

Alpha Divergence Representations of Dynamically Consistent Belief Selection

Rommewinkel, Hendrik Ui, Takashi

Hitotsubashi University & WIAS

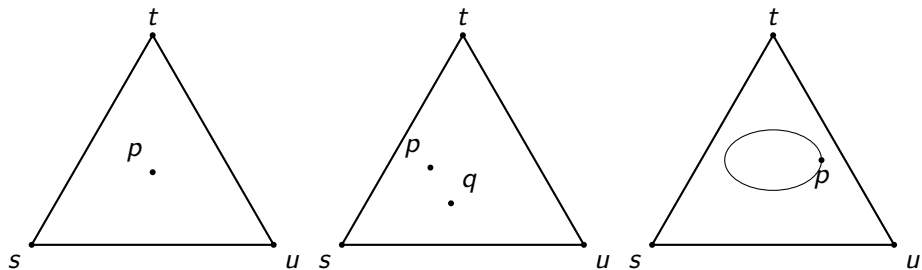
Kanagawa University

July 2026

Table of Contents

- 1 Introduction
- 2 Representation Theorems
- 3 Proof Sketch
- 4 Applications
- 5 Conclusion

Belief Selection



Motivation: Belief Selection Problems

- **Belief Updating:**

- ▶ A decision maker (DM) has a prior belief.
- ▶ DM learns information (events, probabilities).
- ▶ DM chooses new belief via Bayesian (Jeffrey) updating.

- **Classical Statistics:**

- ▶ DM has collected data from a state space.
- ▶ DM is deciding between various parametrizations of a class of models (together with a distribution over the exogeneous variables)
- ▶ DM chooses to fit the parameters using MLE and to use his frequency data as the distribution over exogeneous variables.

- **Machine Learning:**

- ▶ DM has a prior belief
- ▶ DM learns information but cannot compute posterior.
- ▶ DM uses variational inference to approximate posterior.

Example: Maximum a posteriori estimation (MAP)

- $S = X \times Z$
- $X = \{\text{fever+cough, fever w/o cough, no fever}\}$
- $Z = \{\text{influenza, COVID-19, common cold, allergy}\}$

$r(x,z)$	fever, cough	fever, no cough	no fever
influenza	0.1050	0.0300	0.0150
COVID-19	0.0500	0.0100	0.0400
common cold	0.0500	0.0125	0.1875
allergy	0.0100	0.0050	0.4850

Motivation

- **Framework:**

- ▶ Finite state space S .
- ▶ Choices over probability distributions P_T on subsets T of S .
- ▶ All choice sets have a fixed common support for all distributions.

- **Model:**

- ▶ Decision maker has a subjective reference distribution r
- ▶ Decision maker faces constraint set of probability distributions.
- ▶ Decision maker minimizes distance from r as measured by an α -divergence.

Notation

- Finite state space S .
- For nonempty $T \subseteq S$: $P_T = \{p \in P \mid \text{supp}(p) = T\}$.
- $p|E$ is the conditional measure of p on E .
- $\mathcal{A}_T$: set of nonempty compact menus within P_T .
- $\mathcal{A} = \bigcup_{\emptyset \neq T \subseteq S} \mathcal{A}_T$.
- Choice correspondence $C : \mathcal{A} \rightrightarrows P$ with $\emptyset \neq C(A) \subseteq A$.
- Revealed preference on P_T : $p \succsim_T^C q \iff p \in C(\{p, q\})$.

α -Divergence

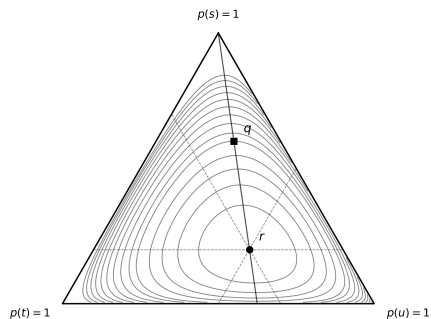
The α -**divergence** from p to q is defined as:

$$D_{\alpha}(p, q) = \frac{1}{\alpha(\alpha - 1)} \left(\sum_{x \in \text{supp}(q)} q(x) \left(\frac{p(x)}{q(x)} \right)^{\alpha} - 1 \right)$$

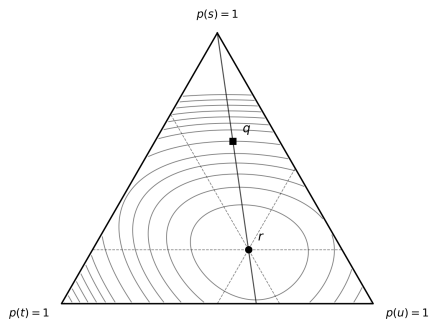
for $\alpha \neq 0, 1$ and p, q with identical support.

- $\alpha = 1$: Kullback-Leibler divergence $KL(p, q)$.
- $\alpha = 0$: Reverse KL divergence $KL(q, p)$.

Indifference curves



(a) $\alpha = 0.5$



(b) $\alpha = 2.5$

Interpretation in Context of Applications

Application	Reference Dist.	Constraint	Behavior
Belief updating	Prior	Information	Bayesian, Jeffrey & relative entropy updating
Classical statistics	Freq. data	Models	MLE, subclass of GEL
Machine Learning	Prior	Information & tractable model	Variational Inference, Mean field approximation
Large Language Models	Utility	Model instances	Unlearning

Dynamic Consistency in Belief Selection

- Rationalizability means the DM is not “paranoid” about constraint set.
- But then also preference for $p|E$ and p should be related!

$r(x,z)$	fever, cough	fever, no cough	no fever
influenza	0.1050	0.0300	0.0150
COVID-19	0.0500	0.0100	0.0400
common cold	0.0500	0.0125	0.1875
allergy	0.0100	0.0050	0.4850

Technical Axioms

Axiom (Within-support rationalizability)

For each nonempty $T \subseteq S$ and all $A \in \mathcal{A}_T$:

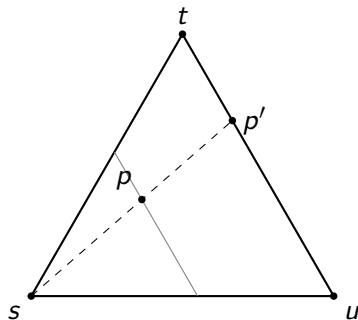
$$C(A) = \{p \in A : p \succsim_T^C q \text{ for all } q \in A\}.$$

Axiom (Continuity)

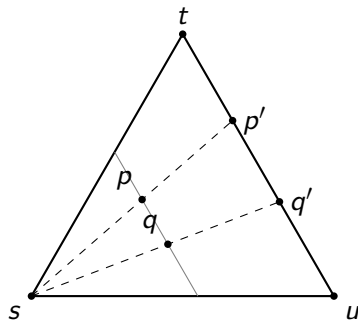
For sequences p^k, q^k within P_T , if $p^k \rightarrow p \in P_T$ and $q^k \rightarrow q \in P_T$, then $p^k \in C(\{p^k, q^k\})$ for all k implies $p \in C(\{p, q\})$.

Updating Consistency

Bayesian Updating



Updating Consistency



Updating consistency

Axiom (Updating Consistency)

Let $\emptyset \neq E \subset T \subseteq S$. For $p, q \in P_T$ with

$$p(E) = q(E) \quad \text{and} \quad p|_{\overline{E}} = q|_{\overline{E}},$$

it holds that

$$p|_E \in C(\{p|_E, q|_E\}) \iff p \in C(\{p, q\}).$$

Strict convexity

Axiom (Strict Convexity)

If $p, q \in P_T$, $p \neq q$, $C(\{p, q\}) = \{p, q\}$, then for $\lambda \in (0, 1)$,

$$C(\{p, \lambda p + (1 - \lambda)q\}) = \{\lambda p + (1 - \lambda)q\}.$$

Essentiality

Definition (Essential event)

A state $s \in S$ is *essential* (in C) if there exist a $t \in S$ and $p, q \in P_T$ with $p(w) = q(w)$ for all $w \neq s, t$ such that

$$C(\{p, q\}) = \{p\}.$$

s and t are then *pairwise essential*.

Main Results

- Main Result 1:

- ▶ Within-Support Rationalizability, Continuity, Updating Consistency, Strict Convexity hold.
- ▶ $\Longleftrightarrow$ DM minimizes on each support T the α -divergence from a *subjective* reference distribution r .

- Main Result 2:

- ▶ Within-Support Rationalizability, Continuity, Updating Consistency hold.
- ▶ $\Longleftrightarrow$ DM partly minimizes, partly maximizes the α -divergence from a *subjective* reference distribution r .

Both require at least 8 (pairwise essential) states.

Result: Divergence Minimization

Theorem

Let S contain at least 8 states. The following are equivalent:

- 1 C satisfies Within-Support Rationalizability, Continuity, Updating Consistency, and Strict Convexity.
- 2 There exist a unique full-support reference distribution $r \in P_S$ and a unique $\alpha \in \mathbb{R}$ such that for every nonempty $T \subseteq S$ and every $A \in \mathcal{A}_T$:

$$C(A) = \arg \min_{p \in A} D_\alpha(p, r|T)$$

Polarized Divergence

Definition (Polarized α -divergence)

Given a partition $\{R^+, R^-\}$ of S a polarized divergence is:

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

- On R^+ : Maximize divergence.
- On R^- : Minimize divergence.
- For $\alpha = 1$ and $\alpha = 0$, one of the two sets is empty.

Result: Divergence Minimization and Maximization

Theorem

Let S contain at least 8 pairwise essential states and all states are essential. The following are equivalent:

- ① *C satisfies Within-Support Rationalizability, Continuity, and Updating Consistency.*
- ② *Either C maximizes nontrivial expected utility, or there exist a unique $r \in P_S$, a unique set $R^+ \subseteq S$ and a unique $\alpha \in \mathbb{R}$ such that for every nonempty $T \subseteq S$ and every $A \in \mathcal{A}_T$:*

$$C(A) = \arg \max_{p \in A} \text{PD}_\alpha^{R^+}(p, r | T)$$

Proof Sketch: From Choices to Preferences

- Within-support rationalizability and Continuity: each revealed relation $\succsim_T^C$ is a continuous weak order rationalizing C on $\mathcal{A}_T$. Let R be the largest essential event.
- Updating consistency generates interior separability of $\succsim_R^C$: for $p, q, p', q' \in P_R$ with
 - ▶ $p(E) = q(E), p'(E) = q'(E),$
 - ▶ $p|\bar{E} = q|\bar{E}, p'|\bar{E} = q'|\bar{E},$
 - ▶ $p|E = p'|E, q|E = q'|E,$ we have that:
$$p \succsim_S^C q \iff p' \succsim_S^C q'$$
- All $\succsim_T^C$ are fully determined by $\succsim_R^C$.

Proof Sketch: Representation

- Use Qin and Rommeswinkel (2022) for additive representation of $\succsim_S^C$ on the (interior of the) simplex.

$$U(p) = \sum_{s \in S} u_s(p(s))$$

- Updating consistency generates fundamental equation of information.
- Use Ebanks et al. (1987) to obtain solution to fundamental equation of information.
- Identify r from rearranging terms: representation of $\succsim_S^C$ by $-D_\alpha(\cdot, r)$.
- Transfer to $T \subsetneq S$: Let $m \in P_{S \setminus T}$; Updating Consistency gives $p \succsim_T^C q \iff \mu p + (1 - \mu)m \succsim_S^C \mu q + (1 - \mu)m$, and

$$D_\alpha(\mu p + (1 - \mu)m, r) = \mu^\alpha r(T)^{1-\alpha} D_\alpha(p, r|T) + K(m, \mu).$$

Variational Inference

- For $\alpha > 0$ we can combine the induced preferences $\succsim_T^C$ into one global continuous preference relation consistent with the preference given each T .
- The case $\alpha = 1$ then corresponds to variational inference.
- Being provided with information $T \subset S$ and a class of admissible models, the decision maker minimizes the KL divergence from the prior belief to the distribution induced by the admissible models.

Application: Bayesian Updating

Suppose DM has belief r . Based on new information $T \subset R$, DM chooses a belief $p \in P_T$ by solving

$$\min_{p \in P_T} D_\alpha(p, r | T)$$

For all α , optimization (FOCs) yields $r|T$ as the maximizer: Bayesian updating of r .

Application: Jeffrey Updating

Suppose DM has belief r . Based on new information

$$\mathcal{Q} = \{p \in P_S \mid p(E_i) = \rho_i > 0 \text{ for all } i \in \{1, \dots, n\}\},$$

where $\Pi = \{E_1, \dots, E_n\}$ is a partition of S and $\rho = \Delta(\{1, \dots, n\})$.

DM chooses a belief $p \in P_S$ by solving

$$\min_{p \in P_S} D_\alpha(p, r \mid T)$$

For all α , this is Jeffrey updating, the updated belief is $\sum_i \rho_i r \mid E_i$

Application: Parametric Estimation

DM observes an i.i.d. sequence $s_1, \dots, s_n \in S$.

DM's reference belief is the empirical distribution $\hat{r}$.

$$\hat{r}(s) = \#\{i \mid s_i = s\} / n$$

Constraint: p is a member of a parametric family (DM's model)

$$\mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\} \subset P_S$$

(possibly including a distribution of exogenous variables).

DM's estimator $\hat{\theta}$ is a solution to

$$\min_{\theta \in \Theta} D_\alpha(p_\theta, \hat{r})$$

Application: Parametric Estimation

When $\alpha = 0$, $\hat{\theta}$ is the MLE.

$$D_0(p_\theta, \hat{r}) = KL(\hat{r}, p_\theta) = -n^{-1} \sum_{i=1}^n \log p_\theta(s_i) - H(\hat{r})$$

$$\min_{\theta \in \Theta} D_0(p_\theta, \hat{r}) \Rightarrow \max_{\theta \in \Theta} \sum_{i=1}^n \log p_\theta(s_i) \Rightarrow \hat{\theta}$$

Upshot: Regression and Updating Consistency

- MLE minimizes α -divergence for $\alpha = 0$.
- α -divergence minimization is equivalent to our axioms.
- OLS is the MLE only under Gaussian errors.
- If you use OLS (or GMM, etc.) when it is not MLE, you probably violate updating consistency.

Comparison to Literature

- Csiszár (1991) characterizes f-divergences from functional properties of inference rules acting on prior models and feasible sets of probability distributions.
- Kannapan and Ng (1973, 1974) characterized KL-divergence from functional properties.
- Shore and Johnson (1980) characterize KL-divergence minimization using a primitive consisting of a prior and information in the form of a convex set of posteriors to choose from.
- Dietrich, List, and Bradley (2016) characterize Bayesian and Jeffrey updating.
- Dominiak, Kovach, and Tserenjigmid (2025) characterize α -divergence minimization for $\alpha \in [0, 1)$ using preferences over Anscombe-Aumann acts and conditional preferences given finite unions of convex sets of posteriors.
- Rommeswinkel (2019) provides Tsallis α -entropy representation using similar axioms but only for $\alpha > 0$.

Conclusion

- Foundation for a wide variety of statistical techniques.
- Decision maker needs to fix one value of α !
- α -divergences with $\alpha \neq 0, 1$ are interesting in case of preference for a “bad fit”.
- Next: convex set choice domain, reverse Bayesianism

Thank You!

Variational Inference

$x \in X$ is an observed data; $z \in Z$ is a latent variable.

$$S = X \times Z$$

DM observes x : the event $E_x \equiv \{x\} \times Z$ occurs.

$r(x|z) = \frac{r(x, z)}{r(z)}$ is intractable since $r(z) = \sum_x r(x, z)$ involves too many additions.

$$\{p_\theta \mid \theta \in \Theta\} \subset \Delta(Z) \text{ and } \mathcal{Q}_{E_x} = \{p_\theta \delta_x \mid \theta \in \Theta\}$$

$$\min_{p \in \mathcal{Q}_{E_x}} KL(p, r) \Rightarrow \min_{\theta \in \Theta} KL(p_\theta \delta_x, r)$$

E.g., mean field approximation

$$Z = \prod_i Z_i \quad \text{and} \quad \mathcal{Q}_{E_x} = \prod_i \Delta(Z_i) \times \{\delta_x\}$$

Application 3: Semi-parametric estimation

DM's reference belief is the empirical distribution $\hat{r}$.

$$\hat{r}(s) = \#\{i \mid s_i = s\} / n$$

Constraint: p satisfies the moment restriction

$$E_p[g(s, \gamma)] = 0,$$

where $g : S \times \Gamma \rightarrow \mathbb{R}^m$ for some $\gamma \in \Gamma$ (e.g., regression coefficients).

$$\mathcal{P}_\gamma = \{p \in \mathcal{P} \mid E_p[g(s, \gamma)] = 0\}$$

$$\mathcal{P}_\Gamma = \bigcup_{\gamma \in \Gamma} \mathcal{P}_\gamma \subsetneq \mathcal{P}$$

DM's estimator $\hat{\gamma}$ is a solution to

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

Newey and Smith (2004) show that the class of minimum α -divergence (i.e. Cressie-Read power divergence) estimators is a subclass of generalized empirical likelihood (GEL) estimators.

GEL is introduced to improve GMM (Hansen, 1982) for small samples.

⇒ updating consistency & strict convexity characterizes this subclass.

Special cases:

- $\alpha = 0$: Empirical Likelihood Estimator (Owen, 1988; Qin and Lawless, 1994)
- $\alpha = 1$: Exponential Tilting Estimator (Imbens et al., 1998; Kitamura and Stutzer, 1997)

When $\alpha = 0$, $\hat{\gamma}$ is the EL estimator.

$$D_0(p, \hat{r}) = KL(\hat{r}, p) = -n^{-1} \sum_{i=1}^n \log p(s_i) - H(\hat{r})$$

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

$\Downarrow$

$$\hat{\gamma} \in \arg \max_{\gamma \in \Gamma} \left(\max_{p \in \mathcal{P}_\gamma} \sum_{i=1}^n \log p(s_i) \right)$$

When $\alpha = 1$, $\hat{\gamma}$ is the ET estimator.

$$D_1(p, \hat{r}) = KL(p, \hat{r})$$

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

The inner optimization problem is the relative entropy updating:

$$\hat{r}_\gamma(s) \equiv \hat{r}(s) \exp(\theta_\gamma^\top g(s, \gamma) - \psi_\gamma(\theta_\gamma)),$$

where

$$\psi_\gamma(\theta_\gamma) = \log E_{\hat{r}}[\exp(\theta_\gamma^\top g(s, \gamma))] \text{ and } \nabla_{\theta} \psi_\gamma(\theta_\gamma) = 0$$

Moreover,

$$\hat{\gamma} \in \arg \max_{\gamma \in \Gamma} \psi_\gamma(\theta_\gamma)$$

Machine Unlearning

Emergent literature on machine unlearning (Zhang et. al., 2023; Wang et. al., 2024; Nguyen et. al., 2025)

For example, in the context of LLMs:

- State space: set of answers to a query.
- r, X^+, X^- : contains both beliefs of what DM thinks is a correct answer and what “should (not) be written”.
- LLM’s are probability distributions over answers
- Polarized divergences for $\alpha \neq 0, 1$ allow to “misfit” a model on some texts.