Information Geometry of Noncooperative Games

Nils Bertschinger
David H. Wolpert
Eckehard Olbrich
Juergen Jost

SFI WORKING PAPER: 2014-06-017

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author’s copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu
INFORMATION GEOMETRY OF NONCOOPERATIVE GAMES

NILS BERTSCHINGER\textsuperscript{a}, DAVID H. WOLPERT\textsuperscript{b}, ECKEHARD OLBRICH\textsuperscript{a} AND JÜRGEN JOST\textsuperscript{a, b}

In some games, additional information hurts a player, e.g., in games with first-mover advantage, the second-mover is hurt by seeing the first-mover's move. What are the conditions for a game to have such negative "value of information" for a player? Can a game have negative value of information for all players? To answer such questions, we generalize the definition of marginal utility of a good (to a player in a decision scenario) to define the marginal utility of a parameter vector specifying a game (to a player in that game). Doing this requires a cardinal information measure; for illustration we use Shannon measures. The resultant formalism reveals a unique geometry underlying every game. It also allows us to prove that generically, every game has negative value of information, unless one imposes \textit{a priori} constraints on the game's parameter vector. We demonstrate these and related results numerically, and discuss their implications.

\textbf{Keywords:} Game theory, Value of information, Shannon information, Information geometry.

1. INTRODUCTION

How a player in a noncooperative game behaves typically depends on what information she has about her physical environment and about the behavior of the other players. Accordingly, the joint behavior of multiple interacting players can depend strongly on the information available to the separate players, both about one another, and about Nature-based random variables. Precisely how the joint behavior of the players depends on this information is determined by the preferences of those players. So in general there is a strong interplay among the information structure connecting a set of players, the preferences of those players, and their behavior.

This paper presents a novel approach to study this interplay, based on generalizing the concept of "marginal value of a good" from the setting of a single decision-maker in a game against Nature to a multi-player setting. This approach uncovers a unique (differential) geometric structure underlying each noncooperative game. As we show, it is this geometric structure of a game that governs the associated "interplay among the information structure of the game, the preferences of the players, and their behavior". Accordingly, we can use this geometric structure to analyze how changes to the information structure of the game affects the behavior of the players in that game, and therefore affects their expected utilities.

This approach allows us construct general theorems on when there is a change to an information structure that will \textit{reduce} information available to a player but \textit{increase} their expected utility. It also allows us to construct extended "Pareto" versions of these theorems, specifying when there is a change to an information structure that will both reduce information available to \textit{all} players and increase \textit{all} of their expected utilities.

\textsuperscript{a}Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
nils.bertschinger@mis.mpg.de; olbrich@mis.mpg.de; jjost@mis.mpg.de
\textsuperscript{b}Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
http://davidwolpert.weebly.com
We illustrate these theoretical results with computer experiments involving the noisy leader-follower game. We also discuss the general implications of these results for well-known issues in the economics of information.

1.1. Value of information

Intuitively, it might seem that a rational decision maker cannot be hurt by additional information. After all, that is the standard interpretation of Blackwell’s famous result that adding noise to an observation by sending it through an additional channel, called *garbling*, cannot improve expected utility of a Bayesian decision maker in a game against Nature (Blackwell, 1953). However games involving multiple players, and/or bounded rational behavior, might violate this intuition.

To investigate the legitimacy of this intuition for general noncooperative games, we first need to formalize what it means to have “additional information”. To begin, consider the simplest case, of a single-player game. We can compare two scenarios: One where the player can observe a relevant state of nature, and another situation that is identical, except that now she cannot observe that state of nature. More generally, we can compare a scenario where the player receives a noisy signal about the state of nature to a scenario that is identical except that the signal she receives is strictly noisier (in a certain sense) than in the first scenario. Indeed, in his seminal paper Blackwell (1953), Blackwell characterized precisely those changes to an information channel, namely adding noise by sending the signal through an additional channel, that can never increase the expected utility of the player. So at least in a game against Nature, one can usefully define the “value of information” as the difference in highest expected utility that can be achieved in a low noise scenario (more information) compared to a high noise scenario (less information), and prove important properties about this value of information.

In trying to extend this reasoning from a single player game to a multi-player game two new complications arise. First, in a multi-player game there can be multiple equilibria, with different expected utilities from one another. All of those equilibria will change, in different ways, when noise is added to an information channel connecting players in the game. Indeed, even the number of equilibria may change when noise is added to a channel. This means there is no well-defined way to compare equilibrium behavior in a “before” scenario with equilibrium behavior in an “after” scenario in which noise has been added; there is arbitrariness in which pair of equilibria, one from each scenario, we use for the comparison. Note that there is no such ambiguity in a game against Nature. (In addition, this ambiguity does not arise in the Cournot scenarios discussed below if we restrict attention to perfect equilibria.)

A second complication is that in a multi-player game all of the players will react to a change in an information channel, if not directly then indirectly, via the strategic nature of the game. This effect can even result in a negative value of information, in that it means a player would prefer less (i.e., noisier) information. Indeed, such negative value of information can arise even when both the “before” and “after” scenarios have unique (subgame perfect) equilibria, so that there is no ambiguity in choosing which two equilibria to compare.
To illustrate this, consider the Cournot duopoly where two competing manufacturers of a good each choose a production level. Assume that one player — the “leader” — chooses his production level first, but that the other player, the “follower”, has no information about the leader’s choice before making her choice. So as far as its equilibrium structure is concerned, this scenario is equivalent to a simultaneous-move game. Assuming that both players can produce the good for the same cost and that the demand function is linear, it is well known that in that equilibrium both players get the same profit.

Now change the game by having the follower observe the leader’s move before she moves. So the only change is that the follower now has more information before making her move. In this new game, the leader can choose a higher production level compared to the production level of the simultaneous move game — the monopoly production level — and the follower has to react by choosing a lower production level. Thus, the follower is actually hurt by this change to the game that results in her having more information.

In this example, the leader changes his move to account for the information that (he knows that) the follower will receive. Then, after receiving this information, the follower cannot credibly ignore it, i.e., cannot credibly behave as in the simultaneous move game equilibrium. So this equilibrium of the new game, where the follower is hurt by the extra information, is subgame-perfect. These and several other examples of negative value of information can be found in the game theory literature (see section 1.7 for references).

In this paper we introduce a broad framework that overcomes these two complications which distinguish multi-player games from single-player games. This framework is based on generalizing the concept of the “marginal value of a good”, to a decision-maker in a game against Nature, so that it can apply to multi-player game scenarios. This means that in our approach, the “before” and “after” scenarios traditionally used to define value of information in games against Nature are infinitesimally close to one another. More precisely, we consider how much the expected utility of a player changes as one infinitesimally changes the conditional distribution specifying the information channel in a game, where one is careful to choose the infinitesimal change to the information channel that maximizes the associated change in the amount of information in the channel. (This is illustrated in Fig. 1.)

In the next subsection we provide a careful motivation for our “marginal value of information” approach. As we discuss in the following subsection, this careful motivation of our approach shows that it requires us to choose both a cardinal measure of amount of information, and an inner product to relate changes in utility to changes in information. We spend the next two subsections discussing how to make those choices. Next we discuss the broad benefits of our approach, e.g., as a way to quantify marginal rates of substitution of different kinds of information arising in a game. After this we relate our framework to previous work. We end this section by providing a roadmap to the rest of our paper.
Figure 1.— Both the expected utility of player $i$ and the amount of information player $i$ receives depend, in part, on the strategy profile of all the players, $\sigma$. Via the equilibrium concept, that profile in turn depends on the specific conditional distributions $\theta$ in the information channel providing data to player $i$. So a change to $\theta$ results in a coupled change to the expected utility of player $i$, $E_{\theta}[u_i]$, and to the amount of information in their channel, $f(\theta)$. The “marginal value of information” to $i$ is how much $E_{\theta}(u_i)$ changes if $\theta$ is changed infinitesimally, in the direction in distribution space that maximizes the associated change in $f(\theta)$.

1.2. From consumers making a choice to multi-player games

To motivate our framework, first consider the simple case of a consumer whose preference function depends jointly on the quantities of all the goods they get in a market. Given some current bundle of goods, how should we quantify the value they assign to getting more of good $j$? The usual economic answer is the marginal utility of good $j$ to the consumer, i.e., the derivative of their expected utility with respect to amount of good $j$.

Rather than ask what the marginal value of good $j$ is to the consumer, we might ask what their marginal value is for some linear combination of $j$ and a different good $j'$. The natural answer is their “marginal value” is the marginal utility of that precise linear combination of goods.

More generally, rather than consider the marginal value to the consumer of a linear combination of the goods, we might want to consider the marginal value to them of some arbitrary, perhaps non-linear function of quantities of each of the goods. What marginal value would they assign to that?

To answer this question, write the vector of quantities of goods the consumer possesses as $\theta$. Then write the consumer’s expected utility as $V(\theta)$, and the amount of good $j$ as the function $g(\theta) = \theta_j$. So in a round-about way, we can formulate the marginal value they assign to good $j$ is the directional derivative of their expected utility $V(\theta)$, in the direction in $\theta$ space of maximal gain in the amount of good $j$. That quantity is just the projection of the gradient (in $\theta$ space) of $V(\theta)$ onto the gradient of $g(\theta)$.

Stated more concisely, the marginal value the consumer assigns to $g(\theta)$ is the projection of the gradient of $V(\theta)$ onto the gradient of $g(\theta)$. Now if instead we set $g(\theta) = \sum_i \alpha_i \theta_i$, then $g$ now specifies a linear combination of the goods. However it is still the case that the marginal value they assign to $g(\theta)$ is the projection of the gradient of $V(\theta)$ onto the gradient of $g(\theta)$. In light of this, it is natural to quantify the marginal value
the consumer assigns to any scalar-valued function \( f(\theta) \) — even a nonlinear one — as the projection of the gradient of \( V(\theta) \) on the gradient of \( f(\theta) \). Loosely speaking, this projection is how much expected utility would change if the value of \( f \) were changed infinitesimally, but to first order no other degree of freedom aside from the value of \( f \) were changed. More formally, it is given by the dot product between the two gradients of \( V \) and \( f \), after the gradient of \( f \) is normalized to have unit length.

We have to be a bit more careful than in our reasoning though, due to unit considerations. To be consistent with conventional terminology, we would like to define how much the consumer values an infinitesimal change to \( f \) expressed per unit of \( f \). Indeed, we would typically say that how much the consumer values a change to good \( j \) is given by the associated change in utility divided by the amount of change in good \( j \). (After all, that tells us change in utility per unit change in the good.)

Based on this reasoning, we propose to measure the value of an infinitesimal change of \( f \) as

\[
\frac{\langle \nabla V, \nabla f \rangle}{||\nabla f||^2} \tag{1}
\]

where the brackets indicate a dot product, and the double-vertical lines are the norm under this dot product. The measure in (1) says that if a small change in the value of \( f \) leads to a big change in expected utility, \( f \) is more valuable than if the same change in expected utility required a bigger change in the value of \( f \).

All of the reasoning above can be carried over from the case of a single consumer to apply to multi-player scenarios. To see how, first note that in the reasoning above, \( \theta \) is simply the parameter vector determining the utility of the consumer. In other words, it is the parameter vector specifying the details of a game being played by a decision maker in a game against Nature. So it naturally generalizes to a multi-player game, as the parameter vector specifying the details of such a game.

Next, replace the consumer player in the reasoning above by a particular player in the multi-player game. The key to the reasoning above is that specifying \( \theta \) specifies the expected utility of the consumer player. In the case of the consumer, that map from parameter vector to expected utility is direct. In a multi-player game, that direct map becomes an indirect map specified in two stages: First by the equilibrium concept, taking \( \theta \) to the mixed strategy profile of all the players, and then from that profile to the expected utility of any particular player. (Cf., Fig. 1.)

As mentioned above though, there is an extra complication in the multi-player case that is absent in the case of the single consumer. Typically multi-player games have multiple equilibria for any \( \theta \), and therefore multiple values of \( V(\theta) \). (In Fig. 1, the map from \( \theta \) to the mixed strategy profile is multi-valued in games with multiple players.) However we need to have the mapping from \( \theta \) to the expected utility of the players be single-valued to use the reasoning above. This means that we have to be careful when calculating gradients to specify which precise branch of the set of equilibria we are

\footnote{While this quantification of value of a change to \( f \) may accord with common terminology, it has the disadvantage that it may be infinite, depending on the current \( \theta \) and the form of \( f \). Thus, in a full analysis, it might be useful to just study the dot product between the gradient of expected utility and the gradient of \( f \), in addition to the measure in (1). For reasons of space though, we do not consider such alternatives here.}
considering. Having done that, our generalization from the definition of marginal utility for the case of a consumer choosing a bundle of goods to marginal utility for a player in a multi-player game is complete.

1.3. General comments on the marginal value approach

There are several aspects of this general framework that are important to emphasize. First, in either the case of a game against Nature (the consumer) or a multi-player game, there is no reason to restrict attention to Nash equilibria (or some appropriate refinement). All that we need is that $\theta$ specifies (a set of) equilibrium expected utilities for all the players. The equilibrium concept can be anything at all.

Second, note that $\theta$, together with the solution concept and choice of an equilibrium branch, specifies the mixed strategy profile of the players, as well as all prior and conditional probabilities. So it specifies the distributions governing the joint distribution over all random variables in the game. Accordingly, it specifies the values of all cardinal functions of that joint distribution. So in particular, however we wish to quantify “amount of information”, so long as it is a function of that joint distribution, it is an indirect function of $\theta$ (for a fixed solution concept and associated choice of a solution branch). This means we can apply our analysis for any such quantification of the amount of information as a function $f(\theta)$.

We have to make several choices every time we use this approach. One is that we must choose what parameters of the game to vary. Another choice we must make is what precise function of the dot products of gradients to use, e.g. whether to consider normalized ratios as in Eq. (1) or a non-normalized ratio. Taken together these choices fix what economic question we are considering. Similar choices (e.g., of what game parameters to allow to vary) arise, either implicitly or explicitly, in any economic modeling.

In addition to these two issues, there are two other issues we must address. First, we must decide what information measures we wish to analyze. Second, we confront an additional, purely formal choice, unique to analyses of marginal values. This is the choice of what coordinate system to use to evaluate the dot products in Eq. (1). The difficulty is that changing the coordinate system changes the values of both dot products and gradients in general — both of which occur in Eq. (1). So different choices of coordinate system would give different marginal values of information. However since the choice of coordinate system is purely a modeling choice, we do not want our conclusions to

---

2To give a simple example that the dot product can change depending on the choice of coordinate system, consider the two Cartesian position vectors $(1, 0)$ and $(0, 1)$. Their dot product in Cartesian coordinates equals 0. However if we translate those two vectors into polar coordinates we get $(1, 0)$ and $(1, \pi/2)$. The dot product of these two vectors is 1, which differs from 0, as claimed.

3To give a simple example that gradients can change depending on the choice of coordinate system, consider the gradient of the function from $\mathbb{R}^2 \to \mathbb{R}$ defined by $h(x, y) = x^2 + y^2$ in Cartesian coordinates. The vector of partial derivatives of $h$ in Cartesian coordinates is the (Cartesian) vector $(2x, 2y)$. However if we express $h$ in polar coordinates, and evaluate the vector of partial derivatives with respect to those coordinates, we get $(\partial h / \partial r, \partial h / \partial \theta) = (2r, 0)$, which when transformed back to Cartesian coordinates is the vector $(2x, 0)$. So the gradients change, as claimed.
change if we change how we parametrize the noise level in a communication channel, for example.

We address this second pair of issues in turn, in the next two subsections.

1.4. How to quantify information in game theory

To use the framework outlined in Sec. 1.2, we must choose a function \( f(\theta) \) that measures the amount of information in a game with parameter \( \theta \). Traditionally, a player’s information is represented in game theory by a signal that the player receives during the game.\(^4\) Thus information is often thought of as an object or commodity. But this general approach does not integrate the important fact that a signal is only informative to the extent that it changes what the player believes about some other aspect of the game. It is the relationship between the value of the signal and that other aspect of the game that determines the “amount of information” in the signal.

More precisely, let \( y \), sampled from a distribution \( p(y) \), be a payoff-relevant variable whose state player \( i \) would like to know before making her move, but which she cannot observe directly. Say that the value \( y \) is used to generate a datum \( x \), and that it is \( x \) that the player directly observes, via a conditional distribution \( P(x \mid y) \). If for some reason the player ignored \( x \), then she would assign the a-priori likelihood \( P(y) \) to \( y \), even though in fact its a-posteriori likelihood is \( P(y \mid x) = \sum_{y'} p(y') p(x \mid y') p(y) / p(x) \). This difference in the likelihoods she would assign to \( y \) is a measure of the information that \( x \) provides about \( y \). Arguably, this change of distribution is the core property of information that is of interest in economic scenarios.

Fixing her observation \( x \) but averaging over \( y \)’s, and working in log space, this change in the likelihood she would assign to the actual \( y \) if she ignored \( x \) (and so used likelihoods \( p(y) \) rather than \( p(y \mid x) \)) is

\[
\sum_{y} p(y \mid x) \ln \left[ \frac{p(y \mid x)}{p(y)} \right].
\]

(2)

Averaging this over possible data \( x \) she might observe gives

\[
\sum_{x,y} p(x) p(y \mid x) \ln \left[ \frac{p(y \mid x)}{p(y)} \right].
\]

(3)

Eq. (3) gives the (average) increase in information that player \( i \) has about \( y \) due to observing \( x \). Note that this is true no matter how the variables \( X \) and \( Y \) arise in the strategic interaction. In particular, this interpretation of the quantity in Eq. (3) does not require that the value \( x \) arise directly through a pre-specified distribution \( p(x \mid y) \). \( x \) and \( y \) could instead be variables concerning the strategies of the players at a particular equilibrium.

In this sense, we have shown that the quantity in Eq. (3) is the proper way to measure the information relating any pair of variables arising in a strategic scenario. None of

\(^4\)This includes information partitions, in the sense that the player is informed about which element of her information partition obtains.
the usual axiomatic arguments motivating Shannon’s information theory (Cover and Thomas, 1991; Mackay, 2003) were used in doing this. However in Sec. 2.1 below we will show that the quantity in Eq. (3) is just the mutual information between $X$ and $Y$, as defined by Shannon.

Note that even once we decide to use the mutual information of a signal to quantify information, we must still make the essentially arbitrary choice of which signal, to which player, concerning which other variable, we are interested in. So for example, we might be interested in the mutual information between some state of Nature and the move of player 1. Or the mutual information between the original move of player 1 and the last move of player 1.

These kinds of mutual informations will be the typical choices of $f$ in our computer experiments presented below. However our analysis will also hold for choices for $f$ that are derived from mutual information, like the “information capacity” described below. Indeed, our general theorems will hold for arbitrary choices of $f$, even those that bear no relation to concepts from Shannon’s information theory.

1.5. Differential geometry’s role in game theory

Recall that in the naive motivation of our approach presented at the end of Sec. 1.2, the value of information depends on our choice of the coordinate system of the game parameters. To avoid this issue we must use inner products, defined in terms of a metric tensor, rather than dot products. Calculations of inner products are guaranteed to be covariant, not changing as we change our choice of coordinate system. For similar reasons we must use the natural gradient rather than the conventional gradient. The metric tensor specifying both quantities also tells us how to measure distance. So it defines a (non-Euclidean) geometry.

Evidently then, very elementary considerations force us to use tensor calculus with an associated metric to analyze the value of information in games. However for many economic questions, there is no clearly preferred distance measure, and no clearly preferred way of defining inner products. For such questions, the precise metric tensor we use should not matter, so long as we use some metric tensor. The analysis below bears this out. In particular, the existence / impossibility theorems we prove below do not depend on which metric tensor we use, only that we use one.

Nonetheless, for making precise calculations the choice of tensor is important. For example, it matters when we evaluate precise (differential) values of information, plot vector fields of gradients of mutual information, etc. To make such calculations in a covariant way we need to specify a precise choice of a metric. We will refer to marginal utility of information when the inner product is defined in terms of such a metric as differential value of information.\footnote{We have chosen to use the term “value” because of well-entrenched convention. The reader should beware that “value” also refers to the output of a function, e.g., in expressions like “the value of $h(x)$ evaluated at $x = 5$”. This can lead to potentially confusing language like “the value of the value of information”.}

In general, there are several choices of metric that can be motivated. In this paper we restrict attention to the Fisher information metric (Amari and Nagaoka, 2000; Cover
and Thomas, 1991), since it is based on information theory, and therefore naturally “matches” the quantities we are interested in. (See Sec. 2.2 for a more detailed discussion of this metric.) However similar calculations could be done using other choices.

1.6. Other benefits of the marginal value approach

This approach of making infinitesimal changes to information channels and examining the ramifications on expected utility is very general and can be applied to any information channel in the game. That means for example that we can add infinitesimal noise to an information channel that models dependencies between different states of nature and examine the resultant change in the expected utility of a player. As another example, we can change the information channel between two of the players in the game, and analyze the implications for the expected utility of a third player in the game.

In fact, the core idea of this approach extends beyond making infinitesimal changes to the noise in a channel. At root, what we are doing is making an infinitesimal change to the parameter vector that specifies the noncooperative game. This differential approach can be applied to other kinds of infinitesimal changes besides those involving noise vectors in communication channels. For example, it can be applied to a change to the utility function of a player in the game. As another example, the changes can be applied to the rationality exponent of a player under a logit quantal response equilibrium (McKelvey and Palfrey, 1998). This flexibility allows us to extend Blackwell’s idea of “value of information” far beyond the scenarios he had in mind, to (differential) value of any defining characteristics of a game. This in turn allows us to calculate marginal rates of substitution of any component of a game’s parameter vector with any other component, e.g., the marginal rate of substitution for player i of (changes to) a specific information channel and of (changes to) a tax applied to player j.

More generally still, there is nothing in our framework that requires us to consider marginal values to a player in the game. So for example, we can apply our analysis to calculate marginal social welfare of (changes to) information channels, etc. Carrying this further, we can use our framework to calculate marginal rates of substitution in noncooperative games to an external regulator concerned with social welfare who is able to change some parameters in the game.

In this context, the need to specify a particular branch of the game is a benefit of the approach, not a necessary evil. To see why, consider how a (toy model of a regulator) concerned with social welfare would set some game parameters, according to conventional economics analysis. The game and associated set of parameter vectors is considered ab initio, and an attempt is made to find the global optimal value of the parameter vector. However whenever the game has multiple equilibrium branches, in general what parameter vector is optimal will depend on which branch one considers — and there is no good generally applicable way of predicting which branch will be appropriate, since that amounts to choosing a universal refinement.

However our framework provides a different way for the regulator to control the parameter vector. The idea is to start with the actual branch that gives an actual, current
player profile for a currently implemented parameter vector $\theta$. We then tell the regulator what direction to incrementally change that parameter vector given that the players are on that branch. No attempt is made to find an ab initio global optimum. So this approach avoids the problem of predicting what branch will arise — we use the one that is actually occurring. Furthermore, the parameters can then be changed along a smooth path leading the players from the current to the desired equilibrium (see (Wolpert, Harre, Olbrich, Bertschinger, and Jost, 2012) for an example of this idea).

1.7. Previous work

In his famous theorem, Blackwell formulated the imperfect information of the decision maker concerning the state of nature as an information channel from the move of Nature to the observation of the decision maker, i.e., as conditional probability distribution, leading from the move of Nature to the observation of the decision maker. This is a very convenient way to model such noise, from a calculational standpoint. As a result, it is the norm for how to formulate imperfect information in Shannon information theory (Cover and Thomas, 1991; Mackay, 2003), which analyses many kinds of information, all formulated as real-valued function of probability distributions. Indeed, use of conditional distributions to model imperfect information is the norm in all of engineering and the physical sciences, e.g., computer science, signal processing, stochastic control, machine learning, physics, stochastic process theory, etc.

There were some early attempts to use Shannon information theory in economics to address the question of the value of information. Except for special cases such as multiplicative payoffs (Kelly gambling (Kelly, 1956)) and logarithmic utilities (Arrow, 1971), where the expected utility will be proportional to the Shannon entropy, the use of Shannon information was considered to provide no additional insights. Indeed, Radner and Stiglitz (1984) rejected the use of any single valued function to measure information because it provides a total order on information and therefore allows for a negative value of information even in the decision case considered by Blackwell.

In multi-player game theory, i.e. multi-agent decision situations, the role of information is even more involved. Here, many researchers have constructed special games, showing that the players might prefer more or less information depending on the particular structure of the game (see (Levine and Ponssard, 1977) for an early example). This work showed that Blackwell’s result cannot directly be generalized to situations of strategic interactions.

Correspondingly, the most common formulation of imperfect information in game theory does not use information channels let alone Shannon information. Instead, states of nature are lumped using information partitions specifying which states are indistinguishable to an agent. In this approach, more (less) information is usually modeled as refining (coarsening) an agent’s information partition. In particular, noisy observations are formulated using such partitions in conjunction with a (common) prior distribution on the states of nature. Even though, this is formally equivalent to conditional distributions, it leads to a fundamentally different way of thinking about information. The formulation of information in terms of information partitions provides a natural partial
order based on refining partitions. Thus, in contrast to Shannon information theory, which quantifies the amount of information, it cannot compare the information whenever the corresponding partitions are not related via refinements. In addition, the avoidance of conditional distributions makes many calculations more difficult.

Recently, some work in game theory has made a distinction between the “basic game” and the “information structure”: The basic game captures the available actions, the payoffs and the probability distribution over the states of nature, while the information structure specifies what the players believe about the game, the state of nature and each other (see for instance (Bergemann and Morris, 2013; Lehrer, Rosenberg, and Shmaya, 2013)). More formally this is expressed in games of incomplete information having each player observing a signal, drawn from a conditional probability distribution, about the state of nature. In principle these signals are correlated. The effects of changes in the information structure were studied by considering certain types of garbling nodes as by Blackwell. While this goes beyond refinements of information partitions, it still only provides a partial order of information channels. Lehrer, Rosenberg, and Shmaya (2013) showed that if two information structures are equivalent with respect to a specific garbling the game will have the same equilibrium outcomes. Thus, they characterized the class of changes to the information channels that leave the players indifferent with respect to a particular solution concept. Similarly, Bergemann and Morris (2013) introduced a Blackwell-like order on information structures called “individual sufficiency” that provides a notion of more and less informative structures, in the sense that more information always shrinks the set of Bayes correlated equilibria. A similar analysis relating the set of equilibria between different information structures has been obtained by Gossner (2000) and is in line with his work (Gossner, 2010) relating more knowledge of the players to an increase of their abilities, i.e. the set of possible actions available to them. As formulated in this work, more information can be seen to increase the number of constraints on a possible solution for the game.

Overall, the goal of these attempts has been to characterize changes to information structures which imply certain properties of the solution set, independent of the particular basic game. This is clearly inspired by Blackwell’s result which holds for all possible decision problems. So in particular, these analyses aim for results that are independent of the details of the utility function(s). Moreover, the analyses are concerned with results that hold simultaneously for all solution points (branches) of a game. Given these constraints on the kinds of results one is interested in, as observed by Radner and Stiglitz, Shannon information (or any other quantification of information) is not of much help.

In contrast, we are concerned with analyses of the role of information in strategic scenarios that concern a particular game with its particular utility functions. Indeed, our analyses focus on a single solution point at a time, since the role of information for the exact same game will differ depending on which solution branch one is on. Arguably, in many scenarios regulators and analysts of a strategic scenario are specifically interested in the actual game being played, and the actual solution point describing the behavior of its players. As such, our particularized analyses can be more relevant than

6 According to Gossner (2000) this terminology goes back to Aumann.
broadly applicable analyses, which ignore such details. While not being much help in the broadly applicable analyses of Bergemann and Morris (2013); Gossner (2000, 2010), etc., we argue below that Shannon information is useful if one wants to analyze the role of information in a particular game with its specific utility functions. In this case, the idea of marginal utility of a good to a decision-maker in a particular game against Nature can be naturally extended “marginal utility” of information to a player in a particular multi-player game on a particular solution branch of that game. Thus, one is naturally lead to a quantitative notion of information and the differential value of information as elaborated above.

1.8. Roadmap

In Sec. 2 we review basic information theory as well as information geometry. In Sec. 3, we review Multi-Agent Influence Diagrams (MAIDs) and explain why they are especially suited to study information in games. Next, we introduce quantal response equilibria of MAIDs and show how to calculate partial derivatives of the associated strategy profile with respect to components of the associated game parameter vector. Based on these definitions, in Sec. 4 we define the differential value of information and in Sec. 5 we prove general conditions for the existence of negative value of information. In particular, the marginal value of information described above is the ratio of the marginal change in expected utility to the marginal change in information, as one makes infinitesimal changes to the channel’s conditional distribution in the direction that maximizes change in information. One can also consider the marginal change in expected utility for other infinitesimal changes to the observation channel conditional distributions. We prove that generically, in all games there is such a direction in which information is decreased. In this sense, we prove that generically, in all games there is (a way to infinitesimally change the channel that has) negative value of information, unless one imposes a priori constraints on how the channel’s conditional distribution can be changed.

This theorem holds for arbitrary games, not just leader-follower games. We establish other theorems that also hold for arbitrary games. In particular we provide necessary and sufficient conditions for a game to have negative value of information simultaneously for all players. (This condition can be viewed as a sort of “Pareto negative value of information”.)

Next, in Sec. 6 we illustrate our proposed definitions and results in a simple decision situation as well as an abstracted version of the duopoly scenario that was discussed above, in which the second-moving player observes the first-moving player through a noisy channel. In particular, we show that as one varies the noise in that channel, the marginal value of information is indeed sometimes negative for the second-moving player, for certain starting conditional noise distributions in the channel (and at a particular equilibrium). However for other starting distributions in that channel (at the same equilibrium), the marginal value of information is positive for that player. In fact, all four pairs of [positive / negative] marginal value of information for the [first / second] – moving player can occur.
Information theory

- **X, Y** Sets
- **x, y** Elements of sets, i.e. \( x \in X \)
- **X, Y** Random variables with outcomes in \( X, Y \)
- **\Delta_X** Probability simplex over \( X \)
- **I(X; Y)** Mutual information between \( X \) and \( Y \)

Differential geometry

- **v, \theta** Vectors
- **v_i** \( i \)-th entry of contra-variant vector
- **v_j** \( i \)-th entry of co-variant vector
- **g_{ij}** Metric tensor. Its inverse is denoted by \( g^{ij} \)
- **\frac{\partial}{\partial \theta} f** Partial derivative wrt \( \theta \)
- **\nabla f** Gradient of \( f \)
- **\nabla^2 f** Hessian of \( f \)
- **\langle v, w \rangle_g** Scalar product of \( v, w \) wrt metric \( g \)
- **|v|_g** Norm of vector \( v \) wrt metric \( g \)

Multi-agent influence diagrams

- **G = (V, E)** Directed acyclic graph with vertices \( V \) and edges \( E \subset V \times V \)
- **X_v** State space of node \( v \in V \)
- **N** Set of nature or change nodes, i.e. \( N \subset V \)
- **D_i** Set of decision nodes of player \( i \)
- **pa(v)** Parents of node \( v \)
- **p(x_v | x_{pa(v)})** Conditional distribution at nature node \( v \in N \)
- **\sigma_i(a_v | x_{pa(v)})** Strategy of player \( i \) at decision node \( v \in D_i \)
- **u_i** Utility function of player \( i \)
- **E(u_i | a_i)** Conditional expected utility of player \( i \)
- **V_i = E(u_i)** Value, i.e. expected utility, of player \( i \)

Differential value of information

- **V_{f, \theta}** Differential value of direction \( \theta \)
- **V_{f, \theta}** Differential value of \( f \) in direction \( \theta \)
- **V_f** Differential value of \( f \)
- **Con(\{v_i\})** Conic hull of nonzero vectors \( \{v_i\} \)
- **Con(\{v_i\})** Conic hull of \( \{v_i\} \)

| TABLE I | Summary of notation used throughout the paper. |

After this we present a section giving more examples. We end with a discussion of future work and conclusions.

A summary of the notation we use is provided in Table. I.

2. REVIEW OF INFORMATION THEORY AND GEOMETRY

As a prerequisite for our analysis of game theory, in this section we review some basic aspects of information theory and information geometry. In doing this we illustrate additional advantages to using terms from Shannon information theory to quantify information for game-theoretic scenarios. We also show how Shannon information theory gives rise to a natural metric on the space of probability distributions.

The following section will start by reviewing salient aspects of game theory, laying the formal foundation for our analysis of differential value of information.
2.1. Review of Information theory

We will use notation that is a combination of standard game theory notation (Fudenberg and Tirole, 1991) and standard Bayes net notation (Koller and Friedman, 2009). (See (Koller and Milch, 2003) for a good review of Bayes nets for game theoreticians.)

The probability simplex over a space $X$ is written as $\Delta_X$. $\Delta_{X|Y}$ is the space of all possible conditional distributions of $x \in X$ conditioned on a value $y \in Y$. For ease of exposition, this notation is adopted even if $X \cap Y \neq \emptyset$. We use uppercase letters $X, Y$ to indicate random variables with the corresponding domains written as $X, Y$. We use lowercase letters to indicate a particular element of the associated random variable’s range, i.e., a particular value of that random variable. In particular, $p(X) \in \Delta_X$ always means an entire probability distribution vector over all $x \in X$, whereas $p(x)$ will typically refer instead to the value of $p(.)$ at the particular argument $x$. Here, we couch the discussion in terms of countable spaces, but much of the discussion carries over to the uncountable case.

Information theory provides a way to quantify the difference between two distributions, as Kullback-Leibler (KL) divergence (Cover and Thomas, 1991). This measure of the difference between probability distributions has now become a standard across statistics and many other fields:

**Definition 1** Let $p, q \in \Delta_X$. The Kullback-Leibler divergence between $p$ and $q$ is defined as

$$D_{KL}(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

The KL-divergence is non-negative and vanishes if and only if $p \equiv q$. Since the KL-divergence is not symmetric, it does not form a metric.

To quantify the information of a signal $X$ about $Y$, Shannon defined the mutual information between $X$ and $Y$ as the average (over $p(X)$) KL-divergence between $p(Y | x)$ and $p(Y)$:

**Definition 2** The **mutual information** between two random variables $X$ and $Y$ is defined as:

$$I(X; Y) = \mathbb{E}_{p(X)} [D_{KL}(p(Y | x) || p(Y))] = \sum_{x,y \in X \times Y} p(x, y) \log \frac{p(y | x)}{p(x)}$$

where the logarithm to base two is commonly chosen. In this case, the information has units of **bits**.

The mutual information, together with the related quantity of entropy, forms the basis of information theory. It not only allows us to quantify information, but has many applications in different areas ranging from coding theory to machine learning to evolutionary biology. Moreover, as we showed in deriving Eq. (3), arguably it provides the proper way to quantify information in game theory.
Here, we only mention some properties of mutual information which are directly relevant to our analysis of the value of information. First, note that $I(X; Y)$ can also be written as $I(X; Y) = \sum_{x, y \in X \times Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$. Thus, it quantifies the divergence between the joint distribution $p(X, Y)$ and the product of the corresponding marginals $p(X)p(Y)$. From this perspective, mutual information can be seen as a general measure of statistical dependency, i.e. a sort of non-linear correlation, and it vanishes if and only if $X$ and $Y$ are independent.

Another important property of mutual information is the following:

**Proposition 1** **Data-processing inequality**: Let $X \rightarrow Y \rightarrow Z$ form a Markov chain, i.e., $p(x, y, z) = p(x)p(y | x)p(z | y)$. Then,

$$I(X; Y) \geq I(X; Z)$$

(Typically we refer to the distributions taking $X \rightarrow Y$ and then taking $Y \rightarrow Z$ as (information) channels.)

The data-processing inequality applies in particular if the channel $p(z | y)$ from $Y$ to $Z$ is a deterministic mapping $f : Y \rightarrow Z$, i.e. $p(z | y) = 1$ if $z = f(y)$ and 0 otherwise. Thus processing $Y$ via some transformation $f$ can never increase the amount of information we have about $X$. (This is the basis for the term “data-processing inequality”).

An information partition $\{A_1, \ldots, A_n\}$ can be viewed as a random variable with values $x \in \{1, \ldots, n\}$, i.e. the signal $x$ reveals which element of the partition was hit. Coarsening that partition can then be viewed as a deterministic map from $x \in X$ to a value $y \in Y$ in the coarser partition. Now when we want to evaluate how much information the agent obtains from the coarser partition $Y$ about some other random variable $N$, corresponding to a state of nature, we see that $N \rightarrow X \rightarrow Y$ is a Markov chain. Thus, the data-processing inequality applies and the mutual information between $N$ and $Y$ cannot exceed the mutual information between $N$ and $X$. So by using mutual information, we can not only state that the amount of information is reduced when an information partition is coarsened, but also quantify by how much.

As another example of the use of the data-processing inequality, in Blackwell’s analysis a channel $p(y | x)$ is said to be “more informative” than a channel $p(z | x)$ if there exists some channel $q(z | y)$ such that $p(z | x) = \sum_{y \in Y} p(y | x)q(z | y)$. Since in this case $X \rightarrow Y \rightarrow Z$ forms a Markov chain, the data-processing inequality can again be applied to prove that $I(X; Y) \geq I(X; Z)$. So again, we can use mutual information to go beyond the partial orders of “amounts of information” considered in earlier analyses, to provide a cardinal value that agrees with those partial orders.

Given the evident importance of mutual information, it is natural to make the following definition:

---

7Importantly, there is not an analog of this result if we quantify the information in one random variable concerning another random variable with their statistical covariance rather than with their mutual information. For some scenarios, post-processing a variable $Y$ can increase its covariance with $X$. (See (Wolpert and Leslie, 2012).)
The channel capacity $C$ of an information channel $p(y \mid x)$ from $X$ to $Y$ is defined as

$$C = \max_{p(X)} I(X; Y)$$

The data processing inequality shows that chaining information channels can never increase the capacity.\(^8\)

Unfortunately, in general we cannot solve the maximization problem defining information capacity analytically. So closed formulas for the channel capacity are only known for special cases. This in turn means that partial derivatives of the channel capacity with respect to the channel parameters are difficult to calculate in general.

One special case where one can make that calculation is the binary (asymmetric) channel (Amblard, Michel, and Morfù, 2005). For this reason, we will use that channel in the examples considered in this paper that involve marginal value of information capacity.\(^9\)

2.2. Information geometry

Consider a distribution over a space of values $x$ which is parametrized with $d$ parameters $\theta = \theta^1, \ldots, \theta^d$ living in a $d$-dimensional differentiable manifold $\Theta$. Write this distribution as $p(x; \theta)$. We will be interested in differentiable geometry over the manifold $\Theta$. Here we use the the convention of differential geometry to denote components of contra-variant vectors living in $\Theta$ by upper indices and components of co-variant vectors by lower indices (see appendix 9 for details).

In general, expected utilities and information quantities depend on the $d$ parameters specifying a position on the manifold. This dependence can be direct, e.g., as with the information capacity of a channel with certain noise parameters is directly given by position on $\Theta$. Alternatively, the dependence may be indirect, e.g., as with the expected utilities of the players who adjust their strategies to match changes in the position on $\Theta$.

Here we will assume that all such functions of interest are differentiable functions of $\theta$ in the interior of $\Theta$. This allows us to evaluate the partial derivatives $\frac{\partial}{\partial \theta}$ of the functions of interest with respect to the parameters specifying the game. As discussed above, in order to obtain results that are independent of the chosen parametrization, we need a metric on the space $d$ parameters. Given that $\theta$ parameterizes a probability distribution, a suitable choice for us is the Fisher information metric. This is given by

$$g_{ij}(\theta) = \sum_x p(x; \theta) \frac{\partial \log p(x; \theta)}{\partial \theta^i} \frac{\partial \log p(x; \theta)}{\partial \theta^j}$$

\(^8\)Fix $P(y \mid x)$ and $P(z \mid y)$. The data processing inequality holds for any distribution $p(X)$ and thus in particular it holds for the distribution $q(X)$ that achieves the maximum of $I(X; Z)$. So $C_{X \rightarrow Y} \geq I_{q}(X; Y) \geq I_{q}(X; Z) = C_{X \rightarrow Z}$.

\(^9\)Another important class of information channels with known capacity are the so called symmetric channels (Cover and Thomas, 1991). In this case, the noise is symmetric in the sense that it does not depend on a particular input, i.e. the channel is invariant under relabeling of the inputs. This class is rather common in practice and includes channels with continuous input, e.g. the Gaussian channel.
where \( p(x; \theta) \) is a probability distribution parametrized by \( \theta \).

The statistical origin of the Fisher metric lies in the task of estimating a probability distribution from a family parametrized by \( \theta \) from observations of the variable \( x \). The Fisher metric expresses the sensitivity of the dependence of the family on \( \theta \), that is, how well observations of \( x \) can discriminate among nearby values of \( \theta \).

With this metric, and using the Einstein summation convention (see appendix 9 again), we can form the scalar product of two (contravariant) tangent vectors \( v = (v^1, \ldots, v^d) \), \( w = (w^1, \ldots, w^d) \) as

\[
(v, w)_g = \sum_{i,j} g_{ij} v^i w^j
\]

The norm of a vector \( v \) is then given as

\[
\|v\|_g = \sqrt{(v, v)_g} = \sqrt{\sum_{i,j} g_{ij} v^i v^j}
\]

The gradient of any functional \( f : \Delta_X(\theta) \to \mathbb{R} \) can then be obtained from the partial derivatives as follows:

\[
\text{grad}(f)_i = g^{ij} \frac{\partial f}{\partial \theta^j}
\]

where \( g^{ij} \) denotes the inverse of the metric \( g_{ij} \) and we have again used Einstein summation for the index \( j \). Thus, the gradient is a contra-variant vector, whose \( d \) components are written as \( (\text{grad}(f)_1, \ldots, \text{grad}(f)_d) \).

As an example, consider a binary asymmetric channel \( p(s|x; \theta) \) with input distribution

\[
p(x) = \begin{cases} 
q & \text{if } x = 0 \\
1 - q & \text{if } x = 1
\end{cases}
\]

and parameters \( \theta = (\epsilon^1, \epsilon^2) \) for transmission errors

\[
p(s|x; \theta) = \begin{cases} 
1 - \epsilon^1 & \text{if } x = 0, s = 0 \\
\epsilon^1 & \text{if } x = 0, s = 1 \\
\epsilon^2 & \text{if } x = 1, s = 0 \\
1 - \epsilon^2 & \text{if } x = 1, s = 1
\end{cases}
\]

In this setup, the Fisher information metric of \( p(x, s; \epsilon^1, \epsilon^2) \) is a \( 2 \times 2 \) matrix with entries

\[
g(\epsilon^1, \epsilon^2) = \begin{pmatrix}
\frac{q}{\epsilon^1(1-\epsilon^1)} & 0 \\
0 & \frac{1-q}{\epsilon^2(1-\epsilon^2)}
\end{pmatrix}
\]

The cross-terms vanish since \( \epsilon^1 \) and \( \epsilon^2 \) parameterize different aspects of the channel. Thus, the sensitivity to changes in \( \epsilon^1 \) does not depend on \( \epsilon^2 \) and vice-versa.

\section{3. Multi-Agent Influence Diagrams}

Bayes nets (Koller and Friedman, 2009) provide a very concise, powerful way to model scenarios where there are multiple interacting Nature players (either automata or inanimate natural phenomena), but no human players. They do this by representing the information structure of the scenario in terms of a Directed Acyclic Graph (DAG) with conditional probability distributions at the nodes of the graph. In particular, the use of conditional distributions rather than information partitions greatly facilitates the
analysis and associated computation of the role of information in such systems. As a result they have become very wide-spread in machine learning and information theory in particular, and in computer science and the physical sciences more generally.

Influence Diagrams (IDs (Howard and Matheson, 2005)) were introduced to extend Bayes nets to model scenarios where there is a (single) human player interacting with Nature players. There has been much analysis of how to exploit the graphical structure of the ID to speed up computation of the optimal behavior assuming full rationality, which is quite useful for computer experiments.

More recently, Multi-Agent Influence Diagrams (MAIDs (Koller and Milch, 2003)) and their variants like semi-net-form games (Backhaus, Bent, Bono, Lee, B., D.H., and Xie, in press; Lee, Wolpert, Backhaus, Bent, Bono, and B., 2013; Lee, Wolpert, Backhaus, Bent, Bono, and Tracey, 2012) and Interactive POMDP’s (Doshi, Zeng, and Chen, 2009) have extended IDs to model games involving arbitrary numbers of players. As such, the work on MAIDs can be viewed as an attempt to create a new game theory representation of multi-stage games based on Bayes nets, in addition to strategic form and extensive form representations.

Compared to these older representations, typically MAIDs more clearly express the interaction structure of what information is available to each player in each possible state.\footnote{In a MAID a player has information at a decision node $A$ about some state of nature $X$ if there is a directed edge from $X$ to $A$.} They also very often require far less notation than those other representations to fully specify a given game. Thus, we consider them as a natural starting point when studying the role of information in games.

A MAID is defined as follows:

**Definition 4** An $n$-player MAID is defined as a tuple $(G, \{X_v\}, \{p(x_v \mid x_{pa(v)})\}, \{u_i\})$ of the following elements:

- A directed acyclic graph $G = (V, E)$ where $V = D \cup N$ is partitioned into
  - a set of nature or chance nodes $N$ and
  - a set of decision nodes $D$ which is further partitioned into $n$ sets of decision nodes $D_i$, one for each player $i = 1, \ldots, n$.
- A set $X_v$ of states for each $v \in V$.
- A conditional probability distribution $p(x_v \mid x_{pa(v)})$ for each nature node $v \in N$, where $pa(v) = \{u : (u, v) \in E\}$ denotes the parents of $v$ and $x_{pa(v)}$ is their joint state.
- A family of utility functions $\{u_i : \prod_{v \in V} X_v \rightarrow \mathbb{R}\}_{i=1,\ldots,n}$.

In particular, as mentioned above, a one-person MAID is an influence diagram (ID (Howard and Matheson, 2005)).
elements of $X$ as $x$. We define $X_{\mathcal{A}} \equiv \prod_{v \in \mathcal{A}} X_v$ for any $\mathcal{A} \subseteq V$, with elements of $X_{\mathcal{A}}$ written as $x_{\mathcal{A}}$. So in particular, $X_D \equiv \prod_{v \in D} X_v$, and $X_N \equiv \prod_{v \in N} X_v$, and we write elements of these sets as $x_D$ (or $a_D$) and $x_N$, respectively.

We will sometimes write an $n$-player MAID as $(G, X, p, \{u_i\})$, with the decompositions of those variables and associations among them implicit. (So for example the decomposition of $G$ in terms of $E$ and a set of nodes $[\bigcup_{i=1,...,n} D_i] \cup N$ will sometimes be implicit.)

A solution concept is a map from any MAID $(P, G, X, p, \{u_i\})$ to a set of conditional distributions $\{\sigma_i(x_v \mid x_{pa(v)}) : v \in D_i, i = 1, \ldots, n\}$. We refer to the set of distributions $\{\sigma_i(x_v \mid x_{pa(v)}) : v \in D_i\}$ for any particular player $i$ as that player’s strategy. We refer to the full set $\{\sigma_i(x_v \mid x_{pa(v)}) : v \in D_i, i = 1, \ldots, n\}$ as the strategy profile. We sometimes write $\sigma_v$ for a $v \in D_i$ to refer to one distribution in a player’s strategy and use $\sigma$ to refer to a strategy profile.

The intuition is that each player can set the conditional distribution at each of their decision nodes, but is not able to introduce arbitrary dependencies between actions at different decision nodes. In the terminology of game theory, this is called the agent representation. The rule for how the set of all players jointly set the strategy profile is the solution concept.

In addition, we allow the solution concept to depend on parameters. Typically there will be one set of parameters associated with each player. When that is the case we sometimes write the strategy of each player $i$ that is produced by the solution concept as $\sigma_i(a_v \mid x_{pa(v)}, \beta)$ where $\beta$ is the set of parameters that specify how $\sigma_i$ was determined via the solution concept.

The combination of a MAID $(G, X, p, \{u_i\})$ and a solution concept specifies the conditional distributions at all the nodes of the DAG $G$. Accordingly it specifies a joint probability distribution

\begin{equation}
\sum_{v \in N} p(x_v) = \prod_{v \in N} p(x_v \mid x_{pa(v)}) \prod_{i=1, \ldots, n} \prod_{v \in D_i} \sigma_i(a_v \mid x_{pa(v)})
\end{equation}

\begin{equation}
\prod_{v \in V} p(x_v \mid x_{pa(v)})
\end{equation}

where we abuse notation and denote $\sigma_i(a_v \mid x_{pa(v)})$ by $p(x_v \mid x_{pa(v)})$ whenever $v \in D_i$.

In the usual way, once we have such a joint distribution over all variables, we have fully defined the joint distribution over $X$ and therefore defined conditional probabilities of the states of one subset of the nodes in the MAID, $\mathcal{A}$, given the states of another subset of the nodes, $\mathcal{B}$:

\begin{equation}
p(x_{\mathcal{A}} \mid x_{\mathcal{B}}) = \frac{p(x_{\mathcal{A} \cup \mathcal{B}})}{p(x_{\mathcal{B}})} = \frac{\sum_{x_{\mathcal{A} \cup \mathcal{B}}} p(x_{\mathcal{A} \cup \mathcal{B}}, x_{\mathcal{B} \setminus \mathcal{A} \cup \mathcal{B}})}{\sum_{x_{\mathcal{B}}} p(x_{\mathcal{B}}, x_{\mathcal{B} \setminus \mathcal{A} \cup \mathcal{B}})}
\end{equation}

Similarly the combination of a MAID and a solution concept fully defines the conditional value of a scalar-valued function of all variables in the MAID, given the values of some other variables in the MAID. In particular, the conditional expected utilities are
given by

\[ \mathbb{E}(u_i \mid x_{\mathcal{R}}) = \sum_{x_{\mathcal{V}\setminus \mathcal{R}}} p(x_{\mathcal{V}\setminus \mathcal{R}} \mid x_{\mathcal{R}}) u_i(x_{\mathcal{V}\setminus \mathcal{R}}, x_{\mathcal{R}}) \]

We will sometimes use the term “information structure” to refer to the graph of a MAID and the conditional distributions at its Nature nodes. (Note that this is a slightly different use of the term from that used in extensive form games.) In order to study the effect of changes to the information structure of a MAID, we will assume that the probability distributions at the nature nodes are parametrized by a set of parameters \( \theta \), i.e., \( p_v(x_v \mid x_{\text{pa}(v)}; \theta) \). We are interested in how infinitesimal changes to \( \theta \) (and other parameters of the MAID like \( \beta \)) affect \( p(x_V) \), expected utilities, mutual information among nodes in the MAID, etc.

### 3.1. Quantal response equilibria of MAIDs

A solution concept for a game specifies how the actions of the players are chosen. In our framework, it is not crucial which solution concept is used (so long as the strategy profile of the players at any \( \theta \) is differentiable in the interior of \( \Theta \)). For convenience, we choose the (logit) quantal response equilibrium (QRE) (McKelvey and Palfrey, 1998), a popular model for bounded rationality.\(^{11}\) Under a QRE, each player \( i \) does not necessarily make the best possible move, but instead chooses his actions at the decision node \( v \in D_i \) from a Boltzmann distribution over his move-conditional expected utilities:

\[ \sigma_i(a_v \mid x_{\text{pa}(v)}) = \frac{1}{Z_i(x_{\text{pa}(v)})} e^{\beta_i \mathbb{E}(u_i \mid a_v, x_{\text{pa}(v)})} \]

for all \( a_v \in X_v \) and \( x_{\text{pa}(v)} \in \prod_{u \in \text{pa}(v)} X_u \). In this expression \( Z_i(x_{\text{pa}(v)}) = \sum_{a \in X_v} e^{\beta_i \mathbb{E}(u_i \mid a, x_{\text{pa}(v)})} \) is a normalization constant, \( \mathbb{E}(u_i \mid a_v, x_{\text{pa}(v)}) \) denotes the conditional expected utility as defined in eq. (10) and \( \beta_i \) is a parameter specifying the “rationality” of player \( i \).

This interpretation is based on the observation that a player with \( \beta = 0 \) will choose her actions uniformly at random, whereas \( \beta \to \infty \) will choose the action(s) with highest expected utility, i.e., corresponds to the rational action choice. Thus, it includes the Nash equilibrium where each player maximizes expected utility as a boundary case.

As shorthand, we denote the (unconditional) expected utility of player \( i \) at some equilibrium \( \{\sigma_i\}_{i=1,...,n} \), \( \mathbb{E}(\{\sigma_i\}_{i=1,...,n}(u_i)) \), by \( V_i \).

### 3.2. Partial derivatives of QREs of MAIDs with respect to game parameters

Our definition of differential value of information depends on the partial derivatives of the strategy profile of the players with respect to parameters of the underlying game. As noted above though, in general there can be multiple equilibria for a given parameter vector, i.e., multiple strategy profiles \( \{\sigma_i\}_{i=1,...,n} \) that simultaneously solves eq. (11) for...
all players. In such a case we have to choose a particular equilibrium branch at which to calculate partial derivatives. Loosely speaking, depending on the equilibrium branch chosen, not only the strategies of the players but also their partial derivatives will be different. This means that players will value changes to the parameters of the game differently depending on which different equilibrium branch they are on. This is just as true for a QRE equilibrium concept as any another. Thus, in the following we implicitly assume that we have chosen an equilibrium branch on which we want to investigate the value of information.

For computations involving the partial derivatives of the players strategies at a QRE (branch) it can help to explicitly introduce the normalization constants as an auxiliary variable. The QRE condition from eq. (11) is then replaced by the following conditions

\[ \sigma_i(a_v | x_{p(v)}; \beta_i, \theta) - \frac{e^{\beta_i u_i(a_v, x_{p(v)}; \beta, \theta)}}{Z_i(x_{p(v)}; \beta_i, \theta)} = 0 \]

\[ Z_i(x_{p(v)}; \beta_i, \theta) - \sum_{v \in X_v} e^{\beta_i u_i(a_v, x_{p(v)}; \beta, \theta)} = 0 \]

for all players \( i \), decision nodes \( v \in D_i \) and all states \( a_v \in X_v, x_v \in \prod_{a \in P(a_v)} X_a \). (Here and throughout this section, subscripts on \( \sigma, Z, \) etc. should not be understood as specifications of coordinates as in the Einstein summation convention.)

Overall, this gives rise to a total of \( M \) equations for \( M \) unknown quantities \( \sigma_i(a_v | x_{p(v)}), Z_i(x_{p(v)}) \).

Using a vector valued function \( f \) we can abbreviate the above by the following equation:

\[ f(\sigma_{\beta, \theta}, Z_{\beta, \theta}, \beta, \theta) = 0 \]

where \( \sigma_{\beta, \theta} \) is a vector of all strategies

\[ \{ \sigma_i(a_v | x_v; \beta_i, \theta) : i = 1, \ldots, n, v \in D_i, a_v \in X_v, x_v \in \prod_{a \in P(a_v)} X_a \} \]

\( Z_{\beta, \theta} \) collects all normalization constants, and \( \theta \) is the \( M \)-dimensional vector of all 0’s. Note that in general, even once the distributions at all decision nodes have been fixed, the distributions at chance nodes affect the value of \( \mathbb{E}(u_i | a_v, x_{p(v)}; \beta, \theta) \). Therefore they affect the value of the function \( f \). This is why \( f \) can depend explicitly on \( \theta \), as well as depend directly on \( \beta \).

The (vector-valued) partial derivative of the position of the QRE in \( (\sigma_{\theta}, Z_{\theta}) \) with respect to \( \theta \) is then given by implicit differentiation of eq. (12):

\[ \left[ \frac{\partial \sigma_x}{\partial \theta} \right] = - \left[ \frac{\partial f}{\partial \sigma_x} \frac{\partial f}{\partial Z_x} \right]^{-1} \frac{\partial f}{\partial \theta} \]

where the dependence on \( \beta \) is hidden for clarity, all partial derivatives are evaluated at the QRE, and we assume that the matrix \( \left[ \frac{\partial f}{\partial \sigma_x} \frac{\partial f}{\partial Z_x} \right] \) is invertible at the point \( \theta \) at which we are evaluating the partial derivatives.

These equations give the partial derivatives of the mixed strategy profile. They apply to any MAID, and allow us to write the partial derivatives of other quantities of interest.
In particular, the partial derivative of the expected utility of any player $i$ is

$$
\frac{\partial V_i}{\partial \theta} = \sum_{x \in X} u_i(x) \sum_{v \in V} \frac{\partial p(x_v | x_{pa(v)}; \theta)}{\partial \theta} \prod_{v' \neq v} p(x_{v'} | x_{pa(v')}; \theta)
$$

where each term $\frac{\partial p(x_v | x_{pa(v)}; \theta)}{\partial \theta}$ is given by the appropriate component of Eq. (13) if $v$ is a decision node. (For the other, chance nodes, $\frac{\partial p(x_v | x_{pa(v)}; \theta)}{\partial \theta}$ can be calculated directly). Similarly, the partial derivatives of other functions of interest such as mutual informations between certain nodes of the MAID can be calculated from Eq. (13).

Evaluating those derivatives and the additional ones needed for the Fisher metric by hand can be very tedious, even for small games. Here, we used automatic differentiation (Pearlmutter and Siskind, 2008) to obtain numerical results for certain parameter settings and equilibrium branches. Note that automatic differentiation is not a numerical approximation, like finite differences or the adjoint method. Rather it uses the chain rule to evaluate the derivative alongside the value of the function.

4. INFORMATION GEOMETRY OF MAIDS

4.1. General Considerations

As explained above, to obtain results that are independent of a particular parametrization, we need to work with gradients instead of partial derivatives, and therefore need to specify a metric. Throughout our analysis we assume that any such space of parameters of a game is considered under a coordinate system such that the associated metric is full rank and in fact Riemannian\(^ {12} \). The analysis here will not depend on that choice of metric, but as discussed above, for concreteness we can assume the Fisher metric on $p(x_V; \theta, \beta)$. With this choice, our analysis reflects how sensitively the equilibrium distribution of the variables in the game depends on the parameters of the game.

We now define several ways within the context of this geometric structure to quantify the differential value of parameter changes in arbitrary directions in $\Theta$, as well as the more particular case of differential value of some function $f$. Furthermore, we state general results (that are independent of the metric) about negative values and illustrate the possible results with several examples.

4.2. Types of differential value

Say that we fix all distributions at nature nodes in a MAID except for some particular Nature-specified information channel $p(x_v | x_{pa(v)})$, and are interested in the differential value of mutual information through that channel. In general, the expected utility of a player $i$ in this MAID is not a single-valued function of the mutual information in that channel $I(X_i; X_{pat(i)})$. There are two reasons for this. First, the same value of $I(X_i; X_{pat(i)})$ can occur for different conditional distributions $p(x_v | x_{pat(v)})$, and therefore that value

\(^ {12} \)This means that the parameters $\theta^j$ are non-redundant in the sense that the family of probability distributions parametrized by $(\theta^1, \ldots, \theta^d)$ is locally a non-singular $d$-dimensional manifold.
of \( I(X_v; X_{\text{path}}) \) can correspond to multiple values of expected utility in general. Second, as discussed above, even if we fix the distribution \( p(x_v | x_{\text{path}}) \), there might be several equilibria (strategy profiles) all of which solve the QRE equations but correspond to different distributions at the decision nodes of the MAID.

Evidently then, if \( v \) is a chance node in a MAID and \( i \) a player in that MAID, there is no unambiguously defined “differential value to \( i \) of the mutual information” in the channel from \( \text{path}(v) \) to \( v \). We can only talk about differential value of mutual information at a particular joint distribution of the MAID, a distribution that both specifies a particular equilibrium of player strategies on one particular equilibrium branch, and that specifies one particular channel distribution \( p(x_v | x_{\text{path}}) \). Once we make such a specification, we can analyze several aspects of the associated value of mutual information.

A central concept in our analysis will be a formalization of the “alignment” between changes in expected utility and changes in mutual information (or some other function \( f(\theta) \)) at a particular \( \theta \) and an associated branch. (Recall the discussion in the introduction.) There are several ways to quantify such alignment. Here we focus on quantifications involving vector norms and the scalar product \( \frac{\partial}{\partial \theta} V \) and \( \frac{\partial}{\partial \theta} I(X; S) \), i.e. the mutual information between certain nodes \( X, S \) of the MAID. As mentioned, for such norms and inner products to be independent of the parametrization of \( \theta \) that we use to calculate them, we must evaluate them under a metric, and here we choose the Fisher information metric. More precisely, we will quantify the alignment using the inner product

\[
\langle \text{grad}(V), \text{grad}(I(X; S)) \rangle \equiv \frac{\partial}{\partial \theta^k} V g(\theta)^{kl} \frac{\partial}{\partial \theta^l} I(X; S)
\]

where as always \( V \) is the expected utility of a particular player (whose index \( i \) is dropped for brevity), \( g^{kl}(\theta) \) denotes the inverse of the Fisher information matrix \( g_{kl}(\theta) \) as defined in eq. (4), and for consistency with the rest of our analysis, we also choose the contravariant vector norm \( |\delta \theta| \equiv \sqrt{\sum g_{kl} \delta \theta^l \delta \theta^l} \) and similarly for covariant vectors.

This inner product involves changes to \( \theta \) along the gradient of mutual information. To see how it can be used to quantify “value of information”, we first consider a more general inner product, namely the differential value of making an infinitesimal change along an arbitrary direction in parameter space:

**Definition 5** Let \( \delta \theta \in \mathbb{R}^d \) be a contravariant vector. The (differential) value of direction \( \delta \theta \) at \( \theta \) is defined as

\[
\mathcal{V}_{\delta \theta}(\theta) \equiv \frac{\langle \text{grad}(V), \delta \theta \rangle}{| \delta \theta |}
\]

This is the length of the projection of \( \text{grad}(V) \) in the unit direction \( \delta \theta \). Intuitively, the direction \( \delta \theta \) is valuable to the player to the extent that \( V \) increases in this direction. This is what the value of direction \( \delta \theta \) quantifies. (Note that when \( V \) decreases in this direction, the value is negative.)

In general, a mixed-index metric like \( g(\theta)^{kl} \) must be the Kronecker delta function...
(regardless of the choice of metric $g$). Therefore we can expand

\[
\mathcal{V}_{\delta\theta}(\theta) = \frac{\frac{\partial}{\partial \theta} V g(\theta)^{ij} g(\theta)_{ij} \delta \theta}{\sqrt{\delta \theta^a g(\theta)_{ab} \delta \theta^b}} = \frac{\frac{\partial}{\partial \theta} V \delta^a}{\sqrt{\delta \theta^a g(\theta)_{ab} \delta \theta^b}}
\]

The absence of the metric in the numerator in Eq. (15) reflects the fact that the vector of partial derivatives $\frac{\partial}{\partial \theta} V$ is a covariant vector, whereas $\delta \theta$ is a contravariant vector.

As discussed above and elaborated below, one important class of directions $\delta \theta$ at a given game vector $\theta$ are gradients of functions $f(\theta)$ evaluated at $\theta$, e.g., the direction $\frac{\partial}{\partial \theta} I(X; S)$. However even when the direction $\delta \theta$ we are considering is not parallel to the gradient of an information-theoretic function $f(\theta)$ like mutual information, capacity or player rationality, we will often be concerned with quantifying the “value” of such a $f$ in that direction $\delta \theta$. We can do this with the following definition, related to the definition of differential value of a direction.

**Definition 6** Let $\delta \theta \in \mathbb{R}^d$ be a contravariant vector. The (differential) value of $f$ in direction $\delta \theta$ at $\theta$ is defined as:

\[
\mathcal{V}_{f,\delta\theta} = \frac{\langle \text{grad}(V), \delta \theta \rangle}{\langle \text{grad}(f), \delta \theta \rangle} = \frac{\langle \text{grad}(V), \delta \theta \rangle}{\langle \text{grad}(f), \delta \theta \rangle}
\]

This quantity considers the relation between how $V$ and $f$ change when moving in the direction $\delta \theta$. If the sign of the differential value of $f$ in direction $\delta \theta$ at $\theta$ is positive, then an infinitesimal step in in direction $\delta \theta$ at $\theta$ will either increase both $V$ and $f$ or decrease both of them. If instead the sign is negative, then such a step will have opposite effects on $V$ and $f$. The size of the differential value of $f$ in direction $\delta \theta$ at $\theta$ gives the rate of change in $V$ per unit of $f$, for movement in that direction. Note that $\mathcal{V}_{f,\delta\theta}$ is independent of the metric because both numerator and denominator are.

Given the foregoing, a natural way to quantify the “value of $f$” without specifying an arbitrary direction $\delta \theta$ is to consider how $V$ changes when stepping in the direction of $\text{grad}(f)$, i.e. the direction corresponding to the steepest increase in $f$. This is captured by the following definition:

**Definition 7** The (differential) value of $f$ at $\theta$ is defined as:

\[
\mathcal{V}_f(\theta) = \frac{\langle \text{grad}(V), \text{grad}(f) \rangle}{\langle \text{grad}(f), \text{grad}(f) \rangle} = \frac{\langle \text{grad}(V), \text{grad}(f) \rangle}{\langle \text{grad}(f), \text{grad}(f) \rangle} = \frac{\langle \text{grad}(V), \text{grad}(f) \rangle}{\langle \text{grad}(f), \text{grad}(f) \rangle} = \frac{\langle \text{grad}(V), \text{grad}(f) \rangle}{\langle \text{grad}(f), \text{grad}(f) \rangle}
\]

In contrast to $\mathcal{V}_{f,\delta\theta}$, the value of $f$, $\mathcal{V}_f$, does depend on the metric. Formally, this is due to the fact that gradients are contravariant vectors:

\[
\mathcal{V}_f(\theta) = \frac{\langle \text{grad}(V), \text{grad}(f) \rangle}{\langle \text{grad}(f), \text{grad}(f) \rangle} = \frac{\frac{\partial V}{\partial \theta} g^{ik} g_{ij} \frac{\partial f}{\partial \theta}}{\frac{\partial f}{\partial \theta} g^{ik} g_{ij} \frac{\partial f}{\partial \theta}} = \frac{\frac{\partial V}{\partial \theta} g^{ik} g_{ij} \frac{\partial f}{\partial \theta}}{\frac{\partial f}{\partial \theta} g^{ik} g_{ij} \frac{\partial f}{\partial \theta}}
\]

where we have used the fact that $g^{ij}$ is the inverse of $g_{ij}$.

Less formally, differential value of information at $\theta$ measures how much $V$ changes as we move along the direction of fastest growth of $f$ starting from $\theta$. That “direction
of fastest growth of $f$ starting from $\theta^*$ is conventionally defined as the vector from $\theta$ to that point a distance $\epsilon$ from $\theta$ that has the highest value of $f(\theta)$. In turn, the set of such points a distance $\epsilon$ from $\theta$ will vary depending on the metric. As a result, the direction of fastest growth of $f$ will vary depending on the metric. That means the directional derivative of $V$ along the direction of fastest growth of $f$ will vary depending on the metric. In fact, changing the metric may even change the sign of $V_f(\theta)$.

By the Cauchy-Schwarz inequality, $V_f(\theta) \leq \frac{|\text{grad}(V)|}{|\text{grad}(f)|}$, with equality if and only either $|\text{grad}(V)(\theta)| = 0$ or $\text{grad}(f)$ is positively proportional to $\text{grad}(V)$ (assuming $|\text{grad}(f)(\theta)|^2 \neq 0$ so that $V_f(\theta)$ is well-defined). In addition the bit-valued variable of whether the upper bound $\frac{|\text{grad}(V)|}{|\text{grad}(f)|}$ of $V_f(\theta)$ is tight or not has the same value at a given $\theta$ in all coordinate systems, since it is a (covariant) scalar. In fact, that bit is independent of the metric.

In particular, the “differential value of mutual information” (between some nodes $X$ and $S$) is

$$V_{I}(X;S)(\theta) = \frac{\frac{\partial}{\partial \theta} V g^{kl} \frac{\partial}{\partial \theta} I(X;S)}{\text{grad}(I(X;S))g^{kl}\text{grad}(I(X;S))}.$$ 

This is the amount that the player would value a change in the mutual information between $X$ and $S$, measured per unit of that mutual information.

To get an intuition for differential value of $f$, consider a locally invertible coordinate transformation at $\theta$ that makes the normalized version of $\text{grad}(f)$ be one of the basis vectors, $\hat{e}$. When we evaluate “(differential) value of $f$ at $\theta$”, we are evaluating the partial derivative of expected utility with respect to the new coordinate associated with that $\hat{e}$. (This is true no matter what we choose for the other basis vectors of the new coordinate system.) More concretely, since the coordinate transformation is locally invertible, moving in the direction $\hat{e}$ in the new coordinate system induces a change in the position in the original game parameter coordinate system, i.e., a change in $\theta$. This change in turn induces a change in the equilibrium profile $\sigma$. Therefore it induces a change in the expected utilities of the players. It is precisely the outcome of this chain of effects that “value of $f$” measures.

Changing the original coordinate system $\Theta$ will not change the outcome of this chain of effects — differential value of $f$ is a covariant quantity. However changing the underlying space of game parameters, i.e. what properties of the game are free to vary, will modify the outcome of this chain of effects. In other words, changing the parametrized family of games that that we are considering will change the value of the differential value of $f$. So we must be careful in choosing the game parameter space; in general, we should choose it to be exactly those attributes of the game that we are interested in varying. For example, if we suppose that some channels are free to vary, their specification must be included. Similarly, if we choose a model in which an overall multiplicative factor equally affecting all utility functions (i.e., a uniform tax rate) is free to vary, then we must also include that factor in our game parameter space. Conversely, if we choose a model in which there is no tax specified exogenously in the game parameter vector, then we must not include such a rate in our game parameter space. All of these choices will affect the dimensionality and structure of the parameter space and thus the formula
we use to evaluate value of $f$.

5. PROPERTIES OF DIFFERENTIAL VALUE

We now present some general results concerning value of a function $f : \Theta \to \mathbb{R}$, in particular conditions for negative values. Throughout this section, we assume that both $f$ and $V$ are twice continuously differentiable. In addition, note that when we randomly and independently choose (the directions of) $n \leq d$ vectors in $\mathbb{R}^d$, they are linearly independent with probability 1. That means, generically $n \leq d$ nonzero vectors span an $n$-dimensional linear subspace. In the sequel, we shall often implicitly assume that we are in such a generic situation and refrain from discussing nongeneric situations, that is, situations with additional linear dependencies among the vectors involved.

5.1. Preliminary definitions

To begin we introduce some particular convex cones (see appendix 9 for the relevant definitions) that we will use in our analysis of differential value of $f$ for a single player:

**Definition 8** Define four cones
\[
C_{++}(\theta) \equiv \{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle > 0, \langle \text{grad}(f), \delta \theta \rangle > 0 \}
\]
\[
C_{+-}(\theta) \equiv \{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle > 0, \langle \text{grad}(f), \delta \theta \rangle < 0 \}
\]
\[
C_{-+}(\theta) \equiv \{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle < 0, \langle \text{grad}(f), \delta \theta \rangle > 0 \}
\]
\[
C_{--}(\theta) \equiv \{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle < 0, \langle \text{grad}(f), \delta \theta \rangle < 0 \}.
\]

and also define
\[
C_{\pm}(\theta) \equiv C_{+-}(\theta) \cup C_{-+}(\theta).
\]

So there are two hyperplanes, $\{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle = 0 \}$ and $\{ \delta \theta : \langle \text{grad}(f), \delta \theta \rangle = 0 \}$, that separate the tangent space at $\theta$ into the four disjoint convex cones $C_{++}(\theta), C_{+-}(\theta), C_{-+}(\theta), C_{--}(\theta)$. These cones are convex and pointed. In fact, each of them is contained in some open halfspace.

By the definition of the differential value of $f$ in the direction $\delta \theta$, it is negative for all $\delta \theta$ in either $C_{++}(\theta)$ or $C_{+-}(\theta) = -C_{-+}(\theta)$, that is, in $C_{\pm}(\theta)$.

5.2. Geometry of negative value of information

In principle, either the pair of cones $C_{++}$ and $C_{--}$ or the pair of cones $C_{+-}$ and $C_{-+}$ could be empty. That would mean that either all directions $\delta \theta$ have positive value of $f$, or all have negative value of $f$, respectively. We now observe that the latter pair of cones is nonempty — so there are directions $\delta \theta$ in which the value of $f$ is negative — iff the value of $f$ is less than its maximum:
**Proposition 2**  Assume that $\text{grad}(V)$ and $\text{grad}(f)$ are both nonzero at $\theta$. Then $C_{+}(-\theta)$ and $C_{-}(\theta)$ are nonempty iff

$$V_{f}(\theta) < \frac{|\text{grad}(V)(\theta)|}{|\text{grad}(f)(\theta)|}$$

(16)

**Proof:**  Eq. (16) is equivalent to

$$\langle \text{grad}(V), \text{grad}(f) \rangle < |\text{grad}(V)| |\text{grad}(f)|,$$

that is, the two vectors $\text{grad}(V)$ and $\text{grad}(f)$ are not positively collinear, that is, they are not positive multiples of each other. It follows from Lemma 6 that two (nonzero) vectors $v_1, v_2$ are not positively collinear iff $\text{Con}([v_1, v_2])$ is pointed, i.e., iff the there are points in neither $\text{Con}([v_1, v_2])$ nor its dual. That in turn is equivalent to there being a third vector $w$ with

$$\langle v_1, w \rangle > 0, \quad \langle v_2, w \rangle < 0.$$  

(17)

With $v_1 = \text{grad}(V), v_2 = \text{grad}(f)$, this means that Eq. (16) implies that $C_{+} \neq \emptyset$, and therefore $C_{-} = -C_{+} \neq \emptyset$.

Q.E.D.

We emphasize that this result (and other results below) are not predicated on our use of the QRE, Fisher metric, or an information-theoretic definition of $f$. It holds even for other choices of the solution concept, metric, and/or definition of “amount of information” $f$. In addition, the requirement in Prop. 2 that $\theta$ be in the interior of $\Theta$ is actually quite weak. This is because often if a given MAID of interest is represented by a $\theta$ on the border of $\Theta$ in one parametrization of the set of MAIDs, under a different parameterization the exact same MAID will correspond to a parameter $\theta$ in the interior of $\Theta$.

Recall from the discussion just below Def. 7 that so long as neither $\text{grad}(f)$ nor $\text{grad}(V)$ equals 0, $V_{f}(\theta) < \frac{|\text{grad}(V)(\theta)|}{|\text{grad}(f)(\theta)|}$ iff $\text{grad}(V)(\theta) \neq \text{grad}(f)(\theta)$. So Prop. 2 identifies the question of whether $\text{grad}(V)(\theta)$ is positively proportional to $\text{grad}(f)(\theta)$ with the question of whether $C_{a}(\theta)$ is empty.

To illustrate Prop. 2, consider a situation where $V_{f}(\theta)$ is strictly less than the upper bound $\frac{|\text{grad}(V)(\theta)|}{|\text{grad}(f)(\theta)|}$, so that $\text{grad}(f) \neq \text{grad}(V)$. Suppose now, the player is allowed to add any vector to the current $\theta$ that has a given (infinitesimal) magnitude. Then she would not choose the added infinitesimal vector to be parallel to $\text{grad}(f)$, i.e., she would prefer to use some of that added vector to improve other aspects of the game’s parameter vector besides increasing $f$. Intuitively, so long as they value anything other than $f$, the upper bound on $V_{f}(\theta)$ is not tight.

Prop. 2 not only means that we would generically expect there to be directions that have negative value of $f$, but also that we would expect directions that have positive value of $f$:
Corollary 3  Assuming that $\operatorname{grad}(V)$ and $\operatorname{grad}(f)$ are both nonzero at $\theta$, 

\begin{equation}
|V_f(\theta)| < \frac{|\operatorname{grad}(V)(\theta)|}{|\operatorname{grad}(f)(\theta)|}
\end{equation}

implies that $C_{++}(\theta)$ and $C_{--}(\theta)$ are both nonempty.

Proof: Define $g(\theta) \equiv -f(\theta)$, and write $C^g$ or $C^f$ to indicate whether we are considering spaces defined by $V$ and $g$ or by $V$ and $f$, respectively.

\begin{align*}
|V_f(\theta)| &< \frac{|\operatorname{grad}(V)(\theta)|}{|\operatorname{grad}(f)(\theta)|} \\
\iff |V_g(\theta)| &< \frac{|\operatorname{grad}(V)(\theta)|}{|\operatorname{grad}(g)(\theta)|} \\
\implies V_g(\theta) &< \frac{|\operatorname{grad}(V)(\theta)|}{|\operatorname{grad}(g)(\theta)|} \\
\iff &C^g_{++}(\theta) \text{ and } C^g_{--}(\theta) \text{ are both non-empty} \\
\iff &C^{f}_{--}(\theta) \text{ and } C^{f}_{++}(\theta) \text{ are both non-empty}
\end{align*}

where Prop. 2 is used to establish the second to last equality. \textit{Q.E.D.}

Note that the converse to Coroll. 3 does not hold. A simple counter-example is where $\operatorname{grad}(V)(\theta) \propto \operatorname{grad}(f)(\theta)$, for which both $C_{++}(\theta)$ and $C_{--}(\theta)$ are nonempty, but $|V_f(\theta)| = \frac{|\operatorname{grad}(V)(\theta)|}{|\operatorname{grad}(f)(\theta)|}$.

Coroll. 3 means that so long as $|V_f(\theta)| < \frac{|\operatorname{grad}(V)(\theta)|}{|\operatorname{grad}(f)(\theta)|}$, at $\theta$ there are directions $\delta\theta$ with positive value of $f$. This has interesting implications for the analysis of value of information in the Braess’ paradox and Cournot example of negative value of information, as discussed below in Sec. 7.2. It also means that even if for some particular $f$ of interest one intuitively expects that increasing $f$ should reduce expected utility, generically there will be infinitesimal changes to $\theta$ that will increase $f$ but increase expected utility. An illustration of this is also discussed in the “negative value of utility” example in Sec. 7.2.

5.3. Genericity of negative value of information

Consider situations where $f$ is a monotonically increasing function of $V$ across a compact $S \subseteq \Theta$. This means that $f$ and $V$ have the same level hypersurfaces across $S$ (although, of course the values of $V$ and $f$ on any such common level hypersurface will in general be different). The monotonicity implies that the linear order induced by values $f(\theta)$ relates the level hypersurfaces in the same way that the linear order induced by values $V(\theta)$ relates those level hypersurfaces. Say that in addition neither $\operatorname{grad}(f)$ nor $\operatorname{grad}(V)$ equals 0 anywhere in $S$. So $\operatorname{grad}(V)$ and $\operatorname{grad}(f)$ are proportional to one
another throughout $S$ (although the proportionality constant may change). This means that $V_f(\theta)$ is maximal throughout $S$, and so by Prop. 2, for no $\theta$ in $S$ is there a direction $\delta \theta$ such that $V_{f, \delta \theta}(\theta) < 0$.

In general though, level hypersurfaces will not match up throughout a region, as the condition that $\text{grad} V$ and $\text{grad} f$ be proportional is very restrictive and special, and so is typically violated. When they do not, we have points in that region that have directions with negative value of $f$. We now derive a criterion involving both the gradients and the Hessians of $V$ and $f$ to identify such a mismatch.\footnote{Hypothesize that the gradients were not parallel at some $\theta \in S$. Since neither gradient equals zero, this would mean that there is a direction $\delta \theta \in S$ that has zero inner product with $\text{grad} f$ at $\theta$ but not with $\text{grad} V$ there. $\delta \theta$ would lie in the level surface of $f$ going through $\theta$ but not the level surface of $V$ going through $\theta$. This would contradict the supposition that those level surfaces are coincident throughout $S$.}

Proposition 4 Assume that both $f$ and $V$ are analytic at $\theta$ with nonzero gradients, and choose some $\epsilon > 0$. Define

\[
\begin{align*}
\nu &\equiv \text{grad} V(\theta) \\
\phi &\equiv \text{grad} f(\theta) \\
B_\epsilon(\theta) &\equiv \{ \sigma : \text{dist}(\sigma, \theta) < \epsilon \}
\end{align*}
\]

Then
\[
(19) \quad \implies \quad C_+(\theta') = \emptyset \quad \forall \theta' \in B_\epsilon(\theta)
\]

\[
(20) \quad \Rightarrow \quad \nu \propto \phi \quad \text{and} \quad \left| \nabla\nabla(f) - \frac{\phi_i}{|\nu|} \nabla\nabla(V) \right| |\nu_j| = \left| \nabla\nabla(f) \right| - \frac{\phi_i}{|\nu|} \left| \nabla\nabla(V) \right|, 
\]

where the proportionality constant in Eq. (20) is greater than zero.

Proof: We start with an observation. Whereas the two vectors $\text{grad} V$ and $\text{grad} f$ depend on the Fisher metric, the question of whether those two vectors are collinear does not depend on the choice of metric. (This is in contrast to the question of whether they are orthogonal, which needs a metric.) However as these two vectors depend on the metric, and as we shall have to take derivatives and therefore cannot stay in a single tangent space, we shall have to employ some of the differential geometry concepts reviewed in the appendix.

Hypothesize that $C_+(\theta') = \emptyset$. From Prop. 2 this is true iff $\text{grad} V(\theta')$ and $\text{grad} f(\theta')$ are positively proportional, that is,

\[
(21) \quad \text{grad} f(\theta') = \rho(\theta') \text{grad} V(\theta')
\]

for some positive function $\rho$. So using the definitions in the proposition, $\rho(\theta) = \frac{|\nu|}{|\phi|}$. In addition, if (19) holds, we can lower both the contravariant vector field $\text{grad} f$ and the contravariant vector field $\text{grad} V$ in (21) into covariant vector fields, and then apply the covariant derivative operator with respect to $\theta'$ to both sides. If we then raise all lower indices back up to upper indices, we derive

\[
(22) \quad [\nabla\nabla f(\theta')]^j_{ij} - \rho(\theta') [\nabla\nabla V(\theta')]^j_{ij} = [\text{grad} \rho(\theta')]^j [\text{grad} V(\theta')]^i
\]
Since both Hessians in this equation are the covariant Hessians, and therefore symmetric with respect to $i, j$, we conclude by symmetrization that

$$\frac{[\text{grad}\rho(\theta')][\text{grad}(V)(\theta')]^T}{[\text{grad}(V)]^T} \cdot [\text{grad}\rho(\theta')][\text{grad}(V)(\theta')]$$

for all pairs $(i, j)$. So in particular, evaluating all terms at the center of the ball, $\theta' = \theta$, we see that

$$\frac{[\text{grad}\rho]^T}{[\text{grad}(V)]^T} = \frac{[\text{grad}\rho]^T}{[\text{grad}(V)]^T}\text{for all } i, j \text{ such that } [\text{grad}(V)]^i \neq 0 \text{ and } [\text{grad}(V)]^j \neq 0.$$ Accordingly $[\text{grad}\rho]^T \propto [\text{grad}(V)]^T$ for all $i$ such that $[\text{grad}(V)]^i \neq 0$ where the proportionality constant is independent of $i$. So evaluating at $\theta' = \theta$ and by the above proportionality, (22) simplifies to

$$[\nabla^2 f]^{ij} - \frac{[\phi v]}{v^T} [\nabla V]^{ij} = k[\text{grad}(V)]^i[\text{grad}(V)]^j$$

for some scalar $k$. Taking norms of both sides, yields

$$k = \frac{\|\nabla^2 f - \frac{[\phi v]}{v^T} \nabla V\|}{v^T},$$

whence (20).

We will refer to the matrix with entries

$$\left| \nabla^2 f - \frac{[\phi v]}{v^T} \nabla V \right|^T \cdot \left( \nabla^2 f ight)_{ij} - \left| \nabla V \right|^2_{ij} \cdot \frac{[\phi v]}{v^T} \nabla V$$

as the second-order mismatch. Prop. 4 means that if there is nonzero second-order mismatch at some $\theta$ then $C_{\rho}(\theta)$ cannot be empty throughout a d-dimensional sphere centered at $\theta$. In other words, it means that for any $\epsilon > 0$, there is some $\theta'$ within distance $\epsilon$ of $\theta$ and associated direction $\delta \theta'$ such that at $\theta'$ there are negative values of $f$ in direction $\delta \theta'$.

---

15 One might think that so long as the Hessian of $V$ at $\theta$ is not a constant times the Hessian of $f$ at $\theta$, then the level curves of $V$ and $f$ will not match up throughout the vicinity of $\theta$. Such mismatch in the level curves would in turn imply that even if the gradients of $V$ and $f$ were parallel at $\theta$, there would be points infinitesimally close to $\theta$ where the gradients of $V$ and of $f$ are not parallel. If this reasoning were valid, it would suggest that the condition in Prop. 4 that the second-order mismatch is non-zero could be tightened (by replacing that condition with the requirement that $D_{\theta} f = \sum D_i V$) and yet the implication given by that proposition — that there are points where the gradients are not parallel — would still follow. This is not the case though: Even if the Hessian of $V$ is not a constant times the Hessian of $f$ at $\theta$, it is possible that the gradients of $f$ and $V$ are parallel throughout a neighborhood of $\theta$. As a simple example, take $d = 2$, $V(\theta) = \theta^T f(\theta) = (1 + \theta^2)^T (1 + \theta^2)^T$. Even though the Hessian of this $V$ is not a constant times the Hessian of this $f$ anywhere near the origin, the gradients of $f$ and $V$ are parallel everywhere near the origin.

16 However it is still possible for $C_{\rho}(\theta)$ to be empty throughout a $(d - 1)$-dimensional hyperplane $T$ going through $\theta$. As an example, choose $d = 2$, $V(\theta) = \theta^T f(\theta) = (1 + \theta^2)^T (1 + \theta^2)^T$. The second order mismatch is nonzero for this $V$ and $f$ at $\theta = (0, 0)$. (The Hessian of $f$ at the origin is twice the identity matrix, the Hessian of $V$ there is the zero matrix, but $v_i$ only has one non-zero entry, for $i = j = 1$.) However everywhere on the line $\{\theta^T t \in \mathbb{R}^2 : t \in \mathbb{R} \}$ going through the origin, $\text{grad} f(\theta^T)$ and $\text{grad} (V)(\theta^T)$ are parallel.
In light of this, arguably the most important issues to analyze in any given game is not whether at some particular \( \theta \) there is a direction in which some particular type of information increases but expected utility decreases. That is almost always the case. Rather our results redirect attention to different issues. One such issue is what fraction of all \( \theta \) and all possible infinitesimal changes to \( \theta \) have negative value of information.

Another important issue arises whenever there are constraints on the allowed directions in \( \Theta \) specifying how \( \theta \) is changed. Such constraints are quite common in the real world. As an example, there might be inequality constraints on the information capacity of a channel, which would manifest themselves in constraints on the allowed change in parameter space for \( \theta \) for which those constraints are tight. Another example of a constraint on the allowed change to the parameter vector is the stipulation that the change must be an insertion (or removal) of a garbling node, in the sense of Blackwell’s theorem (Leshno and Spector, 1992). With such constraints, in general the directions in the parameter space for \( \theta \) of a channel, which would manifest themselves in constraints on the allowed change in parameter space for \( \theta \) for which those constraints are tight. Another example of a constraint on the allowed change to the parameter vector is the stipulation that the change must be an insertion (or removal) of a garbling node, in the sense of Blackwell’s theorem (Leshno and Spector, 1992). With such constraints, in general the directions in the sets \( C_{+}(\theta) \) and \( C_{-}(\theta) \) arising in Prop. 2 may all violate the constraints, and so not be allowed. This is why, for example, the result of Blackwell (that there are no garbling nodes that one can insert just before a decision node in an ID that increase maximal expected utility for a rational player) is consistent with our results; the insertion of such a garbling node corresponds to a constrained change in \( \theta \), a change that lies in a cone that does not intersect \( C_{+}(\theta) \) (see Fig. 4).

### 5.4. Pareto negative values of information

Recall that Prop. 2 established that there are directions in which the value of \( f \) is negative for \( i \) iff \( \text{grad}(f) \) and \( \text{grad}(V) \) are not positively proportional. We can generalize this to give a condition for there to be directions in which the value of \( f \) is negative for all players, in which case we say that we have “Pareto negative values of information”. To make this precise, define \( V_{i;\delta\theta}(\theta) \) as the value of \( f \) in direction \( \delta\theta \) to player \( i \) in a MAID.

**Proposition 5** Given a set of functions \( V_{i} : \Theta \rightarrow \mathbb{R} \), define \( C \equiv \text{Con}((\text{grad}(V_{i}))) \). Assume \( C^{\perp} \neq \emptyset \). Then

\[
\text{grad}(f) \not\in C
\]

(27) \( \iff \exists \, \delta\theta \quad : \quad V_{i;\delta\theta}(\theta) < 0 \, \forall i \)

where all \( N + 1 \) gradients are evaluated at \( \theta \).

**Proof:** Note that the dual cone \( \text{Con}(\text{grad}(f)) = H_{-}(\text{grad}(f)) \). So by Lemma 8 (see appendix 9), since \( C^{\perp} \) is nonempty, \( C^{\perp} \subseteq H_{-}(\text{grad}(f)) \) iff \( \text{grad}(f) \in C \). Thus, \( \text{grad}(f) \not\in C \) means that the set difference \( C^{\perp} \setminus H_{-}(\text{grad}(f)) \) is nonempty. Since those two sets are defined by strict inequalities, this set difference is open. So in fact \( \text{grad}(f) \not\in C \) iff there is some \( \delta\theta \) such that both \( \delta\theta \in C^{\perp} \) and \( \delta\theta \not\in \text{cl}[H_{-}(\text{grad}(f))] \), the closure of \( H_{-}(\text{grad}(f)) \). For that \( \delta\theta \), \( \langle \delta\theta, V_{i} \rangle < 0 \) for all \( i \), but \( \langle \delta\theta, \text{grad}(f) \rangle > 0 \). This means \( V_{i;\delta\theta}(\theta) < 0 \) for all \( i \). This argument also works in the converse direction. \( Q.E.D. \)

The existence condition in Prop. 5 is crucial. To give a simple example, note that if \( \text{grad}(V)^{1} = -\text{grad}(V)^{2} \), then even if \( \text{grad}(f) \not\in C \), there is no direction \( \delta\theta \) such that
both $\nabla_{f,\delta\theta}^1(\theta) < 0$ and $\nabla_{f,\delta\theta}^2(\theta) < 0$.\footnote{To see this, note that if there were such a $\delta \theta$, it could not lie in $H_d(\text{grad}(V_i))$, and so must either lie in $H_{\perp}(\text{grad}(V_i))$ or $H_{\perp}(\text{grad}(V_i)) = H_{\perp}(\text{grad}(V_i))$. Wolog assume the former. Then $\nabla_{f,\delta\theta}^1(\theta) < 0$ would require that $\delta \theta \in H_{\perp}(\text{grad}(f))$. However $\delta \theta \in H_{\perp}(\text{grad}(V_i)) = H_{\perp}(\text{grad}(V_i))$ would then force $\nabla_{f,\delta\theta}^2(\theta) > 0$.} The reason this is not a counterexample to Prop. 5 is that $C^\perp$ is empty in this case.

Similarly, the implication of Prop. 5 can be confounded when $\text{Con}((\text{grad}(V_i)))$ is not pointed, so that by Lemma 7, $C^\perp = \emptyset$. This is the case for many sets of MAIDs. As an illustration, consider a set of $N$-player MAIDs indexed by $\Theta$ and a $\theta \in \Theta$ where there is some ball $B(\theta)$ centered on $\theta$ such that the MAIDs of all $\theta' \in B(\theta)$ have the same joint-move space $X$. Suppose further that the utility functions for all those $\theta'$ specify that the associated game is constant-sum, with the sum of utilities having the same value $\mathcal{U}$ for all those $\theta' \in B(\theta)$. (For example, it may be that the $\theta' \in B(\theta)$ vary only in the conditional distributions they assign to the chance nodes of the MAID — so that the utility functions of the players are the same throughout $B(\theta)$ — and that the utility functions at $\theta$ form a constant-sum game with utility sum $\mathcal{U}$.) Then no matter what equilibrium concept we use, the sum of the expected utilities of the players will equal $\mathcal{U}$ for all $\theta' \in B(\theta)$. Therefore the sum of the vectors $\text{grad}(V_i)$ evaluated at $\theta$ equals zero, i.e., the grad’s there are linearly dependent. So $\text{Con}((\text{grad}(V_i)))$ is not pointed in such sets of MAIDs. This means that there is no direction $\delta \theta$ such that $\nabla_{f,\delta\theta}^i(\theta) < 0 \forall i$ for such sets of MAIDs.

When the condition in Prop. 5 holds, there is a direction from $\theta$ that is Pareto-improving, yet reduces $f$. So for example, if that condition holds for a particular set of MAIDs when $f$ is the information capacity of a channel in the MAID, then all players would benefit from reducing the capacity of that channel.

Similarly, to the second order mismatch condition above which implies that negative value of information is rather generic, the conditions of Prop. 5 are expected to hold if the number of players $N$ is less than the dimension $d$ of the parameter space $\Theta$. Simply observe that $\text{Con}((\text{grad}(V_i)))$ is spanned by $N$ vectors and thus of at most dimension $N < d$. Now, it is rather non-generic that an unrelated vector, e.g. $\text{grad}(f)$, happens to lie in this lower-dimensional subspace. Thus, in this case, Pareto improving changes to $\theta$ which decrease the value of $f$ should be prevalent.

However it is important to remember that Pareto-improving changes to $\theta$ are not the same thing as Pareto-improving changes to a strategy profile in a game specified by a single $\theta$. We can have Pareto-improving changes to $\theta$ even if $\theta$ specifies a game in which there are no Pareto-improving changes to the equilibrium strategy profile. In particular, $\theta$ can specify a constant-sum game — which allows no Pareto-improving changes to strategy profiles — while there are changes to $\theta$ that (result in changes to the equilibrium profile that) are Pareto-improving.

This is reflected in the fact that even if $\theta$ specifies a constant-sum game, it may be that $\text{Con}((\text{grad}(V_i)))$ is non-pointed.\footnote{For example it may be that all games infinitesimally close to $\theta$ are also constant-sum — but with a different sum of utilities from the sum at $\theta$. In this case changing the game from $\theta$ to some infinitesimally close $\theta'$ will change the sum of expected utilities of the players, and so the sum of the $\text{grad}(V_i)$ is non-zero.} Due to this fact, Prop. 5 does not mean that if $\theta$ specifies a constant-sum game, then there is no direction in $\Theta$ that is Pareto-improving
while reducing $f$. The character of games infinitesimally close (in $\Theta$) to $\theta$ is what is important.

Note that the bit of whether or not $\text{grad}(f) \in \text{Con}([\text{grad}(V_i)])$ is a covariant quantity. So the implications of Prop. 5 do not change if we change the coordinate system. In fact, the value of that bit is independent of our choice of the metric, so long as no grad $V_i$ is in the kernel of the metric. This means in particular that the implications of Prop. 5 do not vary with the choice of metric, so long as we stick to Riemannian metrics.

A similar result, can be obtained for Pareto positive value of information. By Def. 6, for any $i$, $\mathcal{V}^i_{f,\delta\theta}(\theta') < 0$ iff $\mathcal{V}^i_{f,\delta\theta}(\theta') > 0$. So an immediate corollary of Prop. 5 is that that whenever $C^\perp \neq \emptyset$, there a direction $\delta\theta$ in which $\mathcal{V}^i_{f,\delta\theta}(\theta') > 0$ iff $-\text{grad}(f) \notin C$. So if $C^\perp \neq \emptyset$, then if in addition the negative of the gradient of $f$ is not contained in the conic hull of the gradients of the players’ expected utilities, there is a direction in which we can change the game parameter vector which will increase both $f$ and expected utility for all players.

6. ANALYSIS OF SINGLE-PLAYER GAMES

To illustrate the foregoing, in this section we work through the formulas given in Sec. 4 for the case of a game against Nature (the Blackwell ID).

6.1. Decision problem

In a decision problem one agent plays against nature. (So this MAID is the special case of an influence diagram.) A simple example, which we will use to illustrate our notion of differential value of information is shown in Fig. 2 as the DAG of an ID.

In this MAID there is a state of nature random variable $X$ taking on values $x$ according to a distribution $p(x)$. The agent observes $x$ indirectly through a noisy channel that produces a value $s$ of a signal random variable $S$ according to a probability distribution $p(s \mid x; \theta)$ parametrized by $\theta$. The agent then takes an action, which we write as the value $a$ of the random variable $A$, according to the distribution $\sigma(a \mid s)$. Finally, the utility $u(x, a)$ is a function that depends only on $x$ and $a$.

![Figure 2.— A simple decision situation.](image)

We will refer to this MAID as the **Blackwell ID**, since it corresponds to the situation analyzed by Blackwell.
In the Blackwell ID, for a given signal $s$ and action $a$, the conditional expected utility is

$$E(u | s, a; \theta) = \sum_x p(x | s; \theta) u(x, a)$$

(28)

and given any distribution $\sigma(a | s)$ by the decision-maker, the associated unconditional expected utility is

$$\sum_s, a \sigma(a | s) \mathbb{E}(u | s, a; \theta)$$

where $p(s; \theta) = \sum_x p(s | x; \theta) p(x)$. A fully rational decision maker will set $\sigma(a | s)$ to maximize this unconditioned expected utility. That would be equivalent to setting $\sigma(a | s)$ for each possible $s$ to maximize $E(u | s, a; \theta)$.

Here, to have a differentiable strategy, we assume that the agent is not fully rational but plays a quantal best response with a finite $\beta$:

$$\sigma(a | s) = \frac{1}{Z(s)} e^{\beta \mathbb{E}(u | s, x; \theta)}$$

where $Z(s) = \sum_a e^{\beta \mathbb{E}(u | s, x; \theta)}$ denotes the normalization constant.

6.1.1. Calculating the gradient of the expected utility

In order to identify the direction in the parameter space of the channel that is relevant for utility changes we have to calculate the gradient of $V(\beta, \theta)$ with respect to $\theta$. The expected utility depends on $\theta$ in two ways: (1) Directly via the change of the channel distribution $\frac{\partial}{\partial \theta} p(s | x; \theta)$ and (2) indirectly via the induced change in the strategy $\frac{\partial}{\partial \theta} \sigma(a | s)$. The last term could be obtained from the implicit differentiation formula eq. (12).

As explained above, in order to obtain a proper contra-variant gradient vector we need a metric. We start from the observation that the joint distribution on the MAID, i.e.

$$p(x, s, a; \theta) = p(x)p(s | x; \theta)\sigma(a | s)$$

depends on the channel parameters $\theta$ via the channel and implicitly the strategy of the agent. According to eq. (4), the Fisher metric is given by

$$g_{il}(\theta) = \sum_{x,s,a} p(x, s, a; \theta) \frac{\partial \log p(x, s, a; \theta)}{\partial \theta^i} \frac{\partial \log p(x, s, a; \theta)}{\partial \theta^l}$$

Thus, even when using a binary channel with parameters $\theta = (\epsilon^1, \epsilon^2)$, the metric is not the one that we calculated in Sec. 2.2, but includes additional terms which reflect that the strategy of the decision maker adapts to changes of the channel parameters as well. Using $p(x, s, a; \theta) = \log p(x) + \log p(s | x; \epsilon^1, \epsilon^2) + \log \sigma(a | s)$ and that $p(x)$
does not depend on $e^1, e^2$, we obtain:

$$g_{ij}(e^1, e^2) = \sum_{x,i,d} p(x)p(s \mid x)\sigma(a \mid s)\frac{\partial \log p(s \mid x)}{\partial e^k} \frac{\partial \log p(s \mid x)}{\partial e^l} + \sum_{x,i,d} p(x)\frac{\partial p(s \mid x)}{\partial e^k} \frac{\partial \log p(s \mid x)}{\partial e^l} + \sum_{x,i,d} \frac{\partial p(s \mid x)}{\partial e^k} \frac{\partial \sigma(a \mid s)}{\partial e^l} \frac{\partial p(s \mid x)}{\partial e^l}$$

$$+ \sum_{x,i,d} \frac{\partial p(s \mid x)}{\partial e^k} \frac{\partial \sigma(a \mid s)}{\partial e^l} + \sum_{x,i,d} \frac{\partial \sigma(a \mid s)}{\partial e^k} \frac{\partial \sigma(a \mid s)}{\partial e^l}$$

Thus, in addition to the first term which corresponds to the result calculated above, we have three additional terms which take the dependence of the players decision into account. The gradient of the expected utility is then obtained by $(\text{grad}(V))^j = \frac{\partial \text{grad}(V)}{\partial \theta} g^{ij}(\theta)$ where $g^{ij}$ is the inverse of the Fisher metric $g_{ij}(\theta)$.

Fig. 3 illustrates the effect of changes in the player’s rationality $\beta$ on the metric $g(e^1, e^2)$. For different channel parameters, the shape of the metric is represented by plotting an ellipse which illustrates how the “unit ball” wrt the Fisher metric over the manifold of channel parameters appears when it is plotted in a coordinate system of that manifold given by the values $(e^1, e^2)$. Note that $\beta = 0$ corresponds to a player that always plays uniformly at random, i.e., a completely non-rational player. Since such a player does not react to changes of the channel noise, the metric reduces to the one calculated in Sec. 2.2 in this case. In contrast, a more rational player reacts to changes of the channel parameters and the metric ellipses get distorted accordingly. The figure shows that a more rational player reacts more strongly to changes of the channel noise (stronger distortion) and does so at successively higher noise levels (that is where the distortion is most pronounced). In the extreme case of a fully rational player, $\beta \to \infty$, the metric becomes singular along the line in parameter space where the best response of the player changes.

Generalizing from this kind of plot, we could also include the rationality parameter $\beta$ of the player into the metric, in which case the metric becomes a $3 \times 3$ matrix, $g(e^1, e^2, \beta)$. This analysis would take into account how the strategy of the player changes when their rationality $\beta$ varies, and how that relates to changes when the channel parameters vary.

Similarly we can include parameters of utility functions, e.g., tax rates, regulatory factors, etc., in the parameter space. (See Ex. 3.) We could then use our framework to evaluate the value of (functions) of those parameters. By comparing these values of functions of $\theta$, our framework would allow us to study the marginal rates of substitution among (value of) information, utility, rationality, etc..

Here though, to focus on the effect of information changes, we restrict attention to changes to the channel parameters $(e^1, e^2)$ and consider $\beta$ and other parameters as fixed.

---

19 As explained above, the Fisher metric quantifies the sensitivity of a distribution $p(\epsilon; \theta)$ to changes of the parameters $\theta$. Thus, when the distribution is insensitive to changes of $\theta$, a unit change to the distribution requires a large parameter change and the unit ball appears stretched in the $i$th coordinate direction.
6.2. Gradient of the mutual information

As argued above, to understand more precisely the role of information we should compare the gradient of the expected utility with the gradient of the mutual information. The mutual information between $S$ and $X$ is

\begin{equation}
I(S; X) = \sum_{s,x} p(s \mid x, \theta) p(x) \log \frac{p(s \mid x, \theta)}{\sum_{x'} p(s \mid x, \theta) p(x)} .
\end{equation}

\begin{equation}
\frac{\partial}{\partial \theta^k} I(S; X) = \sum_{s,x} \frac{\partial}{\partial \theta^k} p(s \mid x; \theta) p(x) \log \frac{p(s \mid x, \theta)}{\sum_{x'} p(s \mid x, \theta) p(x)}
\end{equation}

As usual, the gradient $\text{grad} I(S; X)$ is obtained by multiplying this vector of partial derivatives with the inverse of the Fisher metric.

6.3. Differential value of information

To illustrate the above concepts, we consider a simple setup using a binary state of nature $p(X = 0) = p(X = 1) = \frac{1}{2}$ and a binary channel $p(s| x; \theta)$ as in eq. (6). The player
has two moves \( A = 0, A = 1 \) and obtains utility

\[
u(x, a) = \begin{cases} 
0 & x = 0, a = 0 \\
-2 & x = 0, a = 1 \\
0 & x = 1, a = 0 \\
1 & x = 1, a = 1 
\end{cases}
\]

Thus, the player has a safe action \( A = 0 \), but can obtain a higher utility by playing 1 when she is certain enough that the state of nature is 1.

Fig. 4 show the isoclines of the mutual information \( I(X; S) \) and the expected utility \( V(\beta, \epsilon_1, \epsilon_2) \) for \( \beta = 5 \). Both mutual information and expected utility improve with decreasing channel noise, i.e. with reducing \( \epsilon_1 \) and/or \( \epsilon_2 \). Nevertheless, the isoclines of these quantities do not exactly match. Thus it is possible to change the channel parameters such that the expected utility increases while the mutual information decreases. As an example, consider any parameter combination \( (\epsilon_1, \epsilon_2) \) where the isoclines plotted in Fig. 4 intersect. Moving into the region “above” the \( MI \) isocline and below the \( V \) isocline will increase \( V \) while decreasing \( MI \).

This potential inconsistency between changes to mutual information and changes to expected utility does not violate Blackwell’s theorem. That is because we allow arbitrary changes to the channel parameters; those that result in the inconsistency which cannot be represented as garbling in the sense of Blackwell. As an illustration, for one particular pair of game parameter values, the set of channels which are more or less informative according to Blackwell are visualized as the light and dark gray regions in fig. 4 respectively.

This potential for an inconsistency between changes to information and utility is also illustrated by considering the gradient vectors \( \text{grad} I(X; S) \) and \( \text{grad} V \) which are orthogonal to the isoclines (wrt the Fisher metric). By Prop. 2, only if those gradients are collinear is it the case that every change to the game parameter vector increasing the expected utility must necessarily increase the mutual information. However in Fig. 4 we clearly see that these gradients giving the directions of steepest ascent of \( I(X; S) \) and \( V \) are different.

Conversely, Cor. 3 implies that we can find infinitesimal changes to the game parameters that cause both expected utility and mutual information to increase, in agreement with Blackwell’s theorem. In the present example, such changes arise if we simultaneously reduce both channel noises, e.g. by moving directly towards the origin.

Recall that the differential value of information quantifies how much expected utility \( V \) changes per unit change of the mutual information \( (X; S) \). So the value of information is large when the gradients are almost collinear and can even become negative when their angle is above 90 degrees. Fig. 5 plots the differential value of information as a function of the game parameter (Def. 7), and therefore shows where the gradients \( \text{grad} V \) and \( \text{grad} I(X; S) \) are aligned and where they are aimed in very different directions:
7. ANALYSIS OF MULTI-PLAYER GAMES

In this section we illustrate our approach for games involving more than a single player.

7.1. Leader-follower example

We start by analyzing simple games involving a leader $A^1$ and a follower $A^2$ (see Fig. 6). In contrast to the single-player game, now the distribution $p(X)$ of the state of nature is replaced by the equilibrium strategy $\sigma(a_1)$ of player 1, a strategy that will also depend on the parameters $\theta$. Another difference is that there are now two utility functions, one for each player.

As in the decision problem, we consider binary state spaces and an asymmetric binary channel with parameters $\theta = (\epsilon^1, \epsilon^2)$. We use the utility functions of the players analyzed in (Bagwell, 1995)\(^{20}\)

---

\(^{20}\)The game is a discretization of the Stackelberg duopoly game, with two moves for each player.
Bagwell pointed out that in the pure strategy Nash equilibria the leader can only take advantage of moving first (by playing $L$), when the follower can observe his move perfectly (Stackelberg solution). As soon as the slightest amount on noise is added to the channel only the equilibrium of the simultaneous move game (both playing $R$, the Cournot solution) remains. Here we show that our differential analysis uncovers a much richer structure. In particular, we show there exist a QRE branch and parameters for the noise of the channel such that both players prefer more noise.

In the decision case, we used $I(X;S)$ to quantify the amount of information that is available to the (single) player. In the multi-player game setting, the corresponding quantity is $I(A_i;S)$, and strongly depends on the move of the leader. As an illustration, consider a symmetric channel $p(s|a_1)$ parametrized by the single value $\epsilon \equiv \epsilon_{1,2}$. Fig. 7 shows how the strategy of the leader depends on the channel noise and the rationality $\beta = \beta_{1,2}$ of the players. This shows that for sufficiently rational players ($\beta > 5$) there exist multiple QRE solutions. For $\beta \to \infty$ the three QRE equilibria converge to the pure strategy Nash equilibrium where both play $R$ (Cournot outcome, lower branch in red/green) and to the two mixed strategy Nash equilibria of the original game, respec-

---

<table>
<thead>
<tr>
<th>$u_1/u_2$</th>
<th>follower</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>(5, 2)</td>
</tr>
<tr>
<td>$R$</td>
<td>(6, 3)</td>
</tr>
</tbody>
</table>

---

21There are additional mixed equilibria, which change smoothly with the noise. These are mentioned in Bagwell (Bagwell, 1995), but not discussed further.
Figure 6.— A 2-player game where player $A^1$ (leader) can move before player $A^2$ (follower).

For $\epsilon = 0$, the upper mixed strategy equilibrium coincides with the equilibrium mentioned above where the leader has an advantage. In the following, we focus on the branch that smoothly connects to the origin $\beta = 0$, the so-called “principal branch”, which includes that upper equilibrium.

Fig 8 shows the channel capacity as a function of the conditional distribution in the channel, as well as the mutual information $I(A_1; S)$ that is actually transferred across the channel. As soon as the leader is rational enough, he starts to prefer the move $L$. This means that the mutual information $I(A_1; S)$ decreases when the leader gets rational enough. However, the potential information that could be transferred, i.e., the channel capacity, is independent of player strategies, and so is still high. This illustrates how studying the information capacity rather than the mutual information is perhaps more in line with standard game theory, where the information partition is considered as part of the specification of the game parameters, independent of the resultant player strategies.

In this simple symmetric-channel scenario the space of game parameters concerning the channel noise is one-dimensional, and so analyses of “gradients” over that space are not particularly illuminating. Accordingly, to further investigate the role of information in the leader-follower game we consider an asymmetric channel, so that $p(s|a_1)$ is parametrized by two noise parameters, $\epsilon_1, \epsilon_2$, giving the probability of error for the two inputs $a_1 = L$ and $a_1 = R$ respectively. We also fix $\beta = 10$ for both players. In this case, we again find multiple QRE branches when the channel noise is small enough. We focus on the branch where the leader has the biggest advantage and can achieve the highest utility while the utility of the follower is lowest. In the following, we refer to this QRE

---

22This demonstrates that our analysis can easily be extended to analyze Nash equilibria. In this case, choosing a branch corresponds to choosing a particular equilibrium, and the partial derivatives $\frac{\partial I}{\partial \epsilon}$ vanish, as long as the equilibrium exists.

23Remember that $I(A_1; S) = \mathbb{E}_{p(A_1)}[D_{KL}(p(S|a_1) \parallel p(S))]$ and thus it vanishes if the leader plays a pure strategy, since the average becomes trivial.
solution as the “Stackelberg branch”.

Fig. 9 shows the differential value of channel capacity as well as mutual information $I(A_1; S)$ for the leader and the follower on this Stackelberg branch. Again, we observe a large difference between results for channel capacity and for mutual information. The expected utility of the leader is higher when the channel noise is low than when it is high. So we would expect a positive differential value of channel capacity. For the follower, the situation is reversed (panel C, same figure) and accordingly her differential value of channel capacity should be negative. This is mostly confirmed (Fig. 9, panels A and B), but there are regions of the parameter space where the differential value of channel capacity is negative for both players. This occurs because $\epsilon_1$ is more important (in terms of expected utility) to the leader than is $\epsilon_2$, a distinction which is not reflected by the channel capacity (which is symmetric w.r.t. $\epsilon_1, \epsilon_2$).

Understanding the differential value of mutual information $I(A_1; S)$ is less straightforward. Part of the reason is that we have to take account of the fact that the strategy of the leader becomes more deterministic when the channel noise is low, and accordingly the mutual information can be reduced even though the channel capacity is increased. In fact, the mutual information has a saddle point around $(\epsilon_1 = 0.15, \epsilon_2 = 0.3)$ which leads to a singularity of the differential value of mutual information at this point. For these reasons we now focus our analysis on the channel capacity.

Fig. 10 A shows the isoclines of expected utility for both leader and follower, the isoclines of the channel capacity, and the corresponding gradient fields, all on the Stackelberg branch. We immediately see that the gradients are nowhere collinear. So from Prop. 2 we know that for all values of the game parameters, the channel noises can be

\[ \text{Figure 7.— Surface of QRE equilibria for the symmetric channel leader-follower game.} \]
infinitesimally changed such that the capacity increases whereas the expected utility (of either the leader or follower) decreases. More interestingly, the gradient field shows that everywhere $\nabla \text{Capacity} \notin \text{Con}(\{\nabla V^1, \nabla V^2\})$. Thus by Prop. 5, there must be directions such that both players are better off and yet the channel capacity decreases when the game is moved in that direction.

Simply by redefining $f$ as the negative of the capacity as the function of interest (which amounts to flipping the corresponding gradient vectors in our figures), the same condition in Prop. 5 can be used to identify regions of channel parameter space where (there are directions in which we can move the game parameters in which) both players prefer more capacity. Now $\nabla (-\text{Capacity}) \in \text{Con}(\{\nabla V^1, \nabla V^2\})$, except for a small region in the upper left corner (containing for example the point $(\epsilon^1 = 0.05, \epsilon^2 = 0.35)$). So we can immediately conclude that there are no such directions, for all parameter values outside this region.

To illustrate the effect of our choice of which equilibrium branch to analyze, Fig. 10 shows the corresponding isoclines on the branch just below the Stackelberg branch. Again, the gradient vectors are nowhere aligned. However now $\nabla \text{Capacity} \in \text{Con}(\{\nabla V^1, \nabla V^2\})$. Thus, we can conclude that on this branch there are no directions where both players prefer a decrease in channel capacity, in contrast to the case on the Stackelberg branch.

Summarizing, even in rather simple leader-follower games, there is a surprisingly complicated underlying geometric structure. However just like in the decision-theoretic scenario, our theorems provide general conditions on the gradient vectors in leader-follower games that determine regions of channel parameter space in which there is negative value of information for both players. Moreover, these conditions are easily checked. In particular, they do not require laboriously comparing different information

---

24By Prop. 4 this behavior is rather generic and thus expected.
partitions to get non-numeric partial orders. Instead our conditions are based on cardinal quantifications of how much the players value infinitesimal changes to the information structure of the game, measured in units of utility per bits of information.

This kind of analysis can be applied multiple times at once, to quantify how much the players value different candidate changes to the information structure of the game. These quantifications are all measured in the same units, of utility per bits of information. So we can use their values to evaluate marginal rates of substitution of various kinds of changes to the game’s information structure. More generally, we can apply our analysis to quantify how much the players value different candidate infinitesimal changes to any aspects of the game specification, e.g., to parameters in utility functions. This allows us to evaluate marginal rates of substitution of all aspects of the game specification. This is the subject of future research.

Figure 9.— Top: Differential value of channel capacity for the leader (panel A) and the follower (panel B) in the asymmetric channel scenario. Bottom: Differential value of mutual information $I(X ; S)$ for the leader (panel C) and the follower (panel D).
Figure 10.— Isoclines of the expected utilities for the leader and the follower as well as the channel capacity in the asymmetric channel scenario, on the Stackelberg branch (A) and on the branch just below it in the space of leader strategies (B). The expected utility levels are color-coded and the channel capacity (dotted) increases towards the origin (lower-left corner). The gradient vectors show the corresponding directions of steepest ascent (wrt the Fisher metric).

7.2. Illustrative examples

To illustrate the generality of our framework, in this section we present additional examples, emphasizing the implications of our results in Sec. 5 for several different economic scenarios.

Example 1 Consider a variant of the well-known scenario where a decision by an individual is a costly signal to a prospective employer of their capability as employees. In this variant there is a car repair chain that wants to hire a new mechanic. (So we have two players.) There is an applicant for the position who has some pre-existing ability at car repair. That ability is their type; it is determined by a prior distribution that neither player can affect. The repair chain cannot directly observe the applicant’s ability. So instead, they will give the applicant a written test of their knowledge of cars. The repair chain will then use the outcome of that test to decide whether to offer the job to the applicant and if so, at what salary. (The idea is that giving too low a salary to a new mechanic will raise the odds that after being trained by the repair chain that new mechanic would simply leave for a different repair chain, at a cost to the repair chain.)

The applicant will study before they take the test. The harder they study, the greater the cost to them. There is also a conditional distribution of how they do on the test given their ability to repair cars and how hard they study. The fact that that distribution is not a single-valued map means it is a noisy information channel, from the studying decision of the applicant to the outcome of the test.

Suppose that the repair chain feels frustrated by the fact that the test gives them a
noisy signal of the applicant’s ability to repair cars. Knowing this, a test-design company approaches the repair chain, and offers to sell them a new test that has a double-your-money back guarantee that it is less noisy than the current test (for some functional of how “noisy” a test is that the test company and repair chain both use). Thinking they will get double their money back if they buy the new test but have lower expected utility, the repair chain buys the test. 25

Our results show that generically, there are ways that the repair chain will have lower expected utility with the new test — but not be able to invoke the guarantee to get any money back from the test-design company, since the new test is less noisy than the old one. That is, there are directions $\delta \theta$ in the parameters describing how the test is designed, such that the test is made more accurate, but less useful for the repair chain, i.e. it has a negative value of information.

The MAID corresponding to this noisy signaling game is shown in Fig. 11 A. For comparison, panel B reproduces the MAID for the leader-follower game that was studied in section 7. In the noisy signaling game there is an additional nature node $T$ which player 1 (the applicant) can observe, but that player 2 (the car repair chain) cannot ob-

\[\text{Figure 11.— MAID for the noisy signalling game described in example 1 (panel A) and the leader-follower game from section 7 (panel B).}\]
serve. Another difference is that now the utility of both players depends in part on the outcome of nature’s choice. Since the applicant incurs a cost if she studies, her utility directly depends on her move (study hard or party). However the utility of the repair chain does not depend directly on the move of the applicant.

We now present a variant of Braess’ paradox (Braess, 1968) which provides a particularly striking example of how extra information can simultaneously hurt all players of a game, even in the case of many players. Braess’ paradox is a “paradox” that arises in congestion games over networks. Although it has arisen in the real world in quite complicated transportation networks, it can be illustrated with a very simple network. We illustrate in this next example, before presenting our variant of Braess’ paradox.

**Example 2** Consider a scenario where there are 4000 players, all of whom must commute to work at the same time, from a node “Start” on a network to a node “End”. Say that there are a total of four roadways between Start and End that all the players know exist. (See Fig. 12.)

The move of each player is a choice of what route they will follow to get to work. There are two choices they can make. The first is the route Start-A-End, and the second is the route Start-B-End. The number of minutes it takes each player to follow their chosen route is illustrated in Fig. 12. In that figure, $t$ indicates the amount of time it takes a player to cross the associated road. In addition, $T$ refer to the total number of players (out of 4000) who follow the associated road. So on those two roads where $t = T/100$, the greater the amount of traffic $t$, the slower the players can drive, and so the greater the amount of time $T$ it takes to cross the road. In contrast, it takes 45 minutes to traverse the other two roads, regardless of the amount of traffic on the roads.

The paradox arises, if a new highway from A to B (dashed) is opened which takes only 4 minutes to traverse. Now, each player has a third option, namely to take the route Start-A-B-End. So we have a new game.

The move of each player is a choice of what route they will follow to get to work. There are two choices they can make. The first is the route Start-A-End, and the second is the route Start-B-End. The number of minutes it takes each player to follow their chosen route is illustrated in Fig. 12. In that figure, $t$ indicates the amount of time it takes a player to cross the associated road. In addition, $T$ refer to the total number of players (out of 4000) who follow the associated road. So on those two roads where $t = T/100$, the greater the amount of traffic $t$, the slower the players can drive, and so the greater the amount of time $T$ it takes to cross the road. In contrast, it takes 45 minutes to traverse the other two roads, regardless of the amount of traffic on the roads.

The paradox arises, if a new highway from A to B (dashed) is opened which takes only 4 minutes to traverse. Now, each player has a third option, namely to take the route Start-A-B-End. So we have a new game.

It is straightforward to verify that in all Nash equilibria of the original game, i.e. the game without the new highway, exactly half of the players choose Start-A-End, and half choose Start-B-End. The total cost (i.e., negative utility) to all players is a 65 minute commute. When the new highway is opened, every player will choose Start-A-B-End.
This increases the travel time of all players to 84 minutes. Thus, the expected cost for all players rises when they have more possible actions.

Now consider a variant of Braess’ paradox where both games are changed slightly, so that they only differ in their information structures. In this variant, the new highway exists in both games, but is closed due to construction with prior probability .1. We assume that the cost of 4 minutes is incurred by a player if they try to go down the new highway both if they are successful and get through to B, or if they are blocked by the new construction and have to return to A. In addition, in both games all players are informed about whether the new highway is open, but via a noisy signal (e.g. via a newspaper article from several days before saying when the new highway is schedule to open). The difference between the games is the amount of noise in that signal.

In the new first game, the signal is completely noisy, providing no information at all about whether the new highway is open. So the players have to decide their strategy based purely on the prior probability of .1 that the new highway is open. It is straightforward to verify that in this case, the Nash equilibrium is for half of the players to choose Start-A-End, half to choose Start-B-End, and none to try to go down the new highway. The resultant travel time per player is 65 minutes.

In contrast, in the new version of the second game, the signal is noise-free. So with probability .1 the new highway is open, all players know that, and therefore all take the new highway, for a total travel time of 84 minutes. With probability .9 the new highway is closed, all players know that, and therefore none try to take the new highway, for a total travel time of 65 minutes. So the expected travel time per player in the new second game is \( \frac{1}{2}(84) + \frac{9}{10}(65) \).

So the extra information that is available in the new second game, but not in the new first game, hurts all players.

Note that in contrast to the leader-follower games and noisy signaling games analyzed above, in our variant of Braess’ paradox the extra information concerns a move of nature, not the move of some other player. Nevertheless, our results on negative value of information still apply. In particular, for any particular player \( i \) in that game, and any precise choice for information-theoretic function \( f \), Prop. 2 tells us that there are other directions \( \delta \theta \) in which we infinitesimally reduce the noise in the signal about the state of the highway so that information increases and expected utility for player \( i \) goes up. Prop. 4 then tell us that this property is generic, i.e., it is true for almost all utility functions that differ only slightly from the ones in Ex. 2).

This may have important real-world applications. Computer routing networks (e.g., networks that route communication packets, jobs, etc.), are typically run in a distributed fashion where each router adaptively modifies its behavior to try to optimize its own “utility function” based on noisy signals it receives concerning the state of the overall network. In addition to arising in human transportation traffic networks, Braess’ paradox often arises in such computer routing networks (Roughgarden and Tardos, 2002). This

---

\(26\)In particular, if any of the players in this strategy profile who choose Start-A-End were to change their choice to try to go down the new highway, their new expected travel time would be \( \frac{1}{2}(24 + 20.01) + \frac{9}{10}(24 + 45) = 66.5 \), which is greater than 65, their travel time if they stick with their original strategy.
has led to a large body of work on how to redesign the adaptive algorithms used by the routers to avoid the paradox and associated loss of expected utility (Shoham and Leyton-Brown, 2009; Wolpert and Tumer, 2002). Our analysis suggests that it should be possible to avoid Braess’ paradox — and indeed to increase the expected utility — without redesigning the routing algorithms, but instead changing the signals the routers receive concerning the state of the network.

8. FUTURE WORK

In this paper we primarily used our geometric framework to analyze the relationship between changes in information and associated changes in expected utility. However the framework is far more broadly applicable. It can be used to analyze the relationship between expected utilities and any function \( f(\theta) \) that depends on the parameters specifying the game. \( f \) is not restricted to be an information-theoretic function.

As a particularly simple illustration of this breadth of the applicability of our framework, we can use it to analyze the relationship between expected utility and a function \( f(\theta) \) that simply returns one of the components of \( \theta \). This analysis reveals what might be called scenarios with “negative value of utility”:

**Example 3** This example concerns a simultaneous move game of two players, who have two possible moves each. The bimatrix of the game is as follows:

<table>
<thead>
<tr>
<th></th>
<th>( L )</th>
<th>( R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T )</td>
<td>(1,4)</td>
<td>(4,0,1)</td>
</tr>
<tr>
<td>( B )</td>
<td>(2,2)</td>
<td>(3,3)</td>
</tr>
</tbody>
</table>

where \( \theta \geq 0 \) is a parameter adjusting the utility of the row player.

With \( \theta < 1 \) this game has no pure strategy Nash equilibrium. The unique Nash equilibrium of this game is a mixed equilibrium with the row player playing top with \( p_T = \frac{1}{4} \) and the column player playing left with \( p_L = \frac{1-\theta}{2-\theta} = 1 - \frac{1}{2\theta} \). Since the column player is indifferent between \( L \) and \( R \) at the equilibrium, the expected utility \( V_{col} \) is independent of \( \theta \). While the row player has an expected utility of \( V_{row} = \frac{4-2\theta}{2-\theta} \).

Now, \( \frac{\partial V_{row}}{\partial \theta} = \frac{1}{(2-\theta)^2} \) is strictly positive which means that the row player prefers \( \theta \) to increase. At the same time, this reduces her utility for the outcome \( T, R \) and thus, one could say that the row player has a negative value of utility.\(^{27}\)

As Ex. 3 illustrates, our framework allows us to quantify the value of any infinitesimal change in the change in any function \( f(\theta) \) induced by an infinitesimal change in the game parameter. Indeed, we can evaluate such a value even if the induced change in \( f(\theta) \) is indirect, arising via the effect of the change in \( \theta \) on the player strategy profile that is mediated by the equilibrium concept (as in the leader-follower game from Sec. 7 and as in Ex. 2).

\(^{27}\)As an aside, this phenomenon can be exploited by an external party who can enter a publicly visible binding contract with Row under which Row must pay the tax to the external party, to the benefit of both Row and that external party. This is an example of an external party “mining” a game (Bono and Wolpert, 2014).
This breadth of quantities whose “value” can be evaluated using our framework allows us to analyze the trade-offs, inherent in a game, among multiple changes to the parameters specifying that game and/or functions of those parameters. These trade-offs can be used to calculate marginal rates of substitution, in which we compare the values assigned by a single player to different changes in the game parameter. Assuming transferrable utility, they can also be used to compare the values assigned by different players to the same change in the game parameter. Under that assumption they can even be used to compare the values assigned by different players to different changes in the game parameter. For example, with our framework we can quantify the relationship between the value to player  \( i \) of changes to the information capacity of one information channel in a game, and the value to a separate player  \( j \neq i \) of changes to  \( i \)’s utility function in that game. Similarly, by considering the QRE rationality parameter  \( \beta \) as a component of the game parameter  \( \theta \), we can study the value of rationality and its relations with values of other quantities. This allows us to do things like quantify the relationship between extra rationality by player  \( i \) and extra information available to that player. Colloquially, “\( i \)’s knowing this much more about certain utility-relevant quantities before they make their move is equivalent to their being this much smarter when they make that move”. Future work involves making a detailed investigation of these kinds of trade-offs.

Another issue we intend to investigate in future work is “second-order effects”. Changing the parameter of a game  \( \theta \) infinitesimally can affect the value of every function  \( g : \theta \in \Theta \rightarrow \mathbb{R} \), not just functions like expected utilities, mutual information, etc. In particular, this is true if  \( g \) is the differential value of some  \( f(\theta) \) evaluated at  \( \theta' \). Changing  \( \theta \) will not just change the expected utility of player  \( i \) and what  \( f(\theta) \) is; it will also change the differential value to player  \( i \) of changes to  \( f \). As an example, changing the conditional distribution specifying one information channel in a game will in general change the differential value of a different information channel. In future work we hope to investigate these second order effects and whether they depend on “second-order” properties of the geometric structure of the game, like the Ricci curvature tensor of the Fisher metric.

Since in general games have multiple equilibria, it is often impossible to infinitesimally change the game parameter in way that is Pareto-optimal simultaneously for all equilibria. In other future work, we intend to investigate the generalizations of Prop. 5 that determine when such changes are possible.

Finally, our framework provides important new capabilities for policy making or mechanism design, by providing guidance to a regulator external to a game on how to modify the components of the game parameter vector that are under their control. As discussed above, in games with multiple equilibria — arguably the majority of real-world games of interest to a regulator — it makes sense for the regulator to determine which equilibrium the players have adopted for a current value of  \( \theta \) simply by observing how the players are behaving. Using our framework, it may be possible for that regulator to gradually change  \( \theta \) from that starting value, to move the equilibrium down the branch that has the current equilibrium, to where it intersects with another branch, and then guide the equilibrium along that second branch, back to the starting  \( \theta \). In this way the regulator may be able to gradually change player behavior to go from a cur-
rent equilibrium of a game specified by $\theta$, to a different, Pareto-superior equilibrium for that same $\theta$. In some situations, the regulator should even be able to design that trajectory through the space of game parameter vectors so that each infinitesimal change to $\theta$ along that trajectory is Pareto-improving. (See (Wolpert, Harre, Olbrich, Bertschinger, and Jost, 2012) for an example of this kind of approach.)

9. CONCLUSION

In this paper we introduce a new framework to study the value of information in arbitrary noncooperative games. Our starting point is the well-established concept of marginal utility of a good in a decision problem. We present a very natural way to generalize this, to marginal utility to a player in a noncooperative game of an arbitrary function of a parameter specifying that game. Interestingly, this generalization forces us to introduce a metric over the space of game parameters. In this way we show that geometry is intrinsic to noncooperative game theory, with each game specifying its own associated Riemannian manifold.

In parallel with this analysis, we consider the issue of how best to quantify economically relevant aspects of the information structure in that game. We argue that mutual information is a natural way to do this, using very simple considerations grounded in economics. As we discuss, such a (cardinal) quantification of information also has several advantages over the partial orders commonly used in the past to investigate the role of information in games.

We then combine our two analyses into a unified framework. We do this by taking the “function of a parameter specifying a game” from the first analysis to be the mutual information that was motivated in our second analysis. This combination of our two analyses reveals how a game’s geometry governs the relation between changes to its information structure and changes to the expected utilities of the players.

We then use our framework to derive general conditions for the existence of negative value of information. In particular, we show that for almost any game, there are changes to the information structure of the game that both increase the information available to any particular player in that game and hurt that player. We then extend our analysis to characterize the set of games where there are such changes that simultaneously increase the information available to all players, while hurting all players. We illustrate our framework with computational analyses of a single-player decision scenario as well as a two player leader-follower game.

Finally, we note that our framework can be applied to analyze the effects of arbitrary changes to a game, not just changes to its information structure. As a particularly simple example, we construct a game that has “negative value of utility”, in which the expected value of a player’s utility $u$ increases when we change the game by applying a monotonically decreasing transformation to $u$. More generally, the breadth of applicability of our framework allows us to analyze marginal rates of substitution of different aspects of an information structure, of the utility functions of the players, or of any other parameters specifying the game the players are engaged in.
Acknowledgments

NB acknowledges support by the Klaus Tschira Stiftung. The research of JJ was supported by the ERC Advanced Investigator Grant FP7-267087. DHW acknowledges support of the Santa Fe Institute.

APPENDIX: REVIEW OF CONIC HULLS

In our analysis below we shall work in some tangent space of a parameter manifold that is equipped with the Fisher metric \( \langle \cdot, \cdot \rangle \). Now, whenever we have a scalar product on such a (finite dimensional) vector space, we can perform a linear transformation of that vector space to turn that scalar product into the standard Euclidean one. Thus, in our analysis below, we shall essentially be doing elementary Euclidean geometry, just in a different coordinate system.\(^{28}\) This will be reflected in the fact that simple graphical pictures and intuition can be used to understand many of our results.

We start by reviewing some simple linear algebra in Euclidean space, that is, \( \mathbb{R}^d \) equipped with the Euclidean scalar product \( \langle \cdot, \cdot \rangle \). For a nonzero vector \( v \in \mathbb{R}^d \), the hyperplane \( H_0(v) = \{ w : \langle v, w \rangle = 0 \} \) separates \( \mathbb{R}^d \) into two halfspaces \( \mathcal{H}_+(v) (\mathcal{H}_-(v) \text{ resp.}) \equiv \{ w : \langle v, w \rangle > 0 (< 0 \text{ resp.}) \} \), and for two vectors \( v_1, v_2 \) that are not positively collinear, the associated halfspaces overlap, e.g.,

\[
H_+(v_1) \cap H_-(v_2) \neq \emptyset.
\]

When we have several nonzero vectors \( v_i \), their conic hull (Boyd and Vandenberghe, 2003) is defined as

\[
\text{Con}(\{ v_i \}) \equiv \left\{ \sum_i \alpha^i v_i : \alpha^i \geq 0 \forall i \right\}
\]

Note that \( C = \text{Con}(\{ v_i \}) \) is a cone since whenever \( v \in C \), it follows that \( kv \in C \) for \( k > 0 \). Note also that this cone is convex because whenever \( v, w \in C \), then \( \lambda v + (1 - \lambda)w \in C \) for \( 0 \leq \lambda \leq 1 \).

A cone \( C \) is called pointed if it does not contain any bi-infinite straight line. The following will be used in the sequel\(^{29}\):

**Lemma 6**  
If the vectors \( v_i \) are linearly independent, then \( \text{Con}(\{ v_i \}) \) is pointed.

Of course, the converse of Lemma 6 does not hold in general, because for all \( v \in \text{Con}(\{ v_i \}) \), \( \text{Con}(\{ v_i \}) = \text{Con}(\{ v_i \} \cup v) \).

\(^{28}\) The nonlinear nature of the Fisher or any other Riemannian metric only comes into play when we look at the tangent spaces of several points simultaneously. What linear coordinate transformation turns the Fisher metric into the Euclidean one will depend on the particular tangent space in which we are working for a particular \( \theta \), and in general, there will be no such transformation that works for all tangent spaces simultaneously.

\(^{29}\) For the sake of space we omit the proof of this and the following basic properties.
We will also need to use the definition that the dual to a conic hull $\text{Con}(\{v_i\})$ is

$$(34) \quad \text{Con}(\{v_i\})^\perp \equiv \bigcap_i H_-(v_i).$$

Equivalently,

$$\text{Con}(\{v_i\})^\perp = \{v : \langle v, v_i \rangle < 0 \text{ for all } i\}.$$  

The following elementary property relating conic hulls and pointedness is used in the sequel.

**Lemma 7**  \textit{$\text{Con}(\{v_i\})$ is not pointed $\Rightarrow [\text{Con}(\{v_i\})]^\perp = \emptyset \Rightarrow$ the vectors $v_i$ are linearly dependent.}

For any single vector $v$, $\text{Con}(v)^\perp = H_-(v)$. In general, when we enlarge the set of vectors $v_i$, $\text{Con}(\{v_i\})$ gets larger whereas the dual conic hull $\text{Con}(\{v_i\})^\perp$ becomes smaller. More precisely,

**Lemma 8**  \textit{Let $C_1, C_2$ be nonempty convex cones whose duals are nonempty. Then}

$$(36) \quad C^\perp \subseteq C$$

and

$$(37) \quad C_1 \subset C_2 \quad \Leftrightarrow \quad C_2^\perp \subset C_1^\perp.$$ 

The requirement that the duals be non-empty is crucial in this result. For example, if the second part of Lemma 8 held for empty $C^\perp$, then we would have $C^\perp = \emptyset \Rightarrow C' \subset C$ for any conic hull $C' \subseteq \mathbb{R}^n$. The only way this could be would be if $C = \mathbb{R}^n$. However for any vector $v \in \mathbb{R}^n$ with $n > 2$, $[\text{Con}(\{v, -v\})]^\perp = \emptyset$, even though $[\text{Con}(v, -v)] \neq \mathbb{R}^n$.

**APPENDIX: BASIC CONCEPTS OF DIFFERENTIAL GEOMETRY**

In this appendix, we provide an introduction to the basic concepts of differential geometry as needed and utilized in the main text. References include the monographs (Ay, Jost, Lê, and Schwachhöfer, to appear; Jost, 6th edition, 2011).

Classical differential geometry works with coordinate representations of geometric objects and the transformations of those representations under coordinate changes. The geometric objects themselves are invariantly defined, but their coordinate representations are not, and in order to resolve this tension, the tensor calculus has been developed.

We start with some conventions:
1. Einstein summation convention

\[ a^i b_i := \sum_{i=1}^{d} a^i b_i \]

The content of this convention is that a summation sign is omitted when the same index occurs twice in a product, once as an upper and once as a lower index. The conventions about when to place an index in an upper or lower position will be given subsequently. One aspect of this, however, is

2. When \( G = (g_{ij}) \) is a metric tensor (a notion to be explained below), with with indices \( i, j \) ranging from 1 to \( d \), the inverse metric tensor is written as \( G^{-1} = (g^{ij}) \), that is, by raising the indices. In particular, the fact that the product of a matrix and its inverse is the identity matrix turns into

\[ g^{ij} g_{jk} = \delta^i_k := \begin{cases} 1 & \text{when } i = k \\ 0 & \text{when } i \neq k, \end{cases} \]

the so-called Kronecker symbol.

3. Combining the previous rules, we obtain more generally

\[ v^i = g^{ij} v_j \text{ and } v_i = g_{ij} v^j. \]

A (finite dimensional) manifold \( M \) is locally modeled after \( \mathbb{R}^d \). Thus, locally, it can be represented by coordinates \( \theta = (\theta^1, \ldots , \theta^d) \) taken from some open subset of \( \mathbb{R}^d \). That is, while their global topology may be intricate, local patches of a manifold can be represented by coordinates taken from an open set in \( \mathbb{R}^d \). These coordinates, however, are not canonical, and we may as well choose other ones, \( \eta = (\eta^1, \ldots , \eta^d) \), with \( \theta = F(\eta) \) for some homeomorphism \( F \). When the manifold \( M \) is differentiable, we can cover it by local coordinates in such a manner that all such coordinate transitions are diffeomorphisms, that is, bijective maps that are differentiable and whose inverses are differentiable as well. For simplicity, by “differentiable” we shall mean “infinitely often differentiable” in the sequel. Again, the choice of coordinates is non-canonical.

The basic content of classical differential geometry then is to investigate how various expressions representing objects on \( M \) like tangent vectors transform under coordinate changes. Here and in the sequel, all objects defined on a differentiable manifold will be assumed to be differentiable themselves. This is checked in local coordinates, but since coordinate transitions are diffeomorphic, the differentiability property does not depend on the choice of coordinates.

First of all, we can consider differentiable functions \( \phi \). Their values are, of course, independent of the choice of coordinates, that is, if \( \theta = F(\eta) \), then \( \phi(\theta) = \phi(F(\eta)) \).

Next, there are the tangent vectors. A tangent vector for \( M \) at some point represented by \( \theta_0 \) (in local coordinates \( \theta \)) is an expression of the form

\[ v = v^i \frac{\partial}{\partial \theta^i}. \]
this means that it operates on a function \( \phi(\theta) \) in our local coordinates as

\[
(42) \quad v(\phi)(\theta_0) = \left. \frac{\partial \phi}{\partial \theta} \right|_{\theta = \theta_0}.
\]

The tangent vectors at \( x \in M \) form a vector space, called the tangent space \( T_xM \) of \( M \) at \( x \). The question then is how the same tangent vector is represented in different local coordinates \( \eta \) with \( \theta = F(\eta) \) as before. The answer comes from the requirement that the result of the operation of the tangent vector \( v \) on a function \( \phi \), \( v(\phi) \), should be independent of the choice of coordinates. Applying here and in the sequel always the chain rule, this yields

\[
(43) \quad v = v^i \frac{\partial}{\partial \eta^i}.
\]

Thus, the coefficients of \( v \) in the \( \eta \)-coordinates are \( v^i \frac{\partial}{\partial \eta^i} \). This is verified by the following computation

\[
(44) \quad v \frac{\partial \eta^i}{\partial \theta^j} \frac{\partial}{\partial \eta^k} \phi(F(\eta)) = v \frac{\partial \eta^i}{\partial \theta^j} \frac{\partial \theta^j}{\partial \eta^k} = v \frac{\partial \eta^i}{\partial \theta^j} \frac{\partial \theta^j}{\partial \eta^k} = v^i \frac{\partial \phi}{\partial \theta^j}.
\]

as required.

A vector field then is defined as \( v(x) = v^i(x) \frac{\partial}{\partial \eta^i} \), that is, by having a tangent vector at each point \( x \) of \( M \). As indicated above, we assume here that the coefficients \( v^i(x) \) are differentiable.

Returning to a single tangent vector, \( v = v^j \frac{\partial}{\partial \theta^j} \) at some point \( x_0 \), we consider a covector \( \omega = \omega_i d\theta^i \) at this point as an object dual to \( v \), with the rule

\[
(45) \quad d\theta^j \frac{\partial}{\partial \theta^j} = \delta^j_i
\]

yielding

\[
(46) \quad \omega_i d\theta^j (v^j \frac{\partial}{\partial \theta^j}) = \omega_i v^j \delta^i_j = \omega_i v^i.
\]

This expression depends only on the coefficients \( v^i \) and \( \omega_i \) at the point under consideration and does not require any values in a neighborhood. We can write this as \( \omega(v) \), the application of the covector \( \omega \) to the vector \( v \), or as \( v(\omega) \), the application of \( v \) to \( \omega \).

We have the transformation behavior

\[
(47) \quad d\theta^i = \frac{\partial \theta^j}{\partial \eta^i} d\eta^j
\]

required for the invariance of \( \omega(v) \). Thus, the coefficients of \( \omega \) in the \( \eta \)-coordinates are given by the identity

\[
(48) \quad \omega_i d\theta^i = \omega_i \frac{\partial \theta^i}{\partial \eta^i} d\eta^i.
\]
The transformation behavior of a tangent vector as in (43) is called contravariant, the opposite one of a covector as (48) covariant.

A 1-form then assigns a covector to every point in $M$, and thus, it is locally given as $\omega_i(x) d\theta^i$.

Having derived the transformation of vectors and covectors, we can then also determine the transformation rules for other tensors. A lower index always indicates covariant, an upper one contravariant transformation.

The metric tensor, written as $g_{ij} d\theta^i \otimes d\theta^j$, with $g_{ij} = \langle \frac{\partial}{\partial \theta^i}, \frac{\partial}{\partial \theta^j} \rangle$ being the product of those two basis vectors, operates on pairs of tangent vectors. It therefore transforms doubly covariantly, that is, becomes

$$g_{ij}(F(\eta)) \frac{\partial \theta^i}{\partial \eta^\alpha} \frac{\partial \theta^j}{\partial \eta^\beta} d\eta^\alpha \otimes d\eta^\beta.$$  

We require that the metric tensor be positive definite and symmetric, that is,

$$g_{ij} = g_{ji}$$

The function of the metric tensor is to provide a Euclidean product of tangent vectors,

$$\langle v, w \rangle = g_{ij} v^i w^j$$

for $v = v^i \frac{\partial}{\partial \theta^i}, w = w^i \frac{\partial}{\partial \theta^i}$. As a check, in this formula, $v^i$ and $w^i$ transform contravariantly, while $g_{ij}$ transforms doubly covariantly so that the product as a scalar quantity remains invariant under coordinate transformations.

A differentiable manifold equipped with such a metric tensor is called a Riemannian manifold.

Vectors $v, w \in T_x M$ with $\langle v, w \rangle = 0$ are called orthogonal. The norm of a vector $v \in T_x M$ is defined as

$$|v| = \sqrt{\langle v, v \rangle}.$$  

For a function $\phi$, we have its differential

$$d\phi = \frac{\partial \phi}{\partial \theta^i} d\theta^i,$$

a 1-form; this depends on the differentiable structure, but not on the metric. The gradient of $\phi$, however, involves the metric; it is defined as

$$\text{grad} \phi = g^{ij} \frac{\partial \phi}{\partial \theta^i} \frac{\partial}{\partial \theta^j}.$$  

The gradient of a function $\phi$ is orthogonal to the level hypersurfaces $\phi = c$, in the following sense. When $v \in T_x M$ is tangent to such a level hypersurface, it satisfies

$$v(\phi) = v^i \frac{\partial \phi}{\partial \theta^i} = 0.$$
When \( \mathbf{v} \) then satisfies (55), we have

\[
\langle \nabla \phi, \mathbf{v} \rangle = g^{ij} \frac{\partial \phi}{\partial \theta^i} v^j = 0,
\]

that is, \( \nabla \phi \) and \( \mathbf{v} \) are orthogonal.

We also need the formula for the product of the gradients of two functions \( \phi, \psi \),

\[
\langle \nabla \phi, \nabla \psi \rangle = g^{ij} \frac{\partial \phi}{\partial \theta^i} g^{k\ell} \frac{\partial \psi}{\partial \theta^\ell} = g_{ij} \frac{\partial \phi}{\partial \theta^j} \frac{\partial \psi}{\partial \theta^i}.
\]

In differential geometry, one also needs a notion of second derivatives. The first derivative of a function is easy to compute in local coordinates; it yields a one-form, or dually, a tangent vector. When we wish to compute the second derivatives of a function, we would thus have to compute the first derivatives of a tangent vector field. But how could this be done, as the tangent spaces at different points are not canonically identified? This is in contrast to the Euclidean situation, where any tangent space can be parallelly moved into any other one. Therefore, one also needs a notion of parallel transport in Riemann geometry upon which to build the notion of the derivative of a tangent vector field. This leads to the covariant derivative of Levi-Civita which we now introduce. Let \( (\theta^1, \ldots, \theta^d) \) be local coordinates, as usual. The covariant derivative \( \nabla \) satisfies

\[
\nabla_{\frac{\partial}{\partial \theta^i}} \frac{\partial}{\partial \theta^j} = \Gamma^k_{ij} \frac{\partial}{\partial \theta^k}
\]

for all \( i, j \)

with

\[
\Gamma^i_{jk} = \frac{1}{2} \left( g^{il} g_{jl,k} + g_{lk,j} - g_{lk,j} g^{il} g_{jl,k} \right),
\]

where

\[
(g^{ij})_{i,j=1,\ldots,d} = (g_{ij})^{-1} \quad \text{(i.e. } g^{ij} g_{ij} = \delta_{ij})
\]

and

\[
g_{jk} = \frac{\partial}{\partial \theta^j} g_{jk}.
\]

The expressions \( \Gamma^i_{jk} \) are called the Christoffel symbols. \( \nabla \) is then extended to all vector fields \( \mathbf{v} = v^j \frac{\partial}{\partial \theta^j} \) via the product rule

\[
\nabla_{\frac{\partial}{\partial \theta^i}} \mathbf{v} = \frac{\partial v^j}{\partial \theta^i} \frac{\partial}{\partial \theta^j} + v^j \nabla_{\frac{\partial}{\partial \theta^i}} \frac{\partial}{\partial \theta^j}.
\]

Moreover,

\[
\nabla_{w^j \frac{\partial}{\partial \theta^i}} = w^j \nabla_{\frac{\partial}{\partial \theta^i}} v^j \frac{\partial}{\partial \theta^j}.
\]
The covariant derivative $\nabla$ is set up in such a way that it is compatible with the metric, in the sense it satisfies the product rule

$$ u \langle v, w \rangle = \langle \nabla_u v, w \rangle + \langle v, \nabla_u w \rangle $$

for all vector fields $u, v, w$.

The covariant derivative then allows us to introduce the Riemannian version of the Hessian of a function. The Hessian of a differentiable function $f : M \to \mathbb{R}$ on a Riemannian manifold $M$ is

$$ \nabla df. $$

Here, we have $df = \frac{\partial f}{\partial \theta^i} d\theta^i$ in local coordinates, hence

$$ \nabla_{\partial_i} df = \left( \frac{\partial^2 f}{\partial \theta^i \partial \theta^j} - \frac{\partial f}{\partial \theta^k} \Gamma^k_{ij} \right) d\theta^i \otimes d\theta^j, $$

i.e.

$$ \nabla df = \left( \frac{\partial^2 f}{\partial \theta^i \partial \theta^j} - \frac{\partial f}{\partial \theta^k} \Gamma^k_{ij} \right) d\theta^i \otimes d\theta^j. $$

We also have

$$ \nabla df(v, w) = \langle \nabla \text{grad}(f), w \rangle, $$

since $w(f) = \langle \text{grad} f, w \rangle$ and thus

$$ v(w(f)) = v(\text{grad}(f), w) $$

$$ = \langle \nabla \text{grad}(f), w \rangle + \langle \text{grad}(f), \nabla w \rangle $$

$$ = \langle \nabla \text{grad}(f), w \rangle + \langle \nabla w \rangle(f), $$

and applying (62) to $v$ and $w$ yields

$$ \nabla df(v, w) = v(w(f)) - (\nabla w)(f). $$

In local coordinates, with $e^i = \frac{\partial}{\partial \theta^i}$, we then have the components of the Hessian of $f$ given as

$$ D_{ij}f = \nabla df(e^i, e^j) = \langle \nabla \text{grad}(f), e^i \rangle. $$

Note that the Hessian is symmetric in the sense that

$$ D_{ij}f = D_{ji}f $$

for all $i, j$.

The Hessian $D_{ij}$ is covariant, but (65) indicates that we can also introduce the tensor $\nabla \nabla f$ with components

$$ (\nabla \nabla f)^i_j = g^{ij} D_{ij}f. $$
Let now \([a, b]\) be a closed interval in \(\mathbb{R}\), \(\gamma : [a, b] \rightarrow M\) a differentiable curve. The length of \(\gamma\) then is defined as

\[
L(\gamma) := \int_{a}^{b} \left\| \frac{d\gamma}{dt}(t) \right\| dt
\]

The length of a curve can be computed in local coordinates. Working with the coordinates \((x^1(\gamma(t)), \ldots, x^d(\gamma(t)))\) we use the abbreviation

\[
\dot{x}^i(t) := \frac{d}{dt}(x^i(\gamma(t))).
\]

Then

\[
L(\gamma) = \int_{a}^{b} \sqrt{g_{ij}(x(\gamma(t)))\dot{x}^i(t)\dot{x}^j(t)} dt
\]

With this concept of curve length, we can define the distance function in a Riemannian manifold \(M\) by

\[
\text{dist}(p, q) = \inf\{L(\gamma) : \gamma : [0, 1] \rightarrow M, \gamma(0) = p, \gamma(1) = q\}.
\]

REFERENCES


Bono, J. W., and D. H. Wolpert (2014): “Game Mining: How to Make Money from those about to Play a Game Game Mining: How to Make Money from those about to Play a Game Game Mining: How to Make Money from those about to Play a Game,” Advances in Austrian Economics.


