

ALPHA DIVERGENCE REPRESENTATIONS OF DYNAMICALLY CONSISTENT BELIEF SELECTION

Rommewinkel, Hendrik^{1,2} and Ui, Takashi^{1,3}

¹Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-0004, hendrik.r@r.hit-u.ac.jp

²Waseda Institute for Advanced Study, 1-104 Totsukamachi, Shinjuku City, Tokyo 169-8050

³Kanagawa University, 3-chōme-27-1 Rokkakubashi, Yokohama, Kanagawa 221-0802

THIS VERSION: MARCH 2026

Abstract

We provide a normative foundation for the class of alpha divergences in a parsimonious model of belief selection. If a decision maker is to update beliefs dynamically consistently — in the sense that additional information cannot reverse preferences over distributions that already agreed on the relevant states — has convex preferences, and satisfies a separability condition across subsets of states, then she *ought to* choose her beliefs by minimizing the alpha divergence from a unique subjective reference distribution to the chosen distribution. We show that this axiomatically justifies a wide range of statistical estimators and practices.

KEYWORDS: decision theory, relative entropy, alpha divergence, entropy, dynamic consistency

1 INTRODUCTION

Consider a decision maker who holds a background belief in the form of a prior probability distribution, frequency data, or degrees of confidence of the correctness of various texts. When new information arrives in the form of a set of beliefs that is consistent with this information she then selects, from this feasible set, a belief that reconciles the new information with her background world model. If the decision maker is dynamically consistent with respect to the arrival of arbitrary new information, she should have a complete and transitive preference relation over probability distributions. In particular, it may be plausible to prefer probability distributions closer to the background belief over those further away. In statistics and machine learning, closeness between distributions is most commonly measured by the Kullback–Leibler divergence. When appropriate, the Kullback–Leibler divergence is replaced by a broader class of divergence measures, namely the family of alpha-divergences, special power form cases of f-divergences (Ali & Silvey, 1966).

This paper provides an axiomatic foundation for rankings of beliefs based on divergence from a reference distribution. We impose that if Bayesian conditioning preserves the ranking of probability distributions that are identical outside the updating event, a condition we call *updating consistency*. This condition turns out to greatly restrict the permissible preferences; if a decision maker fulfills standard axioms together with strict convexity of preferences and updating consistency, then the ranking admits an alpha-divergence representation relative to an endogenously determined reference belief with some parameter $\alpha > 0$. Thus, from the decision maker’s choices over distributions, an analyst can infer a unique reference probability distribution and a unique parameter α such that the decision maker minimizes the alpha divergence between the reference distribution and the chosen distribution.

We further show that if we do not impose strict convexity, the decision maker maximizes the divergence on some states and minimizes it on other states, a representation we call a *polarized divergence*. We also weaken updating consistency to a separability condition to characterize alpha divergences for arbitrary $\alpha \in \mathbb{R}$ for preferences on distributions with full support.

This paper is best read as a normative contribution. The axioms we impose —

dynamic consistency of belief rankings, a preference for moderate distributions, and coherence of conditional evaluations — are defensible as requirements on a rational agent. Their conjunction uniquely picks out alpha divergence minimization. In our applications we show that statistical procedures as diverse as Bayesian updating, Jeffrey updating, maximum likelihood estimation, variational inference, and generalized empirical likelihood methods have these axioms as a common normative foundation.

The problem of updating subjective beliefs to partitional information can in our model be understood as choosing from the set of probability distributions in which the event the decision maker has just learned has probability one. Alpha divergence minimization for arbitrary α corresponds to Bayesian updating in this case. More generally, when information is provided in the form of the exact probability of an event, our decision maker chooses from the resulting set of probability distributions according to Jeffrey updating. If Bayesian updating is computationally intractable and the decision maker has a parameter $\alpha = 1$, then the decision maker employ variational inference: Variational inference, which is commonly used in machine learning, minimizes the KL divergence between the prior and the posterior given that the posterior must be within a computationally tractable class of models. Thus, the standard methods of updating beliefs can be understood as choosing a posterior probability distribution that is dynamically consistent, avoids extreme distributions due to strict convexity, and fulfills updating consistency.

Classical statistics can also be understood as a problem of choosing a probability distribution: namely the distribution generated by different models and parametrizations in response to an empirical distribution of observations. In our framework, when the reference distribution is interpreted as the empirical distribution and the model fixes the available probability distributions, then minimizing the alpha divergence with $\alpha = 0$ corresponds to maximum likelihood estimation. In semi-parametric estimation, both the Empirical Likelihood estimator (Owen, 1988) and the Exponential Tilting estimator (Kitamura & Stutzer, 1997) correspond to alpha divergence minimization with different values of α .

The paper proceeds as follows. In Section 2 we introduce the formal model and the alpha divergence. In Section 3 we provide our characterization result. We provide an overview over the above mentioned applications in section 5. Proofs are stated in the appendices.

2 NOTATION

Let S be a set of states and Σ be a σ -algebra on S . Let $\mathcal{P}$ be the set of probability measures on S . If p and q are probability measures with $p(E) = q(E)$ for some event $E \in \Sigma$, then $p_E q$ is the probability measure that agrees with p on E and with q on its complement, $\bar{E}$. We endow $\mathcal{P}$ with the product topology and subsets thereof with the subset topology. The *interior* of a set S is denoted $\text{int}(S)$.

We denote the *restriction* of p to E by p_E , i.e., $p_E(F) = p(F)$ for all $F \subseteq E$. If $p(E) > 0$, we denote by $p|E$ the *conditional measure* on event E , i.e., for all $F \subseteq E$, $p|E(F) = p_E(F)/p(E)$. At times it will be useful to *compose* restrictions: $p_E q(x) = p(x)$ if $x \in E$ and otherwise $p_E q(x) = q(x)$ (for q with $p(E) = q(E)$). Also, for all $\mu \in [0, 1]$ we define the *mixture of two measures* p and q by: $(\mu p + (1 - \mu)q)(x) = \mu p(x) + (1 - \mu)q(x)$.

Occasionally, where it will be unambiguous, we will find it useful to refer to binary Bernoulli distributions with probabilities p and $1 - p$ simply by $(p, 1 - p)$.

We consider binary relations on $\mathcal{P}$. A generic relation is denoted by $\succsim$, its symmetric part by $\sim$, and its asymmetric part by $\succ$. A function $u : \mathcal{P} \rightarrow \mathbb{R} \cup \{\infty, -\infty\}$ represents a relation $\succsim$ if for all $p, q \in \mathcal{P}$: $p \succsim q$ if and only if $u(p) \geq u(q)$.

An event E is *essential* (in $\succsim$) if for some p, q , $p|E \succ q|E$.

The key functional form we study in this paper is the alpha divergence of two probability distributions:

Definition 1 (α -divergence). The α -divergence from p to q , $D_\alpha : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$ for $\alpha \in \mathbb{R} \setminus \{0, 1\}$ is defined as

$$\begin{aligned} D_\alpha(p, q) &= \frac{1}{\alpha(\alpha - 1)} \left(\sum_{x \in \text{supp}(q)} q(x) \left(\frac{p(x)}{q(x)} \right)^\alpha - 1 \right), \\ D_1(p, q) &= \sum_{x \in \text{supp}(q)} p(x) \ln p(x)/q(x), \\ D_0(p, q) &= \sum_{x \in \text{supp}(q)} q(x) \ln q(x)/p(x), \end{aligned} \tag{1}$$

whenever these terms are well-defined.

The cases $\alpha = 1$ and $\alpha = 0$ are known as the Kullback-Leibler divergence and reverse Kullback-Leibler divergence. There are several variants of α divergences.

The Tsallis α divergence differs by a factor of α and has a different limiting case for $\alpha \rightarrow 0$. The Renyi α divergence is a logarithmic transformation of the Tsallis α divergence. Also, we used the convention that summation is over the support of q to avoid excessive case distinctions.

3 CHARACTERIZATION

The following two definitions are standard:

Axiom 1 (Weak Order). $\succsim$ is a *weak order* if it is complete and transitive.

In the case of choice of probability distributions as posterior beliefs, weak order can be interpreted as dynamic consistency as the decision maker receives additional information. For example, the decision maker might at first receive the information that the true distribution is within a set A and later that it is within a subset B . Sen's conditions alpha and beta can naturally be interpreted as the decision maker being dynamically consistent, e.g., if the optimal probability distribution from A is within B , it remains optimal. If the decision maker violates weak order, they must violate Sen's conditions alpha and beta for some combination of A and B , thus be dynamically inconsistent under the revelation of additional information.

Axiom 2 (Continuity). $\succsim$ is *continuous* on $P \subseteq \mathcal{P}$ if the sets $\{p | p \succ q\}$ and $\{p | q \succ p\}$ are open in P . If $P = \mathcal{P}$, we simply say $\succsim$ is continuous.

Continuity of preferences is plausible in the interior of the set of probability distributions, i.e., if all distributions have strictly positive probabilities. On the boundary however, it is plausible that the decision maker's preferences are discontinuous. We will cover such cases by simply restricting our entire preference on interior distributions, allowing us to maintain continuity.

Axiom 3 (Updating Consistency). $\succsim$ is *updating consistent* if whenever $p(E) = q(E) > 0$ and $p|\bar{E} = q|\bar{E}$, then

$$p \succsim q \quad \Leftrightarrow \quad p|E \succsim q|E \tag{2}$$

This axiom can be interpreted as follows: if the decision maker learns that event E is true, and two probability distributions agreed on the probability measure outside E , then the decision maker should not revise her opinion on the two probability measures. If the measure p was preferred to q prior to learning that E obtains, then p must also be preferred to q after learning that E obtains. Violating updating consistency is normatively problematic: the decision maker revises her ranking of two distributions upon learning information that, by hypothesis, is irrelevant to their difference — their conditional measures on E are unchanged, as is their agreement outside E . A rational agent should have no grounds for this revision.

We can view updating consistency as complementing the dynamic consistency notion of our weak order axiom. The choice over probability distributions has additional structure that weak order does not exploit.

Before stating the characterization of updating consistent preferences in its full generality, we consider the special case of convex preferences. We define:

Axiom 4 (Convexity/Concavity). $\succsim$ is *strictly convex* (*strictly concave*) if whenever $p \sim q$, then $\theta p + (1 - \theta)q \succ (\prec)p$ for all $\theta \in (0, 1)$.

Convexity can be understood as a preference for moderate probability measures while concavity is a preference for extreme distributions. If two probability measures are equally attractive, the moderate will prefer a convex combination of the two while the extremist will prefer either of the two measures to any convex combination. From a normative standpoint, strict convexity encodes epistemic humility: without positive evidence for extremes, a rational agent should not prefer a distribution that concentrates all probability on a small set of states over a balanced mixture of equally-attractive distributions.

Proposition 1 (Convex updating consistent preferences). *Let Σ contain at least 4 mutually disjoint essential events. Then the following statements are equivalent:*

1. $\succsim$ fulfills weak order, continuity, updating consistency, and strict convexity.
2. There exists a unique probability measure r and a unique $\alpha > 0$ such that $\succsim$ is represented by the negative of the α -divergence from r :

$$U(p) = -D_\alpha(p, r). \quad (3)$$

In other words, a decision maker whose preference fulfills updating consistency and strict convexity behaves as if they have a reference probability measure r in mind and when confronted with the choice between different measures they minimize the α -divergence from r . Thus, the decision maker's preference is maximized at r and the decision maker tries to choose a probability measure as close to r as possible. Naturally, we obtain the reverse result for concave preferences:

Proposition 2 (Concave updating consistent preferences). *Let Σ contain at least 4 mutually disjoint essential events. Then the following statements are equivalent:*

1. $\succsim$ fulfills weak order, continuity, updating consistency, and concavity.
2. There exists a unique probability measure r and a unique $\alpha > 0$ such that $\succsim$ is represented by the α -divergence from r :

$$U(p) = D_\alpha(p, r). \quad (4)$$

Convexity and concavity therefore decide the direction of the preference – divergence maximization or minimization. It turns out that the functional form of the α -divergence and the existence of a unique reference measure r is largely implied by updating consistency. Without imposing strict convexity or strict concavity, it can however be the case that the decision maker maximizes the divergence for some states and minimizes it for other states. We therefore define:

Definition 2 (Polarized Divergence). Given a partition $\{X^+, X^0, X^-\}$ of S the polarized α -divergence from p to q , $PD_\alpha : \text{int}(\mathcal{P}) \times \text{int}(\mathcal{P}) \rightarrow \mathbb{R}$ is defined as

$$\begin{aligned} PD_\alpha(p, q) = & \frac{p(X^+)^\alpha}{q(X^+)^{\alpha-1}} D_\alpha(p|X^+, q|X^+) \\ & - \frac{p(X^-)^\alpha}{q(X^-)^{\alpha-1}} D_\alpha(p|X^-, q|X^-) \\ & + \frac{1}{\alpha(\alpha-1)} \left(\frac{p(X^+)^\alpha}{q(X^+)^{\alpha-1}} - \frac{p(X^-)^\alpha}{q(X^-)^{\alpha-1}} - 1 \right) \end{aligned} \quad (5)$$

for all $p, q \in \text{int}(\mathcal{P})$.

The interpretation of the representation is that the decision maker has in mind a reference distribution q . The decision maker partitions the state space into states

on which the decision maker prefers a conditional probability as close as possible to their reference distribution, X^- , states into which they prefer conditional probabilities to be far from their reference distribution, X^+ , and irrelevant states, X^0 . Thus, on X^+ , the decision maker maximizes the divergence, and on X^- minimizes the divergence. In addition, the decision maker accounts for the relative probability mass assigned to the events X^+ and X^- . For negative α , the decision maker tries to assign as much probability mass to X^- as possible, given fixed conditional distributions. For positive α , the decision maker tries to assign as much probability mass to X^+ as possible.

One may wonder why the polarized divergence does not just linearly aggregate the conditional divergences. The aggregation method for the conditional divergences is natural, given the following observation for α -divergences:

Remark 1 (Decomposition of a Divergence). For any α -divergence, event E , and any $p, q \in \text{int}(\mathcal{P})$ the following decomposition holds:

$$\begin{aligned} D_\alpha(p, q) &= \frac{q(E)^\alpha}{p(E)^{\alpha-1}} D_\alpha(p|E, q|E) \\ &\quad + \frac{q(\bar{E})^\alpha}{p(\bar{E})^{\alpha-1}} D_\alpha(p|\bar{E}, q|\bar{E}) \\ &\quad + D_\alpha((p(E), p(\bar{E})), (q(E), q(\bar{E}))) \end{aligned} \tag{6}$$

Our polarized divergence follows the same pattern except for changes in signs for all terms about states contained in X^- .

It turns out that aside from weak order and continuity, updating consistency is sufficient to characterize uniquely the beliefs of the decision maker and that they act as if they optimize a divergence relative to their own belief:

Theorem 1. *Let Σ contain at least 4 mutually disjoint essential events. Then the following statements are equivalent:*

1. $\succsim$ fulfills weak order, continuity, and updating consistency.
2. $\succsim$ is an α -polarized divergence for some unique $\alpha > 0$.

The characterization in Theorem 1 only addresses the case in which $\alpha > 0$. It turns out that the remaining parameter space $\alpha \leq 0$ violates updating consistency

only at the boundary when some states have zero probability. For such probability measures, the divergence can become (negative) infinite. This violates updating consistency because one measure may be strictly preferred to another but after conditioning both measures on E , $\bar{E}$ has zero probability and the divergence of both measures becomes infinite. However, this limitation is quite contrived as the α -divergences for $(-\infty, 0]$ are otherwise well behaved.

We therefore weaken updating consistency by only requiring consistent evaluation of conditional probabilities for probability measures that assign nonzero probability to all states.

Axiom 5 (Interior Separability I). $\succsim$ is *interior separable* if for all $p, p', q, q' \in \text{int}(\mathcal{P})$, whenever

- $p|E = p'|E$ and $q|E = q'|E$,
- $p(E) = q(E)$ and $p'(E) = q'(E)$, and
- $p|\bar{E} = q|\bar{E}$ and $p'|\bar{E} = q'|\bar{E}$ (allowing $p|\bar{E} \neq p'|\bar{E}$),

then:

$$p \succsim q \iff p' \succsim q'. \quad (7)$$

This condition is implied by updating consistency because under the stated conditions:

$$p \succsim q \iff p|E \succsim q|E \iff p'|E \succsim q'|E \iff p' \succsim q'.$$

The following reformulation is equivalent:

Axiom 6 (Interior Separability II). $\succsim$ is *interior separable* if for all $p, q \in \text{int}(\mathcal{P})$,

$$\begin{aligned} p &\succsim p(E)q|E + (1 - p(E))p|\bar{E} \\ \iff q(E)p|E + (1 - q(E))q|\bar{E} &\succsim q. \end{aligned} \quad (8)$$

In other words, if two probability measures disagree only conditional on an event E , then the preference only depends on the conditional measure on that event. In particular, the preference does not depend on the probability of the event they agree on and does not depend on the conditional probability measure on what they agree on.

We can interpret interior separability as a preference axiom for imposing Jeffrey updating: let's say the decision maker initially holds belief $p(E)$ and given that prefers the conditional distributions $p|\bar{E}$ to $q|\bar{E}$. Now the decision maker is forced to update their beliefs of the likelihood of E to $q(E)$ and is informed about various other probabilities unrelated to states in $\bar{E}$. Then the decision maker should still prefer $p|\bar{E}$ over $q|\bar{E}$. Jeffrey updating is however a statement about the optimal distribution and our statement is about preferences.

Theorem 2. *Let Σ contain at least 4 mutually disjoint essential events. Then the following statements are equivalent:*

1. *The relation $\succsim$ on $\text{int}(\mathcal{P})$ fulfills weak order, continuity, and interior separability.*
2. *$\succsim$ can be represented by an α -polarized divergence for some unique $\alpha \in \mathbb{R}$.*

Thus, after weakening updating consistency to Interior Separability, we recover the full parameter space of α -divergences in our characterization. Of course, strict convexity and concavity can be imposed to force a direction for the representation, divergence minimization or maximization, respectively.

4 COMPARATIVE STATICS

We provide comparative statics results for preferences with a negative alpha divergence representation, i.e., preferences that fulfill weak order, continuity, interior separability, and convexity.

The following definition is standard:

Definition 3 (Comparative Convexity). $\succsim_1$ is *more convex than* $\succsim_2$ with respect to p and q if for all $\alpha \in [0, 1]$:

$$\begin{aligned} & \alpha p \oplus (1 - \alpha)q \succsim_1 q \\ \Rightarrow & \alpha p \oplus (1 - \alpha)q \succsim_2 q \end{aligned} \tag{9}$$

We now employ this definition to define what it means that a preference is more sensitive towards lower or higher probabilities. Consult Figure 1: the preferences on the left figure have much more curved indifference curves near the corners of

the simplex while the preferences on the right are almost linear there. The dashed lines partition the graph into areas in which the left decision maker has either more or less convex preferences. This means that the decision maker in the left figure cares much more about how the probability among the low probability states are allocated than the decision maker on the right.

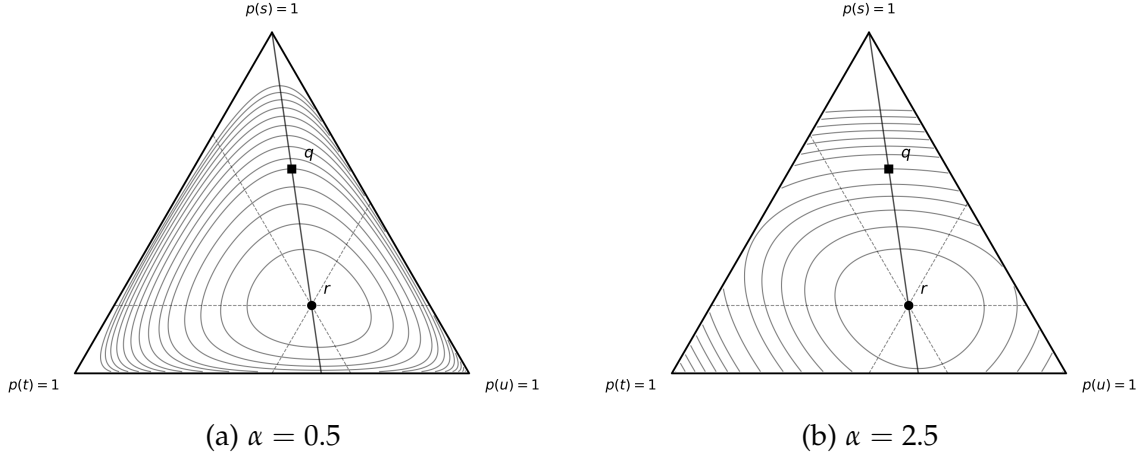


Figure 1: Level curves of $D_\alpha(p||r)$ on the simplex for $r = (0.2, 0.3, 0.5)$. The solid line connects the distributions p for which the ratio $p(t)/p(u)$ equals $r(t)/r(u)$. Dashed lines mark the distributions for which $p(s) = r(s)$, $p(t) = r(t)$, and $p(u) = r(u)$.

Definition 4 (Low Probability Sensitivity). $\succsim_1$ is *more sensitive than* $\succsim_2$ with respect to low probabilities if $\succsim_1$ is more convex than $\succsim_2$ with respect to all p and q such that:

- $p_{\{s,t,u\}} = q_{\{s,t,u\}}$,
- $p(s) > r(s)$, $p(t) < r(t)$, $p(u) < r(u)$,
- $q(s) > r(s)$, and
- $r(t)/r(u) = q(t)/q(u)$

for some $s, t, u \in S$ such that $r(t), r(u) > 0$.

Intuitively, the proposition states that the parameter α describes the degree of convexity of preferences from the reference distribution in the direction of the corners of the simplex: the first restriction of comparative convexity restricts the analysis to the simplex of states s, t, u . The second condition states that the

probability of s is “too high” in p and q compared to that in r and the reverse for t and u . The final restriction states that mixing p and q brings the likelihood ratio of s and t in p and q closer to the likelihood ratio in r .

We now state the corresponding definition for high probabilities. In Figure 1, the decision maker on the left has much flatter indifference curves near the sides of the triangle than the decision maker on the right. We capture this with the following definition:

Definition 5 (High Probability Sensitivity). $\succsim_1$ is *more sensitive than* $\succsim_2$ with respect to high probabilities if $\succsim_1$ is more convex than $\succsim_2$ with respect to all p, q such that:

- $p_{\overline{\{s,t,u\}}} = q_{\overline{\{s,t,u\}}}$,
- $p(s) < r(s), p(t) > r(t), p(u) > r(u)$,
- $q(s) < r(s)$, and
- $r(t)/r(u) = q(t)/q(u)$

for some $s, t, u \in S$ such that $r(t), r(u) > 0$.

In this definition, preferences are more convex when state s has a low probability in p and q compared to r and thus convex combinations combine lotteries with large probabilities.

Proposition 3 (Sensitivity). *Let $\succsim_1$ and $\succsim_2$ be represented by a negative alpha divergence with parameters r, α_1 and r, α_2 , respectively. Then the following statements are equivalent:*

1. $\alpha_1 \geq \alpha_2$.
2. $\succsim_1$ is more sensitive than $\succsim_2$ with respect to large probabilities.
3. $\succsim_2$ is more sensitive than $\succsim_1$ with respect to small probabilities.

5 APPLICATIONS

Alpha divergences are used in a variety of machine learning and statistics applications (e.g., Mironov, 2017). We highlight a few applications here.

5.1 GENERALIZED BAYESIAN UPDATING

Our framework provides a unified perspective on belief updating. Suppose a decision maker holds a prior belief $r \in \mathcal{P}$ and receives information that restricts the set of feasible beliefs to a subset $\mathcal{Q} \subset \mathcal{P}$. If the decision maker chooses the posterior $p \in \mathcal{Q}$ to minimize the α -divergence $D_\alpha(p, r)$, this procedure rationalizes several standard updating rules for $\alpha > 0$:

- Bayesian Updating: If the information comes in the form of an event $E \subseteq S$ such that $\mathcal{Q} = \{p \in \mathcal{P} \mid p(E) = 1\}$, the solution to the minimization problem is the Bayesian posterior $r|E$.
- Jeffrey Updating: If the information is a partition $\{E_1, \dots, E_n\}$ with associated probabilities $\rho_1, \dots, \rho_n$, so that $\mathcal{Q} = \{p \in \mathcal{P} \mid p(E_i) = \rho_i \text{ for all } i\}$, the solution corresponds to Jeffrey updating: $p(s) = \sum_i \rho_i(r|E_i)(s)$.
- Relative Entropy Updating: For $\alpha = 1$ (Kullback-Leibler divergence), this framework recovers the Principle of Minimum Cross-Entropy. For example, if $\mathcal{Q}$ is defined by moment constraints $\mathcal{Q} = \{p \in \mathcal{P} \mid E_p[f(s)] = \eta\}$, the solution is of the exponential form $p(s) = r(s) \exp(\theta f(s) - \psi(\theta))$, where $\psi(\theta) = \ln E_r[\exp(\theta f(s))]$ and θ is chosen such that $\psi'(\theta) = \eta$ (Shore & Johnson, 1980).

A notable feature of the first two cases is that the updated belief is independent of α : the same posterior obtains for any $\alpha > 0$. This α -independence arises because both Bayesian and Jeffrey conditioning impose hard constraints on probability masses, so the curvature of the divergence plays no role in determining the solution. From an economic perspective, this means that any decision maker characterized by Theorem 1— regardless of the particular α governing her utility—will apply Bayesian conditioning when an event is ruled out, and Jeffrey conditioning when partition probabilities are revised. Our characterization thus provides a unified axiomatic foundation for these updating rules grounded in dynamic consistency, without requiring the modeler to take a stand on α .

The framework also rationalizes *maximum a posteriori* (MAP) estimation, a standard inference procedure in Bayesian statistics and machine learning. Suppose $S = X \times Z$, where X is the set of observable outcomes and Z is a set of latent

states. After observing $x \in X$, a decision maker with set of feasible posteriors $\mathcal{Q}_x = \{\delta_x \delta_z \mid z \in Z\}$ selects a point belief $\delta_x \delta_z$. For any such degenerate belief and $\alpha > 0$, the α -divergence $D_\alpha(\delta_x \delta_z, r)$ is strictly decreasing in the joint probability $r(x, z)$, so the minimizer places all weight on

$$z^* \in \arg \max_{z \in Z} r(x, z) = \arg \max_{z \in Z} r(z|x).$$

This is precisely the MAP estimate. An important practical consequence is that the decision is determined entirely by the joint prior $r(x, z)$ —computing the marginal $r(x)$ or the conditional $r(z|x)$ is unnecessary. The result holds for every $\alpha > 0$, so MAP estimation is likewise α -independent.

5.2 VARIATIONAL INFERENCE

In many statistical and machine learning applications, computing the exact posterior $r|E$ is intractable, often due to a high-dimensional integral required for the normalizing constant. Variational Inference (VI) addresses this by approximating the posterior with a distribution from a tractable family $\mathcal{Q}$ (e.g., the family of product distributions in the mean-field approximation).

The standard VI approach finds the member of $\mathcal{Q}$ that minimizes the Kullback-Leibler divergence to the true posterior, $KL(p, r|E)$. Minimizing $KL(p, r|E)$ over $p \in \mathcal{Q}$ is equivalent to minimizing $KL(p, r)$ over $\mathcal{Q}$ because $KL(p, r|E) = KL(p, r) + \ln r(E)$, and $\ln r(E)$ is constant with respect to p . Thus, VI can be viewed as an instance of our divergence minimization framework with $\alpha = 1$. Recent literature (e.g., Hernandez-Lobato et al., 2016) suggests that using general α -divergences can yield better approximations and robustness properties.

5.3 PARAMETRIC AND SEMI-PARAMETRIC ESTIMATION

Our model also provides a decision-theoretic foundation for frequentist estimation methods, where the reference distribution r is taken to be the empirical distribution $\hat{r}_n(s) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{s_i=s}$ generated by an i.i.d. sample. Thus, for the applications in this subsection, our characterization shows that the frequency data r and the robustness preferences, α can be elicited from a decision maker’s model choices.

PARAMETRIC ESTIMATION Consider a parametric family of distributions $\mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$. An estimator $\hat{\theta}$ can be defined as the parameter that minimizes the divergence $D_\alpha(p_\theta, \hat{r}_n)$. For $\alpha = 0$ (the reverse Kullback-Leibler divergence), minimizing $D_0(p_\theta, \hat{r}_n)$ is equivalent to maximizing $\sum_{i=1}^n \ln p_\theta(s_i)$, which is the Maximum Likelihood Estimator (MLE). This connects the standard MLE procedure to the minimization of a reverse KL divergence from the empirical distribution.

SEMI-PARAMETRIC ESTIMATION Consider a model defined by moment restrictions $E_p[g(s, \gamma)] = 0$ for some parameter $\gamma \in \Gamma$. Let $\mathcal{P}_\gamma = \{p \in \mathcal{P} \mid E_p[g(s, \gamma)] = 0\}$. The estimator $\hat{\gamma}$ is chosen to minimize the profile divergence:

$$\hat{\gamma} \in \arg \min_{\gamma \in \Gamma} \left(\min_{p \in \mathcal{P}_\gamma} D_\alpha(p, \hat{r}_n) \right)$$

This formulation corresponds to the class of Generalized Empirical Likelihood (GEL) estimators Newey and Smith (2004). Important special cases include (see Nishimura & Komaki, 2008):

- Empirical Likelihood (EL): For $\alpha = 0$, the estimator corresponds to the Empirical Likelihood estimator (Owen, 1988).
- Exponential Tilting (ET): For $\alpha = 1$, the estimator corresponds to the Exponential Tilting estimator (Kitamura & Stutzer, 1997).

It is worth noting that the exponential tilting estimator corresponds to the reverse updating of the empirical likelihood estimator, and vice versa. Our axiomatic characterization implies that updating consistent preferences over these constrained sets of distributions lead naturally to these estimation procedures.

5.4 MACHINE UNLEARNING

In some cases, the decision maker does not seek to approximate data with a model but has the reverse preference: that the model does not contain certain data that is sensitive or may legally not be used. This corresponds the problem of selectively removing information from a trained model. The emerging literature on machine unlearning provides methods to achieve this and our axioms for the polarized divergence provide a normative foundation for this:

In polarized divergence maximization/minimization, the decision maker maximizes the alpha divergence on some states and minimizes it on others. In large language models, the different states are the output texts of the model given an input text. If an output text is undesirable, it is natural to try to minimize its probability to occur but the model should nonetheless remain functional. Thus, other output probabilities should still remain close to the original probabilities of the model before unlearning.

Standard methods partition the training data into a retain set and a forget set, and combine gradient descent on the retain set with gradient ascent on the forget set (Maini et al., 2024). Zhang et al. (2024) and Wang et al. (2025) replace the gradient ascent with explicit divergence objectives by selectively minimizing and maximizing KL and reverse KL divergence objectives, thus employing effectively a polarized divergence objective. Wang et al. (2025) shows that there is a substantial performance difference between KL and reverse KL objectives in this context, showing that different values of α lead to differences in performance. It is noteworthy that in this case, the subjective component of our representation, r , effectively contains a combination of both the preferences over which data should be forgotten or retained and the empirical probabilities with which the model has been trained. It would be interesting to investigate models in which a separation of the two is possible.

6 RELATED LITERATURE

Our work is related to characterizations of information measures as surveyed by Csiszar (2008). In comparison to this literature we use a preference-based approach instead of directly imposing axioms on a function. Moreover, in our setting one distribution is endogeneously determined while functional characterizations impose assumptions on how the function responds with respect to exogeneous changes in both distributions.

There is a large literature on belief updating rules and their characterizations that usually assume a more complicated primitive. Most closely related is the work of Dominiak et al. (2025) who discuss and characterize updating rules in response to non-partitional information and characterize (among other functional forms) alpha-divergence minimization. They impose conditions on preferences over

Anscombe-Aumann acts and conditional preferences over Anscombe-Aumann acts given convex sets of distributions. The preferences over acts fix the prior (our subjective reference distribution) and the decision maker chooses among posteriors. Our characterization uses a more parsimonious setting requiring only preferences over posteriors and determining the reference distribution endogeneously using effectively only the second part of their primitive. Thus, we simplify the primitive by removing the Anscombe-Aumann structure and allow to elicit the reference distribution of the decision maker from the decision data over distributions.

Zhao (2022) characterizes KL divergence minimization — our $\alpha = 1$ case — using an axiomatic framework for belief revision of an exogenous prior in response to qualitative probability statements of the form “event A is at least as likely as event B .” Within the class of pseudo-Bayesian updating rules (characterized by Exchangeability, Stationarity, and Reduction), the additional axiom of Conservatism—requiring that probability mass be adjusted by just enough to make the two events equally likely—selects KL-divergence minimization. In comparison, while we allow the prior/reference distribution to be endogeneous, we elicit a full preference profile over all distributions and not just probability choices in response to comparative likelihood statements. Having access to preferences over all distributions allows us to determine the prior endogeneously and distinguish whether the decision maker minimizes a KL-divergence or another alpha divergence.

Our results also relate to geometric aggregation of probability distributions (Dietrich, 2021; Dietrich & List, 2016): geometric aggregation can be interpreted as minimizing a weighted average of KL-divergences of the aggregated probability distribution to the individual beliefs.¹ Thus, if each individual’s preference relation over the posterior belief fulfills our representation with $\alpha = 1$, geometric aggregation is simply a weighted utilitarian aggregation.

7 CONCLUSION

This paper provides a normative foundation for alpha divergence minimization. Three axioms — weak order (dynamic consistency of belief rankings), strict convexity, and updating consistency (coherence of conditional evaluations) — are

¹We are grateful to Kensei Nakamura for pointing this out to us.

each individually defensible as requirements on a rational agent provided the state space is complete. Their conjunction uniquely characterizes alpha divergence minimization relative to an endogenously determined reference distribution. Weakening strict convexity to allow concavity and weakening updating consistency to interior separability extends the characterization to polarized divergences and to all $\alpha \in \mathbb{R}$.

As a descriptive model the distinguishing feature of our characterization is that we can elicit the subjective reference distribution r . In the context of updating problems (Bayesian updating, Jeffrey updating, Variational Inference), this corresponds to eliciting the prior. In the context of classical statistics, this corresponds to eliciting the empirical distribution that the decision maker has observed. In the context of machine unlearning, this corresponds to the desirability of different output texts. The comparative statics results show that the elicited parameter α corresponds to whether the decision maker emphasises a better fit on high or low values of r .

A PROOF OF THEOREM 2

Our proof proceeds in three steps. In the first step, we prove that $\succsim$ can be represented by an additively separable utility representation. In the second step, we derive the fundamental equation of information and use it to derive a representation in the form of a Tsallis entropy and an expectation with probability distortions similar to Rommeswinkel (2019). In the third step, we obtain the various functional forms represented in this paper.

A.1 ADDITIVE REPRESENTATION

To obtain $\succsim$ being representable by $U(p) = \sum_x u_x(p(x))$, we employ Theorem 1 of Qin and Rommeswinkel (2022). While they cover an extended real-valued representation on the full simplex, we only require a representation on the interior. Their proof first constructs a real-valued representation on the interior, thus for our purposes it is sufficient to show that their assumptions on the preference relation hold on the interior. On the interior, their separability I condition is implied by interior separability. Our continuity is equivalent to their continuity. What remains to be shown is that our axioms imply their comeasurability and strong essentiality.

Definition 6 (Essentiality). States $s, t \in S$ are *essential* in $\succsim$ if for some p, q , $p|_{\overline{\{s,t\}}} = q|_{\overline{\{s,t\}}}$, we have that $p \succ q$. s, t are *strongly essential* in $\succsim$ if for all $\alpha \in (0, 1)$ there exist p, q with $p|_{\overline{\{s,t\}}} = q|_{\overline{\{s,t\}}}$ such that

$$\alpha \cdot p|_{\{s,t\}} + (1 - \alpha) \cdot p|_{\overline{\{s,t\}}} \succ \alpha \cdot q|_{\{s,t\}} + (1 - \alpha) \cdot q|_{\overline{\{s,t\}}}. \quad (10)$$

We say that a set of states $S = \{s, t, \dots, x\}$ is *pairwise (strongly) essential* if for all pairs $s, t \in S$ such that $s \neq t$, s, t is (strongly) essential.

Since Qin and Rommeswinkel (2022) requires strong essentiality, we prove first that under continuity and interior separability, essentiality and strong essentiality are equivalent:

Lemma 1. *If $\succsim$ fulfills weak order, continuity and interior separability, then s, t are essential if and only if they are strongly essential.*

Proof. $\Leftarrow$ is trivial, we prove $\Rightarrow$: By essentiality, there exist $p \succ q$ with $p|_{\overline{\{s,t\}}} = q|_{\overline{\{s,t\}}}$. Thus, $p|_{\overline{\{s,t\}}} = q|_{\overline{\{s,t\}}}$. If p, q are not interior, then by continuity, we can find

interior $p' \succ q'$ in neighborhoods of p and q . Thus, without loss of generality, p, q are interior. By interior separability, $p \succ q$ if and only if (10) holds. $\square$

Comeasurability and coseparability are defined as follows:

Definition 7 (Comeasurability). $\succsim$ fulfills *comeasurability* if for all $p, p', \bar{p}, \bar{p}', q, q', \bar{q}, \bar{q}', r \in \text{int}(\mathcal{P})$ satisfying

$$p(E) = p'(E) = r(E), \quad q(F) = q'(F) = r(F),$$

$$\bar{p}(G) = \bar{p}'(G) = r(G), \quad \bar{q}(H) = \bar{q}'(H) = r(H)$$

and all mutually disjoint events $E, F, G, H \in \Sigma$,

$$\begin{aligned} &\text{if } p_E q_F r \sim \bar{p}_G \bar{q}_H r, \\ &\quad p'_E q'_F r \sim \bar{p}'_G \bar{q}'_H r, \text{ and} \\ &\quad p_E q'_F r \sim \bar{p}_G \bar{q}'_H r, \\ &\text{then } p'_E q'_F r \sim \bar{p}'_G \bar{q}'_H r. \end{aligned} \tag{11}$$

The special case in which $E = G$ and $F = H$ is called *coseparability* (Qin & Rommeswinkel, 2023). If comeasurability (coseparability) only holds for specific events $E, F, G, H, (E, F)$, we say that it holds with respect to those events.

It will be useful to define for all events E the *conditional relation* $\succsim_E$ over conditional measures on E by: $p|E \succsim_E q|E$ if $r(E)p|E + r(\bar{E})r|\bar{E} \succsim r(E)q|E + r(\bar{E})r|\bar{E}$ for some r such that $r(E) > 0$.

We say that a property of $\succsim$ holds *order-locally* if for all q not $\succsim$ -extremal in $\text{int}(\mathcal{P})$ we can find p, r such that $p \succ q \succ r$ and the property is fulfilled for all measures that are at least as good as r and at most as good as p .

We next prove that order-local comeasurability (coseparability) everywhere implies global comeasurability (coseparability).

Lemma 2 (Extensions of Order-local Comeasurability). *If $\succsim$ fulfills weak order and continuity, then Order-local comeasurability (with respect to E, F, G, H) implies comeasurability (with respect to E, F, G, H). Order-local coseparability (with respect to E, F) implies coseparability (with respect to E, F).*

Proof. Suppose comeasurability with respect to E, F, G, H does not hold for $p, q, r, \bar{p}, \bar{q}$. By continuity and interior separability we can find p''_E strictly between p_E and p'_E in $\succsim_E$ -preference and $\bar{p}''_G$ such that:

$$p''_E q_F r \sim \bar{p}''_G \bar{q}_H r \quad (12)$$

and analogously defined q''_F and $\bar{q}''_H$. From Figure 2 it is easy to see that comeasurability must also be violated for at least one other combination of: p, p', p'', q, q', q'' and $\bar{p}, \bar{p}', \bar{p}'', \bar{q}, \bar{q}', \bar{q}''$. Since $\mathcal{P}$ is topologically separable, finite repetitions of this step yields a violation of comeasurability for any arbitrarily small interval of indifference curves.

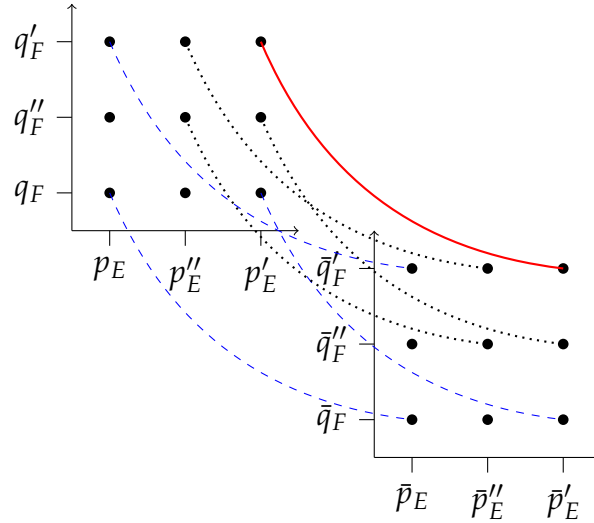


Figure 2: Sketch of Comeasurability Proof

Comeasurability states that the three blue dashed indifferences imply the red solid indifference. If this is violated for the “large squares” involving p, p', q, q' and $\bar{p}, \bar{p}', \bar{q}, \bar{q}'$, then it must also be violated for one of the “smaller squares” involving also p'', q'' and $\bar{p}'', \bar{q}''$. For example, if comeasurability holds for all small squares except the red indifference, then the marked black dotted indifferences hold. But then comeasurability of this smaller square implies that also the red indifference must hold.

□

Further, we show that comeasurability holds for four (two pairs) of dimensions when sufficiently many additional dimensions are available.

Lemma 3 (Coseparability). *Assume that $\succsim$ satisfies weak order, continuity, and interior separability. Let $E = (s, t)$ and $F = (v, w)$ be two pairs of essential states and let there be at least 8 essential states that are pairwise essential. Then $\succsim$ fulfills coseparability with respect to E and F .*

Proof. Assume the antecedent of (11) holds. Since there are at least 8 pairwise essential states, we can find in the complement of $E \cup F$ two other pairs of essential states which we gather into events G and H . By continuity, weak order, and interior separability we can find utility representations V_E and V_F of $\succsim_E$ and $\succsim_F$, respectively. If $V_E(p_E) - V_E(p'_E)$ and $V_F(p_F) - V_F(p'_F)$ are sufficiently small, then by continuity we have restricted solvability, i.e., we can find r' such that $r_{EUFUGUH} = r'_{EUFUGUH}$ and:

$$\begin{aligned} p'_E q_F r &\sim p_E q_F r'_G \sim \bar{p}'_E \bar{q}_F r \\ p_E q'_F r &\sim p_E q_F r'_H \sim \bar{p}_E \bar{q}'_F r \end{aligned} \quad (13)$$

It follows from interior separability that:

$$p'_E q'_F r \sim p_E q_F r'_G r'_H \sim \bar{p}_E \bar{q}_E r'_G r'_H \sim \bar{p}'_E \bar{q}'_F r \quad (14)$$

as desired. By Lemma 2, order-local coseparability implies global coseparability. $\square$

Having obtained coseparability, we next prove comeasurability for events consisting of pairs of essential states.

Lemma 4 (Coseparability implies Comeasurability). *If $\succsim$ fulfills coseparability for events with two states and there are 8 or more pairwise essential states, then $\succsim$ fulfills comeasurability on these events.*

Proof. We provide the proof for exactly 8 states, the proof becomes easier the more states are available as we will see shortly.

Suppose $E \cup F \cup G \cup H$ contains 6 states or less and I contains two of the remaining states. (Consider Figure 3 as an example). Then order-local comeasurability with respect to E, F, G, H is equivalent to order-local comeasurability with respect to E, I, G, I : Via continuity, locally we can find $q''_I, \bar{q}''_I$ such that $p_E q_F q''_I r \sim p_E q'_F r \sim \bar{p}_G \bar{q}_H \bar{q}''_I r$ and by interior separability (11) holds for E, F, G, H

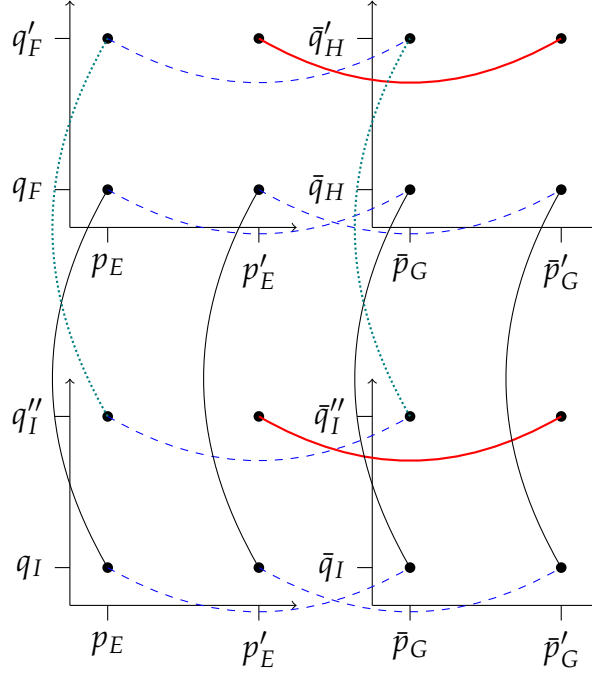


Figure 3: Translating Comeasurability Across Events

The black solid lines connect identical alternatives in the different coordinate systems. The green dotted indifference curves hold by construction using solvability. The blue dashed indifference curves therefore hold at the bottom if they hold at the top. By interior separability the red indifference curve holds at the bottom if and only if it holds at the top.

if and only if it holds for E, I, G, I . Now $E \cup G \cup I$ contains less than 6 states and therefore by the same argument order-local comeasurability with respect to E, I, G, I is equivalent to order-local comeasurability with respect to J, I, J, I for some J , which is implied by coseparability.

If $E \cup F \cup G \cup H$ contains 7 states or more, at most one of E, F can intersect with at most one of G, H . Suppose E and G intersect. Then order-local combination with respect to E, F, G, H is equivalent to order-local comeasurability with respect to E, H, G, H : By continuity, locally we can find q''_H such that $p_E q_F q''_H r \sim p_E q'_F r$ and by interior separability (11) holds for E, F, G, H if and only if it holds for E, H, G, H . Since $E \cup G \cup H$ has at most 6 states, order-local comeasurability with respect to E, H, G, H is equivalent to order-local comeasurability with respect to I, H, I, H for some I .

We have thus proven order-local comeasurability which by Lemma 2 is equivalent

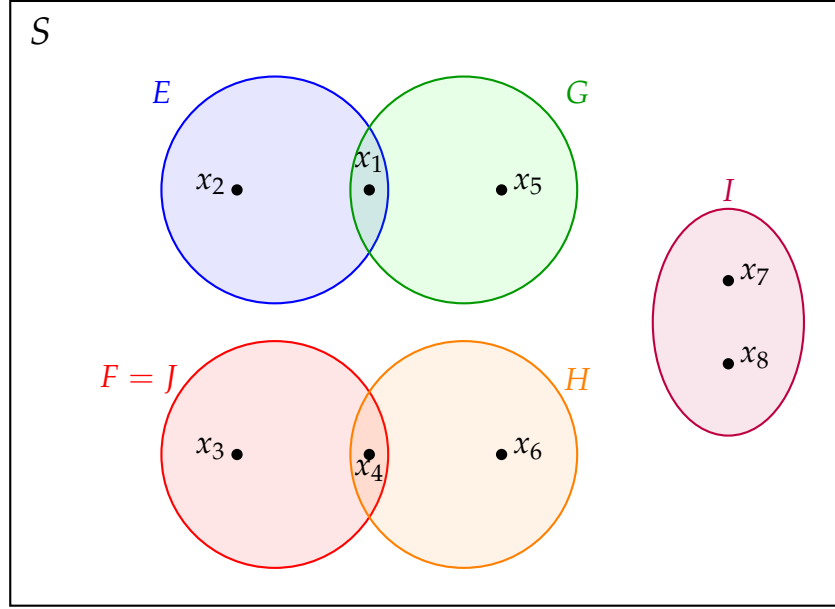


Figure 4: Local Coseparability to Local Comeasurability
 Example of case with two “free” states x_7 and x_8 . Comeasurability over E, F, G, H can be first translated into Comeasurability over E, I, G, I and then into Coseparability over I and J .

lent to comeasurability □

We now show that comeasurability holds for arbitrary events.

Lemma 5 (Comeasurability). *If $\succsim$ fulfills weak order, continuity, and interior separability, and there are four disjoint essential events, then $\succsim$ fulfills comeasurability.*

Proof. If (11) is nontrivial, then E, F, G , and H must be essential. If an event is essential, by interior separability it must contain an essential subevent with only two states. Let E', F', G', H' be such subevents, respectively. Then we can find locally $p''_{E'}, q''_{F'}, \bar{p}''_{G'}, \bar{q}''_{H'}$ and r'' such that $p'_E q'_F r' \sim p''_{E'} q_{F'} r'' \sim \bar{p}''_{G'} q_{F'} r''$. By interior separability (11) holds for E', F', G', H' if and only if it holds for E, F, G, H . Thus, order-local comeasurability with respect to two-state events implies order-local comeasurability. By Lemma 2, comeasurability of $\succsim$ follows. □

We have thus proven that if $\succsim$ fulfills weak order, continuity, and interior separability, then it fulfills the assumptions of Theorem 1 of Qin and Rommeswinkel (2022). Thus, on the interior $\succsim$ can be represented by an additive utility representation.

A.2 FUNDAMENTAL EQUATION OF INFORMATION

Lemma 6 (Affinity of Additive Components). *In the representation $U(p) = \sum_s f_s(p(s))$, if $E \subset S$ contains at least two essential, disjoint subevents, then for all $\psi \in (0, 1)$:*

$$\sum_{s \in E} f_s(p(s)) = B(\psi) + \psi^\alpha \sum_{s \in E} f_s(\psi \cdot p(s)) \quad (15)$$

for some $\alpha \in \mathbb{R}$.

Proof. Suppose $E \subset S$ contains at least two essential, disjoint subevents. By interior separability, if $p(E) = q(E) < 1$, $p'(E) = q'(E) = \psi p(E)$ and $p|\bar{E} = q|\bar{E} = p'|\bar{E} = q'|\bar{E}$, then

$$\begin{aligned} \sum_{s \in E} f_s(p(s)) \geq \sum_{s \in E} f_s(q(s)) &\Leftrightarrow U(p) \geq U(q) \\ &\Leftrightarrow p \succsim q \\ &\Leftrightarrow p' \succsim q' \\ &\Leftrightarrow U(p') \geq U(q') \\ \Leftrightarrow \sum_{s \in E} f_s(p'(s)) \geq \sum_{s \in E} f_s(q'(s)) &\Leftrightarrow \sum_{s \in E} f_s(\psi \cdot p(s)) \geq \sum_{s \in E} f_s(\psi \cdot q(s)) \end{aligned} \quad (16)$$

And thus the first and last line are additive representations over the same relation on the set $\{\hat{p} \in \text{int}(\mathcal{P}) | \hat{p}(E) = p(E)\}$. By the uniqueness of additive representations (Theorem 1 of Qin & Rommeswinkel, 2022), if E contains two essential subevents, then these representations must be affine transformations of another, thus for all ψ :

$$\sum_{s \in E} f_s(q(s)) = B(\psi) + A(\psi) \cdot \sum_{s \in E} f_s(\psi \cdot q(s)) \quad (17)$$

for some continuous functions A, B . Applying this condition twice yields:

$$B(\psi\zeta) + A(\psi\zeta) \cdot \sum_{s \in E} f_s(\psi\zeta \cdot q(s)) = B(\psi) + A(\psi)B(\zeta) + A(\psi) \cdot A(\zeta) \sum_{s \in E} f_s(\psi\zeta \cdot q(s)). \quad (18)$$

Thus, $A(\psi\zeta) = A(\psi)A(\zeta)$. This Cauchy equation has the solution $A(\psi) = \psi^\alpha$ for some $\alpha \in \mathbb{R}$. $\square$

Notice that if E, F are pairwise disjoint, essential events and $E \cup F \subset S$, then the previous lemma holds for $E \cup F$, and thus from (15):

$$\sum_{s \in E} f_s(q(s)) = B(\psi) + \psi^\alpha \cdot \sum_{s \in E} f_s(\psi \cdot q(s)) - \sum_{s \in F} f_s(q(s)) - \psi^\alpha \sum_{s \in F} f_s(\psi \cdot q(s)) \quad (19)$$

Holding fixed $q|_F$, we see that (15) holds for any essential event such that there exists a disjoint subevent F with $E \cup F \subset S$.

Lemma 7. *If $\succsim$ fulfills Weak Order, Continuity, and Interior Separability and has at least 4 disjoint, essential subevents, then $\succsim$ can be represented on $\text{int}(\mathcal{P})$ by:*

$$V(p) = \begin{cases} \sum_s p(s) \eta_s + \eta p(s) \ln p(s), & \alpha = 1 \\ \sum_s \eta_s \ln p(s), & \alpha = 0 \\ \sum_s p(s)^\alpha \eta_s, & \text{else.} \end{cases} \quad (20)$$

For all $s \in X$, $\eta_s \in \mathbb{R}$, $\eta \in \mathbb{R}$, $\alpha \in \mathbb{R}$. V is unique up to linear transformations except if $\alpha = 1$ and $\eta = 0$ in which case it is unique up to affine transformations.

Proof. Let $\{s, t\}$ and $\{w, x\}$ be essential. By Interior Separability, without loss of generality, three states are pairwise essential, let's say s, t, w . We first derive the fundamental equation of information by using Lemma 6 twice with $\psi = \frac{1}{1-p(s)}$ and $\psi = \frac{1}{1-p(t)}$. Since $\succsim$ contains 4 essential subevents, we can apply 6 to both $\{s, w\}$ and $\{t, w\}$ to obtain:

$$\begin{aligned} & f_s(p(s)) + (1 - p(s))^{-\alpha} \left(f_t\left(\frac{p(t)}{1 - p(s)}\right) + f_w\left(\frac{1 - p(s) - p(t)}{1 - p(s)}\right) \right) + B_s(1 - p(s)) \\ &= f_t(p(t)) + (1 - p(t))^{-\alpha} \left(f_s\left(\frac{p(s)}{1 - p(t)}\right) + f_w\left(\frac{1 - p(s) - p(t)}{1 - p(t)}\right) \right) + B_t(1 - p(t)) \end{aligned} \quad (21)$$

Defining $f(q) = f_s(q) + B_s(1 - q)$, $g(q) = f_t(q) + f_w(1 - q)$, $h(q) = f_t(q) + B_t(1 - q)$, and $k(q) = f_s(q) + f_w(1 - q)$, the continuous solutions (e.g., Ebanks et al.,

1987) to the resulting functional equation are:

$$\begin{aligned}
 f(q) &= \begin{cases} q\eta_6 \ln q + (1-q)\eta_6 \ln(1-q) + \eta_3 q - \eta_2(1-q) + \eta_5, & \alpha = 1 \\ \eta_4 \ln 1-q + \eta_5 \ln 1-q + \eta_6 \ln q + \eta_2 + \eta_3, & \alpha = 0 \\ \eta_4 q^\alpha - \eta_3(1-q)^\alpha + \eta_6, & \text{else,} \end{cases} \\
 g(q) &= \begin{cases} q\eta_6 \ln q + (1-q)\eta_6 \ln(1-q) + \eta_1 q + \eta_2, & \alpha = 1 \\ \eta_5 \ln 1-q + \eta_4 \ln q + \eta_1, & \alpha = 0 \\ \eta_1 q^\alpha - \eta_2(1-q)^\alpha + \eta_3, & \text{else.} \end{cases}
 \end{aligned} \tag{22}$$

Analogous functional forms hold for g, h, k . Substituting the solution yields for the different cases:

— $\alpha = 1$:

$$\begin{aligned}
 U_{s,t,w}(p) &= p(s) ((\eta_3 + \eta_5) + \eta_6 \ln p(s)) \\
 &\quad + p(t) ((\eta_1 + \eta_5) + \eta_6 \ln p(t)) \\
 &\quad + p(w) (\eta_5 + \eta_6 \ln p(w))
 \end{aligned} \tag{23}$$

— $\alpha = 0$:

$$\begin{aligned}
 U(p) &= \eta_6 \ln p(s) + \eta_5 \ln p(w) + \eta_4 \ln p(t) \\
 &\quad + \eta_1 + \eta_2 + \eta_3
 \end{aligned} \tag{24}$$

— $0 \neq \alpha \neq 1$:

$$U(p) = \eta_4 p(s)^\alpha + \eta_1 p(t)^\alpha + \eta_2 p(w)^\alpha + \eta_6 \tag{25}$$

Thus,

$$f_s(p(s)) = \begin{cases} p(s)u_s + \eta p(s) \ln p(s), & \alpha = 1 \\ \eta_s \ln p(s), & \alpha = 0 \\ \eta_s p(s)^\alpha, & \text{else.} \end{cases} \tag{26}$$

where we have omitted constant terms that do not affect utility comparisons and renamed the parameters. $\square$

A.3 FUNCTIONAL FORMS

Lemma 8 (Divergence Decomposition). *If*

$$V(p) = \begin{cases} \sum_s p(s) \eta_s + \eta p(s) \ln p(s), & \alpha = 1 \\ \sum_s \eta_s \ln p(s), & \alpha = 0 \\ \sum_s p(s)^\alpha \eta_s, & \text{else,} \end{cases} \quad (27)$$

represents $\succsim$ uniquely up to increasing linear transformations on $\text{int}(\mathcal{P})$, then we can find a partition $\{X^-, X^0, X^+\}$ of X and a unique probability measure q on X such that for some $\alpha \in \mathbb{R}$

$$U(p : q) = \begin{cases} \sum_{s \in X^+} p(s) \ln \frac{p(s)}{q(s)} - \sum_{s \in X^-} p(s) \ln \frac{p(s)}{q(s)}, & \alpha = 1 \\ \sum_{s \in X^+} q(s) \ln \frac{q(s)}{p(s)} - \sum_{s \in X^-} q(s) \ln \frac{q(s)}{p(s)}, & \alpha = 0 \\ \frac{1}{\alpha(\alpha-1)} \left(\sum_{s \in X^+} \frac{p(s)^\alpha}{q(s)^{\alpha-1}} - \sum_{s \in X^-} \frac{p(s)^\alpha}{q(s)^{\alpha-1}} - 1 \right), & \text{else,} \end{cases} \quad (28)$$

represents $\succsim$. q is unique on X^+ and X^- . If $\alpha \in (0, 1)$ then either $X^+ = X$ or $X^- = X$.

Proof. We distinguish three cases for α .

— $\alpha = 1$: Define:

$$\begin{aligned} \theta &= -\ln \sum_{x \in X} \exp(-\eta_x / \eta) \\ q(s) &= \exp(\theta - \eta_s / \eta) \end{aligned} \quad (29)$$

If $\eta > 0$, define

$$\begin{aligned} X^+ &= X \\ X^- &= \emptyset \\ U &= V / |\eta| - \theta \end{aligned} \quad (30)$$

to obtain the desired representation U . If $\eta < 0$, define

$$\begin{aligned} X^+ &= \emptyset \\ X^- &= X \\ U &= V/|\eta| + \theta \end{aligned} \tag{31}$$

to obtain the desired representation U .

If $\eta = 0$, then the representation is not unique up to linear transformations but only unique up to affine transformations. Thus, $\eta = 0$ is excluded by the uniqueness condition.

— $\alpha = 0$: Define

$$\begin{aligned} \theta &= 1 / \sum_{x \in X} |\eta_x| \\ q(s) &= |\eta_s| \cdot \theta \\ X^+ &= \{s | \eta_s < 0\} \\ X^- &= \{s | \eta_s > 0\} \\ X^0 &= \{s | \eta_s = 0\} \\ U &= V \cdot \theta + \sum_{x \in X^+} |\eta_x| \ln |\eta_x| - \sum_{x \in X^-} |\eta_x| \ln |\eta_x|. \end{aligned} \tag{32}$$

to obtain the desired representation.

— $0 \neq \alpha \neq 1$: Define:

$$\begin{aligned} \theta &= \frac{1}{(\sum_s (\alpha(\alpha - 1) \eta_s)^{1/(1-\alpha)})^{1-\alpha}} \\ q(s) &= (\alpha(\alpha - 1) \theta |\eta_s|)^{1/(1-\alpha)} \\ X^+ &= \{s | \eta_s < 0\} \\ X^- &= \{s | \eta_s > 0\} \\ X^0 &= \{s | \eta_s = 0\} \\ U &= V \cdot \theta + \sum_{x \in X^+} |\eta_x| \ln |\eta_x| - \sum_{x \in X^-} |\eta_x| \ln |\eta_x|. \end{aligned} \tag{33}$$

to obtain the desired representation.

The uniqueness properties follow from the uniqueness of V up to linear transformations. Transforming V in any of the above definitions by multiplication with a constant has no effect on q , U , or the partition of the state space.

□

Lemma 9 (Conditional Divergence Decomposition).

$$U(p : q) = \begin{cases} p(X^+)D_1(p|X^+, q|X^+) - p(X^-)D_1(p|X^-, q|X^-) + p(X^+) \ln \frac{p(X^+)}{q(X^+)} - p(X^-) \ln \frac{p(X^-)}{q(X^-)}, \\ q(X^+)D_0(p|X^+, q|X^+) - q(X^-)D_0(p|X^-, q|X^-) + q(X^+) \ln \frac{q(X^+)}{p(X^+)} - q(X^-) \ln \frac{q(X^-)}{p(X^-)}, \\ \frac{p(X^+)^\alpha}{q(X^+)^{\alpha-1}}D_\alpha(p|X^+, q|X^+) - \frac{p(X^-)^\alpha}{q(X^-)^{\alpha-1}}D_\alpha(p|X^-, q|X^-) + \frac{1}{\alpha(\alpha-1)} \left(\frac{p(X^+)^\alpha}{q(X^+)^{\alpha-1}} - \frac{p(X^-)^\alpha}{q(X^-)^{\alpha-1}} - 1 \right) \end{cases} \quad (34)$$

represents $\succsim$. q is unique on X^+ and X^- . If $\alpha \in (0, 1)$ then either $X^+ = X$ or $X^- = X$.

Proof. This lemma straightforward follows from substituting $p(x) = p(X^+)p|X^+(x)$ for $x \in X^+$ and analogous definitions for q and for $x \in X^-$, etc.. □

B PROOF OF THEOREM 1

Updating consistency implies interior separability. Thus we can use Theorem 2 to obtain a polarized α -divergence representation with $\alpha \in \mathbb{R}$. We show that $\alpha > 0$ under updating consistency:

Suppose $p, p' \in \text{int}(\mathcal{P})$ and $p \succ p'$. For some x such that $r(x) > 0$ let $\bar{p}|\overline{\{x\}} = p|\overline{\{x\}}$ and $\bar{p}'|\overline{\{x\}} = p'|\overline{\{x\}}$ but $\bar{p}(x) = 0 = \bar{p}'(x)$. By updating consistency $\bar{p} \succ \bar{p}'$. But for $\alpha \leq 0$, $PD_0(\bar{p}) = PD_0(\bar{p}') \in \{-\infty, \infty\}$, a contradiction. Thus, $\alpha > 0$.

C PROOFS OF PROPOSITIONS 1 AND 2

Sketch of Proposition 1. We first use Theorem 2 to obtain a polarized divergence representation. Since $\alpha \leq 0$ violates updating consistency, it follows that $\alpha > 0$.

Note that $D_\alpha(p, q)$ is convex in p for all $\alpha \in \mathbb{R}$. It follows that D_α is quasiconvex in its first argument, i.e., $D_\alpha(\mu p + (1 - \mu)p', q) \leq D_\alpha(p, q)$ for all p, p' such that $D_\alpha(p, q) = D_\alpha(p', q)$.

If q is nontrivial, then for some p, p' such that $D_\alpha(p, q) = D_\alpha(p', q)$ we have that $D_\alpha(\mu p + (1 - \mu)p', q) < D_\alpha(p, q)$ and thus D_α is not quasiconcave.

It follows that if $\alpha = 1$ and $\succsim$ is strictly convex, then $X^+ = \emptyset$. If $\alpha \neq 1$, whenever X^+ is not empty, then strict convexity is violated. $\square$

Sketch of Proposition 2. The proof is analogous to the proof of Proposition 1, replacing strict convexity with strict concavity. Since $D_\alpha(p, q)$ is strictly convex in p but not quasiconcave, strict concavity of $\succsim$ forces $X^- = \emptyset$, yielding the representation $U(p) = D_\alpha(p, r)$. $\square$

D PROOF OF PROPOSITION 3

Proof. Suppose for the moment that r, q have unit probability on $E = \{s, t, u\}$. Suppose $q(s) > r(s)$, i.e., s is a state with a “too high” probability. Further, $q(t)/r(t) = q(u)/r(u)$, thus states t and u have “too low” probabilities.

The indifference curves of both decision makers are tangent to the line of points p' such that $p'(s) = q(s)$. Thus, all improvements p fulfill $p(s) < q(s)$. Because preferences are convex for both decision makers, we only need to show that for decision maker 1 the indifference curve is more curved than for decision maker 2. Concretely, we show that for every $p \sim_2 q$, the MRS between adding additional probability to state s (reducing u) versus adding probability to state t (reducing u) is greater for decision maker 1 than for decision maker 2 if and only if $p(t) > q(t)$, and $p(u) < q(u)$.

Let $p(u) = 1 - p(s) - p(t)$, then:

$$MRS_i^{st}(p) = \frac{\partial U_i(p)/\partial p(s)}{\partial U_i(p)/\partial p(t)} = \frac{\left(\frac{q(s)}{r(s)}\right)^{\alpha_i-1} - \left(\frac{q(u)}{r(u)}\right)^{\alpha_i-1}}{\left(\frac{q(t)}{r(t)}\right)^{\alpha_i-1} - \left(\frac{q(u)}{r(u)}\right)^{\alpha_i-1}} \quad (35)$$

Note that the sign changes at the top and bottom due to $1 \geq \alpha_i \geq 0$ always cancel. Thus, the MRS is nonnegative if and only if $p(t)/r(t) \geq p(u)/r(u)$. We show that the MRS decreases in α_i in this case.

Since $p(s) > r(s)$, $p(u) < r(u)$, the numerator is increasing in α . The denomi-

nator is decreasing in α if and only if:

$$\left(\frac{q(s)}{r(s)}\right)^{\alpha_i-1} \ln \frac{q(s)}{r(s)} < \left(\frac{q(u)}{r(u)}\right)^{\alpha_i-1} \ln \frac{q(u)}{r(u)} \Leftrightarrow \frac{p(t)}{r(t)} < \frac{p(u)}{r(u)}. \quad (36)$$

It follows that when $p(t)/r(t) > q(t)/r(t)$:

$$\begin{aligned} \alpha_1 &\geq \alpha_2 \\ \Leftrightarrow MRS_1^{st}(p) &< MRS_2^{st}(p) \end{aligned} \quad (37)$$

Define the set of probability distributions $P_r^q = \{p | p(E) = 1 \wedge p(s) < q(s) \wedge p(t) < r(t) \wedge p(u) < r(u)\}$ in which the probabilities of s are smaller than $q(s)$ and those of the other two states are bounded by r .

Thus, applying a symmetric argument with the relative likelihoods of t and u reversed, we obtain:

$$\begin{aligned} \alpha_1 &\geq \alpha_2 \\ \Leftrightarrow \{p \in P_r^q | p \succsim_2 q\} &\subseteq \{p \in P_r^q | p \succsim_1 q\} \end{aligned} \quad (38)$$

and thus, whenever $p \succsim_2 q$, then $p \succsim_1 q$ given the imposed constraints on the probabilities.

We have only shown this result for three states with positive probability. Note that if α_i is greater than zero, the result directly extends via updating consistency of the preferences. Similarly, if α_i is less than zero the results extend because the utility representation over all states is (for interior probabilities) increasing in the sum of the additive components over three states.

The equivalence of comparative high probability sensitivity and the comparison of α follows by a symmetric argument; we can directly see in equation (36) that the sign of the MRS and comparative statics with respect to α are exactly opposite when $q(s) < r(s)$, $p(t) > r(t)$, and $p(u) > r(u)$.

We finally make the following observation: If $\alpha_1 > \alpha_2$, then except when the relative probability of $p(t)/p(u)$ exactly matches $r(t)/r(u)$, we have that $MRS_1^{st}(p) \neq MRS_2^{st}(p)$, and thus, indifference curves always cross. This means that convexity comparisons between decision makers are indeed only possible for

distributions in which $q(t)/q(u) = r(t)/r(u)$. Conditions that do not assume a common r and/or only involve arbitrary probability distributions would therefore need to be considerably more complicated. $\square$

REFERENCES

- Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J.R. Stat. Soc. Ser. B*, 28, 131–142.
- Csiszar, I. (2008). Axiomatic characterizations of information measures. *Entropy*, 10, 261–273.
- Dietrich, F. (2021). Fully bayesian aggregation. *Journal of Economic Theory*, 194.
- Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. In A. Hajek & C. Hitchcock (Eds.), *Oxford handbook of philosophy and probability*. Oxford University Press.
- Dominiak, A., Kovach, M., & Tserenjigmid, G. (2025). *Inertial updating with general information* [Working Paper].
- Ebanks, B. R., Kannappan, P., & Ng, C. T. (1987). Generalized fundamental equation of information of multiplicative type. *Aequationes Mathematicae*, 32(1), 19–31. <https://doi.org/10.1007/BF02311295>
- Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernandez-Lobato, D., & Turner, R. (2016). Black-box alpha divergence minimization. *Proceedings of the 33rd International Conference on Machine Learning*, 48, 1511–1520.
- Kitamura, Y., & Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4), 861–874.
- Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., & Kolter, J. Z. (2024). TOFU: A task of fictitious unlearning for LLMs. *Conference on Language Modeling (COLM)*.
- Mironov, I. (2017). Rényi differential privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 263–275. <https://doi.org/10.1109/CSF.2017.11>
- Newey, W., & Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood. *Econometrica*, 65, 861–874.

- Nishimura, T., & Komaki, F. (2008). The information geometric structure of generalized empirical likelihood estimators. *Communications in Statistics - Theory and Methods*, 37(12), 1867–1879.
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2), 237–249.
- Qin, W.-Z., & Rommeswinkel, H. (2022). Additive representations on a simplex. *Journal of Mathematical Economics*, 103.
- Qin, W.-z., & Rommeswinkel, H. (2023). Quasi-separable Preferences [forthcoming]. *Theory and Decision*.
- Rommeswinkel, H. (2019). *Procedural Mixture Sets*.
- Shore, J. E., & Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*.
- Wang, B., Zi, Y., Sun, Y., Zhao, Y., & Qin, B. (2025). RKLD: Reverse KL-divergence-based knowledge distillation for unlearning personal information in large language models. *Proceedings of the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.
- Zhang, R., Lin, L., Bai, Y., & Mei, S. (2024). Negative preference optimization: From catastrophic collapse to effective unlearning. *Conference on Language Modeling (COLM)*.
- Zhao, C. (2022). Pseudo-Bayesian updating. *Theoretical Economics*, 17(2), 709–742. <https://doi.org/10.3982/TE4849>