $\overline{\mathbf{K}} \equiv (K_t = \overline{K})_{t\geq 0}$.[10] That is, $\overline{\mathbf{K}}$ is the cutoff rule claimed to be optimal in the original analysis. First, we show that the stationary rule $\overline{\mathbf{K}}$ is not optimal for the voter when the incumbent cannot return to office after being ousted.

[10]Note that this is a fixed point of Equation 4 as V_{t+1}^I, V_{t+1}^O, and θ_t^* are all functions of the sequence of performance thresholds $(K_s)_{s\geq 0}$.

Proposition A. *Assume that* $\lambda = 0$. *Then* $\frac{\partial U(\overline{\mathbf{K}})}{\partial K_0} = 0$ *and* $\frac{\partial U(\overline{\mathbf{K}})}{\partial K_t} < 0$ *for all* $t > 0$. *In addition, the optimal retrospective rule is not stationary.*

In words, the stationary retrospective rule $\overline{\mathbf{K}}$ identified in the original analysis is in fact suboptimal. The voter can improve on $\overline{\mathbf{K}}$ by marginally lowering the cutoff (i.e., being more lenient toward the incumbent) from the second period onward. Intuitively, the effect of the proposed deviation at a date $t > 0$ on the voter's contemporaneous flow payoff is nearly neutral, but it allows the voter to extract strictly higher effort from the incumbent in earlier periods.

An implication of Proposition A is that the equilibrium cutoffs θ_t^* for the state variable are also non-stationary. Moreover, we can show that, if the probability distribution of the state variable satisfies the monotone hazard rate property,[11] then $\theta_t^* < \theta_0^*$ for all $t > 0$. That is, the incumbent is less likely to be reelected in the initial period than in any subsequent period:

Corollary B. *Assume that* $\lambda = 0$ *and* $\frac{1-F(x)}{f(x)}$ *is a decreasing function. Then, in equilibrium, the probability of office turnover in period* $t = 0$ *is greater than in any subsequent period.*

Another implication of Proposition A is that *even if* we force the voter to use stationary cutoffs, the stationary cutoff derived in the original analysis, $\overline{K}$, is *still* not optimal for the voter. Indeed, reducing the cutoff by some small ϵ in every period improves voter welfare while preserving stationarity. This follows from the fact that, letting $\mathbf{K}$ be a sequence with $K_t \equiv K$, it is the case that $\frac{dU(\overline{\mathbf{K}})}{dK} = \sum_{t \geq 0} \frac{\partial U(\overline{\mathbf{K}})}{\partial K_t} < 0$.

The preceding logic suggests that under the actual optimal retrospective rule, cutoffs K_t should be monotonically decreasing in t. After all, if the benefit of lowering K_t below $\overline{K}$ is that the officeholder's motivation improves before t, the resulting gains accrue over more periods the higher t is. We can prove this analytically for the special case of $\lambda = 0$,

[11] This is a standard technical assumption (it is also used in Ferejohn's Proposition 3).

$\phi(a) \equiv a$, and θ_t uniformly distributed on $[0,1]$. This same set of assumptions also appears in Ferejohn (see the Corollary on p. 17).

Proposition C. *Suppose that* $\lambda = 0$, $\phi(a) \equiv a$ *and* $\theta_t \sim U[0,1]$. *Then*

1. *The optimal retrospective rule* $(K_t)_{t\geq 0}$ *decreases in* t *towards a limit* $K_\infty < \overline{K}$, *where* $\overline{K}$ *is the stationary cutoff obtained in Ferejohn's original analysis. More precisely,* $K_0 > \overline{K} > K_1 > \ldots \searrow K_\infty$.

2. *The corresponding sequence of state variable cutoffs,* $(\theta_t^*)_{t\geq 0}$, *decreases in* t *towards a limit* $\theta_\infty^* < \frac{1}{2}$, *where* $\frac{1}{2}$ *is the cutoff obtained in Ferejohn's original analysis. More precisely,* $\theta_0^* = \frac{1}{2} > \theta_1^* > \ldots \searrow \theta_\infty^*$.

3. θ_∞^* *is the unique solution of the equation* $1 - 2\theta_\infty^* + \theta_\infty^* \ln(\theta_\infty^*) = 0$ *between 0 and 1 (*$\theta_\infty^* \approx 0.318$*), and* $K_\infty = \frac{\delta\theta_\infty^* W}{1-\delta\theta_\infty^*}$.

The precise values of θ_t^* and K_t can be obtained by solving a recursive system of equations (Equations 13, 14, 18 in Appendix A). Also, part 2 of the Proposition holds for a broader class of distributions of θ_t than the uniform distribution (see Corollary D in Appendix A).

Proposition C reflects the general intuition about increasing leniency: because setting lower targets in later periods incentivizes the incumbent to work more in early periods, the voter prefers to be lenient and set $K_t < \overline{K}$ for large t. In fact, she does this for all $t \geq 1$. In contrast, she sets an initial target $K_0 > \overline{K}$: because the expectation of future leniency is enough to motivate the incumbent early on, the voter can afford to extract high effort in the first period.

Corresponding to declining cutoffs in performance, θ_t^* decreases over time—that is, the incumbent is more likely to meet the performance target at later times—for two reasons. First, because K_t is decreasing in t, the target becomes easier to meet over time. Second, as the voter becomes more lenient, the incumbent's continuation value from retaining

office increases, which further incentivizes the incumbent to seek reelection. Hence, the incumbent's probability of reelection increases period over period.

2.3 Additional remarks

More on stationarity It is useful to examine the issue of stationarity from a methodological perspective, for it illuminates a subtle pitfall when applying the principle of dynamic programming. In the Remark on p. 17 of Ferejohn (1986), it is argued that the optimal retrospective rule is stationary because one can rewrite the maximization of voter welfare as a dynamic programming problem. The general if-then logic is correct but the premise turns out not to hold in this instance. One cannot apply dynamic programming techniques to maximizing voter welfare in Ferejohn's setting because they are valid only if the maximization problem can be separated into two isolated problems: one of optimizing future choices, and one of optimizing the current choice, given (optimal) future choices. Such separation is impossible in Ferejohn's model since, as we argued above, future choices of cutoffs impact current flow payoffs.

One can construct a variant of Ferejohn's model in which the stationary rule $\overline{\mathbf{K}}$ from the original analysis is indeed optimal. Consider a modified game in which, at the beginning of each period, the voter can announce and *commit* to a cutoff for that period (the voter cannot commit to cutoffs in future periods.) Besides one-period commitment power, let us also impose Markov Perfect Equilibirum (MPE) as the solution concept, so that the voter's announced cutoffs must be the same at every history (as they are all payoff-equivalent) on or off the equilibrium path. In this version of the game, the voter announces $K_t = \overline{K}$ for all t in the unique MPE. Indeed, the game has been modified in a way that allows us to apply the dynamic programming principle: when the voter announces K_t at time t, she has no reason to worry how this choice might have affected the incumbent's incentives in past periods, since the past has already run its course. Her

problem at any time t is then equivalent to the problem she faces at time 0, and thus it is optimal to induce θ_t^* to satisfy the same first-order condition that θ_0^* satisfies, as given in Equation 4. Modifying the game in this way, however, goes against Ferejohn's stated intention of not prescribing exogenous commitment power to the voter (see p. 9 of Ferejohn).

Barro (1973) Another seminal paper on electoral accountability, often mentioned alongside Ferejohn's, is Barro (1973). The two models differ in the details but they are both concerned with how voters should optimally set a performance threshold for reelecting politicians. Strikingly, the stationarity of performance cutoffs is also a feature of the equilibrium in Barro's model (with the exception of the last period, as Barro considers a game of finite horizon). While Barro does not spell out a solution concept, or whether the voter's choice of cutoff is a credible commitment, his solution is also only valid *if* we assume that the voter has one-period commitment power, as sketched above. A detailed examination of Barro's analysis is beyond the scope of the current paper.

Comparative statics In addition to characterizing the equilibrium, Ferejohn also analyzes comparative statics, finding that voter welfare is increasing in the value of office W and decreasing in the probability of returning to office after being ousted, λ. Both claims are intuitive, but the original proofs no longer go through given the correct characterization of equilibrium. Nonetheless, the comparative statics results do continue to hold, under some assumptions. See the Online Appendix for details.

3 Conclusion

In this paper, we identify and correct an error in the original analysis in Ferejohn. Our solution predicts time-series patterns of politician behavior and political turnover that

differ qualitatively from those implied by the original solution. These patterns translate into a set of testable implications that ought to inform empirical studies of accountability, and our exercise shows that Ferejohn's model yields even richer conceptual insights than previously thought.

Specifically, the voter's increasing leniency toward politicians has a flavor of so-called democratic fatigue (or backsliding). Declining accountability in existing models of democratic fatigue arises as the result of a changing institutional environment due to the actions of strategic politicians (Luo and Przeworski, 2023; Howell, Shepsle and Wolton, 2023), or as the result of changing voter preferences (Grillo and Prato, 2023). Here, "democratic fatigue" arises in spite of the underlying environment being stationary. Another point of distinction is that democratic fatigue is typically viewed in a negative light, i.e., voters are worse off because of it. Here, the incumbent is less accountable as time progresses purely because this is the most efficient way for the voter to incentivize her. The voter is better off relative to a situation where accountability remains stable! Putting it simply, democratic fatigue is the result of the voter's "deal with the devil."

Notably, the voter-optimal retention rule in ALR also features incentive backloading for intuitively similar reasons as our corrected solution of Ferejohn's model. The two main departures of ALR's setting from Ferejohn's are: 1) ALR assume the state variable θ_t takes binary values rather than being drawn from a continuous distribution, and 2) ALR allow the voter to choose fully history-dependent reelection rules, as opposed to rules that condition only on current performance and calendar time (cf. Footnote 4). In their setting, the voter can, and indeed will, base reelection decisions on the incumbent's entire performance history. At a minimum, in all non-trivial cases, the voter-optimal retention rule will condition on the incumbent's seniority.[12] The upshot is that Ferejohn prohibits the voter from choosing the class of strategies that would be considered optimal by ALR. For this reason, ALR's results, while suggestive, do not have a direct bearing on

[12]Seniority is distinct from calendar time since the former is reset whenever the incumbent is replaced.

the (correct) solution of the exact problem laid out by Ferejohn.

References

Acemoglu, Daron, Michael Golosov, and Aleh Tsyvinski. 2008. "Political Economy of Mechanisms." *Econometrica*, 76(3): 619–641.

Acharya, Avidit, Elliot Lipnowski, and Joao Ramos. 2022. "Political Accountability Under Moral Hazard." *working paper*.

Anesi, Vincent, and Peter Buisseret. 2022. "Making Elections Work: Accountability with Selection and Control." *American Economic Journal: Microeconomics*, 14(4): 616–44.

Barro, Robert J. 1973. "The Control of Politicians: an Economic Model." *Public Choice*, 19–42.

Burdett, Ken, and Melvyn Coles. 2003. "Equilibrium wage-tenure contracts." *Econometrica*, 71(5): 1377–1404.

Duggan, John, and César Martinelli. 2017. "The Political Economy of Dynamic Elections: Accountability, Commitment, and Responsiveness." *Journal of Economic Literature*, 55(3): 916–84.

Ferejohn, John. 1986. "Incumbent Performance and Electoral Control." *Public Choice*, 5–25.

Grillo, Edoardo, and Carlo Prato. 2023. "Reference points and democratic backsliding." *American Journal of Political Science*, 67(1): 71–88.

Holmström, Bengt. 1979. "Moral Hazard and Observability." *The Bell Journal of Economics*, 74–91.

Howell, William G, Kenneth A Shepsle, and Stephane Wolton. 2023. "Executive absolutism: the dynamics of authority acquisition in a system of separated powers." *Quarterly Journal of Political Science*, 18(2): 243–275.

Lazear, Edward P. 1981. "Agency, Earnings Profiles, Productivity, and Hours Restrictions." *The American Economic Review*, 71(4): 606–620.

Luo, Zhaotian, and Adam Przeworski. 2023. "Democracy and its Vulnerabilities: Dynamics of Democratic Backsliding." *Quarterly Journal of Political Science*, 18(1): 105–130.

Meirowitz, Adam. 2007. "Probabilistic Voting and Accountability in Elections with Uncertain Policy Constraints." *Journal of Public Economic Theory*, 9(1): 41–68.

Ray, Debraj. 2002. "The Time Structure of Self-Enforcing Agreements." *Econometrica*, 70(2): 547–582.

Appendix

A Proofs

Proof of Proposition A. From Equation (2), we have that $\frac{\partial \theta_t^*(\mathbf{K})}{\partial K_t} = \frac{\theta_t^*(\mathbf{K})}{K_t}$. We can then write

$$\frac{\partial U}{\partial K_t} = \delta^t(1 - F(\theta_t^*(\mathbf{K}))) - \delta^t f(\theta_t^*(\mathbf{K}))\theta_t^*(\mathbf{K}) + \sum_{s=0}^{t-1} \delta^s K_s(-f(\theta_s^*(\mathbf{K})))\frac{\partial \theta_s^*(\mathbf{K})}{\partial K_t}. \tag{5}$$

For $t = 0$, the third term vanishes, so setting $\frac{\partial U}{\partial K_0} = 0$ implies

$$0 = 1 - F(\theta_0^*) - f(\theta_0^*)\theta_0^*, \tag{6}$$

which is equivalent to (4) due to (2).

For $t > 0$, the two conditions will differ unless the third term were to vanish. To calculate $\frac{\partial \theta_s^*(\mathbf{K})}{\partial K_t}$ ($s < t$), observe that θ_s^* depends on K_t only through V_{s+1}^I, which depends on K_t only through V_{s+2}^I, ..., which depends on K_t only through V_t^I. (Here we leverage our assumption that $\lambda = 0$, as otherwise both V_{s+1}^I and V_{s+1}^O, etc. would matter.) Denote $V_t^I = V_t$. Then we can write

$$\frac{\partial \theta_s^*(\mathbf{K})}{\partial K_t} = \frac{\partial \theta_s^*}{\partial V_{s+1}} \left(\prod_{l=s+1}^{t-1} \frac{\partial V_l}{\partial V_{l+1}} \right) \frac{\partial V_t}{\partial K_t}. \tag{7}$$

Differentiating (2), which under the assumption $\lambda = 0$ becomes $\theta_t^* = \frac{K_t}{\phi^{-1}(\delta V_{t+1})}$, we obtain

$$\frac{\partial \theta_s^*}{\partial V_{s+1}} = -\frac{K_s}{(\phi^{-1}(\delta V_{s+1}))^2}(\phi^{-1})'(\delta V_{s+1})\delta = -\delta \frac{\theta_s^{*2}}{K_s} \frac{1}{\phi'\left(\frac{K_s}{\theta_s^*}\right)},$$

where in the last step we have again substituted (2) and used the inverse function theorem. Next we aim to calculate the remaining derivatives in (7). By definition of the

officeholder's payoffs,

$$V_t = W + \int_{\theta_t^*}^{m} \left[\delta V_{t+1} - \phi\left(\frac{K_t}{\theta}\right) \right] f(\theta) d\theta, \tag{8}$$

whence

$$\frac{\partial V_t}{\partial V_{t+1}} = \delta(1 - F(\theta_t^*)) \tag{9}$$

$$\frac{\partial V_t}{\partial K_t} = -\int_{\theta_t^*}^{m} f(\theta)\phi'\left(\frac{K_t}{\theta}\right)\frac{1}{\theta} d\theta. \tag{10}$$

Then we can rewrite (7) and (5) respectively as

$$\frac{\partial \theta_s^*(\mathbf{K})}{\partial K_t} = \delta^{t-s}\frac{\theta_s^{*2}}{K_s}\frac{1}{\phi'\left(\frac{K_s}{\theta_s^*}\right)} \prod_{l=s+1}^{t-1}(1 - F(\theta_l^*)) \int_{\theta_t^*}^{m} f(\theta)\phi'\left(\frac{K_t}{\theta}\right)\frac{1}{\theta} d\theta \tag{11}$$

$$\begin{aligned} \delta^{-t}\frac{\partial U}{\partial K_t} &= 1 - F(\theta_t^*) - f(\theta_t^*)\theta_t^* \\ &\quad - \sum_{s=0}^{t-1} f(\theta_s^*)\frac{\theta_s^{*2}}{\phi'\left(\frac{K_s}{\theta_s^*}\right)} \prod_{l=s+1}^{t-1}(1 - F(\theta_l^*)) \int_{\theta_t^*}^{m} f(\theta)\phi'\left(\frac{K_t}{\theta}\right)\frac{1}{\theta} d\theta. \end{aligned} \tag{12}$$

Recall that $\theta^* \equiv \theta_t^*(\overline{\mathbf{K}})$ solves the equation $0 = 1 - F(\theta^*) - f(\theta^*)\theta^*$ (this follows from our discussion in 2.1 and is also shown on p. 16 of the original paper). Hence, when evaluating $\frac{\partial U}{\partial K_t}$ at $\mathbf{K} = \overline{\mathbf{K}}$, the first line of the expression in (12) vanishes. The second line is an empty summation for $t = 0$, but clearly negative for all $t > 0$. This proves the first part of the proposition.

For the second, let $\mathbf{K}$ be an optimal retrospective rule, and suppose it is stationary, i.e., $K_t \equiv K$ for all t. Since the officeholder's problem is stationary, it follows that $\theta_t^* \equiv \tilde{\theta}$ is constant in t. Since an optimal rule must satisfy $\frac{\partial U}{\partial K_0} = 0$, (12) implies that $0 = 1 - F(\tilde{\theta}) - f(\tilde{\theta})\tilde{\theta}$. (Alternatively, we could have $\frac{\partial U}{\partial K_0} \leq 0$ and $K_0 = 0$ in a corner solution. But then, as $\mathbf{K}$ is stationary, $K_t = 0$ for all t, which is clearly suboptimal.) This again implies that $\frac{\partial U}{\partial K_t} < 0$ for all $t > 0$, so $\mathbf{K}$ is suboptimal (unless $K_t = 0$ for all $t > 0$, which

violates either the optimality or the stationarity of **K**). □

Proof of Corollary B. Let **K** be an optimal retrospective rule. Note that, if $K_0 = 0$, the voter could do better with a rule $\tilde{\mathbf{K}}$ given by $\tilde{K}_t \equiv K_{t+1}$, as all the voter's (positive) payoffs are moved forward by one period.[13] Thus $K_0 > 0$. Then, by (12), θ_0^* must solve the equation $0 = 1 - F(\theta) - f(\theta)\theta$, so $\theta_0^* = \frac{1-F(\theta_0^*)}{f(\theta_0^*)}$. For $t > 0$, if $K_t > 0$, we must have $\frac{\partial U}{\partial K_t} = 0$, whence $1 - F(\theta_t^*) - f(\theta_t^*)\theta_t^* > 0$ by (12). Then $\theta_t^* < \frac{1-F(\theta_t^*)}{f(\theta_t^*)}$. Under the assumption that $\frac{1-F(x)}{f(x)}$ is decreasing in x, this implies $\theta_t^* < \theta_0^*$. On the other hand, if $K_t = 0$, then $\theta_t^* = 0 < \theta_0^*$ by (2). □

Proof of Proposition C. The proof has four main parts. First, we simplify and rearrange (12) to obtain a recursive system of equations which pins down a unique sequence of thresholds $(\theta_t^*)_{t\geq 0}$ satisfying (12) for all t. Second, we use various properties of the system to show that this sequence is decreasing. Third, we show that the optimal rule cannot involve corner solutions (in particular, θ_t^* must be positive for all t), so it must indeed satisfy (12) for all t, and hence it must be the sequence we have characterized. Fourth, we show that the optimal cutoffs $(K_t)_{t\geq 0}$ can be written in terms of the thresholds $(\theta_t^*)_{t\geq 0}$, and use the resulting expressions to show that the sequence $(K_t)_{t\geq 0}$ is also decreasing.

Pinning down θ_t^*. Under the assumption $\phi(a) \equiv a$, (12) simplifies to

$$0 = 1 - F(\theta_t^*) - f(\theta_t^*)\theta_t^* - \sum_{s=0}^{t-1} f(\theta_s^*)\theta_s^{*2} \prod_{l=s+1}^{t-1} (1 - F(\theta_l^*)) \int_{\theta_t^*}^{m} \frac{f(\theta)}{\theta} d\theta$$

for all t such that $K_t > 0$. Note that, while we are differentiating with respect to K_t, the resulting conditions can all be stated in terms of the θ_t^*. This should be interpreted as: a sequence of performance targets is optimal for the voters iff it **induces** the officeholder to only exert effort for states above these thresholds.

[13] This is a strict improvement unless $K_t \equiv 0$ is optimal, which is impossible.

Under the assumption $\theta_t \sim U[0,1]$, (12) further simplifies to

$$0 = 1 - 2\theta_t^* + \ln(\theta_t^*) \sum_{s=0}^{t-1} \theta_s^{*2} \prod_{l=s+1}^{t-1} (1-\theta_l^*).$$

Let $A_t = \sum_{s=0}^{t-1} \theta_s^{*2} \prod_{l=s+1}^{t-1}(1-\theta_l^*)$. Then we can calculate $(\theta_t^*)_{t\geq 0}$ and $(A_t)_{t\geq 0}$ as the solutions to a recursive system given by $A_0 = 0$ and

$$0 = 1 - 2\theta_t^* + A_t \ln(\theta_t^*) \tag{13}$$

$$A_{t+1} = (1-\theta_t^*)A_t + \theta_t^{*2}. \tag{14}$$

For $A_t > 0$, (13) may in general have two values of θ_t^* that serve as solutions. If so, note that, because the right-hand side represents $\frac{\partial U}{K_t}$, and it is concave in θ_t^* (which is proportional to K_t), the lower solution would correspond to a local minimum of U (as $\frac{\partial U}{K_t}$ crosses zero from below), while the higher one would correspond to a local maximum (as $\frac{\partial U}{K_t}$ crosses zero from above). So the only the higher solution is valid.

θ_t^* is decreasing in t. Let $T : [0, \overline{x}] \to \mathbb{R}$ be a function implicitly defined as follows: for each $x \in [0, \overline{x}]$, $y = T(x)$ is the highest solution to the equation $0 = 1 - 2y + x\ln(y)$. $\overline{x}$ is the highest value of x for which the equation has any solutions $y \in [0,1]$.

Lemma 1. *There is a unique valid solution to the system (13–14). In it, $A_t \nearrow A_\infty$ and $\theta_t^* \searrow \theta_\infty^*$. Moreover, $\theta_\infty^* = A_\infty$.*

Proof. Note first that T must be decreasing: as x increases, the expression $1 - 2y + x\ln(y)$ decreases for all $y \in (0,1)$. Since $1 - 2y + x\ln(y)$ is decreasing around the highest solution $y = T(x)$, y must decrease to compensate.

Let S be a mapping defined by $S(A) = (1 - T(A))A + T(A)^2$. In particular, $A_{t+1} = S(A_t)$ for all t. By the implicit function theorem, $T'(A) = -\frac{\ln(T)}{-2+\frac{A}{T}} = \frac{\frac{2T-1}{A}}{2-\frac{A}{T}} = \frac{T}{A}\frac{2T-1}{2T-A}$.

Then

$$S' = -T'A + (1-T) + 2TT' = T'(2T-A) + 1 - T = \frac{T}{A}\frac{2T-1}{2T-A}(2T-A) + 1 - T$$
$$= \frac{T}{A}(2T-1) + 1 - T = \frac{T}{A}A\ln(T) + 1 - T = T\ln(T) + 1 - T.$$

This expression is positive for all $T \in (0,1)$: indeed, its derivative with respect to T is $\ln(T) < 0$ for all $T \in (0,1)$, and its value at $T = 1$ is 0. Hence $S'(A) > 0$ for all A in the domain of T.

Since $A_0 = 0$, $\theta_0^* = \frac{1}{2}$, so $A_1 = \frac{1}{4} > 0 = A_0$. Then, because S is increasing, $A_2 = S(A_1) > S(A_0) = A_1$. Iterating, $A_{t+1} > A_t$ for all t. $(A_t)_{t\geq 0}$ must then converge to a limit A.

Moreover, because T is decreasing, and $\theta_t^* = T(A_t)$, the fact that $(A_t)_{t\geq 0}$ is increasing implies that $(\theta_t^*)_{t\geq 0}$ is decreasing, towards some limit θ_∞^*.

As $t \to \infty$, $A_{t+1} \to A_\infty$, while $(1-\theta_t^*)A_t + \theta_t^{*2} \to (1-\theta_\infty^*)A_\infty + \theta_\infty^{*2}$. Since $\theta_\infty^* > 0$ (otherwise (13) cannot hold for high t), (14) implies $A_\infty = \theta_\infty^*$. □

It follows that $\theta_\infty^* = T(\theta_\infty^*)$, so θ_∞^* solves the equation $1 - 2\theta_\infty^* + \theta_\infty^* \ln(\theta_\infty^*) = 0$. As the left-hand side of this equation is decreasing over $\left(0, \frac{1}{2}\right)$, and it goes from positive to negative over this interval, it has a unique solution in this range, which is approximately 0.318. With this observation, we conclude the proof of Proposition C.2, and the first half of Proposition C.3.

No corner solutions. Because there is a unique solution to the system of FOCs from (12), the local optimum we have found must be *the* solution to the voter's problem, if said solution is interior. Next, we rule out optimal retrospective rules involving corner solutions, i.e., $K_t = 0$ and $\frac{\partial U}{\partial K_t} \leq 0$ for some values of t.

Suppose for the sake of contradiction that such a rule $\tilde{\mathbf{K}}$ is optimal, and that t_0 is the first period in which $\frac{\partial U(\tilde{\mathbf{K}})}{\partial K_t} < 0$. Then the sequence $(\tilde{\theta}_t^*)_{t\geq 0}$ solves (13–14) for $t < t_0$. Let us calculate the impact on the voter's welfare of modifying $\tilde{\mathbf{K}}$ to a new rule $\hat{\mathbf{K}}$ such that the

sequence $(\hat{\theta}^*_t)_{t\geq 0}$ solves (13–14) for $t \leq t_0$, and $\hat{\theta}^*_t \equiv \tilde{\theta}^*_t$ for $t > t_0$ (i.e., shifting $\tilde{\theta}^*_t$ from 0 to the interior critical point of (12) and leaving other thresholds unchanged).

To do this, define $\tilde{\mathbf{K}}(\theta)$ to be a rule inducing $\theta^*_{t_0}(\tilde{\mathbf{K}}(\theta)) = \theta$ and $\theta^*_t(\tilde{\mathbf{K}}(\theta)) = \tilde{\theta}^*_t = \hat{\theta}^*_t$ for other t.[14] In particular, $\hat{\mathbf{K}} = \tilde{\mathbf{K}}(\theta^*_{t_0})$. Then

$$\frac{dU(\tilde{\mathbf{K}}(\theta))}{d\theta} = \sum_{t=0}^{t_0} \frac{\partial U(\tilde{\mathbf{K}}(\theta))}{\partial K_t} \frac{d\tilde{K}_t(\theta)}{d\theta} = \frac{\partial U(\tilde{\mathbf{K}}(\theta))}{\partial K_{t_0}} \delta V_{t_0+1},$$

where we have used that $\frac{\partial U(\tilde{\mathbf{K}}(\theta))}{\partial K_t} = 0$ for $t < t_0$ because $\tilde{\mathbf{K}}(\theta)$ solves (12) for $t < t_0$; that (2) simplifies to $\theta^*_{t_0} = \frac{K_{t_0}}{\delta V_{t_0+1}}$, so $\tilde{K}_{t_0}(\theta) = \theta \delta V_{t_0+1}$; and that the voter's payoffs after t_0, as well as V_{t_0+1}, are constant in θ because $\tilde{\mathbf{K}}(\theta)$ is independent of θ for $t > t_0$. Hence

$$\begin{aligned} U(\hat{\mathbf{K}}) - U(\tilde{\mathbf{K}}) &= \delta V_{t_0+1} \int_0^{\theta^*_{t_0}} \frac{\partial U(\tilde{\mathbf{K}}(\theta))}{\partial K_{t_0}} d\theta = \delta V_{t_0+1} \int_0^{\theta^*_{t_0}} (1 - 2\theta + A_{t_0} \ln(\theta)) d\theta \\ &= \delta V_{t_0+1} \left[\theta^*_{t_0} - \theta^{*2}_{t_0} + A_{t_0}(\theta^*_{t_0} \ln(\theta^*_{t_0}) - \theta^*_{t_0})\right] = \delta V_{t_0+1} \theta^*_{t_0} \left[\theta^*_{t_0} - A_{t_0}\right], \end{aligned}$$

applying (13) in the last step. Hence $U(\hat{\mathbf{K}}) > U(\tilde{\mathbf{K}})$, as Lemma 1 implies $\theta^*_{t_0} > A_{t_0}$, which shows that $\tilde{\mathbf{K}}$ was not optimal.

K_t is decreasing in t. Next, we provide a characterization of the optimal retrospective rule $\mathbf{K} = (K_t)_{t\geq 0}$ that induces the optimal sequence of thresholds $(\theta^*_t)_{t\geq 0}$ characterized in Lemma 1.

Recall that, under the assumptions $\lambda = 0$, $\phi(a) \equiv a$ and $\theta_t \sim U[0,1]$, (2) simplifies to $\theta^*_t = \frac{K_t}{\delta V_{t+1}}$, and (8), the incumbent's Bellman equation, simplifies to

$$\begin{aligned} V_t &= W + \int_{\theta^*_t}^{m} \left[\phi\left(\frac{K_t}{\theta^*_t}\right) - \phi\left(\frac{K_t}{\theta}\right)\right] f(\theta) d\theta = \int_{\theta^*_t}^{1} \left[\frac{K_t}{\theta^*_t} - \frac{K_t}{\theta}\right] d\theta \\ &= W + (1 - \theta^*_t)\frac{K_t}{\theta^*_t} + K_t \ln(\theta^*_t). \end{aligned} \tag{15}$$

[14] We provide tools below to produce rules inducing any sequence of thresholds, in particular allowing the construction of $\tilde{\mathbf{K}}(\theta)$. Indeed, reversing the steps between (15) and (18) yields that, for any sequence of thresholds $(\theta^*_t)_{t\geq 0}$, the sequence of K_t as defined by (18) makes $(\theta^*_t)_{t\geq 0}$ the equilibrium thresholds.

Rearranging (2) and (15),

$$\frac{K_t}{\delta\theta_t^*} = V_{t+1} = W + K_{t+1}\left(\frac{1-\theta_{t+1}^*}{\theta_{t+1}^*} + \ln(\theta_{t+1}^*)\right)$$
$$\frac{K_t}{\theta_t^*} = \delta W + \delta\left(1 - \theta_{t+1}^* + \theta_{t+1}^*\ln(\theta_{t+1}^*)\right)\frac{K_{t+1}}{\theta_{t+1}^*}. \tag{16}$$

Iteratively applying (16), we obtain, for any $\tau > t$,

$$\frac{K_t}{\theta_t^*} = \delta W \sum_{s=t}^{\tau-1}\delta^{s-t}\prod_{l=t+1}^{s}(1-\theta_l^*+\theta_l^*\ln(\theta_l^*)) + \frac{K_\tau}{\theta_\tau^*}\delta^{\tau-t}\prod_{l=t+1}^{\tau}(1-\theta_l^*+\theta_l^*\ln(\theta_l^*)). \tag{17}$$

As argued above, $1 - x + x\ln(x) \in [0,1]$ for $x \in (0,1]$; and we have shown $\theta_t^* \in (0, \frac{1}{2}]$ for all t. Hence the expression $\sum_{s=t}^{\tau-1}\delta^{s-t}\prod_{l=t+1}^{s}(1-\theta_l^*+\theta_l^*\ln(\theta_l^*))$ increases towards a limit between zero and $\frac{1}{1-\delta}$ as $\tau \to \infty$. Moreover, $\delta^{\tau-t}\prod_{l=t+1}^{\tau}(1-\theta_l^*+\theta_l^*\ln(\theta_l^*)) \leq \delta^{\tau-t}$ which converges to zero as $\tau \to \infty$, and $\frac{K_\tau}{\theta_\tau^*} = \delta V_{\tau+1} \leq \frac{\delta W}{1-\delta}$ for all τ by (2). Thus, taking the limit as $\tau \to \infty$, we obtain

$$K_t = \delta\theta_t^* W \sum_{s=t}^{\infty}\delta^{s-t}\prod_{l=t+1}^{s}(1-\theta_l^*+\theta_l^*\ln(\theta_l^*)). \tag{18}$$

Next, we will show that K_t is decreasing in t, towards a limit K_∞. (This will also imply that V_t is increasing in t.)

We aim to show that, for all t,

$$K_t > K_{t+1}$$
$$\iff \theta_t^*\sum_{s=t}^{\infty}\delta^{s-t}\prod_{l=t+1}^{s}(1-\theta_l^*+\theta_l^*\ln(\theta_l^*)) > \theta_{t+1}^*\sum_{s=t+1}^{\infty}\delta^{s-(t+1)}\prod_{l=t+2}^{s}(1-\theta_l^*+\theta_l^*\ln(\theta_l^*)).$$

It suffices to show that the inequality holds term by term, that is, that for all $s \geq t$,

$$\theta_t^* \delta^{s-t} \prod_{l=t+1}^{s} (1 - \theta_l^* + \theta_l^* \ln(\theta_l^*)) > \theta_{t+1}^* \delta^{s+1-(t+1)} \prod_{l=t+2}^{s+1} (1 - \theta_l^* + \theta_l^* \ln(\theta_l^*))$$
$$\iff \theta_t^*(1 - \theta_{t+1}^* + \theta_{t+1}^* \ln(\theta_{t+1}^*)) > \theta_{t+1}^*(1 - \theta_{s+1}^* + \theta_{s+1}^* \ln(\theta_{s+1}^*)).$$

Recall that $1 - x + x\ln(x)$ is positive and decreasing in x for $x \in (0,1)$, and that θ_{s+1}^* is decreasing in s. It follows that the tightest case, in which the right-hand side is as high as possible, is as $s \to \infty$, so it is enough to show that

$$\theta_t^*(1 - \theta_{t+1}^* + \theta_{t+1}^* \ln(\theta_{t+1}^*)) > \theta_{t+1}^*(1 - \theta_\infty^* + \theta_\infty^* \ln(\theta_\infty^*))$$
$$\iff \theta_t^* \left(\frac{1 - \theta_{t+1}^*}{\theta_{t+1}^*} + \ln(\theta_{t+1}^*) \right) > \theta_\infty^*$$

for all t, using that $1 - 2\theta_\infty^* + \theta_\infty^* \ln(\theta_\infty^*) = 0$.

We prove this in two steps. First, we check manually that this inequality holds for $t = 0$ and $t = 1$. We obtain $\theta_0^* = \frac{1}{2}$, $\theta_1^* \approx 0.3786$, $\theta_2^* \approx 0.3380$ and $\theta^* \approx 0.3178$, so the cases $t = 0$ and $t = 1$ boil down to $0.3350 > 0.3178$ and $0.3308 > 0.3178$.

Next, we show something more general than needed: that, given a value of A, and letting $\theta = T(A)$, $\tilde{A} = S(A)$ and $\tilde{\theta} = T(\tilde{A}) = T(S(A))$, we have

$$\theta \left(\frac{1 - \tilde{\theta}}{\tilde{\theta}} + \ln(\tilde{\theta}) \right) > \theta_\infty^*$$

whenever A is such that $\theta \in (\theta_\infty^*, \theta_2^*]$, i.e., for $A \in [A_2, \theta_\infty^*)$. To show this, rewrite the left-hand side of the inequality as $T(A)\left(\frac{1-T(S(A))}{T(S(A))} + \ln(T(S(A)))\right) =: Z(A)$. Note that

$Z(\theta^*_\infty) = \theta^*_\infty$. Then it is enough to show that $Z'(A) < 0$ for $A \in (A_2, \theta^*_\infty)$. Now

$$\begin{aligned}
Z'(A) &= T'(A)\left(\frac{1-\tilde\theta}{\tilde\theta} + \ln(\tilde\theta)\right) + T(A)\left(-\frac{1}{\tilde\theta^2} + \frac{1}{\tilde\theta}\right) T'(S(A))S'(A) \\
&= -\frac{\ln(\theta)}{-2+\frac{A}{\theta}}\left(\frac{1-\tilde\theta}{\tilde\theta} + \ln(\tilde\theta)\right) + \theta\left(-\frac{1}{\tilde\theta^2} + \frac{1}{\tilde\theta}\right)\left(-\frac{\ln(\tilde\theta)}{-2+\frac{\tilde A}{\tilde\theta}}\right)(1-\theta+\theta\ln(\theta)) \\
&= \frac{\ln(\theta)}{2-\frac{2\theta-1}{\theta\ln(\theta)}}\left(\frac{1-\tilde\theta}{\tilde\theta} + \ln(\tilde\theta)\right) + \theta\left(-\frac{1}{\tilde\theta^2} + \frac{1}{\tilde\theta}\right)\frac{\ln(\tilde\theta)}{2-\frac{2\tilde\theta-1}{\tilde\theta\ln(\tilde\theta)}}(1-\theta+\theta\ln(\theta)),
\end{aligned}$$

which is negative iff

$$\begin{aligned}
&\theta\left(1-\frac{1}{\tilde\theta}\right)\frac{\ln(\tilde\theta)}{2-\frac{2\tilde\theta-1}{\tilde\theta\ln(\tilde\theta)}}(1-\theta+\theta\ln(\theta)) < -\frac{\ln(\theta)}{2-\frac{2\theta-1}{\theta\ln(\theta)}}\left(1-\tilde\theta+\tilde\theta\ln(\tilde\theta)\right) \\
&\quad\iff \theta\frac{2-\frac{2\theta-1}{\theta\ln(\theta)}}{-\ln(\theta)}(1-\theta+\theta\ln(\theta)) < \frac{\tilde\theta}{1-\tilde\theta}\left(1-\tilde\theta+\tilde\theta\ln(\tilde\theta)\right)\frac{2-\frac{2\tilde\theta-1}{\tilde\theta\ln(\tilde\theta)}}{-\ln(\tilde\theta)} \qquad (19)
\end{aligned}$$

where we have used that $\frac{2-\frac{2x-1}{x\ln(x)}}{-\ln(x)}$ is positive for $x \in (0.2, 1]$ (note that $2-\frac{2x-1}{x\ln(x)} > 0$ iff $2x\ln(x) - 2x + 1 < 0$; this function is decreasing in $x \in (0,1)$ and vanishes at $x \approx 0.1867$).

Define $R(x) = x\frac{2-\frac{2x-1}{x\ln(x)}}{-\ln(x)}(1-x+x\ln(x)) = \frac{f(x)g(x)}{h(x)}$ where $f(x) = 2x - \frac{2x-1}{\ln(x)}$, $g(x) = 1-x+x\ln(x)$, $h(x) = -\ln(x)$. We will show that $R(x)$ is increasing in x for $x \in [0.3178, 0.3380]$.

$$\begin{aligned}
R'(x) &= R(x)\left(\frac{f'}{f} + \frac{g'}{g} - \frac{h'}{h}\right) \\
&= R(x)\left(\frac{2-\frac{2\ln(x)-\frac{2x-1}{x}}{\ln(x)^2}}{2x-\frac{2x-1}{\ln(x)}} + \frac{\ln(x)}{1-x+x\ln(x)} - \frac{1}{x\ln(x)}\right) \\
&= R(x)\left(\frac{2\ln(x)-2+\frac{2x-1}{x\ln(x)}}{2x\ln(x)-2x+1} + \frac{\ln(x)}{1-x+x\ln(x)} - \frac{1}{x\ln(x)}\right) \\
&= R(x)\left(\frac{2\ln(x)}{2x\ln(x)-2x+1} + \frac{\ln(x)}{1-x+x\ln(x)} - \frac{2}{x\ln(x)}\right).
\end{aligned}$$

Clearly $\frac{-2}{x\ln(x)} > 0$ for $x \in (0,1)$. In addition, for $x \in (0.2, 1)$,

$$\begin{aligned}\frac{2\ln(x)}{2x\ln(x)-2x+1}+\frac{\ln(x)}{1-x+x\ln(x)}&>0\\ \iff \frac{2}{2x\ln(x)-2x+1}&<-\frac{1}{1-x+x\ln(x)}\\ \iff 3-4x+4x\ln(x)&>0,\end{aligned}$$

where we have used that $2x\ln(x)-2x+1<0$ for $x\in(0.2,1)$. Since $3-4x+4x\ln(x)$ is decreasing in x for $x\in(0,1)$ and vanishes at $x\approx 0.3824$, it is positive for all $x\in[0.3178,0.3380]$. Thus $R'(x)>0$ in this interval, as we wanted to show.

Now, because R is increasing, and $\frac{1}{1-x}$ is increasing in x, it is enough to show that (19) holds even if we replace θ with θ_2^* and $\tilde{\theta}$ with θ_∞^*. In that case we obtain

$$\begin{aligned}\theta_2^*\frac{2-\frac{2\theta_2^*-1}{\theta_2^*\ln(\theta_2^*)}}{-\ln(\theta_2^*)}(1-\theta_2^*+\theta_2^*\ln(\theta_2^*))&<0.103,\\ \frac{\theta_\infty^*}{1-\theta_\infty^*}\left(1-\theta_\infty^*+\theta_\infty^*\ln(\theta_\infty^*)\right)\frac{2-\frac{2\theta_\infty^*-1}{\theta_\infty^*\ln(\theta_\infty^*)}}{-\ln(\theta_\infty^*)}&>0.129\end{aligned}$$

which proves the result.

Taking the limit of (18) as t goes to infinity, and using that $1-2\theta_\infty^*+\theta_\infty^*\ln(\theta_\infty^*)=0$, we obtain

$$K_\infty=\delta\theta_\infty^*W\sum_{s=0}^{\infty}\delta^s\theta_\infty^{*s}=\frac{\delta\theta_\infty^*W}{1-\delta\theta_\infty^*}=\frac{\delta W}{\frac{1}{\theta_\infty^*}-\delta}=\frac{\delta W}{2-\delta-\ln(\theta_\infty^*)},$$

which finishes the proof of Proposition C.3.

Finally we show that $K_0>\overline{K}>K_1$, which is the only missing part of Proposition C.1. Recall that, in Ferejohn (1986)'s stationary solution, $\theta_t^*=\frac{1}{2}$ for all t. Then, by (2), $\overline{K}=\frac{\delta V}{2}$, where $V=W+\int_{\frac{1}{2}}^1\left(\delta V-\frac{\overline{K}}{\theta}\right)d\theta$. Rearranging, $V=\frac{W}{1-\delta\left(\frac{1}{2}+\frac{1}{2}\ln\left(\frac{1}{2}\right)\right)}$ and $\overline{K}=\frac{\delta W}{2-\delta\left(1+\ln\left(\frac{1}{2}\right)\right)}$.

Taking $t = 0$ in Equation 18,

$$K_0 = \delta \frac{1}{2} W \sum_{s=0}^{\infty} \delta^s \prod_{l=1}^{s} (1 - \theta_l^* + \theta_l^* \ln(\theta_l^*)) >$$
$$> \delta \frac{1}{2} W \sum_{s=0}^{\infty} \delta^s \prod_{l=1}^{s} (1 - \theta_0^* + \theta_0^* \ln(\theta_0^*)) = \frac{\delta W}{2} \sum_{s=0}^{\infty} \delta^s \left(\frac{1}{2} + \frac{1}{2} \ln \left(\frac{1}{2} \right) \right)^s = \overline{K},$$

where we have used that $1 - x + x \ln(x)$ is decreasing over $\left(0, \frac{1}{2}\right)$.

On the other hand,

$$K_1 = \delta \theta_1^* W \sum_{s=1}^{\infty} \delta^s \prod_{l=2}^{s} (1 - \theta_l^* + \theta_l^* \ln(\theta_l^*)) <$$
$$< \delta \theta_1^* W \sum_{s=0}^{\infty} \delta^s \prod_{l=1}^{s} (1 - \theta_\infty^* + \theta_\infty^* \ln(\theta_\infty^*)) =$$
$$= \delta \theta_1^* W \sum_{s=0}^{\infty} \delta^s \theta_\infty^{*s} = \frac{\delta \theta_1^* W}{1 - \delta \theta_\infty^*} \leq \frac{0.379 \delta W}{1 - 0.317 \delta}.$$

To show that $K_1 < \overline{K} \approx \frac{\delta W}{2 - 0.306\delta}$, it is enough to check that $0.379 \times (2 - 0.306\delta) < 1 - 0.317\delta$ for all $\delta \in [0, 1]$, which is true.

□

Corollary D. *Suppose $\lambda = 0$, $\phi(a) \equiv a$, and f can be written as follows: $f(\theta) = Ce^{\int_0^\theta \frac{\eta(t)}{t^2} dt}$, for any $C > 0$ and any differentiable function η such that $\frac{\eta(t)}{t^2}$ is bounded and $\eta'(t) > -2$ for all t. Then, in the optimal retrospective rule, the sequence of cutoffs (θ_t^*) decreases in t towards a positive limit θ_∞^*.*

Proof. The same proof goes through as in Proposition C. To see why, note that, for a general density f, the system (13-14) becomes

$$0 = 1 - F(\theta_t^*) - f(\theta_t^*)\theta_t^* - A_t B(\theta_t^*) \tag{20}$$

$$A_{t+1} = (1 - F(\theta_t^*))A_t + f(\theta_t^*)\theta_t^{*2} \tag{21}$$

with $B(x) = \int_x^m \frac{f(\theta)}{\theta} d\theta$. As before, define $T(A)$ to be the largest solution y of $1 - F(y) - f(y)y - AB(y)$. If $f(1) > 0$, which our conditions on f imply, this expression is negative for $y = 1$, so $1 - F(y) - f(y)y - AB(y)$ must cross zero from above at $y = T(A)$, whence $T'(A) < 0$.

Define $S(A) = (1 - F(T(A))A + f(T(A))T(A)^2$. After analogous algebra, $S'(A) = 1 - F(T) - B(T)T$. The expression $1 - F(T) - B(T)T$ has derivative $-B(T) < 0$ with respect to T and evaluates to 0 at $T = 1$, so it is positive for $T < 1$, whence S is increasing. Then, since $A_1 > 0 = A_0$, $(A_t)_{t\geq 0}$ is increasing, and since $\theta_t^* = T(A_t)$ and $T' < 0$, $(\theta_t^*)_{t\geq 0}$ is decreasing.

To finish the proof, we need to verify that $T(A)$ is in fact the U-maximizing value of θ_t^*. We do this in two parts. First, we show that our conditions on f ensure that the right-hand side of (20), as a function of θ_t^*, crosses zero from above only once, so there are no other local maxima of U. Second, we show that choosing $\theta_t^* = 0$ cannot be better than the interior optimum $T(A)$.

For the first part, we will argue that, for an arbitrary value of A, all critical points θ_0 of $1 - F(\theta) - f(\theta)\theta - AB(\theta)$ must be local maxima. Because there must be a local minimum between two local maxima, this will imply that in fact there is at most only one local maximum, whence the expression can only cross zero from above once. Indeed, suppose θ_0 is a critical point, i.e., $-2f(\theta_0) - f'(\theta_0)\theta_0 - AB'(\theta_0) = -2f(\theta_0) - f'(\theta_0)\theta_0 + A\frac{f(\theta_0)}{\theta_0} = 0$. Then $A = 2\theta_0 + \frac{f'(\theta_0)}{f(\theta_0)}\theta_0^2$. θ_0 is a local maximum if

$$\begin{aligned}
&-3f'(\theta_0) - f''(\theta_0)\theta_0 - AB''(\theta_0) < 0 \\
&\quad\iff 3f'(\theta_0) + f''(\theta_0)\theta_0 > \left(2\theta_0 + \frac{f'(\theta_0)}{f(\theta_0)}\theta_0^2\right)\frac{f'(\theta_0)\theta_0 - f(\theta_0)}{\theta_0^2} \\
&\quad\iff 3\frac{f'(\theta_0)}{f(\theta_0)} + \frac{f''(\theta_0)}{f(\theta_0)}\theta_0 > \left(2 + \frac{f'(\theta_0)}{f(\theta_0)}\theta_0\right)\left(\frac{f'(\theta)}{f(\theta)} - \frac{1}{\theta}\right)
\end{aligned}$$

We will verify that this condition holds for *all* θ for the class of densities f we have

specified. Write $\lambda(x) = \int_0^x \frac{\eta(t)}{t^2} dt$, so $f(x) = e^{\lambda(x)}$. Then $f'(x) = \lambda'(x)f(x)$, and $f''(x) = \lambda''(x)f(x) + (\lambda'(x))^2 f(x)$, so we need to verify that

$$\begin{aligned} & 3\lambda'(x) + (\lambda''(x) + (\lambda'(x))^2)x > (2 + \lambda'(x)x)\left(\lambda'(x) - \frac{1}{x}\right) \\ \iff & 3\lambda'(x) + (\lambda''(x) + (\lambda'(x))^2)x > (\lambda'(x))^2 x - \lambda'(x) + 2\lambda'(x) - \frac{2}{x} \\ \iff & 2\lambda'(x) + \lambda''(x)x > -\frac{2}{x} \end{aligned}$$

Since $\lambda'(x) = \frac{\eta(x)}{x^2}$ and $\lambda''(x) = \frac{\eta'(x)x^2 - \eta(x)2x}{x^4}$, the inequality simplifies to $\eta'(x) > -2$, as we assumed.

For the second part, as in our original argument, it is enough to verify that

$$\begin{aligned} & 0 < \int_0^{\theta_t^*} \left[1 - F(\theta) - f(\theta)\theta - A_t \int_\theta^m \frac{f(\tilde{\theta})}{\tilde{\theta}} d\tilde{\theta}\right] d\theta \\ \iff & 0 < \int_0^{\theta_t^*} \left[1 - F(\theta) - f(\theta)\theta\right] d\theta - A_t \int_0^m \frac{f(\theta)}{\theta} \min(\theta, \theta_t^*) d\theta \\ \iff & 0 < \theta(1 - F(\theta))\big|_0^{\theta_t^*} - A_t \left(F(\theta_t^*) + \theta_t^* B(\theta_t^*)\right) \end{aligned}$$

Applying (20), the right-hand side equals $-A_t F(\theta_t^*) + f(\theta_t^*)\theta_t^{*2}$, which by (21) equals $A_{t+1} - A_t$, which is positive by our previous argument. □

Note that we obtain $\theta \sim U[0,1]$, i.e. $f(x) \equiv 1$, by choosing $C = 1$ and $\eta \equiv 0$. Intuitively, the condition $\eta'(x) > -2$ ensures that f is not too steeply concave in any part of its domain. It is equivalent to requiring that $\left(x^2 \frac{f'(x)}{f(x)}\right)' > -2$ for all x.

Online Appendix

B Comparative Statics

Consider now the general case in which the incumbent might come back to power after being ousted. More specifically, as in Ferejohn (1986), assume that, when a politician is out of office, she returns to power with probability $\lambda \in [0,1]$ whenever the new incumbent is ousted.

The following proposition provides a partial analog to Ferejohn's Propositions 4 and 5.

Proposition D (Comparative Statics)**.**

(i) Assume either $\lambda = 0$ or $\phi(a) \equiv a$. Then the voter's payoff U from the optimal retrospective rule is increasing in W. In the latter case, it is exactly proportional to W.

(ii) The voter's maximized payoff U is higher if $\lambda = 0$ than if λ takes any positive value in a neighborhood of 0.

Proof. For part (i), suppose first that $\lambda = 0$. Then, for any rule **K**, it is clear that the officeholder's value function V_t at any t is increasing in W, as her payoff is increasing in W for any fixed strategy she may follow. By (2), it follows that $\theta_t^*(\mathbf{K})$ is a decreasing function of W for all t, and hence $U(\mathbf{K})$ is increasing in W from any rule (in particular the optimal one).

Suppose now instead that $\phi(a) \equiv a$. In this case, the problem faced by the officeholder given a pair $(W, \mathbf{K})$ is homothetic in W and **K**: if $(W, \mathbf{K})$ is multiplied by $\alpha > 0$, the officeholder's set of payoffs attainable by different strategies is also multiplied by α (as multiplying all effort choices by α achieves exactly this, and the process is reversible); thus the optimal payoffs in the continuation at each t, V_t, are also multiplied by α, and

the best-response efforts $a_t(\theta)$ are multiplied by α for all θ, with the thresholds $\theta_t^*(\mathbf{K})$ remaining fixed. It follows that the voter's attainable payoffs are also proportional to W: if W is multiplied by α, she can multiply her equilibrium payoff by α by also scaling the rule $\mathbf{K}$ by the same factor.

For part (ii), let us rewrite the officeholder's Bellman equation (15) for the general case of $\lambda \geq 0$. Her utility when in and out of office, respectively, is

$$V_t^I = W + \int_{\theta_t^*}^{m} \left[\delta V_{t+1}^I - \phi\left(\frac{K_t}{\theta}\right)\right] f(\theta)d\theta + F(\theta_t^*)\delta V_{t+1}^O, \tag{22}$$

$$\begin{aligned} V_t^O &= F(\theta_t^*)(\lambda\delta V_{t+1}^I + (1-\lambda)\delta V_{t+1}^O) + (1-F(\theta_t^*))\delta V_{t+1}^O = \\ &= \delta V_{t+1}^O + F(\theta_t^*)\lambda\delta(V_{t+1}^I - V_{t+1}^O), \end{aligned} \tag{23}$$

where $F(\theta_t^*)$ is the probability that the new incumbent forfeits her position, giving the officeholder a chance λ to return. Denoting $\Delta V_t = V_t^I - V_t^O$, and combining (22) and (23),

$$\begin{aligned} \Delta V_t &= W - \int_{\theta_t^*}^{m} \phi\left(\frac{K_t}{\theta}\right) f(\theta)d\theta + (1-F(\theta_t^*))\delta V_{t+1}^I + F(\theta_t^*)\delta V_{t+1}^O - \delta V_{t+1}^O - F(\theta_t^*)\lambda\delta\Delta V_{t+1} \\ &= W - \int_{\theta_t^*}^{m} \phi\left(\frac{K_t}{\theta}\right) f(\theta)d\theta + (1-(1+\lambda)F(\theta_t^*))\delta\Delta V_{t+1}. \end{aligned} \tag{24}$$

Assume that $\lambda \in [0, \frac{1-\delta}{\delta}]$. This assumption guarantees that $\Delta V_t \geq 0$ for all t.[15] Indeed, $\phi\left(\frac{K_t}{\theta}\right) \leq \delta\Delta V_{t+1}$ for all $\theta \geq \theta_t^*$ by (2), so (24) implies

$$\begin{aligned} \Delta V_t &\geq W - (1-F(\theta_t^*))\delta\Delta V_{t+1} + (1-(1+\lambda)F(\theta_t^*))\delta\Delta V_{t+1} = \\ &= W - \lambda F(\theta_t^*)\delta\Delta V_{t+1} \geq W - \lambda\delta\frac{W}{1-\delta} \geq 0. \end{aligned}$$

Denote by $V_t^I(\mathbf{K},\lambda)$, $V_t^O(\mathbf{K},\lambda)$, $\Delta V_t(\mathbf{K},\lambda)$ the officeholder's payoffs as a function of the

[15] For high values of λ, it is in principle possible to have $\Delta V_t < 0$, so that the officeholder in period $t-1$ might prefer *not* to be reelected. Intuitively, this could happen if K_t is very high, $K_{t'}$ is low for $t' > t$, and λ is high. Then being in power in period t likely means being out of power forever after, whereas being out of power in period t allows the officeholder to come to power in period $t+1$ and stay there for a long time.

parameters **K**, λ. We will now argue that a higher λ weakens the officeholder's incentives for effort:

Lemma 2. *If* $\lambda_0 \leq \frac{1-\delta}{\delta}$, $\Delta V_t(\boldsymbol{K}, \lambda_0) \leq \Delta V_t(\boldsymbol{K}, 0)$ *for all* t, $\boldsymbol{K}$.

Proof. We begin by showing that, if $\lambda_0 \leq \frac{1-\delta}{\delta}$ and $\Delta V_{t+1}(\mathbf{K}, \lambda_0) \leq \Delta V_{t+1}(\mathbf{K}, 0)$, then $\Delta V_t(\mathbf{K}, \lambda_0) \leq \Delta V_t(\mathbf{K}, 0)$.

Let $\tilde{V}_t^I(\mathbf{K}, 0)$ be the officeholder's continuation payoff in period t, in the case $\lambda = 0$, if she followed the optimal strategy for $\lambda = \lambda_0$ (i.e., choosing $\theta_t^* = \theta_t^*(\mathbf{K}, \lambda_0)$). Of course, $\tilde{V}_t^I(\mathbf{K}, 0) \leq V_t^I(\mathbf{K}, 0) = \Delta V_t(\mathbf{K}, 0)$ because this is in general suboptimal. By (24),

$$\begin{aligned}
\tilde{V}_t^I(\mathbf{K}, 0) &= W - \int_{\theta_t^*(\mathbf{K}, \lambda_0)}^{m} \phi\left(\frac{K_t}{\theta}\right) f(\theta) d\theta + (1 - F(\theta_t^*(\mathbf{K}, \lambda_0)))\delta \Delta V_{t+1}(\mathbf{K}, 0) \geq \\
&\geq W - \int_{\theta_t^*(\mathbf{K}, \lambda_0)}^{m} \phi\left(\frac{K_t}{\theta}\right) f(\theta) d\theta + (1 - F(\theta_t^*(\mathbf{K}, \lambda_0)))\delta \Delta V_{t+1}(\mathbf{K}, \lambda_0) \geq \\
&\geq W - \int_{\theta_t^*(\mathbf{K}, \lambda_0)}^{m} \phi\left(\frac{K_t}{\theta}\right) f(\theta) d\theta + (1 - (1 + \lambda_0) F(\theta_t^*(\mathbf{K}, \lambda_0)))\delta \Delta V_{t+1}(\mathbf{K}, \lambda_0) = \\
&= \Delta V_t(\mathbf{K}, \lambda_0),
\end{aligned}$$

where we have used that $\Delta V_{t+1}(\mathbf{K}, \lambda_0) \leq \Delta V_{t+1}(\mathbf{K}, 0)$ by assumption and that $\Delta V_{t+1}(\mathbf{K}, \lambda_0) \geq 0$ because λ_0 is low enough. More generally, the same argument shows that, if $0 < \Delta V_{t+1}(\mathbf{K}, \lambda_0) - \Delta V_{t+1}(\mathbf{K}, 0) = M$, then $\Delta V_t(\mathbf{K}, \lambda_0) - \Delta V_t(\mathbf{K}, 0) \leq \delta M$. Since $\Delta V_{t'} \leq \frac{W}{1-\delta}$ in all cases, we can conclude that either $\Delta V_t(\mathbf{K}, \lambda_0) \leq \Delta V_t(\mathbf{K}, 0)$ (if the same inequality holds for any $t' > t$) or, if not, then

$$\Delta V_t(\mathbf{K}, \lambda_0) - \Delta V_t(\mathbf{K}, 0) \leq \delta^{t'-t}(V_{t'}(\mathbf{K}, \lambda_0) - \Delta V_{t'}(\mathbf{K}, 0)) \leq \delta^{t'-t} \frac{W}{1-\delta}$$

for arbitrarily high t', which also implies the $\Delta V_t(\mathbf{K}, \lambda_0) \leq \Delta V_t(\mathbf{K}, 0)$. □

The result now follows immediately from Lemma 2: if ΔV_t is lower at all t for $\lambda \in (0, \frac{1-\delta}{\delta})$ than for $\lambda = 0$, then any fixed rule **K** extracts less effort from the officeholder in the

former case, and so the voter's payoff must also be lower when comparing the respective optimal rules. □

C Teaching Guide for Proposition C

Ferejohn's model is often the first formal model of accountability taught to graduate students in political science. With that in mind, this Section provides a guide to proving the main claims of Proposition C that should make the analysis digestible for first or second-year graduate students.

(i) Note that the voter's welfare given a retrospective rule **K** can be written as $U(\mathbf{K}) = \sum_{t=0}^{\infty} \delta^t K_t(1 - F(\theta_t^*(\mathbf{K})))$. (Equation 3')

Note that, under the assumptions of Proposition C, this simplifies to $U(\mathbf{K}) = \sum_{t=0}^{\infty} \delta^t K_t(1 - \theta_t^*(\mathbf{K}))$.

(ii) Differentiate the expression for $U(\mathbf{K})$ with respect to K_t for each t to obtain the relevant FOC for each performance threshold:

$$0 = \frac{\partial U}{\partial K_t} = \delta^t(1 - \theta_t^*(\mathbf{K})) - \delta^t K_t \frac{\partial \theta_t^*(\mathbf{K})}{\partial K_t} - \sum_{s=0}^{t-1} \delta^s K_s \frac{\partial \theta_s^*(\mathbf{K})}{\partial K_t}.$$

(iii) Note that Equation 2 reduces to $\theta_t^* = \frac{K_t}{\delta V_{t+1}^I}$. Show that this implies $\frac{\partial \theta_t^*(\mathbf{K})}{\partial K_t} = \frac{\theta_t^*(\mathbf{K})}{K_t}$.

(iii') Note that Equation 8 reduces to

$$V_t^I = W + \int_{\theta_t^*}^{1} \left[\delta V_{t+1}^I - \frac{K_t}{\theta}\right] d\theta = W + \delta V_{t+1}^I(1 - \theta_t^*) + K_t \ln(\theta_t^*).$$

Show that this, combined with Equation 2, implies

$$\frac{\partial \theta_t^*(\mathbf{K})}{\partial K_{t+1}} = -\frac{K_t}{\delta V_{t+1}^{I2}} \ln(\theta_{t+1}^*) = -\theta_t^{*2} \frac{\delta}{K_t} \ln(\theta_{t+1}^*).$$

More generally, for $s < t$, repeated application of Equation 8 yields $\frac{\partial V^I_{s+1}}{\partial V^I_t} = \delta^{t-s-1}(1-\theta^*_{s+1}) \times \ldots \times (1-\theta^*_{t-1})$, so

$$\frac{\partial \theta^*_s(\mathbf{K})}{\partial K_t} = -\theta^{*2}_s \frac{\delta}{K_s}\frac{\partial V^I_{s+1}}{\partial K_t} = -\theta^{*2}_s \frac{\delta}{K_s}\frac{\partial V^I_{s+1}}{\partial V^I_t}\ln(\theta^*_t) = -\theta^{*2}_s \frac{\delta^{t-s}}{K_s}\ln(\theta^*_t)\prod_{l=s+1}^{t-1}(1-\theta^*_l).$$

(iv) Combine (ii), (iii) and (iii') to obtain

$$\begin{aligned} 0 = \frac{\partial U}{\partial K_t} =& \delta^t(1-\theta^*_t) - \delta^t\theta^*_t + \sum_{s=0}^{t-1}\delta^t \ln(\theta^*_t)\theta^{*2}_s \prod_{l=s+1}^{t-1}(1-\theta^*_l) \\ \iff 0 =& 1 - 2\theta^*_t + \sum_{s=0}^{t-1}\ln(\theta^*_t)\theta^{*2}_s \prod_{l=s+1}^{t-1}(1-\theta^*_l) \end{aligned}$$

for all t.

(v) Write down the equations obtained for the first few values of t:

$$\begin{aligned} 0 &= 1 - 2\theta^*_0 \\ 0 &= 1 - 2\theta^*_1 + \ln(\theta^*_1)\theta^{*2}_0 \\ 0 &= 1 - 2\theta^*_2 + \ln(\theta^*_2)(\theta^{*2}_0(1-\theta^*_1) + \theta^{*2}_1) \\ 0 &= 1 - 2\theta^*_3 + \ln(\theta^*_3)(\theta^{*2}_0(1-\theta^*_1)(1-\theta^*_2) + \theta^{*2}_1(1-\theta^*_2) + \theta^{*2}_2) \\ &\ldots \end{aligned}$$

Rewrite the system by defining $A_t = \sum_{s=0}^{t-1}\theta^{*2}_s \prod_{l=s}^{t-1}(1-\theta^*_l)$ to obtain Equations (13)–(14) as shown in the proof of Proposition C:

$$
\begin{aligned}
A_0 &= 0 && 0 = 1 - 2\theta_0^* \\
A_1 &= \theta_0^{*2} + A_0(1-\theta_0^*) = \theta_0^{*2} && 0 = 1 - 2\theta_1^* + \ln(\theta_1^*)A_1 \\
A_2 &= \theta_1^{*2} + A_1(1-\theta_1^*) && 0 = 1 - 2\theta_2^* + \ln(\theta_2^*)A_2 \\
A_3 &= \theta_2^{*2} + A_2(1-\theta_2^*) && 0 = 1 - 2\theta_3^* + \ln(\theta_3^*)A_3 \\
&\dots
\end{aligned}
$$

Convince yourself that this recursive system pins down θ_t^* and A_t for all t. Moreover, defining T implicitly by $0 = 1 - 2T(x) + x\ln(T(x))$ and defining S by $S(x) = (1 - T(x))x + T(x)^2$, convince yourself that $\theta_t^* = T(A_t)$ for all t, and $A_{t+1} = S(A_t)$ for all t.

The steps up to this point cover the preliminary results before Proposition C as well as the first main step of the proof of this proposition ("pinning down θ_t^*"). The next step is to show that the sequence $(\theta_t^*)_{t\geq 0}$ is decreasing in t.

(vi) To prove this result analytically, show by using the definitions of S and T that (a) S is a stricty increasing function; (b) T is a strictly decreasing function; and (c) $A_1 > A_0$. Deduce that $A_{t+1} > A_t$ for all t and hence $\theta_{t+1}^* < \theta_t^*$ for all t.

You may like to check the result numerically. Here are two ways. First, using the recursive system from (v), you may solve numerically for as many elements of the sequence $(\theta_t^*)_{t\geq 0}$ as desired, and check that the sequence is decreasing. Second, you may plot the functions S and T to verify that they are increasing and decreasing, respectively. Both are simple coding exercises.

(vii) Take the limit of Equations (13)–(14) to characterize θ_∞^*.

(viii) The analytical proof that $(K_t)_{t\geq 0}$ is decreasing is involved. The interested reader may follow the argument given in Proposition C.

However, the result is easy to check numerically. Indeed, you only need to follow the logic from Equation (15) up to Equation (18). If you have solved for $(\theta_t^*)_{t\geq 0}$ numerically on a computer, you can use Equation (18) to then compute K_t for as many values of t as you like and check that the sequence is decreasing.

The only complication is that the formula for K_t involves values of θ_s^* for all $s \geq t$. Of course, you can only calculate a finite number of values of θ_s^*. However, if you are using a value of δ not too close to 1, you can approximate K_t arbitrarily well by replacing tail values of θ_s^* for $s >> t$ with θ_∞^*.