# A Novel Information Fusion Framework Based on a Simple Stochastic Aggregation Operator with Applications in Decision-Making

**Anonymous submission**


## Abstract

Information fusion plays a critical role in decision-making, particularly in complex scenarios. When formalizing information fusion, information is commonly represented as probability distributions, with the Bayesian framework serving as a powerful and predominant method. However, despite its theoretical strength, Bayesian methods often involve intricate computations, even when using exponential family priors, which can limit their practicality in certain decision contexts. This paper introduces a novel information fusion framework based on a simple stochastic aggregation operator. Our key insight lies in customizing appropriate measurable spaces for specific decision-making scenarios, a crucial distinction from the Bayesian approach that allows for more intuitive and computationally efficient information fusion. We demonstrate this customization in various settings, including counterfactual-based personalization and Direct Preference Optimization (DPO). The framework's core is a mathematically simple yet rigorous aggregation operator with elegant properties, including an Abelian group structure. While complementing rather than replacing Bayesian methods, our framework presents a distinctive alternative characterized by its mathematical simplicity and intuitive nature, potentially broadening the toolkit for information fusion in decision-making contexts.


## Introduction

In decision-making, representing information with probability distributions is essential. Clemen and Winkler(**?**) highlight the integration challenge of different information sources, such as scientific models and forecasting methods, especially when data is scarce. Aggregating these distributions improves our understanding of the available knowledge and uncertainties, aiding in informed decision-making. The aggregation of probabilities typically follows either mathematical or behavioral methods. Mathematical approaches combine individual probability distributions into a unified whole. These methods, particularly Bayesian ones, are noted for their systematic updating of probabilities with new data (**??**). In contrast, behavioral methods involve creating a consensus among experts through interaction (**?**). While they might be less exact than mathematical methods, they effectively harness collective expertise for better decision-making, filtering out repetitive or less relevant information.

In this paper, we propose a novel approach to information fusion that combines mathematical elegance with practical applicability. Our method introduces a simple stochastic aggregation operator that forms the foundation of a highly interpretable and intuitive framework for combining probabilistic information.

The key innovation of our approach lies in the careful selection of appropriate measurable spaces tailored to specific decision-making contexts. This crucial step allows us to represent diverse types of information as probability distributions within a unified framework, enabling straightforward aggregation through our proposed operator. Our framework offers several notable advantages:

1. **Simplicity and Interpretability**: The core aggregation operation is mathematically simple and intuitively interpretable, facilitating its adoption and understanding in various decision-making scenarios.

2. **Mathematical Elegance**: Our framework exhibits beautiful mathematical properties, including an Abelian group structure. This elegance provides a solid theoretical foundation while maintaining simplicity.

We demonstrate the utility of our framework through applications in diverse fields, including: 1) Personalized incentive optimization; 2) Aggregation of multiple expert predictions; 3) DPO in alignment of large language models. These case studies illustrate how our approach can provide new insights and potentially improve decision-making processes across various domains. While our framework does not aim to replace established Bayesian methods, it offers a complementary perspective that may be particularly valuable in scenarios where computational simplicity and intuitive interpretability are paramount.

The remainder of this paper is organized as follows: Section 2 provides preliminaries on stochastic aggregation. Section 3 introduces our core stochastic aggregation operation. Section 4 explores the Abelian group structure of our framework. Section 5 discusses adaptations and extensions of our framework. Section 6 demonstrates an application in Direct Preference Optimization. Finally, Section 7 concludes the paper with a discussion of implications and future research directions.

## Preliminaries on Stochastic Aggregation

Traditional aggregation methods in statistics, such as mean, median, maximum, and minimum, typically involve simple mathematical operations applied either to different instances of a single variable or across multiple variables. These methods, while effective for summarizing central tendencies or range of data, do not directly address the aggregation of entire probability distributions.

The Bayesian paradigm, directly operates on distributions, presents a powerful framework for the fusion of information from various sources (See e.g. (**??**)). It is fundamentally based on Bayes' theorem, which offers a mechanism to update a probability distribution by combining prior knowledge with new evidence. Given a set of information $e_1, e_2, \ldots, e_n$ regarding an event or quantity of interest $U$, the updated probability distribution $p^*$ can be calculated using the following formula:

$$p^* \triangleq p(u|e_1, \ldots, e_n) \propto p(u)L(e_1, \ldots, e_n|u), \qquad (1)$$

where $L$ represents the likelihood function associated with the observed information, and the symbol $\propto$ denotes proportionality. This principle can be applied to aggregate any type of information represented by probability distributions.

However, the computation of $p^*$ is often very challenging, particularly in complex scenarios with intricate dependency structures. The most common approach to simplify this calculation is the use of exponential family prior distributions. However, this method has several drawbacks, including limited flexibility, interpretability issues, and lack of robustness. An alternative and flexible approach involves utilizing statistical tools such as Copulas. Copulas are used to describe/-model the dependence (inter-correlation) between random variables, which have been used widely in quantitative finance. Sklar's theorem states that any multivariate joint distribution can be written in terms of univariate marginal distribution functions and a copula which describes the dependence structure between the variables.

**Theorem 1** (Sklar's Theorem (**?**))**.** *Let $H$ be a joint distribution function of random variables with marginal distribution functions $F_1, F_2, \ldots, F_n$. Then there exists a copula $C$ such that for all $u_1, u_2, \ldots, u_n$,*

$$H(u_1, u_2, \ldots, u_n) = C(F_1(u_1), \ldots, F_n(u_n)).$$

*If $H$ has a density $h$, and the marginals $F_i$ have densities $f_i$, then*

$$h(u_1, \ldots, u_n) = c(F_1(u_1), \ldots, F_n(u_n)) \cdot f_1(u_1) \cdot \ldots \cdot f_d(u_n),$$

*where $c$ is the density of $C$. If all marginals $F_i$ are continuous, then $C$ is unique.*

It is proposed that by using copula to describe dependence structure, the posterior probability in Eq. (1) can be simplified to:

$$p^* \propto c[1 - F_1(u), \ldots, 1 - F_n(u)] \prod_{i=1}^{n} f_i(u), \qquad (2)$$

where

$$f_i(u) = p(u|e_i) \qquad (3)$$

represents the posterior probability given the information $e_i$ (**?**), which implicitly assumes that there exists a posterior $f_i$ given the prior $p(u)$ represents the information $e_i$ for any $i$. Hence, we refer this as to the "prior-dependent" distribution representation for information. The underlying logic of this method provides a mathematical framework for fusing prior information and multiple pieces of evidence. In this framework, each piece of evidence generates a distribution, and the fused information is constructed as a distribution that incorporates all these individual distributions. Consequently, it can be argued that the copula-based Bayesian method is mathematically founded on the principle of *Stochastic Aggregation for Distributions*.

However, the key formula (2) involves concepts and computations of considerable complexity. In the following sections, we introduce an intuitive information fusion framework that offers significant mathematical simplicity.

## The Stochastic Aggregation Operation

When utilizing probability distributions to represent information, the fusion of information is formalized through the stochastic aggregation of probability distributions. This naturally leads to a fundamental question: "What is the simplest mathematical framework for aggregating two distributions?" In response, we propose the following mathematical operator:

**Definition 2** (Stochastic Product Operator for Probability Measures)**.** Let $(\Omega, \mathcal{F})$ be a measurable space, and let $P_1$ and $P_2$ be probability measures on $(\Omega, \mathcal{F})$. The Stochastic Product of $P_1$ and $P_2$, denoted as $P_1 \odot P_2$, is defined as a new probability measure $P$ on $(\Omega, \mathcal{F})$ such that:

$$P(A) \propto P_1(A)P_2(A) \qquad (4)$$

, for any atom event $A$. where the symbol $\propto$ denotes proportionality, implying that the right-hand side should be normalized to ensure $P$ is a valid probability measure.

This stochastic product operator shares some similarities with the Product of Experts model introduced by **?**, although our formulation and application differ in several key aspects. While Hinton's work focuses on machine learning model construction, our operator is formulated in a more general measure-theoretic framework for probability fusion.

**Example 3** (Stochastic Aggregation for Probability Distributions)**.** *Consider a set of $K$ independent standard normal distributions, each denoted by $\mathcal{N}(0, 1)$. Applying the Stochastic Product Operator, we derive an aggregated distribution $\mathcal{N}^*$ as follows:*

$$\mathcal{N}^* = \underbrace{\mathcal{N}(0,1) \odot \cdots \odot \mathcal{N}(0,1)}_{K \text{ times}} \qquad (5)$$

*where $\odot$ denotes the Stochastic Product Operator.*

*It can be demonstrated that the resulting distribution $\mathcal{N}^*$ is a normal distribution with mean 0 and variance $\frac{1}{K}$, expressed as $\mathcal{N}(0, \frac{1}{K})$. This result yields two significant interpretations:*

1. ***Information Fusion:*** *The combination of $K$ independent opinions, each represented by $\mathcal{N}(0,1)$, results in a collective opinion with a reduced variance of $\frac{1}{K}$. This variance reduction can be interpreted as an increase in the precision of the aggregated information.*
2. ***Sampling Theory Parallel:*** *This result aligns with a fundamental concept in sampling theory. If one were to draw $K$ independent samples from $\mathcal{N}(0,1)$ and compute their arithmetic mean, the distribution of this sample mean would follow $\mathcal{N}(0, \frac{1}{K})$. This parallel highlights the connection between our aggregation method and established statistical principles.*

The concept of stochastic aggregation naturally extends from probability distributions to random variables. This extension is formalized in the following definition:

**Definition 4** (Stochastic Aggregation for Random Variables)**.** Let $X_1$ and $X_2$ be two random variables defined on the same probability space $(\Omega, \mathcal{F}, P)$. The Stochastic Aggregation of $X_1$ and $X_2$, denoted as $X_1 \odot X_2$, is defined as a new random variable $X$ whose probability function $p_X$ is given by the Stochastic Product of the probability functions of $X_1$ and $X_2$. Specifically, we have:

$$p_X(x) \propto p_1(x)p_2(x) \tag{6}$$

where $p_1$ and $p_2$ are the probability functions of $X_1$ and $X_2$, respectively. Here, $p(x)$ represents the probability density function for continuous random variables or the probability mass function for discrete random variables.

The stochastic aggregation operation defined in Eq. (6) exhibits two key differences from Eq. (2). First, it does not involve prior information. Second, the copula-related density is constant, implying a form of conceptual independence among the pieces of evidence. Consequently, our information fusion method can be considered the most intuitive and mathematically simplest approach available.

This approach offers two significant advantages: excellent mathematical properties and strong interpretability. The mathematical elegance of this method is evident in its simplicity and the well-defined properties of the stochastic product operator. Furthermore, the straightforward nature of the aggregation process enhances its interpretability, allowing for clear and intuitive understanding of how different pieces of information are combined, which is particularly valuable in decision-making contexts.

## The Abelian Group Structure

A fundamental question arises regarding our proposed simple information fusion framework: In which practical decision-making scenarios can this framework be effectively applied, and through what methods can its advantages be fully leveraged? To address this question, we turn our attention to a critical concern in contemporary internet platforms: the optimization of user retention within budget constraints through various incentive strategies. Personalized incentives based on causal relationships have emerged as a prevalent approach to address this challenge (**?**). A simplified illustrative example would be:

**Example 5** (The Personalized Decision-making Setting)**.** *Consider a causal model tailored for personalized incentives, encompassing observable variables: $S$, $\mathbf{X}$, $T$, and $Y$, as depicted in the causal diagram (Figure 1). The causal mechanisms for each participant $u$ are described below:*

1. *$S$: Assigns users to one of three experiment groups. The random group ($S = 0$) receives incentives based purely on chance; the pure strategy group ($S = 1$) has incentives tailored according to specific user characteristics; and the mixed strategy group ($S = 2$) combines random allocation with user-specific strategies.*
2. *$\mathbf{X}$: Denotes the pre-treatment features of the user that influence both the treatment and outcome. This includes demographic details, historical engagement levels, and other relevant factors.*
3. *$T$: A binary incentive treatment variable. For users in the random group ($S = 0$), this decision is made with uniform probability; For users in the pure strategy group ($S = 1$), the incentive is a deterministic function of the pre-treatment features; In the mixed strategy group ($S = 2$), the decision is influenced by the user's features but retains some randomness.*
4. *$Y$: The outcome variable of the user's reaction to the incentive, e.g. conversion, purchase or retention.*

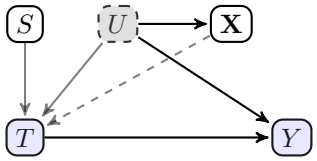


Figure 1: Causal Model for Personalized Incentives: This diagram illustrates the causal relationships among group assignment $S$, incentive treatment $T$, pre-treatment features $\mathbf{X}$, and the outcome variable $Y$. The model integrates a unit representation $U$, capturing all relevant endogenous information (excluding $T$) that determines the Layer valuations regarding to $(T, Y)$.

In this scenario, we conceptualize the entire user base of the platform as a measurable space $\mathcal{U}$. Any random variable that selects an individual from this space corresponds to a probability distribution. We represent information using these probability distributions. For instance, consider the evidence or observation $Y = y$. Regardless of the causal modeling framework employed (Potential Outcomes(**?**), Structural Causal Models (**?**), or Distribution-consistency Structural Causal Models (**?**)), the process of evidence abduction yields a distribution over $\mathcal{U}$, which we denote as $U(y)$.

This approach naturally extends to multiple pieces of evidence, addressing scenarios that pose challenges for mainstream causal frameworks. Consider a subpopulation exhibiting homogeneity across individuals, where we have observations $y_1, ..., y_n$.The homogeneity assumption allows us to treat these samples as independent and identically distributed (i.i.d.). The fusion of information from these observations can be conceptualized as the aggregation of information obtained from each individual piece of evidence.

Mathematically, this can be expressed as:

$$U(y_1, ..., y_n) = U(y_1) \odot \cdots \odot U(y_n)$$

where $\odot$ represents our stochastic product operator for information fusion. This formulation elegantly captures the process of combining multiple observations within our proposed framework. It demonstrates how our simple information fusion method can be applied to aggregate evidence in a principled manner.

This example illustrates the fundamental logic of our information fusion framework. The process involves designing an appropriate measurable space for a specific scenario, representing information as probability distributions within this space, and then employing the Stochastic Product Operator to fuse these distributions. We will now formalize the elegant properties of this information fusion framework in the context of our decision-making setting.

**Definition 6** (Information Fusion)**.** Let $S_1$ and $S_2$ be two unit selection variables of the population $\mathcal{U}$. The Information fusion of $S_1$ and $S_2$, denoted as $S_1 \odot S_2$, is defined as a unit selection variable $S$ whose distribution $p(u)$ satisfies:

$$p(u) \propto p_1(u)p_2(u) \tag{7}$$

where $p_1(u)$ and $p_2(u)$ are the respective distributions of $S_1$ and $S_2$. [1]

The Information Fusion operation, denoted as $\odot$, inherits several desirable properties from the Stochastic Product Operator. These properties collectively form an algebraic structure known as an Abelian Group. We formally define these properties below:

**Property 7** (Commutativity)**.** *For all unit selection variables $S_1$ and $S_2$ in the population $\mathcal{U}$, the operation $\odot$ is commutative if:*

$$S_1 \odot S_2 = S_2 \odot S_1 \tag{8}$$

**Property 8** (Associativity)**.** *For all unit selection variables $S_1$, $S_2$, and $S_3$ in the population $\mathcal{U}$, the operation $\odot$ is associative if:*

$$(S_1 \odot S_2) \odot S_3 = S_1 \odot (S_2 \odot S_3) \tag{9}$$

**Property 9** (Identity Element)**.** *There exists an element $U$, such that for any unit selection variable $S$ in the population $\mathcal{U}$:*

$$S \odot U = S \tag{10}$$

**Property 10** (Inverse Element)**.** *For every unit selection variable $S$ with non-zero probability on every unit $u \in \mathcal{U}$[2], there exists a unique inverse $S^*$, such that:*

$$S \odot S^* = U \tag{11}$$

**Theorem 11.** *The Information Fusion operation $\odot$ forms an Abelian Group on the set of unit selection variables of the population $\mathcal{U}$.*

[1]These distributions represent posteriors obtained by updating a uniform prior, such as distributions resulting from evidence abduction.

[2]For unit selection variables with zero probability on some units, special techniques are required.

*Proof.* Commutativity and associativity are directly evident from the operation's definition in Eq. (7), where $S_1 \odot S_2 = S_2 \odot S_1$ and $(S_1 \odot S_2) \odot S_3 = S_1 \odot (S_2 \odot S_3)$, due to the commutative and associative nature of multiplication in probability densities.

The uniform distribution over $\mathcal{U}$ acts as the identity element, as $S \odot U = S$ for any unit selection variable $S$, maintaining the original probability density unchanged.

Lastly, the inverse element for each unit selection variable $S$, denoted as $S^*$, can be constructed using inverse probabilities as weights. For a unit selection variable $S$ with probability density $p(u)$, its inverse $S^*$ has a probability density proportional to $\frac{1}{p(u)}$. This ensures that $S \odot S^*$ aligns with the uniform distribution, thus fulfilling the requirement for an inverse element.

Therefore, the Information Fusion operation $\odot$ satisfies all the properties of an Abelian Group. □

This group structure endows the operation with several significant advantages:

1. **Flexible Information Combination**: The commutative and associative properties allow for the fusion of information from multiple sources in any order, providing flexibility in the aggregation process.
2. **Consistency in Aggregation**: The identity element ensures that combining a piece of information with no information (represented by the uniform distribution) does not alter the original information.
3. **Information Reversal**: The existence of inverse elements allows for the theoretical "removal" of a piece of information from an aggregated whole, which can be crucial in certain analytical scenarios.
4. **Mathematical Rigor**: The group structure provides a solid mathematical foundation for operations on information, allowing for the application of well-established algebraic techniques in information processing.

These properties collectively establish a robust framework for information fusion in our decision-making context, particularly in scenarios involving counterfactual-based personalized incentives as described in Example 5.

## Adaptation of the Information Fusion Framework

Our proposed information fusion framework, initially introduced for specific decision-making scenarios, shows promising potential for adaptation to a wider range of information types and probability distributions. While acknowledging potential limitations, we explore how this framework can be extended to accommodate diverse forms of information, thereby possibly enhancing its versatility and applicability. One significant extension is the incorporation of set-based information. This adaptation allows us to integrate subset information into our fusion framework, potentially broadening its scope. To formalize this concept, we introduce the notion of Set Information Fusion:

**Definition 12** (Set Information Fusion)**.** Let $\mathcal{U}$ be the population space, and consider a set $A \subseteq \mathcal{U}$. We denote the uniform random variable supported on $A$ as $U_A$, defined by:

$$P(U_A = u) = \frac{1}{|A|}\mathbf{1}_A(u) \tag{12}$$

where $\mathbf{1}_A(u)$ is the indicator function of set $A$. This representation allows us to define the aggregation of a random variable $X$ with a set $A$ as:

$$X \odot A \triangleq X \odot U_A, \tag{13}$$

where $\odot$ denotes our stochastic product operator for information fusion.

This definition provides a mechanism to incorporate set-based information into our fusion framework. By treating a set $A$ as a uniform distribution over its elements, we can seamlessly integrate it with other forms of information represented by random variables.

Furthermore, our information fusion approach may not be limited to the measurable space $\mathcal{U}$. We suggest that it could be extended to more general measurable spaces, which could significantly expand the framework's applicability. In the following sections, we will explore this generalization, with a particular focus on exponential family probability distributions.

**Definition 13** (Exponential Family Distribution)**.** A probability distribution belongs to the exponential family if its density function can be expressed as:

$$f(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta})\right) \tag{14}$$

where $\mathbf{x}$ is the random vector, $\boldsymbol{\theta}$ is the parameter vector, $h(\mathbf{x})$ is a function of $\mathbf{x}$, $\boldsymbol{\eta}(\boldsymbol{\theta})$ is the natural parameter, $\mathbf{T}(\mathbf{x})$ is the sufficient statistic, and $A(\boldsymbol{\theta})$ is the log-partition function.

Information can be represented using exponential family distributions. This naturally leads to the question of how to fuse two such information sources according to our definition. We propose the following:

**Definition 14** (Exponential Family Information Fusion)**.** Consider two information sources, represented by exponential family distributions $f_1(\mathbf{x}|\boldsymbol{\theta}_1)$ and $f_2(\mathbf{x}|\boldsymbol{\theta}_2)$ respectively:

$$f_1(\mathbf{x}|\boldsymbol{\theta}_1) = h_1(\mathbf{x}) \exp\left(\boldsymbol{\eta}_1(\boldsymbol{\theta}_1)^T \mathbf{T}_1(\mathbf{x}) - A_1(\boldsymbol{\theta}_1)\right)$$
$$f_2(\mathbf{x}|\boldsymbol{\theta}_2) = h_2(\mathbf{x}) \exp\left(\boldsymbol{\eta}_2(\boldsymbol{\theta}_2)^T \mathbf{T}_2(\mathbf{x}) - A_2(\boldsymbol{\theta}_2)\right)$$

Their information fusion can be defined as:

$$(f_1 \odot f_2)(\mathbf{x}) \propto h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta})\right) \tag{15}$$

where $\odot$ denotes our stochastic product operator, and:

$$h(\mathbf{x}) = h_1(\mathbf{x})h_2(\mathbf{x})$$
$$\mathbf{T}(\mathbf{x}) = (\mathbf{T}_1(\mathbf{x})\ \mathbf{T}_2(\mathbf{x}))$$
$$\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1(\boldsymbol{\theta}_1) \\ \boldsymbol{\eta}_2(\boldsymbol{\theta}_2) \end{pmatrix}$$
$$A(\boldsymbol{\theta}) = A_1(\boldsymbol{\theta}_1) + A_2(\boldsymbol{\theta}_2)$$

The proposed definition of Exponential Family Information Fusion reveals several intriguing properties of our framework. Primarily, it addresses the fusion of disparate distributions by concatenating their respective sufficient statistics and natural parameters into higher-dimensional vectors. A notable characteristic is that the resultant fused distribution maintains the exponential family structure, demonstrating closure under this fusion operation. , let us consider an example using multiple independent Bernoulli trials, which belong to the exponential family.

**Example 15** (Bernoulli Trials Fusion)**.** *Consider a sequence of $n$ independent Bernoulli trials. We have two different information sources about these trials, each represented by a product of Bernoulli distributions with parameters $p_1$ and $p_2$ respectively. Both distributions are defined on the same measure space $\{0,1\}^n$. The probability mass functions for each information source are given by:*

$$f_1(\mathbf{x}|p_1) = p_1^{\sum_{i=1}^n x_i}(1-p_1)^{n-\sum_{i=1}^n x_i}$$
$$f_2(\mathbf{x}|p_2) = p_2^{\sum_{i=1}^n x_i}(1-p_2)^{n-\sum_{i=1}^n x_i}$$

*where $\mathbf{x} = (x_1, ..., x_n)$ and $x_i \in \{0,1\}$ for $i = 1, ..., n$. These can be rewritten in the exponential family form as:*

$$f_1(\mathbf{x}|p_1) = \exp\left(\sum_{i=1}^n x_i \log\left(\frac{p_1}{1-p_1}\right) + n\log(1-p_1)\right)$$
$$f_2(\mathbf{x}|p_2) = \exp\left(\sum_{i=1}^n x_i \log\left(\frac{p_2}{1-p_2}\right) + n\log(1-p_2)\right)$$

*where $\eta_j = \log\left(\frac{p_j}{1-p_j}\right)$ is the natural parameter, $T(\mathbf{x}) = \sum_{i=1}^n x_i$ is the sufficient statistic, and $A_j(\eta_j) = -n\log(1-p_j) = n\log(1+e^{\eta_j})$ is the log-partition function for $j = 1, 2$.*

*Now, let's fuse these two distributions using our Exponential Family Information Fusion:*

$$\begin{aligned}&(f_1 \odot f_2)(\mathbf{x}) \\ &\propto \exp\left(T(\mathbf{x})(\eta_1+\eta_2) - (A_1(\eta_1) + A_2(\eta_2))\right) \\ &= \left(\frac{p_1 p_2}{(1-p_1)(1-p_2)}\right)^{\sum_{i=1}^n x_i} (1-p_1)^n(1-p_2)^n\end{aligned}$$

*This result demonstrates that our fusion operation preserves the exponential family structure. However, it's important to note that the fused distribution is not a product of independent Bernoulli trials with a single parameter. Instead, it's a more complex distribution over the same measure space $\{0,1\}^n$, where the probability of observing $k$ successes in $n$ trials follows a different structure than a simple Bernoulli product.*

Moreover, we can prove that exponential family information fusion preserves the **Abelian group structure** previously observed in our framework on the $\mathcal{U}$ space. It exhibits

commutativity and associativity, possesses an identity element (representable as a uniform distribution), and each distribution has an inverse (constructible by negating the natural parameters). This consistency in mathematical structure, coupled with the potential to handle a broader spectrum of distribution types, enhances the applicability of our framework.

The fusion of multiple information sources or expert opinions presents a critical challenge in decision-making and forecasting. While Hinton's work (**?**) established a foundation with the Product of Experts model, our proposed method introduces a key innovation: the explicit incorporation and subsequent removal of a common information source representing fundamental trends or shared background knowledge.

**Example 16** (Analyst Forecast Aggregation)**.** *Consider a financial market scenario where multiple analysts provide forecasts for a particular outcome. The stochastic product operator $\odot$ synthesizes these predictions as follows:*

1. *Let $F_i$ represent each analyst's forecast as a probability distribution.*
2. *Define $S$ as the distribution representing the common information source.*
3. *Apply the stochastic product operator to combine these forecasts:*

$$F_{agg} = S \odot \bar{F}_1 \odot \cdots \odot \bar{F}_n \tag{16}$$

*where $S^*$ is the inverse of $S$, $\bar{F}_i = F_i \odot S^*$, and $n$ is the number of analysts.*

*The $\bar{F}_i$ terms provide direct insight into how each expert's view deviates from the consensus. Removal of common background allows for clearer identification of each analyst's unique perspectives.*

The resulting $F_{agg}$ represents a refined consensus forecast that aggregates collective expertise while adjusting for underlying market trends. This example demonstrates the intuitive aspects of our framework in specific decision-making scenarios where synthesizing diverse expert opinions while accounting for common information is crucial.

## Application in Direct Preference Optimization

The generality of information fusion as a concept suggests its potential for wide-ranging applications. We have identified a particularly intriguing and relevant application in the context of Direct Preference Optimization (DPO). DPO is a cutting-edge post-training technique employed in the alignment of large language models, including the LLaMA series (**?**). The core mechanism of DPO involves optimizing a policy $\pi$ to maximize the expected reward while constraining the divergence from a reference policy $\pi_{ref}$. This can be expressed mathematically as:

$$\max_{\pi} \mathbb{E}_{x\sim d, y\sim\pi(\cdot|x)}[r(x,y)] - \beta D_{KL}(\pi(\cdot|x)||\pi_{ref}(\cdot|x)) \tag{17}$$

where $x$ represents the input state, $y$ the output action, $r(x, y)$ the reward function, and $\beta$ a temperature parameter controlling the trade-off between reward maximization and policy divergence. The solution to this optimization problem takes the form:

$$\pi'(y|x) = \frac{\pi_{ref}(y|x)\exp(\frac{1}{\beta}r(x,y))}{Z(x)} \tag{18}$$

where $Z(x)$ is a normalization factor. We propose to reinterpret this solution using our stochastic aggregation operation:

$$\pi(\cdot|x) = \pi_{ref}(\cdot|x) \odot p_r(\cdot|x;\beta) \tag{19}$$

Here, $p_r(\cdot|x;\beta)$ represents the Boltzmann exploration probability distribution generated by applying the softmax function to the reward $r(\cdot|x)$ with temperature parameter $\beta$:

$$p_r(\cdot|x;\beta) \propto \exp(\frac{1}{\beta}r(x,\cdot)) \tag{20}$$

This reinterpretation offers several valuable insights:

- It clearly delineates the roles of the reference policy $\pi_{ref}(\cdot|x)$ as prior information and $p_r(\cdot|x;\beta)$ as new preference-based information.
- The parameter $\beta$ controls the balance between exploitation of reward information and exploration based on the reference policy:
  - As $\beta \to 0$, the policy tends towards pure exploitation, prioritizing actions that maximize the reward.
  - As $\beta \to \infty$, the policy approaches the reference policy, emphasizing exploration.
- In the case where $r(x, y)$ is constant for all $y$ given $x$, $p_r(\cdot|x;\beta)$ becomes a uniform distribution. Consequently, $\pi(\cdot|x) = \pi_{ref}(\cdot|x)$, indicating that in the absence of preference information, the optimal policy reverts to the reference policy.

This novel perspective on DPO provides a more intuitive understanding of the algorithm's behavior and its relationship to concepts in information fusion. By framing DPO within our information fusion framework, we gain new insights into the interplay between prior knowledge (represented by the reference policy) and learned preferences (captured by the reward-based distribution). The application of our information fusion framework to DPO not only demonstrates the framework's versatility but also provides a fresh perspective on a state-of-the-art machine learning technique. This connection between information fusion and preference optimization may inspire more efficient and interpretable preference-based learning algorithms.

## Conclusion and Discussions

While representing information as probability distributions is a common practice in the field of information fusion, our work introduces a critical insight: the necessity of representing information as probability distributions within an appropriately chosen measurable space, tailored to the specific decision-making context.

The challenge of aggregating probability assessments from diverse information sources is a longstanding problem

in statistics and stochastic modeling. The Bayesian framework, which adopts the probabilistic representation of information, is a powerful approach to this challenge. However, its computational complexity and conceptual intricacy often limit its practical applicability. Our proposed framework offers a novel perspective on this problem. By carefully selecting the measurable space based on the decision-making setting, we enable the definition of information fusion through a mathematically simple and intuitive operation. This simplicity confers significant advantages to our new framework in terms of computational efficiency and interpretability.

The key contributions of our work are multifaceted. At its core, we introduce a mathematically simple yet rigorous stochastic product operator for aggregating probability distributions. This operator forms the foundation of our framework, which we demonstrate possesses an Abelian group structure. This structure provides a solid mathematical foundation for information fusion operations, lending our approach both theoretical elegance and practical utility. Beyond its theoretical merits, we showcase the framework's applicability to real-world scenarios. Specifically, we demonstrate its effectiveness in personalized incentive optimization and DPO in machine learning, illustrating how our approach can address complex decision-making challenges across diverse domains. Through these contributions, our work not only advances the theoretical understanding of information fusion but also provides a practical tool for tackling real-world decision-making problems.

While our approach does not aim to replace the Bayesian framework, it offers a complementary perspective that may be particularly useful in scenarios where computational simplicity and intuitive interpretability in the fusion stage are paramount. The framework's adaptability to various measurable spaces and its extension to exponential family distributions further enhance its potential applicability across diverse domains.

**Discussions and Future Research.** The core idea of our paper can be succinctly summarized: the representation of information fundamentally determines the process of information fusion. Information fusion and information representation are deeply intertwined problems. A key distinction of our approach lies in the redistribution of complexity within the information fusion process. While the Bayesian framework involves a complex fusion process without extensive consideration of information representation, our framework takes the opposite approach. The challenge in our method lies in identifying an appropriate measurable space and representing various types of information as probability distributions within this space. Once this representation is achieved, the actual fusion of information is executed through a straightforward and intuitive operator. Specifically, our framework emphasizes the effort required to identify an appropriately chosen measurable space and represent diverse types of information as probability distributions within this space. This process, while simplifying the fusion operation, introduces a rich set of challenges in the representation phase.

The representation of information encompasses numerous dimensions, particularly evident in causal inference scenarios. As outlined by **?**, information in causal reasoning can be stratified into three layers: association, intervention, and counterfactual. Further distinctions arise between population-level and individual-level information **?**, evidence and observations, events and sets, and structural equations versus noise distribution information. The complexity of these representations is detailed discussion provided in Appendix. This multifaceted nature of information representation raises a critical question for future research: How can we effectively fuse such diverse types of information within a unified framework? While we do not claim to have an answer to this question, our work suggests a promising approach: customizing information representation based on the specific decision-making setting at hand.

## Formalization of Causal Information

Information is a broad concept that can be formalized in various ways from a causal perspective. The DiscoSCM (**?**) is an extended causal modeling framework of both potential outcomes (PO) (**??**) and structural causal models (SCMs) (**?**). The PO approach begins with a population of units. There is a treatment/cause $T$ that can take on different values for each unit. Corresponding to each treatment value, a unit is associated with a set of potential outcomes, represented as $Y(t)$. Only one of these potential outcomes, corresponding to the treatment received, can be observed. The causal effect is related to the comparison between potential outcomes, of which at most one corresponding realization is available, with all the others missing. (**?**) refers to this missing data nature as the "fundamental problem of causal inference". In constrast, the SCM framework starts with structural equations that represents the underlying causal mechanisms of observed phenomena.

**Definition 17** (**Structural Causal Models** (**?**))**.** A structural causal model is a tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F} \rangle$, where

- $\mathbf{U}$ is a set of background variables, also called exogenous variables, that are determined by factors outside the model, and $P(\cdot)$ is a probability function defined over the domain of $\mathbf{U}$;
- $\mathbf{V}$ is a set $\{V_1, V_2, \ldots, V_n\}$ of (endogenous) variables of interest that are determined by other variables in the model – that is, in $\mathbf{U} \cup \mathbf{V}$;
- $\mathcal{F}$ is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each $f_i$ is a mapping from (the respective domains of) $U_i \cup Pa_i$ to $V_i$, where $U_i \subseteq U$, $Pa_i \subseteq \mathbf{V} \setminus V_i$, and the entire set $\mathcal{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. That is, for $i = 1, \ldots, n$, each $f_i \in \mathcal{F}$ is such that

$$v_i \leftarrow f_i(pa_i, u_i),$$

  i.e., it assigns a value to $V_i$ that depends on (the values of) a select set of variables in $U \cup V$.

Potential outcomes are derivatives of the $do$-operator.

**Definition 18** (**Submodel-"Interventional SCM"** (**?**))**.** Consider an SCM $\langle \mathbf{U}, \mathbf{V}, \mathcal{F} \rangle$, with a set of variables $X$ in $V$, and a particular realization $x$ of $X$. The $do(x)$ operator, representing an intervention (or action), modifies the set of structural equations $\mathcal{F}$ to $\mathcal{F}_x := \{f_{V_i} : V_i \in V \setminus X\} \cup \{f_X \leftarrow x : X \in X\}$ while maintaining all other elements constant. Consequently, the induced tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}_x \rangle$ is called as *Interential SCM* , and potential outcome $Y(x)$ (or denoted as $Y_x(u)$) is defined as the set of variables $Y \subseteq V$ in this submodel.

These two frameworks are considered equivalent, as most statements in these causal frameworks are generally translatable. One of the most important statements is the consistency rule, which is an assumption in the PO framework and a theorem in the SCM framework.

**Assumption 19** (**Consistency** (**??**))**.** *The potential outcome $Y(t)$ precisely matches the observed variable $Y$ given observed treatment $T = t$, i.e.,*

$$T = t \Rightarrow Y(t) = Y. \tag{21}$$

However, this consistency rule may lead to capacity limitations for counterfactual inference. Consider a hypothetical scenario:"If an individual with average ability scores exceptionally high on a test due to good fortune, what score would the individual achieve had he retaken the test under the identical conditions? An exceptionally high score or an average one?" Intuitively, predicting an average score seems more practical since luck is typically non-replicable. To accommodate this "uncontrollable good fortune", the distribution-consistency assumption is proposed.

**Assumption 20** (**Distribution-consistency**)**.** *For any individual represented by $U = u$ with an observed treatment $X = x$, the counterfactual outcome $Y(x)$ is equivalent in distribution to the observed outcome $Y$. Formally,*

$$X = x, U = u \Rightarrow Y(x) \overset{d}{=} Y \tag{22}$$

*where $\overset{d}{=}$ denotes equivalence in distribution.*

To explicitly incorporate individual semantics $U$, the Distribution-consistency Structural Causal Model (DiscoSCM) framework is proposed as follows.

**Definition 21** (**Distribution-consistency Structural Causal Model (DiscoSCM)**)**.** A DiscoSCM is a tuple $\langle U, \mathbf{E}, \mathbf{V}, \mathcal{F} \rangle$, where

- $U$ is a unit selection variable, where each instantiation $U = u$ denotes an individual. It is associated with a probability function $P(u)$, uniformly distributed by default.
- $\mathbf{E}$ is a set of exogenous variables, also called noise variables, determined by factors outside the model. It is independent to $U$ and associated with a probability function $P(e)$;
- $\mathbf{V}$ is a set of endogenous variables of interest $\{V_1, V_2, \ldots, V_n\}$, determined by other variables in $\mathbf{E} \cup \mathbf{V}$;

- $\mathcal{F}$ is a set of functions $\{f_1(\cdot,\cdot;u), f_2(\cdot,\cdot;u), \ldots, f_n(\cdot,\cdot;u)\}$, where each $f_i$ is a mapping from $E_i \cup Pa_i$ to $V_i$, with $E_i \subseteq \mathbf{E}$, $Pa_i \subseteq \mathbf{V} \setminus V_i$, for individual $U = u$. Each function assigns a value to $V_i$ based on a select set of variables in $\mathbf{E} \cup \mathbf{V}$. That is, for $i = 1, \ldots, n$, each $f_i(\cdot,\cdot;u) \in \mathcal{F}$ is such that

$$v_i \leftarrow f_i(pa_i, e_i; u),$$

i.e., it assigns a value to $V_i$ that depends on (the values of) a select set of variables in $E \cup V$ for each individual $U = u$.

**Definition 22.** For a DiscoSCM $\langle U, \mathbf{E}, \mathbf{V}, \mathcal{F}\rangle$, $X$ is a set of variables in $V$ and $x$ represents a realization, the $do(\mathbf{x})$ operator modifies: 1) the set of structural equations $\mathcal{F}$ to

$$\mathcal{F}_{\mathbf{x}} := \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} \leftarrow x\},$$

and; 2) noise $\mathbf{E}$ to couterfactual noise $\mathbf{E}(\mathbf{x})$ maintaining the same distribution. [3] The induced submodel $\langle U, \mathbf{E}(x), \mathbf{V}, \mathcal{F}_{\mathbf{x}}\rangle$ is called the *interventional DiscoSCM*.

**Definition 23 (Counterfactual Outcome).** For a DiscoSCM $\langle U, \mathbf{E}, \mathbf{V}, \mathcal{F}\rangle$, $X$ is a set of variables in $V$ and $x$ represents a realization. The counterfactual outcome $Y^d(x)$ (or denoted as $Y(x)$, $Y_x(e_x)$ when no ambiguity concerns) is defined as the set of variables $Y \subseteq V$ in the submodel $\langle U, \mathbf{E}(x), \mathbf{V}, \mathcal{F}_{\mathbf{x}}\rangle$. In the special case that $X$ is an empty set, the corresponding submodel is denoted as $\langle U, \mathbf{E}^d, \mathbf{V}, \mathcal{F}\rangle$ and its counterfactual noise and outcome as $E^d$ and $Y^d$, respectively.

This framework introduces a novel lens – individual/population – to address causal questions, when climbing the Causal Hierarchy: associational, interventional, and counterfactual layers. Specifically, consider a DiscoSCM where $e$ represents the observed trace or evidence (e.g., $X = x, Y = y$), the following conclusions can be drawn.

**Definition 24 (Layer Valuation with DiscoSCM).** A DiscoSCM $\langle U, \mathbf{E}, \mathbf{V}, \mathcal{F}\rangle$ induces a family of joint distributions over counterfactual outcomes $Y(x), \ldots, Z(w)$, for any $Y, Z, \ldots, X, W \subseteq V$:

$$P(y_x, \ldots, z_w; u) = \sum_{\{e_x \ldots, e_w \mid Y^d(x)=y, \quad \ldots, \ Z^d(w)=z, U=u\}} P(e_x, ..., e_w). \tag{23}$$

is referred to as Layer 3 valuation. In the specific case involving only one intervention [4], e.g., $do(x)$:

$$P(y_x; u) = \sum_{\{e_x \mid Y^d(x)=y, U=u\}} P(e_x), \tag{24}$$

is referred to as Layer 2 valuation. The case when no intervention:

$$P(y; u) = \sum_{\{e \mid Y=y, U=u\}} P(e), \tag{25}$$

is referred to as Layer 1 valuation. Here, $y$ and $z$ represent the observed outcomes, $x$ and $w$ the observed treatments, $u$ the noise instantiation, and we denote $y_x$ and $z_w$ as the realization of their corresponding potential outcomes, $e_x$, $e_w$ as the instantiation of their corresponding counterfactual noises.

**Theorem 25 (Individual-Level Valuations).** *For any given individual $u$,*

$$P(y_x|e; u) = P(y_x; u) = P(y|x; u)$$

*indicating that the (individual-level) probabilities of an outcome at Layer 1/2/3 are equal.*

Individual-level valuations (e.g. $P(y_x|e; u)$) are primitives while population-level valuations (e.g. $P(y_x|e)$) are derivations.

**Theorem 26 (Population-Level Valuations).** *Consider a DiscoSCM wherein $Y(x)$ is the counterfactual outcome, and $e$ represents the observed trace or evidence. The Layer 3 valuation $P(Y(x)|e)$ is computed through the following process:*

***Step 1 (Abduction):** Derive the posterior distribution $P(u|e)$ of the unit selection variable $U$ based on the evidence $e$.*

***Step 2 (Valuation):** Compute individual-level valuation $P(y_x; u)$ in Def. 24 for each unit $u$.*

***Step 3 (Reduction):** Aggregate these individual-level valuations to obtain the population-level valuation as follows:*

$$P(Y(x) = y|e) = \sum_u P(y_x; u)P(u|e), \tag{26}$$

Notice that in the DiscoSCM framework, the counterfactual outcome $Y_u^d(t)$ is still a random variable that equals in distribution to $Y_u$, rather than a constant $y$, when observing $X_u = x, Y_u = y$ for an individual $u$. In other words, it can conceptually be seen as an extension of the two preceding frameworks, achieved by replacing the traditional consistency rule with a distribution-consistency rule.

[3] Note that $\mathbf{E}(\mathbf{x})$ is not a function of $x$, but rather a random variable indexed by $x$. Importantly, it shares the same distribution as $\mathbf{E}$.

[4] When $X = \emptyset$, we simplify the notation $Y^d(x)$ to $Y^d$ and $E_x$ to $E^d$.