A Theory of *Ex Post* Rationalization[*]

Erik Eyster[†], Shengwu Li[‡], and Sarah Ridout[§]

July 16, 2021

Abstract

Human beings attempt to rationalize their past choices, even those that were mistakes in hindsight. We propose a formal theory of this behavior. The theory predicts that agents commit the sunk-cost fallacy. Its model primitives are identified by choice behavior and it yields tractable comparative statics.

1 Introduction

Human beings seek to rationalize their past choices. Upon discovering that, in hindsight, they have made a mistake, people can adapt their attitudes or beliefs instead of conceding the error. In seeking to rationalize yesterday's choices, they distort their choices today. We propose a theory of this behavior.

We start with a motivating example from Thaler (1980).[1]

> **Example 1**: Bob pays $100 for a ticket to a basketball game to be played 60 miles from his home. On the day of the game there is a snowstorm. He decides to go anyway. If the ticket had been free-of-charge, he would have stayed home.

Bob has committed the sunk-cost fallacy. It is not worth going to the basketball game during a snowstorm. In hindsight, it was a mistake to have bought the ticket. But if Bob goes to the game, then he can avoid acknowledging the

[*] We thank Mohammad Akbarpour, Ben Golub, Matthew Rabin, Philipp Strack, Bruno Strulovici, and Tomasz Strzalecki for valuable comments. All errors remain our own.

[†] UC Santa Barbara, erikeyster@ucsb.edu

[‡] Harvard University, shengwu_li@fas.harvard.edu

[§] Harvard University, ridout@g.harvard.edu

[1] The ticket cost $40 in Thaler's example; we have raised the price due to inflation.

mistake, by exaggerating his enthusiasm for basketball or by downplaying the hazards of driving through a snowstorm. If he stays home, then he is inarguably worse off than if he had not bought a ticket in the first place.

There are two key ingredients for this behavior. First, Bob must have made a choice that was an *ex post* mistake. Hence, this modified example is far-fetched:

> **Example 2**: Bob receives a free ticket to a basketball game and loses \$100 due to an unusually high utility bill. On the day of the game there is a snowstorm. He decides to go anyway. If he had not lost the \$100, he would have stayed home.

Second, there must be plausible preferences that, if adopted, would justify Bob's earlier decision. To illustrate this, we replace the physical consequences in Example 1 with monetary gains and losses.

> **Example 3**: Bob pays \$100 for a financial option that can only be exercised on the day of the basketball game. It yields \$180 if exercised in good weather and *loses* \$20 if exercised in a snowstorm. On the day of the game there is a snowstorm. He decides to exercise the option anyway, for a net loss of \$100 + \$20.

Example 3 is unnatural because there is no way for Bob to rationalize his initial purchase. Letting the option expire results in a net loss of \$100, whereas exercising the option results in a net loss of \$120. More money is better, so Bob has to acknowledge the mistake and cut his losses.

Even at high stakes, decision-makers sometimes seek to rationalize sunk costs rather than acknowledge error. For instance, a senior Irish Republican Army leader was asked in 1978 whether the cost of violent resistance had been worth it. He replied, "Virtually nothing has been achieved. We can't give up now and admit that men and women who were sent to their graves died for nothing." (Smith, 1997, p. 225)[2]

Motivated by these examples, we propose a theory about agents who seek to rationalize their past choices by adapting their preferences. We model an agent facing a decision problem with this structure:

1. The agent chooses action a_1 from menu A_1.
2. The agent learns the state of the world $s \in S$.

[2] For further reading on how rationalizations by combatants prolonged the Troubles, see Chapter 3 of Alonso (2007).

3. The agent chooses action a_2 from menu $A_2(a_1)$, which can depend on his first action.

A utility function takes as arguments a_1, a_2, and s. The agent's *material utility function* is denoted u. The agent may adopt any utility function in the set $\mathcal{V}$, which we call *rationales.* u and $\mathcal{V}$ are primitives of the model. We assume that $u \in \mathcal{V}$.

We start by describing the agent's choice from menu $A_2(a_1)$, after having chosen a_1 from menu A_1 and learned that the state is s. The agent maximizes a weighted sum of his material utility and a regret term that is assessed according to the agent's chosen rationale. Formally, the agent chooses action a_2 from menu $A_2(a_1)$ and rationale v from $\mathcal{V}$ to maximize *total utility*, that is

$$(1-\gamma)\underbrace{u(a_1,a_2,s)}_{\substack{\text{material}\\\text{utility}}}+\gamma\underbrace{\left[v(a_1,a_2,s)-\max_{\substack{\hat{a}_1\in A_1\\\hat{a}_2\in A_2(\hat{a}_1)}} v(\hat{a}_1,\hat{a}_2,s)\right]}_{\textit{ex post}\text{ optimality under chosen rationale}}$$

where $\gamma \in [0,1)$ is the weight on the regret term.

Observe that if a_1 is *ex post* optimal, that is

$$\max_{\hat{a}_2\in A_2(a_1)} u(a_1,\hat{a}_2,s)=\max_{\substack{\hat{a}_1\in A_1\\\hat{a}_2\in A_2(\hat{a}_1)}} u(\hat{a}_1,\hat{a}_2,s),$$

then the theory predicts that the agent chooses a_2 to maximize material utility. Moreover, if A_1 is a singleton then a_1 is trivially *ex post* optimal. Hence, the theory departs from the classical prediction only when the agent has made a choice and that choice was an *ex post* mistake.

When the agent has made an *ex post* mistake, he may be able to reduce regret by choosing rationale $v \neq u$. By construction, a_2 maximizes a weighted sum of his material utility $u(a_1,a_2,s)$ and his chosen rationale $v(a_1,a_2,s)$, which distorts his choices compared to the classical benchmark.

We now apply the model to the earlier examples. The rationales $\mathcal{V}$ are parameterized by $\theta \in [0,400]$. Utility function v_θ specifies that the agent gets θ utils for attending the game, -200 utils for driving through a snowstorm, and $-p$ utils for paying p dollars. Material utility is $u = v_{180}$, so a classical agent ($\gamma = 0$) is willing to pay \$180 to attend the game in good weather, but will stay home in a snowstorm.

In Example 1, the menu A_1 has two alternatives; the agent can get a ticket and lose \$100 or he can decline. After buying the ticket for \$100, the agent learns that there is a snowstorm. If he stays home and adopts rationale v_θ, then his total utility is

$$(1-\gamma)\underbrace{(-100)}_{\substack{\text{material}\\\text{utility}}}+\gamma\underbrace{(-100-\max\{0,\theta-200-100\})}_{\text{regret term}}\leq -100.$$

If he attends the game and adopts rationale v_θ for $\theta \geq 300$, then his total utility is

$$(1-\gamma)\underbrace{(180-200-100)}_{\substack{\text{material}\\\text{utility}}}+\gamma\underbrace{(\theta-200-100-(\theta-200-100))}_{\text{regret term}}=(1-\gamma)(-120).$$

By attending the game and exaggerating his enthusiasm, the agent is able to reduce regret at the cost of material utility. If $\gamma > \frac{1}{6}$, his total utility is strictly higher when he attends the game.

Suppose instead that the ticket was free-of-charge. Then staying home in a snowstorm leads to no regret under the agent's material utility function u. Hence, the agent maximizes total utility by adopting rationale $v = u$ and staying home. The agent's behavior exhibits the sunk-cost fallacy; his choice on the day of the basketball game depends on upfront costs that he cannot recover.

In Example 2, the agent has no choice initially, so the menu A_1 contains only one alternative: the agent gets a ticket and loses \$100. This is trivially *ex post* optimal, so the agent maximizes total utility by staying home in the snowstorm. Hence, removing unchosen alternatives from the menu A_1 can alter the agent's later choice from A_2.

In Example 3, we have taken the agent's material utility for the outcomes in Example 1 and converted utils to dollars. The agent bought the financial option for \$100. Exercising the option in good weather yields \$180, and exercising it in a snowstorm loses \$20. But every available rationale agrees about money, so there is no room to reduce regret and the agent does not exercise the option in a snowstorm.

The theory predicts not only that the agent counts sunk costs as reasons to act, but also that he counts *unsunk benefits* as reasons to refrain, as the next example illustrates.

Example 4: Bob has an opportunity to buy a discounted ticket

> to the basketball game for \$20. He declines, believing that there will be a snowstorm. On the day of the game the weather is warm and sunny. An acquaintance offers to sell him a ticket for \$100. Bob decides to stay home. If the tickets had never been on discount, he would have gone to the game.

Under our utility specification, accepting the acquaintance's offer and attending the game yields a material utility of $180 - 100 = 80$ but a regret term of -80. Staying home and adopting the rationale v_0 yields a material utility of 0 and a regret term of 0. For $\gamma > \frac{1}{2}$, the agent declines the offer, even though he would have accepted had the tickets never been on discount.[3]

To complete the model, we specify the agent's behavior when choosing from the first menu A_1. At this point, the agent has no earlier choices to rationalize, so we assume that the agent evaluates choices from A_1 according to his expected material utility under some prior on the states S. This depends on the agent's beliefs about his future choice from A_2. We study two benchmarks: a *naïve* agent believes he will maximize material utility when choosing from A_2; a *sophisticated* agent correctly foresees his choices from A_2, but evaluates them according to material utility.

Given material utility u and the available rationales $\mathcal{V}$, the theory predicts the agent's choices. But how is the modeler to specify the primitives? In some situations, we can use standard restrictions on the preferences agents may plausibly hold. For instance, for an agent choosing between money lotteries, we could assume that $\mathcal{V}$ is a class of preferences with constant relative risk aversion. Similarly, for an agent bidding in an auction, we could assume that the rationales $\mathcal{V}$ have different valuations for the object, but are all quasi-linear in money.

If we do not make *a priori* restrictions on $\mathcal{V}$, can we nonetheless deduce the available rationales? We prove that the model primitives u and $\mathcal{V}$ are fully identified by choice behavior. To be specific, starting from a setting with finitely many actions, we extend the setting so that the agent faces decision problems of the following form:

1. The agent chooses a lottery over first actions and a menu of a_1-contingent lotteries over second actions.

2. The agent learns the state of the world $s \in S$.

[3] Tykocinski et al. (1995) find evidence of such behavior in a vignette study.

3. The agent chooses a contingent lottery over second actions from the menu.

We extend every rationale $v \in \mathcal{V}$ to evaluate lotteries according to their expected utility. Under a regularity condition, we prove that whenever the agent's choice correspondence is consistent with material utility u and rationales $\mathcal{V}$, those primitives are unique up to a state-specific affine transformation.[4] Hence, statements about the agent's available rationales can be reduced to statements about the agent's choice behavior.

Next, we impose more structure on u and $\mathcal{V}$ to yield comparative statics, by assuming that first actions, second actions, and rationales are complements. Let the first actions, the second actions, and parameter set Θ be totally ordered sets.[5] We assume that the rationales $\mathcal{V}$ have the form $\{w(a_1, a_2, \theta, s) : \theta \in \Theta\}$, for some function w that is supermodular in (a_1, a_2, θ). For instance, this includes the rationales we posited for Example 1, if we impose that buy a ticket for $40 is a higher action than don't buy a ticket, and that go to the game is a higher action than stay home. It also includes time-separable utility functions, of the form $w(a_1, a_2, \theta, s) = w_1(a_1, \theta, s) + w_2(a_2, \theta, s)$, with w_t supermodular in (a_t, θ). We assume that the menu $A_2(a_1)$ is monotone non-decreasing in a_1.

We prove that if the agent's first action was *ex post* too high, then his second action is distorted upwards compared to the classical benchmark. Symmetrically, if the agent's first action was *ex post* too low, then his second action is distorted downwards compared to the classical benchmark, as happens in Example 4.

This result yields comparative statics for a variety of settings. It predicts sunk-cost effects in multi-part projects — when first-period effort and second-period effort are complements, the agent responds to first-period cost shocks by exaggerating the value of the project and raising second-period effort. It predicts that agents repeatedly facing identical decisions will have 'sticky' choice behavior, responding too little to new information. In particular, lab subjects who make incentivized reports of priors and posteriors will report posteriors biased towards their priors, and will underweight informative signals compared to subjects who report only posteriors.

As a further application, we study a consumer buying from a monopolist

[4] Even classical models of state-dependent utility are only unique up to a state-specific affine transformation, because beliefs and utilities are not separately identified.

[5] We later state our results under weaker order-theoretic assumptions, requiring that the first actions are a partially ordered set and that the second actions are a lattice.

via a two-part tariff, specifying an upfront payment and a per-unit price. The consumer first chooses whether to make the upfront payment and then has a taste shock that affects the marginal utility of consumption. If he made the upfront payment, he then chooses how much to buy at the per-unit price. The available rationales all have quasi-linear utility, but differ in their taste for the good.

We assume that the monopolist can produce the good at constant marginal cost. Thus, under the classical model, the profit-maximizing tariff sets a per-unit price equal to marginal cost, yields the efficient level of consumption, and extracts all surplus via the upfront payment. This simple benchmark removes all classical motives for pricing above marginal cost, isolating the effect of *ex post* rationalization.

When the taste shock is low, making the upfront payment was an *ex post* mistake. In this case, a rationalizer distorts his demand upwards compared to the classical benchmark. The higher the upfront payment, the greater the distortion. By contrast, conditional on participating, a classical consumer's demand does not depend on the upfront payment, since this is a sunk cost.

We derive the profit-maximizing two-part tariff, finding that it exploits the rationalizer's behavior by setting the per-unit price strictly above marginal cost. This result holds both for naïfs and for sophisticates, even though sophisticates fully foresee the loss in material utility from distorted demand. Hence, the rationalizing behavior of consumers yields a novel motive for firms to price above marginal cost.

2 Literature review

A novel element of our approach is that the agent distorts today's choices so as to justify yesterday's choices; the utility function includes a desire to reduce *retrospective* regret. This contrasts with the canonical approach to regret theory, in which the agent makes choices today so as to reduce regret tomorrow (Savage, 1951; Loomes and Sugden, 1982; Bell, 1982; Wong and Kwong, 2007; Sarver, 2008), which is a desire to reduce *prospective* regret. Our theory agrees with the classical model for one-shot choices (singleton A_2), unlike prospective regret theories.

Our theory posits that the agent seeks to justify her choice *ex post*. She does not adopt the perspective that her choice, while mistaken in hindsight, was sensible given what she knew at the time. After an event has occurred, people are overconfident that they could have predicted it in advance — this phenomenon, known as hindsight bias, is the subject of a vast literature (Fischhoff, 1975; Blank et al., 2007). People even misremember their own *ex ante* predictions, falsely believing that they correctly predicted what came to pass (Fischhoff and Beyth, 1975; Fischhoff, 1977). Moreover, lab subjects are sometimes unaware of the actual factors that determined their decisions, and generate spurious explanations when queried by experimenters (Nisbett and Wilson, 1977).

Ex post rationalization is thematically related to theories of cognitive dissonance (Festinger, 1957; Cooper, 2007). These theories posit that people adapt their actions and cognitions to achieve internal consistency. They predict that the more a person suffers to obtain a reward, the more they will value that reward. For instance, some experiments find that unexpectedly severe initiation rituals cause new members to evaluate the group more positively, a result that is also consistent with *ex post* rationalization (Aronson and Mills, 1959; Gerard and Mathewson, 1966).

Sunk-cost effects have been documented in many settings, such as business decisions[6], usage of durable goods[7], professional sports[8], and auctions[9]. Some studies do not find evidence of sunk-cost effects[10]. For a meta-analysis, see Roth et al. (2015).

There is no broad consensus as to why human behavior exhibits sunk-cost effects. Some studies have offered explanations based on prospect theory (Thaler, 1980; Arkes and Blumer, 1985). These explanations depend on losses relative to a reference point, whereas the theory of *ex post* rationalization predicts that sunk-cost effects depend on whether the agent made a choice to incur those costs. Other studies have suggested that sunk-cost effects may be rational due to reputation concerns (Prendergast and Stole, 1996; McAfee et al., 2010), limited memory (Baliga and Ely, 2011), and self-signaling motives (Hong et al., 2019).

Most directly, the present study builds on ideas from Eyster (2002) and Ridout (2020).

[6] McCarthy et al. (1993); Schoorman (1988); Staw et al. (1997); Guenzel (2020).

[7] Ho et al. (2018).

[8] Staw and Hoang (1995); Camerer and Weber (1999); Keefer (2017).

[9] Herrmann et al. (2015); Augenblick (2016).

[10] Ashraf et al. (2010); Friedman et al. (2007); Ketel et al. (2016); Negrini et al. (2020).

Eyster (2002) studies a two-period model in which an agent wishes to maximize a weighted average of material utility and *ex post* regret according to a fixed utility function, but limits attention to alternative first actions that are consistent with the chosen second action. To illustrate, the theory of Eyster (2002) would explain Example 1 by positing that if Bob attends the game, then only buy a ticket is consistent, so he feels no regret. On the other hand, if Bob stays home, then both buy a ticket and don't buy a ticket are consistent, so he feels regret for having bought the ticket. One limitation of this approach is that the modeler's intuitions about consistency may vary with how the actions are framed – stay home seems consistent with don't buy a ticket, but stay home with a ticket in hand does not seem consistent with don't buy a ticket, so Bob can also avoid regret by staying home with a ticket in hand. One advantage of our present approach is that it overcomes this framing objection. Instead of using a frame-dependent consistency relation, we explain behavior via the adoption of rationales, *i.e. post hoc* reasons for the agent's choice, and the rationales are identified from choice data.

Ridout (2020) studies a model of one-shot choice, with an agent who has a set of 'justifiable' preferences and a material preference. As in the present study, the agent sometimes fails to maximize his material preference, but the mechanism is different. Ridout's agent may forego an alternative he prefers because he does not consider his material preference justifiable. By contrast, our agent's material preference belongs to the set of rationales, so there is no distortion in the absence of past mistakes. The agent foregoes a preferred alternative only if acting on his preference would lead him to regret a past decision.

3 Statement of theory

In our model, an agent chooses an action from a menu, then learns the state of the world, and then finally chooses an action from a second menu, which can depend on the first action.

We now define the model primitives. $\mathcal{A}_1$ denotes the *first actions*; $\mathcal{A}_2$ denotes the *second actions*; and S denotes the *states of the world*, with representative elements a_1, a_2, and s, respectively.

A *decision problem* $D \equiv (A_1, A_2, F)$ consists of

1. a first-period menu $A_1 \subseteq \mathcal{A}_1$,

2. and a second-period menu correspondence $A_2 : A_1 \rightrightarrows \mathcal{A}_2$.
3. a prior over states $F \in \Delta S$,

We require that A_1 and A_2 be non-empty.

A *utility function* is a function $v : \mathcal{A}_1 \times \mathcal{A}_2 \times S \to \Re$. The *rationales* are denoted $\mathcal{V}$; these are a set of utility functions that the agent may adopt to justify her actions. The agent's *material utility function* is denoted u, and we assume that $u \in \mathcal{V}$.

The set of rationales $\mathcal{V}$ captures the preferences that the agent regards as reasonable. For instance, the rationales could specify the agent's utility from consuming a good or service. Alternatively, rationales could specify the agent's subjective beliefs about some payoff-relevant event, with the observed state s being a noisy signal about that event.

We start by describing choice in the second period. The agent facing decision problem D has chosen a_1 from menu A_1 and learned that the state is s. She chooses $a_2 \in A_2(a_1)$ and $v \in \mathcal{V}$ to maximize

$$U_D(a_2, v \mid a_1, s) \equiv$$

$$(1-\gamma)\underbrace{u(a_1, a_2, s)}_{\substack{\text{material}\\ \text{utility}}} + \gamma \underbrace{\left[v(a_1, a_2, s) - \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} v(\hat{a}_1, \hat{a}_2, s) \right]}_{\textit{ex post} \text{ optimality under chosen rationale}} \quad (1)$$

for parameter $\gamma \in [0, 1)$. Equation (1) states that the agent places weight $(1-\gamma)$ on maximizing material utility u, and γ on rationalizing her choice *ex post*. The second term in (1) measures how close her course of action is to the *ex post optimum* under her chosen rationale v. When a_1 was *ex post* sub-optimal according to u, the second term might be increased by adopting rationale $v \neq u$. This distorts the agent's choice of a_2, which maximizes $(1 - \gamma)u(a_1, a_2, s) + \gamma v(a_1, a_2, s)$.

We restrict attention to decision problems for which the relevant maxima are well-defined. This is implied, for instance, if every $v \in \mathcal{V}$ is continuous in the actions, and the sets $A_2(a_1)$, $\{(a'_1, a'_2) : a'_1 \in A_1 \text{ and } a'_2 \in A_2(a'_1)\}$, and $\mathcal{V}$ are compact.

There are two natural benchmarks for first-period behavior. A *naïf* chooses

a_1 to maximize $\mathbb{E}_F[u(a_1, a_2^*(a_1, s), s)]$ where $a_2^*(a_1, s)$ is a selection from

$$\operatorname*{argmax}_{a_2 \in A_2(a_1)} u(a_1, a_2, s).$$

A *sophisticate* chooses a_1 to maximize $\mathbb{E}_F[u(a_1, \tilde{a}_2(a_1, s), s)]$ where $\tilde{a}_2(a_1, s)$ is a selection from

$$\operatorname*{argmax}_{a_2 \in A_2(a_1)} \max_{v \in \mathcal{V}} U_D(a_2, v \mid a_1, s).$$

Naïfs and sophisticates both maximize expected material utility *ex ante*, albeit with different beliefs about *ex post* behavior. Maximization of *ex ante* material utility removes any concern for prospective regret of the kind studied by Loomes and Sugden (1982), Bell (1982), and Sarver (2008), isolating the novel effects of retrospective regret.[11] This assumption captures the idea that even though the sophisticate can foresee how rationalization will alter her *ex post* behavior, she views this as a mistake *ex ante* — hindsight bias is only compelling in hindsight.

3.1 Discussion of modeling choices

Plausibly, the agent's rationalization motive depends on the kind of *ex ante* uncertainty she faced. Choosing a risky investment is not like choosing a bet in roulette. It is easier to remember the *ex ante* perspective when evaluating choices with objective risks. By contrast, people are more likely to say, "I should have known it!" for decisions that involved Knightian uncertainty or required deliberation to weigh competing considerations. Our model abstracts from this nuance, representing uncertainty using only a distribution over states. However, we interpret the scope of the theory to be confined to those kinds of uncertainty which seem predictable in hindsight.[12]

For the theory to depart from the classical prediction, the available rationales $\mathcal{V}$ must be limited. For instance, if $\mathcal{V}$ includes a 'stoic' rationale that is indifferent between all action sequences, then the second term in (1) can always be set to zero, and the theory predicts material utility maximization. Thus, the theory's novel predictions depend on plausible restrictions on the rationales

[11] Another natural benchmark is the *empathetic sophisticate*, who maximizes $\mathbb{E}_F\left[\max_{a_2 \in A_2(a_1)} \max_{v \in \mathcal{V}} U_D(a_2, v \mid a_1, s)\right]$. This agent both correctly foresees his *ex post* behavior, and weighs regret directly when choosing *ex ante*.

[12] We suggest that experimental tests of the theory use forms of uncertainty that require the subject to exercise judgment, rather than objective risks such as coin flips or dice rolls.

that the agent can adopt.

Other economic exercises also take restrictions on preferences as given. For instance, in mechanism design, positive results often depend on restricting the agent's preferences to lie within certain *a priori* limits (Hurwicz, 1972). As another example, structural estimation methods often require functional-form restrictions on preferences.

We suggest that $\mathcal{V}$ should be some standard class of preferences for the setting under consideration, if such standards exist. This serves to prevent *ad hoc* explanations and to make the theory a portable extension of existing models, in the sense of Rabin (2013).

Even without *a priori* restrictions on $\mathcal{V}$, the theory has empirical content. In particular, one can pin down the preference relation induced in each state by the material utility function u. One method is to present the agent with binary choices between complete plans of action (*i.e.* singleton $A_2(a_1)$), with no uncertainty about the state (*i.e.* degenerate F). This method works for both naïfs and sophisticates. The theory requires that whenever a_1 was *ex post* optimal according to u, the agent then maximizes u when choosing a_2.

Nonetheless, because the theory's predictions depend on $\mathcal{V}$, its general applicability depends on whether $\mathcal{V}$ can be identified from choice behavior. We take up this challenge in the next section.

4 Identification of preference parameters

In this section, we find that in settings with finitely many actions, the theory's preference parameters are identified from choice behavior.

In particular, let Z_1 be the first actions and let $Z_2(z_1)$ be the second actions that are possible following action z_1. A first-period lottery is a distribution over Z_1. A second-period lottery specifies, for each $z_1 \in Z_1$, a contingent distribution over $Z_2(z_1)$.[13]

We extend the environment as follows:

1. At $t = 1$, the agent chooses between tuples, each consisting of a first-period lottery and a menu of second-period lotteries.

[13] In Section 3, we allowed the second-period menu A_2 to depend on the chosen action a_1, but assumed that the possible second actions $\mathcal{A}_2$ did not depend on the chosen action a_1. Here we are allowing the set of possible second actions to depend on the chosen first action. This is a more general model; the effect is to forbid the analyst to observe choices from impossible menus.

2. The agent learns the state s.

3. At $t = 2$, the agent chooses a second-period lottery from the menu.

We extend each utility function to evaluate lotteries according to their expected utility. We then show that, under a regularity condition, choice behavior pins down the parameter γ, the material utility function u, and the rationales $\mathcal{V}$. u and $\mathcal{V}$ are unique up to the usual affine transformation.

We assume that at $t = 2$, the agent knows the state, but does not yet know the realization of the first-period lottery. This construction is appropriate when the state captures objective information that can be provided to the agent directly, such as a weather forecast or a corporate earnings report. It is inappropriate (except as a thought experiment) when the agent learns the state only by experiencing the first action, as when an agent finds out his propensity for seasickness only by sailing at sea.

Another important caveat is that we assume that the analyst observes the agent's second-period behavior contingent on each first-period choice, including for choices that do not maximize *ex ante* material utility. Even under the classical model, full identification requires that we know how the agent would choose at $t = 2$ after statewise-dominated choices at $t = 1$. Trembles by the agent might justify access to this contingent choice data. Alternatively, the analyst may be able to induce different choices by manipulating the agent's prior, provided that each action profile is optimal in some state.

4.1 Extending the environment

The possible actions at $t = 1$ are given by a finite set Z_1. The possible actions at $t = 2$, given that z_1 was taken at $t = 1$, are given by a finite set $Z_2(z_1)$. Let $Z \equiv \{(z_1, z_2) : z_1 \in Z_1, z_2 \in Z_2(z_1)\}$. Let $\mathcal{U}$ denote the set of expected-utility preferences on $\Delta(Z)$. For any $a \in \Delta(Z)$, let a_1 denote the marginal of a on Z_1, and let a_2 denote the marginal on $(Z_2(z_1))_{z_1 \in Z_1}$.

For any set B, let $\mathcal{K}(B)$ denote the collection of nonempty subsets of B. Let $\mathcal{K}_f(B)$ denote the collection of finite nonempty subsets of B. Let

$$\mathcal{A} \equiv \{A \in \mathcal{K}_f(\Delta(Z)) : a_1 = b_1 \text{ for all } a, b \in A\}$$
$$\mathcal{D}_2 \equiv \{(A_2, A_1) \in \mathcal{A} \times \mathcal{K}_f(\Delta(Z)) : A_2 \subseteq A_1\}$$

At $t = 1$, the agent selects a menu $A_2 \in \mathcal{A}$ from a collection of menus in $\mathcal{K}_f(\mathcal{A})$. At $t = 2$, the agent selects one or more lotteries from A_2. The agent learns the state between $t = 1$ and $t = 2$, so choices at $t = 2$ may be state-contingent. For each state s, the primitive is $c_2^s : \mathcal{D}_2 \to \mathcal{A}$ such that $c_2^s(A_2|A_1) \subseteq A_2$ for all $(A_2, A_1) \in \mathcal{D}_2$. To interpret this: A_1 is the union of all menus the agent was offered at $t = 1$, A_2 is the menu the agent chose at $t = 1$, and $c_2^s(A_2|A_1)$ is the agent's selection from A_2 in state s at $t = 2$.

Definition 4.1 (Representation). *$(\gamma, u^s, \mathcal{V}^s) \in [0,1] \times \mathcal{U} \times \mathcal{K}(\mathcal{U})$ represents $c_2^s : \mathcal{D}_2 \to \mathcal{A}$ if*

$$c_2^s(A_2|A_1) = \operatorname*{argmax}_{a \in A_2} \left((1-\gamma)u^s(a) + \gamma \max_{v \in \mathcal{V}^s} \left(v(a) - \max_{b \in A_1} v(b) \right) \right)$$

for all $(A_2, A_1) \in \mathcal{D}_2$.

As with classical models of state-dependent utility, the agent's behavior in both periods remains the same if we scale up the utility functions in state s and scale down the prior probability on s to compensate. Hence, in general, utility can at best be identified up to a *state-specific* affine transformation. We will shortly state conditions under which this is possible.

4.2 Identification result

For the identification result, we restrict attention to choice functions with regular representations, defined as follows.

Definition 4.2 (Regularity). *$(\gamma, u^s, \mathcal{V}^s) \in [0,1] \times \mathcal{U} \times \mathcal{K}(\mathcal{U})$ is **regular** if it satisfies the following conditions:*

1. *$\gamma \in (0, 1)$.*
2. *$\mathcal{V}^s$ is compact, convex, and non-singleton.*
3. *$u^s \in relint(\mathcal{V}^s)$.*
4. *(No Free Distortion) If $v, v' \in \mathcal{V}^s$ and $v' = \alpha v + \beta$ for some $\alpha > 0$ and $\beta \in \mathcal{U}$ such that $\beta(z_1, z_2) = \beta(z_1, z_2')$ for all $z_1 \in Z_1$ and all $z_2, z_2' \in Z_2(z_1)$, then $v = v'$.*

Regularity rules out the agent always maximizing material utility u^s in state s. In that case, it is not possible to fully identify the rationales.

Clause 3 of Definition 4.2 requires that u^s is in the relative interior of $\mathcal{V}^s$. Essentially, this means that if the agent can distort his rationale in one direction, then he can distort his rationale (at least slightly) in the opposite direction.

No Free Distortion says that the agent cannot distort his utility from first-period actions without also distorting his utility over second-period actions. To see why this is helpful, consider a decision maker who has access to rationales v and αv, for $\alpha > 1$. By inspection of Definition 4.2, the agent's total utility is weakly increased by switching from αv to v, so the availability of the scaled-up rationale is not reflected in choice behavior. Clause 4 rules out pairs of rationales that are positive affine transformations of each other, but it is stronger. It also rules out pairs of rationales that are positive affine transformations of each other when the first-period lottery is held fixed. Intuitively, this is needed because we only observe choice between lotteries with the same marginal over first-period actions, so we cannot get full identification unless different rationales disagree on this space.

Suppose that the agent's second-period choice correspondence in state s has a regular representation. The next theorem states that this representation is unique up to a positive affine transformation.

Theorem 4.3. *For any state s: if c_2^s has regular representations $(\gamma, u^s, \mathcal{V}^s)$ and $(\hat{\gamma}, \hat{u}^s, \hat{\mathcal{V}}^s)$, then*

$$\begin{aligned}\hat{\gamma} &= \gamma \\ \hat{u}^s &= \alpha^s u^s + \beta^s(u^s) \\ \hat{\mathcal{V}}^s &= \{\alpha^s v + \beta^s(v) : v \in \mathcal{V}^s\}\end{aligned}$$

for some $\alpha^s > 0$ and some $\beta^s : \mathcal{V}^s \to \mathbb{R}$.

The proof is in Appendix A.

4.2.1 Proof sketch for Theorem 4.3

The key step in the proof is to recover $\mathcal{V}^s$ from choice data; we sketch the method here.

We start with a simple example to illustrate the idea. At $t = 1$, the agent either starts a project ($z_1 = 1$) or declines ($z_1 = 0$). If he starts the project, then at $t = 2$ he has the option to continue ($z_2 = 1$) or stop ($z_2 = 0$). Otherwise the only option is to stop ($z_2 = 0$). The rationales have the form $\theta z_1 z_2 - s z_1 - s z_2$, where $\theta \in [0, 3]$ and s is a cost shock.

The agent's choices between lotteries determine, in each state, the probability p_{11} that $z_1 = 1$ and $z_2 = 1$, the probability p_{00} that $z_1 = 0$ and $z_2 = 0$, and the probability $p_{10} = 1 - p_{11} - p_{00}$ that $z_1 = 1$ and $z_2 = 0$.

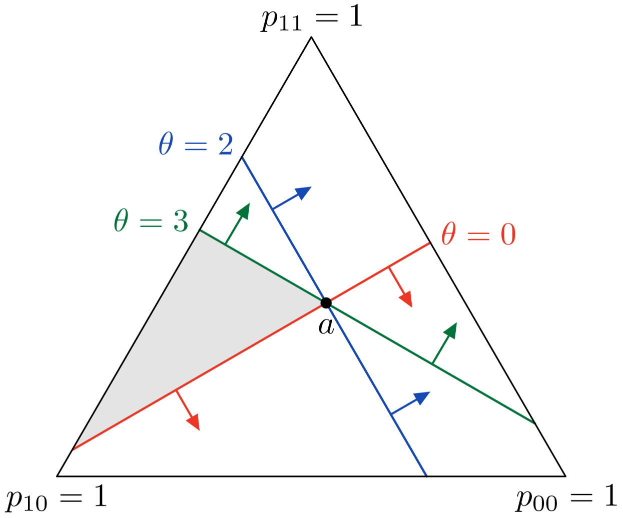

Figure 1: Indifference curves for lotteries over action profiles when $s = 1$. Arrows indicate direction of increasing utility for each rationale.

Let us suppose that $s = 1$, and fix some fully mixed lottery a (over action profiles). In state $s = 1$, a has a marginal distribution over action profiles, which is in the unit 3-simplex. Figure 1 depicts the indifference curves that pass through this point, for rationales with $\theta = 0$, $\theta = 2$, and $\theta = 3$. The shaded area consists of lotteries that *every* rationale regards as weakly worse than a. This set is a convex cone with vertex at a, and its supporting hyperplanes are the indifference curves of the rationales.

Returning to generality, let us fix some fully mixed lottery over action profiles $a = (a_1, a_2)$ and some state s. Let us consider the lotteries that every rationale regards as weakly worse than a, that is

$$\bar{W}_{\text{inner}}(a, s) \equiv \bigcap_{v \in \mathcal{V}^s} \{b \in \Delta(Z) : v(a, s) \geq v(b, s)\} \tag{2}$$

which corresponds to the shaded area in Figure 1. Since each rationale evaluates lotteries according to expected utility, (2) is a convex cone with vertex at a. It will turn out that the cone's supporting hyperplanes pin down $\mathcal{V}^s$.

Recall from Example 2 in Section 1 that a rationalizer's second-period choice can depend on the availability of *unchosen* first-period alternatives. We use this phenomenon to fully identify $\bar{W}_{\text{inner}}(a, s)$. Let p be some lottery over action profiles. Let us say that p *matters for* a (in state s) if there exists another lottery over action profiles x and a second-period lottery b_2 such that

1. after choosing $\{(a_1, a_2), (a_1, b_2)\}$ and foregoing $\{x\}$, and learning the state is s, the agent chooses (a_1, a_2),
2. but after choosing $\{(a_1, a_2), (a_1, b_2)\}$ and foregoing $\{x\}$ and $\{p\}$, and learning the state is s, the agent *does not* choose (a_1, a_2).

Notice that if $p \in \bar{W}_{\text{inner}}(a, s)$, then *every* rationale weakly prefers a to p in state s, so the availability of p does not alter the regret term, which implies that p does not matter for a in state s. Thus, $p \notin \bar{W}_{\text{inner}}(a, s)$ if p matters for a in state s.

It turns out that the converse almost holds. Intuitively, if $p \notin \bar{W}_{\text{inner}}(a, s)$, then we can find another alternative x such that a is dominated by $\{p, x\}$ in state s. That is, every rationale prefers p or x to a in state s. If the agent has access to an equally materially desirable alternative b, and if b is not itself dominated by $\{p, x\}$, the agent will choose b over a after foregoing p and x. It turns out that such a b can always be found if p is close enough to a. We prove Lemma A.4: For any choice correspondence with a regular representation, $p \notin \bar{W}_{\text{inner}}(a, s)$ if and only if there exists $\epsilon \in (0, 1]$ such that $\epsilon p + (1 - \epsilon) a$ matters for a in state s.

Lemma A.4 establishes that $\bar{W}_{\text{inner}}(a, s)$ is identified from choice behavior. With this in hand, we then establish that the supporting hyperplanes of $\bar{W}_{\text{inner}}(a, s)$ are the indifference curves of the utility functions in $\mathcal{V}^s$. Thus, the preferences represented by the rationales can be deduced from choice data. We leave further details (including the identification of material utility, and the scaling of the rationales relative to material utility) to Appendix A.

5 Comparative statics for complements

In this section, we impose additional structure on the setting to yield comparative statics results. With this structure, the theory predicts systematic deviations from the classical model.

We now assume that the first actions $\mathcal{A}_1$ are a partially ordered set and the second actions $\mathcal{A}_2$ are a lattice.[14] We further assume that the rationales $\mathcal{V}$ have the form $\{w(a_1, a_2, \theta, s) : \theta \in \Theta\}$, for some totally ordered parameter set Θ and

[14] For more detail on comparative statics, see Milgrom and Shannon (1994).

some function w. Hence, $\mathcal{A}_2 \times \Theta$ is a lattice under the component-wise order. We use θ^* to denote the parameter value that corresponds to material utility, so $u(a_1, a_2, s) = w(a_1, a_2, \theta^*, s)$. Hence, the agent facing some decision problem D, having chosen a_1 from menu A_1 and observed state s, chooses $a_2 \in A_2(a_1)$ and $\theta \in \Theta$ to maximize

$$U_D(a_2, \theta \mid a_1, s) \equiv \\ (1-\gamma) w(a_1, a_2, \theta^*, s) + \gamma \left[w(a_1, a_2, \theta, s) - \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} w(\hat{a}_1, \hat{a}_2, \theta, s) \right]. \quad (3)$$

We assume that w and U_D have non-empty maxima with respect to (a_1, a_2, θ), and similarly for subsets of these arguments.

For our next theorem, we will assume that the choice variables are complements — the marginal return of raising one variable is non-decreasing in the other variables. If $\mathcal{A}_1$, $\mathcal{A}_2$, and Θ are intervals of $\Re$ and w is differentiable in (a_1, a_2, θ), then our assumption is equivalent to the requirement that the cross partial derivatives are all non-negative.

Formally, w has *increasing differences* between a_1 and (a_2, θ) if for any $\tilde{a}_1' \geq \tilde{a}_1$ and $(\tilde{a}_2', \tilde{\theta}') \geq (\tilde{a}_2, \tilde{\theta})$, we have

$$w(\tilde{a}_1', \tilde{a}_2', \tilde{\theta}', s) - w(\tilde{a}_1', \tilde{a}_2, \tilde{\theta}, s) \geq w(\tilde{a}_1, \tilde{a}_2', \tilde{\theta}', s) - w(\tilde{a}_1, \tilde{a}_2, \tilde{\theta}, s).$$

w is *supermodular* in (a_2, θ) if for any $(\tilde{a}_2, \tilde{\theta})$ and $(\tilde{a}_2', \tilde{\theta}')$, we have

$$\begin{aligned} w(a_1, (\tilde{a}_2, \tilde{\theta}) \vee (\tilde{a}_2', \tilde{\theta}'), s) + w(a_1, (\tilde{a}_2, \tilde{\theta}) \wedge (\tilde{a}_2', \tilde{\theta}'), s) \\ \geq w(a_1, \tilde{a}_2, \tilde{\theta}, s) + w(a_1, \tilde{a}_2', \tilde{\theta}', s). \end{aligned}$$

Given any lattice X, we order subsets Y and Z with the *strong set order*, writing $Y \ll Z$ if for any $y \in Y$ and $z \in Z$, we have $y \wedge z \in Y$ and $y \vee z \in Z$. Given a partially ordered set Q, we say that a correspondence $J : Q \rightrightarrows X$ is *monotone non-decreasing* if $q \leq q'$ implies that $J(q) \ll J(q')$.

The next theorem shows that under appropriate complementarities, the theory yields systematic deviations from the classical benchmark. In the theory, sunk-cost effects are one instance of a more general phenomenon, stated as follows.

Theorem 5.1. *Suppose that w has increasing differences between a_1 and (a_2, θ)*

and is supermodular in (a_2, θ). Suppose that $A_2(a_1)$ is monotone non-decreasing in a_1. If the agent's initial choice was ex post weakly higher than optimal, i.e. $\bar{a}_1 \geq a_1^*$ for

$$a_1^* \in \operatorname*{argmax}_{a_1 \in A_1} \left\{ \max_{a_2 \in A_2(a_1)} w(a_1, a_2, \theta^*, s) \right\} \tag{4}$$

then the agent's choice from menu A_2 is weakly higher than materially optimal, i.e.

$$\operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \left\{ \max_{\theta \in \Theta} U(\bar{a}_1, a_2, \theta, s, \gamma) \right\} \gg \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*, s), \tag{5}$$

and for any selection $\bar{a}_2$ from the left-hand side of (5), there exists $\bar{\theta} \geq \theta^*$ such that

$$\bar{a}_2 \in \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} U_D(a_2, \bar{\theta} \mid \bar{a}_1, s). \tag{6}$$

Symmetrically, if the agent's initial choice was ex post weakly lower than optimal, then the agent's choice $\bar{a}_2 \in A_2$ is weakly lower than materially optimal, and there exists $\bar{\theta} \leq \theta^*$ such that $\bar{a}_2 \in \operatorname{argmax}_{a_2 \in A_2(\bar{a}_1)} U_D(a_2, \bar{\theta} \mid \bar{a}_1, s)$.

The proof is in Appendix B.

5.1 Applications of Theorem 5.1

Here are some applications that satisfy the assumptions of Theorem 5.1.

5.1.1 The sunk-cost effect for two-part projects

The agent chooses effort levels $a_1 \in [0, 1]$ and $a_2 \in [0, 1]$. The project succeeds with probability $a_1 a_2$, he receives a reward valued at $\theta \geq 0$ if it succeeds, and he pays effort cost $sc_1(a_1) + c_2(a_2)$, for continuous non-decreasing cost functions c_1 and c_2, where s is a cost shock for first-period effort. Hence

$$w(a_1, a_2, \theta) = \theta a_1 a_2 - sc_1(a_1) - c_2(a_2)$$

The agent chooses a_1 before learning s, so the materially optimal choice of a_2 does not depend on the realized s. Theorem 5.1 implies that when s has a high enough realization, so that a_1 was *ex post* too high, the rationalizer's choice of a_2 is distorted upwards compared to the classical benchmark. Hence, the agent persists more than is materially optimal for projects that turned out to be unexpectedly costly.

5.1.2 Encountering the same problem twice

The agent faces a decision problem, chooses an action, then learns the state, and then faces the same problem again. That is, $A_1 = A_2 = A$ and

$$w(a_1, a_2, \theta, s) = \phi(a_1, \theta, s) + \phi(a_2, \theta, s),$$

for some function $\phi : A \times \Theta \times S \to \Re$ that is supermodular in (a, θ). Let $a^*(s)$ be an ex post optimal choice in state s, *i.e.* $a^*(s) \in \operatorname{argmax}_a \phi(a, \theta^*, s)$.

Utility is time-separable, so upon learning the state, a classical agent chooses $a_2 = a^*(s)$; his second-period choice does not depend on his first-period choice. By contrast, Theorem 5.1 implies that a rationalizer chooses a_2 to maximize $(1 - \gamma)\phi(a_2, \theta^*, s) + \gamma\phi(a_2, \bar{\theta}, s)$, with $\bar{\theta} \geq \theta^*$ when $a_1 \geq a^*(s)$ and $\bar{\theta} \leq \theta^*$ when $a_1 \leq a^*(s)$. Attempting to rationalize the earlier decision creates a link between otherwise-separate decisions, pulling the rationalizer's second-period choice away from $a^*(s)$ in the direction of his initial choice a_1. Thus, the rationalizer's choice is 'stickier' than a classical agent's choice, responding less to the new information about the state.

5.1.3 Belief elicitation

In many laboratory experiments, subjects provide point estimates of some quantity, then learn some information, and finally report updated estimates. They are paid for one decision drawn at random, so they encounter the same problem twice, in the sense of Section 5.1.2.[15]

The outcome Y is a real-valued random variable with countable support.[16] The subject makes an incentivized report of $\mathbb{E}[Y]$, then observes a signal X with known conditional distribution $g(x \mid y)$, then makes an incentivized report of $\mathbb{E}[Y \mid X = x]$. The available rationales are priors on Y; these are a set of probability mass functions indexed by θ, denoted $(h_\theta)_{\theta \in \Theta}$. We assume that this set is totally ordered by the monotone likelihood ratio property (MLRP) (Milgrom, 1981), that is, for any $\theta > \theta'$ and any $y > y'$

$$h_\theta(y) h_{\theta'}(y') > h_{\theta'}(y) h_\theta(y').$$

This restriction is without loss of generality if Y is a Bernoulli random variable,

[15] Azrieli et al. (2018) study the merits of paying one decision drawn at random.

[16] A parallel construction works if Y has support in some interval and each available rationale is an atomless distribution with strictly positive density.

i.e. when the agent is being asked to report the probability of some event. We assume that each h_θ has full support, so that no rationale is ruled out by some signal realization.

Given prior h_θ and signal realization x, we denote the posterior probability mass function $h_\theta(y \mid X = x)$. The agent reports a_1, then observes the signal realization, then reports a_2. For each report a_t, the agent faces quadratic loss (conditional on the signal realization)

$$\phi(a_t, \theta, x) = -\sum_y (a_t - y)^2 h_\theta(y \mid X = x).$$

This captures the interim expected utility of a risk-neutral agent facing a quadratic scoring rule. It also captures the interim expected utility of an agent with general risk preferences facing an appropriate binarized scoring rule (Hossain and Okui, 2013).

Given the same signal realization, MLRP-ordered priors induce posteriors that are ordered by first-order stochastic dominance (Milgrom, 1981; Klemens, 2007). Thus, if $\theta > \theta'$ then $h_\theta(y \mid X = x)$ first-order stochastically dominates $h_{\theta'}(y \mid X = x)$. It follows that ϕ is supermodular in (a_t, θ).

Our analysis in Section 5.1.2 implies that when $a_1 \geq \mathbb{E}[Y \mid X = x]$, then the agent's reported posterior beliefs are distorted upwards, $a_2 \geq \mathbb{E}[Y \mid X = x]$.[17] Such preference for consistency in belief elicitation is folk wisdom amongst experimenters.[18]

6 Sunk costs and two-part tariffs

We study a monopolist selling a divisible good to a consumer using a two-part tariff, as in Thaler (1980). Thaler proposes that raising the upfront payment can increase the quantity demanded, because the consumer falls victim to the sunk-cost fallacy. We formalize this observation in the context of our model.

The timing is as follows:

1. The monopolist offers a two-part tariff with per-unit price $p \geq 0$ and

[17] Also in that case, when $a_1 \in \text{argmax}_{a \in A_1} \{\phi(a, \theta', x)\}$ for some θ', then $a_2 \leq a_1$. (If $a_2 > a_1$, then $\phi(a_1, \theta^*, x) > \phi(a_2, \theta^*, x)$, in which case by reducing a_2 to a_1 and adopting the rationale θ', the agent can achieve higher material utility and zero regret, which contradicts the optimality of a_2.) This assumption is satisfied whenever the rationales include all full-support priors on Y.

[18] See Falk and Zimmermann (2018) for controlled experiments which show that laboratory subjects report beliefs that are distorted towards their prior reports.

upfront payment L.

2. The consumer accepts and pays L or rejects.
3. The consumer learns the state $s \in [0,1]$, which is drawn according to strictly increasing atomless CDF $F : [0,1] \to [0,1]$.
4. If the consumer accepted the contract, the consumer chooses quantity $q \in \Re_0^+$ and pays pq.

The monopolist produces the good at constant marginal cost $c > 0$. The consumer has utility 0 if he rejects the contract. The consumer's utility from tariff (p, L) is

$$u(q, s, \theta, p, L) = (s + \theta)2\sqrt{q} - pq - L$$

where $\Theta = [-1, 1]$ and $\theta^* = 0$.

If the monopolist faces a classical consumer ($\gamma = 0$), then the profit-maximizing tariff sets $p = c$ and chooses L so that the consumer's participation constraint binds.

For $p > 0$, let us define

$$q^*(s, p, L) \equiv \underset{q}{\operatorname{argmax}}\, u(q, s, \theta^*, p, L) = \frac{s^2}{p^2}$$

and

$$u^*(s, p, L) \equiv \max_q u(q, s, \theta^*, p, L) = \frac{s^2}{p} - L.$$

Let $\bar{s}$ be the state at which a classical consumer breaks even, that is

$$\bar{s} \equiv \begin{cases} \sqrt{Lp} & \text{if } L \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

If $\bar{s} > 1$, then in every state the consumer's material utility is strictly negative. Hence, the consumer rejects the contract, regardless of whether he is naïve, sophisticated, or classical. Now we study the case $\bar{s} \leq 1$.

Suppose a rationalizing consumer accepts contract (p, L). Then his problem is to choose $q \in \Re_0^+$ and $\theta \in \Theta$ to maximize

$$(1-\gamma)[s2\sqrt{q} - pq - L] \\ + \gamma\left[(s+\theta)2\sqrt{q} - pq - L - \max\left\{\max_{\hat{q}}(s+\theta)2\sqrt{\hat{q}} - p\hat{q} - L, 0\right\}\right] \quad (7)$$

If $s \geq \bar{s}$, then Equation (8) is maximized by setting $s = q^*(s,p,L)$ and $\theta = 0$, since that maximizes material utility $[s2\sqrt{q} - pq - L]$ and sets the regret term to 0.

We now cover the case $s < \bar{s}$. The consumer's problem reduces to choosing θ to maximize

$$\max_q [(s+\gamma\theta)2\sqrt{q} - pq - L] - \gamma \max\left\{\max_{\hat{q}}(s+\theta)2\sqrt{\hat{q}} - p\hat{q} - L, 0\right\}$$

$$= \frac{(s+\gamma\theta)^2}{p} - L - \gamma \max\left\{\frac{(s+\theta)^2}{p} - L, 0\right\} \quad (8)$$

which is continuous in θ, increasing in θ for $\theta < \bar{s} - s$, and decreasing in θ for $\theta > \bar{s} - s$. Hence Equation (8) attains its maximum at $\theta = \bar{s} - s = \sqrt{Lp} - s$, and the rationalizing consumer demands

$$q_\gamma(s,p,L) \equiv \begin{cases} \frac{(s+\gamma(\sqrt{Lp}-s))^2}{p^2} & \text{if } s < \sqrt{Lp} \\ \frac{s^2}{p^2} & \text{if } s \geq \sqrt{Lp} \end{cases}. \quad (9)$$

Compared to the classical consumer, the rationalizing consumer's quantity demanded is distorted upwards. For $s < \bar{s}$, the rationalizing consumer behaves in state s as a classical consumer would in state $s+\gamma(\bar{s}-s)$, that is, $q_\gamma(s,p,L) = q^*(s+\gamma(\bar{s}-s),p,L)$. For $s \geq \bar{s}$, the rationalizing consumer behaves the same as a classical consumer, $q_\gamma(s,p,L) = q^*(s,p,L)$.

Equation (9) implies that rationalizing consumer's behavior exhibits sunk-cost effects; raising the upfront payment L increases demand conditional on accepting the contract, but only below the break-even state.

Proposition 6.1. *Suppose* $0 \leq L < L'$. *For* $s < \sqrt{L'p}$, $q_\gamma(s,p,L) < q_\gamma(s,p,L')$. *For* $s \geq \sqrt{L'p}$, $q_\gamma(s,p,L) = q_\gamma(s,p,L')$.

Of course, raising the upfront payment too far leads the consumer to reject the contract. We now characterize the profit-maximizing two-part tariff for a naïf, which sets the per-unit price strictly above marginal cost.

Theorem 6.2. *For a naïf, the monopolist's profit-maximizing tariff sets the per-unit price* p *to solve*

$$\frac{p}{c} = \frac{\int_0^1 s^2 dF(s) + \lambda}{\int_0^1 s^2 dF(s) + \frac{\lambda}{2}}$$

where $\bar{s}^2 = \int_0^1 s^2 dF(s)$ *and*

$$\lambda = \int_0^{\bar{s}} 2\gamma s(\bar{s} - s) + \gamma^2(\bar{s} - s)^2 dF(s)$$

and sets the upfront payment L *so that the naïf's participation constraint binds.*

The proof is in Appendix C.

Why does the profit-maximizing tariff for the naïf set the per-unit price above marginal cost? The naïf does not foresee that he will distort his demand upwards when the state is low, so he underestimates the second-period payments he will make upon accepting the contract. The monopolist exploits this by setting $p > c$.

A sophisticated consumer correctly foresees his demand when deciding whether to accept the contract. Nonetheless, the profit-maximizing contract still sets the per-unit price above marginal cost.

Theorem 6.3. *For a sophisticate, the monopolist's profit-maximizing tariff sets the per-unit price* p *to solve*

$$\frac{p}{c} = \frac{\int_0^1 s^2 dF(s) + \int_0^{\tilde{s}} 2\gamma s(\tilde{s} - s) + \gamma^2(\tilde{s} - s)^2 dF(s)}{\int_0^1 s^2 dF(s) + \int_0^{\tilde{s}} \gamma s(\tilde{s} - s) dF(s)}$$

where $\tilde{s}$ *solves*

$$\int_0^1 s^2 dF(s) - \int_0^{\tilde{s}} \gamma^2(\tilde{s} - s)^2 dF(s) - \tilde{s}^2 = 0$$

and sets the upfront payment L *so that the sophisticate's participation constraint binds.*

The proof is in Appendix C.

The key intuition for Theorem 6.3 is that the sophisticate would rather pay the monopolist via the per-unit price than via the upfront payment, because the upfront payment is a sunk cost that distorts demand and reduces material utility.

For the monopolist to extract surplus, we must have either $p > c$ or $L > 0$. In the classical case, $p = c$ and $L > 0$ can yield the first-best outcome, whereas $p > c$ reduces material efficiency. By contrast, for a rationalizer, $L > 0$ also reduces material efficiency compared to the first-best, and sophisticated rationalizers foresee this. The profit-maximizing tariff trades off these two distortions, extracting surplus via a combination of $p > c$ and $L > 0$.

7 Conclusion

Standard economic theory holds that human beings make decisions to satisfy their preferences. In our theory, human beings adopt preferences to rationalize their previous decisions. Reassuringly, this is also tractable for formal economic analysis.

Having come this far, it is for the reader to decide whether the exercise was worthwhile.

References

Alonso, R. (2007): *The IRA and Armed Struggle*, Political Violence, Taylor & Francis.

Arkes, H. R. and C. Blumer (1985): "The psychology of sunk cost," *Organizational Behavior and Human Decision Processes*, 35, 124–140.

Aronson, E. and J. Mills (1959): "The effect of severity of initiation on liking for a group." *The Journal of Abnormal and Social Psychology*, 59, 177.

Ashraf, N., J. Berry, and J. M. Shapiro (2010): "Can higher prices stimulate product use? Evidence from a field experiment in Zambia," *American Economic Review*, 100, 2383–2413.

Augenblick, N. (2016): "The sunk-cost fallacy in penny auctions," *The Review of Economic Studies*, 83, 58–86.

Azrieli, Y., C. P. Chambers, and P. J. Healy (2018): "Incentives in experiments: A theoretical analysis," *Journal of Political Economy*, 126, 1472–1503.

Baliga, S. and J. C. Ely (2011): "Mnemonomics: the sunk cost fallacy as a memory kludge," *American Economic Journal: Microeconomics*, 3, 35–67.

Bell, D. E. (1982): "Regret in decision making under uncertainty," *Operations research*, 30, 961–981.

Blank, H., J. Musch, and R. F. Pohl (2007): "Hindsight bias: On being wise after the event," *Social Cognition*, 25, 1–9.

Camerer, C. F. and R. A. Weber (1999): "The econometrics and behavioral economics of escalation of commitment: A re-examination of Staw and Hoang's NBA data," *Journal of Economic Behavior & Organization*, 39, 59–82.

Cooper, J. (2007): *Cognitive dissonance: 50 years of a classic theory*, Sage.

Eyster, E. (2002): "Rationalizing the past: A taste for consistency," *Nuffield College Mimeograph.*

Falk, A. and F. Zimmermann (2018): "Information Processing and Commitment," *The Economic Journal*, 613, 1983–2002.

Festinger, L. (1957): *A Theory of Cognitive Dissonance.*, California: Stanford University Press.

Fischhoff, B. (1975): "Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty." *Journal of Experimental Psychology: Human perception and performance*, 1, 288.

——— (1977): "Perceived informativeness of facts." *Journal of Experimental Psychology: Human Perception and Performance*, 3, 349.

Fischhoff, B. and R. Beyth (1975): "I knew it would happen: Remembered probabilities of once—future things," *Organizational Behavior and Human Performance*, 13, 1–16.

Friedman, D., K. Pommerenke, R. Lukose, G. Milam, and B. A. Huberman (2007): "Searching for the sunk cost fallacy," *Experimental Economics*, 10, 79–104.

Gerard, H. B. and G. C. Mathewson (1966): "The effects of severity of initiation on liking for a group: A replication," *Journal of Experimental Social Psychology*, 2, 278–287.

Guenzel, M. (2020): "In too deep: The effect of sunk costs on corporate investment," Tech. rep., University of Pennsylvania Working Paper.

Herrmann, P. N., D. O. Kundisch, and M. S. Rahman (2015): "Beating irrationality: does delegating to IT alleviate the sunk cost effect?" *Management Science*, 61, 831–850.

Ho, T.-H., I. P. Png, and S. Reza (2018): "Sunk cost fallacy in driving the world's costliest cars," *Management Science*, 64, 1761–1778.

Hong, F., W. Huang, and X. Zhao (2019): "Sunk cost as a self-management device," *Management Science*, 65, 2216–2230.

Hossain, T. and R. Okui (2013): "The binarized scoring rule," *Review of Economic Studies*, 80, 984–1001.

Hurwicz, L. (1972): "On informationally decentralized systems," in *Decision and Organization*, ed. by C. B. McGuire and R. Radner, Amsterdam: North-Holland, chap. 14, 297–336.

Keefer, Q. A. (2017): "The sunk-cost fallacy in the National Football League: Salary cap value and playing time," *Journal of Sports Economics*, 18, 282–297.

Ketel, N., J. Linde, H. Oosterbeek, and B. van der Klaauw (2016): "Tuition fees and sunk-cost effects," *The Economic Journal*, 126, 2342–2362.

Klemens, B. (2007): "When Do Ordered Prior Distributions Induce Ordered Posterior Distributions?" *Available at SSRN 964720*.

Loomes, G. and R. Sugden (1982): "Regret theory: An alternative theory of rational choice under uncertainty," *The economic journal*, 92, 805–824.

McAfee, R. P., H. M. Mialon, and S. H. Mialon (2010): "Do sunk costs matter?" *Economic Inquiry*, 48, 323–336.

McCarthy, A. M., F. D. Schoorman, and A. C. Cooper (1993): "Reinvestment decisions by entrepreneurs: rational decision-making or escalation of commitment?" *Journal of business venturing*, 8, 9–24.

Milgrom, P. and C. Shannon (1994): "Monotone comparative statics," *Econometrica*, 157–180.

Milgrom, P. R. (1981): "Good news and bad news: Representation theorems and applications," *The Bell Journal of Economics*, 380–391.

Negrini, M., A. Riedl, and M. Wibral (2020): "Still in search of the sunk cost bias," Tech. rep., CESifo Working Paper.

Nisbett, R. E. and T. D. Wilson (1977): "Telling more than we can know: Verbal reports on mental processes." *Psychological review*, 84, 231.

Prendergast, C. and L. Stole (1996): "Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning," *Journal of political Economy*, 104, 1105–1134.

Rabin, M. (2013): "An approach to incorporating psychology into economics," *American Economic Review*, 103, 617–22.

Ridout, S. (2020): "A Model of Justification," *arXiv preprint arXiv:2003.06844*.

Roth, S., T. Robbert, and L. Straus (2015): "On the sunk-cost effect in economic decision-making: a meta-analytic review," *Business research*, 8, 99–138.

Sarver, T. (2008): "Anticipating regret: Why fewer options may be better," *Econometrica*, 76, 263–305.

Savage, L. J. (1951): "The theory of statistical decision," *Journal of the American Statistical association*, 46, 55–67.

Schoorman, F. D. (1988): "Escalation bias in performance appraisals: An unintended consequence of supervisor participation in hiring decisions." *Journal of Applied Psychology*, 73, 58.

Smith, M. L. R. (1997): *Fighting for Ireland?: the military strategy of the Irish Republican movement*, London: Routledge.

Staw, B. M., S. G. Barsade, and K. W. Koput (1997): "Escalation at the credit window: A longitudinal study of bank executives' recognition and write-off of problem loans." *Journal of Applied Psychology*, 82, 130.

Staw, B. M. and H. Hoang (1995): "Sunk costs in the NBA: Why draft order affects playing time and survival in professional basketball," *Administrative Science Quarterly*, 474–494.

THALER, R. (1980): "Toward a positive theory of consumer choice," *Journal of Economic Behavior & Organization*, 1, 39–60.

TYKOCINSKI, O. E., T. S. PITTMAN, AND E. E. TUTTLE (1995): "Inaction inertia: Foregoing future benefits as a result of an initial failure to act." *Journal of personality and social psychology*, 68, 793.

WONG, K. F. E. AND J. Y. KWONG (2007): "The role of anticipated regret in escalation of commitment." *Journal of Applied Psychology*, 92, 545.

A Proof of Theorem 4.3

We establish some notation.

- Since we will only compare choices within a given state, we drop the superscript s throughout.
- For any $V \subseteq \mathcal{U}$, let U_{pref} denote the set of preferences with representations in V.
- For any $S \subseteq \Delta(Z)$ and any $a_1 \in \Delta(Z_1)$, let $S^{a_1} \equiv \{s \in S : s_1 = a_1\}$. For any $v \in \mathcal{U}$, let v^{a_1} denote the restriction of v to $\Delta(Z)^{a_1}$. For any $\succsim \in \mathcal{U}_{\text{pref}}$, let $\succsim^{a_1}$ denote the restriction of $\succsim$ to $\Delta(Z)^{a_1}$. For any $V \subseteq \mathcal{U}$, let $V^{a_1} \equiv \{v^{a_1} : v \in V\}$.
- For any set S, let $\text{co}(S)$ denote the convex hull, and let $\bar{S}$ denote the closure.[19]
- For any $a \in \Delta(Z)$ and any $A_1 \in \mathcal{K}_f(\Delta(Z))$ such that $a \in A_1$, let

$$U(a|A_1) \equiv (1-\gamma)u(a) + \gamma \max_{v \in \mathcal{V}} \left(v(a) - \max_{b \in A_1} v(b) \right).$$

- For any $a \in \Delta(Z)$, let

$$\begin{aligned} W_{\text{inner}}(a) &\equiv \bigcap_{v \in \mathcal{V}} \{b \in \Delta(Z) : v(a) > v(b)\} \\ W_{\text{outer}}(a) &\equiv \bigcap_{v \in \mathcal{V}} \{b \in \Delta(Z)^{a_1} : ((1-\gamma)u + \gamma v)(a) > ((1-\gamma)u + \gamma v)(b)\}. \end{aligned}$$

[19] When S is a set of preferences, a preference $\succsim \in \text{co}(S)$ if some representation of $\succsim$ is a convex combination of preferences in S. A non-constant preference $\succsim \in \bar{S}$ if some representation of $\succsim$ is a convex combination of preferences in S.

The next lemma provides a convenient alternative formulation for NFD, which we use throughout the proof.

Lemma A.1. *For any $\mathcal{V} \in \mathcal{K}(\mathcal{U})$: $\mathcal{V}$ satisfies NFD if and only if, for all $a_1 \in int(\Delta(Z_1))$, no two distinct utilities in $\mathcal{V}$ represent the same preference on $\Delta(Z)^{a_1}$.*

Proof. The "if" direction is easy, so we prove "only if." Fix any $a_1 \in \text{int}(\Delta(Z_1))$. Suppose v, v' are distinct but represent the same preference $\succsim^{a_1}$ on $\Delta(Z)^{a_1}$. For any $z_1 \in Z_1$, we can vary $a(\cdot|z_1)$ while holding a_1 constant and $a(\cdot|z_1')$ constant for each $z_1' \in Z_1 \setminus \{z_1\}$. This delivers an expected-utility preference on $\Delta(\{z_1\} \times Z_2(z_1))$. Since the representation of this preference is unique up to a positive affine transformation, we have

$$v'(z_1, \cdot) = \alpha(z_1)v(z_1, \cdot) + \beta(z_1)$$

for some $\alpha(z_1) > 0$ and $\beta(z_1) \in \mathbb{R}$.

Suppose that, for some $z_1^* \in Z_1$, $v(z_1, \cdot)$ is constant for all $z_1 \in Z_1 \setminus \{z_1^*\}$. Set $\alpha \equiv \alpha(z_1^*)$. Then, for each $z_1 \in Z_1 \setminus \{z_1^*\}$, replace $\beta(z_1)$ with the constant

$$v'(z_1, \cdot) - \alpha v(z_1, \cdot).$$

We now have

$$v'(z_1, \cdot) = \alpha v(z_1, \cdot) + \beta(z_1)$$

for all $z_1 \in Z_1$, so NFD is violated.

Now suppose that $v(z_1, \cdot)$ and $v(z_1', \cdot)$ are non-constant for distinct $z_1, z_1' \in Z_1$. For each $z_1 \in Z_1$, fix some element of $\text{argmax}_{Z_2(z_1)} v(z_1, \cdot)$, and denote it $\bar{Z}_2(z_1)$. Similarly, fix some element of $\text{argmin}_{Z_2(z_1)} v(z_1, \cdot)$, and denote it $\underline{Z}_2(z_1)$. Take any element of

$$\underset{z_1 \in Z_1}{\text{argmax}}\, a_1(z_1)\left(v(z_1, \bar{Z}_2(z_1)) - v(z_1, \underline{Z}_2(z_1))\right),$$

and call it z_1^*. Take any $z_1^{**} \in Z_1 \setminus \{z_1^*\}$ such that $v(z_1^{**}, \cdot)$ is non-constant. Let

$$\lambda \equiv \frac{a_1(z_1^{**})}{a_1(z_1^*)} \times \frac{v(z_1^{**}, \bar{Z}_2(z_1^{**})) - v(z_1^{**}, \underline{Z}_2(z_1^{**}))}{v(z_1^*, \bar{Z}_2(z_1^*)) - v(z_1^*, \underline{Z}_2(z_1^*))}.$$

Notice that $0 < \lambda \leq 1$. We modify a as follows to create two new lotteries

$a^*, a^{**} \in \Delta(Z_1)^{a_1}$:

$$\begin{aligned}
&a_1^* = a_1^{**} = a_1\\
&a^*(\cdot|z_1^*) = \frac{\lambda}{1+\lambda}\delta_{\bar{Z}_2(z_1^*)} + \frac{1}{1+\lambda}a(\cdot|z_1^*)\\
&a^{**}(\cdot|z_1^*) = \frac{\lambda}{1+\lambda}\delta_{\underline{Z}_2(z_1^*)} + \frac{1}{1+\lambda}a(\cdot|z_1^*)\\
&a^*(\cdot|z_1^{**}) = \frac{\lambda}{1+\lambda}a(\cdot|z_1^{**}) + \frac{1}{1+\lambda}\delta_{\underline{Z}_2(z_1^{**})}\\
&a^{**}(\cdot|z_1^{**}) = \frac{\lambda}{1+\lambda}a(\cdot|z_1^{**}) + \frac{1}{1+\lambda}\delta_{\bar{Z}_2(z_1^{**})}\\
&a^*(\cdot|z_1) = a^{**}(\cdot|z_1) = a(\cdot|z_1) \text{ for all } z_1 \in Z_1 \setminus \{z_1^*, z_1^{**}\}.
\end{aligned}$$

The definition of λ ensures that $v(a_1^*) = v(a_1^{**})$. Since v and v' represent the same preference on $\Delta(Z)^{a_1}$, we must also have $v'(a_1^*) = v'(a_1^{**})$. This implies

$$\begin{aligned}
\lambda &= \frac{a_1(z_1^{**})}{a_1(z_1^*)} \times \frac{v'(z_1^{**}, \bar{Z}_2(z_1^{**})) - v'(z_1^{**}, \underline{Z}_2(z_1^{**}))}{v'(z_1^*, \bar{Z}_2(z_1^*)) - v'(z_1^*, \underline{Z}_2(z_1^*))}\\
&= \frac{\alpha(z_1^{**})}{\alpha(z_1^*)} \times \frac{a_1(z_1^{**})}{a_1(z_1^*)} \times \frac{v(z_1^{**}, \bar{Z}_2(z_1^{**})) - v(z_1^{**}, \underline{Z}_2(z_1^{**}))}{v(z_1^*, \bar{Z}_2(z_1^*)) - v(z_1^*, \underline{Z}_2(z_1^*))}\\
&= \frac{\alpha(z_1^{**})}{\alpha(z_1^*)}\lambda.
\end{aligned}$$

Since $\lambda > 0$, we have $\alpha(z_1^*) = \alpha(z_1^{**})$. Set $\alpha \equiv \alpha(z_1^*)$. We can write

$$v'(z_1, \cdot) = \alpha v(z_1, \cdot) + \beta(z_1)$$

for all $z_1 \in Z_1$ such that $v(z_1, \cdot)$ is non-constant. Now, for each $z_1 \in Z_1$ such that $v(z_1, \cdot)$ is constant, replace $\beta(z_1)$ with the constant

$$v'(z_1, \cdot) - \alpha v(z_1, \cdot).$$

We now have

$$v'(z_1, \cdot) = \alpha v(z_1, \cdot) + \beta(z_1)$$

for all $z_1 \in Z_1$, which violates NFD. □

The next lemma establishes some properties of $W_{\text{inner}}(a)$ and $W_{\text{outer}}(a)$.

Lemma A.2. *For any regular representation $(u, \mathcal{V})$ and any $a \in int(\Delta(Z))$,*

$W_{inner}(a)$ and $W_{outer}(a)$ are open and nonempty, and satisfy

$$\bar{W}_{inner}(a) = \bigcap_{v\in\mathcal{V}} \{b \in \Delta(Z) : v(a) \geq v(b)\} \tag{10}$$

$$\bar{W}_{outer}(a) = \bigcap_{v\in\mathcal{V}} \{b \in \Delta(Z)^{a_1} : ((1-\gamma)u + \gamma v)(a) \geq ((1-\gamma)u + \gamma v)(b)\}. \tag{11}$$

Proof. For openness of $W_{\text{inner}}(a)$: Consider a sequence $b_i \to b$ such that $b_i \notin W_{\text{inner}}(a)$ for all i. By definition of $W_{\text{inner}}(a)$, for each b_i, there exists some $v_i \in \mathcal{V}$ such that $v_i(b_i) \geq v_i(a)$. Since $\mathcal{V}$ is compact, we can pass to a convergent subsequence $\bar{v} \in \mathcal{V}$. We must have $\bar{v}(b) \geq \bar{v}(a)$, so $b \notin W_{\text{inner}}(a)$. The argument for openness of $W_{\text{outer}}(a)$ is the same since $\{(1-\gamma)u + \gamma v : v \in \mathcal{V}\}$ inherits compactness from $\mathcal{V}$.

For nonemptiness of $W_{\text{inner}}(a)$: Fix $a \in \Delta(Z)$ and $B \in \mathcal{K}_f(\Delta(Z))$. Suppose there does not exist $v \in \mathcal{V}$ such that $v(a) \geq \max_{b\in B} v(b)$. We show that there exists $\bar{b} \in \text{co}(B)$ such that $v(a) < v(\bar{b})$ for all $v \in \mathcal{V}$. For each $v \in \mathcal{V}$, we have some $b \in B$ such that $v(b) < v(a)$. It without loss to assume that $v(a) = 0$ for all $v \in \mathcal{V}$. We want to find a set of weights α such that

$$\sum_{b\in B} \alpha(b)v(b) > 0 \tag{12}$$

for all $v \in \mathcal{V}$. Enumerate the elements of B: $(b_1, \ldots, b_{|B|})$. For each $v \in \mathcal{V}$, let v_B denote the vector with $v(b_i)$ as its i-th entry. Let $\mathcal{V}_B : \{v_B : v \in \mathcal{V}\}$. Like $\mathcal{V}$, $\mathcal{V}_B$ is nonempty, compact and convex. Let $N \equiv \mathbb{R}_-^{|B|}$, which is nonempty, closed and convex. Notice that no element of $\mathcal{V}_B$ can be weakly negative. (Otherwise, some $v \in \mathcal{V}$ would rank a weakly higher than each member of B.) Thus, $\mathcal{V}_B$ and N are disjoint, and we can apply the separating hyperplane theorem. This delivers a nonzero $\alpha \in \mathbb{R}^{|B|}$ and $c \in \mathbb{R}$ such that $\alpha' n < c < \alpha' v_B$ for all $n \in N$, $v_B \in \mathcal{V}_B$. Since the zero vector belongs to N, we must have $c > 0$. Suppose the ith element of α is strictly negative. By choosing n with a sufficiently negative number in ith position and zeros elsewhere, we get $\alpha' n > c$, a contradiction. Thus, each element of α is weakly positive. If we rescale α to a unit sum, we still have $\alpha' v_B > 0$ for all $v_B \in \mathcal{V}_B$. This inequality is equivalent to (12).

Now fix some $a \in \text{int}(\Delta(Z))$, and suppose that $W_{\text{inner}}(a) = \emptyset$. Since there is no $b \in \Delta(Z)$ such that $v(a) > v(b)$ for all $v \in \mathcal{V}$, there cannot be any $\bar{b} \in \Delta(Z)$ such that $v(\bar{b}) > v(a)$ for all $v \in \mathcal{V}$. Fix any $B \in \mathcal{K}_f(\Delta(Z))$ such that $a \in \text{int}(\text{co}(B))$. By the previous argument, there must be some $v \in \mathcal{V}$

such that $v(a) \geq \max_{b \in B} v(b)$. Since $a \in \text{int}(\text{co}(B))$, v must be a constant. By NFD and $|\mathcal{V}| > 1$, $\mathcal{V}$ must also contain some non-constant utility v'. Since $\mathcal{V}$ is convex, it must contain all convex combinations of v' and v. But these convex combinations have the same restriction to $\Delta(Z)^{a_1}$ that v' does, which violates NFD. Conclude that $\mathcal{V}$ does not contain a constant preference, so $W_{\text{inner}}(a) \neq \emptyset$.

We can use the same arguments to establish that the restriction of $W_{\text{inner}}(a)$ to $\Delta(Z)^{a_1}$ is nonempty. Since $W_{\text{outer}}(a)$ is a superset of the restriction of $W_{\text{inner}}(a)$ to $\Delta(Z)^{a_1}$, $W_{\text{outer}}(a)$ is also nonempty.

For (10): Fix some $b \in \text{int}(\Delta(Z))$ such that $v(a) \geq v(b)$ for all $v \in \mathcal{V}$. We want to show that b is a limit of points in $W_{\text{inner}}(a)$. Take any $b' \in W_{\text{inner}}(b)$. (We have already seen that such a b' exists.) Now fix a sequence $\epsilon_i \to 0$, and let $b_i \equiv \epsilon_i b' + (1 - \epsilon_i)b$. We have $v(a) \geq v(b) > v(b_i)$ for all $v \in \mathcal{V}$, so $b_i \in W_{\text{inner}}(a)$ for all i. Since $b_i \to b$, this is the desired result. Now fix some $b \in \bar{W}_{\text{inner}}(a)$. We have a sequence $b_i \to b$ such that, for each b_i, $v(a) > v(b_i)$ for all $v \in \mathcal{V}$. By continuity of v, $v(a) \geq v(b)$ for all $v \in \mathcal{V}$. The same arguments work for (11). □

Next, we show that $W_{\text{inner}}(a)$ can be fully identified from choice data.

Definition A.3. *Fix $a, p \in \Delta(Z)$. Say that p matters for a if there exist $x \in \Delta(Z)$ and $b \in \Delta(Z)^{a_1}$ such that*

$$a \in c_2(a, b | a, b, x)$$
$$a \notin c_2(a, b | a, b, x, p).$$

Lemma A.4. *For any regular representation $(u, \mathcal{V})$, any $a \in int(\Delta(Z))$, and any $p \in \Delta(Z)$: $p \notin \bar{W}_{inner}(a)$ if and only if, for some $\epsilon \in (0, 1]$, $\epsilon p + (1 - \epsilon)a$ matters for a.*

Proof. Suppose that $p \notin \bar{W}_{\text{inner}}(a)$. We show that $\epsilon p + (1 - \epsilon)a$ matters for a for some $\epsilon \in (0, 1]$. Since $p \notin \bar{W}_{\text{inner}}(a)$, we have $v^*(p) > v^*(a)$ for some $v^* \in \mathcal{V}$. It is without loss to assume that v^* is extreme. Fix $\hat{x} \in W_{\text{inner}}(a)$, which is nonempty by Lemma A.2. For $\hat{\delta} > 0$ sufficiently small, we have $a \in W_{\text{inner}}(a + \hat{\delta}(a - \hat{x}))$ and $v^*(p) > v^*(a + \hat{\delta}(a - \hat{x}))$. Let $\bar{x} \equiv a + \hat{\delta}(a - \hat{x})$ for some such $\hat{\delta}$. Since $v^*(p) > v^*(\bar{x}) > v^*(a)$, we will have $v^*(\bar{x}) + \lambda(v^*(\bar{x}) - v^*(p)) = v^*(a)$ for some $\lambda > 0$. For some $\bar{\delta} > 0$, we will have $\bar{\delta}((1 + \lambda)\bar{x} - \lambda p) + (1 - \bar{\delta})a \in \Delta(Z)$, since a is interior. Let $x \equiv \bar{\delta}((1 + \lambda)\bar{x} - \lambda p) + (1 - \bar{\delta})a$ for some such $\bar{\delta}$. Notice that

$v^*(x) = v^*(a)$. Substituting for x yields

$$\frac{1}{1+\lambda\bar{\delta}}x + \frac{\lambda\bar{\delta}}{1+\lambda\bar{\delta}}p = \frac{\bar{\delta}(1+\lambda)}{1+\lambda\bar{\delta}}\bar{x} + \frac{1-\bar{\delta}}{1+\lambda\bar{\delta}}a.$$

Since every $v \in \mathcal{V}$ strictly prefers $\bar{x}$ to a, every $v \in \mathcal{V}$ strictly prefers the right-hand side to a. Thus, there is a convex combination of x and p that every $v \in \mathcal{V}$ strictly prefers to a. This implies $\max_{x,p} v > v(a)$ for all $v \in \mathcal{V}$. For any $\epsilon \in (0, 1]$, let $x_\epsilon \equiv \epsilon x + (1-\epsilon)a$, and let $p_\epsilon \equiv \epsilon p + (1-\epsilon)a$. We still have $\max_{x_\epsilon,p_\epsilon} v > v(a)$ for all $v \in \mathcal{V}$, so

$$U(a|a, x_\epsilon, p_\epsilon) < (1-\gamma)u(a).$$

Since $|\mathcal{V}| > 1$ and u is non-extreme, there is some extreme $\bar{v} \in \mathcal{V}$ that is neither v^* nor u. By NFD, $\bar{v}$ does not represent the same preference on $\Delta(Z)^{a_1}$ that u or v^* does. Thus, there must exist $b \in \Delta(Z)^{a_1}$ such that

$$\begin{aligned} u(b) &= u(a) \\ v^*(b) &\leq v^*(a) \\ \bar{v}(b) &> \bar{v}(a). \end{aligned}$$

Suppose not. Then, $(v^*)^{a_1}$ and $\bar{v}^{a_1}$ must agree when restricted to $\{b \in \Delta(Z)^{a_1} : u(b) = u(a)\}$. But this implies that $(v^*)^{a_1} \in \text{co}(\{\bar{v}^{a_1}, u^{a_1}\})$ or that $\bar{v}^{a_1} \in \text{co}(\{(v^*)^{a_1}, u^{a_1}\})$. In the first case, v^* must have the same restriction to $\Delta(Z)^{a_1}$ as some convex combination of u and $\bar{v}$. Since $\mathcal{V}$ is convex, that combination actually belongs to $\mathcal{V}$. By NFD, it must equal v^*, which contradicts the assumption that v^* is extreme. A parallel argument handles the second case.

Since $\bar{v}(b) > \bar{v}(a)$, we have

$$\bar{v}(b) = \max_{a,b,x_\epsilon,p_\epsilon} \bar{v}$$

for ϵ sufficiently close to 0. This implies

$$\begin{aligned} U(b|a, b, x_\epsilon, p_\epsilon) &= (1-\gamma)u(b) \\ U(b|a, b, x_\epsilon) &= (1-\gamma)u(b). \end{aligned}$$

Since $U(a|A_1) \leq U(a|B_1)$ whenever $a \in B_1 \subset A_1$,

$$\begin{aligned} U(a|a,b,x_\epsilon,p_\epsilon) &\leq U(a|a,x_\epsilon,p_\epsilon) \\ &< (1-\gamma)u(a) \\ &= (1-\gamma)u(b) \\ &= U(b|a,b,x_\epsilon,p_\epsilon) \\ \implies \quad & a \notin c_2(a,b|a,b,x_\epsilon,p_\epsilon). \end{aligned}$$

Since $v^*(a) = v^*(x_\epsilon) \geq v^*(b)$,

$$\begin{aligned} U(a|a,b,x_\epsilon) &= (1-\gamma)u(a) \\ &= (1-\gamma)u(b) \\ &= U(b|a,b,x_\epsilon) \\ \implies \quad & a \in c_2(a,b|a,b,x_\epsilon). \end{aligned}$$

Conclude that p_ϵ matters for a. □

The next lemma shows that we can use $W_{\text{inner}}(a)$ to fully recover $\mathcal{V}_{\text{pref}}$.

Lemma A.5. *Fix any regular representation $(u, \mathcal{V})$ and any $a \in int(\Delta(Z))$.*

1. *For any $V_{pref} \subset \mathcal{U}_{pref}$ such that $W_{inner}(a) = \bigcap_{\succsim \in V_{pref}} \{b \in \Delta(Z) : a \succ b\}$,*

$$\bar{co}(V_{pref}) = \{\succsim \in \mathcal{U}_{pref} : a \succ b \textit{ for all } b \in W_{inner}(a)\}.$$

2. *$\mathcal{V}_{pref} = \{\succsim \in \mathcal{U}_{pref} : a \succ b$ for all $b \in W_{inner}(a)\}$.*

Proof. First part: We show that $\succsim^* \in \text{co}(V_{\text{pref}})$ if $a \succ^* b$ for all $b \in \bar{W}_{\text{inner}}(a) \setminus \{a\}$. By definition of V_{pref}, we have

$$\{b \in \Delta(Z) : b \succsim^* a\} \setminus \{a\} \subset \bigcup_{\succsim \in \mathcal{V}_{\text{pref}}} \{b \in \Delta(Z) : b \succ a\}.$$

Fix some $\epsilon > 0$ such that $B_\epsilon(a) \subset \text{int}(\Delta(Z))$. We have

$$\{b \in \Delta(Z) : b \succsim^* a\} \setminus B_\epsilon(a) \subset \bigcup_{\succsim \in \mathcal{V}_{\text{pref}}} \{b \in \Delta(Z) : b \succ a\}.$$

By the Heine-Borel theorem, there exists some finite $S_{\text{pref}} \subset V_{\text{pref}}$ such that

$$\{b \in \Delta(Z) : b \succsim^* a\} \setminus B_\epsilon(a) \subset \bigcup_{\succsim \in S_{\text{pref}}} \{b \in \Delta(Z) : b \succ a\} . \tag{13}$$

This implies

$$\{b \in \Delta(Z) : b \succsim^* a\} \setminus \{a\} \subset \bigcup_{\succsim \in S_{\text{pref}}} \{b \in \Delta(Z) : b \succ a\} . \tag{14}$$

Suppose not. Then, there must be some $b \in B_\epsilon(a) \setminus \{a\}$ such that $b \succsim^* a$, but $b \precsim a$ for all $\succsim \in S_{\text{pref}}$. For some $\lambda > 0$, we will have $b + \lambda(b - a) \in \Delta(Z) \setminus B_\epsilon(a)$. By linearity, $b + \lambda(b - a) \succsim^* a$, but $b + \lambda(b - a) \precsim a$ for all $\succsim \in S_{\text{pref}}$. Since $b + \lambda(b - a) \notin B_\epsilon(a)$, this contradicts (13).

Assign $\succsim^*$ a representation v^* such that $v^*(a) = 0$. For each $\succsim \in S_{\text{pref}}$, assign a representation v such that $v(a) = 0$, and let S denote the resulting set of utilities. By Farkas' lemma, we can find $\alpha > 0$ such that $\alpha v^* \in \text{co}(S)$ provided there does not exist $p \in \mathbb{R}^{|Z|}$ such that $p'v^* < 0$ and $p'v \geq 0$ for all $v \in S$. Suppose there exists some such p. Consider the case $\sum_{i:p_i>0} p_i \geq \sum_{i:p_i<0}(-p_i)$. Normalize p by dividing each p_i by $\sum_{i:p_i>0} p_i$. (Since $p \neq 0$, this term must be strictly positive.) We have

$$\begin{aligned} \sum_{i:p_i>0} p_i v^*(a_i) &< \sum_{i:p_i<0} (-p_i) v^*(a_i) \\ \sum_{i:p_i>0} p_i v(a_i) &\geq \sum_{i:p_i<0} (-p_i) v(a_i) \text{ for all } v \in S. \end{aligned}$$

Since the normalization ensures that $\sum_{i:p_i>0} p_i = 1$, the left-hand side is the valuation of a lottery, which we label p^+. Since $v^*(a) = v(a) = 0$,

$$\begin{aligned} v^*(p^+) &< \sum_{i:p_i<0} (-p_i) v^*(a_i) + \left(1 - \sum_{i:p_i<0} (-p_i)\right) v^*(a) \\ v(p^+) &\geq \sum_{i:p_i<0} (-p_i) v(a_i) + \left(1 - \sum_{i:p_i<0} (-p_i)\right) v(a) \text{ for all } v \in S. \end{aligned}$$

Now the right-hand side is also the valuation of a lottery, which we label p^-.

We have

$$\begin{aligned} v^*(p^+) &< v^*(p^-) \\ v(p^+) &\geq v(p^-) \text{ for all } v \in S. \end{aligned}$$

For $\epsilon > 0$ sufficiently small, we have

$$\begin{aligned} v^*(a) &< v^*(a + \epsilon(p^- - p^+)) \\ v(a) &\geq v(a + \epsilon(p^- - p^+)) \text{ for all } v \in S. \end{aligned}$$

This contradicts (14). An exactly similar argument covers the case $\sum_{i:p_i>0} p_i \geq \sum_{i:p_i<0}(-p_i)$. Conclude that $\alpha v^* \in \text{co}(S)$ for some $\alpha > 0$, so $\succsim^* \in \text{co}(S_{\text{pref}})$. Since $S_{\text{pref}} \subset V_{\text{pref}}$, $\succsim^* \in \text{co}(V_{\text{pref}})$.

Now take any $\bar{\succsim} \in \mathcal{U}_{\text{pref}}$ such that $a \mathrel{\bar{\succ}} b$ for all $b \in W_{\text{inner}}(a)$. Suppose that there is some $\succsim^* \in \mathcal{U}_{\text{pref}}$ such that $a \succ^* b$ for all $b \in \bar{W}_{\text{inner}}(a) \setminus \{a\}$. We already showed that $\succsim^* \in \text{co}(V_{\text{pref}})$. Assign $\bar{\succsim}$ a representation $\bar{v}$, and assign $\succsim^*$ a representation v^*. Fix any sequence $\epsilon_i \to 0$ such that $\epsilon_i \in (0,1)$ for each i. Each $\epsilon_i v^* + (1-\epsilon_i)\bar{v}$ represents some $\succsim_i \in \mathcal{U}_{\text{pref}}$. For each i, $a \succ_i b$ for all $b \in \bar{W}_{\text{inner}}(a) \setminus \{a\}$. Thus, $\succsim_i \in \mathcal{V}_{\text{pref}}$ for all i. Since the utilities that represent $\succsim_i$ converge to $\bar{v}$, which represents $\bar{\succsim}$, we have $\bar{\succsim} \in \overline{\text{co}}(V_{\text{pref}})$.

Finally, suppose there is no $\succsim^* \in \mathcal{U}_{\text{pref}}$ such that $a \succ^* b$ for all $b \in \bar{W}_{\text{inner}}(a) \setminus \{a\}$. For any $R \subset Z$, we will use the superscript R to denote the restriction of a set or preference to $\Delta(R)$. We will find $R^* \subset Z$ such that $|R^*| \geq 3$ for which there exists an EU preference $\succsim^*$ on $\Delta(R^*)$ with the following property: for any $a^* \in \Delta(R^*)$,

$$b^* \in \bar{W}^{R^*}_{\text{inner}}(a^*) \setminus \{a^*\} \quad \Longrightarrow \quad a^* \succ^* b^*. \tag{15}$$

Since there is no preference with this property when $R^* = Z$, there must be $a^0, b^0 \in \Delta(Z)$ such that

$$\begin{aligned} b^0 &\in \bar{W}_{\text{inner}}(a^0) \setminus \{a^0\} \\ a^0 + (a^0 - b^0) &\in \bar{W}_{\text{inner}}(a^0) \setminus \{a^0\}. \end{aligned}$$

This implies $b^0 \sim a^0$ for all $\succsim \in V_{\text{pref}}$. Since $b^0 \neq a^0$, we can always relabel the elements of Z so that $b^0(z_1) \neq a^0(z_1)$. Let $R^1 \equiv Z \setminus \{z_1\}$, and check whether there is any preference satisfying (15) with R^1 in place of R^*. If not, then we can repeat the argument, deriving a new indifference condition $a^1 \sim b^1$ and setting $R^2 \equiv Z \setminus \{z_1, z_2\}$. We repeat the process until we arrive at the desired R^*. To see why some such R^* must exist, suppose that we have iterated the process to $R^{n-3} = \{z_{n-2}, z_{n-1}, z_n\}$. Suppose there is no preference satisfying (15) with R^{n-3} in place of R^*. This can only happen if $\bar{W}^{R_{n-3}}_{\text{inner}}(a^{n-3})$ is a half-plane, which in turn only happens if $|\mathcal{V}^{R_{n-3}}_{\text{pref}}| = 1$. Since each preference in $\mathcal{V}_{\text{pref}}$ is pinned down by its restriction to $\mathcal{V}^{R_{n-3}}_{\text{pref}}$ (and the indifference conditions), we have $|\mathcal{V}_{\text{pref}}| = 1$. This contradicts NFD and $|\mathcal{V}| > 1$. Conclude that R^* exists.

Fix any $a^* \in \text{int}(\Delta(R^*))$. The first part of the proof implies, for any EU preference $\succsim^{R^*}$ on $\Delta(R^*)$,

$$\succsim^{R^*} \in V^{R^*}_{\text{pref}} \quad \Longleftrightarrow \quad a^* \succ^{R^*} b^* \text{ for all } b^* \in W^{R^*}_{\text{inner}}(a^*). \tag{16}$$

Take any $\succsim \in \mathcal{U}_{\text{pref}}$ such that, for all $a \in Z$, $a \succ b$ for all $b \in W_{\text{inner}}(a)$. By (16), the restriction of $\succsim$ to R^* belongs to $V^{R^*}_{\text{pref}}$. Thus, there is some preference in V_{pref} that agrees with $\succsim$ in its restriction to R^*. This preference is pinned down by the indifference conditions derived at the last step: it must be indifferent between a^0 and b^0, a^1 and b^1, and so on. Notice that $\succsim$ must satisfy all these indifference conditions. (For instance, $a^0 \sim b^0$ because $b^0 \in \bar{W}_{\text{inner}}(a^0)$ implies $a^0 \succsim b^0$, while $a^0 + (a^0 - b^0) \in \bar{W}_{\text{inner}}(a^0)$ implies $b^0 \succsim a^0$. An exactly similar argument covers any remaining indifference conditions.) Conclude that $\succsim \in V_{\text{pref}}$.

Second part: By definition,

$$W_{\text{inner}}(a) = \bigcap_{\succsim \in \mathcal{V}_{\text{pref}}} \{b \in \Delta(Z) : a \succ b\}.$$

By the first part, $\succsim \in \bar{\text{co}}(\mathcal{V}_{\text{pref}})$ if and only if $a \succ b$ for all $b \in W_{\text{inner}}(a)$. Since $\mathcal{V}_{\text{pref}}$ is closed and convex, $\succsim \in \mathcal{V}_{\text{pref}}$ if and only if $a \succ b$ for all $b \in W_{\text{inner}}(a)$. □

We now recover $\succsim_u$. For any $A \in \mathcal{A}$, we have $c_2(A|A) = \text{argmax}(A, \succsim_u)$. Fix any $a \in \text{int}(\Delta(Z))$. For any $b, d \in \Delta(Z)^{a_1}$, we have

$$b \succsim_u d \quad \Longleftrightarrow \quad b \in c_2(b, d|b, d).$$

This allows us to recover $\succsim^{a_1}_u$. By NFD, $\succsim_u$ is the unique member of $\mathcal{V}_{\text{pref}}$ that has restriction $\succsim^{a_1}_u$. As usual, the representation of $\succsim_u$ is pinned down up to a positive affine transformation.

We now show that, for any two representations with the same u and γ, for any $\succsim \in \mathcal{V}_{\text{pref}}$, the utilities that represent $\succsim$ differ (at most) by an additive constant. For this step, we need to know $W_{\text{outer}}(a)$. The following lemma shows that $W_{\text{outer}}(a)$ can be fully recovered from choice data.

Definition A.6. *Fix $a \in \Delta(Z)$ and $p \in \Delta(Z)^{a_1}$. Say that p is never chosen over a if there is no $x \in \Delta(Z)$ such that*

$$\{p\} = c_2(a, p | a, p, x).$$

Lemma A.7. *For any regular representation $(u, \mathcal{V})$, any $a \in int(\Delta(Z))$, and any $p \in \Delta(Z)^{a_1}$: $p \in \bar{W}_{outer}(a)$ if and only if, for all $\epsilon \in (0, 1]$, $\epsilon p + (1-\epsilon)a$ is never chosen over a.*

Proof. First, we show: if $p \in \bar{W}_{\text{outer}}(a)$, then p_ϵ is never chosen over a for any $\epsilon \in (0, 1]$. If $p \in \bar{W}_{\text{outer}}(a)$, then $p_\epsilon \in \bar{W}_{\text{outer}}(a)$ as well. Thus, it suffices to show: if $p \in \bar{W}_{\text{outer}}(a)$, then p is never chosen over a. We have

$$\begin{aligned} U(p|a, p, x) &= (1-\gamma)u(p) + \gamma \left(v_p(p) - \max_{a,p,x} v_p \right) \\ &\leq (1-\gamma)u(a) + \gamma \left(v_p(a) - \max_{a,p,x} v_p \right) \\ &\leq U(a|a, p, x), \end{aligned}$$

which is the desired result.

Now we show: if $p \notin \bar{W}_{\text{outer}}(a)$, then there is some $\epsilon \in (0, 1]$ for which p_ϵ is sometimes chosen over a. This is clearly the case if $u(p) > u(a)$. Suppose $u(p) = u(a)$. By assumption, there is some $v \in \mathcal{V}$ such that $((1-\gamma)u + \gamma v)(p) > ((1-\gamma)u + \gamma v)(a)$, so $v(p) > v(a)$. Choose any $\hat{x}$ such that $a \in W_{\text{inner}}(\hat{x})$. Since $v(p) > v(a)$, we have

$$v(p) > v(\epsilon \hat{x} + (1-\epsilon)a)$$

for ϵ close enough to 0. Let $x \equiv \epsilon \hat{x} + (1-\epsilon)a$ for some such ϵ. Since $v(p) > \max_{x,a} v$,

$$U(p|a, p, x) = (1-\gamma)u(p) = (1-\gamma)u(a).$$

Since $a \in W_{\text{inner}}(\hat{x})$, we have $a \in W_{\text{inner}}(x)$, so

$$U(a|a, p, x) < (1-\gamma)u(a) = U(p|a, p, x).$$

Conclude that p is sometimes chosen over a.

Now we cover the final case: $u(p) < u(a)$. Let

$$S \equiv \{v \in \mathcal{V} : ((1-\gamma)u + \gamma v)(a) \geq ((1-\gamma)u + \gamma v)(p)\}.$$

We want to find $\hat{x}$ on the boundary of $W_{\text{inner}}(a)$ such that $v(a) > v(\hat{x})$ for all $v \in S$. Suppose there is no such $\hat{x}$. Then,

$$W_{\text{inner}}(a) = \bigcup_{\succsim \in S_{\text{pref}}} \{b \in \Delta(Z) : a \succ b\}.$$

By Lemma A.5, $S_{\text{pref}} = \bar{\text{co}}(S_{\text{pref}}) = \mathcal{V}_{\text{pref}}$. That is, every $v \in \mathcal{V}$ represents the same preference as some $s \in S \subseteq \mathcal{V}$. By NFD, we must have $v = s$, so $\mathcal{V} = S$. But since $((1-\gamma)u + \gamma v)(p) > ((1-\gamma)u + \gamma v)(a)$ for some $v \in \mathcal{V}$, this cannot be the case. Conclude that there exists $\hat{x}$ on the boundary of $W_{\text{inner}}(a)$ such that $v(a) > v(\hat{x})$ for all $v \in S$. Since $\hat{x} \in \bar{W}_{\text{inner}}(a)$, we must have $v(a) = v(\hat{x})$ for some $v \in \mathcal{V} \setminus S$. For some $\hat{\delta} > 0$, we have $a + \hat{\delta}(a - \hat{x}) \in \Delta(Z)$. Set $x \equiv a + \hat{\delta}(a - \hat{x})$ for some such $\hat{\delta}$. Notice that

$$\begin{aligned} &v(x) > v(a) \text{ for all } v \in S \\ &v(a) = v(x) \text{ for some } v \in \mathcal{V} \setminus S. \end{aligned}$$

Since no $v \in \mathcal{V}$ prefers a to both p and x, we have

$$U(a|a, p_\epsilon, x) = (1-\gamma)u(a) + \gamma \left(v_\epsilon(a) - \max_{p_\epsilon, x} v_\epsilon \right)$$

for all $\epsilon > 0$, where $v_\epsilon \in \text{argmax}_{v \in \mathcal{V}} \{v(a) - \max_{p_\epsilon, x} v\}$. Take a sequence $\{\epsilon_i\} \to 0$. Since $\mathcal{V}$ is compact, we can pass to a convergent subsequence of v_{ϵ_i}. Denote the limit v_0. Fix some $v^* \in \mathcal{V}$ such that $v^*(a) = v^*(x)$. We have

$$\begin{aligned} v_{\epsilon_i}(a) - v_{\epsilon_i}(x) &\geq v_{\epsilon_i}(a) - \max_{p_{\epsilon_i}, x} v_{\epsilon_i} \\ &= \max_{v \in \mathcal{V}} \left\{ v(a) - \max_{p_{\epsilon_i}, x} v \right\} \\ &\geq v^*(a) - \max_{p_{\epsilon_i}, x} v^* \\ &= \min \{v^*(a) - v^*(p_{\epsilon_i}), v^*(a) - v^*(x)\} \\ &= \min \{\epsilon_i (v^*(a) - v^*(p)), 0\}. \end{aligned}$$

Moreover, the limit of the left-hand side is

$$\lim_{i\to\infty} (v_{\epsilon_i}(a) - v_{\epsilon_i}(x)) = v_0(a) - v_0(x)$$

and the limit of the right-hand side is

$$\lim_{i\to\infty} \min\{\epsilon_i(v^*(a) - v^*(p)), 0\} = 0$$

so $v_0(a) \geq v_0(x)$, which implies $v_0 \notin S$, and hence $((1-\gamma)u + \gamma v_0)(p) > ((1-\gamma)u + \gamma v_0)(a)$. For i sufficiently large, the same will be true with v_{ϵ_i} in place of v_0. Using this fact,

$$\begin{aligned} U(a|a, p_{\epsilon_i}, x) &= (1-\gamma)u(a) + \gamma \left(v_{\epsilon_i}(a) - \max_{p_\epsilon, x} v_{\epsilon_i} \right) \\ &\leq (1-\gamma)u(a) + \gamma \left(v_{\epsilon_i}(a) - v_{\epsilon_i}(p_{\epsilon_i}) \right) \\ &= (1-\gamma)(1-\epsilon_i)u(a) + \epsilon_i \left[(1-\gamma)u(a) + \gamma \left(v_{\epsilon_i}(a) - v_{\epsilon_i}(p) \right) \right] \\ &< (1-\gamma)\left[(1-\epsilon_i)u(a) + \epsilon_i u(p) \right] \\ &= (1-\gamma)u(p_{\epsilon_i}) \end{aligned}$$

for i sufficiently large. Since $v_0(p) > v_0(a) = v_0(x)$, we have

$$U(p_{\epsilon_i}|a, p_{\epsilon_i}, x) = (1-\gamma)u(p_{\epsilon_i}).$$

Conclude that, for some $\epsilon \in (0, 1]$, p_ϵ is sometimes chosen over a. □

Fix some $a \in \text{int}(\Delta(Z))$, and let

$$G_{\text{pref}} \equiv \{\succsim \in \mathcal{V}_{\text{pref}} : a \sim b \text{ for some } b \in \bar{W}_{\text{inner}}(a) \text{ s.t. } u(a) > u(b)\}.$$

We will show that, for any two representations with the same u, for any $\succsim \in G_{\text{pref}}$, the utilities that represent $\succsim$ differ (at most) by an additive constant.

Lemma A.8. *Fix some $\succsim \in G_{pref}$. For any regular representation $(u, \mathcal{V})$, a representation v of $\succsim$ belongs to $\mathcal{V}$ only if*

$$((1-\gamma)u + \gamma v)(a) > ((1-\gamma)u + \gamma v)(b) \textit{ for all } b \in W_{outer}(a) \quad (17)$$

$$((1-\gamma)u + \gamma v)(a) = ((1-\gamma)u + \gamma v)(b) \textit{ for some } b \in \bar{W}_{outer}(a) \textit{ s.t. } u(a) > u(b). \quad (18)$$

Proof. Each $v \in \mathcal{V}$ must satisfy (17) by definition of $W_{\text{outer}}(a)$. For (18), suppose that $\hat{v} \in \mathcal{V}$ represents $\succsim \in G_{\text{pref}}$ but $((1-\gamma)u + \gamma\hat{v})(a) > ((1-\gamma)u + \gamma\hat{v})(b)$ for all $b \in \bar{W}_{\text{outer}}(a)$ such that $u(a) > u(b)$. Then, there exists[20] some $\alpha > 1$ such that

$$((1-\gamma)u + \alpha\gamma\hat{v})(a) > ((1-\gamma)u + \alpha\gamma\hat{v})(b) \text{ for all } b \in W_{\text{outer}}(a). \tag{19}$$

We can use exactly the same arguments in the proof of Lemma A.5 to show the following: any preference on $\Delta(Z)^{a_1}$ that strictly prefers a to everything in $W_{\text{outer}}(a)$ must have a representation in $\{(1-\gamma)u^{a_1} + \gamma v^{a_1} : v \in \mathcal{V}\}$. By (19), the preference represented by $(1-\gamma)u^{a_1} + \alpha\gamma\hat{v}^{a_1}$ satisfies this condition. Thus, there exist $\beta_1 > 0$ and $\beta_2 \in \mathbb{R}$ such that

$$\beta_1((1-\gamma)u^{a_1} + \alpha\gamma\hat{v}^{a_1}) + \beta_2 \in \{(1-\gamma)u^{a_1} + \gamma v^{a_1} : v \in \mathcal{V}\}.$$

This implies

$$(\beta_1 - 1)\left(\frac{1-\gamma}{\gamma}\right)u^{a_1} + \alpha\beta_1\hat{v}^{a_1} + \beta_2 \in \mathcal{V}^{a_1}, \tag{20}$$

which is conveniently rewritten

$$\left(\beta_1\left(\frac{1}{\gamma} - 1 + \alpha\right) - \left(\frac{1}{\gamma} - 1\right)\right) \times$$
$$\left(\frac{(\beta_1 - 1)\left(\frac{1}{\gamma} - 1\right)}{\beta_1(\frac{1}{\gamma} - 1 + \alpha) - \left(\frac{1}{\gamma} - 1\right)}u^{a_1} + \frac{\alpha\beta_1}{\beta_1(\frac{1}{\gamma} - 1 + \alpha) - \left(\frac{1}{\gamma} - 1\right)}\hat{v}^{a_1}\right) + \beta_2 \in \mathcal{V}^{a_1}. \tag{21}$$

Suppose $\beta_1 \geq 1$. Since $\mathcal{V}$ is convex and both u and $\hat{v}$ belong to it,

$$\frac{(\beta_1 - 1)\left(\frac{1}{\gamma} - 1\right)}{\beta_1(\frac{1}{\gamma} - 1 + \alpha) - \left(\frac{1}{\gamma} - 1\right)}u^{a_1} + \frac{\alpha\beta_1}{\beta_1(\frac{1}{\gamma} - 1 + \alpha) - \left(\frac{1}{\gamma} - 1\right)}\hat{v}^{a_1} \in \mathcal{V}^{a_1}. \tag{22}$$

Together with NFD, (21) and (22) imply that

$$\beta_1\left(\frac{1}{\gamma} - 1 + \alpha\right) - \left(\frac{1}{\gamma} - 1\right) = 1.$$

Rearranging gives

$$\frac{1}{\gamma}(\beta_1 - 1) + \beta_1(\alpha - 1) = 0.$$

[20] This may not be obvious. For a proof, see the end of this document.

But since $\alpha > 1$ and $\beta_1 \geq 1$, this cannot hold.

Now suppose $\beta_1 < 1$. By (20), we know there exists $\bar{v} \in \mathcal{V}$ such that

$$\begin{aligned} &\bar{v}^{a_1} = (\beta_1 - 1)\left(\frac{1}{\gamma} - 1\right) u^{a_1} + \alpha\beta_1 \hat{v}^{a_1} + \beta_2 \\ \implies \quad &\frac{1}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}\bar{v}^{a_1} + \frac{(1-\beta_1)\left(\frac{1}{\gamma}-1\right)}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}u^{a_1} = \frac{\alpha\beta_1}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}\hat{v}^{a_1} + \frac{\beta_2}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}. \end{aligned}$$

Since u and $\bar{v}$ both belong to $\mathcal{V}$, and since $\mathcal{V}$ is convex,

$$\frac{1}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}\bar{v} + \frac{(1-\beta_1)\left(\frac{1}{\gamma}-1\right)}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}u \in \mathcal{V}.$$

Since this utility represents the same preference on $\Delta(Z)^{a_1}$ that $\hat{v}$ does, and since $\hat{v} \in \mathcal{V}$, NFD implies

$$\frac{1}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}\bar{v} + \frac{(1-\beta_1)\left(\frac{1}{\gamma}-1\right)}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}u = \hat{v}. \tag{23}$$

Since $\hat{v}$ represents a preference in G_{pref}, by definition of G_{pref} there exists $b \in \bar{W}_{\text{inner}}(a)$ such that $u(a) > u(b)$ but $\hat{v}(a) = \hat{v}(b)$. Plugging this into (23) and rearranging gives

$$\bar{v}(a) - \bar{v}(b) = (\beta_1 - 1)\left(\frac{1}{\gamma} - 1\right)(u(a) - u(b)) < 0.$$

We have $\bar{v}(a) < \bar{v}(b)$ for some $b \in \bar{W}_{\text{inner}}(a)$, which contradicts $\bar{v} \in \mathcal{V}$. Conclude that (18) cannot be violated. □

Fix some $\succsim \in G_{\text{pref}}$ and an arbitrary representation $\hat{v}$ of $\succsim$. Suppose that $\alpha\hat{v} + \beta$ and $\alpha'\hat{v} + \beta'$ represent $\succsim$ in $(\gamma, u, \mathcal{V})$ and $(\gamma, u, \mathcal{V}')$ respectively. By Lemma A.8,

$$\begin{aligned} &((1-\gamma)u + \alpha\gamma\hat{v})(a) = ((1-\gamma)u + \alpha\gamma\hat{v})(b) \text{ for some } b \in \bar{W}_{\text{outer}}(a) \text{ s.t. } u(a) > u(b) \\ &((1-\gamma)u + \alpha'\gamma\hat{v})(a) = ((1-\gamma)u + \alpha'\gamma\hat{v})(b') \text{ for some } b' \in \bar{W}_{\text{outer}}(a) \text{ s.t. } u(a) > u(b'). \end{aligned}$$

If $\alpha' > \alpha$, we have

$$\begin{aligned} & (1-\gamma)(u(a)-u(b)) = \alpha\gamma(\hat{v}(b)-\hat{v}(a)) < \alpha'\gamma(\hat{v}(b)-\hat{v}(a)) \\ \implies \quad & ((1-\gamma)u + \alpha'\gamma\hat{v})(a) < ((1-\gamma)u + \alpha'\gamma\hat{v})(b). \end{aligned}$$

Since $b \in \bar{W}_{\text{outer}}(a)$, this contradicts (17). The case $\alpha > \alpha'$ is exactly the same, so we have $\alpha = \alpha'$. Conclude that, conditional on u and γ, two representations of $\succsim$ differ at most by an additive constant.

Now we extend the argument from G_{pref} to the remainder of $\mathcal{V}_{\text{pref}}$. Let

$$G \equiv \{v \in \mathcal{V} : v \text{ represents } \succsim \in G_{\text{pref}}\}.$$

Lemma A.9. *For any regular representation* $(\gamma, u, \mathcal{V})$*,*

1. $\bar{co}(G_{pref}) = \mathcal{V}_{pref}$.
2. $\bar{co}(G) = \mathcal{V}$.

Proof. We first show that $\bar{\text{co}}(G_{\text{pref}} \cup \{\succsim_u\}) = \mathcal{V}_{\text{pref}}$. By Lemma A.5, it suffices to show

$$W_{\text{inner}}(a) = \bigcap_{\succsim \in G_{\text{pref}} \cup \{\succsim_u\}} \{b \in \Delta(Z) : a \succ b\}. \tag{24}$$

Since $G_{\text{pref}} \cup \{\succsim_u\} \subseteq \mathcal{V}_{\text{pref}}$, the left-hand side of (24) is a subset of the right-hand side. Suppose the set inclusion is strict. Since the right-hand side is convex, it includes a boundary point x of $W_{\text{inner}}(a)$. Notice that $u(a) > u(x)$. Since $W_{\text{inner}}(a)$ is an open convex cone, there exists some utility $\hat{v}$ such that $\hat{v}(a) = \hat{v}(x) > \hat{v}(b)$ for all $b \in W_{\text{inner}}(a)$. By Lemma A.5, $\hat{v}$ represents a preference that belongs to $\mathcal{V}_{\text{pref}}$. Call this preference $\hat{\succsim}$. Since $a \mathrel{\hat{\sim}} x$ for some $x \in \bar{W}_{\text{inner}}(a)$ such that $u(a) > u(x)$, $\hat{\succsim} \in G_{\text{pref}}$. But this contradicts the assumption that x belongs to the right-hand side of (24).

Next, we show that $\bar{\text{co}}(G \cup \{u\}) = \mathcal{V}$. Since $\mathcal{V}$ is closed and convex and $G \cup \{u\} \subset \mathcal{V}$, we have $\bar{\text{co}}(G \cup \{u\}) \subseteq \mathcal{V}$. Now take any $v \in \mathcal{V}$. By the first part, v represents a preference in $\bar{\text{co}}(G_{\text{pref}} \cup \{\succsim_u\})$. Every such preference has a representation in $\bar{\text{co}}(G \cup \{u\})$. By NFD, that representation can only be v. Thus, $\mathcal{V} \subseteq \bar{\text{co}}(G \cup \{u\})$.

Fix any $v^* \in \text{ext}(\mathcal{V})$. Since $\mathcal{V} = \bar{\text{co}}(G \cup \{u\})$, v^* can be written

$$v^* = \lim_{i \to \infty} v^i,$$

where each $v^i \in \mathrm{co}(G \cup \{u\})$. In turn, each v^i can be written

$$v^i = \left(1 - \sum_{j=1}^{J^i} \lambda_j^i\right) u + \sum_{j=1}^{J^i} \lambda_j^i g_j^i$$

where the λ_j^i are positive and sum to no more than unity, and each $g_j^i \in G$. Pass to a subsequence such that $\lim_i \sum_{j=1}^{J^i} \lambda_j^i$ converges, and let $\bar{\lambda}$ denote the limit. If $\bar{\lambda} = 0$, then $v^* = u$. This contradicts $u \in \mathrm{relint}(\mathcal{V})$, so $\bar{\lambda} > 0$. Substituting for each v^i gives

$$\begin{aligned}
v^* &= (1-\bar{\lambda})u + \lim_{i\to\infty} \sum_{j=1}^{J^i} \lambda_j^i g_j^i \\
&= (1-\bar{\lambda})u + \lim_{i\to\infty} \left(\sum_{j=1}^{J^i} \lambda_j^i \times \sum_{j=1}^{J^i} \left(\frac{\lambda_j^i}{\sum_{k=1}^{J^i} \lambda_k^i} \right) g_j^i \right) \\
&= (1-\bar{\lambda})u + \bar{\lambda} \lim_{i\to\infty} \sum_{j=1}^{J^i} \left(\frac{\lambda_j^i}{\sum_{k=1}^{J^i} \lambda_k^i} \right) g_j^i \\
&= (1-\bar{\lambda})u + \bar{\lambda}\bar{g} \text{ for some } \bar{g} \in \bar{\mathrm{co}}(G).
\end{aligned}$$

Since $u, \bar{g} \in \mathcal{V}$, both can be written as convex combinations of extremes of $\mathcal{V}$. Since u is non-extreme, it must place weight on at least two distinct extremes. If $\bar{\lambda} < 1$, then the same is true of v^*. Since v^* is extreme, it must be that $\bar{\lambda} = 1$, so $v^* \in \bar{\mathrm{co}}(G)$. Since v^* was an arbitrary member of $\mathrm{ext}(\mathcal{V})$, we have $\mathcal{V} = \bar{\mathrm{co}}(G)$. This implies $\mathcal{V}_{\text{pref}} = \bar{\mathrm{co}}(G_{\text{pref}})$. □

Fix any $\succsim \in \mathcal{V}_{\text{pref}}$. Suppose that v and v' represent $\succsim$ in $(\gamma, u, \mathcal{V})$ and $(\gamma, u, \mathcal{V}')$ respectively. Since $\mathcal{V} = \bar{\mathrm{co}}(G)$, each $v \in \mathcal{V}$ can be written

$$v = \lim_{i\to\infty} v^i$$

where each $v^i \in \mathrm{co}(G)$. In turn, each v^i can be written

$$v^i = \sum_{j=1}^{J^i} \lambda_j^i g_j^i$$

where λ^i is a vector of weights and g^i is a vector of utilities in G. Substituting for each v_i, we have

$$v = \lim_{i \to \infty} \sum_{j=1}^{J^i} \lambda_j^i g_j^i$$

where each $g_j^i \in G$. We already showed that, for each $g_j^i \in G$, there exists a constant β_j^i such that $g_j^i + \beta_j^i \in G'$. Since $\mathcal{V}'$ is convex, it must contain each $\sum_{j=1}^{J^i} \lambda_j^i (g_j^i + \beta_j^i) \in \mathcal{V}'$. Since $\mathcal{V}'$ is compact, we can pass to a convergent subsequence of $\sum_{j=1}^{J^i} \lambda_j^i (g_j^i + \beta_j^i)$. Let v^∞ denote the limit. Since $\mathcal{V}'$ is closed, we have $v^\infty \in \mathcal{V}'$. Moreover,

$$v^\infty - v = \lim_{i \to \infty} \sum_{j=1}^{J_i} \lambda_j^i \beta_j^i \in \mathbb{R}$$

so v^∞ represents $\succ$. Since v^∞ and v' both represent $\succ$ and belong to $\mathcal{V}'$, NFD implies that they are identical. Substituting v' for v^∞ gives

$$v' - v = \lim_{i \to \infty} \sum_{j=1}^{J_i} \lambda_j^i \beta_j^i \in \mathbb{R},$$

so, conditional on u and γ, v' and v differ by an additive constant at most.

Finally, we show that γ is identified. Suppose that $(\gamma, u, \mathcal{V})$ and $(\gamma', u', \mathcal{V}')$ are both regular representations. Since we can always construct a new regular representation by dividing u' and $\mathcal{V}'$ by a positive constant and adding another constant to u', it is without loss to assume that $u = u'$. Since $\mathcal{V}_{\text{pref}}$ and W_{inner} do not differ across representations, we must also have $G_{\text{pref}} = G'_{\text{pref}}$. Fix some $\succsim \in G_{\text{pref}}$. Suppose that g and $g' \equiv \alpha g + \beta$ represent $\succsim$ in G and G' respectively. By Lemma A.8, we must have

$$((1-\gamma')u+\gamma' g')(a) = ((1-\gamma')u+\gamma' g')(b') \text{ for some } b' \in \bar{W}_{\text{outer}}(a) \text{ s.t. } u(a) > u(b').$$

This implies

$$((1-\gamma)u + \gamma k g)(a) = ((1-\gamma)u + \gamma k g)(b') \tag{25}$$

where

$$k \equiv \alpha \frac{\gamma'}{\gamma} \frac{1-\gamma}{1-\gamma'}.$$

Again by Lemma A.8, we must have

$$((1-\gamma)u+\gamma g)(a) = ((1-\gamma)u+\gamma g)(b) \text{ for some } b \in \bar{W}_{\text{outer}}(a) \text{ s.t. } u(a) > u(b). \tag{26}$$

We showed below Lemma A.8 that (25) and (26) together imply $k = 1$. Solving for α gives

$$g' = \frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'}g + \beta.$$

Since $\mathcal{V} = \bar{\text{co}}(G)$ and $u \in \mathcal{V}$, we have

$$u = \sum_i \lambda_i g_i$$

for some weight vector λ and some vector g of utilities in G. By the previous argument, $g_i \in G$ implies

$$\frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'}g_i + \beta_i \in G'$$

for some constant β_i. Since $\mathcal{V}'$ is convex,

$$\sum_i \lambda_i \left(\frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'}g_i + \beta_i\right) = \frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'}u + \sum_i \lambda_i \beta_i \in G'.$$

Since $u \in \mathcal{V}'$, NFD implies

$$\frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'} = 1,$$

which in turn implies $\gamma = \gamma'$.

A.1 Proof of claim in Lemma A.8

Lemma A.10. *Fix a regular representation $(u, \mathcal{V})$ and $a \in int(\Delta(Z))$. Suppose that $\hat{v} \in \mathcal{V}$ represents $\succsim \in G_{pref}$, but $((1-\gamma)u + \gamma\hat{v})(a) > ((1-\gamma)u + \gamma\hat{v})(b)$ for all $b \in \bar{W}_{outer}(a)$ such that $u(a) > u(b)$. Then, there exists some $\alpha > 1$ such that (19) holds.*

Proof. The result is clearly true if $u = \hat{v}$, so suppose $u \neq \hat{v}$. Let

$$I \equiv \{b \in \Delta(Z)^{a_1} : u(a) = u(b) \text{ and } \hat{v}(a) = \hat{v}(b)\}.$$

Let

$$S \equiv \mathrm{co}(I \cup \bar{W}_{\text{outer}}(a)).$$

Using the fact that I and $\bar{W}_{\text{outer}}(a)$ are compact and convex, it is straightforward to show that S is closed.

Since $u \neq \hat{v}$ and $a \in \mathrm{int}(\Delta(Z))$, there must exist $b^* \in \mathrm{int}(\Delta(Z)^{a_1})$ such that

$$u(a) > u(b^*)$$
$$((1-\gamma)u + \gamma\hat{v})(a) = ((1-\gamma)u + \gamma\hat{v})(b^*).$$

Suppose $b^* \in S$. Then, by definition of I and $\bar{W}_{\text{outer}}(a)$,

$$b^* \in \mathrm{co}(\{b \in S : ((1-\gamma)u + \gamma\hat{v})(a) = ((1-\gamma)u + \gamma\hat{v})(b)\}).$$

Moreover, b^* must place positive weight on some $\bar{b} \in \bar{W}_{\text{outer}}(a)$ with $u(a) > u(\bar{b})$ as well as $((1-\gamma)u+\gamma\hat{v})(a) = ((1-\gamma)u+\gamma\hat{v})(\bar{b})$. This contradicts the assumption about $\hat{v}$. Conclude that $b^* \notin S$.

We extend S beyond the simplex as follows:

$$S^{\text{ext}} \equiv \left\{ s \in \mathbb{R}^Z : \sum_Z s(z) = 1 \text{ and } \epsilon s + (1-\epsilon)a \in S \text{ for some } \epsilon > 0 \right\}.$$

It is straightforward to show that S^{ext} inherits closedness and convexity from S. Moreover, since $b^* \in \mathrm{int}(\Delta(Z)^{a_1})$ and S and S^{ext} agree on $\Delta(Z)^{a_1}$, $b^* \notin S^{\text{ext}}$. By the separating hyperplane theorem, there exists some utility w and some constant $\bar{w}$ such that

$$w(b^*) > \bar{w} \geq w(s) \text{ for all } s \in S^{\text{ext}}. \tag{27}$$

Suppose that $w(i) > w(a)$ for some $i \in I$. We have $i + \lambda(i - a) \in S^{\text{ext}}$ for all $\lambda > 0$. But for λ sufficiently large, $w(i + \lambda(i-a)) > \bar{w}$, which contradicts (27). Now suppose $w(i) < w(a)$ for some $i \in I$. By definition of i, we must have $a + \epsilon(a - i) \in I$ for $\epsilon > 0$ sufficiently small. Thus, $a + \lambda(a - i) \in S^{\text{ext}}$ for all $\lambda > 0$. But for λ sufficiently large, $w(a+\lambda(a-i)) > \bar{w}$, which contradicts (27). Conclude that $w(a) = w(i)$ for all $i \in I$. This implies that w can be written as a linear combination of $(1-\gamma)u + \gamma\hat{v}$ and u: for some $\alpha_1, \alpha_2 \in \mathbb{R}$,

$$w = \alpha_1((1-\gamma)u + \gamma\hat{v}) + \alpha_2(u).$$

Since $w(b^*) > w(a)$ but $((1-\gamma)u+\gamma\hat{v})(b^*) = ((1-\gamma)u+\gamma\hat{v})(a)$ and $u(a) > u(b^*)$, we must have $\alpha_2 < 0$.

Let

$$w^* \equiv (1-\epsilon)((1-\gamma)u + \gamma\hat{v}) + \epsilon w$$

for $\epsilon > 0$ small enough that $\epsilon(1-\alpha_1-\alpha_2) < 1$. We can rewrite w^* as follows:

$$\begin{aligned} w^* &= (1-\gamma)(1-\epsilon(1-\alpha_1-\alpha_2))u + \gamma(1-\epsilon(1-\alpha_1))\hat{v} \\ &\propto (1-\gamma)u + \gamma\frac{1-\epsilon(1-\alpha_1)}{1-\epsilon(1-\alpha_1-\alpha_2)}\hat{v} \\ &= (1-\gamma)u + \alpha\gamma\hat{v} \text{ for some } \alpha > 1 \end{aligned}$$

where the final equality uses $\alpha_2 < 0$. Since w^* is a convex combination of $(1-\gamma)u+\gamma v$ and w, and since both of these utilities weakly prefer a to everything in $\bar{W}_{\text{outer}}(a)$, (19) holds. □

B Proof of Theorem 5.1

Our proof strategy is to show that when $\bar{a}_1$ was *ex post* too high, the agent acts according to a rationale that no less than θ^*, and this distorts a_2 upwards compared to the classical case $\gamma = 0$. The key difficulty is that the pairs (a_2, θ) that maximize total utility are not monotone in γ; for instance, all rationales maximize total utility when $\gamma = 0$. In general, there may be rationales strictly below θ^* that maximize total utility.

Suppose $\bar{a}_1$ was *ex post* too high. We prove that if some rationale $\bar{\theta} < \theta^*$ maximizes total utility, then the accompanying action a_2 must maximize material utility. Then we show that whenever rationale $\bar{\theta}$ maximizes total utility, then rationale θ^* also maximizes total utility. Thus, when we are only concerned with actions a_2 that maximize total utility, it is without loss of generality to restrict the rationales to be above θ^*. An argument using Topkis's theorem then yields Theorem 5.1.

In this proof, we fix the state s and the decision problem D with $A_2(a_1)$ monotone non-decreasing, and we suppress the dependence of w on s to reduce notation. We define $\tilde{U}(a_1, a_2, \theta) \equiv U_D(a_2, \theta \mid a_1, s)$.

Let a_1^* be as defined in Theorem 5.1 and let $\bar{a}_1 \geq a_1^*$. We now define

$$\begin{aligned}
&\bar{\Theta} \equiv \operatorname*{argmax}_{\theta \in \Theta} \max_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \theta), \\
&\bar{\Theta}_0^+ \equiv \bar{\Theta} \cap \{\theta \geq \theta^*\}, \\
&\bar{\Theta}^- \equiv \bar{\Theta} \setminus \bar{\Theta}_0^+ = \bar{\Theta} \cap \{\theta < \theta^*\}.
\end{aligned}$$

Observe that

$$\operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \max_{\theta \in \Theta} \tilde{U}(\bar{a}_1, a_2, \theta) = \bigcup_{\bar{\theta} \in \bar{\Theta}} \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}). \tag{28}$$

Lemma B.1. *If w has increasing differences between a_1 and (a_2, θ) and is supermodular in (a_2, θ), then*

$$\bigcup_{\bar{\theta} \in \bar{\Theta}^-} \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) \subseteq \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*). \tag{29}$$

Proof. Take any $\bar{\theta} \in \bar{\Theta}^-$. By a_1^* *ex post* optimal, we have

$$\begin{aligned}
&\max_{a_2 \in A_2(a_1^*)} w(a_1^*, a_2, \bar{\theta}) - \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}) \leq 0 \\
&= \max_{a_2 \in A_2(a_1^*)} w(a_1^*, a_2, \theta^*) - \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \theta^*).
\end{aligned} \tag{30}$$

Let us select some

$$a_2^{**} \in \operatorname*{argmax}_{a_2 \in A_2(a_1^*)} w(a_1^*, a_2, \theta^*) \text{ and } \tilde{a}_2 \in \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \bar{\theta}).$$

Recall that $a_1^* \leq \bar{a}_1$ and $\bar{\theta} < \theta^*$. Observe that by $A_2(a_1)$ monotone non-decreasing, we have $a_2^{**} \wedge \tilde{a}_2 \in A_2(a_1^*)$ and $a_2^{**} \vee \tilde{a}_2 \in A_2(\bar{a}_1)$. Now we have

$$\begin{aligned}
&\max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \theta^*) - \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}) \\
\leq & \max_{a_2 \in A_2(a_1^*)} w(a_1^*, a_2, \theta^*) - \max_{a_2 \in A_2(a_1^*)} w(a_1^*, a_2, \bar{\theta}) && \text{by rearranging (30)} \\
\leq & w(a_1^*, a_2^{**}, \theta^*) - w(a_1^*, a_2^{**} \wedge \tilde{a}_2, \bar{\theta}) && \text{by construction of } a_2^{**} \\
\leq & w(\bar{a}_1, a_2^{**}, \theta^*) - w(\bar{a}_1, a_2^{**} \wedge \tilde{a}_2, \bar{\theta}) && \text{by increasing differences} \\
\leq & w(\bar{a}_1, a_2^{**} \vee \tilde{a}_2, \theta^*) - w(\bar{a}_1, \tilde{a}_2, \bar{\theta}) && \text{by supermodular} \\
\leq & \max_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*) - \max_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \bar{\theta}). && \text{by construction of } \tilde{a}_2
\end{aligned}$$

Rearranging terms yields

$$\begin{aligned}
&\max_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \bar{\theta}) - \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}) \\
\leq & \max_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*) - \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \theta^*).
\end{aligned} \tag{31}$$

Take any

$$\bar{a}_2 \in \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) \text{ and } a_2^* \in \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*).$$

$\bar{\theta} \in \bar{\Theta}^-$ implies that $\tilde{U}(\bar{a}_1, \bar{a}_2, \bar{\theta}) \geq \tilde{U}(\bar{a}_1, a_2^*, \theta^*)$, that is,

$$\begin{aligned}
&(1-\gamma) w(\bar{a}_1, \bar{a}_2, \theta^*) + \gamma \left[w(\bar{a}_1, \bar{a}_2, \bar{\theta}) - \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}) \right] \\
\geq & (1-\gamma) w(\bar{a}_1, a_2^*, \theta^*) + \gamma \left[w(\bar{a}_1, a_2^*, \theta^*) - \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \theta^*) \right].
\end{aligned} \tag{32}$$

By (31), (32), and $\gamma \in [0, 1)$, we have $w(\bar{a}_1, \bar{a}_2, \theta^*) \geq w(\bar{a}_1, a_2^*, \theta^*)$, so

$$\bar{a}_2 \in \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*). \tag{33}$$

We have proved that for any $\bar{\theta} \in \bar{\Theta}^-$,

$$\operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) \subseteq \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*).$$

Taking the union over $\bar{\Theta}^-$ yields (29). $\square$

Lemma B.2. *If w has increasing differences between a_1 and (a_2, θ), then $\tilde{U}$ has increasing differences between a_1 and (a_2, θ).*

Proof. Take any $a_1' \geq a_1$ and any $(a_2', \theta') \geq (a_2, \theta)$. Canceling terms, we have

$$\begin{aligned}
&\tilde{U}(a_1', a_2', \theta') - \tilde{U}(a_1', a_2, \theta) - \left[\tilde{U}(a_1, a_2', \theta') - \tilde{U}(a_1, a_2, \theta)\right] \\
=&(1-\gamma)\left[w(a_1', a_2', \theta^*) - w(a_1', a_2, \theta^*) - \left[w(a_1, a_2', \theta^*) - w(a_1, a_2, \theta^*)\right]\right] \\
&+ \gamma\left[w(a_1', a_2', \theta') - w(a_1', a_2, \theta) - \left[w(a_1, a_2', \theta') - w(a_1, a_2, \theta)\right]\right].
\end{aligned} \tag{34}$$

By increasing differences for w, the term multiplied by $(1-\gamma)$ and the term multiplied by γ are both non-negative, so the right-hand side of (34) is non-negative. $\square$

Lemma B.3. *If w is supermodular in (a_2, θ), then $\tilde{U}$ is supermodular in (a_2, θ).*

Proof. Take any (a_2, θ) and (a_2', θ'). Canceling terms, we have

$$\begin{aligned}
&\tilde{U}(a_1, (a_2, \theta) \vee (a_2', \theta')) + \tilde{U}(a_1, (a_2, \theta) \wedge (a_2', \theta')) \\
&- \tilde{U}(a_1, a_2, \theta) - \tilde{U}(a_1, a_2', \theta') \\
=&\gamma[w(a_1, (a_2, \theta) \vee (a_2', \theta')) + w(a_1, (a_2, \theta) \wedge (a_2', \theta')) \\
&- w(a_1, a_2, \theta) - w(a_1, a_2', \theta')].
\end{aligned} \tag{35}$$

Supermodularity of w in (a_2, θ) implies that the right-hand side of (35) is non-negative. $\square$

Lemma B.4. *Under the assumptions of Theorem 5.1, if $\bar{\Theta}^-$ is non-empty, then*

$$\operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*) \subseteq \bigcup_{\bar{\theta} \in \bar{\Theta}_0^+} \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}). \tag{36}$$

Proof. Take any $\bar{\theta} \in \bar{\Theta}^-$ and let

$$a_2^* \in \operatorname*{argmax}_{a_2 \in A_2(a_1^*)} \tilde{U}(a_1^*, a_2, \theta^*) \text{ and } \bar{a}_2 \in \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}).$$

(a_2^*, θ^*) maximizes $\tilde{U}$ at a_1^*, $(\bar{a}_2, \bar{\theta})$ maximizes $\tilde{U}$ at $\bar{a}_1$, and $a_1^* \leq \bar{a}_1$. By Lemma B.2 and Lemma B.3, $\tilde{U}$ has increasing differences between a_1 and (a_2, θ) and is supermodular in (a_2, θ). Recall that $A_2(a_1)$ is monotone non-decreasing.

Thus, by Topkis's theorem $(a_2^*, \theta^*) \vee (\bar{a}_2, \bar{\theta})$ maximizes $\tilde{U}$ at $\bar{a}_1$, so $\theta^* \in \bar{\Theta}_0^+$. Thus we have

$$\begin{aligned}\underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, w(\bar{a}_1, a_2, \theta^*) &= \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, \tilde{U}(\bar{a}_1, a_2, \theta^*) \\ &\subseteq \bigcup_{\bar{\theta} \in \bar{\Theta}_0^+} \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, \tilde{U}(\bar{a}_1, a_2, \bar{\theta}),\end{aligned}$$

which completes the proof. □

Lemma B.5. *Under the assumptions of Theorem 5.1, we have*

$$\bigcup_{\bar{\theta} \in \bar{\Theta}} \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) = \bigcup_{\bar{\theta} \in \bar{\Theta}_0^+} \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, \tilde{U}(\bar{a}_1, a_2, \bar{\theta})$$

Proof. Recall $\bar{\Theta} = \bar{\Theta}_0^+ \cup \bar{\Theta}^-$, so if $\bar{\Theta}^-$ is empty then we are done. Suppose $\bar{\Theta}^-$ is non-empty. Then Lemma B.1 and Lemma B.4 together imply that

$$\bigcup_{\bar{\theta} \in \bar{\Theta}^-} \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) \subseteq \bigcup_{\bar{\theta} \in \bar{\Theta}_0^+} \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, \tilde{U}(\bar{a}_1, a_2, \bar{\theta})$$

which completes the proof. □

Lemma B.6. *Under the assumptions of Theorem 5.1, we have*

$$\bigcup_{\bar{\theta} \in \bar{\Theta}_0^+} \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) \gg \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, w(\bar{a}_1, a_2, \theta^*). \tag{37}$$

Proof. Take any $\bar{\theta} \in \bar{\Theta}_0^+$. We define

$$g_{\bar{\theta}}(a_2, \gamma) \equiv (1 - \gamma) w(\bar{a}_1, a_2, \theta^*) + \gamma w(\bar{a}_1, a_2, \bar{\theta}).$$

By $\bar{\theta} \geq \theta^*$ and w supermodular, $g_{\bar{\theta}}(a_2, \gamma)$ is supermodular. Thus, by Topkis's theorem, $\operatorname{argmax}_{a_2 \in A_2(\bar{a}_1)} g_{\bar{\theta}}(a_2, \gamma)$ is monotone non-decreasing in γ. Thus we have,

$$\begin{aligned}\underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) = \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, g_{\bar{\theta}}(a_2, \gamma) &\gg \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, g_{\bar{\theta}}(a_2, 0) \\ &= \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}}\, w(\bar{a}_1, a_2, \theta^*).\end{aligned} \tag{38}$$

(38) holds for all $\bar{\theta} \in \bar{\Theta}_0^+$, which implies (37). □

By Lemma B.5, Lemma B.6, and (28), we have

$$\underset{a_2 \in A_2(\bar{a}_1)}{\text{argmax}} \max_{\theta \in \Theta} \tilde{U}(\bar{a}_1, a_2, \theta) \gg \underset{a_2 \in A_2(\bar{a}_1)}{\text{argmax}} \, w(\bar{a}_1, a_2, \theta^*). \tag{39}$$

Moreover, by Lemma B.5, for any a_2 in the left-hand side of (39), there exists $\bar{\theta} \geq \theta^*$ such that

$$\bar{a}_2 \in \underset{a_2 \in A_2(\bar{a}_1)}{\text{argmax}} \, \tilde{U}(\bar{a}_1, a_2, \bar{\theta}).$$

which completes the proof of Theorem 5.1 for the case $\bar{a}_1 \geq a_1^*$. Finally, note that if w has increasing differences and is supermodular, then the function $\tilde{w}(\tilde{a}_1, \tilde{a}_2, \tilde{\theta}) \equiv w(-\tilde{a}_1, -\tilde{a}_2, -\tilde{\theta})$ has increasing differences and is supermodular. Moreover, if $A_2(a_1)$ is monotone non-decreasing, this is preserved when we reverse the order relations on $\mathcal{A}_1$ and $\mathcal{A}_2$. Thus, our proof covers the case $\bar{a}_1 \leq a_1^*$. □

C Proofs of Theorems 6.2 and 6.3

C.1 Proof of Theorem 6.2

A naïve rationalizer facing contract (p, L) predicts that he will choose $q^*(s, p, L)$ in state s. Hence the monopolist's problem is to choose p and L to maximize

$$\int_0^1 (p - c) q_\gamma(s, p, L) dF(s) + L \tag{40}$$

subject to

$$\int_0^1 u^*(s, p, L) dF(s) \geq 0 \tag{41}$$

which is equivalent to

$$\int_0^1 \frac{s^2}{p} dF(s) - L \geq 0. \tag{42}$$

By Equation (9) and substituting $\bar{s} = \sqrt{Lp}$, we transform the monopolist's problem to choose p and $\bar{s} \in [0, 1]$ to maximize

$$\left(\frac{1}{p} - \frac{c}{p^2}\right) \left(\int_0^1 s^2 dF(s) + \int_0^{\bar{s}} 2\gamma s(\bar{s} - s) + \gamma^2 (\bar{s} - s)^2 dF(s)\right) + \frac{\bar{s}^2}{p} \tag{43}$$

subject to

$$\int_0^1 s^2 dF(s) - \bar{s}^2 \geq 0. \tag{44}$$

Suppose Equation (44) does not bind and $p < c$. The derivative of Equation (43) with respect to p is at least

$$\left(-\frac{1}{p^2} + 2\frac{c}{p^3}\right)\int_0^1 \frac{s^2}{p} dF(s) - \frac{1}{p^2}\int_0^1 \frac{s^2}{p} dF(s) \\ = \int_0^1 \frac{s^2}{p} dF(s) \frac{2}{p^2}\left(\frac{c}{p} - 1\right) > 0$$

Hence, if Equation (44) does not bind, then $p \geq c$. For $p \geq c$, Equation (43) is strictly increasing in $\bar{s}$, so Equation (44) binds. Hence, the monopolist's problem reduces to choosing p to maximize

$$\left(\frac{1}{p} - \frac{c}{p^2}\right)\left(\int_0^1 s^2 dF(s) + \int_0^{\bar{s}} 2\gamma s(\bar{s} - s) + \gamma^2(\bar{s} - s)^2 dF(s)\right) \\ + \frac{1}{p}\int_0^1 s^2 dF(s) \quad (45)$$

where $\bar{s}^2 = \int_0^1 s^2 dF(s)$. Taking the first-order condition for Equation (45) and rearranging yields:

$$\frac{p}{c} = \frac{\int_0^1 s^2 dF(s) + \lambda}{\int_0^1 s^2 dF(s) + \frac{\lambda}{2}} \quad (46)$$

where

$$\lambda = \int_0^{\bar{s}} 2\gamma s(\bar{s} - s) + \gamma^2(\bar{s} - s)^2 dF(s). \quad (47)$$

This implies that $p > c$.

To build intuition, consider the derivative of Equation (45) with respect to p, evaluated at $p = c$, which is equal to

$$\frac{1}{p^2}\left(\int_0^1 s^2 dF(s) + \int_0^{\bar{s}} 2\gamma s(\bar{s} - s) + \gamma^2(\bar{s} - s)^2 dF(s)\right) \\ - \frac{1}{p^2}\int_0^1 s^2 dF(s) > 0 \quad (48)$$

Raising p results in more revenue from the per unit price, which is captured by the first term. But it also necessitates a lower upfront payment to satisfy the participation constraint, captured by the second term. The consumer is naïve, so he underestimates his demand, and accepts a reduction in the upfront payment that is less than the rise in revenue from the per-unit price.

C.2 Proof of Theorem 6.3

A sophisticated rationalizer facing contract (p, L) predicts that he will choose $q_\gamma(s, p, L)$ in state s. Hence the monopolist's problem is to choose p and L to maximize

$$\int_0^1 (p-c)q_\gamma(s,p,L)dF(s) + L \tag{49}$$

subject to

$$\int_0^1 2s\sqrt{q_\gamma(s,p,L)} - pq_\gamma(s,p,L)dF(s) - L \geq 0. \tag{50}$$

A parallel argument establishes that the monopolist's problem is equivalently to choose p and $\tilde{s} \in [0,1]$ to maximize Equation (43) but now subject to

$$\int_0^1 s^2 dF(s) - \int_0^{\tilde{s}} \gamma^2(\tilde{s}-s)^2 dF(s) - \tilde{s}^2 \geq 0. \tag{51}$$

The same steps as before establish that Equation (51) binds, which pins down $\tilde{s}$. Substituting the binding constraint into Equation (43) yields

$$\left(\frac{1}{p} - \frac{c}{p^2}\right)\left(\int_0^1 s^2 dF(s) + \int_0^{\tilde{s}} 2\gamma s(\tilde{s}-s) + \gamma^2(\tilde{s}-s)^2 dF(s)\right) \\ + \frac{1}{p}\left(\int_0^1 s^2 dF(s) - \int_0^{\tilde{s}} \gamma^2(\tilde{s}-s)^2 dF(s)\right). \tag{52}$$

Taking the first-order condition for Equation (52) and rearranging yields

$$\frac{p}{c} = \frac{\int_0^1 s^2 dF(s) + \int_0^{\tilde{s}} 2\gamma s(\tilde{s}-s) + \gamma^2(\tilde{s}-s)^2 dF(s)}{\int_0^1 s^2 dF(s) + \int_0^{\tilde{s}} \gamma s(\tilde{s}-s) dF(s)} \tag{53}$$

which implies $p > c$.

To build intuition, the sophisticated consumer's participation constraint is

$$\frac{1}{p}\int_0^1 s^2 dF(s) - \frac{1}{p}\int_0^{\tilde{s}} \gamma^2(\tilde{s}-s)^2 dF(s) - \frac{\tilde{s}^2}{p} \geq 0 \tag{54}$$

where the first term is *ex ante* value to a rational agent of being able to buy the good at per-unit price p, the second term reflects the losses from distorted quantity choices, and the last term is equal to the upfront price. The middle term $-\frac{1}{p}\int_0^{\tilde{s}} \gamma^2(\tilde{s}-s)^2 dF(s)$ is negative, so when the participation constraint binds, the sophisticate pays a lower upfront price than a classical consumer. However, raising p reduces the loss from the distortion. Hence, when we raise

p, the reduction in the upfront payment needed to compensate the sophisticate is *smaller* than for the classical consumer.

The key intuition is to distinguish between levels and compensating variation. The sophisticate foresees and accounts for his later distortion, so he requires a lower upfront price than the classical consumer. But if we increase the per-unit price, the compensating fall in the upfront price is smaller for the sophisticate than for the classical consumer. And it is this compensating variation that matters for optimal pricing.

Now consider the derivative of Equation (52) with respect to p evaluated at $p = c$. This is

$$\frac{1}{p^2}\left(\int_0^1 s^2 dF(s) + \int_0^{\tilde{s}} 2\gamma s(\tilde{s}-s) + \gamma^2(\tilde{s}-s)^2 dF(s)\right) \\ - \frac{1}{p^2}\left(\int_0^1 s^2 dF(s) - \int_0^{\tilde{s}} \gamma^2(\tilde{s}-s)^2 dF(s)\right) > 0. \quad (55)$$

The first term captures the direct effect on profit due to the rise in the per-unit price. The second term reflects the fact that the upfront price has to be lower to compensate the sophisticated consumer.