Bühl, Vitus; Schmidt, Robert C.

Conference Paper

Coordinating to avoid the catastrophe

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2020: Gender Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Coordinating to avoid the catastrophe

Vitus Bühl* Robert C. Schmidt†

March 1, 2020

Abstract

In the presence of a tipping point for dangerous climate damages, the cooperation problem of climate protection can be transformed into a coordination problem that is much easier to deal with (Barrett, 2013). This holds in particular if the amount of greenhouse gas emissions that triggers the catastrophe is precisely known, while the well-known free-rider problem re-appears if the location of the threshold is sufficiently uncertain. In this paper, we focus on the question how the non-signatories (outsiders of a climate agreement) coordinate to avoid the catastrophe, if the tipping point is known. In particular, in light of a multiplicity of equilibria in this coordination problem, the assumption that outsiders will always successfully coordinate to avoid the threshold, even if this is in their collective interest, seems overly optimistic. We analyze how the probability that the outsiders coordinate on an equilibrium in which the threshold is avoided, affects the incentives of countries to join the climate coalition. In some cases, there are multiple equilibria at the participation stage: an equilibrium with full participation, and an equilibrium in which a much smaller coalition forms – just large enough to achieve an outcome in which the catastrophe is avoided with positive probability.

Keywords: tipping point, climate catastrophe, coordination game, international environmental agreement, climate cooperation

*Institute for Microeconomics, University of Hagen, Universitätsstr. 11, 58097 Hagen, Germany; E-mail: vitus.buehl@fernuni-hagen.de

†Institute for Microeconomics, University of Hagen, Universitätsstr. 11, 58097 Hagen, Germany; E-mail: robert.schmidt@fernuni-hagen.de

1 Introduction

Whilst anthropogenic climate change itself gains more and more attention, also the idea of possible tipping points is getting spread. The latter can be reached when thresholds for a maximal increase in temperature, which is often linked to the CO_{2eq}-concentration in the atmosphere, are exceeded so that catastrophic or at least dangerous climate damages occur (Lenton 2011).

In contrast to the standard model, where climate damages and abatement costs are continuous functions, as for example in Barrett (1994), and the international problem of abating emissions is seen as a social dilemma, Barrett (2013) argues that in the presence of tipping points the social dilemma may be reduced to a mere coordination problem.

In the latter paper this is true if the potential damage of exceeding the threshold is high and the position of the threshold is known with sufficient precision. In this situation a climate treaty would only serve as a coordination device, to pin down the amount of abatement each country has to contribute.

Under those circumstances, cooperation turns out to be very stable. A coalition acting as a Stackelberg leader can, by committing to abate only a certain amount thus leaving a gap to the needed threshold, punish countries which are not in the coalition to abate their maximum feasible amount of emissions. This results if the gap, left by the coalition, to reach the threshold is exactly the sum of the outsiders maximal abatements they are willing to contribute to deter the catastrophe. Therefore every outsider would be better off by joining the coalition (or indifferent in the knife-edge-case).

In our paper we drop the assumption made in Barrett (2013) that outsiders are always able to coordinate on the equilibrium to close the gap (if this is a feasible equilibrium) and thereby prevent the catastrophic climate damages when the coalition on its own does not abate enough to stay on the save side of the threshold.

The reason for doing this is that Barrett's assumption (Barrett 2013) seems overly optimistic (from the perspective of the coalition members), whilst an assumption that the outsiders always fail to coordinate on the equilibrium to prevent the catastrophe, if such an equilibrium exists, seems overly pessimistic. The first case would imply, that coalition members only abate the minimal amount for which it still holds, that the outsiders fill the gap to reach the threshold. In this case the outsiders would be better off than letting the catastrophe happen, or in the knife-edge-case, they would be indifferent between abating, therefore preventing the catastrophe or contributing nothing. Latter would then trigger the catastrophe. Considering this indifference it seems implausible to assume that the outsiders are always able to coordinate on the equilibrium to prevent the catastrophe.

On the other hand the assumption that the outsiders never coordinate on the equilibrium which prevents the catastrophe leads to small coalition sizes, as there will be only enough members to prevent the catastrophe on their own while the outsiders are

freeriding and abating nothing.

This paper seeks to bridge the gap between these two, kind of extreme, assumptions, where on the one hand outsiders never coordinate and on the other hand always coordinate to prevent the catastrophe. To achieve this, we assume that the outsiders may fail to coordinate on the preventing equilibrium if such an equilibrium exists, with a certain probability. This probability is assumed to depend on the amount of abatement the outsiders have to achieve to prevent the catastrophe and the number of outsiders.

2 The Model

There are N ex ante symmetric countries, facing a potential climate catastrophe. The tipping point for reaching the catastrophe is located at $\bar{Q}$, so that the catastrophe occurs if the total abatement of all countries, $Q = \sum_i q_i$ satisfies $Q < \bar{Q}$, where $q_i \geq 0$ is the abatement of emissions of country i. If the catastrophe occurs, each country incurs damages of $X > 0$. The tipping point is avoided and no damages are incurred if $Q \geq \bar{Q}$. Country i's abatement costs are given by $C(q_i)$, with $C'(q_i) > 0$, and $C''(q_i) \geq 0$ for all $q_i > 0$ and $C(q_i) = 0$ for $q_i = 0$.

The game comprises three stages: in stage 1, each country decides whether or not to join a climate coalition; in stage 2, the coalition members collectively decide about their abatement targets, so as to maximize the joint welfare of the coalition members; in stage 3, the outsiders choose their abatement levels individually and non-cooperatively. The game is analyzed by backwards induction. We denote by k the number of signatories that join the coalition in stage 1, and by Q^c the total abatement of all coalition members chosen in stage 2. The number of non-signatories is denoted by $n = N - k$, and their total abatement is denoted Q^n.

The following observation will facilitate the characterization of equilibria: there is a maximum abatement effort per country, denoted by $\tilde{q}$, that any country would still be willing to invest in order to prevent the catastrophe, assuming that this country is pivotal. This maximum effort is given by:

$$C(\tilde{q}) = X \Leftrightarrow \tilde{q} = C^{-1}(X).$$

We assume throughout that

$$\tilde{q} < \bar{Q} < N\tilde{q}. \tag{1}$$

If $\bar{Q} > N\tilde{q}$, there can never exist an equilibrium in which the catastrophe is avoided with strictly positive probability, as the total abatement required for this would exceed the maximum willingness to abate aggregated over all countries. If, by contrast, $\bar{Q} < \tilde{q}$, there is clearly no equilibrium (in pure strategies) in which the catastrophe may occur, as a

single country would be willing to prevent the catastrophe on its own, if no other country invests any effort to prevent it.

Stage 3:

We focus on equilibria in pure strategies. However, following Karp and Sakamoto (2019), we assume that in case of a multiplicity of equilibria, the equilibrium that is actually played is chosen randomly. If $Q^c \geq \bar{Q}$, then the joint efforts of the coalition members are sufficient to prevent the catastrophe. Then there is a unique equilibrium in stage 3, such that the abatement of all outsiders is zero: $q_i^n = 0$. Any deviation by an outsider that entails a positive abatement effort leads to a lower welfare for this country. If $Q^c < \bar{Q}$, there can be two equilibrium types in stage 3: either the outsiders coordinate to avoid the catastrophe, such that their joint abatement effort Q^n satisfies: $Q^n = \bar{Q} - Q^c$, or they fail to avoid the catastrophe, in which case none of these countries chooses any positive abatement effort: $q_i^n = 0$ for all outsiders. It is easy to see, that there can be no other equilibrium outcome in stage 3: if $Q^n > \bar{Q} - Q^c$, then at least one outsider can gain a strictly higher welfare by (slightly) lowering its effort, without triggering the catastrophe. And if $0 < Q^n < \bar{Q} - Q^c$, then the catastrophe occurs with certainty, so that any non-signatory that chooses a strictly positive effort, can reach a higher welfare by deviating to an effort level of zero.

Clearly, if $\bar{Q} - Q^c > n\tilde{q}$, then the unique equilibrium in stage 3 entails $q_i^n = 0$ for all outsiders. Hence, it remains to characterize the set of equilibria for the intermediate case where $\bar{Q} - n\tilde{q} \leq Q^c < \bar{Q}$. In other words, this is the case where the total effort required for the non-signatories to avoid the catastrophe satisfies: $Q^n \in (0, n\tilde{q}]$. Unless $Q^n = n\tilde{q}$ holds exactly, there is a multiplicity of equilibria of the type where the catastrophe is avoided, as the exact distribution of the burden of abatement among the non-signatories is not fully determined. In the simplest case, this burden is shared equally, so that $q_i^n = q^n$ holds for all non-signatories, where $q^n = Q^n/n$. With strictly convex abatement costs, this is also the most efficient outcome.

Lemma 1. *If $\bar{Q} - Q^c \in (0, \tilde{q})$, the equilibrium where $q_i^n = 0$ for all outsiders fails to exist, so in any pure-strategy Nash equilibrium in stage 3, the catastrophe is avoided with certainty. If $\bar{Q} - Q^c \in [\tilde{q}, n\tilde{q}]$, both types of equilibria co-exist.*

Proof of Lemma 1. Considering the first case and imagining the situation, that non of the outsiders would abate any emissions. In this situation at least one of the outsiders would be willing to abate the whole amount $\bar{Q} - Q^c$ as it is smaller then $\tilde{q}$, so that this country would still be better off than not abating and therefore incurring the damage X. In the second case however, if no outsider abates any emissions, no outsider on its own can gain any increase in welfare, by deviating, thus abating up to $\tilde{q}$, as the catastrophe would

still happen with probability 1, therefore this is one type of equilibrium for this case. On the other hand, if the outsiders collectively would abate $\bar{Q} - Q^c$, while no outsider has to abate more than $\tilde{q}$, the catastrophe would be avoided with certainty. If any outsider deviates in this situation by reducing its emissions, the catastrophe would happen with probability 1 again, therefore the deviating outsider would be strictly worse off. Hence in the second case the two types of equilibria coexist.

□

Let us denote by $F_n(\bar{Q} - Q^c)$ the probability that countries coordinate on the equilibrium where $q_i^n = 0$ for all outsiders. Figure 1 illustrates the finding of Lemma 1. If $\bar{Q} - Q^c < \tilde{q}$, then in stage 3, there is no equilibrium in which the catastrophe occurs, hence, $F_n(\bar{Q} - Q^c) = 0$. Conversely, if $\bar{Q} - Q^c > n\tilde{q}$, there is no equilibrium in which the catastrophe is avoided, so that $F_n(\bar{Q} - Q^c) = 1$. However, in the intermediate range: $\bar{Q} - Q^c \in [\tilde{q}, n\tilde{q}]$, according to Lemma 1, both types of equilibria co-exist. This corresponds to the light green area in Figure 1. For values of $\bar{Q} - Q^c$ in this range, the probability with which the outsiders fail to prevent the catastrophe, can lie anywhere between 0 and 1.

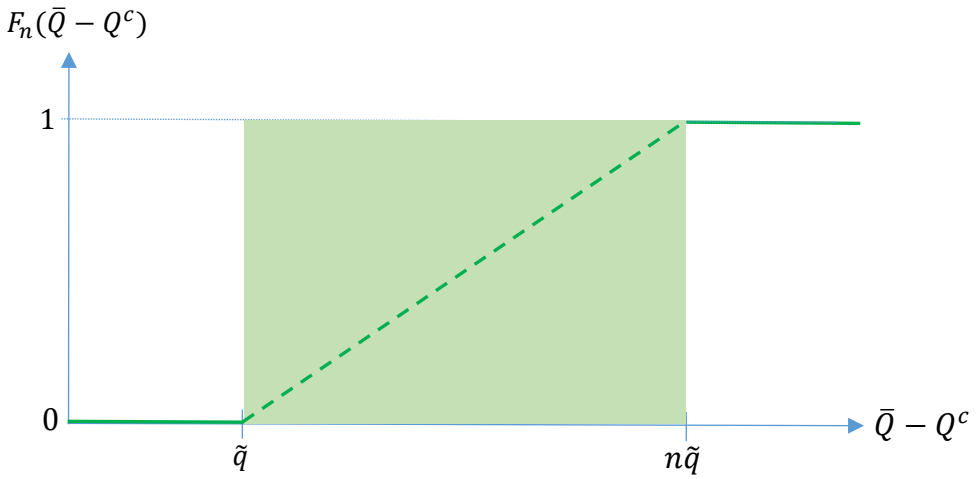

Figure 1: Probability of outsiders to *not* prevent catastrophe, as function of $\bar{Q} - Q^c$

We follow Karp and Sakamoto (2019) by assuming that a randomization device selects among the set of equilibria. This selection may depend on the size of the "gap" (to prevent the catastrophe), $\bar{Q} - Q^c$, as well as on the number of outsiders that need to coordinate, n. The notation $F_n(\bar{Q} - Q^c)$ highlights this dependency, but suppresses the information on how the burden of abatement is split among the non-signatories, in case they coordinate on an equilibrium in which the catastrophe is avoided (with certainty). Again, in the simplest case, one may assume that the burden is split equally among the non-signatories, but there are other equilibria in which the burden is split differently. While this does not affect the (expected) welfare of the coalition members, it does affect

the welfare of the non-signatories, and hence, the stable coalition size (to be determined in stage 1). We will return to this issue later.

The chances of coordinating on an equilibrium in which the catastrophe is avoided, may depend also on the (average) burden per outsider, if the catastrophe is to be avoided. For example, it may be plausible to assume that the outsiders are more likely to coordinate to avoid the catastrophe, if the burden per outsider is lower. Hence, it is convenient to define:

$$f_n(q_n) \equiv F_n(\bar{Q} - Q^c = nq^n),$$

where $q^n \equiv \frac{\bar{Q}-Q^c}{n}$ is the average burden per outsider, if the catastrophe is avoided (under an equal burden sharing, it holds that $q_i^n = q^n$).

A simple case is if the outsiders always coordinate to avoid the catastrophe, if such an equilibrium exists. In Figure 1, the function $F_n(\bar{Q} - Q^c)$ would then be at the bottom of the light green rectangle in the intermediate range. However, the assumption that the outsiders always successfully coordinate to avoid the catastrophe seems overly optimistic. Especially with many outsiders, coordination may easily fail. On the other hand, to assume that outsiders never avoid the catastrophe, if such an equilibrium exists, seems overly pessimistic. A simple way to bridge these extreme views is to assume, that there is a smooth transition from $F_n = 0$ to $F_n = 1$ as the "gap" $\bar{Q} - Q^c$ is raised gradually from $\tilde{q}$ to $n\tilde{q}$. A linear transition is indicated in Figure 1, see the dashed green line in the center of the figure. In terms of the function f_n, this linear relation is depicted in Figure 2, for two different values of n, denoted n_1 and n_2, such that $n_1 > n_2$.

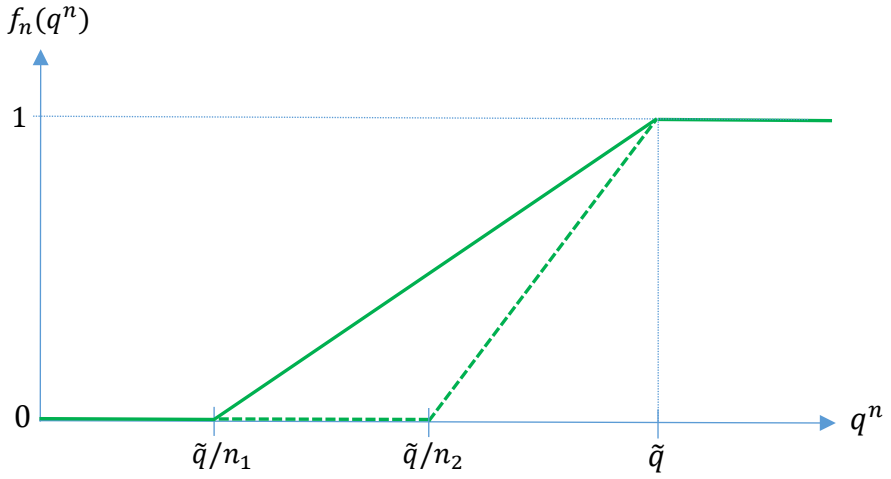

Figure 2: Probability of outsiders to *not* prevent catastrophe, as function of q^n

Clearly, under the above linear assumption, the function f_n increases more steeply with the average burden per outsider, q^n, if n is smaller.

Below, we will analyze coalitional stability for this linear relation, but before we get

there, it might be useful to characterize the two extreme cases which our model wants to bridge the gap in between.

Border cases

As mentioned there are two rather extreme cases, where the outsiders either never coordinate on the equilibrium to avoid the catastrophe or always coordinate on this equilibrium, always provided such an equilibrium exists. To illustrate these two cases seems useful to later analyse the somewhat intermediate approach of our model. For those border cases we assume a quadratic cost function $C(q) = \frac{q^2}{2}$.

Case 1: Outsiders always coordinate on catastrophe-preventing equilibrium if possible The first case is where the outsiders always coordinate on the equilibrium to prevent the catastrophe, if this exists. The existence of such an equilibrium is dependent on the abatement a possible coalition would leave the outsiders. As mentioned in Lemma 1, an equilibrium where the catastrophe is avoided exists for $\bar{Q} - Q^c \in [0, n\tilde{q}]$. Therefore a coalition could always push each of the outsiders to abate $\tilde{q}$ by leaving a gap of $\bar{Q} - Q^c = n\tilde{q}$. As assumed, $\bar{Q} < N\tilde{q}$, which means that in a situation with a coalition size of $k = 0$ at least one outsider would prefer to form a singleton coalition, thereby abating $\bar{Q} - n\tilde{q} < \tilde{q}$. It is easy to see, that also the remaining outsiders want to join the coalition, as they always end up abating $\frac{\bar{Q}-(N-k)\tilde{q}}{k} < \tilde{q}$, so that their welfare is strictly higher in the coalition. This process only ends when the grand coalition is reached. Latter is stable as a deviation of an coalition member, by abating less than $\bar{Q}/N$ would trigger the catastrophe so that each country would have an additional welfare loss of $X = C(\tilde{q}) > C(\frac{\bar{Q}}{N})$ where the right hand side is the maximum saving a deviating country could gain, by abating 0.

Case 2: Outsiders never coordinate on catastrophe-preventing equilibrium if another equilibrium exists The other extreme case occurs, when the outsiders never coordinate on the catrastophe-preventing equilibrium, if this and the catastrophe-inducing equilibrium exist. As this case is a little bit more complicated, we will solve it using backwards induction also, but being a little bit more explicit, starting with stage 3: The outsiders either abate $Q^n = \bar{Q} - Q^c$ for $\bar{Q} - Q^c \leq \tilde{q}$ or they abate $Q^n = 0$ for $\bar{Q} - Q^c \leq 0$. While the latter is clear, the former comes from Lemma 1, where it is proven, that for $Q^n \leq \tilde{q}$ there is no equilibrium where the outsiders do not abate, therefore the only remaining equilibrium is abating $Q^n \leq \tilde{q}$. This brings us to Stage 2, where the coalition either abates $Q^c = \bar{Q} - \tilde{q}$ for $\bar{Q} - \tilde{q} \leq k\tilde{q}$ and $k < N$, $Q^c = 0$ for $\bar{Q} - \tilde{q} > k\tilde{q}$ or $Q^c = \bar{Q}$ for $\bar{Q} \leq k\tilde{q}$ and $k = N$. It is easy to see that the size of a coalition which abates a positive amount $Q^c = \bar{Q} - \tilde{q} > 0$ is given by $k \geq \frac{\bar{Q}-\tilde{q}}{\tilde{q}}$. As a deviation in stage 2, by a country lowering abatement, would immediately trigger the catastrophe, thus incurring

a damage of $X = C(\tilde{q}) \geq C(q^c)$, there is no incentive for such a behaviour in this stage. Stage 1 comprises the coalition formation. Lets first look at the grand coalition, where $k = N$. Such a coalition avoids the catastrophe with certainty, as $\bar{Q} < N\tilde{q}$ and therefore $C(\frac{\bar{Q}}{N}) \leq C(\tilde{q}) = X$. Deviating at this Stage would give a coalition size of $k = N - 1$, whereas the one outsider will be left with a gap of $\tilde{q}$. This abatement $q^n = \tilde{q}$ is strictly larger than what the deviating country would have had to abate in the coalition $q^c(k^* = N) = \frac{\bar{Q}}{N}$, so that no country has an incentive to leave the grand coalition. This also means that a coalition which comprises $k = N - 1$ members is never stable, because the remaining outsider always prefers to join the coalition. This leaves us with coalition sizes, for which it holds that $k \geq \frac{\bar{Q}-\tilde{q}}{\tilde{q}}$ and $k < N - 1$. This coalition, together with the outsiders would prohibit the catastrophe with certainty. To describe what coalition sizes are possible let us start with the minimal viable coaltion size $k_{min} = \frac{\bar{Q}-\tilde{q}}{\tilde{q}}$. Should one of the coalition members decide to deviate by leaving the coalition, then the remaining coalition stays inactive in stage 2, by not abating, thus triggering the catastrophe. The proof is straight forward. If it holds exactly that $k_{min} = \frac{\bar{Q}-\tilde{q}}{\tilde{q}} \Leftrightarrow Q^c = k\tilde{q} = \bar{Q} - \tilde{q}$, than any country leaving the coalition would cause the coalition to remain inactive, because to prevent the catastrophe the remaining coalition members would still have to abate $Q^c = \bar{Q} - \tilde{q} = k\tilde{q}$ but with $k - 1$ members, thus each member would have to abate strictly more than $\tilde{q}$. In the following we will use $r \equiv \frac{\bar{Q}-\tilde{q}}{\tilde{q}}$ as a short-hand notation. For a coalition of the size $k = r$ to be externally stable, we have to compare the welfares an outsider has in this situation with the welfare a coalition member has at $k^* = k + 1$. If it holds that $\pi^c(k^*) \geq \pi^n(k)$ than the coalition is externally not stable. Formally the condition for external instability can be written as:

$$\pi^c(k^*) = -\frac{(\frac{\bar{Q}-\tilde{q}}{k+1})^2}{2} \geq \pi^n(k) = -\frac{(\frac{\tilde{q}}{N-k})^2}{2}$$

After rearranging we find that the coalition is externally not stable if:

$$r \leq \frac{k+1}{N-k}$$

Please note that the right-hand side of this equation is growing in k. Replacing k with r, as the minimal possible value, we get:

$$r \leq \frac{r+1}{N-r}$$

so that it depends on r and N if the coaltion is externally stable. What can be seen though, is that when the coalition of size k is not externally stable, then only the grand coalition is stable, because the right-hand side of the equation grows in k while the left-hand side stays at r. If on the other hand the coalition is externally stable, we have two

stable coalition sizes. First $k = k_{min}$ and secondly $k = N$.

Stage 2:

In stage 2, there are three types of equilibrium outcomes that may arise. The simplest case is the case where the coalition members decide not to abate any emissions, and the outsiders also don't abate, so that the catastrophe occurs with certainty. In each of the other two cases, the signatories choose positive efforts, and the catastrophe is avoided with strictly positive probability. One possibility is that the signatories choose positive efforts, while outsiders invest a small effort (equal to $Q^n = \tilde{q}$) and the catastrophe is avoided with certainty, hence: $Q^c = \bar{Q} - \tilde{q}$. The other possibility is where $Q^c < \bar{Q} - \tilde{q}$, so that a larger effort of the non-signatories is needed to avoid the catastrophe, and the catastrophe occurs with strictly positive probability (smaller than 1).[1]

We maintain our earlier convention that $q^n \equiv \frac{\bar{Q}-Q^c}{n}$ denotes the average burden per outsider, IF the catastrophe is avoided in stage 3. This still allows for the case where the outsiders only select such an equilibrium with a certain probability, and do not avoid the catastrophe with the remaining probability, $f_n(q^n)$.

We assume that signatories allocate their efforts equally among them. This is the most plausible assumption, given ex ante symmetric countries. In stage 2, the expected welfare per signatory is, thus:

$$\pi^c = -C(q^c) - f_n(q^n)X. \tag{2}$$

The coalition members, acting cooperatively, seek to maximize π^c over q^c, anticipating the behavior of the non-signatories in the subsequent stage 3. Unlike in stage 3, there is no coordination problem among the signatories. Because they take their decisions jointly and cooperatively, they simply select the value of q^c (burden per signatory) that maximizes the welfare per signatory. The welfare maximum is obviously unique. The maximizer, q^c, is generally also unique, except for some knife-edge cases where two (or more) values of q^c lead to the same maximum welfare.

To characterize the outcome of stage 2, we thus proceed in the following steps, corresponding to the three possible types of equilibrium outcome mentioned at the beginning of this subsection. We first calculate the welfare per signatory in the two "extreme cases". One extreme case is where signatories do not abate at all, and non-signatories do not fill the gap. The resulting welfare per country is obviously given by $-X$. The other "extreme" case is where signatories avoid the catastrophe mostly on their own, such that $Q^c = \bar{Q} - \tilde{q}$. This implies $q^c = (\bar{Q} - \tilde{q})/k$, so that $\pi^c = -C((\bar{Q} - \tilde{q})/k)$. Welfare in

[1] Another possible case for some general distribution of the probability-function $f_n(q^n)$ might be $Q^c = 0$, and the outsiders nevertheless avoid the catastrophe with probability 1. To rule this out, we assume that $f_n(q^n = 0) \geq 0$ and $f_n(q^c > 0) > 0$.

these two cases is then compared with the welfare in case of an "interior solution", such that $\bar{Q} - n\tilde{q} < Q^c < \bar{Q} - \tilde{q} \Leftrightarrow \bar{Q} - Q^c \in (\tilde{q}, n\tilde{q})$. If such an interior solution exists, it is characterized by the following condition:

$$C'(q^c) = -f_n'(q^n) X \frac{\partial q^n}{\partial q^c}.$$

We will discuss uniqueness of an interior solution (if such exists) below. Existence of an interior solution requires that the corresponding value of q^c that solves the above first-order condition, indeed lies in the "interior range", i.e., that it satisfies $Q^c = kq^c < \bar{Q} - \tilde{q}$.[2] Otherwise, a corner solution is obtained such that $Q^c = \bar{Q} - \tilde{q}$, and the catastrophe is avoided with certainty.

In a final step, welfare in these cases is compared, and signatories select the value of q^c that delivers the maximum welfare.

In an interior solution, the condition $q^n = \frac{\bar{Q} - Q^c}{n}$ is satisfied, which (together with $Q^c = kq^c$) implies

$$\frac{\partial q^n}{\partial q^c} = -\frac{k}{n}.$$

Furthermore, from the definition of $\tilde{q}$, we have: $X = C(\tilde{q})$. The above first-order condition can thus be rewritten as:

$$C'(q^c) = \frac{k}{n} C(\tilde{q}) f_n' \left(\frac{\bar{Q} - kq^c}{n} \right).$$

If an interior solution exists, this condition implicitly defines the corresponding value of q^c.

If $f_n(q^n)$ is either linear or convex in q^n in the interior range, and $C(q)$ is convex in q, then the solution to the above first-order condition (if there is a solution) is unique. To see this, notice that an increase in q^c leads to a rise in the left-hand side of the condition, $C'(q^c)$, while the right-hand side is either declining in q^c, or constant (in the linear case).

In order to obtain more specific results, we will in the following adopt the linear specification of the function f_n that corresponds to the dashed line in Figure 1 (middle part of the figure). It is straight-forward to verify that the linear specification corresponds to:

$$F_n(\bar{Q} - Q^c) = \frac{\frac{\bar{Q} - Q^c}{\tilde{q}} - 1}{n - 1}.$$

Hence:

$$f_n(q^n) = \frac{nq^n - \tilde{q}}{(n-1)\tilde{q}},$$

[2] The other condition, $Q^c > \bar{Q} - n\tilde{q} \Leftrightarrow \bar{Q} - Q^c < n\tilde{q}$, is never binding, as will be argued below.

which yields:

$$f_n'(q^n) = \frac{n}{(n-1)\tilde{q}}.$$

The above first-order condition then simplifies to:

$$C'(q^c) = \frac{k}{n-1}\frac{C(\tilde{q})}{\tilde{q}}.$$

Let us impose even more structure on the model, by assuming that the abatement cost function has the following quadratic specification: $C(q) = q^2/2$. Then $C'(q^c) = q^c$ and $C(\tilde{q}) = \tilde{q}^2/2$, so that we immediately obtain:

$$q^c = \frac{k\tilde{q}}{2(n-1)} \text{ , and } q^n = \frac{\bar{Q} - kq^c}{n} = \frac{2(n-1)\bar{Q} - k^2\tilde{q}}{2n(n-1)}. \tag{3}$$

By inserting the above expressions for q^c and q^n in (2), the resulting welfare per signatory in case of an interior solution in stage 2 can be written as:

$$\pi^c = -\frac{\tilde{q}^2}{8(n-1)^2}\left(-k^2 + 4(n-1)\frac{\bar{Q}-\tilde{q}}{\tilde{q}}\right).$$

using $r \equiv \frac{\bar{Q}-\tilde{q}}{\tilde{q}}$ as a short-hand notation, and replacing n by $N-k$, this welfare can be expressed as a function of the coalition size k:

$$\pi^c(k) = -\frac{\tilde{q}^2}{8(N-k-1)^2}\left(4(N-k-1)r - k^2\right). \tag{4}$$

Recall that this is the welfare per signatory in case of an interior solution. After some rearrangements, we can also express the probability with which the outsiders fail to prevent the catastrophe (in case of an interior solution) as a function of k:

$$f_n = \frac{2(n-1)r - k^2}{2(n-1)^2},$$

where $n = N - k$.

Let us now compare the above results for the different types of outcomes that may arise. Let us start by investigating when the case where the coalition members decide not to abate any emissions, and the outsiders also don't abate, may arise. Such a case can only emerge as an equilibrium outcome in stage 2, if it holds true that – given $Q^c = 0$ – in stage 3, the non-signatories indeed choose not to abate with probability 1. This amounts to the following condition: $\bar{Q} \geq n\tilde{q}$, hence, if the coalition is sufficiently large: $k \geq N - \bar{Q}/\tilde{q}$.[3] Notice, that this is a necessary, not a sufficient condition, for the case to arise where no country abates any positive amount. Intuitively, given our earlier

[3]It is easy to verify that this condition is equivalent to $r \geq n-1$.

assumptions, non-signatories have the advantage that if they choose to abate a positive amount, they always avoid the catastrophe with certainty. In other words, none of their abatement efforts are "wasted". By contrast, signatories cannot recoup their abatement efforts invested in stage 2, in case the outsiders fails to coordinate on an equilibrium in which the catastrophe is avoided in stage 3. In such a case, the signatories' efforts are indeed wasted. Anticipating this possibility, the signatories may decide not to invest into abatement at all.

Also for the case where the catastrophe is avoided with certainty, and $Q^c = \bar{Q} - \tilde{q}$ so that outsiders contribute only a small amount of $Q^n = \tilde{q}$ to the aggregate abatement, a simple necessary condition can be formulated. Namely, $Q^c = \bar{Q} - \tilde{q}$ can only be optimal if $q^c = Q^c/k = \frac{\bar{Q}-\tilde{q}}{k} < \tilde{q}$, hence, if $k > \frac{\bar{Q}-\tilde{q}}{\tilde{q}} = r$. Otherwise, signatories would rather not abate at all, and let the catastrophe occur with certainty. Again, the condition $k > r$ is just a necessary, not a sufficient condition for this case to arise in equilibrium.

Next, we investigate under what conditions an interior solution can emerge, such that the catastrophe occurs with strictly positive probability (smaller than 1), and signatories as well as non-signatories choose positive abatement effort, given by (3). For this to be the case, it must hold that $q^n \in (\tilde{q}/n, \tilde{q})$ (see Figure 2). The case where $q^n \geq \tilde{q} \Leftrightarrow Q^c \leq \bar{Q} - n\tilde{q}$ cannot emerge. This would imply that signatories leave a "gap" for the non-signatories that is so large, that each of them prefers not to abate at all, rather than to avoid the catastrophe. The catastrophe would, thus, occur with certainty, so that any positive abatement efforts of the signatories would be wasted, amounting to a welfare per signatory below $-X$. Hence, this outcome is from the signatories' perspective strictly inferior to the outcome where no country abates any positive amount at all. By contrast, the condition $q^n > \tilde{q}/n \Leftrightarrow Q^c < \bar{Q} - \tilde{q}$ plays an important role. If it is violated, an interior solution does not exist, as the welfare per signatory is higher in the case where $Q^c = \bar{Q} - \tilde{q}$, so that the catastrophe is avoided with certainty (corner solution). The condition $q^n > \tilde{q}/n$ can be rewritten as follows, using (3):

$$2(n-1)\frac{\bar{Q} - \tilde{q}}{\tilde{q}} \geq k^2,$$

or equivalently:

$$r \geq \frac{k^2}{2(N-k-1)} = \frac{k^2}{2(n-1)} \equiv r_1^{crit}. \tag{5}$$

The other crucial condition for an interior solution to exist is $\pi^c \geq -X$, where π^c is the welfare of a signatory given in (4). If this condition is violated, then signatories prefer not to abate any positive amount, thereby triggering the catastrophe with certainty, rather than to invest an effort of q^c per country according to (3). The condition $\pi^c \geq -X$ can

be rewritten as:

$$4(n-1)\frac{\bar{Q}-\tilde{q}}{\tilde{q}} - k^2 \leq 4(n-1)^2,$$

or equivalently

$$r \leq \frac{4(N-k-1)^2+k^2}{4(N-k-1)} = n-1+\frac{k^2}{4(n-1)} \equiv r_2^{crit}. \tag{6}$$

Next, we analyze for what coalition sizes k an interior solution can exist at all. An interior solution exists if r lies in between the two critical values stated in (5) and (6). Hence, an interior solution can only exist if $r_2^{crit} > r_1^{crit}$, or equivalently:

$$k < \frac{2(N-1)}{3}.$$

If k exceeds this value, which implies that the coalition comprises around two thirds of all countries or more, then there is no interior solution. Then if r is sufficiently small, which means that the damages are large relative to the necessary efforts to avoid them, then the coalition avoids the catastrophe mostly on its own (i.e., $\bar{Q}-Q^c=\tilde{q}$). Otherwise, the signatories prefer not to abate at all, and let the catastrophe occur with probability 1. The indifference point between these two cases is where $r=k$, as this corresponds to $q^c=\tilde{q}$ given $\bar{Q}-Q^c=\tilde{q}$ (see above).

Figure 3 illustrates the values of the crucial parameter r, capturing the damages and necessary efforts to avoid them, for which the three types of outcomes indicated at the beginning of this section occur, depending on the coalition size k. Note, that the critical values for r shown in this figure only depend on the parameter N. They are independent of the parameters X and $\bar{Q}$, reflecting the damages if the catastrophe occurs, resp. the total efforts needed to avoid the catastrophe. These enter the above conditions ((5) and (6)) only via the (combined) parameter r.

Interestingly, as Figure 3 illustrates, the formation of a coalition of an intermediate size can prevent countries from avoiding the catastrophe, in situations where the catastrophe would be avoided with positive probability under a smaller (or a larger) coalition size. This effect is again related with our earlier assumption that in case of an interior solution, coalition members cannot be sure if their efforts are fruitful or wasted, whereas in the absence of any coalition, efforts are never wasted. This leads to an efficiency loss under partial cooperation (i.e., if there is a coalition with relatively few members). By contrast, in the absence of any coalition, countries only avoid the catastrophe with a certain probability, but IF they invest in abatement, the efforts are never wasted and $Q=\bar{Q}$ is precisely satisfied. In the presence of a sufficiently large coalition, this efficiency loss vanishes. But in this case, the outsiders free-ride to a large extent on the efforts of the signatories, who avoid the catastrophe (mostly) on their own. This explains, why for

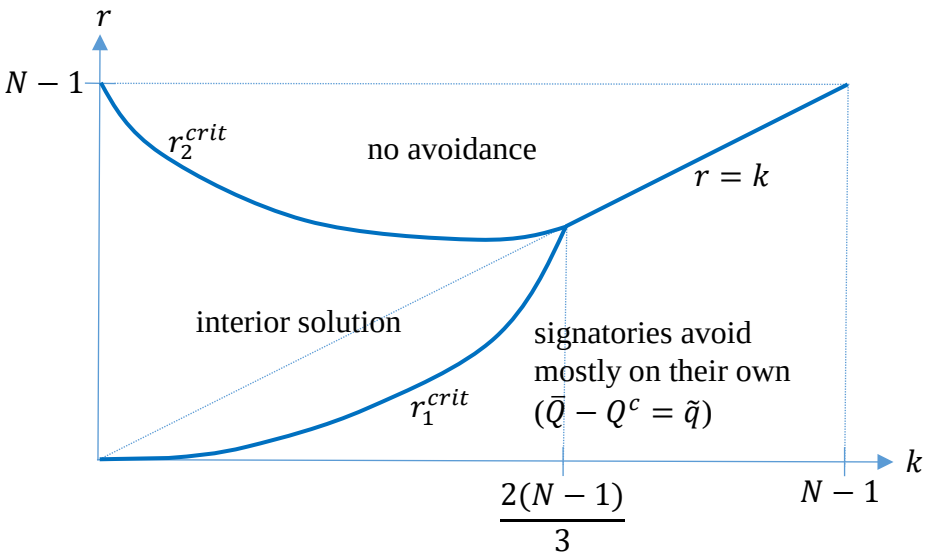

Figure 3: Possible types of outcomes, depending on coalition size k

fairly large values of k ($k > \frac{2(N-1)}{3}$), the catastrophe may still occur with probability 1 for values of r, such that in the absence of any cooperation, the catastrophe would be avoided with strictly positive probability. Only if there is a grand coalition ($k = N$), or a coalition of $N-1$ countries, all of these effects are absent. To see this, notice that if $k = N-1$, the catastrophe is avoided if and only if $r \leq N-1$, which is equivalent to $\bar{Q}/N \leq \tilde{q}$. We are now ready to analyze the equilibrium coalition size(s) in this model.

Stage 1:

There are often multiple 'stable' (i.e., equilibrium) coalition sizes in this model. Let us directly begin with the following observation:

Proposition 1. *Under a known location of the threshold $\bar{Q}$, the grand coalition ($k^* = N$) is always stable, while a coalition of size $N-1$ is never stable.*

Proof of Proposition 1. Recall that our earlier assumption that $\bar{Q} < N\tilde{q}$ implies $r < N-1$. Hence, if a grand coalition forms, it always avoids the catastrophe with certainty, and the welfare per country strictly exceeds $-X$. Under a coalition size of $k = N-1$, by contrast, the coalition members always set $Q^c = \bar{Q} - \tilde{q}$, so that the only remaining outsider obtains a welfare of $-X$. This is because there is no coordination problem in this case, so that the signatories can assure that the catastrophe is avoided, and yet, force the outsider to invest her maximum effort ($\tilde{q}$) to prevent the catastrophe. The outsider, thus, strictly prefers to deviate at the participation stage, thereby joining the coalition. □

The result builds heavily on the assumption that the location of the threshold is precisely known from the outset. In an extension, we will consider also the case where

the location of the threshold is not known with certainty from the outset. Then different results will be obtained, as the coalition members generally cannot be sure that the outsider avoids the catastrophe with certainty, if they leave too much of the burden to the remaining outsider. We will return to this issue later.

The next observation is, that for coalition sizes $k < N$, but sufficiently close to N, any remaining outsiders strictly prefer to deviate and join the coalition. In Proposition 1, this was already indicated for $k = N - 1$, but often, it also holds for somewhat smaller values of k. As k decreases, the incentives to join for the outsiders are, however, declining. This is because (for sufficiently large values of k) the coalition always leaves a 'gap' of the same size: $\bar{Q} - Q^c = \tilde{q}$ for the outsiders to avoid the catastrophe, but as k declines, there are more outsiders ($n = N - k$) who share that burden. Therefore, there is a critical coalition size, below which the incentive for the outsiders to join the coalition are reversed, and instead, it becomes profitable for insiders to leave the coalition (or, more precisely: to deviate at the participation stage by not joining the coalition). To determine the critical coalition size that separates the case where outsiders prefer to deviate by joining from the case where insiders prefer to deviate by not joining the coalition, it is necessary to specify how outsiders share the burden of "closing the gap" to avoid the catastrophe.

Lemma 2. *Under the assumption that outsiders split the burden equally, so that each outsider abates $q^n = (\bar{Q} - Q^c)/n$ to close the gap, provided that $\bar{Q} - Q^c \in (0, \tilde{q}]$, the critical coalition size that separates the case where outsiders prefer to deviate by joining from the case where insiders prefer to deviate by not joining the coalition, is characterized by:*

$$r = \frac{k}{n}.$$

Proof of Lemma 2. Note that for sufficiently large values of k, the catastrophe is always avoided with certainty and the signatories avoid it mostly on their own (so $\bar{Q} - Q^c = \tilde{q}$). Hence, the welfare per signatory equals the welfare per non-signatory if and only if $q^c = q^n$, where $q^c = (\bar{Q} - \tilde{q})/k$, and $q^n = \tilde{q}/n$. This immediately yields $r = k/n$. □

Lemma 2 has an interesting policy implication. Namely, if the international community seeks to coordinate on a cooperative agreement that comprises most (in our model: all) countries, it is necessary to start building a sufficiently large coalition. The coalition size needs to reach at least a certain "threshold size" (in our model, it is implied by the condition $r = k/n$). Once that is achieved, the coalition "stabilizes itself", as any remaining outsiders actually prefer to deviate and join the coalition. This is due to the threshold in the climate system. The signatories leave a burden for the outsiders that is on average larger than the burden per insider. By contrast, if countries fail to coordinate on a coalition that has at least this threshold size, cooperation unravels as countries prefer to exit the coalition.

The "threshold coalition size" that separates the case where outsiders prefer to deviate by joining from the case where insiders prefer to deviate by not joining the coalition, is illustrated in Figure 4 (see the red curve), for the case where the outsiders share their burden of abatement equally. For $k = N - 1$, the critical value $r = k/n$ equals $N - 1$. However, if r is very close to $N-1$, exiting the coalition is not a profitable deviation, as a coalition of size $N-2$ is already too small to prevent the catastrophe at all (see Figure 4). For smaller values of r, given a coalition size below the critical value, a deviation "to the left" (by not joining) is profitable, while for a coalition size above the critical value, a deviation "to the right" by joining is profitable.

If outsiders do not share their burden of abatement equally, there must be at least one outsider that finds deviating at the participation stage by joining the coalition more attractive than implied by the red curve ($r = k/n$) in Figure 4. Given our assumption of ex ante symmetric countries, the most natural case is where the identities of countries that carry a higher burden than others, are determined randomly.[4] Due to the convexity of the abatement cost function, this means that the critical coalition size above which outsiders prefer to deviate and join the coalition, is then smaller. In Figure 4, this would correspond to a (non-parallel) shift of the red curve to the left. To simplify the exposition, we will in the following focus on the case where the outsiders share the burden of abatement equally. Nevertheless, the key insights of our analysis do not depend on this simplifying (and rather natural) assumption. Qualitatively, our main results are preserved also under alternative assumptions about this burden sharing.

To illustrate this, we can look at a rather extreme case, where only one of the outsiders, which is determined randomly, has to abate the missing $\tilde{q}$. In this case each of the outsiders faces a chance of $p = 1/n$ to abate $\tilde{q}$ and a chance of $(1-p) = 1-1/n$ to abate nothing. Therefore the expected cost of an outsider can be written as $E(c^n) = -\frac{\tilde{q}^2}{2n}$. As it still holds, that $f_n = 0$, because $\bar{Q} - Q^c \leq \tilde{q}$, $E(c^n)$ is also the expected welfare of an outsider. As before, it must hold, that $\pi^c = \pi^n$, so $-\frac{-\tilde{q}^2}{2n} = -\frac{1}{2}(\frac{\bar{Q}-\tilde{q}}{\tilde{q}})^2$. Solving this equation yields that the separation line now can be described as $r = \frac{k}{\sqrt{n}} \geq \frac{k}{n}$ for $n, k \geq 1$ resulting in the above described (non-parallel) shift to the left.

As Proposition 1 indicates, the grand coalition is always stable (given a known threshold). Furthermore, as argued above, given a large value of r (very close to $N-1$), even if $k < N$, a deviation of an insider by not joining the coalition is not profitable, if the deviation induces the entire coalition to become inactive in terms of abatement, while it is active without the deviation. The crucial condition here is that a deviation "to the left" (by not joining) can only be profitable, if the critical line $r = k$ is not crossed. This brings us to the next result:

[4]For example, half of the outsiders may not invest in abatement at all, while the others split the burden equally.

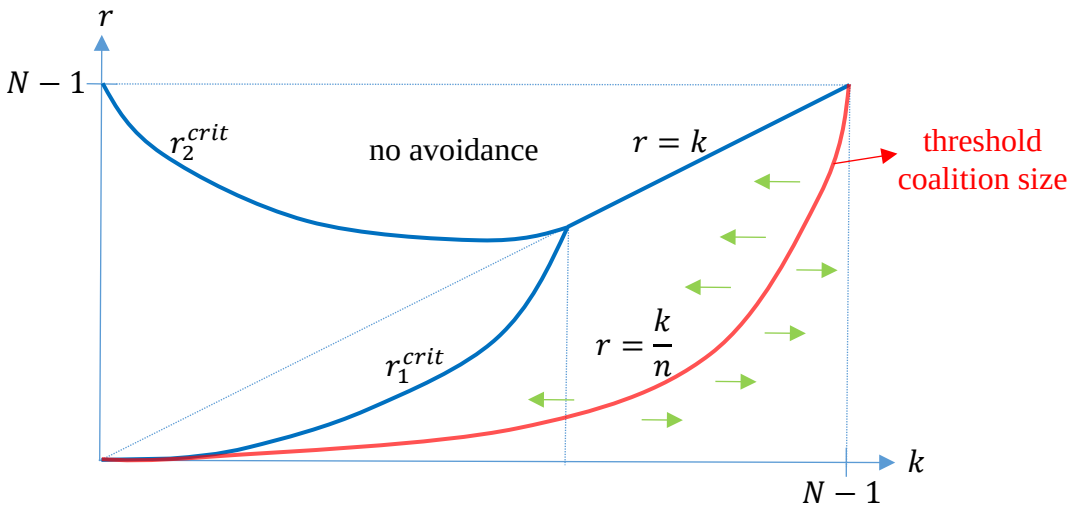

Figure 4: Direction of profitable deviations at stage 1 (green arrows), depending on k

Proposition 2. *If* $r \geq \frac{2(N-1)}{3}$, $k^* = \lceil r \rceil$, *where* $\lceil r \rceil$ *is the smallest integer at least as large as* r, *is a stable coalition size, if* $\lceil r \rceil < N-2$.

Proof of Proposition 2. Given a coalition size that satisfies $k = \lceil r \rceil$, a deviation of a (prospective) coalition member at the participation stage, thereby not joining the coalition, is never profitable, as this deviation leads to a coalition size that falls short of the critical size $k = r$, which renders the entire coalition inactive. The country, thus, strictly prefers to stay inside the coalition (by joining in stage 1). Furthermore, a deviation by an outsider, thereby joining the coalition, is also not profitable, if $\lceil r \rceil < N-2$, because the deviation is insufficient to cross the critical curve $r = k/n$ (a necessary but not sufficient condition for such deviation to be profitable) – see Figure 4. To show that this also holds for small values of N and what actually is the lowest value of N for which such a situation can arise, we go into a little bit more detail here. First we pin down the lowest value N for which it holds that $r \geq \frac{2(N-1)}{3}$ and $\lceil r \rceil < N-2$, so that the described situation can arise. The lowest N clearly comes from $r = \frac{2(N-1)}{3} = N-3$. Thus $N_{min} = 7$. Next we show, that the intersection of $k = r/n$ and a vertical line at $k = N-2$ is lower than $\frac{2(N-1)}{3}$ for $N \geq 7$. To achieve this, we solve $r = k/n = \frac{k}{N-k}$ for k and get the intersection with $k = N-2$. This can be noted as $\frac{Nr}{1+r} = N-2$ and yields $r = \frac{N-2}{2}$, which is clearly smaller than $\frac{2(N-1)}{3}$ for $N > 2$. Therefore the critical line can never be crossed in this situation as it always lies right of $N-2$ for $r = \frac{2(N-1)}{3}$. Only if $\lceil r \rceil = N-2$, deviating from $k = \lceil r \rceil$ to $k' = N-1$ crosses the critical curve $r = k/n$, so that the deviation might be profitable. To see if such a deviation actually is profitable we compare the welfare of an outsider, where $k = N-2$ and the welfare of a coalition member, when $k = N-1$. Former yields $\pi^n(k = N-2) = -\frac{\tilde{q}^2}{2n^2} = -\frac{\tilde{q}^2}{8}$, while the latter can be written as $\pi^c(k = N-1) = -\frac{1}{2}(\frac{\bar{Q}-\tilde{q}}{N-1})^2$. For a deviation of an outsider at $k = N-2$ it must

hold that $\pi^c(k = N-1) > \pi^n(k = N-2)$. Solving this inequality leads to $r \leq \frac{(N-1)}{4}$. Inserting the lowest possible value $r = \frac{2(N-1)}{3}$ yields:

$$\frac{2(N-1)}{3} \leq \frac{N-1}{4}$$

which only holds for $N = 1$. As the described situation only arises if $N > 6$, this poses a contradiction, so that a deviation of an outsider at $k = N-2$ is never profitable for $r \geq \frac{2(N-1)}{3}$. For $\lceil r \rceil = N-1$, a deviation to $k' = N-1$ is always profitable, see Proposition 1. □

Figure 5 illustrates the results of Propositions 1 and 2.[5] The green lines indicate the stable coalition sizes for the given value of the (combined) parameter r.[6] For any value of r (with $r < N-1$, which is equivalent to $\bar{Q} < N\tilde{q}$), $k^* = N$ is a stable coalition size (Proposition 1). Furthermore, for $\frac{2(N-1)}{3} \leq \lceil r \rceil < N-2$ (where the right inequality is equivalent to $r \leq N-3$), also $k^* = \lceil r \rceil$ is a stable coalition size, where $\lceil r \rceil$ is the smallest integer at least as large as r.

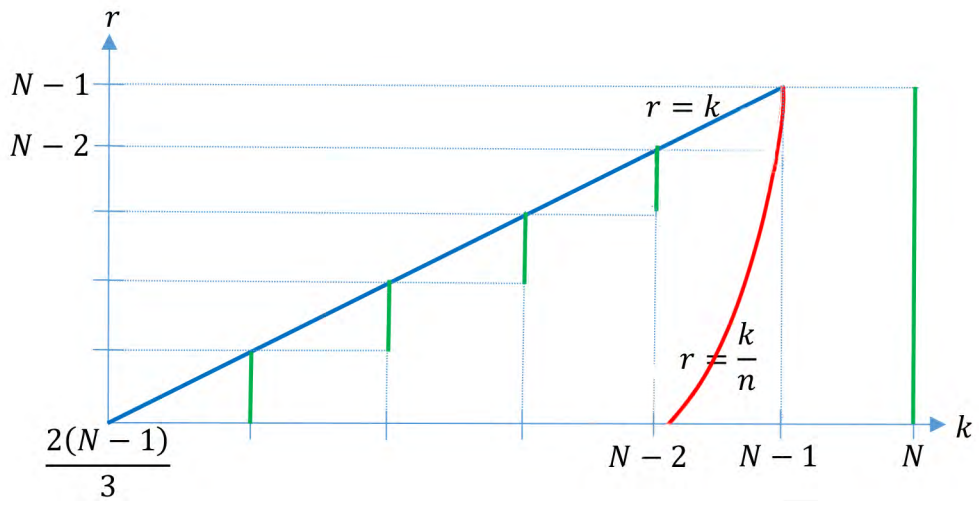

Figure 5: Stable coalition sizes k^* (green lines) for $r \geq \frac{2(N-1)}{3}$

Note, that the equilibrium where $k^* = \lceil r \rceil$ (for $r \geq \frac{2(N-1)}{3}$) is of a "threshold type". This means that the stable coalition size is just large enough to lie above some critical threshold. Here, this threshold is the critical coalition size (characterized by the condition $k \geq r$), below which the signatories would remain inactive, so that the catastrophe occurs for sure. To prevent this inaction of the coalition, the signatories prefer to stay inside

[5] The range of values shown in the figure correspond to a rectangle in the top right corner of Figure 4.

[6] Of course, the stable coalition size k^* depends on the parameter r (and not vice versa). For better comparability with our earlier figures, we maintain the convention that k is on the horizontal, and r on the vertical axis.

the coalition, rather than "leaving it" (by not entering in stage 1). By contrast, the equilibrium with $k^* = N$ is not of a threshold type.

It is interesting to note, that the equilibrium with $k^* = N$ corresponds to the outcome that is obtained under the assumption that outsiders *always* coordinate to avoid the catastrophe, if avoidance is an equilibrium (this variant of the model is analyzed in detail in the section "Border cases"). In that case, the signatories can extract the maximum willingness to abate from the outsiders by lowering their own efforts, so that they obtain a higher welfare than the outsiders. Any remaining outsiders, thus, always prefer to join the coalition. Similarly, the equilibrium where $k^* = \lceil r \rceil$ (for $r \geq \frac{2(N-1)}{3}$) corresponds to the outcome that is obtained under the assumption that outsiders *never* coordinate to avoid the catastrophe, if non-avoidance is an equilibrium (also found in the section "Border cases"). In that case, the outsiders free-ride on the efforts of the signatories, and a coalition forms that is just large enough to invest in abatement. In this sense, our model "inherits" properties of each of these extreme cases and, at least for a range of parameter values, displays both of these equilibrium types. However, the analysis of equilibria in the present model is not completed yet.

Let us now turn to the analysis of stable coalition sizes $k^* \leq \lceil \frac{2(N-1)}{3} \rceil$ (if such exist). Here, the threshold $r = k$ no longer plays a role. Instead, the type of outcome (no avoidance, interior solution, avoidance with $\bar{Q} - Q^c = \tilde{q}$) now depends on the location in the k-r-space relative to the critical values r_1^{crit} and r_2^{crit}, as well as $r = k/n$ (see Figure 3).

Let us define k_1^{crit} as the value of k that solves $k = r_1^{crit}$ for $k \in [0, \frac{2(N-1)}{3}]$. Additionally let us define $k_{2,left}^{crit}$ as the value of k that solves $k = r_2^{crit}$ for k being left of the minimum of the r_2^{crit}-curve and $k_{2,right}^{crit}$ as the value of k that solves $k = r_2^{crit}$ for k being right of said minimum.

In case of an equilibrium in the interior range, hence, a value of k such that an interior solution is obtained in stage 2, the stable coalition size can be approximated by the condition: $\pi_n(k) = \pi_c(k)$. In most cases, this will give a reasonable approximation of the actual coalition size(s) that satisfy the conditions of internal and external stability. The expected payoff of a non-signatory is:

$$\pi^n(k) = -(1 - f_n(q^n))C(q^n) - f_n(k)X.$$

Equalizing the payoff per outsider with the payoff per signatory yields the condition:

$$C(q^c) = C(q^n)(1 - f_n(q^n)).$$

Replacing q^c and q^n by the expressions given in (3) yields a higher-order polynomial in the coalition size k. Hence, we cannot provide a closed-form solution for the stable coalition size, if there is a stable coalition size that yields an interior solution, but this coalition

size can be computed numerically. Still we can provide some conditions for which certain kinds of outcomes have to arise.

Proposition 3. *For any value of $r \in (\frac{2}{N-2}, N-1)$, there exists at least one more stable coalition size, in addition to $k^* = N$, and $k^* = \lceil r \rceil$ (for $r \geq \frac{2(N-1)}{3}$). That additional stable coalition size is strictly smaller than $k^* = \lceil r \rceil$ (for $r \geq \frac{2(N-1)}{3}$), resp. weakly smaller than $k^* = \lceil k_1^{crit} \rceil$ (for $r < \frac{2(N-1)}{3}$).*

Proof of Proposition 3. Consider first a (given) coalition size of $k = 0$. Then either it holds true that at least a single country prefers to form a (singleton) coalition, or no country prefers to cooperate. In the latter case, $k^* = 0$ is a stable coalition size. This case can only arise if r is sufficiently large. Otherwise it must hold that the expected welfare of the single coalition member would be higher than the expected welfare the country would have received, when staying an outsider. Therefore we have to compare those welfares, where we can write the expected welfare for every country, if the coalition size equals zero, to be

$$\pi^n(k=0) = -\tilde{q}^2 \frac{N^2 r + N(1+r)^2 - (1+r)^3}{2N^2(N-1)}.$$

and the expected welfare of the single coalition member to be

$$\pi^c(k=1) = -\frac{\tilde{q}^2(4(N-2)r - 1)}{8(N-2)^2}$$

From those two equations we can calculate if a single coalition member can achieve a higher welfare, then by staying outside of the coalition and therefore gaining the welfare where there is no coalition at all. By setting $\pi^n(k=0) = \pi^c(k=1)$ and solving for r we can get the value of r, where at least one country would be indifferent to form a singleton coalition. This value can be written as:

$$\begin{aligned} &\frac{\sqrt[3]{N^2\left(8N^3 - 68N^2 + 3\sqrt{3}\sqrt{64N^4 - 696N^3 + 2747N^2 - 4302N + 2187} + 239N - 243\right)}}{6\,(N-2)^{2/3}} \\ &\qquad + \frac{2\,(N-5)\,N^{\frac{8}{3}}\sqrt[3]{N^2 - 4N + 4}}{\sqrt[3]{3N-6}} \\ &\cdot \frac{1}{\sqrt[3]{8N^3 - 68N^2 + 3\sqrt{3}\sqrt{64N^4 - 696N^3 + 2747N^2 - 4302N + 2187} + 239N - 243}} \\ &\qquad + \frac{N}{3} - 1 = r_{max} \end{aligned} \tag{7}$$

For all values of r above this r_{max}, no country has an incentive to form a singleton

coalition, which means, that if additionally $r \geq \frac{2(N-1)}{3}$ holds, the only stable coalition size in addition to $k^* = N$, and $k^* = \lceil r \rceil$ is $k^* = 0$. This is area A in Figure 7.

If on the other hand $r \leq r_{max}$ and $r \geq \frac{2(N-1)}{3}$ (Area B in Figure 7) then there has to be an interior solution with $k^* \geq 1$. This is because even if the singleton coalition is not externally stable, the coalition size may only grow so large that the r_2^{crit}-curve is not crossed ($k \leq \lfloor k_{2,left}^{crit} \rfloor$). This area can only exist if $N \geq 6$ as only then r_{max} is strictly larger than $\frac{2(N-1)}{3}$ as can be seen in see Figure 6.

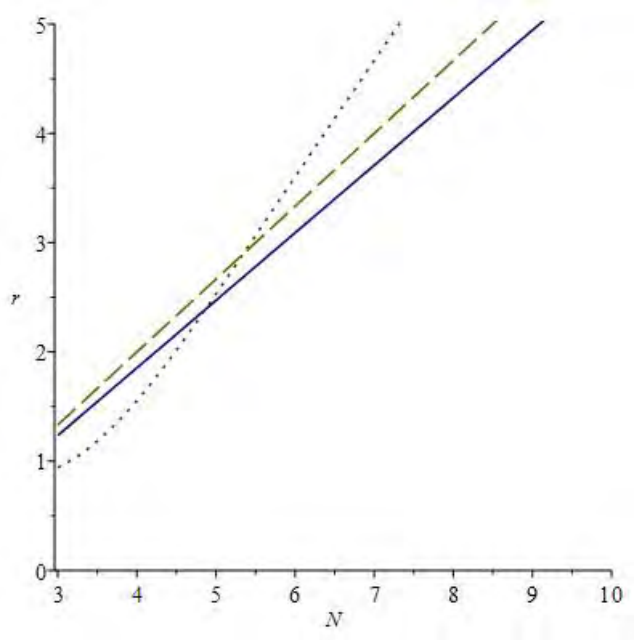

Figure 6: r_{max} (dotted), $r_{2,min}^{crit}$ (solid) and $r = \frac{2(N-1)}{3}$ (dashed) depending on N

The next area we look at is where it holds that $r \leq r_{max}$ and $r < \frac{2(N-1)}{3}$. This area can be further divided into a small part where r lies between $\frac{2(N-1)}{3}$ and above the minimum of the r_2^{crit}-curve (Area C in Figure 7) and the larger rest below, but where $r > \frac{2}{N-2}$ still holds.

To get the r-value of the minimum of the r_2^{crit}-curve we calculate the first order condition through differentiation with respect to k:

$$\frac{\delta r_2^{crit}}{\delta k} = \frac{-4N^2 + (10k+8)N - 5k^2 - 10k - 4}{4(N-k-1)^2} = 0$$

This yields two solutions for k, of which only the following constitutes the k for a minimum, which can be proven easy by the second derivative (not shown).

$$k^{min} = \left(1 - \frac{\sqrt{5}}{5}\right)(N-1)$$

Inserting this back into the equation for r_2^{crit} yields:

$$r_{2,min}^{crit} = \frac{(\sqrt{5}-1)(N-1)}{2}$$

In this area it is certain, that there are two additional stable coalition sizes. This situation alwas occurs for $N \geq 3$ as can be seen in Figure 6, because it must hold that $r_{2,min}^{crit} < \frac{2(N-1)}{3}$. One has to be a true interior solution, for the same reasons discussed for area B. The second stable coalition size can be of the interior kind or it may be at $k^* = \lceil k_1^{crit} \rceil$. This is due to the fact, that for coalition sizes below the values of k that satisfy $r = k/n$, countries have an incentive to leave the coalition, unless this directly triggers the no-avoidance or they reach the interior zone. Therefore these incentives to exit are not reversed for $r < \frac{2(N-1)}{3}$, as long as $k > \lceil k_1^{crit} \rceil$ continues to hold. This follows from the fact that in this range, it always holds that $\bar{Q} - Q^c = \tilde{q}$, but the more outsiders there are, the lower the burden per outsider. This implies that the coalition size falls, at least until we enter the range of the interior solution. Hence, there must be a stable coalition size weakly smaller than $k^* = \lceil k_1^{crit} \rceil$ (for $r < \frac{2(N-1)}{3}$), and greater than $k_{2,right}^{crit}$. Only if $k_{2,right}^{crit} \geq \lceil k_1^{crit} \rceil - 1$ it is clear, that the second additional equilibrium is exactly at $k^* = \lceil k_1^{crit} \rceil$, because any coalition member leaving at this k^* would immediately put the remaining coalition in the no avoidance zone.[7]

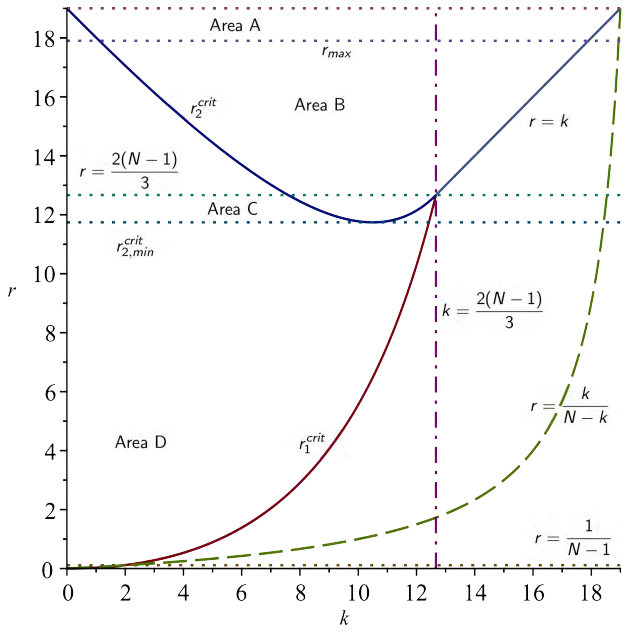

Figure 7: Different areas of Proposition 3 for $N = 20$

[7] Actually, there is a very small area directly above the minimum of the r_2^{crit}-curve, where the wideness of the avoidance zone is smaller than 1. Due to the integer constraint on k, it could happen that the no-avoidance-zone is simply jumped when the coalition size k grows or shrinks by 1. However, up to today no case of this actually happening could be found numerically.

For the last area D in Figure 7 of this Proposition it can only be concluded, that there is a stable coalition size k for which it must hold that $1 \leq k^* \leq \lceil k_1^{crit} \rceil$. This reasoning is quite analogue to the right-hand side interior equilibrium in area C. For coalition sizes below the values of k that satisfy $r = k/n$ and $r \leq r_{2,min}^{crit}$, countries will always leave the coalition, unless they reach the interior zone. Therefore these incentives to exit are not reversed for $r < \frac{2(N-1)}{3}$, as long as $k > \lceil k_1^{crit} \rceil$ continues to hold. Hence, there must be a stable coalition size weakly smaller than $k^* = \lceil k_1^{crit} \rceil$ and strictly greater than 0.

□

Proposition 4. *For any value of $r \in (0, \frac{1}{N-1})$, there exists only $k^* = N$ as a stable coalition size.*

To be continued: This has to be checked for $r \in [\frac{1}{N-1}, \frac{2}{N-2}]$ as the singleton coalition $k^* = 1$ is feasible and $k' = k^* + 1$ would cross the critical curve $r = k/n$ of $k = r/n$ as a necessary but not sufficient condition for this deviation to be profitable. If it is profitable only the grand coalition would be stable. Otherwise additionally $k^* = 1$ would be a stable coaliton size.

Proof of Proposition 4. This proof is straight forward. As for $r < \frac{1}{N-1}$ it always holds, that $r < r_{max}$ so the singleton coalition will be formed. Upon forming this coalition the critical curve $r = k/n$ is already crossed, so that the coalition grows until the grand coalition is reached. Therefore, there can not be any other equilibrium for this range of r-values.

□

2.1 Preliminary numerical equilibrium and welfare analysis

As mentioned before, the actual stable coalition sizes in the described model depend on the used parameters and the form of the cost- and probability-function used in the model. Also the different welfare outcomes therefore depend on those. In the following we provide a numerical example using the above specified forms for the cost- and probability-function (quadratic, linear). Additionally we will set $\bar{Q} = 1000$ and $N = 20$, so that Figure 7 still applies. X and therefore $\tilde{q}$ will be implicitly determined by the level of r. To keep the analysis compact, we will only provide one example for every area B-D, so that we will look at 3 different levels of r in the course of this section. For each example we will calculate the stable coalition sizes, the abatement levels for coalition members and outsiders, their resulting welfares and the global welfare Π^{global}. The aim of this section is to show what welfare level in comparison to the grand coalition can be achieved at the additional stable coalition sizes.

Starting in area B, we set $r = 13$ and get the results shown in Table 1. Due to space limitations we only show coalition sizes relevant for the analysis of stable equilibria.

Table 1: Results for $r = 13$ (Area B)

r	13						
k	0	1	5	6	7	13	20
q_n	50.0	52.53	62.41	64.36	65.71	10.2	—
q_c	—	1.98	12.76	16.48	20.83	71.43	50
$\tilde{q}$	71.43	71.43	71.43	71.43	71.43	71.43	71.43
$f_n(q^n)$	0.68	0.72	0.86	0.89	0.91	0	0
π^n	-2140.17	-2223.8	-2469.46	-2499.93	-2516.95	-52.06	—
pi^c	—	-1840.44	-2287.46	-2415.17	-2546.59	-2551.02	-1250
Π^{global}	-42803.44	-44092.72	-48479.26	-49490.08	-50546.55	-33527.7	-25000

Beginning with $k = 0$ it can be seen that this coalition size is externally unstable, so that at least the singleton coalition will be formed. Upon forming the singleton coalition the only member significantly reduces its abatement efforts, while the outsiders raise abatement levels (if they coordinate on the catastrophe-preventing equilibrium) to compensate this reduction. The described external instability continues (not shown) up until $k = 6$. here we find that this coalition size is externally and internally stable. Interestingly at this coalition size the expected global welfare is lower than in the case of $k = 0$, where there is no cooperation at all. [8]. Also the probability of the catastrophe to occur is significantly larger then with $k = 0$. The second stable coalition size is the threshold equilibrium (non-interior) at $k = 13$. Here we have a significant higher welfare than for the interior solution. The last stable equilibrium for this parameter constellation is the grand coalition ($k = 20$), which also yields the highest global welfare.

Table 2: Results for $r = 12$ (Area C)

r = 12								
k	0	1	6	7	8	12	13	20
q_n	50.0	52.52	63.82	64.84	64.69	26.1	10.99	—
q_c	—	2.14	17.75	22.44	27.97	65.93	71.01	50
$\tilde{q}$	76.92	76.92	76.92	76.92	76.92	76.92	76.92	76.92
$f_n(q^n)$	0.63	0.67	0.82	0.83	0.83	0.24	0	0
π^n	-2329.1	-2429.66	-2789.45	-2812.89	-2808.2	-981.72	-60.38	—
π^c	—	-1970.1	-2573.44	-2706.9	-2836.32	-2898.2	-2520.92	-1250
Π^{global}	-46582.06	-48133.62	-54492.93	-55515.78	-56388.97	-42632.17	-33194.62	-25000

Moving to area C, we set $r = 12$ and obtain the results shown in Table 2. Analogous to area B, the singleton coalition is formed and the coalitions reimain externally unstable until $k = 7$, which is the first interior solution for this parameter combination. As in area B this solution yields a lower global benefit and a higher probability of the catastrophe to happen than the situation where $k = 0$. This time there is a second interior solution for $k = 12$. This equilibrium yields a slightly higher global welfare than for $k = 0$, but

[8] Latter is just used as a benchmark, as it is not a stable coalition size. The other benchmark is the grand coalition which is a stable equilibrium.

Table 3: Results for $r = 4$ (Area D)

r	4					
k	0	1	6	7	8	20
q_n	50.0	52.34	51.65	45.51	34.85	—
q_c	—	5.56	46.15	58.33	72.73	50
$\tilde{q}$	200.0	200.0	200.0	200.0	200.0	200.0
$f_n(q^n)$	0.21	0.22	0.2	0.16	0.1	0
π^n	-5197.37	-5481.01	-5089.11	-4130.58	-2530.46	—
π^c	—	-4429.01	-5088.76	-4965.28	-4628.1	-1250
Π^{global}	-103947.37	-108568.25	-101780.1	-88454.43	-67390.32	-25000

still is quite far from the result of the grand coalition. Interestingly the probability for the catastrophe to occur is significantly lower.

Last we look at area D, therefore setting $r = 4$, which yields the results shown in Table 3. Again we find a stable interior solution, this time at $k = 7$. In this case it holds, that the global welfare is slightly higher and $f_n(q^n)$ is slightly reduced than for $k = 0$, but also this interior solution falls short, when compared to the results of the grand coalition at $k = 20$.

To be continued: Maybe we can provide an analytical solution, when interior solutions are worse than no coalition at all.

3 Preliminary Conclusion

First of all we could show, that, by introducing a probability for the outsiders to not coordinate on the catastrophe-preventing equilibrium if there exists more than one feasible equilibrium, there may be an additional kind of stable coalition sizes, which we describe as interior solutions. Those interior solutions are characterized by the fact that, even as the threshold is known with certainty, the catastrophe is only avoided with a certain probability strictly larger than 0 and lower than 1. Generally it seems to hold, that equilibria in the interior range (if such exist), do not yield significant welfare gains, relative to the fully non-cooperative benchmark. In fact, the resulting welfare under this type of cooperation can be even lower than in the absence of any cooperation, as we showed with some numerical examples. The reason for this finding is that for very low values of k, the few signatories only join the coalition in order to commit themselves not to abate a lot. Then the non-signatories raise their efforts in response. So in a small coalition, the signatories are actually the free-riders, while the outsiders abate more (in case they coordinate to prevent the catastrophe). Only if the coalition size is sufficiently large, the coalition members move from "free-riding" towards "contributing a lot", while the outsiders more and more play the role of free-riders. Hence, a small coalition (in the interior range) is only profitable to a rather small extent, if at all.

Literature

Barrett, Scott (1994): Self-Enforcing International Environmental Agreements. In: Oxford Economic Papers 46, p. 878–894.

Barrett, Scott (2013): Climate treaties and approaching catastrophes. In: Journal of Environmental Economics and Management 66 (2), p. 235–250.

Karp, Larry S.; Sakamoto, Hiroaki (2019): Sober optimism and the formation of international environmental agreements. Available at SSRN 3337139.

Lenton, T. M. (2011): Beyond 2 C: redefining dangerous climate change for physical systems. In: Wiley Interdisciplinary Reviews: Climate Change 2 (3), p. 451–461.

Schmidt, R. C. (2017). Dynamic cooperation with tipping points in the climate system. Oxford Economic Papers, 69(2), 388-409.