Superconditioning*

Simon M. Huttegger

January 2024

Abstract

When can a shift from a prior to a posterior be represented by conditionalization? A well-known result, known as "superconditioning" and going back to work by Diaconis and Zabell, gives a sharp answer. This paper extends the result and connects it to the reflection principle and common priors. I show that a shift from a prior to a set of posteriors can be represented within a conditioning model if and only if the prior and the posteriors are connected via a general form of the reflection principle. Common priors can be characterized by principles that require a certain kind of coherence between distinct sets of posteriors. I discuss the implications these results have for diachronic and synchronic modes of updating, learning experiences, the common prior assumption of game theory, and time-slice epistemology.

1 Introduction

According to the reflection principle, a Bayesian agent's prior probability is a weighted average of her future probability assignments, with weights representing the agent's beliefs about future probabilities.[1] It's widely agreed that the reflection principle is not in general tenable as a rationality constraint on belief change. The common thread in many problem cases is that changes in belief need not come about by responding to new evidence. Rather, probabilities may shift because the agent is subject to forgetting, manipulation, or misinformation. For that reason, the reflection principle is often restricted to situations where information is gained.[2]

But what does it mean to gain information? Bayesians usually represent an agent's credal state by a probability space: a triple consisting of a set of worlds, a set of propositions (subsets of worlds), and a probability assignment to propositions. If the agent learns a proposition to be true, the Bayesian norm of updating says to conditionalize probabilities on that propositions. In this case, there is an obvious answer to the question of whether information is gained: just consider the conditioning proposition that expresses the content of the presumed learning experience. But the main point of the reflection principle is that it also applies in situations where an agent does not

*Thanks to Brad Armendt, Melissa Fusco, Bill Harper, Daniel Herrmann, Calum McNamara, Eric Pacuit, Jan-Willem Romeijn, Brian Skyrms, Snow Zhang, Kevin Zollman, and two anonymous referees for helpful comments on earlier versions of this paper. Parts of this work were presented at the Conference Combining Probability and Logic (Munich 2021), a colloquium at the University of Groningen (2022), the Workshop on Learning, Randomness and Complexity (CMU 2022), and the Philosophy of Science Association meeting in Pittsburgh (2022). I would like to thank the audiences at these venues for their feedback.

[1] See Goldstein [1983], van Fraassen [1984], and Skyrms [1990].

[2] Jeffrey [1988], Skyrms [1990], and Huttegger [2017].

update by conditioning on a proposition in her probability space. The direct route of arguing that something is learned by saying what is learned is blocked.

My answer to the problem of capturing updating in repsonse to evidence is based on the idea of representing shifts from a prior to posteriors by conditionalization in a more fine-grained probability space, thus connecting cases of opaque learning (no proposition that expresses the content of a learning experience) to a model that resembles the structure of transparent cases of learning (conditionalization). It only traces that structure because the propositions conditionalized upon in the representation are formal idealizations: they need not be propositions in the Bayesian agent's probability space.

The strategy of representing updating by conditionalization is not new. It goes back to Persi Diaconis and Sandy Zabell and has become known as "superconditioning".[3] Diaconis and Zabell consider a shift from a prior to one posterior. I extend their account in two ways. The first extension considers shifts from a prior to a set of posteriors. In order to be representable by conditionalization, such a shift has to satisfy a generalized form of the reflection principle, and vice versa. The second extension considers the question of when shifts from one prior to two, or more, distinct sets of posteriors can be represented by conditionalization simultaneously within one large space. That is possible if and only if the prior is a weighted average of each set of posteriors.

The two extensions are technically straightforward. But I think they do lead to some interesting insights about updating and learning. The second result speaks to some issues that arise for common priors in game theory and for a particular model of time-slice rationality.[4] The first result gives rise to a genuinely diachronic understanding of updating in light reflection. It also answers the question of whether there could in principle be a learning event that prompts a shift. Neither of these points can be adequately addressed within higher-order Bayesian models that involve probabilities about posteriors.

I start with a brief summary of Bayesian updating in §2. After developing and exploring the two extensions of the Diaconis and Zabell result in §3 and §4, I conclude by connecting superconditioning to some other strands in the literature on updating in §5, especially the role of partitions and the contrast between superconditioning and higher-order Bayesian models.

2 Bayesian Updating

Let an agent's credal space be given by a finite set Ω of doxastic possibilities, a Boolean algebra $\mathfrak{B}$ of subsets (propositions) of Ω, and a probability function, P, that assigns probabilities to elements of $\mathfrak{B}$.[5] If A and E are propositions, then $P(A \mid E) = \frac{P(A\&E)}{P(E)}$ is the conditional probability of A given E whenever $P(E) > 0$.

Conditional probabilities are the cornerstone of Bayesian updating. The standard norm for adjusting one's beliefs in response to new evidence is *conditionalization*, of which we consider two versions. The first one takes one evidential proposition, E, as input and says that upon learning E one should adopt $P(A \mid E)$ as one's new probability for any proposition A in $\mathfrak{B}$. The second version takes a partition of evidential propositions, $E_1, \ldots, E_n$, as inputs. It says that upon learning which proposition among $E_1, \ldots, E_n$ is true, one should adopt as one's new beliefs conditional probabilities

[3] Diaconis and Zabell [1982, Theorem 2.1].

[4] See Williamson [2002] and Moss [2015].

[5] The results in this paper hold if Ω is countably infinite. Whether they can be extended to uncountable Ω is an open problem.

given the true member of the partition.

Both versions of conditionalization can be understood *diachronically* as expressing an agent's credal state at two different times. The first version describes an actual update, while the second one describes a number of potential such updates. The first version can be understood, moreover, as a tacit special case of the second one: presumably, upon learning that E is false one should update to conditional probabilities given the negation of E.

Conditionalization requires that there exist evidential propositions in the agent's credal space that fully capture the agent's learning experience. As Richard Jeffrey pointed out, this is too restrictive already in the case of observational learning.[6] In a dimly lit room one might be unable to determine with certainty the color of a piece of furniture. Jeffrey introduced probability kinematics, or Jeffrey conditioning, to model these types of uncertain observational learning events. Instead of having a partition $E_1, \ldots, E_n$ of propositions in $\mathfrak{B}$, one of which will be learned for certain, the agent undergoes a shift in probability assignments for the members of the partition that shies away from allocating posterior probability one to any E_i. None of the E_i describes the learning experience in full detail; rather, the shift as a whole captures the learning experience.

Jeffrey takes the credal space $(\Omega, \mathfrak{B})$ as a psychological model of what's going on in an agent's head. The agent undergoes a learning experience that doesn't correspond to conditioning on any proposition in $\mathfrak{B}$ but does correspond to a Jeffrey shift. Is there a weaker notion of rationality that we can apply here? The same question arises for behavioral interpretations of probabilities. Suppose $(\Omega, \mathfrak{B})$ is an experimenter's model of the agent and that she has measured the agent's credences. The agent updates by Jeffrey conditioning. Have we thus seen proof of irrationality?

The main point here is that Jeffrey conditioning incorporates an element of "black box learning".[7] The prior probability assignment to each element, $P(E_i)$, of the partition is transformed into a posterior probability assignment, $P'(E_i)$, by a process whose inner workings remain hidden. This may be due to the agent being unaware of how she arrives at posterior probability judgments or because an external modeler lacks information about how the agent processes information. Once the posterior probabilities are given, the new probabilities of any other proposition B in $\mathfrak{B}$ is calculated mechanically as a weighted average of conditional probabilities given E_i, with posterior probabilities serving as weights.

Black box learning is not restricted to Jeffrey conditioning. An agent may shift from a prior to a posterior by a less tidy process than taking weighted averages of conditional probabilities over a partition of evidential propositions. In the most extreme case, the shift is represented by pairs of prior and posterior probabilities for any proposition in the credal space.

The obvious question for any kind of learning that involves some black box aspect is this: Why should one think of a black box shift as being due to new information rather than to arbitrary factors? In contrast to conditionalization, there is no proposition in the credal space that describes the learning experience. The very generality of the black box model makes it difficult to distinguish between epistemically legitimate and capricious shifts.

Some commentators think of the reflection principle as a constraint on epistemically legitimate black box shifts.[8] There are several versions of the reflection principle.[9] The version most relevant

[6] Jeffrey [1965].

[7] I borrow this term from Skyrms [1990].

[8] Jeffrey [1988], Skyrms [1990], and Huttegger [2017]. There are many counterexamples in the literature that show quite clearly that the reflection principle fails for general shifts. See, *inter alia*, Levi [1987], Talbott [1991], Bovens [1995], Arntzenius [2003], and Briggs [2009].

[9] Reflection principles were introduced by Goldstein [1983] and van Fraassen [1984]. See Huttegger [2013] for an overview and logical connections between different versions of the principle.

for us is the following. Consider a proposition, A, in the agent's credal space. Suppose that $P_1(A), \ldots, P_n(A)$ are the agent's possible posterior probabilities. Let A_i be the proposition that the agent's posterior probability is given by P_i, $i = 1, \ldots, n$. Then, according to the reflection principle,

$$P(A) = \sum_i P_i(A) P(A_i). \tag{1}$$

In words: the prior is a weighted average of the posteriors. The weights are given by prior beliefs about posteriors.

I shall say more about the reflection principle and what it has to do with learning in §5. In the meantime, note the higher-order structure inherent in the reflection principle (1). Each state is required to determine a posterior probability so that priors over posteriors are well defined. This stands in contrast to conditionalization. Conditionalization can be thought of as diachronic. The reflection principle has a synchronous character: the agent entertains beliefs about her own posterior probabilities. She is pondering plans, or has certain dispositions, to update in accordance with reflection. Whether or not she will *actually* adopt the posteriors is a different story. That's not the case for conditionalization. There is no presumption that the agent plans, or is disposed to, update by conditioning on an evidential proposition.[10]

The next section develops an alternative route to the reflection principle. The basic idea, which goes back to the work of Diaconis and Zabell, is to represent probability shifts within a *superconditioning space*.[11] By taking Diaconis and Zabell's framework a few steps further, we shall see that a generalized version of the reflection principle applies to updating on new evidence in the diachronic mode.

3 Superconditioning: Single Shift

Let $(\Omega, \mathfrak{B}, P)$ and $(\Omega, \mathfrak{B}, P')$ be two credal spaces. In following Diaconis and Zabell, we say that P' *comes from* P *by conditionalization* if there exists a probability space $(\Lambda, \mathfrak{A}, Q)$ of the following kind: for each $\omega \in \Omega$ there is an $E_\omega \in \mathfrak{A}$, and there exists an $E \in \mathfrak{A}$ with $Q(E) > 0$ such that

(i) $Q(E_\omega) = P(\omega)$ and

(ii) $Q(E_\omega \mid E) = P'(\omega)$.

If P' comes from P by conditionalization, the shift can be represented by conditionalization on a proposition E in some probability space even if in $\mathfrak{B}$ there is no such proposition. One may think of $(\Omega, \mathfrak{B})$ as a model of an agent's "object language", which consists of sentences whose meanings the agent can grasp. The space $(\Lambda, \mathfrak{A})$, then, is an idealization for representing shifts in terms of propositions that may go beyond the agent's conceptual framework.

Diaconis and Zabell proved the following theorem:[12]

Theorem 1. *P' comes from P by conditionalization if and only if there exists a $B \geq 1$ such that, for all $\omega \in \Omega$,*

$$P'(\omega) \leq B P(\omega). \tag{2}$$

[10] Despite there being no such presumption, conditionalization is compatible with update plans. Conditional probabilities can, and often are, understood as random variables in a credal space, which requires values of conditional probabilities to be specified at each state of the world.

[11] Diaconis and Zabell [1982].

[12] Diaconis and Zabell [1982, Theorem 2.1].

Condition (2) says that the prior, P, and the posterior, P', cannot be too different. If Ω is finite, (2) is equivalent to the following requirement:

$$\text{For all } \omega \in \Omega, \text{ if } P'(\omega) > 0, \text{then } P(\omega) > 0.$$

This is a simple form of the reflection principle. It says that if you think something has positive probability for you tomorrow, it must already have positive probability today: you should not expect to learn things that currently have probability zero.

Can Theorem 1 be extended to something like the reflection principle (1)? Let's consider a shift from a prior, P, to a posterior in the set $P_1, \ldots, P_n$. I'll refer to this as "a shift from a prior to a set of posteriors" in what follows. Let $(\Omega, \mathfrak{B}, P)$ and $(\Omega, \mathfrak{B}, P_i), i = 1, \ldots, n$ be the corresponding credal spaces. We say that the shift from P to $P_1, \ldots, P_n$ *can be embedded in a conditioning model* if there exists a probability space $(\Lambda, \mathfrak{A}, Q)$ with the following features: for every $\omega \in \Omega$ there is an $E_\omega \in \mathfrak{A}$, and there is a partition $E_1, \ldots, E_n \in \mathfrak{A}$ of Λ such that

(i) $Q(E_\omega) = P(\omega)$ and

(ii) If $Q(E_i) > 0$, then $Q(E_\omega \mid E_i) = P_i(\omega)$.

We can again think of the superconditioning space $(\Lambda, \mathfrak{A}, Q)$ as an idealization. The elements of the partition $E_1, \ldots, E_n$ are not assumed to be in the agent's credal space. If a superconditioning space exists, any shift from P to $P_1, \ldots, P_n$ can be thought of as conditionalization on a proposition that may be inaccessible to the agent.

The following theorem shows that a shift is embeddable in a conditioning model if and only if the prior is a weighted average of the posteriors. The proof is a slight modification of Diaconis and Zabell's proof of Theorem 1.

Theorem 2. *The shift from P to $P_1, \ldots, P_n$ can be embedded in a conditioning model if and only if there exist constants $0 \leq B_i \leq 1, i = 1, \ldots, n$ such that $\sum_i B_i = 1$ and*

$$P(\omega) = \sum_i B_i P_i(\omega) \quad \textit{for all } \omega \in \Omega \tag{3}$$

Moreover, in both cases the superconditioning space is such that $Q(E_i) = B_i$.

Proof. Suppose that $P_1, \ldots, P_n$ can be embedded in a conditioning model. Let $B_i = Q(E_i)$. Then $\sum_i B_i = 1$ and for all $\omega \in \Omega$,

$$P(\omega) = Q(E_\omega) = \sum_{i:Q(E_i)>0} Q(E_i)\ Q(E_\omega \mid E_i) = \sum_{i:Q(E_i)>0} Q(E_i)\ P_i(\omega) = \sum_i B_i P_i(\omega).$$

Clearly, $Q(E_i) > 0$ if and only if $B_i > 0$.

Conversely, suppose there are $0 \leq B_i \leq 1$ such that $\sum_i B_i = 1$ and

$$P(\omega) = \sum_i B_i P_i(\omega).$$

Let $\Lambda = \Omega \times \{b_1, \ldots, b_n\}$, $E_\omega = \bigcup_i (\omega, b_i)$ and $E_i = \bigcup_\omega (\omega, b_i)$. Define Q by

$$Q(\omega, b_i) = B_i P_i(\omega).$$

Then, for all $\omega \in \Omega$,

$$Q(E_\omega) = Q\left(\bigcup_i (\omega, b_i)\right) = \sum_i Q(\omega, b_i) = \sum_i B_i P_i(\omega) = P(\omega).$$

Moreover,

$$Q(E_i) = Q\left(\bigcup_\omega (\omega, b_i)\right) = \sum_\omega Q(\omega, b_i) = \sum_\omega B_i P_i(\omega) = B_i,$$

and so the $Q(E_i)$ are well defined. If $Q(E_i) > 0$, then

$$Q(E_\omega \mid E_i) = \frac{Q(E_\omega \cap E_i)}{Q(E_i)} = \frac{Q(\omega, b_i)}{B_i} = \frac{B_i P_i(\omega)}{B_i} = P_i(\omega)$$

for all $\omega \in \Omega$. □

Equation (3) is a generalization of the reflection principle (1).[13] The latter involves priors over posteriors and is a special case of (3) in the following sense. Suppose each state ω in Ω determines a posterior, and let $A_i = \{\omega \in \Omega : P_i \text{ is the posterior at } \omega\}$. Suppose also that $P_i(A_i) = 1$ and $P_j(A_i) = 0$ for all $j \neq i$. This "luminosity" condition, which is often used in discussions of the reflection principle, says that posteriors locate themselves perfectly. In agential terms: the agent's future personae have full introspective access to their beliefs.[14] The reflection principle (1) then follows from (3) since

$$P(A_i) = \sum_j B_j P_j(A_i) = B_i.$$

Somewhat more generally, suppose there exists a partition $A_1, \ldots, A_n$ of Ω such that $P_i(A_i) = 1$ and $P_j(A_i) = 0$ for all $j \neq i$. Then $P(A_i) = B_i$, by the same reasoning, and the agent's beliefs are *as if* worlds determine posteriors and luminosity holds relative to the partition $A_1, \ldots, A_n$.

Importantly, though, the general reflection principle (3) also applies in situations in which there is no such partition, either because states don't specify posteriors or because the agent is confused about posteriors. Whether or not we have priors over posteriors (implicitly or explicitly), the general reflection principle (3) is a necessary and sufficient condition that a shift *could in principle be given by conditioning on propositions.* As a result, the generalized reflection principle (3) can be understood in a genuinely diachronic way, which sets the present account apart from other analyses of the reflection principle (mentioned in the previous section).[15]

What does this tell us about learning? Jeffrey conditioning and black box learning are motivated by the thought that having no proposition in one's credal space that describes the exact content of a learning experience is not irrational. The impossibility of superconditioning, however, is a different matter. In that case, the shift cannot, in principle, be represented by learning a proposition. The lack of being representable by conditioning indicates, at least to a first approximation, that the shift is not supported by a learning event and is thus epistemically irrational. As a consequence, the generalized reflection principle, being equivalent to superconditioning, is also necessary for rational learning. (I shall discuss this issue in more detail in §5.)

[13] The generalized reflection principle was introduced in van Fraassen [1995].

[14] See, e.g., Weisberg [2007] for a critical discussion.

[15] As in the case of conditionalization, the posteriors can be understood synchronically as well.

Let me close this section with two further points. Not everyone agrees with what I just said: that having limited conceptual resources is no sign of irrationality. David Lewis thought of ideally rational agents as having no restrictions of that sort. He maintained, accordingly, that ideally rational updating always proceeds by conditionalization and contrasted conditionalization with less than ideal, bounded ways of updating, such as Jeffrey conditioning.[16] If a Jeffrey shift respects the generalized reflection principle (3), the contrast between ideal and bounded Bayesian agents loses some of its edge. For then superconditioning says that a Jeffrey conditionalizer has an ideal counterpart who updates by conditionalizing on propositions.[17]

A similar issue arises in certain models of cognitive agents, like Brian Skyrms' two-layered model of a probabilistic automaton.[18] According to this model, the automaton's probability assignments shift in response to learning events that need not come about by conditionalization. These shifts proceed at a level that is conceptually accessible to the engineer of the automaton. Skyrms suggests that the high-level shift may really come about by conditionalization at the finer grained level of the automaton's machine language, which is not accessible to the engineer. This model also corresponds loosely to human agents who shift probabilities after, say, perceptual experiences without being able to express them as propositions—but such propositions might exist at a subconscious level. The superconditioning theorem provides a sharp constraint for when this two-layered model of cognitive agents is feasible: shifts must obey the generalized reflection principle.

4 Superconditioning: Multiple Shifts

The superconditioning theorem can be extended to distinct sets of posteriors that come from the same prior. The sets of posteriors may represent the credal states of two or more agents or the credal states of one agent at different times.

Let $(\Omega, \mathfrak{B}, P)$, $(\Omega, \mathfrak{B}, P_{1i}), i = 1, \ldots, n$ and $(\Omega, \mathfrak{B}, P_{2j}), j = 1, \ldots m$, be credal spaces. We say that the shifts from P to $P_{11}, \ldots, P_{1n}$ and $P_{21}, \ldots, P_{2m}$ *can be embedded in a conditioning model* if there exists a probability space $(\Lambda, \mathfrak{A}, Q)$ that has the following structure: for each ω there is an $E_\omega \in \mathfrak{A}$, and there exist two partitions $E_{11}, \ldots, E_{1n}$ and $E_{21}, \ldots, E_{2m}$ of Λ with each element being a member of $\mathfrak{A}$ such that

(i) $Q(E_\omega) = P(\omega)$;

(ii) if $Q(E_{1i}) > 0$ then $Q(E_\omega \mid E_{1i}) = P_{1i}(\omega)$; and

(iii) if $Q(E_{2j}) > 0$ then $Q(E_\omega \mid E_{2j}) = P_{2j}(\omega)$.

[16] See Lewis [1999, p. 404]:

> Richard Jeffrey has suggested that we should respond to experiential evidence not by conditionalizing, but rather by a less extreme redistribution of degrees of belief. Despite appearances, I do not disagree. He and I are considering different cases. My advice is addressed to a severely idealized, superhuman subject who runs no risk of mistaking his evidence, and who therefore can only lose if he hedges against that risk. Jeffrey's advice is addressed to a less idealized, fallible subject who has no business heeding counsels of perfection that he is unable to follow.

As I understand Lewis, an agent never mistaking their evidence requires them to have perfect conceptual resources. In order to mistake or not mistake some evidence, a proposition representing that evidence must be given.

[17] Williamson goes further than Lewis, arguing that the absence of evidential propositions is a serious, and perhaps fatal, shortcoming of Jeffrey conditionalization [Williamson, 2002]. Superconditioning softens that worry, too, insofar as the general reflection principle holds.

[18] Skyrms [1984, pp. 117-18].

The following result is a slight extension of Theorem 2. Generalizing the proof to shifts from one prior to three or more sets of posteriors is straightforward.

Theorem 3. *The shifts from P to $P_{11}, \ldots, P_{1n}$ and $P_{21}, \ldots, P_{2m}$ can be embedded in a conditioning model if and only if there exist constants $B_{11}, \ldots, B_{1n}$ and $B_{21}, \ldots, B_{2m}$ with $0 \leq B_{1i}, B_{2j} \leq 1, i = 1, \ldots, n, j = 1, \ldots, m$ and $\sum_i B_{1i} = 1,\ \sum_j B_{2j} = 1$ such that*

$$P(\omega) = \sum_i B_{1i} P_{1i}(\omega) = \sum_j B_{2j} P_{2j}(\omega).$$

In both cases the superconditioning space is such that $Q(E_{1i}) = B_{1i}$ and $Q(E_{2j}) = B_{2j}$.

Proof. Suppose the shift can be embedded. Then

$$\begin{aligned} P(\omega) &= Q(E_\omega) \\ &= \sum_{i:Q(E_{1i})>0} Q(E_{1i})\ Q(E_\omega \mid E_{1i}) \\ &= \sum_{i:Q(E_{1i})>0} Q(E_{1i})\ P_{1i}(\omega) = \sum_i Q(E_{1i}) P_{1i}(\omega). \end{aligned}$$

Likewise for the other sum. Setting $B_{1i} = Q(E_{1i})$ and $B_{2j} = Q(E_{2j})$ establishes the conclusion.

Conversely, suppose that

$$P(\omega) = \sum_i B_{1i} P_{1i}(\omega) = \sum_j B_{2j} P_{2j}(\omega).$$

Let $\Lambda = \Omega \times \{b_{11}, \ldots, b_{1n}\} \times \{b_{21}, \ldots, b_{2m}\}$. Also, let $E_\omega = \bigcup_{i,j}(\omega, b_{1i}, b_{2j})$, $E_{1i} = \bigcup_{\omega,j}(\omega, b_{1i}, b_{2j})$, and $E_{2j} = \bigcup_{\omega,i}(\omega, b_{1i}, b_{2j})$. Set

$$Q(\omega, b_{1i}, b_{2j}) = \begin{cases} \frac{B_{1i} B_{2j} P_{1i}(\omega) P_{2j}(\omega)}{P(\omega)} & \text{if } P(\omega) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $P(\omega) > 0$. Then

$$Q\left(\bigcup_j (\omega, b_{1i}, b_{2j})\right) = \sum_j \frac{B_{1i} B_{2j} P_{1i}(\omega) P_{2j}(\omega)}{P(\omega)} = B_{1i} P_{1i}(\omega) \sum_j \frac{B_{2j} P_{2j}(\omega)}{P(\omega)} = B_{1i} P_{1i}(\omega).$$

Hence

$$Q(E_\omega) = Q\left(\bigcup_i \bigcup_j (\omega, b_{1i}, b_{2j})\right) = \sum_i Q\left(\bigcup_j (\omega, b_{1i}, b_{2j})\right) = \sum_i B_{1i} P_{1i}(\omega) = P(\omega).$$

If $P(\omega) = 0$, then $Q(E_\omega) = \sum_{i,j} Q(\omega, b_{1i}, b_{2j}) = 0$.

Furthermore, suppose that $B_{1i} > 0$. Then

$$\begin{aligned} Q(E_{1i}) &= Q\left(\bigcup_{\omega}\bigcup_{j}(\omega, b_{1i}, b_{2j})\right) \\ &= \sum_{\omega}\sum_{j} Q(\omega, b_{1i}, b_{2j}) \\ &= \sum_{\omega:P(\omega)>0}\sum_{j} Q(\omega, b_{1i}, b_{2j}) + \sum_{\omega:P(\omega)=0}\sum_{j} Q(\omega, b_{1i}, b_{2j}) \\ &= \sum_{\omega:P(\omega)>0} B_{1i}P_{1i}(\omega) \\ &= B_{1i} \end{aligned}$$

The fourth identity follows since $Q(\omega, b_{1i}, b_{2j}) = 0$ whenever $P(\omega) = 0$. For the last identity observe that, since $P(\omega) = \sum_i B_{1i}P_{1i}(\omega)$ and $B_{1i} > 0$, $P_{1i}(\omega) = 0$ whenever $P(\omega) = 0$; thus $\sum_{\omega:P(\omega)>0} P_{1i}(\omega) = 1$. If $B_{1i} = 0$, then $Q(\omega, b_{1i}, b_{2j}) = 0$ for all ω, j, and so $Q(E_{1i}) = 0 = B_{1i}$.

Hence, if $Q(E_{1i}) > 0$

$$Q(E_\omega \mid E_{1i}) = \frac{Q(E_\omega \cap E_{1i})}{Q(E_{1i})} = \frac{Q(\bigcup_j(\omega, b_{1i}, b_{2j}))}{Q(E_{1i})} = \frac{B_{1i}P_{1i}(\omega)}{B_{1i}} = P_{1i}(\omega).$$

An analogous argument holds for $Q(E_\omega \mid E_{2j})$. It follows that the shifts can be embedded in a conditioning model. □

The *convex span* of $P_{11}, \ldots, P_{1n}$ is the set of all probability assignments P such that for some constants $B_{11}, \ldots, B_{1n}$ that are non-negative and sum to one, $P(\omega) = \sum_i B_{1i}P_{1i}(\omega)$ for all $\omega \in \Omega$ (*mutatis mutandis* for $P_{21}, \ldots, P_{2m}$). Paraphrasing Theorem 3, the shift from a prior to two distinct sets of posteriors can be embedded in a conditioning model if and only if the prior is in the convex span of *both* sets of posteriors.

Let's say that $P_{11}, \ldots, P_{1n}$ and $P_{21}, \ldots, P_{2m}$ have a *common prior* if there is a P such that the shift from P to $P_{11}, \ldots, P_{1n}$ and $P_{21}, \ldots, P_{2m}$ can be embedded in a conditioning model. It immediately follows from Theorem 3 that $P_{11}, \ldots, P_{1n}$ and $P_{21}, \ldots, P_{2m}$ have a common prior if and only if the intersection of their convex spans is nonempty. There is no talk here about a shift from an actual prior to a set of posteriors. Instead, there are two sets of posteriors that are structured *as if* they come from a common prior. The existence of an otherwise unknown common prior is often enough for the theoretical models I'm going to discuss in the remainder of this section.

Common priors play an important role in game theory.[19] According to John Harsanyi, differences in opinion among rational players trace back to differences in information.[20] This is of special significance for games of *incomplete information*. In a game of incomplete information at least one player is uncertain about which game is being played, that is, there is uncertainty about who the opponents are, or what the utilities are. Harsanyi established a connection between games of incomplete information and certain kinds of games of *imperfect information*. A game of imperfect information is an extensive form game in which at least one player has uncertainty about prior

[19] See Morris [1995] for an overview of the foundations and the methodological significance of the common prior assumption.

[20] Harsanyi [1967-1968].

moves of other players. The special kind of games of imperfect information at play in Harsanyi's theorem is one where nature moves first and assigns a type to each player. Harsanyi proved that a game of incomplete information can be converted into an equivalent game of imperfect information of that type if and only if players have a common prior.[21] Such games of imperfect information are considerably simpler and more tractable than games of incomplete information, leading to numerous applications of Harsanyi's result. One famous application is the rationality of cooperation in the finitely repeated Prisoner's Dilemma when common knowledge of rationality fails.[22] The common prior assumption also plays an important role in Aumann's 'agreeing to disagree' result and in the epistemic foundations of game theoretic solution concepts.[23]

Suppose that each player in a game has a range of posterior beliefs about the structure of the game after having received private information. Theorem 3 tells us whether the sets of posteriors can come from a common prior by conditionalizing on private evidence. This is possible if and only if the convex spans generated by the sets of posteriors have nonempty intersection. The existence of a common prior requires a certain level of agreement among the players' posteriors.

A similar result has been noted by Dov Samet for Harsanyi type spaces.[24] A type space specifies possible posteriors for each agent that vary across states of the world and a partition representing their private knowledge. The superconditioning approach, by contrast, is a surface approach that doesn't specify types. However, the main result of this section can be thought of as characterizing the existence of a type space in which the agents' posteriors come from a common prior by conditionalization.

Samet also gives an alternative characterization of the existence of a common prior, which translates straightforwardly into the superconditioning framework. Let $f : \Omega \to \mathbb{R}$ be a real-valued function. We may think of f as a gamble that pays $f(\omega)$ (in some units) at world ω. Define the following expectations:

$$\mathbb{E}_{1i}(f) = \sum_{\omega} f(\omega) P_{1i}(\omega) \quad \text{and} \quad \mathbb{E}_{2j}(f) = \sum_{\omega} f(\omega) P_{2j}(\omega).$$

The expectation $\mathbb{E}_{1i}(f)$ is the expected value of the gamble f from the point of view of the posterior P_{1i} and $\mathbb{E}_{2j}$ is the expected value of f from the point of view of P_{2j}. It can be shown that two sets of posteriors have a common prior if and only if there is no gamble f such that, for all i, $\mathbb{E}_{1i}(f) > 0$ and, for all j, $\mathbb{E}_{2j}(f) < 0$.[25] In conjunction with Theorem 3, this implies:

Theorem 4. *There exists a prior P such that the shift from P to $P_{11}, \ldots, P_{1n}$ and $P_{21}, \ldots, P_{2m}$ can be embedded in a conditioning model if and only if there exists no function $f : \Omega \to \mathbb{R}$ such that, for $i = 1, \ldots, n$, $\mathbb{E}_{1i}(f) > 0$ and, for $j = 1, \ldots, m$, $\mathbb{E}_{2j}(f) < 0$.*

Suppose there is a common prior and that the gamble f has positive expected utility for all possible posteriors of an agent. Then f cannot be evaluated as uniformly negative by the other agent's posteriors: according to at least one of the second agent's posteriors the expected value of f must be nonnegative. In other words, there is no guarantee that the two agents will view as advantageous the different sides of f, considered as a bet.

[21] Harsanyi [1967-1968].
[22] Kreps et al. [1982].
[23] Aumann [1976] and Aumann and Brandenburger [1995].
[24] Samet [1998]. See also Morris [1994], Bonanno and Nehring [1999], Feinberg [2000] and Halpern [2002].
[25] The proof carries over word for word from the proof of the claim in Samet [1998, p. 173]. Samet's extension of the result to any finite number of agent's also applies to our setting.

These results are well known in game theory. Superconditioning adds a slightly new twist because it doesn't assume a type space. But the same results can also be applied to a model of time-slice epistemology that was introduced by Timothy Williamson and developed by Sarah Moss.[26] According to this model, posterior probabilities are neither constrained by conditionalizing an earlier probability assignment nor by reflecting it. Instead, there is an underlying prior probability measure that is updated at different times by conditionalization on the evidence that is available at that time.[27]

To present the full picture, let me start with an adaptation of Diaconis and Zabell's superconditioning result to the present setting. Consider shifts from P to two posteriors P_1 and P_2. We say that P_1 and P_2 come from P *by conditionalization* if there exists a probability space $(\Lambda, \mathfrak{A}, Q)$ such that, for each $\omega \in \Omega$, there is an $E_\omega \in \mathfrak{A}$ and there are $E_1, E_2 \in \mathfrak{A}$ with $Q(E_1), Q(E_2) > 0$ for which the following is true:

(i) $Q(E_\omega) = P(\omega)$;

(ii) $Q(E_\omega \mid E_1) = P_1(\omega)$; and

(iii) $Q(E_\omega \mid E_2) = P_2(\omega)$.

Two posteriors come from a common prior by conditionalization if there are two conditionalizing events in an idealized probability space that represent both shifts. The following result is an easy extension of Theorem 1:

Theorem 5. *P_1 and P_2 come from P by conditionalization if and only if there exists a $B \geq 1$ such that, for all $\omega \in \Omega$,*

$$P_1(\omega),\ P_2(\omega) \leq BP(\omega). \tag{4}$$

Proof. Suppose P_1 and P_2 come from P by conditionalization. Then, for $i = 1, 2$,

$$P(\omega) = Q(E_\omega) \geq Q(E_\omega \cap E_i) = Q(E_i)P_i(\omega).$$

The conclusion follows by setting B equal to the maximum of $1/Q(E_1)$ and $1/Q(E_2)$.

Conversely, suppose that $P_1(\omega), P_2(\omega) \leq BP(\omega)$ for all $\omega \in \Omega$. If $B = 1$, then $P_1 = P_2 = P$. Otherwise, $B > 1$ and we define the following two probability distributions:

$$Q_1(\omega) = \frac{B}{B-1}P(\omega) - \frac{1}{B-1}P_1(\omega)$$
$$Q_2(\omega) = \frac{B}{B-1}P(\omega) - \frac{1}{B-1}P_2(\omega)$$

[26] Williamson [2002] and Moss [2015].

[27] Moss [2015, p. 175]:

> According to Williamson, your current credences are not constrained by your past credences, but by your current evidence. At any given time, your current credence in a proposition should match the prior conditional probability of that proposition, conditional on your current evidence. The prior probability distribution is a distinguished measure of "something like the intrinsic plausibility of hypotheses prior to investigation" [...], and your current evidence is just your current knowledge [..]. Since knowledge is not necessarily cumulative for rational agents, this proposal answers challenges involving rational memory loss. Since your current mental states include your current knowledge, this proposal advances time-slice epistemology.

For present purposes, the prior probability distribution does not need to have the character required by Williamson ("the intrinsic plausibility of hypotheses prior to investigation"). It's enough that it is a constitutive part of the agent. The superconditioning framework seems to be well-suited for this view of updating since it does not assume that the agent has access to evidence. For more on time-slice rationality see Hedden [2015].

It follows that

$$P(\omega) = \frac{1}{B}P_1(\omega) + \left(1 - \frac{1}{B}\right) Q_1(\omega) = \frac{1}{B}P_2(\omega) + \left(1 - \frac{1}{B}\right) Q_2(\omega).$$

Let $\Lambda = \Omega \times \{b_{11}, b_{12}\} \times \{b_{21}, b_{22}\}$ and define the following events:

$$E_\omega = \bigcup_{i,j}(\omega, b_{1i}, b_{2j}), \quad E_1 = \bigcup_{\omega,j}(\omega, b_{11}, b_{2j}), \quad E_2 = \bigcup_{\omega,i}(\omega, b_{1i}, b_{21})$$

Let Q be the following probability assignment, where probabilities are assumed to be equal to zero whenever $P(\omega) = 0$:

$$Q(\omega, b_{11}, b_{21}) = \frac{\frac{1}{B^2}P_1(\omega)P_2(\omega)}{P(\omega)}, \quad Q(\omega, b_{11}, b_{22}) = \frac{\frac{1}{B}\left(1 - \frac{1}{B}\right) P_1(\omega)Q_2(\omega)}{P(\omega)}$$

$$Q(\omega, b_{12}, b_{21}) = \frac{\left(1 - \frac{1}{B}\right) \frac{1}{B}Q_1(\omega)P_2(\omega)}{P(\omega)}, \quad Q(\omega, b_{12}, b_{22}) = \frac{\left(1 - \frac{1}{B}\right)^2 Q_1(\omega)Q_2(\omega)}{P(\omega)}$$

It is easy to verify that for these assignments $Q(E_\omega) = P(\omega)$, $Q(E_\omega \mid E_1) = P_1(\omega)$, and $Q(E_\omega \mid E_2) = P_2(\omega)$. □

Let's suppose that P is the agent's underlying prior and that P_1 and P_2 are her probability assignments at two successive points in time. If the inequalities (4) obtain, then P_1 and P_2 come from P by conditionalization. We can think of P_1 as the result of conditionalizing on the agent's evidence at one time and P_2 as resulting from conditionalization on a different piece of evidence at a later time. It is not assumed that the agent knows how to say what she has learned.

The constraints imposed by the inequalities (4) are not strong enough to entail any kind of interesting dynamic coherence between P_1 and P_2. In particular, the following two extreme scenarios are compatible with (4) in case $B > 1$:

(i) For some proposition E, $P_1(E) = 1$ and $P_2(E) = 0$.

(ii) For some proposition E, $P_1(E) = 0$ and $P_2(E) = 1$.

Suppose the inequalities in (4) are true. Then in (i) $P_1(E) = 1$ implies $P(E) > 0$, which is compatible with $P_2(E) = 0$. In (ii) the roles of P_1 and P_2 are reversed. So both cases can arise as long as $P(E)$ is positive.

The first case can be thought of as an instance of forgetting past evidence. In the second case the agent undergoes a radical change of opinion. In both cases, P_2 cannot come from P_1 by conditionalization: in each case there exists a proposition that is assigned probability zero by P_1 but has positive probability according to P_2 (E in the second case and the complement of E in the first case). This constitutes a basic failure of dynamic coherence between the agent's time-slices, and so does sit well with the epistemological picture suggested by Moss and Williamson: that the rationality of the agent's time-slices does not depend on how they are related to one another, but does depend on how they are related to the underlying prior probability.

Returning to the setting of Theorem 3 shows that the question of dynamic compatibility is more nuanced. Considering only how shifts actually occur across time does not provide a lot of structure. In general we wish to know how the possible ways in which a shift might proceed are connected to one another.

Let $P_{11}, \ldots, P_{1n}$ be the agent's possible probability assignments at a particular time and $P_{21}, \ldots, P_{2m}$ the possible probabilities at a later time. Theorem 3 does not say that the two sets of probabilities are dynamically coherent in the sense that each P_{1i} is a convex combination of $P_{21}, \ldots, P_{2m}$. In that special case, updating from the prior P to the two sets of posteriors can be represented by a step-wise process of updating to P_{1i} first and to P_{2j} next (at both times by superconditioning). That's compatible with Theorem 3. But it is not necessary. What is required for the two sets of probabilities to arise by conditionalization from an underlying prior probability P is that the intersection of their convex spans is nonempty. This is a weak type of dynamic compatibility. It says that the agent's time-slices cannot be completely divorced from each other if they come from an underlying prior probability by conditionalization. The underlying prior acts as a glue that binds time-slices.

Theorem 4 sheds light on the kind of dynamic compatibility that is at play here. Consider a gamble f that has positive expected utility according to all of an agent's possible posteriors at time 1: $\mathbb{E}_{1i}(f) > 0$ for all i. If the two sets of posteriors have a common prior, it cannot be the case that f is assigned a negative expectation by all posteriors at time 2. If that were the case, then $\mathbb{E}_{2j}(-f) > 0$ for all j, and the agent's two personae would be certain to consider as advantageous the different sides of f, viewed as a bet. Such a radical switch from evaluating a gamble as uniformly positive to evaluating it as uniformly negative at a later time is impossible if posteriors come from a common prior.

This mild form of inter-temporal comptability is a feature of Moss's and Williamson's time-slice model. Even in a situation in which I don't know my underlying prior nor my possible future posteriors, I'm certain that my future self won't completely reverse what I consider to be uniformly advantageous gambles.

5 Conclusions

Extending the superconditioning framework of Diaconis and Zabell in a natural way demonstrates the significance of the reflection principle for Bayesian updating. In what follows, I discuss in more detail what these results tell us about rational learning.

5.1 Partitionality

The results in this paper rely on having a *partition* of proposition in the superconditioning space. This assumption is controversial in settings in which an agent may not have access to her evidence.[28] The nature of evidence is a complex subject matter that I cannot discuss exhaustively here. But a few remarks on why it is plausible to assume partitions, at least for superconditioning, are in order.

First, even if evidential propositions don't always constitute a partition, they do so in many paradigm cases of inquiry. Scientific experiments, for example, often involve partitions, although infinite outcome spaces require some care (in what follows I restrict myself to finite settings).[29] Learning scenarios, as a matter of fact, often are partitional.

[28] See, *inter alia*, Williamson [2002, 2011], Christensen [2010], Elga [2013], Bronfman [2014], Lasonen-Aarnio [2015], Salow [2018], Gallow [2019], and Dorst [2020].

[29] Partitions are straightforward in finite and countably infinite probability spaces. Experiments in uncountable probability spaces are often represented by sub-sigma-algebras, which need not be generated by countable partitions. However, in many applications they are.

Situations where evidence propositions fail to form a partition typically involve agents with imperfect introspective access to their evidence. Examples include the unmarked clock or good-case bad-case scenarios such as meeting an old friend who might be mistaken for a stranger. The unmarked clock only provides evidence up to a margin of error. For instance, if it is n minutes past the full hour, your evidence is assumed to be given by a proposition that includes n but also $n-1$ and $n+1$ as possible true states of the clock. The resulting evidential propositions do not overlap. Similarly, for meeting an old friend it is stipulated that you know that he is your friend when he is in fact your friend; otherwise you are uncertain between the person entering the room being a stranger or your friend. Here, evidence in the second state (friend or not friend) overlaps with the evidence in the first one (friend). Examples like the unmarked clock or meeting s stranger are usually motivated by failures of the KK principle or related access postulates.

By way of responding, note first that non-partitional settings have consequences that are difficult to square with thinking of them as genuine learning events. In an important paper, Bernhard Salow demonstrates that cases like the unmarked clock or meeting an old friend result in intentionally biased investigations: inquiries in which an agent manipulates the evidence they receive so that they are only exposed to desirable information.[30] This is a powerful argument against the view that non-partitional settings are learning scenarios. For example, in good-case bad-case situations good cases provide clearly delineated evidence, but bad cases do not. The situation involves both prospects of learning but also possibilities of gaining no information, calling into question whether the overall setting should be considered a learning scenario. Salow illustrates this point in an especially sharp way.

In the same vein, it has been shown that non-partitional evidence gives rise to dynamic Dutch books and failures of maximizing expected accuracy.[31] More specifically, if propositions don't form a partition, conditonalizing on propositions is dynamically incoherent and is not an update strategy that maximizes expected accuracy. If evidence propositions form a partition, on the other hand, conditioning is both dynamically coherent and maximizes expected accuracy.

Different lessons may be drawn from the close connection between dynamic coherence and expected accuracy, on the one hand, and conditioning on a partition, on the other. One may think that it casts doubt on dynamic coherence and expected accuracy, or on conditioning on a partition, or on both. However, if one already is committed to dynamic coherence, expected accuracy, and conditionalization as an update rule (not necessarily on a partition), these results raise doubts about non-partitional settings. Commitments to dynamic coherence, expected accuracy, and conditionalization can be based on heuristic considerations. Conditionalization requires that one's implicit attitudes towards updating (conditional probabilities) are the same as one's explicit update strategies (conditionalization). Expected accuracy requires that one thinks of information gain as bringing one closer to the truth on average. Dynamic incoherence, in turn, reveals incompatibilities in one's underlying evaluations about a learning situation. These criteria are, I think, as intuitively appealing as the idea that all evidence is given by, perhaps overlapping, propositions. But if one starts with these intuitions about conditionalization, expected accuracy, and dynamic coherence, then one will conclude that non-partitional settings are perhaps not genuine learning situations since conditionalizing in those settings leads to violations of dynamic coherence and expected accuracy. If one takes the latter two as characteristic of learning, there is something wrong with conditionalizing on propositions that are not a partition. As mentioned above, what's wrong is not that non-partitional sitations have no evidential aspects at all, but that, overall, evidence is not represented

[30] Salow [2018].

[31] See Gallow [2019] on Dutch books and Schoenfield [2017] on accuracy.

sufficiently sharply.[32]

I don't take the reasons for the partitionality of evidence given in the preceding paragraphs to be decisive. Obviously, the nature of evidence is a complex topic. However, the reasons given do suggest that evidential partitions occupy a privileged place when it comes to learning in finite settings.

Finally, let me note that the shift from a prior to a set of posteriors, as set forth in the superconditioning framework, involves no assumption about evidence being given by a partition. The shift itself is arbitrary and can proceed in all kinds of ways. The agent may, for instance, update by conditioning on overlapping propositions. The superconditioning theorem simply tells us when such a shift can be represented as conditioning on a partition, with no presumption that the agent has access to to the propositions in the superconditioning space.

5.2 Is Superconditioning Too Permissive?

Conditionalization does not say much about the nature of the propositions conditioned upon. Does it therefore fall prey to the same type of problem that haunts black box learning? In principle, one can condition on any proposition, however misleading or otherwise unworthy it is. Such a process does not constitute learning.

While that is certainly possible, there is one big difference between conditionalization and black box learning. Conditionalization is transparent in that the content of the alleged learning experience is given by a proposition that is assumed to be accessible to the agent. It will, as a result, be usually quite straightforward to determine whether the proposition in question should be conditioned on. This is, by definition, not true for black box learning.

Superconditioning is closer to black box learning. In asserting that a shift from a prior to a set of posteriors can be represented by conditioning on a partition of propositions, nothing is said about the content of those propositions. They might be mere mathematical artifacts that don't to pick out genuine evidential distinctions. What's crucial about superconditioning, though, is the *in-principle existence* of a superconditioning model. The existence of a superconditioning space can be taken, prima facie, as a necessary condition for genuine, or rational, learning. For otherwise conditioning on a a partition of evidential propositions is simply impossible. Considering the existence of a superconditioning space as necessary doesn't presuppose that the propositions conditioned on are indeed evidentially legitimate (I sketch an example of an arguably evidentially illegitimate situation in which superconditioning works in the next section). Since the existence of a superconditioning space is equivalent to the generalized reflection principle, the latter can also be thought of as a necessary condition for genuine, or rational, learning. Superconditioning thus traces the structure of conditioning without there being a guarantee that the evidential legitimacy of the conditioning propositions can be examined by considering their content.

5.3 Higher-Order Models

I mentioned higher-order probability spaces earlier in the paper, but it's worth discussing a few further differences between the superconditioning framework and Bayesian approaches that use

[32]It is worth noting that Schoenfield [2017] argues for a way to conceive of evidence in non-partitional settings that restore partitionality.

higher-order probabilities, such as Harsanyi type spaces and similar models in economics and epistemic logic.[33]

Consider Jeffrey shifts or black box learning events. There are no propositions in the agent's probability space that would express exactly what is being learned. But if agents can avail themselves of their own anticipated future probabilities, propositions about one's new probabilities capture the learning event indirectly. For instance, if learning prompts me to change my belief in A from $1/2$ to $3/4$, the proposition "My new probability for A is $3/4$" may serve as a surrogate for an otherwise ineffable learning experience. Such propositions regulate the connection between my posteriors and my prior.[34] The central principle is the reflection principle in its conditional manifestation, which requires coherence between prior conditional probabilities and posterior probabilities: the probability of a proposition, conditional on its posterior taking on a certain value, is equal to that value.

With that principle in place, Bayesian models have the resources to characterize Jeffrey shifts and black box learning without saying what's being learned. In particular, the reflection principle follows naturally from, and is in a certain sense equivalent to, three distinct desiderata for black box learning: learning should increase expected accuracy, it should be such that one expects to make better decisions, and it should be dynamically coherent.[35] These are plausible necessary conditions for any genuine, or rational, learning event. Since they imply the reflection principle, the latter also is a necessary condition for any genuine, or rational, learning event. The chain of reasoning here is the same as for superconditioning and the generalized reflection principle.

What, if anything, does the superconditioning framework add? First, propositions about an agent's posteriors describe the *effect* a learning experience has on the agent's beliefs and not the learning of any piece of information that *causes* a belief change. Higher-order Bayesian models lack the conceptual resources to say whether or when propositions exist that describe the learning experience in this latter sense since they assume the existence of higher-order surrogates (like "my posterior for A is $3/4$") from the get-go. The superconditioning framework makes no such assumption. That's why we can answer the question whether a proposition that could describe the content of a learning experience exists in principle.[36]

That higher-order models cannot capture the cause of a belief change is one kind of restriction. They are restricted in another way. By taking anticipated future probabilities to be random variables in a probability space, higher-order models treat updating in the synchronous mode (see §2). That's not the case for superconditioning. Superconditioning models follow the template of conditionalization by shifting a prior probability space to a set of posterior probability spaces as opposed to modeling posteriors as parts of the prior probability space.

I take these points to be the two main advantages the superconditioning framework has over Bayesian models that involve future credences as random variables. That's not to say that the superconditioning framework has no limitations. For starters, let me explain how the classical reflection principle can be derived from the generalized reflection principle that characterizes the existence of a superconditioning space. Recall, from §3, the proposition $A_i = \{\omega \in \Omega : P_i \text{ is the posterior at } \omega\}$,

[33] There is a large literature on this. See, among others, Gaifman [1986], Samet [1999], Skyrms [2012, Chapter 5], Huttegger [2017], and Dorst et al. [2021].

[34] They can be used to characterize Jeffrey shifts [Skyrms, 1990].

[35] See Huttegger [2017] for a discussion.

[36] The higher-order proposition describing a learning experience's effect and the proposition describing the experience are of course extensionally equivalent (they pick out the same possible worlds), but they have different philosophical interpretations.

which can be formed if each state determines a posterior. As pointed out there,

$$P(B) = \sum_i P(A_i)P_i(B)$$

if $P_i(A_i) = 1$ for all i and $P_j(A_i) = 0$ for all $j \neq i$. From this, one can derive the classical reflection principle in the following form:

$$P(B \mid A_i) = P_i(B)$$

whenever $P(A_i) > 0$.[37]

Thus there is a close connection between the generalized reflection principle and the classical reflection principle *if* we have propositions about posteriors *and* those posteriors are certain about themselves (luminosity). The latter assumption is not universally accepted. For instance, Kevin Dorst recently argued that posteriors should be allowed to be modest, which is modeled as a failure of introspection.[38] I think there is more to say about modesty *qua* failure of introspection and the settings in which it is a good modeling assumption. But I set that topic aside here to explain the connection between modest posteriors and the superconditioning theorem.

To start with, note that if posteriors are not perfectly introspective, the equation $P(A_i) = B_i$ will typically be false. For then the generalized reflection principle only entails that $P(A_i)$ is a weighted average of the posterior probabilities of A_i:

$$P(A_i) = \sum_j B_j P_j(A_i)$$

But even in these cases the shift can be represented by a conditioning model as long as the generalized reflection principle holds. As a result, the superconditioning framework may not be sufficiently sensitive to higher-order considerations. To see why, consider a much discussed example.[39] Suppose two posteriors, P_1 and P_2, with $P_1(A_1) = \frac{1}{10}$, $P_1(A_2) = \frac{9}{10}$, $P_2(A_1) = \frac{9}{10}$, and $P_2(A_2) = \frac{1}{10}$. These posteriors are defective by being overly modest: each posterior thinks the other one is more likely to be true at worlds at which it is true. They are, in a word, "anti-experts".[40] But a shift to those anti-expert posteriors can be represented by a conditioning model so long as the prior, P, is a weighted average of P_1 and P_2, something that can be achieved effortlessly.

So, we are faced with the following situation. The shift from an appropriate prior to the anti-expert posteriors can be represented in a superconditioning space, even though there is clearly something wrong with the posteriors. They misjudge the truth about themselves, indicating their general unreliability. We should, therefore, not think of the propositions in the superconditioning space that represents the shift as representing a legitimate learning event.

This example pours cold water on the superconditioning model. But it is cold water we are familiar with. To repeat, the superconditioning space is a purely formal construct. It says nothing about the content of the propositions conditionalized on. It only asserts their existence. Thus, the existence of a superconditioning representation can only be taken (at best) as a necessary condition for a shift to be a rational learning scenario. The above example reinforces this point. Sometimes a

[37] Proof: For all B, $P(B \cap A_i) = \sum_j P(A_j)P_j(B \cap A_i) = P(A_i)P_i(B \cap A_i) = P(A_i)P_i(B)$, where the last two inequalities follow from $P_j(A_i) = 0, j \neq i$ and $P_i(A_i) = 1$. If $P(A_i) > 0$, then $P(B \mid A_i) = P_i(B)$. This also holds whenever there is a partition $A_1, \ldots, A_n$ such that $P_j(A_i) = 0, j \neq i$ and $P_i(A_i) = 1$ for all i.

[38] Dorst [2020]. See also Christensen [2010] and Elga [2013].

[39] Adapted from Dorst [2020] and Dorst et al. [2021].

[40] See Dorst et al. [2021] for a discussion.

shift that carries a prior to defective posteriors can be represented by superconditioning. What we should conclude from this is that the existence of a superconditioning space ought to be considered with caution: it does not guarantee the shift's epistemic legitimacy.

Picking up an earlier thread, the same caution should be applied to the reflection principle in higher-order Bayesian models and to some alternatives recently discussed in the literature.[41] By way of intermediaries (Dutch books, value of knowledge, expected accuracy) such principles are plausible necessary conditions for rational learning scenarios. But they do not guarantee legitimacy. The higher-order framework, too, is purely formal. Once we think of concrete learning situations, it may turn out that these principles only hold accidentally.

5.4 Tying Things Together

Now that we have looked at some issues surrounding superconditioning, let me summarize what has been achieved. A generalized version of the reflection principle is equivalent to shifts being representable by conditionalizing on propositions in an enriched probability space. The existence of such a superconditioning space is arguably a necessary condition for shifts to be epistemically legitimate, for otherwise there can be no evidential partition that describes the learning event. The superconditioning framework can be thought of as genuinely diachronic, setting it apart from higher-order Bayesian models and going beyond what can be demonstrated within that setting. Superconditioning is especially relevant when what an agent learns need not be accessible to them, as in the Moss-Williamson model of time-slice rationality. Here, superconditioning reveals that an agent's time-slices need to be dynamically compatible in order to come from conditionalization on an underlying prior.

References

Frank Arntzenius. Some problems for conditionalization and reflection. *Journal of Philosophy*, 100: 356–370, 2003.

Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4:1236–1239, 1976.

Robert J. Aumann and Adam Brandenburger. Epistemic conditions for Nash equilibrium. *Econometrica*, 63:1161–1180, 1995.

Giacomo Bonanno and Klaus Nehring. How to make sense of the common prior assumption under incomplete information. *International Journal of Game Theory*, 28:409–434, 1999.

Luc Bovens. 'P and I will believe that not-P': Diachronic constraints on rational belief. *Mind*, 104: 737–760, 1995.

Ray Briggs. Distorted reflection. *Philosophical Review*, 118:59–85, 2009.

Aaron Bronfman. Conditionalization and not knowing that one knows. *Erkenntnis*, 79:871–892, 2014.

David Christensen. Rational reflection. *Philosophical Perspectives*, 24:121–140, 2010.

[41] I'm referring to the trust principles studied in Dorst et al. [2021], who establish connections between these principles in terms expected accuracy and the value of knowledge in settings where introspection may fail.

Persi Diaconis and Sandy L. Zabell. Updating subjective probability. *Journal of the American Statistical Association*, 77:822–830, 1982.

Kevin Dorst. Evidence: A guide for the uncertain. *Philosophy and Phenomenological Research*, 100:586–632, 2020.

Kevin Dorst, Benjamin A. Levinstein, Bernhard Salow, Brooke E. Husic, and Branden Fitelson. Deference done better. *Philosophical Perspectives*, 35:99–150, 2021.

Adam Elga. The puzzle of the unmarked clock and the new rational reflection principle. *Philosophical Studies*, 164:127–139, 2013.

Yossi Feinberg. Characterizing common priors in the form of common posteriors. *Journal of Economic Theory*, 91:127–179, 2000.

Haim Gaifman. A theory of higher order probabilities. In *Theoretical Aspects of Reasoning about Knowledge*, pages 275–292. Elsevier, 1986.

J Dmitri Gallow. Diachronic dutch books and evidential import. *Philosophy and Phenomenological Research*, 99:49–80, 2019.

Michael Goldstein. The prevision of a prevision. *Journal of the American Statistical Association*, 78:817–819, 1983.

Joseph Y. Halpern. Characterizing the common prior assumption. *Journal of Economic Theory*, 106:316–355, 2002.

John C. Harsanyi. Games With Incomplete Information Played by Bayesian Players. Parts 1-3. *Management Science*, 14:159–183, 320–334, 486–502, 1967-1968.

Brian Hedden. *Reasons Without Persons: Rationality, Identity, and Time*. Oxford University Press, Oxford, 2015.

Simon M. Huttegger. In defense of reflection. *Philosophy of Science*, 80:413–433, 2013.

Simon M. Huttegger. *The Probabilistic Foundations of Rational Learning*. Cambridge University Press, Cambridge, 2017.

Richard C. Jeffrey. *The Logic of Decision*. McGraw-Hill, New York, 1965. 3rd revised edition Chicago: University of Chicago Press, 1983.

Richard C. Jeffrey. Conditioning, kinematics, and exchangeability. In B. Skyrms and W. L. Harper, editors, *Causation, Chance, and Credence*, volume 1, pages 221–255. Kluwer, Dordrecht, 1988.

David M. Kreps, Paul Milgrom, John Roberts, and Robert Wilson. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27:245–252, 1982.

Maria Lasonen-Aarnio. New rational reflection and internalism about rationality. *Oxford studies in epistemology*, 5:145–171, 2015.

Isaac Levi. The demons of decision. *The Monist*, 70:193–211, 1987.

David Lewis. *Papers in Metaphysics and Epistemology*. Cambridge University Press, Cambridge, 1999.

Stephen Morris. Trade with heterogeneous prior beliefs and asymmetric information. *Econometrica*, 62:1327–1347, 1994.

Stephen Morris. The common prior assumption in economic theory. *Economics and Philosophy*, 11:227–253, 1995.

Sarah Moss. Time-slice epistemology and action under indeterminacy. *Oxford Studies in Epistemology*, 5:172–94, 2015.

Bernhard Salow. The externalist's guide to fishing for compliments. *Mind*, 127:691–728, 2018.

Dov Samet. Common priors and separation of convex sets. *Games and Economic Behavior*, 24: 172–174, 1998.

Dov Samet. Bayesianism without learning. *Research in Economics*, 53:227–242, 1999.

Miriam Schoenfield. Conditionalization does not (in general) maximize expected accuracy. *Mind*, 126:1155–1187, 2017.

Brian Skyrms. *Pragmatics and Empiricism*. Princeton University Press, Princeton, 1984.

Brian Skyrms. *The Dynamics of Rational Deliberation*. Harvard University Press, Cambridge, MA, 1990.

Brian Skyrms. *From Zeno to Arbitrage: Essays on Quantity, Coherence, and Induction*. Oxford University Press, Oxford, 2012.

William Talbott. Two principles of Bayesian epistemology. *Philosophical Studies*, 62:135–150, 1991.

Bas van Fraassen. Belief and the problem of Ulysses and the Sirens. *Philosophical Studies*, 77:7–37, 1995.

Bas C. van Fraassen. Belief and the will. *Journal of Philosophy*, 81:235–256, 1984.

Jonathan Weisberg. Conditionalization, reflection, and self-knowledge. *Philosophical Studies*, 135: 179–197, 2007.

Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, Oxford, 2002.

Timothy Williamson. Improbable knowing. *Evidentialism and its Discontents*, 147:164, 2011.