Losing your Religion: The Inversion of Revealed Preferences under Social Pressure[*]

Moti Michaeli[†]& Daniel Spiro[‡]

Abstract

Can social pressure create a situation where individuals with extreme opinions make statements that are closer to the norm than those made by more moderate people? By analyzing a trade-off between being true to one's private opinion and conforming to a social norm, we show that such inversion of revealed preferences arises rather generally when the utility loss from deviating from one's private opinion increases concavely with the size of the deviation. We demonstrate this result in three model variations. In the first, there is social pressure to conform to an endogenously located norm. In the second, each individual does not want to upset another person whose opinion is unknown to her. Although no real norm exists, we show that this leads to the emergence of a virtual norm. In the third, each person pressures others in declaring a public stance. Here, one particularly notable form of inversion arises, where a fictitious norm may be upheld by those who privately despise it the most. They sanction others for deviating from this norm, thus reinforcing it. Interestingly, the less representative the fictitious norm is of the privately held opinions, the more people publicly support it.

[*] We wish to thank Ram Ben-Shalom, Bård Harstad, Sergiu Hart, John Hassler, Michele Micheletti, Torsten Persson, Kjetil Storesletten, Jörgen Weibull, Robert Östling and seminar participants at Hebrew University, Oslo University, BI, Stockholm University and Stockholm School of Economics for valuable comments.

[†] Department of Economics, and Center for the Study of Rationality, the Hebrew University, motimich@gmail.com. Moti receives funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [249159].

[‡] Corresponding author, Dept of Economics, Oslo University, daniel.spiro@econ.uio.no, Tel: +47 22855137, Fax: +47 22855035.

1 Introduction

It is by now well established that social norms, and social pressure to conform to those norms, influence individual decision making in a wide spectrum of situations. An early experiment showing the potency of social pressure was done by Asch (1955), who showed that even for seemingly objective issues, such as comparing two lines and stating which is the longest, social pressure can have strong effects. In economics, models of social norms have been applied to a variety of issues such as choices of neighborhood (Schelling, 1971), herd behavior (Granowetter, 1978) and unemployment (Lindbeck et al, 2003).[1] This paper investigates the terms under which social pressure leads to *inversion of revealed preferences*. By that we mean a situation where an individual whose private opinion is far from the norm declares in public an opinion that is closer to the norm than the declared opinions of others whose private opinions are closer to it.

Imagine a social or political issue that is under some controversy. To help fix ideas, assume a social norm, which is exogenous from the point of view of an individual, exists. Suppose now that each individual in society has some private opinion regarding this issue, and everyone needs to declare their stance in public. An individual whose private opinion differs from the social norm will then need to take into account the social pressure for violating it and the psychological cost of stating an opinion different than her private one.

We analyze this simple trade-off under three model variations. Firstly, the case where there is one clear norm that serves as the single source of social pressure. One can think of this norm as a tradition or just as a consensual opinion. In the model its location is determined endogenously. We show that if individuals have sufficiently concave disutility from deviating from their private opinions, inversion of revealed preferences will arise.[2] We refer to concave disutility from deviating from one's private opinion as *perfectionism*, since it can be seen as representing the attitude that "only the best is good enough". Here skewed norms that are not representative of the private opinions can emerge, and we characterize the societal traits leading to this.

In our second model variant, there is no one clear norm that the individual feels pressure to conform to. Instead, we analyze the case where the individual does not want to upset another person whose opinion is unknown to her, while feeling disutility from not declaring her

[1] For other examples see Goffman (1959) for early research in sociology, Kandel & Lazear (1992) for an application in economics and management, Kuran (1995) for political revolutions and Holbrook et al (2003) for effects on political survey making.

[2] By *sufficiently* concave we mean more concave than the social pressure function.

private opinion. Before choosing what to state, she will first need to aggregate the pressures from all individuals she can possibly meet into a societal pressure function. We occasionally refer to this society as *pluralistic*, since there are many opinion makers.[3] Here, a *virtual* social norm arises endogenously. By that we refer to some opinion that only few may actually believe in, but such that the aggregated pressure increases monotonically in the distance from it. Inversion of revealed preferences now arises quite generally if individuals are perfectionists. In fact, perfectionism is both a necessary and sufficient condition if opinions in society are heterogenous enough.

Thirdly, we analyze the case where a person declares a stance while simultaneously and indirectly putting pressure on others who deviate from that stance.[4] This is a non-trivial fix point problem where the stances of individuals map to an aggregate social pressure function that affects the stances that created it and so on. Here we resort to numerical analysis of the model. We show that in societies with perfectionist individuals there may well appear a situation of inversion of revealed preferences with a *fictitious norm*: a norm that few actually agree with, but is upheld largely by those most opposing it. That is, those who despise the norm the most sanction others for deviating from it. Meanwhile, people whose private opinions are close to the norm declare those opinions in public. Interestingly, the less representative the fictitious norm is of the privately held opinions, the more people publicly support it and hence the more stable it becomes. Thus, skewed norms and inversion of preferences are mutually reinforcing.

The main conclusion we draw from the whole analysis is that perfectionism is what drives the inversion of revealed preferences. The result that links perfectionism and inversion is not model dependent, but recurs throughout the modeling variations we analyze. The intuition for this is best conveyed in the case of one norm. Perfectionists will tend to stick to their exact private opinions as far as they can. But once a perfectionist does deviate from her private opinion, she may as well do so to the extent necessary to lower the social pressure substantially. Consequentially, a perfectionist will either stick completely to her inner opinion or adapt a great deal. Then, since the greatest social pressure is applied to stances far from the norm, it will be those perfectionists with

[3] Our way of modeling implies that individuals both affect the world and adapt themselves. In psychology this is sometimes referred to as primary and secondary control (Rothbaum et al, 1982).

[4] Note that we do not model this as a conscious or strategic action. It simply means that when a person declares a stance, with the objective of lowering the pressure on herself, this declared stance serves as a source of pressure on other individuals.

extreme opinions who will fail to stick to their private opinions. As a result, these types will often go a long way towards the norm (sometimes the whole way), which means that their declarations will pass those of perfectionist people whose views are more moderate. Thus, we will get a change in the ordering when going from private to revealed preferences.

Finding real-life examples of inversion of revealed preferences requires to know the private opinions of people, which are usually unknown. Yet there exist some real cases where private opinions are obtainable, and that are, at least observationally, consistent with our description of inversion. By this we do not mean to say that other explanations for these observations are not viable too. One example where inversion can be observed is in sexual orientation. It is well known that many (if not most) societies hold negative attitudes towards homosexuality.[5] It is also well documented that not all homosexuals disclose their true sexual preferences (D'Augelli, 2006). Psychological experiments show that disproportionately many of those claiming to be homophobic have homosexual tendencies (Weinstein et al, 2012; Adams et al, 1996). One of the studies further showed that among those stating to be heterosexual, but then measured as homosexual, all also stated to be homophobic (Adams et al, 1996). If one considers homophobia to be antithetic to homosexuality, then these observations constitute an inversion of revealed preferences. Additional casual observations, of religious leaders or members of conservative parties first supporting an anti-gay agenda and then caught conducting homosexual activities, are readily available to anyone surfing the Internet. One interpretation of these observations is that a person, in trying to avoid social pressure, puts pressure on others, and thereby upholds a norm that is disadvantageous to her. This interpretation would be in line with our third model. A related observation of inversion of preferences is regarding women's stated sexual preferences and actual sexual arousal (Morokoff, 1985). We, of course, do not know for certain whether sexual preferences are subject to concave disutility of misrepresenting oneself, neither for homosexuals nor for women. Our model simply predicts that if this would be the case, then we may see inversion as observed.[6]

[5] The Pew research center (2007) documents societies' attitudes towards homosexuals, while Savin-Williams & Ream (2003) document the attitudes of parents towards their own children being homosexual. Many of the attitudes documented in both papers are negative.

[6] In the psychological literature, the term *reaction formation*, one of Freud's classical defense mechanisms, has some resemblance to our term of inversion of revealed preferences. There are, however, differences along a few dimensions. Firstly, reaction formation has the focus on an individual dealing with her own tastes vis-à-vis her thoughts of what is right. Secondly, reaction formation seldom compares indi-

A case where it may, a-priori, be reasonable to expect people to be perfectionists is where religious beliefs are concerned. That is, it seems likely that losing one's religion even by just a little bit comes at a rather high personal cost, while declaring a more distant religious stance feels only slightly worse. An example that may be interpreted as demonstrating inversion of revealed preferences indeed comes from the persecution of Jews in Spain in 1391-1492 AD. During this period, the Christian rulers persecuted Jews fiercely while the Muslim population was persecuted considerably less (Baer, 1965, p.286;[7] Ruiz, 2008, p.160). While Judaism today may seem much closer to Christianity than to Islam, this was not the case in 15th century Spain (e.g., polygamy was practised by both Jews and Muslims). In terms of religious closeness, the consensus among historians seems to be that, in that time and place, Islam stood somewhere in between Christianity and Judaism, and possibly closer to the latter.[8] This suggests a social pressure that is convex, linear, or at least not very concave. Jews wanting to stay in Spain, or who could not move, essentially had three options. Either to stick to their faith, thereby continuing to be persecuted; or to convert to Islam, a religion that was perceived by them to be much closer to Judaism than Christianity was, thereby relieving much of the pressure; or to convert to Christianity, thereby essentially eliminating the persecution, but exacting a higher toll in terms of confessing to something that is very far from their private belief. There is no obvious answer to this choice problem, but by all accounts Jews converted in masses to Christianity or stayed with the original faith, while the conversion of Jews to Islam is essentially unheard of (Baer, 1965). Also conversion of Muslims to Christianity was essentially non-existent during this historical period (Ruiz, 2008, p. 160). This short historical account essentially describes an inversion of revealed preferences. Individuals close to the Christian norm (Muslims) declared their intrinsic opinions, while those far from it (Jews) chose full

viduals with different tastes, while such comparison is in the heart of our concept of inversion. Finally, as a consequence of the focus on the individual's interactions with herself, reaction formation theory stays silent on what norms and pressure are bound to arise in a society. For a brief description and some empirical observations of reaction formation see Baumeister et al (1998).

[7] The page number is given for the original Hebrew version of the book.

[8] To make such a ranking we need to show that Christianity is closer to Islam than to Judaism and that Judaism is closer to Islam than to Christianity. Judaism diverts from both Christianity and Islam in central aspects such as the divinity of Christ and his second coming, the means of salvation and God's role in it and in views on afterlife. See Ben-Shalom, 2001, p.252; Grossman, 1998, pp.30-34; Ben-Sasson, 1990, p.20. for why it would have been more natural for a Jew to convert to Islam rather than to Christianity.

conformity.[9]

There are many ways of modeling social norms. The most common formal approach is to let the stances of individuals be binary (e.g., Brock & Durlauf, 2001; Lopez-Pintado & Watts, 2006). This naturally limits any investigation of the relation between heterogenous private opinions and heterogenous revealed stances. An exception is Bernheim's (1994) work on conformity. Just like in our paper, he investigates not only a continuum of inner blisspoints, but also a continuum of stances from which individuals may choose. Perhaps the most important difference between our model and that of Bernheim is that he assumes people are judged by the opinions they are perceived to privately hold, while we assume that people are judged by their actions regardless of their privately held opinions. Plausibly, there is merit to both approaches, and they lead to substantially different results. That social pressure is applied, like in our model, to actions rather than to hidden types, is assumed also by Jones (1984), who studies the coordination of work effort between workers. Although the setup of his model is quite similar to ours, the results diverge, as he limits the functions to be convex, and looks at essentially only two types of people instead of a continuum. Jones' (1984) framework is more directed at looking at the strategic interaction between people, while our model is mainly focused on situations where the number of individuals is large enough for them to ignore their effects on others (similar to the assumption of price taking firms). Finally, Clark & Oswald (1998) have a model that has some similar structure to ours. As their main question regards emulation between individuals, they focus on how an individual is affected by movements of the group while staying silent about how individuals with different tastes compare.

The paper is structured as follows. First we present a model with general functional forms and outline some general results in section 2. Then we analyze the model with a single norm in section 3. Here we also suggest a system of labels that helps us link modeling assumptions to societal traits. In section 4, as a preparation for the richer models

[9] As already stated, we cannot rule out other mechanisms. It may, for instance, be that partial conversion would not have been acceptable. That is, Muslims of Jewish descent would not have been treated as Muslims but as unconverted Jews. This alternative description may seem reasonable only if the persecutors were able to track individuals and their religion over time. Thus, it may explain the harassment by Christian neighbors, but is less likely to account for the persecution pursued by the central Christian authorities conducting the Spanish Inquisition. Likewise, we cannot rule out that Jews thought the Muslims would be next in line for persecution, thus considered it useless to convert to Islam. Had they correct expectations about this, they need not have worried personally, as persecution of Muslims affected only later generations.

that follow, we analyze the aggregation from individual pressure sources to one societal pressure function. Section 5 then analyzes the stances of individuals when the societal pressure is based on the private opinions of individuals. Finally, in section 6, we analyze the stances of individuals when the societal pressure is based on the stated opinions of individuals. Section 7 concludes.

2 General setup

An individual has a private preference or opinion, referred to also as the individual's *type* $t \in (t_l, t_h)$.[10] The publicly declared stance of type t is her choice variable, denoted by $s(t)$. The inner disutility of an individual of type t declaring some stance s in public is given by

$$D(|t-s|),\ \frac{dD}{d(|t-s|)} > 0.$$

If a person minimizes D only, it is immediate that $s(t) = t$. This way t represents the bliss point of an individual in fulfilling her inner preferences and D can be interpreted as the cognitive dissonance or displeasure felt by taking a stance that is not in line with this bliss point. We can, for example, think of t as the position on a political scale.

Now, assume that an individual that takes s as a public stance feels (social or other) pressure $P(s)$. We implicitly define $\bar{s}$, which can be understood as a social norm, by

$$\frac{dP}{d(|s-\bar{s}|)} > 0.$$

That is, $\bar{s}$ is the stance that induces the lowest social pressure. At this point we abstract from the issue of where $\bar{s}$ comes from. The only restrictions that this formulation implies are that there is a unique global and local min point, and that the social pressure is rising symmetrically and monotonically with the size of the deviation from it. This formulation is imposed as an assumption in the basic model with one exogenous norm, but in later sections of the paper it arises endogenously.

The total disutility (or loss) of an individual is then the sum of the cognitive dissonance and the social pressure,

$$L(t,s) = D(t,s) + P(s,\bar{s}). \tag{1}$$

[10] This formulation implies that we treat t as scalar, thus imposing a structure where preferences are ordered on one axis. We do so for simplicity, but our results are applicable to multidimensional representations too, as the example of religions in Medieval Spain illustrates.

So, on the one hand, the individual feels an increasing inner dissonance from taking a stance different than the bliss point. On the other hand, more social pressure is exerted the further the stance is from the norm. Thus, it is immediate that each individual will take a stance somewhere in between (and including) her inner blisspoint and the social norm. That is,

$$\forall t, s^* (t) \in \begin{cases} [\bar{s}, t], \text{ if } \bar{s} \leq t \\ [t, \bar{s}], \text{ if } t > \bar{s} \end{cases},$$

where $s^* (t)$ is the stance that minimizes the loss for type t.[11] For the sake of brevity, we restrict our analysis to the range $\bar{s} \leq t$. The analysis for $\bar{s} > t$ is similar. The first-order condition

$$L' = P' (s - \bar{s}) - D' (t - s), \tag{2}$$

is equal to zero in inner extreme points while the second-order condition,

$$L'' = P'' (s - \bar{s}) + D'' (t - s), \tag{3}$$

is positive in minimum points. Denoting the optimal stance by s^*, we then have that in inner solutions

$$P' (s^* - \bar{s}) = D' (t - s^*). \tag{4}$$

We now turn to look at the function describing the inner solution (if it exists) for every t. More specifically, we concentrate on ranges of t for which the inner solution exists, and where $s^*(t)$ is continuous and twice differentiable.[12] Then, using the implicit function theorem, we get that

$$\frac{ds^*}{dt} = \frac{D'' (t - s^*)}{P'' (s^* - \bar{s}) + D'' (t - s^*)}. \tag{5}$$

To compare the extent of conformity to the norm we will use the following measure.

Definition 1 *The conformity of* t *is* $- |s^* (t) - \bar{s}|$.

The conformity measures how close to the norm an individual's stance is. The closer to zero (i.e., the larger the value), the more a person conforms. Thus, we say that t conforms more than t' if $|s^* (t) - \bar{s}| \leq |s^* (t') - \bar{s}|$.

[11] A sufficient condition for the validity of the upcoming analysis is that both P and D are three times continuously differentiable.

[12] This implies that we only look at ranges either where the solution is unique or where there are no discrete jumps between solutions.

Definition 2 *Inversion of revealed preferences constitutes a situation where for some t and t', with $|t-\bar{s}| < |t'-\bar{s}|$, t' has a strictly higher level of conformity than t.*

That is, we attach the label 'inversion of revealed preferences' to situations where a person whose private preference is far from the norm takes a stance that is closer to the norm than the stance taken by some other person whose private preference is closer to it.

Following these definitions, Lemma 1 states a sufficient condition for having inversion of revealed preferences for types above the norm.[13]

Lemma 1 *For $t \geq \bar{s}$, inversion of revealed preferences occurs if there exists some type t with an inner solution s^*, such that $L(t, s^*) < L(t, s)$ $\forall s \neq s^*$ and $D''(t - s^*) < 0$.*

Proof. *If t has an inner solution s^* and $L(t, s^*) < L(t, s)$ $\forall s \neq s^*$, then there is a neighborhood of t where all types have inner solutions, $D'' < 0$, and $s^*(t)$ is continuous. Thus, in this neighborhood equation 5 applies. In min points, the denominator, $L'' = P''(s^* - \bar{s}) + D''(t - s^*)$, is positive. Hence, since $D''(t - s^*) < 0$, we get that $\frac{ds^*}{dt}$ is negative. Inversion of revealed preferences in this neighborhood of t then trivially follows from definition 2.* ■

This Lemma alludes to one of our main results. That is, concave displeasure from pretence *may* lead to inversion of revealed preferences. The word 'may' is emphasized to highlight that at this point it is not clear yet whether types with concave displeasure from pretence indeed have an inner solution to the optimization problem, and in case they do not, whether corner solutions can induce the same phenomenon.

3 A model with one norm

We start with a basic model in which there is one norm that is determined by the average declared stance in society. That is, we explicitly assume the existence of a social norm, $\bar{s}$, and a predetermined social pressure function to conform to that norm, but the location of this norm is endogenous. $\bar{s}$ may represent either a perceived social norm, such as a consensual opinion, together with a social pressure to conform to it, or some institution that sanctions deviations from a rule of conduct.

For conservation of space, we use the symmetry of the functions $D(t, s)$ and $P(s, \bar{s})$ to present the problem and solution only for $t \geq \bar{s}$. In addition, for tractability and to facilitate the interpretation, we

[13] Equivalent statements apply to the range $t < \bar{s}$.

will now assume that cognitive dissonance and social pressure are power functions.[14]

$$\begin{aligned} D\left(t,s\right) &= A\left|t-s\right|^{\alpha} \ \ , \ \alpha \geq 0 \\ P\left(s,\bar{s}\right) &= B\left|s-\bar{s}\right|^{\beta} \ \ , \ \beta \geq 0. \end{aligned}$$

At this point, it may be useful for the intuition to provide a loose interpretation of the parameters. While this, of course, does not affect the mathematical validity of the results, it eases interpretation by characterizing the societal traits under which certain phenomena arise. For that purpose, we normalize the coefficient of $D\left(t,s\right)$ by setting $A=1$, and set $K \equiv B/A$. Thus K represents the weight of social pressure relative to the cognitive dissonance. For example, if P represents legal sanctioning and not social pressure per se, then K captures the harshness of the judicial punishment system in general. In comparison, β captures how different deviations from the norm are sanctioned in relation to each other. When $\beta < 1$ the pressure is concave, hence already small deviations from the established norm or rule are fairly heavily sanctioned, but only a minor distinction is made between small and large deviations. We believe that this kind of punctiliousness represents "orthodox" societies, since they often emphasize being "true to the book" but do not distinguish so much between large and small wrongdoings[15].[16] As a counter-label, we use the term "liberal" to represent societies with convex social pressure ($\beta > 1$), i.e., societies that are not very meticulous about small non-normative expressions, as long as they are not too far from the consensus. We do not claim that the conventional usage of the term 'liberal' has to do with convex social pressure. However, in the absence of a better word,

[14] For a treatment of more general functional forms see Michaeli & Spiro (2012). There it is shown that most of the upcoming results hold also more generally.

[15] One practical example for concave pressure is the Taliban society, where there are numerous accounts of capital punishment for both misdemeanor and larger crimes. Another example is some Jewish Ultraorthodox communities in Jerusalem, where substantially less harsh yet concave punishment is used when ostracizing individuals for both small and large deviations from their norms. Concave punishment can also be observed when using the data in Herrmann et al (2008), that shows that the patterns of social punishment in Muscat and Riyadh for deviations in a Public Good Game experiment are concave. For a more detailed account of these examples see Michaeli & Spiro (2013).

[16] We acknowledge that, in its everyday use, the term 'orthodox society' may refer not only to a society that is very particular about small norm deviations, but also to society that has harsh punishments. However, we show in section 3.4 that if norms in an orthodox society ($\beta < 1$) are determined endogenously and yet are skewed with respect to the average private opinion in society, K must be large. Hence, we may in these situations expect a correlation between a harsh punishment in general and the orthodoxy of the society.

we use this label as liberal democracies usually do not have any formal sanctioning of expressed opinions as long as the opinion does not deviate too much from the consensus, in which case a person may be subject to surveillance or put in jail.[17] As for α, if $\alpha < 1$ then individuals gain concave displeasure of stating something else than their private opinion. We label such individuals "perfectionists", i.e., once a perfectionist deviates even slightly from her bliss point, it makes little marginal difference for her to deviate more. As an opposite label, we use the term "lax" to represent $\alpha > 1$: as long as a person does not "lie" too much, by taking a stance far from her true type, the psychological cost stays rather small. We will assume throughout that the only difference between individuals in society is their blisspoint t.[18]

In the case of one norm, a first result is the following:

Lemma 2 *Only if $\alpha < 1$ we have inversion of revealed preferences.*

Proof. See the appendix. ■

That is, perfectionism is a necessary condition for inversion of opinions in the presence of one norm.[19] However, this condition is not sufficient. We therefore turn now to investigate the behavior of perfectionist individuals in orthodox and liberal societies.

3.1 Orthodox society with perfectionist individuals

When $\beta \leq 1$, society is intolerant to small deviations from the consensus, but does not distinguish much between moderate and large deviations. Likewise, when $\alpha \leq 1$, people are already sensitive to small deviations from their inner blisspoint, but additional deviation does not add much to their dissonance.

It is now immediate from the second-order condition (3) that any local extremum is a maximum, implying that optimality will be found at either of the corners (at $s^*(t) = t$ or at $s^*(t) = \bar{s}$). This is also intuitive, since when the functions are concave, taking a stance in between t and $\bar{s}$

[17] Examples of this are the prohibition of Nazi parties in Germany, and the usage, in various European countries, of police surveillance and interrogation against people expressing opinions that are too leftists, too rightist or too Islamic. In Hermann et al's (2008) study, we also note that the patterns of social punishment against deviations in the Public Good Game in places such as Melbourne, Copenhagen and Bonn are convex. For a more detailed account see Michaeli & Spiro (2013).

[18] Thus , we abstract from the possibility of varying the parameters α, β and K at the individual level.

[19] This result can be generalized. If P is convex then the Lemma holds for any D. If P is concave, then roughly speaking the lemma holds for any D as long as there is at most one local minimum of L $\forall t$.

inflicts both great dissonance and heavy pressure on the individual. The results are summarized in the following proposition.[20]

Proposition 1 *If $\beta \leq 1$ and $\alpha \leq 1$:*

1. *Individuals either declare the norm or their private opinion as their stance.*

2. *If $\beta \leq \alpha$ then conformity is weakly decreasing with the type's distance from the norm. There is no inversion of revealed preferences.*

3. *If $\beta > \alpha$ then conformity is non-monotonic as a function of the types' distance from the norm. If the range of types is broad enough then there is inversion of revealed preferences.*

Proof. *1) The second-order condition (equation 3) is positive, implying no inner solution. The corner solutions are then either $L(s = \bar{s}) = |t - \bar{s}|^{\alpha}$ or $L(s = t) = K|t - \bar{s}|^{\beta}$. 2) If $\beta < \alpha$ then $L(s = \bar{s}) < L(s = t)$ iff $|t - \bar{s}| < K^{\frac{1}{\alpha - \beta}}$ which implies that t close to $\bar{s}$ choose $s^*(t) = \bar{s}$ while those far from $\bar{s}$ choose $s^*(t) = t$. If $\beta = \alpha$, then $s^*(t) = \bar{s}\ \forall t$ iff $K > 1$ and $s^*(t) = t\ \forall t$ iff $K < 1$. Hence follow the statements on conformity and inversion. 3) If $\beta > \alpha$ then $L(s = \bar{s}) < L(s = t)$ iff $|t - \bar{s}| > K^{\frac{1}{\alpha - \beta}}$ which implies that t far from $\bar{s}$ choose $s^*(t) = \bar{s}$ while those close to $\bar{s}$ choose $s^*(t) = t$. Hence follow the statements on conformity and inversion.* ■

Part 1 of Proposition 1 highlights that perfectionist individuals in orthodox societies display no compromise. People will either "follow their heart" or conform fully to the norm. This represents the "either-or mentality" of such individuals living in orthodox society: for them there is no point in stating preferences close to the truth unless it is the exact truth, and there is no point in moving towards the norm unless they fully conform.

Moreover, this proposition shows (in part 2 and 3) that it is the concavity of the dissonance - i.e. perfectionism - that drives the inversion of revealed opinions. Here, when both functions are concave and all individuals choose corner solutions, the dissonance needs to be more concave than the social pressure for inversion to emerge.

The results of part 2 of the above proposition are depicted in Figure 1, where society is "more orthodox than people are perfectionist". Individuals with opinions close enough to the social norm ($t \in$

[20] We ignore here the degenerate case of $\alpha = \beta$ and $K = 1$, in which all individuals are indifferent between at least two stances.

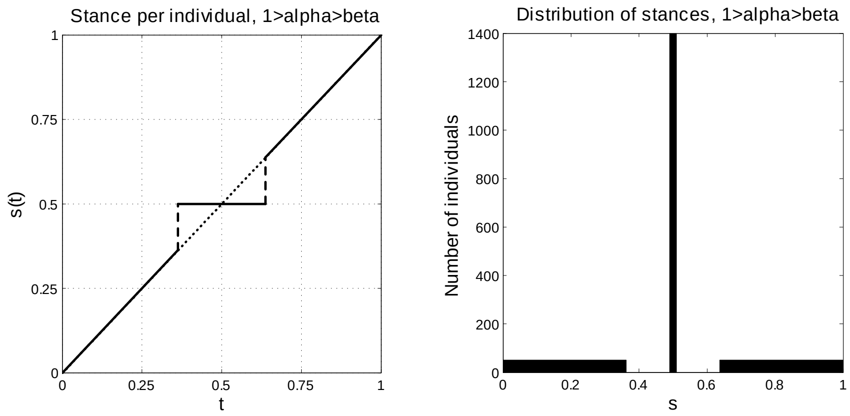

Figure 1: $\beta < \alpha \leq 1$ with $\bar{s} = .5$, $t \in [0,1]$. The left-hand schedule depicts $s^*(t)$ (full line) and $s = t$ (dashed line). The right-hand schedule depicts the probability distribution function of stances under a uniform distribution of types.

$\left[\bar{s} - K^{\frac{1}{\alpha-\beta}}, \bar{s} + K^{\frac{1}{\alpha-\beta}}\right]$) choose to fully conform, while individuals with opinions far enough from the norm simply cope with the social pressure and choose the inner bliss points as their stances. The intuition is that these "extreme" people hold private opinions so distinct from the norm, that they are unwilling to take stances that are close enough to the norm to alleviate the pressure.[21] Then, since deviating even a little from one's inner bliss point is very painful, they might as well take stances that are completely in line with their private opinions. Altogether, this creates alienation in society, where one either conforms fully to the norm or follows one's heart. This way a society that is not tolerant to small deviations from the norm will tend not to succeed in moderating the stances of extreme and perfectionist people.[22]

The mirror image of the previous case – a social pressure which is "less orthodox" than individuals are perfectionist (part 3 of proposition 1) - is depicted in Figure 2 . The observable outcome of this case is a distribution that somewhat resembles a standard bell-shape with mass towards the middle.[23] But there is an important twist. The concentration of stances at $\bar{s}$ consists of individuals with extreme inner blisspoints. That is, the extreme types' declarations are more moderate than those of the moderates. This means that conformity is increasing with the type's distance from the norm. That is, there is inversion of revealed

[21] Remember that in an orthodox society the stance should be very close to the norm in order to alleviate the pressure.

[22] Unless, of course, K is sufficiently large.

[23] At least when the initial distribution of types is uniform.

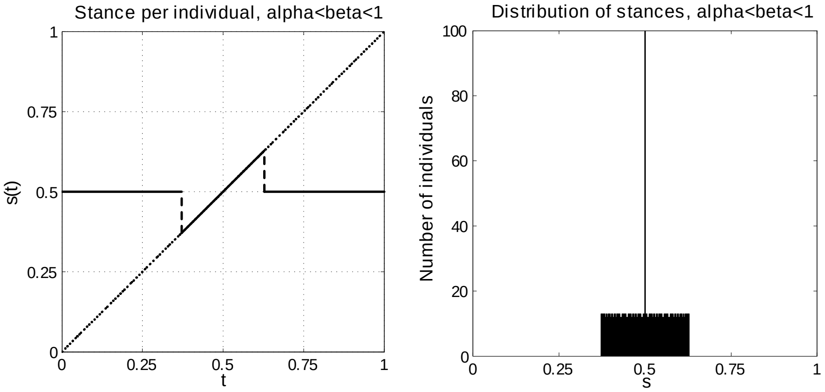

Figure 2: $\alpha < \beta \leq 1$ with $\bar{s} = .5$, $t \in [0,1]$. The left-hand schedule depicts $s^*(t)$ (full line) and $s = t$ (dashed line). The right-hand schedule depicts the probability distribution function of stances under a uniform distribution of types.

preferences (if the range of types is broad enough to include those extremists). The intuition is that when the dissonance is relatively more concave, moderates are not willing to conform since this would inflict too great displeasure. For extremists, however, not conforming implies much greater social pressure, since $P(t, \bar{s})$ is increasing relative to $D(t, \bar{s})$ with the distance from the norm ($|t - \bar{s}|$).

3.2 Liberal society with perfectionist individuals

When $\beta > 1$ and $\alpha < 1$, we get a combination of corner and inner solutions, and inversion of revealed preferences may emerge.[24]

Proposition 2 *If $\alpha < 1 < \beta$ then:*

1. *A type sufficiently close to the norm states her private opinion as a stance, while types sufficiently far from the norm partially conform.*
2. *If the range of types is broad enough, then conformity is non-monotonic in types' distance from the norm, and there is inversion of revealed preferences.*

Proof. *See the appendix.* ■

Since social pressure is convex, it hardly affects moderates, who can openly declare their private opinions at a small social cost. Extremists,

[24] Inversion in line with the upcoming proposition arrives under more general functional forms as long as $P'(x) > D'(x)$ for some $x > 0$.

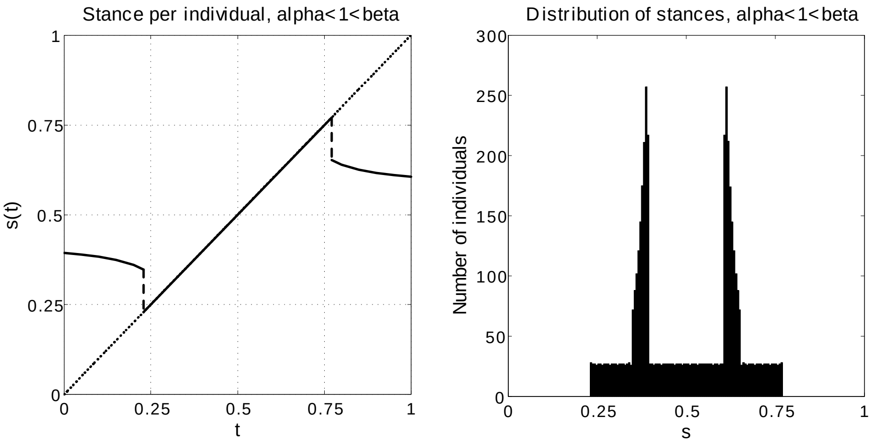

Figure 3: $\alpha < 1 < \beta$ with $\bar{s} = .5$, $t \in [0, 1]$. The left-hand schedule depicts $s^* (t)$ (full line) and $s = t$ (dashed line). The right-hand schedule depicts the probability distribution function of stances under a uniform distribution of types.

on the other hand, would feel too much pressure if they chose to openly declare their private opinions. Hence, and since individuals are perfectionist, once they deviate from their private opinions they might as well conform a great deal.

As illustrated in figure 3 (left-hand schedule), the proposition implies that the extremists conform to the norm more than some moderates do. Furthermore, within the group of those conforming, the more extreme is the individual, the more she conforms. Thus, as in the case of $\alpha < \beta \leq 1$, we get inversion of revealed preferences.[25] But now we get it at two levels – both *between* extremists and moderates and *within* the group of extremists. If the distribution of types is uniform, then the distribution of stances is bimodal.

3.3 Conditions for inversion of revealed preferences

Having gone through all possible cases in the basic model with one social norm, we are now ready to state a general result that describes under what circumstances inversion of revealed preferences arises.

Corollary 3 *Iff $\alpha < 1$, $\alpha < \beta$, and the range of types is broad enough, then there is inversion of revealed preferences.*
Proof. *Follows from Lemma 2 and Propositions 1-2.* ■

[25] That is, if the range of types is broad enough to include those extremists.

This corollary establishes the general pattern: in societies with one norm, perfectionism is a necessary and sufficient condition for the inversion of revealed preferences, as long as society is not too orthodox. The intuition for this is quite simple. A perfectionist will tend to stick to her exact private opinions, but if stating something else, she may as well do so to the extent necessary to lower the social pressure substantially. But in order to get a perfectionist to deviate from her private opinion, a great social pressure is needed. When pressure increases monotonically with the distance from the norm, like has been assumed here, it will be harder for people with extreme opinions to declare them. As a result, they will often go a long way towards the norm (sometimes the whole way), which means that their declarations will pass those of perfectionist people whose views are more moderate. To see how robust this result is, in later sections we will investigate whether a monotonically increasing social pressure function arrives endogenously in more complex models of social interaction.

3.4 Endogenizing the social norm

In order to establish which social norms can arise endogenously, we assume now that the norm is determined by the average declared stance in society. Naturally, there may be other forces shaping the equilibrium position of a social norm, but the average stance seems like a reasonable first case to investigate.

To simplify things, we assume that the distribution of types is uniform. Then the endogenous social norm is determined as follows.

$$\bar{s} = \frac{1}{t_h - t_l} \int_{t_l}^{t_h} s^* (\tau) \, d\tau \tag{6}$$

Focusing our attention on the situations in which inversion of revealed preferences emerges, we get the following result.[26]

Proposition 4 *If $\bar{s}$ is the average stance in society, $\alpha < 1$ and $\alpha < \beta$, then:*

1. *If $\alpha < 1 < \beta$, then the average private opinion, i.e. $\frac{t_h + t_l}{2}$, is the only value of $\bar{s}$ that can be sustained as a social norm in equilibrium.*

2. *If $\alpha < \beta \leq 1$, then $\bar{s}$ can be sustained as a social norm in equilibrium iff $\bar{s} \in \left\{ \frac{t_h + t_l}{2} \right\} \cup \left[t_l + K^{\frac{1}{\alpha - \beta}}, t_h - K^{\frac{1}{\alpha - \beta}} \right]$.*

[26] For a complete analysis see Michaeli & Spiro (2012).

Proof. *See the appendix.* ■

Part 1 of the proposition singles out the average private opinion $\frac{t_h+t_l}{2}$ as the only possible norm in liberal societies, while part 2 of the proposition, about orthodox societies, says that if the weight of the social pressure is large enough, there is a continuous range of possible norms.

How can a norm that is, say, skewed to the right, be the average of declared opinions? First recall that in an orthodox society with perfectionist individuals, each person either fully conforms or declares her exact private opinion. Next, note that in a society that is "less orthodox" than people are perfectionist, types close to the norm express their exact opinions (see Proposition 1). The extreme types do not play any role here since they choose to totally conform, thus giving up their effect in determining the norm. The only aspect of importance for the stability of the norm is whether it is in the center of those who speak their mind. As such, it can be heavily biased, though it cannot be located at the very extreme edges of the range of private opinions.[27] This cannot happen in a liberal society precisely because extreme individuals do not conform fully to the norm. Thus, if the norm is skewed, there will be more people on one side of it, which is unsustainable.

A pattern emerges from the above proposition, where liberal societies are bound to eventually have norms representing the average inner opinions in society.[28] Only orthodox societies can sustain social norms that are not representative of the private opinions of the people.[29] Thus, orthodox societies are bound to be history dependent, as the initial common rule also determines the long-run equilibrium outcome. This predicts that orthodox societies, to a larger extent than liberal ones, will have rules that are not representative of people's opinions on average. Furthermore, it rationalizes why orthodox societies with extremist rules would more often resort to harsh punishments than liberal societies – only in the former is it possible to sustain skewed norms with the help of pressure. Therefore, we should observe a correlation between orthodox societies, harsh punishments and skewed norms.

[27] Unless K is sufficiently large.

[28] Michaeli & Spiro (2012) show that this result holds for every liberal society with symmetric distribution of types, regardless of the curvature of the dissonance function of individuals (i.e., also if there is no inversion).

[29] This is true for any orthodox society under some additional conditions (see Michaeli & Spiro, 2012). If we extend the analysis to other distributions of types, then, if $\alpha < \beta < 1$, it is essentially sufficient that the distribution of types around the endogenous norm is uniform. Outside of this range we can have any symmetric or assymetric distribution conceivable as these individuals will conform fully anyway.

4 Aggregating individual pressure

Up till now we assumed the existence of a unique norm in society. This should not necessarily always be the case. Some societies have a couple of norms, while in others the number of norms is quite large. Moreover, in pluralistic societies it is often customary to argue on debatable issues, where both sides are putting some pressure on each other in their effort to persuade. In this case we can think of individuals as the sources of pressure.

Another related issue concerns the kind of pressure that is felt by an individual in the presence of a representative group of her society: does she feel a pressure to conform to the *average* stance of the group members, as we modeled in the previous section, or should the aggregation of pressures be performed differently? Hence we turn here to answer the following question: if people are pressuring each other, what will the aggregated social pressure function look like?

This analysis focuses on the aggregation of pressure sources into a social pressure function, while ignoring the actual stances individuals take. It will be used as a building block for the next two models. While cases with only a few norms require a separate analysis, once there are sufficiently many norms a continuum of pressure sources is a good approximation. In particular, we will use a uniform distribution of sources of pressure to model a pluralistic society, while staying agnostic about the exact identity of these sources.[30] Since a uniform distribution of pressure sources serves as an opposite benchmark to that of having a single norm, we then turn to investigate a combination of an authority and a uniform distribution of pressure sources.

Consider an individual who declares stance s in a society with multiple sources of pressure. Each source of pressure x inflicts on the individual a pressure that is increasing in the distance from s to x:

$$\begin{aligned} p(s,x) &= p(|s-x|) \\ p(\cdot)' &> 0 \end{aligned}$$

4.1 Uniform sources of pressure

We start with the benchmark case in which x is uniformly distributed from x_l to x_h. This case is particularly useful for investigating society in which the private opinions of people are the sources of pressure, as we explicitly analyze in the next section. If x is uniformly distributed, an

[30] One can think of these sources as representing the private opinions of people, their public stances, any mixture of those, or just a multitude of social norms along the axis of possible stances.

individual with stance s can expect to perceive the following pressure:

$$P_{aggr}(s) \equiv E\left[p\left(|s-x|\right)\right] = \frac{1}{x_h - x_l}\int\limits_{x_l}^{x_h} p\left(|s-x|\right)dx$$
$$= \frac{1}{x_h - x_l}\left[P\left(x_h - s\right) + P\left(s - x_l\right) - 2P(0)\right],\ s \in [x_l, x_h]$$

where, by convention, $P' \equiv p$.[31] If one thinks of the sources of pressure as stemming from individuals in society, then this is the pressure that an individual with stance s can expect to feel when meeting people randomly. What are the properties of this aggregated pressure function? By differentiating P_{aggr} with respect to s, we get

$$P'_{aggr}(s) = p\left(s - x_l\right) - p\left(x_h - s\right) \quad (7)$$
$$P''_{aggr}(s) = p'\left(s - x_l\right) + p'\left(x_h - s\right). \quad (8)$$

Now, define $\bar{s} \equiv \frac{x_l + x_h}{2}$. The following lemma then follows.

Lemma 3 *If x is uniformly distributed, then the aggregated pressure function $P_{aggr}(s)$:*

1. *Is strictly increasing if $s > \bar{s}$ and strictly decreasing if $s < \bar{s}$.*

2. *Is strictly convex if $s \neq \bar{s}$ and $s \in]x_l, x_h[$.*

3. *Has a zero derivative at $\bar{s}$.*

4. *Is symmetric around $\bar{s}$.*

Proof. *1), 2) and 3) follow trivially from the first and second derivatives of P_{aggr}, from $p(\cdot)' > 0$, and from substituting $\bar{s}$ in equation 7. 4) To see the symmetry, let $\tilde{s}$ be the mirror image of s, i.e. $(s + \tilde{s})/2 = \bar{s}$, hence $\tilde{s} = 2\bar{s} - s = x_h + x_l - s$. Then we get $P_{aggr}(\tilde{s}) = P\left(\tilde{s} - x_l\right) + P\left(x_h - \tilde{s}\right) = P\left(x_h + x_l - s - x_l\right) + P\left(x_h - \left(x_h + x_l - s\right)\right) = P\left(x_h - s\right) + P\left(s - x_l\right) = P_{aggr}(s)$.* ■

$P_{aggr}(s)$ has a unique minimum point at $\bar{s} \equiv \frac{x_l + x_h}{2}$ around which it is symmetric. This suggests that qualitatively, the aggregation of punishment from multiple sources of pressure is similar to having a "virtual" social norm at $\frac{x_l + x_h}{2}$.

That social pressure is increasing with the distance from the average x is perhaps not very surprising: the more extreme is one's stance, the

[31] We assume that p is integrable.

more pressure one feels. But that the aggregated pressure function is *convex* may be less obvious, given that we have not specified whether the pressure stemming from each source is convex or concave. This means that under a uniform distribution of sources of pressure, the aggregate social pressure will be convex even if the one-on-one pressure is concave. So a society made up of "orthodox" individuals with uniform tastes who pressure each other will be "liberal" on aggregate. The intuition is that even if an individual is pressured by another person in a concave manner, the fact that there are other people pressuring from another direction undermines the concavity coming from the first person (when you move away from one person you also move towards someone else). It implies that orthodoxy cannot be maintained under such conditions. We know, however, that if there is a unique norm, society *can* be orthodox. This raises a natural question, to which we turn now, about the curvature of the social pressure when combining these two stylized cases.

4.2 Combining uniform pressure with a single norm

When modeling a combination of uniform pressure and a single norm, we think of a society with two types of pressure coexisting together. First, aggregate pressure of individuals who pressure each other, and second, an institutionalized norm that is the average of those pressuring individuals. In this section, it is useful to think of an authority that sets and enforces the institutionalized norm, or a group of people doing it. This analysis turns out to be particularly useful for our third model in Section 6.

In order to model orthodoxy, we look at the case where both individual and institutional pressures have the same concave functional form, p. The total pressure function is then a weighted average of the two:

$$P_{combi}\left(s\right)=\left(1-A\right)P_{aggr}+Ap\left(\left|s-\frac{x_h+x_l}{2}\right|\right) \tag{9}$$

where A is the relative weight of the institutional pressure.

Lemma 4 *If x is uniformly distributed, and p is symmetric around $\bar{s}=\frac{x_l+x_h}{2}$, concavely increasing in $|s-\bar{s}|$, and has the following properties: $\lim_{y\to0^+}p'\left(y\right)=\infty$, $\lim_{y\to0^+}p''\left(y\right)=-\infty$ and $p'''\left(y\right)>0$ for all $y>0$,*[32] *then:*

1. *$P_{combi}\left(s\right)$ is symmetric around $\bar{s}$.*

[32] This includes, but is not confined to, all the concave power functions we analyze in Section 3.

2. $P_{combi}(s)$ is strictly increasing if $s > \bar{s}$ and strictly decreasing if $s < \bar{s}$.

3. $\exists \Delta \in]0, \frac{x_h - x_l}{2}[$, such that $P_{combi}(s)$ is strictly concave in the range $(\bar{s} - \Delta, \bar{s} + \Delta)$ and strictly convex outside it.

Proof. See the appendix. ■

The lemma shows that the authority is more dominant near the norm (where the aggregated individual pressure is rather flat, while the authority is intolerant to small deviations), whereas the aggregated individual pressure is more dominant near the edges of the distribution (where the authority does not distinguish much between close stances). In total, this creates a society that is orthodox at least close to the institutionalized norm.

4.3 Interpretation

The previous two lemmas show that with uniform sources of pressure, orthodoxy is undermined, but adding an authority is sufficient to restore it. Then the question that remains is whether an authority is necessary to create orthodoxy or whether it is sufficient to have a concentration of pressure sources. At the end of the appendix we look at two examples that illustrate societies whose sources of pressure are concentrated around the social norm, with slowly vanishing tails of extreme sources of pressure at each side of it (Gaussian and Exponential distributions). An example of this could be a situation where the pressure sources represent the private opinions of people and these opinions are very, but not fully, homogenous. Here, although there is a clear peak around the norm and each individual inflicts pressure on others in a concave manner, stances close to the norm are pressured in a convex manner, while stances far from the norm are pressured in a concave way.[33] Moreover, in order to have that switch from convex to concave pressure we need a broad enough range of types.

Although it is hard to make any general statements about this, it seems that we need a *single* authority to get an orthodox-type society, at least when considering types close to the norm. Otherwise, the accumulation of individual pressure gives a liberal aggregate. This suggests that orthodoxy will not emerge in a pluralistic society where heterogenous individuals pressure each other, unless there is also an authority present. Moreover, it predicts that societies that are orthodox in the aggregate perception are also authoritarian. Note however, that we use

[33] This is in contrast to the case of a combination of authoritarian and uniformly distributed individual pressure, where slight deviations from the norm are punished in a concave manner, while extreme stances feel convex pressure.

the term ‘authority’ here quite broadly: it is either someone with powers vested by the others, or a person with more clout than others in general, or a group of individuals with identical opinions.[34]

5 Private opinions as the sources of pressure

The overall purpose of this section is to see whether the result we got in the basic model, that perfectionism drives inversion of revealed preferences, is something that generalizes to other settings. More precisely, we now answer the question: if each person pressures others for deviating from her own *private* opinion, under what circumstances will inversion of revealed preferences arise? Thus, we now assume that the entire pressure comes from the individuals in society, that the sources of pressure are the individuals' types (i.e. $x = t$), and that the types are distributed uniformly. We already analyzed the aggregation of uniform pressure sources in section 4, but now we further analyze what stances people will actually take given this social pressure.

We can think of the case analyzed here as resembling casual interactions in society, say at a bar or on the bus. That is, a person i meets another random person j, not knowing the real opinion of j. They start talking, and so i needs to make some statement. Person i would like to diminish the gap between what she says and what j thinks, in order for j not to think badly of her. Now, i knows only the distribution of opinions in society, but not what j actually thinks. Thus i would like to minimize the expected negative perceptions her peers may get, while also considering the displeasure of deviating from her own private opinion.[35]

If types are uniformly distributed from t_l to t_h, then an individual

[34] Recall that the gaussian and exponential distribution analysis in the appendix shows that it is not sufficient for these sources of pressure to be similar, they have to be identical.

[35] Alternatively, we can think of an individual who is at the same time a norm setter and a norm taker. This has some common ground with the terminology of primary and secondary control used in psychology (Rothbaum et al, 1982). Primary control refers to people trying to change the world, while secondary refers to people changing their opinions to be more in line with the world. Note, however, that those terms allow for alternative modeling interpretations. For example, in line with the next section, a person may put pressure on others to follow her *statement*, as opposed to following her private opinion. When this is the case, it may also include a strategic element. Although we abstract from strategic considerations in this paper, our intuition is that such considerations would lead people to use their private opinions as the sources of pressure whenever they get the chance to wield primary control.

with stance s can expect to perceive the following pressure,

$$\begin{aligned} P_{type}(s) &\equiv E\left[p\left(|s-t|\right)\right] = \frac{1}{t_h - t_l}\int_{t_l}^{t_h} p\left(|s-t|\right)dt \\ &= \frac{1}{t_h - t_l}\left[P\left(t_h - s\right) + P\left(s - t_l\right) - 2P(0)\right],\ s \in \left[t_l, t_h\right], \end{aligned}$$

where $P' \equiv p$. The optimization problem of a single individual of type t is then

$$\min_s L = P_{type}(s) + D\left(|t-s|\right).$$

For tractability and conciseness, we once again revert to using power functions for the dissonance and social pressure functions.[36]

$$\begin{aligned} D\left(|t-s|\right) &= |t-s|^{\alpha}\ \ ,\ \alpha > 0 \\ p\left(|s-t|\right) &= K\left|s-t\right|^{\beta}\ \ ,\ \beta > 0. \end{aligned} \tag{10}$$

Rewriting the minimization problem we get:

$$\min_s L = \frac{1}{t_h - t_l}\left[\frac{K}{\beta+1}\left(s-t_l\right)^{\beta+1} - \frac{K}{\beta+1}\left(t_h - s\right)^{\beta+1}\right] + |t-s|^{\alpha}.$$

The following proposition characterizes the outcome of the model with a focus on inversion.

Proposition 5 *Let $\bar{s} \equiv \frac{t_h+t_l}{2}$, and let p and D be given by equation 10. If private opinions are uniformly distributed and are the sources of pressure, then:*

1. *The aggregate pressure, $P_{type}(s)$, has a unique minimum point, a "virtual" norm, at $s = \bar{s}$.*

2. *$P_{type}(s)$ is convex in the distance to the virtual norm.*

3. *If and only if $\alpha < 1$ and the range of types is broad enough, then types sufficiently close to $\bar{s}$ state their private opinions in public and types sufficiently far choose $s^*(t) \in \left]\bar{s}, t\right[$.*

4. *If and only if $\alpha < 1$ and the range of types is broad enough, then conformity is non-monotonic in the distance from $\bar{s}$ and there is inversion of revealed preferences.*

[36] The upcoming results can easily be generalized beyond power functions.

Proof. *See the appendix.* ■

The proposition states that in pluralistic societies, a single "virtual" norm will be established, and society as a whole will be liberal. Furthermore, along with heterogenous private opinions, perfectionism is now both a sufficient and a necessary condition for inversion to arise. Moderate individuals will tend to speak their mind truthfully, while those who are extreme enough will tilt their stances in the direction of the "virtual" norm. It creates inversion at two levels, both between the moderates and the extremists and within the group of extremists (much like in the basic model with perfectionist individuals in a liberal society). The virtual norm becomes like an unspoken consensus in society. If you are close enough to the consensus, you declare your private opinions, but if you are far from it you make concessions to seem to be moderate.[37]

We see that in this model, where the individual wishes not to upset another person of an unknown type, inversion arises as long as people are perfectionists when declaring their opinions. Note that we do not need to impose any restrictions on the curvature of the pressure stemming from individuals.[38] Roughly speaking, the only thing that can prevent inversion from arising when individuals are perfectionists is a society that is sufficiently orthodox. This will hold in any model that considers our basic individual trade-off. In this particular model variation, the aggregation of individual pressures tends to undermine orthodoxy, hence perfectionism becomes sufficient for inversion to arise.

6 Declared opinions as the sources of pressure

We now add a layer of complexity to the model of aggregated individual pressure. Assume that when the individual declares a stance that minimizes the dissonance and pressure she feels, this declaration also puts pressure on anyone else stating a different opinion.[39] Further assume

[37] The connection found here between inversion of revealed preferences and perfectionism can be traced back to equation 5, where we saw that when perfectionist individuals chose inner solutions we got inversion. Therefore this result holds quite generally also for other functional forms. An interesting thing to note is that under this extension of the model we cannot get a situation of no compromise (such as in section 3.1), since the aggregation of social pressure inevitably leads society to be liberal. This also means that extremists will not take the "virtual" norm as their stance, they will merely move towards it.

[38] This result is robust to any situation where the functional form of p and the distribution of types in society are such that the resultant aggregate social pressure P_{type} satisfies the conditions given in Lemma 11 in the appendix. If the distribution of types is uniform, all we need in order to get inversion is that the pressure p exerted by type t_l on type t_h (and vice verse) will be large enough.

[39] This way of modeling is similar to Manski & Mayshar (2003).

that there are sufficiently many individuals for each one not to behave strategically. Hence, people do *not* take into account how their stances affect others' stances, and how that feeds back into affecting them. That is, individuals are essentially taking the distribution of stances as given. We are interested in seeing whether inversion of preferences arises in this setting as well.

We now denote the pressure function by P_{st} and the set of stances in society by S.

$$P_{st}\left(s,S\right)\equiv E\left[p\left(\left|s-\sigma\right|\right)\right]=\frac{1}{t_h-t_l}\int\limits_{t_l}^{t_h}p\left(\left|s-\sigma\left(t\right)\right|\right)dt \tag{11}$$

Here $\sigma\left(t\right)\in S$. The individual wishes to minimize the sum of the expected (or aggregated) pressure and the dissonance,

$$\min_s P_{st}\left(s\right)+D\left(s,t\right). \tag{12}$$

Solving equation 12 for all types produces a function $s^*\left(t\right)$. The general equilibrium is a fixed point such that substituting $s^*\left(t\right)$ for $\sigma\left(t\right)$ in 11 and then solving 12 reproduces the same function $s^*\left(t\right)$.

This is not an easy problem to solve: as we have seen earlier, even with an exogenous norm it is often hard to obtain a closed form solution to $s^*\left(t\right)$, the function that maps types to stances. On top of this, there is clearly a potential for multiple equilibria.

However, luckily, we are interested in the inversion of revealed preferences, and as it turns out, P_{st} is easy to solve and analyze under the scenario of an orthodox society with very perfectionist individuals, i.e., $\alpha<\beta<1$. So we will now take the outcome $s^*\left(t\right)$ of part 3 of Proposition 1 (where the distribution of stances has a uniform part with a peak at $\bar{s}$ on top of it), augment it slightly, and then use it as a starting guess for $\sigma\left(t\right)$.We will then verify that this is a fixed point.

Suppose we have a distribution of types $t\sim U\left(-\frac{1}{2},\frac{1}{2}\right)$ and a distribution of stances such that all types within some range $\left[-s_h,s_h\right]\subset\left[-\frac{1}{2},\frac{1}{2}\right]$ declare their private opinions as stances, while the rest, beyond this range, take a stance of zero (i.e., the center). This creates a distribution of stances that contains a single peak at zero with a uniform part around it. The aggregate pressure function for $s\geq 0$ is then (using power functions):

$$P_{st}\left(s,S\right)=K\int\limits_{-s_h}^{s_h}\left(\left|s-t\right|\right)^{\beta}dt+K\left(1-2s_h\right)\left(s-0\right)^{\beta}$$
$$=\begin{cases}K\left[\left(1-2s_h\right)s^{\beta}+\frac{\left(s_h-s\right)^{\beta+1}+\left(s_h+s\right)^{\beta+1}}{\beta+1}\right] & \text{for } s\leq s_h\\ K\left[\left(1-2s_h\right)s^{\beta}+\frac{\left(s_h+s\right)^{\beta+1}-\left(s-s_h\right)^{\beta+1}}{\beta+1}\right] & \text{for } s\geq s_h\end{cases}.$$

Note that the structure of this aggregated pressure function is similar to that of P_{combi} in equation 9: a combination of a uniform distribution of sources of pressure (stemming from individuals who declare their private opinions) and one norm (stemming form the mass of individuals who declare the same stance).

The next steps are to verify that all types have corner solutions, that types close to zero declare their private opinions as their stances, and that types far from the norm take zero to be their stance. Finally, we need to determine the cutoff between these two groups in order to find the endogenous s_h. To do all of this we need to resort to numerical simulations.[40]

One exemplary simulation result can be seen in Figure 4. Here we have individuals who are rather perfectionist when stating their own opinions ($\alpha = 0.4$), while the pressure they put on others who declare stances away from their own is a bit less concave ($\beta = 0.6$). We find that there is an equilibrium with inversion, where types between roughly -0.1 and 0.1 declare their private opinions, while those beyond this range take zero as their stance. Thus we have a situation where an endogenous norm at $s = 0$ is upheld by those who disagree with it the most. These types, who uphold the norm, not only declare it as their stance, but indirectly also put pressure on each other to keep declaring it - a fictitious social norm.[41]

This model can produce multiple equilibria for a given set of parameters. But these equilibria do not differ in kind. They all imply inversion of revealed preferences with corner solutions (see some examples in Figure 5).[42] What distinguishes between the equilibria is the location of the fictitious norm. Each simulation starts with an initial condition that

[40] While doing some intermediate steps is possible analytically, these render too little economic meaning to motivate detailing them here.

[41] One may find some similarities between the concept of a fictitious norm and the story of The Emperor's New Clothes.

[42] As far as our simulation experiments go, given a set of parameters that leads to inversion for some initial conditions, any other initial condition will lead to inversion too, at least if the distribution of types is uniform.

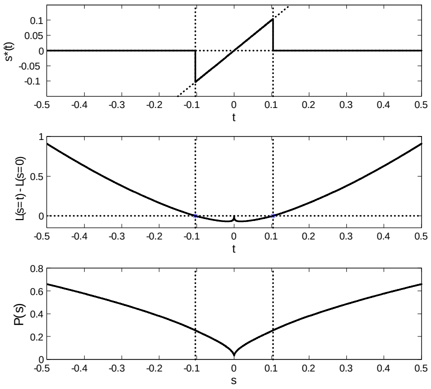

Figure 4: Simulation results with stances as the sources of pressure, and $\alpha = 0.4$, $\beta = 0.6$, $K = 1.83$, $t \sim U(-0.5, 0.5)$. The dashed horizontal lines depict the zero line, the dashed vertical line depict the cutoffs $-s_h$ and s_h, and, in the upper figure, the diagnoal line depicts the 45 degree line ($s = t$).

includes a distribution of types and one norm as a source of pressure. Then, at each period, every individual updates her stance according to the aggregated pressure that results from the choices of all other individuals. We find that if we start with an exogenous norm at some location, and then let people constantly update their stances under the influence of P_{st}, the fictitious norm will stay exactly where the exogenous norm was initially located. But now, instead of being exogenous, it will be maintained by those whose private opinions are far from it.

Moreover, even if the norm is located initially in one corner of the distribution of private opinions, it will survive and be chosen by those with private opinions at the other corner. Unlike the case where the endogenous norm was the average of the stated opinions (Section 3.4), now the distribution of stances in equilibrium need not be symmetric around the fictitious norm. In fact, even if the distribution of types is very far from uniform, a skewed fictitious norm can be maintained.[43] It means that societies in which individuals are perfectionist and social pressure is shaped by all stances will be highly history dependent. The person or the group that manage to establish a first dominant opinion can count on it to remain a status quo also after they are gone.[44]

[43] Some irregularities may arise here if, for instance, the distribution is not smooth so that there are clusters of types. Inversion still exists, but $\frac{ds^*(t)}{dt}$may change sign many times.

[44] This resembles the experimental result where monkeys were punished initially for

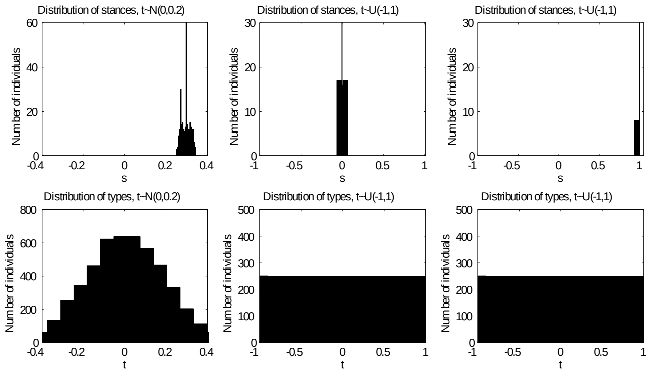

Figure 5: Simulations of the model with P_{st} for the parameters $\alpha = 0.4$, $\beta = 0.6$, $K = 1.8$. Each simulation starts with an initial condition that includes a distribution of types and one norm as a source of pressure. The upper graphs depict the resultant distributions of stances in equilibrium for the corresponding distributions of types in the lower graphs. The initial norm is at 0.3 (left), 0 (middle) and 1 (right) respectively. The distribution of stances in the upper left graph is asymmetric, resulting from an initial asymmetric (random) distribution of types. Note that the axes have been truncated for visibility of the results.

This "stickiness" of skewed norms can be partially explained by the skewness itself. If one thinks of a norm as a tool to enforce people with opinions far from it to conform, then the more skewed is the norm, the more people are enforced to declare it. Consequentially, such a norm will be stronger in its influence on all society, thus more stable than a norm that is located at the center of the distribution of types (especially if this distribution is bell-shaped). From the point of view of those with opinions far from the norm, this creates a vicious circle where they themselves uphold an undesirable and skewed social norm.

We have also considered many other parameter settings, and the result of inversion reappears as long as α is sufficiently far below 1, and β is not smaller than α and not larger than 1.[45] So the result of inversion of revealed preferences seems to be driven by perfectionism here too.

climbing a ladder (Stephenson, 1967). They were then replaced one after the other, until no monkey that has ever seen the original punishment take place was around. Still, all the new monkeys made sure that no one violated the rules by climbing the ladder.

[45] If $\alpha = \beta < 1$, then there is an equilibrium that can possibly be solved analytically. However, this equilibrium is not stable. Small perturbations will lead all types to either declare the norm (midpoint) as their opinions or declare their private opinions.

7 Concluding remarks

The purpose of this paper has been to analyze the circumstances under which social pressure leads to inversion of revealed preferences – a situation where individuals with extreme private opinions declare stances that are more normative than the stances declared by moderate people. Throughout our basic model and its extensions, a common conclusion emerges: *perfectionism* on an individual level is what drives this inversion, where perfectionism means a concave disutility of deviating from one's private opinion. This conclusion holds true in the most basic case, when there is one common norm in society; it holds true when all individuals pressure others for saying things they do not like; and finally, it holds true when individuals indirectly put pressure on others while declaring their own stances. In that final case, we get a situation where people with private opinions far from the consensus in society are the ones who sanction others for deviating from this consensus.

The basic intuition that goes through all the variants of the model is that perfectionism makes people either say exactly what they think, or, in case they deviate from it, say things that lower the social pressure substantially. Since the social pressure is heavier on extreme opinions, only individuals privately holding these opinions misrepresent their opinions publicly. Once these people deviate from their private opinions, they may as well go all the way (or almost all the way) to get rid of the social pressure. At the same time, moderate people publicly declare their private opinions. As a result, those with extreme opinions "pass" the moderates in their statements, appearing in public to be more compliant with the norm - be it institutionalized, virtual, or fictitious.

Is this a realistic result? Is it empirically prevalent? The stability of the results to model extensions suggests that it could be, but these questions are largely left unanswered. We have provided an example that we believe our model can explain: the patterns of conversion of Jews and Muslims in Medieval Spain under the rule of the Christian Church. This example is attractive to use, since, unlike most real world situations, we have historical evidence of the privately held beliefs and opinions of Jewish converts under the persecution of the Christian rulers. Likewise, our model may explain the discrepancies between stated and monitored sexual preferences.

This main result, as well as other conclusions we have made along the way, may have important implications and predictions. One is that pluralism will tend to undermine orthodoxy. Another is that skewed norms will be correlated with orthodoxy and heavy punishments on deviators. Yet another is that liberalism will tend to lead to norms that are representative of average opinions. A final implication that the model highlights is the problem of interpreting stated opinions as a signal of real inner preferences. Under the possibility of preference inversion, such inference cannot be made without deeper scrutiny of whether people are lax or perfectionist about deviations from their bliss-points.

References

[1] Adams, H. E., Wright, L. W. Jr. and Lohr, B. A., (1996), "Is Homophobia Associated With Homosexual Arousal?," *journal of Abnormal Psychology*, Vol. 105, No. 3, pp. 440-445.

[2] Asch, S. E., (1955), "Opinions and Social Pressure," *Scientific American*, Vol. 193, No. 5, pp. 31-35.

[3] Baer, Y., (1965), A history of the Jews in Christian Spain; translated from the Hebrew by Louis Schoffman. Philadelphia: Jewish Publication Society of America, third edition.

[4] Baumeister, R. F., Dale, K. and Sommer, K. L., (1998), "Freudian Defense Mechanisms and Empirical Findings in Modern Social Psychology: Reaction Formation, Projection, Displacement, Undoing, Isolation, Sublimation, and Denial," *Journal of Personality*, Vol. 66, No. 6, pp. 1081-1124.

[5] Ben-Shalom, R., (2001), "Kiddush Hashem and Jewish Martyrology in Aragon and Castile in the Year 1391: Between Spain and Ashkenaz," *Tarbitz*, Vol. 70, No. 2, pp. 227-282. [in Hebrew]

[6] Ben-Sasson, M., (1990), "To the Jewish Identity of the Anusim: an Advisement in the Hishtamdut at the Period of the Almohad Caliphate," *Peamim*, Vol. 42, pp. 16-37. [in Hebrew]

[7] Bernheim, D.B., (1994), "A Theory of Conformity," *Journal of Political Economy,* Vol. 102, No. 5, pp. 841-877.

[8] Brock, W.A., Durlauf, S.N., (2001), "Discrete Choice with Social Interactions," *Review of Economic Studies* Vol. 68, pp. 235–260.

[9] Clark, A. E., Oswald, A. J., (1998), "Comparison-concave utility and following behaviour in social and economic settings," *Journal of Public Economics*, 70, 133–155.

[10] D'Augelli, A. R., (2006), Developmental and contextual factors and mental health among lesbian, gay, and bisexual youths. In A. E. Omoto & H. M. Kurtzman (Eds.), Sexual orientation and mental health: Examining identity and development in lesbian, gay, and bisexual people (pp. 37–53). Washington, DC: APA Books. doi:10.1037/11261-002.

[11] Goffman, E., (1959), Presentation of Self in Everyday Life. New York: The Overlook Press.

[12] Grossman, A., (1998), "Kiddush Hashem in the 11th and 12th Centuries: Between Ashkenaz and the Muslim Countries," *Peamim*,

Vol. 75, pp. 30-34. [in Hebrew]

[13] Granovetter, M., (1976), “Threshold Models of Collective Behavior,” *The American Journal of Sociology,* Vol. 83, No. 6, pp. 1420-1443.

[14] Herrmann, B., Thöni, C. and Gächter, S. (2008), “Antisocial Punishment Across Societies,” *Science*, Vol. 319, pp. 1362–1367.

[15] Holbrook, A. L., Green, M. C. and Krosnick, J. A., (2003) “Telephone Versus Face-to-face Interviewing of National Probability Samples with Long Questionnaires,” *Public Opinion Quarterly*, Vol. 67, pp. 79–125.

[16] Jones, S. R. G., (1984), The Economics of Conformism. Oxford: Basil Blackwell.

[17] Kandel E., Lazear, E. P., (1992), “Peer Pressure and Partnerships,” *The Journal of Political Economy*, Vol. 100, No. 4, pp. 801-817.

[18] Kuran, T., (1995), “The Inevitability of Future Revolutionary Surprises,” *The American Journal of Sociology*, Vol. 100, No. 6, pp. 1528-1551.

[19] Lindbeck, A., Nyberg, S. and Weibull, J. W., (2003), “Social norms and Welfare State Dynamics,” *Journal of the European Economic Association*, Vol 1, Iss 2-3, pp. 533–542.

[20] López-Pintado, D., Watts, D.J., (2008), "Social Influence, Binary Decisions and Collective Dynamics", *Rationality and Society*, Vol. 20, no. 4, pp. 399-443.

[21] Manski, C.F.,Mayshar, J. (2003)“Private Incentives and Social Interactions: Fertility Puzzles in Israel,” *Journal of the European Economic Association*, Vol. 1, No.1, pp. 181-211.

[22] Michaeli, M., Spiro, D., (2012), “The Distribution of Revealed Preferences under Social Pressure,” Discussion Paper no. 609, the Center for the Study of Rationality. http://ratio.huji.ac.il/dp_files/dp609.pdf

[23] Michaeli, M., Spiro, D(2013), “Cultural traits and conformity to norms,” mimeo.

[24] Morokoff, P. J., (1985), “Effects of Sex Guilt, Repression, Sexual "Arousability," and Sexual Experience on Female Sexual Arousal During Erotica and Fantasy,” *Journal of Personality and Social Psychology*, Vol. 49, No. 1, pp. 177-187.

[25] The Pew Research Center. (2007). World publics welcome global trade—but not immigration. Washington, DC.

[26] Rothbaum, F., Weisz, J. R., (1982), “Changing the World and Changing the Self: A Two-Process Model of Perceived Control,” *Journal of Personality and Social Psychology*, Vol. 42, No. 1, pp. 5-37.

[27] Ruiz, T. F., (2008), Spain's Centuries of Crisis: 1300-1474.

[28] Savin-Williams, R. C., Ream, G. L., (2003), "Sex Variations in the Disclosure to Parents of Same-Sex Attractions," *Journal of Family Psychology*, Vol. 17, No. 3, pp. 429–438.

[29] Schelling, T., (1971), "Dynamic Models of Segregation," *Journal of Mathematical Sociology*, Vol. 1, Iss. 2, pp.143–186.

[30] Stephenson, G. R. (1967). "Cultural acquisition of a specific learned response among rhesus monkeys". In: Starek, D., Schneider, R., and Kuhn, H. J. (eds.), Progress in Primatology, Stuttgart: Fischer, pp. 279-288.

[31] Weinstein, N., Ryan, W. S., DeHaan, C. R., Przybylski, A. K. , Legate, N. and Ryan,R. M., (2012), "Parental Autonomy Support and Discrepancies Between Implicit and Explicit Sexual Identities: Dynamics of Self-Acceptance and Defense," *Journal of Personality and Social Psychology*, Vol. 102, No. 4, pp. 815–832.

8 Appendix - Proofs and derivations

8.1 Proof of Lemma 2

For $t \geq \bar{s}$ we get the following minimization problem:

$$\min_s \left\{ (t-s)^\alpha + K(s-\bar{s})^\beta \right\},$$

with a first-order condition

$$L' = -\alpha (t-s)^{\alpha-1} + \beta K (s-\bar{s})^{\beta-1} = 0, \tag{13}$$

and a second-order condition for an internal local minimum point

$$L'' = (\alpha-1)\alpha(t-s)^{\alpha-2} + (\beta-1)\beta K(s-\bar{s})^{\beta-2} > 0. \tag{14}$$

Lemma 5 *Let $t \geq \bar{s}$. If $\alpha > 1$ and $\beta \geq 1$ then conformity is strictly decreasing in t.*

Proof. The SOC is positive $\forall t \geq \bar{s}$, hence all types have inner solutions. Using the FOC, the inner solution for type $t \geq \bar{s}$ is implicitly given by $t = (\beta K/\alpha)^{\frac{1}{\alpha-1}} (s^* - \bar{s})^{\frac{\beta-1}{\alpha-1}} + s^*$. From this expression it is clear that $t(s^*)$ is continuous in s^* and that $dt/ds^* > 0$, which by the implicit function theorem implies that $s^*(t)$ is continuous in t and that $ds^*/dt > 0$, $\forall t \geq \bar{s}$. Thus, each type has a unique solution, and conformity is strictly decreasing (no inversion of revealed preferences). ■

Lemma 6 *If $\beta < 1 < \alpha$, and: i) $\forall t \geq \bar{s}$, L has at most one local minimum in $]\bar{s}, t[$. ii) If for some type $\hat{t} \geq \bar{s}$ L has a local minimum in $]\bar{s}, t[$, then L has a local minimum in $]\bar{s}, t[$ for every type $t > \hat{t}$.*

Proof. i) The FOC for $t \geq \bar{s}$ implies that

$$\alpha\left(t-s\right)^{\alpha-1} = \beta K\left(s-\bar{s}\right)^{\beta-1} \Rightarrow \beta K/\alpha = \left(t-s\right)^{\alpha-1}\left(s-\bar{s}\right)^{1-\beta} \equiv f\left(s\right).$$

Note that $f(s)$ is strictly positive in $]\bar{s},t[$, and that $f(s)=0$ at both edges of the range (i.e. at $s=\bar{s}$ and at $s=t$). This means that $f(s)$ has at least one local maximum in $]\bar{s},t[$. We need this maximum to be larger than $\beta K/\alpha$ for the FOC to hold at some point.

We now proceed to check whether this local maximum of $f\left(s\right)$ is unique:

$$f'\left(s\right) = \left(t-s\right)^{\alpha-2}\left(s-\bar{s}\right)^{-\beta}\left[\left(1-\beta\right)\left(t-s\right)-\left(\alpha-1\right)\left(s-\bar{s}\right)\right].$$

Since $\left(t-s\right)^{\alpha-2}\left(s-\bar{s}\right)^{-\beta}$ is strictly positive in $]\bar{s},t[$, and $[\left(1-\beta\right)\left(t-s\right)-\left(\alpha-1\right)\left(s-\bar{s}\right)]$ is linear in s, positive at $s=\bar{s}$ and negative at $s=t$, $f'(s)=0$ at exactly one point at this range (i.e. a unique local maximum of $f(s)$ in $]\bar{s},t[$). From the continuity of $f(s)$, we get that if the value of $f(s)$ at this local maximum is greater than $\beta K/\alpha$, then $L(t,s)$ has exactly two extrema in the range $]\bar{s},t[$. From the positive values of $L'(t,s)$ at the edges of this range, we finally conclude that the first extremum (where $f(s)$ is rising) is a maximum point of $L(t,s)$, and the second extremum (where $f(s)$ is falling) is a minimum point of $L(t,s)$, i.e. $L(t,s)$ has a unique local minimum. Conversely, if the value of $f(s)$ at its local maximum point is smaller than $\beta K/\alpha$, there is no local extremum to $L(t,s)$ in the range $]\bar{s},t[$ and therefore $s\left(t\right)=\bar{s}$.

ii) Holding s fixed and differentiating $f\left(s\right)$ with respect to t yields $\frac{df(s)}{dt} = \left(\alpha-1\right)\left(t-s\right)^{\alpha-2}\left(s-\bar{s}\right)^{1-\beta} > 0$. That is, if for some pair $\left(\hat{t},\hat{s}\right)$ we have $f(\hat{s})|_{\hat{t}} > \beta K/\alpha$, then we get that $f(\hat{s})|_{t} > \beta K/\alpha$ for every $t > \hat{t}$. Hence, if for some type $\hat{t} \geq \bar{s}$ L has a local minimum in $]\bar{s},t[$, then L has a local minimum in $]\bar{s},t[$ for every type $t > \hat{t}$. ■

We now need to consider corner solutions too. Substituting $s=\bar{s}$ and $s=t$ in equation 13, we get that only $s=\bar{s}$ is a potential solution for the minimization problem of type t. If type t has an inner local minimum too, then we need to compare the loss implied by the corner solution $s=\bar{s}$ to that implied by the inner local minimum, denoted by $\hat{s}$ (if there is no local minimum then $\bar{s}$ is of course the global minimum).

Lemma 7 *Let $Diff \equiv L\left(s\left(t\right)=\bar{s}\right)-L\left(s\left(t\right)=\hat{s}\right)$. If $\beta < 1 < \alpha$, then $Diff$ is monotonically increasing in t, $\forall t \geq \bar{s}$.*

Proof. $Diff = \left(t-\bar{s}\right)^{\alpha} - \left[\left(t-\hat{s}\right)^{\alpha} + K\left(\hat{s}-\bar{s}\right)^{\beta}\right]$. Differentiating $Diff$ with respect to t we get $\frac{dDiff}{dt} = \alpha\left(t-\bar{s}\right)^{\alpha-1} - \left[\alpha\left(t-\hat{s}\right)^{\alpha-1}\left(1-\frac{d\hat{s}}{dt}\right)+\beta K\left(\hat{s}-\bar{s}\right)^{\beta-1}\frac{d\hat{s}}{dt}\right]$. Since $\hat{s}$ satisfies the FOC, we finally get that $\frac{dDiff}{dt} = \alpha\left(t-\bar{s}\right)^{\alpha-1} - \alpha\left(t-\hat{s}\right)^{\alpha-1} > 0$. ■

Lemma 8 *If $\beta < 1 < \alpha$ and $s^*\left(\hat{t}\right) = \bar{s}$ for some $\hat{t} \geq \bar{s}$, then $s^*(t) = \bar{s}$ for every $t < \hat{t}$.*

Proof. If $s^*\left(\hat{t}\right) = \bar{s}$ and $t < \hat{t}$, then either L has no local minimum in $]\bar{s}, t[$, so that t has only a corner solution at $\bar{s}$, or by Lemma 7 and by the chosen solution of $\hat{t}$ we get that t chooses the corner solution at $\bar{s}$. ■

That is, if type t fully conforms, then any type with private opinion closer to the norm fully conforms too, hence no inversion of the order of opinions of these two types when moving from private to public opinions.

Lemma 9 *$\forall t_1, t_2 \geq \bar{s}$, if $t_1 < t_2$ and $\beta < 1 < \alpha$, then $s^*(t_1) \leq s^*(t_2)$, i.e., no inversion of revealed preferences.*

Proof. Given Lemma 8, we need to show this only for t_1, t_2 with inner solutions. The uniqueness of the inner solution for every type t and the continuity of L imply that $s^*(t)$ is continuous, and from equation 5 we get that at the range of inner solutions $\frac{ds^*}{dt} = \frac{D''(t-s^*)}{P''(s^*-\bar{s})+D''(t-s^*)} > 1$, thus $s^*(t)$ is strictly monotonic in t at the range $t \geq \bar{s}$. ■

Lemma 10 *Let $t \geq \bar{s}$. If $\alpha = 1$ then conformity is weakly decreasing in t.*

Proof. Let $\mathring{t}$ solve the equation $\beta K \left(\mathring{t} - \bar{s}\right)^{\beta-1} = 1$. If $\beta > 1$ then $\forall t < \mathring{t}$ we have $L' = -\alpha (t-s)^{\alpha-1} + \beta K (s-\bar{s})^{\beta-1} = -1 + \beta K (s-\bar{s})^{\beta-1} < 0$ for any $s \in [\bar{s}, t]$. So types at the range $\left[\bar{s}, \mathring{t}\right]$ have a corner solution at t. Further, $\forall t \geq \mathring{t}$ we have $L'|_{s=\mathring{t}} = -\alpha \left(t - \mathring{t}\right)^{\alpha-1} + \beta K \left(\mathring{t} - \bar{s}\right)^{\beta-1} = 0$ and $L'' = (\beta-1)\beta K (s-\bar{s})^{\beta-2} > 0$ $\forall s \in (\bar{s}, t]$, thus types with $t \geq \mathring{t}$ have an inner solution at $\mathring{t}$. It thus follows from definition 1 that if $\beta > 1$ conformity weakly decreases. Otherwise, if $\beta \leq 1$, it follows from Proposition 1 (2) that conformity is decreasing in t. This completes the proof. ■

Proof of Lemma 2

We know that inversion of revealed preferences does not arise if $\alpha > 1$ and $\beta \geq 1$ (Lemma 5), if $\beta < 1 < \alpha$ (Lemma 9), and if $\alpha = 1$ (Lemma 10), thus $\alpha < 1$ is a necessary condition for inversion.■

8.2 Other proofs

Lemma 11 *Let $t \geq \bar{s}$. If D is a continuous, increasing and strictly concave function with $\lim_{x\to 0} D'(x) = \infty$ and $D'(x) < \infty$ $\forall x > 0$; and P is a continuous function where $\exists\ \bar{s}$ such that $P(|s-\bar{s}|)$ is increasing and strictly convex in $|s-\bar{s}|$, $P'(0) = 0$, and $\exists x > 0$ s.t. $P'(x) > D'(x)$; then:*

1. *A type sufficiently close to $\bar{s}$ states her private opinion as a stance*
2. *Types sufficiently far from $\bar{s}$ choose $s^*(t) \in]\bar{s}, t[$.*
3. *If the range of types is broad enough, then conformity is non-monotonic in types' distance from the norm, and there is inversion of revealed preferences.*

Proof. Suppose the conditions hold. When D is concave and P is convex, we have that in the corners $L'(s = t) = P'(t - \bar{s}) - \lim_{x \to 0} D'(x) = -\infty$ while $L'(s = \bar{s}) = P'(0) - D'(t - \bar{s}) = -D'(t - \bar{s}) < 0$. This implies that potential corner solutions must be at $s = t$. It also implies that we either have zero or an even number of inner extreme points (e.g., if there are two extreme points, one is a min and the other is a max). Thus, since $L'(s = \bar{s}) < 0$ and $L'(s = t) < 0$, for the existence of a local min point it is sufficient to show that $L' > 0$ for some t and $s \in [\bar{s}, t]$.

We will now show that an inner local min point exists for a type sufficiently far from $\bar{s}$. Define implicitly $\dot{t}$ by $D'(\dot{t} - \bar{s}) = P'(\dot{t} - \bar{s})$. I.e. $\dot{t}$ is the type whose maximal marginal pressure (when choosing $s = \dot{t}$) is exactly equal to the minimal marginal dissonance (when choosing $s = \bar{s}$). We know that $\dot{t} > \bar{s}$ (since $P'(0) = 0$ and $D'(x) > 0$ by construction). This means types close to $\bar{s}$ must choose $s^*(t) = t$. This proves statement 1). By the intermediate value theorem we also know $\dot{t}$ exists in a broad enough range since $P'(0) - D'(0) < 0$ and $P'(t - \bar{s}) - D'(t - \bar{s})$ increases continuously in t, and since by assumption $P'(t - \bar{s}) - D'(t - \bar{s}) > 0$ for a large enough t. Note that $L(\dot{t}, s)$ will not have an inner local min, since this requires $P'|_{s=s^*} = D'|_{s=s^*}$ for some $s^* \in]\bar{s}, \dot{t}[$, while here $\dot{t} \neq \bar{s}$ and $P'|_{s=\dot{t}} = D'|_{s=\bar{s}}$ is the only way to equate D' and P'.

Let us now look at the type $\ddot{t} = \bar{s} + 2(\dot{t} - \bar{s}) + \varepsilon$ where $\varepsilon \geq 0$. This is the type that is just beyond twice as far from the norm as $\dot{t}$. If $\ddot{t}$ chooses $s = \dot{t}$, we have

$$L'\left(\bar{s} + 2\left(\dot{t} - \bar{s}\right) + \varepsilon, \dot{t}\right) = P'(\dot{t} - \bar{s}) - D'(\bar{s} + 2\left(\dot{t} - \bar{s}\right) + \varepsilon - \dot{t}) = \\ P'(\dot{t} - \bar{s}) - D'(\dot{t} - \bar{s} + \varepsilon) = D'(\dot{t} - \bar{s}) - D'(\dot{t} - \bar{s} + \varepsilon).$$

Since D is concave, $\varepsilon > 0$ gives a strictly positive L'. This proves the existence of an inner local min point for types sufficiently but not infinitely far from $\bar{s}$. To prove statement 2), we still need to show that for types sufficiently far from $\bar{s}$ the local min point is also a global one. We have shown that a local min point exists. Let us now examine $Diff(t)$, the difference between the loss at a local min point (denoted by $s = \hat{s}$) and the loss at the corner solution $s = t$.

$$Diff(t) \equiv L(s(t) = t) - L(s(t) = \hat{s}) = P(t - \bar{s}) - [P(\hat{s} - \bar{s}) + D(t - \hat{s})]$$

$$\frac{dDiff}{dt} = P'(t - \bar{s}) - P'(\hat{s} - \bar{s})\frac{d\hat{s}}{dt} - D'(t - \hat{s})\left(1 - \frac{d\hat{s}}{dt}\right).$$

$\hat{s}$ is a local extremum, thus it satisfies equation 4, i.e., $P'(\hat{s} - \bar{s}) = D'(t - \hat{s})$. We then get that

$$\frac{dDiff}{dt} = P'(t - \bar{s}) - P'(\hat{s} - \bar{s}) > 0.$$

Differentiating once more, we get that

$$\frac{d^2Diff}{dt^2} = P''(t - \bar{s}) - P''(\hat{s} - \bar{s})\frac{d\hat{s}}{dt} > 0,$$

since $d\hat{s}/dt$ is negative in inner solutions when inserting a convex P and a concave D in equation 5.

Hence, there exists a t sufficiently far from $\bar{s}$ such that $Diff(t)$ is positive, so that the global minimum point for t is an inner solution. $\frac{dDiff}{dt} > 0$ ensures that all types beyond this point have global inner min points too. This completes the proof of statement 2).

We now prove statement 3). For the range of types close to $\bar{s}$ who state their type we have $\frac{ds^*}{dt} = 1 > 0$, thus conformity is decreasing. If the range of types is broad enough to include also types with global min points, then by the concavity of D and by Lemma 1 there is inversion of revealed preferences, hence overall conformity is non-monotonic in types' distance from $\bar{s}$. ■

Proof of proposition 2

Follows directly from Lemma 11 with $D(x) = A|x|^{\alpha}$, $\alpha \in (0,1)$ and $P(x, \bar{s}) = B|x - \bar{s}|^{\beta}$, $\beta > 1$.■

Proof of proposition 4

Let $d \equiv \min\{t_h - \bar{s}, \bar{s} - t_l\}$. Since the solution for any type's optimization problem depends only on the distance from $\bar{s}$, we know that the distribution of the stances of all the types in the range $[\bar{s} - d, \bar{s} + d]$ is symmetric around $\bar{s}$. Thus $\bar{s}$ is the average stance for this range of types.

If $\bar{s} = \frac{t_h + t_l}{2}$, then $[\bar{s} - d, \bar{s} + d] = [t_l, t_h]$, and so $\bar{s}$ is the average stance for all types in society. It thus follows that $\bar{s} = \frac{t_h + t_l}{2}$ can be sustained as a social norm in equilibrium for any values of α and β. Otherwise, if $\bar{s} \neq \frac{t_h + t_l}{2}$, then there are types that reside outside the range $[\bar{s} - d, \bar{s} + d]$, all of whom either to the left of $\bar{s}$, such that for each of them $s^*(t) \leq \bar{s}$, or to the right of it (such that for each of them $s^*(t) \geq \bar{s}$). Hence, for $\bar{s}$ to be the average of *all* stances, we must have $s^*(t) = \bar{s}$ for all those types with $|t - \bar{s}| > d$. From Proposition 2, we know such types do not exist if $\alpha < 1 < \beta$, thus follows statement 1).

Otherwise, if $\alpha < \beta \leq 1$, then from the proof of part (3) of Proposition 1, we know that $s^*(t) = \bar{s}$ for types sufficiently far from $\bar{s}$. That same proof states that $s^*(t) = \bar{s}$ iff $|t - \bar{s}| > K^{\frac{1}{\alpha-\beta}}$. Thus, for $\bar{s}$ to be the average of all stances, we must have $|t - \bar{s}| > K^{\frac{1}{\alpha-\beta}}$ for any type with $|t - \bar{s}| > d$. It thus follows that $d = \min\{t_h - \bar{s}, \bar{s} - t_l\} \geq K^{\frac{1}{\alpha-\beta}}$, that is, $\bar{s} \in \left[t_l + K^{\frac{1}{\alpha-\beta}}, t_h - K^{\frac{1}{\alpha-\beta}}\right]$, which completes the proof of statement 2).■

Proof of Lemma 4

1) and 2) follow directly from the fact that both p and P_{aggr} are symmetric around $\bar{s}$ and are increasing in $|s - \bar{s}|$.

3) For tractability, we perform the analysis for $s > \frac{x_h+x_l}{2}$. By symmetry, the analysis for $s < \frac{x_h+x_l}{2}$ has the same properties. By differentiating P_{combi} with respect to s and using equations 7 and 8 , we get

$$P'_{combi} = (1-A)\left[p(s-x_l) - p(x_h-s)\right] + Ap'\left(s - \frac{x_h+x_l}{2}\right) \tag{15}$$

$$P''_{combi} = (1-A)\left[p'(s-x_l) + p'(x_h-s)\right] + Ap''\left(s - \frac{x_h+x_l}{2}\right) \tag{16}$$

$$P'''_{combi} = (1-A)\left[p''(s-x_l) - p''(x_h-s)\right] + Ap'''\left(s - \frac{x_h+x_l}{2}\right) \tag{17}$$

Since $\lim_{y\to 0} p''(y) = -\infty$, we get that

$$\lim_{s\to\frac{x_h+x_l}{2}} P''_{combi} = (1-A)\,p'\left(\frac{x_h-x_l}{2}\right) + A \lim_{s\to\frac{x_h+x_l}{2}} p''\left(s - \frac{x_h+x_l}{2}\right) = -\infty,$$

i.e., $P_{combi}(s)$ is concave around $\bar{s}$. Since $\lim_{y\to 0} p'(y) = \infty$, we get that

$$\lim_{s\to x_l^+} P''_{combi} = (1-A)\,p'(x_h-x_l) + Ap''\left(\frac{x_h-x_l}{2}\right) + (1-A)\lim_{s\to x_l^+} p'(s-x_l) = \infty.$$

Similarly, $\lim_{s\to x_h^-} P''_{combi} = \infty$. That is, $P_{combi}(s)$ is convex as s approaches either of the extreme stances, x_h or x_l.

By assumption, $p'''(y) > 0$ for all $y > 0$, hence if $s \geq \frac{x_h+x_l}{2}$, we have $p''(s-x_l) > p''(x_h-s)$ and $p'''\left(s - \frac{x_h+x_l}{2}\right) > 0$, and so $P'''_{combi} > 0$. This means that $P''_{combi}(s)$ is strictly increasing in the interval $\left(\frac{x_h+x_l}{2}, x_h\right)$, and therefore it changes signs exactly once, at some point $\bar{s}+\Delta \in \left(\frac{x_h+x_l}{2}, x_h\right)$. A mirror image applies to the range $\left(x_l, \frac{x_h+x_l}{2}\right)$.■

Proof of proposition 5

Statements 1) and 2) follow from Lemma 3.

We now prove sufficiency in statements 3) and 4). By Lemma 3 we also know that $P'_{type}(0) = 0$. Furthermore, using equation 7 with $x_l = t_l$

and $x_h = t_h$, we get that $\lim\limits_{s \to t_h -} P'_{type}(s) = \lim\limits_{s \to t_h -} \left[p(s - t_l) - p(t_h - s) \right] = p(t_h - t_l) = K(t_h - t_l)^{\beta}$. So for a broad enough range of types condition (ii) in Lemma 11 holds. All other conditions in Lemma 11 hold, and sufficiecny in statements 3) and 4) thus follow from that lemma.

For necessity in statements 3) and 4) note that P is convex with $P\prime(0) = 0$. If $\alpha \geq 1$ then L" is positive and a unique inner solution exists for each type. In particular, types close to $\bar{s}$ have an inner solution. This proves necessity in 3). Equation 5 describes the properties of the inner solution. With a weakly convex D it is weakly positive and hence by definitions 1 and 2 conformity is weakly decreasing for all t and there is no inversion. This proves necessity in 4).■

8.3 Exponential and Gaussian distribution of pressure sources

An interesting complement to the previous analysis is looking at other distributions of pressure sources but continuing with individual pressure being concave. It turns out that it is hard to say anything in general about this, so we will analyze two specific types of distributions, the Exponential and the Gaussian, with individual pressure being a power function. These two distributions both have a clear peak and sharply declining tails. We are interested in seeing whether they can produce orthodox aggregate pressure.

Posit a distribution of pressure sources $f(x)$ which symmetrically has an exponential shape peaking towards the social norm from each side. W.l.o.g. let the social norm be at $s = 0$, i.e. $E(x) = 0$. The minimum and maximum pressure source in society are at $\pm\infty$.

$$\begin{aligned} P_{\exp}(s) &= \int_{t_l}^{t_h} p(|x - s|) f(x)\, dx \\ &= \frac{\lambda}{2} \int_{-\infty}^{\infty} |x - s|^{\alpha} e^{-\lambda |x|} dx = \frac{\lambda}{2} \int_{-\infty}^{\infty} |x|^{\alpha} e^{-\lambda |x + s|} dx \end{aligned}$$

where $0 < \alpha < 1$. Differentiating we get

$$\begin{aligned} P'_{\exp}(s) &= -\frac{\lambda^2}{2} \int_{-\infty}^{\infty} |x|^{\alpha} e^{-\lambda |x + s|} sgn(x + s)\, dx \\ P''_{\exp}(s) &= \frac{\lambda^3}{2} \int_{-\infty}^{\infty} |x|^{\alpha} e^{-\lambda |x + s|} dx. \end{aligned}$$

To see the behavior of this function around the social norm, we now look at

$$\begin{aligned} P''_{\text{exp}}(0) &= \frac{\lambda^3}{2}\int\limits_{-\infty}^{\infty} |x|^{\alpha} e^{-\lambda|x|}dx = \lambda^3 \int\limits_{0}^{\infty} x^{\alpha} e^{-\lambda x} dx \\ &= \lambda^{2-\alpha}\Gamma(\alpha+1) > 0 \end{aligned}$$

where $\Gamma(\alpha+1)$ is an incomplete Gamma function. This implies that total pressure is convex near the norm. Let us now investigate the asymptotic behavior of $P_{\text{exp}}(s)$ for $s \to \infty$. To this end, let us use the "dimensionless" integration variable $z = x/s$, so that $P_{\text{exp}}(s) = \frac{\lambda}{2}s^{\alpha+1}\int\limits_{-\infty}^{\infty} |z-1|^{\alpha} e^{-K|t|}dt$, where $K = \lambda s$. The integral can be written as

$$\begin{aligned} P_{\text{exp}}(s) &= \frac{\lambda}{2}s^{\alpha+1}\left[\int\limits_{-\infty}^{0}(1-z)^{\alpha} e^{Kt}dt + \int\limits_{0}^{1}(1-z)^{\alpha} e^{-Kt}dt + \int\limits_{0}^{1}(z-1)^{\alpha} e^{-Kt}dt\right] \\ &= \frac{\lambda}{2}s^{\alpha+1}\left[K^{-\alpha-1}e^{K}\Gamma(\alpha+1,K) + K^{-\alpha-1}e^{-K}\gamma(\alpha+1,-K) + K^{-\alpha-1}e^{-K}\Gamma(\alpha+1)\right] \end{aligned}$$

where $\Gamma(a,b)$ are incomplete Gamma functions. For large K, $\Gamma(\alpha+1,K) \approx K^{\alpha}e^{-K}$, so the second and third terms of the sum, which contain the rapidly decreasing exponent e^{-K}, can be neglected and we finally obtain for $s \to \infty$:

$$P(s) \approx \frac{\lambda}{2}s^{\alpha+1}K^{-1} = \frac{s^{\alpha}}{2},$$

whence $P'' \approx \frac{1}{2}\alpha(\alpha-1)s^{\alpha-2} \to -0$.

This means that total pressure is convex near the norm and concave for extreme stances (at least when the extreme stances are sufficiently extreme for the limit case to be relevant). Unfortunately, it is hard to say anything about when and how many times it switches from convex to concave.

Let us now in a similar fashion analyze the case where the pressure sources follow a Gaussian distribution, so that $f(x) = \sqrt{\frac{\lambda}{\pi}}e^{-\lambda x^2}$, $t_l = -\infty$, $t_l = -\infty$. The pressure is a concave power function

$$\begin{aligned} P_{gauss}(s) &= \int\limits_{t_l}^{t_h} p(|x-s|)\, f(x)\, dx \\ &= \sqrt{\frac{\lambda}{\pi}}\int\limits_{-\infty}^{\infty}|x-s|^{\alpha} e^{-\lambda x^2}dx = \sqrt{\frac{\lambda}{\pi}}\int\limits_{-\infty}^{\infty}|x|^{\alpha} e^{-\lambda(x+s)^2}dx \end{aligned}$$

$$P'_{gauss}\left(s\right)=-2\lambda\sqrt{\frac{\lambda}{\pi}}\int\limits_{-\infty}^{\infty}\left|x\right|^{\alpha}\left(x+s\right)e^{-\lambda\left(x+s\right)^{2}}dx$$

$$P''_{gauss}\left(s\right)=-2\lambda\sqrt{\frac{\lambda}{\pi}}\int\limits_{-\infty}^{\infty}\left|x\right|^{\alpha}\left[1-2\lambda\left(x+s\right)^{2}\right]e^{-\lambda\left(x+s\right)^{2}}dx$$

$$P''_{gauss}\left(0\right)=-4\lambda\sqrt{\frac{\lambda}{\pi}}\int\limits_{0}^{\infty}x^{\alpha}\left[1-2\lambda x^{2}\right]e^{-\lambda x^{2}}dx$$

Substituting the integration variable with $u=x^2$, we have

$$\begin{aligned}P''_{gauss}\left(0\right)&=-2\lambda\sqrt{\frac{\lambda}{\pi}}\int\limits_{0}^{\infty}u^{\left(\alpha-1\right)/2}\left[1-2\lambda u\right]e^{-\lambda u}du\\&=-\frac{2}{\sqrt{\pi}}\lambda^{3/2}\lambda^{-\left(\alpha+1\right)/2}\left[\Gamma\left(\frac{\alpha+1}{2}\right)-2\Gamma\left(\frac{\alpha+3}{2}\right)\right]\\&=-\frac{2}{\sqrt{\pi}}\lambda^{\left(2-\alpha\right)/2}\Gamma\left(\frac{\alpha+1}{2}\right)\left[1-2\frac{\alpha+1}{2}\right]=\frac{2\alpha}{\sqrt{\pi}}\lambda^{\left(2-\alpha\right)/2}\Gamma\left(\frac{\alpha+1}{2}\right)>0.\end{aligned}$$

Here, we have used the property of the Gamma function, $\Gamma\left(z+1\right)=z\Gamma\left(z\right)$. Let us now investigate the asymptotic behavior of $P\left(s\right)$ for $s\rightarrow\infty$. To this end, let us use the "dimensionless" integration variable $z=x/s$, so that $P\left(s\right)=\sqrt{\frac{\lambda}{\pi}}s^{\alpha+1}\int\limits_{-\infty}^{\infty}\left|z-1\right|^{\alpha}e^{-\lambda z^{2}}dz$, where $K=\lambda s^{2}$. The integral above has a saddle point at $t=0$, so for $K\rightarrow\infty$, $P\left(s\right)\approx\sqrt{\frac{\lambda}{\pi}}s^{\alpha+1}\int\limits_{-\infty}^{\infty}\left[1-\alpha z+O\left(z^{2}\right)\right]e^{-Kz^{2}}dz=\sqrt{\frac{\lambda}{\pi}}s^{\alpha+1}\sqrt{\frac{\pi}{K}}\left[1+O\left(\frac{1}{K}\right)\right]\approx s^{\alpha}$. From here, $P''\approx\alpha\left(\alpha-1\right)s^{\alpha-2}\rightarrow-0$.

So the total pressure is convex around the norm and concave towards the extremes. In this case, it is once more hard to say anything about where and how many times the shift between convex and concave forms takes place.

Under Gaussian and Exponential distributions, it seems that the switch of pressure from convex to concave towards the extremes is dependent on the pressure sources virtually vanishing. Then, from the point of view of someone taking an extreme stance, the perception is that there is just a mass of punishing individuals located at the norm. What truly is a distribution of pressure sources then looks like one authority for someone standing sufficiently far away. On the other hand, for someone close to the norm the variation in the pressure sources becomes visible since she is standing within the main mass of people.