Theoretical Foundations of Predictive Bayes

Sandra Fortini

Bocconi University

Post-Bayes seminar series
May 20 2025

Joint work with:
Sonia Petrone, *Bocconi University*

Exchangeability, Prediction and Predictive Modeling in Bayesian Statistics. Stat. Science, 40(1):40 – 67, 2025.

Predictive Bayes foundations

This seminar explores a predictive approach to Bayesian inference, in which the learning process is specified directly through **predictive distributions**, rather than by placing priors on parameters.

The goal is twofold:

- To provide conceptual motivation for modeling learning through prediction rather than priors;
- To highlight the practical advantages of this approach, including flexibility, interpretability, and computational simplicity.

To explain the predictive perspective, we will first revisit the foundational principles of Bayesian statistics—specifically, what it means to specify a Bayesian model.

In this talk, I focus on the case of **independent observations from a common distribution**, while noting that some results extend to more general settings.

1. Bayesian statistics as a learning process
2. Exchangeability and prediction
3. Conditionally identically distributed random variables
4. Predictive-based approximations of the posterior distribution.
5. A general framework for defining learning rules
6. Discussion

Bayesian vs Frequentist

Assume $X_1, X_2, \ldots$ are independent observations from an unknown distribution F.

Frequentist approach:

- Estimates F through a data-based estimator.
- Distinguishes between:
  - *Aleatory uncertainty*: randomness (e.g., coin toss outcomes).
  - *Epistemic uncertainty*: lack of knowledge (e.g., how the coin is weighted).
- Parameters are fixed but unknown — uncertainty is epistemic.

Bayesian approach (Ramsey (1926), Savage (1954), Dubins and Savage (1965), de Finetti (1931)):

- Assigns a *prior distribution* to F.
- Updates beliefs via Bayes' rule to get the *posterior*.
- Probability models all types of uncertainty.
- Parameters (e.g., coin bias) are treated as random variables.

Key point: The two approaches may look similar, but differ conceptually.

Bayesian Statistics as a Learning Process

Bayesian statistics models both **observations** and **parameters** within a single probabilistic framework.

Learning is represented by conditioning: beliefs are updated via conditional distributions.

In the setting of independent observations from an unknown distribution:

- The unknown distribution F is treated as a random variable $\tilde{F} \sim \pi$, expressing prior uncertainty.
- Given $\tilde{F} = F$, the data X_n are i.i.d. from F: $\quad (X_n) \mid \tilde{F} \sim \tilde{F}^{\infty}$

This defines the joint distribution of $(\tilde{F}, X_1, X_2, \dots)$.

Key point: The Bayesian model is a probability measure on $\mathcal{P}(\mathbb{X}) \times \mathbb{X}^{\infty}$.

There is **no formal distinction** between parameters and observations:

- All unknowns are treated probabilistically.
- Inference is simply conditioning on observed data.

This is why there are **no separate estimators** in Bayesian statistics:

- Once the model is specified, inference is an exercise in probability.

Properties of empirical and predictive distributions

Let $(X_n) \mid \tilde{F} \sim \tilde{F}^{\infty}, \quad \tilde{F} \sim \pi$, and

$$\hat{F}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(A) \quad \text{empirical distribution}$$

$$P_n(A) = \mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) \quad \text{predictive distribution}$$

- For every bounded continuous function g

$$\int g(x)\hat{F}_n(dx) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{\mathbb{P}(\cdot \mid \tilde{F})} \mathbb{E}(g(X_1) \mid \tilde{F}) = \int g(x)\tilde{F}(dx).$$

- Marginalizing with respect to $\tilde{F}$, we see that **the empirical distribution $\hat{F}_n$ converges weakly to $\tilde{F}$,** $\mathbb{P}$-a.s.
- In particular, $\tilde{F}$ **is a function of** (X_n).
- P_n **converges weakly to** $\tilde{F}$:

$$\begin{aligned} P_n(A) &= \mathbb{E}(\mathbb{P}(X_{n+1} \in A \mid \tilde{F}, X_1, \ldots, X_n) \mid X_1, \ldots, X_n) \\ &= \mathbb{E}(\tilde{F}(A) \mid X_1, \ldots, X_n) \to \mathbb{E}(\tilde{F}(A) \mid X_1, X_2, \ldots) = \tilde{F}(A). \end{aligned}$$

A Bayesian model via the distribution of the observations

Consider again $(X_n) \mid \tilde{F} \sim \tilde{F}^{\infty}, \quad \tilde{F} \sim \pi$, which defines a probability measure on $\mathcal{P}(\mathbb{X}) \times \mathbb{X}^{\infty}$.

By marginalizing over $\tilde{F}$, we obtain the joint distribution of the observation sequence $(X_1, X_2, \dots)$.

However, $\tilde{F}$ **is a function of the sequence** (X_n):

$$\tilde{F}(A) = \lim_{n\to\infty} \hat{F}_n(A) = \lim_{n\to\infty} P_n(A),$$

where

$$\hat{F}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(A) \quad \text{(empirical distribution)},$$

$$P_n(A) = \mathbb{P}(X_{n+1} \in A \mid X_1, \dots, X_n) \quad \text{(predictive distribution)}.$$

Thus, this Bayesian model is **fully characterized by the joint distribution of** (X_n).

de Finetti's Theorem

Key question: When does the distribution of (X_n) admit a representation

$$(X_n) \mid \tilde{F} \sim \tilde{F}^{\infty} \text{ for some random } \tilde{F}?$$

de Finetti's Theorem (1930s)

There exists $\tilde{F}$ such that $(X_n) \mid \tilde{F} \sim \tilde{F}^{\infty}$
if and only if (X_n) is **exchangeable**, i.e.,

$$(X_{\sigma(1)}, X_{\sigma(2)}, \dots) \overset{d}{=} (X_1, X_2, \dots)$$

for every finite permutation σ of $\mathbb{N}$.

Interpretation: de Finetti's theorem suggests an alternative way to define a Bayesian model:

- **Specify directly an exchangeable distribution for the observations**.

Contrast with frequentist modeling:

- The probabilistic model accounts for both *aleatory* and *epistemic* uncertainty.

Even under random sampling, the X_i are not independent:
observing $X_1, \dots, X_n$ **gives information about** $X_{n+1}, X_{n+2}, \dots$.

Assigning a Bayesian model through a joint distributions of the observations

What is the advantage of directly specifying an exchangeable distribution for the observations?

It eliminates the need to assign a prior distribution to parameters that **lack a clear interpretation** in terms of observables.

However, **a prior distribution is still implicitly** defined.

- By de Finetti's theorem, the distribution of (X_n) **implicitly defines a random distribution** $\tilde{F}$.
- $\tilde{F}$ is the almost sure **limit of both the empirical and the predictive distributions**.

Specifying a Bayesian Model via Predictive Distributions

In practice, assigning the full law of $(X_1, X_2, \dots)$ in a way that captures both intrinsic aleatory and epistemic uncertainty is challenging.

Predictive distributions offer a solution.

It is well known that the joint distribution of a sequence (X_n) can be constructed from its sequence of predictive distributions:

Theorem (Ionescu-Tulcea)

Let P_0 be a probability measure on a measurable space $(\mathbb{X}, \mathcal{X})$, and for each $n \geq 1$, let P_n be a function satisfying:

(i) *For each $A \in \mathcal{X}$, $P_n(A \mid x_1, \dots, x_n)$ is measurable in $(x_1, \dots, x_n)$,*

(ii) *For each fixed $(x_1, \dots, x_n)$, $P_n(\cdot \mid x_1, \dots, x_n)$ is a probability measure on $\mathcal{X}$.*

Then there exists a stochastic process (X_n) such that:

$$X_1 \sim P_0, \quad X_{n+1} \mid X_1, \dots, X_n \sim P_n(\cdot \mid X_1, \dots, X_n)$$

Moreover, the joint distribution of (X_n) is uniquely determined.

Advantage: Predictive distributions naturally reflect epistemic uncertainty: P_n encodes our **beliefs** about X_{n+1} given the data $X_1, \dots, X_n$.

Predictive distributions as learning rules

In this perspective, a predictive distribution is *not* a physical mechanism generating X_{n+1} from past data.

Rather, $\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n)$ is a **learning rule**:
a conditional probability that formalizes how we update our beliefs about future events based on current information.

We refer to this predictive perspective as **predictive modeling**.

In predictive modeling, reasoning focuses on observable quantities
and on how the sample informs prediction.

Predictive modeling is also intriguing as a form of **Bayesian learning without a prior**:
no explicit prior is required, although a prior may be **implied** by the predictive structure.

Example: Predictive Learning for Binary Outcomes

Example (Eggenberger and Pólya (1923); Pólya (1931))

Let (X_n) be a sequence of random variables taking values in $\{0, 1\}$.

We aim to learn the probability of observing a 1 by directly modeling predictive probabilities.

Initially, the predictive probability for X_1 is given by:

$$\mathbb{P}(X_1 = 1) = \frac{\alpha_1}{\alpha}, \quad \text{with } \alpha_1 < \alpha.$$

After observing $X_1, \ldots, X_n$, the prediction for X_{n+1} is:

$$\mathbb{P}(X_{n+1} = 1 \mid X_1, \ldots, X_n) = \frac{\alpha_1 + \sum_{i=1}^{n} X_i}{\alpha + n}.$$

This defines a sequence of predictive distributions that **reinforce** the observed outcomes: each time a 1 is observed, the predicted probability of future 1's increases.

Interpretation: Learning occurs through recursive prediction updates—
each observation incrementally adjusts the belief about future outcomes.

1. Bayesian statistics as a learning process
2. Exchangeability and prediction
3. Conditionally identically distributed random variables
4. Predictive-based approximations of the posterior distribution.
5. A general framework for defining learning rules
6. Discussion

Predictive distributions and exchangeability

The Ionescu-Tulcea theorem imposes no constraints on the predictive distributions: for any sequence (P_n) such that each P_n is a measurable function of $X_1, \dots, X_n$ and a probability measure, there exists a sequence (X_n) with P_n as its conditional distributions.

Question: What constraints must the (P_n) satisfy in order for (X_n) to be *exchangeable*?

Theorem (Fortini et al. (2000))

Let $(X_n)_{n\geq 1}$ be a sequence of random variables with predictive rule $(P_n)_{n\geq 0}$. Then (X_n) is exchangeable if and only if, for all $n \geq 0$:

i) $\forall n$, $\forall A$, $\mathbb{P}(X_{n+1} \in A \mid X_1, \dots, X_n)$ *is symmetric in* $X_1, \dots, X_n$.

ii) $\forall n$, $\mathbb{P}(X_{n+1} \in A, X_{n+2} \in B \mid X_1, \dots, X_n)$ *is symmetric in* (A, B).

Example: Verifying Exchangeability

Example (Binary outcomes)

Let (X_n) be binary random variables in $\{0, 1\}$ with predictive rule:

$$\mathbb{P}(X_1 = 1) = \frac{\alpha_1}{\alpha}, \quad \mathbb{P}(X_{n+1} = 1 \mid X_1, \dots, X_n) = \frac{\alpha_1 + \sum_{i=1}^n X_i}{\alpha + n}.$$

Condition (i): P_n depends only on $\sum X_i \Rightarrow$ symmetry in the past.
Condition (ii):

$$\mathbb{P}(X_{n+1} = 1, X_{n+2} = 0 \mid X_{1:n}) = \frac{\alpha_1 + \sum_{i=1}^n X_i}{\alpha + n} \cdot \frac{(\alpha + n + 1) - (\alpha_1 + \sum_{i=1}^n X_i + 1)}{\alpha + n + 1}$$
$$= \mathbb{P}(X_{n+1} = 0, X_{n+2} = 1 \mid X_{1:n}).$$

Conclusion: (X_n) is exchangeable.

Moreover the unknown probability of outcome 1 is:

$$\tilde{F}(1) = \lim_{n\to\infty} \frac{\alpha_1 + \sum_{i=1}^n X_i}{\alpha + n} = \lim_{n\to\infty} \frac{\sum_{i=1}^n X_i}{n}$$

(Strong law of large numbers for exchangeable sequences).

Necessity of the second condition

Notice that the symmetry of $\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n)$ is *not sufficient* for exchangeability.

Example

Let $\mathbb{X} = \{0, 1\}$ and define:

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_2 = 1 \mid X_1) = \alpha \in (0, 1),$$

$$\mathbb{P}(X_{n+1} = 1 \mid X_1, \ldots, X_n) = \frac{\sum_{j=1}^{n} X_j}{n} \text{ for } n > 1.$$

Then, for example,

$$\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 1) = 0 \neq \alpha(1 - \alpha)/2 = \mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 0),$$

so the joint distribution is not symmetric.

Assigning predictive distributions that satisfy conditions (i) and (ii) of the above theorem is **not an easy task** in general.

Sufficient statistics

A useful strategy to simplify the assignment of predictive distributions is to use **predictive sufficient statistics**.

First, recall condition (i) in the predictive characterization of exchangeability

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) \text{ is symmetric in } X_1, \ldots, X_n.$$

which implies that the predictive distribution

$$P_n \textbf{ is a function of } \widehat{F}_n = \tfrac{1}{n} \textstyle\sum_{k=1}^{n} \delta_{X_k}.$$

If P_n depends on $\widehat{F}_n$ only through a statistic $T_n = T(\widehat{F}_n)$, i.e.,

$$\exists q_n : P_n = q_n(T(\hat{F}_n)),$$

then T_n is called a **predictive sufficient statistic**.

Sufficient statistics and parametric models

Theorem (Fortini et al. (2000))

Under regularity conditions, if $T(\hat{F}_n)$ is predictive sufficient, then

$$\exists q : P_n = q_n(T(\widehat{F}_n)) \longrightarrow q(T(\tilde{F})).$$

This implies that

$$\tilde{F} = q(T(\tilde{F})).$$

Thus,

- *the model $\tilde{F} = q(\tilde{\theta})$ is* **parametric**
- *the parameter $\tilde{\theta} = T(\tilde{F})$ is the* **limit of the sufficient statistic***.*

The regularity conditions ensure that small changes in the predictive sufficient statistic T_n do not produce abrupt changes in the predictive distribution P_n.

Dominated parametric models:
If the predictive distributions P_n are:

- absolutely continuous with respect to a measure λ, and
- converge in total variation,

then the model $\tilde{F} = q(\tilde{\theta})$ is **dominated** (Fortini and Petrone, 2025).

Examples of predictive constructions

A natural framework for modeling learning is **reinforcement learning** (Pemantle, 2007), where the probability assigned to an event increases whenever the event occurs.

Example (Pólya sequence, Blackwell and MacQueen (1973))

Let $\mathbb{X}$ be a Polish space.
The sequence (X_n) is defined by the predictive rule:

$$X_1 \sim P_0, \quad X_{n+1} \mid X_1, \dots, X_n \sim P_n = \frac{\alpha P_0 + \sum_{i=1}^{n} \delta_{X_i}}{\alpha + n}, \quad \text{for } \alpha > 0.$$

This predictive construction corresponds to a Bayesian model with:

$$\tilde{F} \sim \mathrm{DP}(\alpha, P_0), \quad X_i \mid \tilde{F} \sim \tilde{F}.$$

It also admits an urn interpretation where the X_n represent colors (Hoppe (1984, 1987), Aldous (1985))
Description: Start with an urn containing $\alpha > 0$ black balls. At each step:

- If a colored ball is drawn, return it and add another of the same color.
- If a black ball is drawn, a new color is drawn from P_0 and added to the urn together with the black ball.

Examples of predictive constructions

Example (*Kernel-based Dirichlet sequences, Berti et al. (2023)*)

Kernel-based Dirichlet sequences are exchangeable sequences where the predictive distribution replaces the point masses δ_{X_i} of the Pólya urn scheme with a smoothed version via a probability kernel K:

$$P_n(\cdot) = \frac{\alpha P_0 + \sum_{i=1}^{n} K(\cdot \mid X_i)}{\alpha + n}.$$

Exchangeability holds if and only if the kernel satisfies

$$K(\cdot \mid x) = P_0(\cdot \mid \mathcal{G})(x)$$

for some σ-algebra $\mathcal{G}$ on $\mathbb{X}$ (Berti et al. (2023), Sariev and Savov (2024)).

In particular, if $K(\cdot \mid x) \ll P_0$ for all $x \in \mathbb{X}$, the associated de Finetti measure $\tilde{F}$ corresponds to a mixture model where the component distributions (the kernels) have known disjoint supports—for instance, a histogram with fixed bins.

This example highlights that **exchangeability can be a strong constraint**, especially when aiming for both tractable prediction rules and specific modeling flexibility.

1. Bayesian statistics as a learning process
2. Exchangeability and prediction
3. Conditionally identically distributed random variables
4. Predictive-based approximations of the posterior distribution.
5. A general framework for defining learning rules
6. Discussion

Beyond Exchangeability

There are several reasons to **relax the exchangeability assumption**, even when modeling homogeneous observations:

- To derive predictive distributions with simpler or closed-form expressions;
- To increase modeling flexibility;
- To reduce computational cost and avoid MCMC-based methods;
- To bypass the need for explicitly specifying a prior.

Key question: How can we relax exchangeability while preserving the idea that **observations arise from the same unknown distribution**?

Exchangeability and its structural components

An exchangeable sequence (X_n) satisfies, for every $n \geq 2$:

$$(X_{n+1}, X_{n+2}, \dots) \mid \tilde{F} \overset{d}{=} (X_1, X_2, \dots) \mid \tilde{F} \sim \tilde{F}^{\infty}.$$

Marginalizing:

$$(X_{n+1}, X_{n+2}, \dots) \overset{d}{=} (X_1, X_2, \dots),$$

so the sequence (X_n) is **strongly stationary**.

Moreover, an exchangeable sequence (X_n) satisfies $\forall n, k$

CID condition

$$X_{n+k} \mid (X_1, \dots, X_n) \overset{d}{=} X_{n+1} \mid (X_1, \dots, X_n) \sim \mathbb{E}(\tilde{F} \mid X_1, \dots, X_n),$$

meaning that the X_n are **conditionally identically distributed** (Berti et al., 2004).

Characterization of Exchangeability (Kallenberg, 1988)

$$\textbf{Exchangeability} = \textbf{Stationarity} + \textbf{CID}$$

The proof is based on the ergodic theorem.

Conditionally Identically Distributed (CID) Random Variables

We can relax the assumption of exchangeability by requiring only the **CID** condition (conditionally identically distributed), thereby giving up strong stationarity.

Under the CID condition, **all future observations share the same conditional distribution given the past**.

In fact, it is sufficient to assume that just **two consecutive** future observations have the same conditional distribution:

Sufficient Condition for CID

$$X_{n+2} \mid X_1, \ldots, X_n \overset{d}{=} X_{n+1} \mid X_1, \ldots, X_n \quad \text{for all } n.$$

This is **equivalent** to a martingale condition on the predictive distributions:

Martingale Characterization

$$\mathbb{E}(P_{n+1}(A) \mid X_1, \ldots, X_n) = P_n(A) \quad \text{for all } A \text{ and all } n.$$

Proof sketch:

$$\mathbb{E}(P_{n+1}(A) \mid X_{1:n}) = \mathbb{E}\Big[\mathbb{P}(X_{n+2} \in A \mid X_{1:n+1}) \,\Big|\, X_{1:n}\Big] = \mathbb{P}(X_{n+2} \in A \mid X_{1:n}) = P_n(A).$$

Properties of CID sequences

How close is CID to exchangeability?

- The X_n are **identically distributed**
- The sequence (P_n) of the **predictive** distributions is a **martingale measure** and **converges to a random probability measure** $\tilde{F}$**.**
- By martingale methods, it is possible to prove that **if** P_n **converge to** $\tilde{F}$**, then also the empirical distributions** $\hat{F}_n$ **converge to the same** $\tilde{F}$**.**
- A sequence (X_n) whose predictive distributions converge to a random probability measure $\tilde{F}$ is **asymptotically exchangeable** (Aldous, 1985):

  $$(X_{n+1}, X_{n+2}, \dots) \xrightarrow{d} (Z_1, Z_2, \dots),$$

  where (Z_n) is exchangeable with directing measure $\tilde{F}$.

Thus, a CID sequence **implicitly defines a prior distribution**: the law of $\tilde{F}$.

An example of CID sequence

Example (Weighted Pólya Urn, Fortini et al. (2018))

Let $X_1 \sim P_0$, and for any $n \geq 1$,

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1, \ldots, X_n) = \frac{\alpha P_0(\cdot) + \sum_{k=1}^{n} W_k \delta_{X_k}(\cdot)}{\alpha + \sum_{k=1}^{n} W_k},$$

where the **weights** W_k **are positive random variables**, and for every $n \geq 0$, W_{n+1} is conditionally independent of X_{n+1} given $X_1, \ldots, X_n$.

This weighted Pólya urn scheme models a learning process in which some observations exert more influence than others.

This scheme accounts for varying levels of reliability across observations, assigning greater influence to those considered less noisy or more accurate.

1. Bayesian statistics as a learning process
2. Exchangeability and prediction
3. Conditionally identically distributed random variables
4. **Predictive-based approximations of the posterior distribution.**
5. A general framework for defining learning rules
6. Discussion

A Predictive-Based Simulation Scheme

For CID sequences, the associated prior and posterior distributions are **generally unknown in closed form**.

However, we can **simulate** (Fortini and Petrone, 2020, 2023) from the posterior by exploiting the convergence of the predictive distribution P_n to the directing random measure $\tilde{F}$.

To generate a sample from the posterior given observations $(x_1, \ldots, x_n)$:

- Select a grid of points $t_1, \ldots, t_k$;
- **Sequentially generate future observations** $(x_{n+1}, \ldots, x_N)$**, where each** x_{i+1} **is drawn from the predictive distribution** $P_i(\cdot \mid x_1, \ldots, x_i)$ for $i = n, \ldots, N-1$;
- Using the extended sample $(x_1, \ldots, x_N)$, compute $[P_N(t_1), \ldots, P_N(t_k)]$;
- Since N is large, this vector approximates a sample from the posterior distribution of $[\tilde{F}(t_1), \ldots, \tilde{F}(t_k)]$;
- Repeating the procedure M times yields a Monte Carlo sample of size M from the posterior.

A Predictive-Based Asymptotic Approximation of the Posterior Distribution

While posterior simulation is possible, for large n a direct asymptotic approximation becomes available.

Theorem (Fortini and Petrone (2023))

Let (α_n) be such that

$$\frac{\mathbb{E}\left[(P_n(t) - P_{n-1}(t))^2 \mid X_1, \ldots, X_{n-1}\right]}{\alpha_n^2} \longrightarrow U_{\mathbf{t}} > 0 \quad \mathbb{P}\text{-}a.s.$$

Let $b_n = \left(\sum_{k \geq n} \alpha_k^2\right)^{-1}$ *and* $V_{n,t} = \dfrac{1}{n} \sum_{m=1}^{n} \dfrac{(P_m(t) - P_{m-1}(t))^2}{\alpha_m^2}$.
Under mild assumptions,

$$\tilde{F}(t) \mid X_1, \ldots, X_n \approx \mathcal{N}\left(P_n(t), \frac{V_{n,t}}{b_n}\right).$$

The result extends to $[\tilde{F}(t_1), \ldots, \tilde{F}(t_k)]$.

Implications of the Predictive-Based Approximation

This result allows for the construction of asymptotic credible intervals for $\tilde{F}(t)$:

$$\left[P_n(t) \pm z_{1-\alpha/2} \sqrt{\frac{V_{n,t}}{b_n}} \right],$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

Since $P_n(t) = \mathbb{E}(\tilde{F}(t) \mid X_1, \dots, X_n)$, the approximation is centered in the **posterior mean** of $\tilde{F}(t)$.

Since $b_n = \left(\sum_{k \geq n} \alpha_k^2 \right)^{-1}$, with $\alpha_n^2 \approx \mathbb{E}\left[(P_n(t) - P_{n-1}(t))^2 \mid X_1, \dots, X_{n-1} \right]$, then **the size of the interval** depends on how fast the **predictive updates**

$$\Delta_{t,n} = P_n(t) - P_{n-1}(t) \to 0.$$

Typically, $|P_n(t) - P_{n-1}(t)| \approx |\hat{F}_n(t) - \hat{F}_{n-1}(t)| \approx 1/n$ so that $b_n \approx n$.

Not a Bernstein–von Mises Result

The approximation

$$\tilde{F}(t) \mid X_{1:n} \approx \mathcal{N}\left(P_n(t), \frac{V_{n,t}}{b_n}\right)$$

is **not a consistency** result in the classical sense.

Classical posterior consistency (as in Bernstein–von Mises theorems) studies the asymptotic behavior of the posterior distribution assuming the data are i.i.d. from a *true* distribution F_0.

In our setting, the asymptotic result is given under the **distribution of the** (X_n) defined by the sequence of the **predictive distribution** (P_n).

The goal is to provide an asymptotic **predictive-based approximation** of the posterior distribution of $\tilde{F}(t)$.

Consistency and contraction rate are interesting open problems.

1. Bayesian statistics as a learning process

2. Exchangeability and prediction

3. Conditionally identically distributed random variables

4. Predictive-based approximations of the posterior distribution.

5. A general framework for defining learning rules

6. Discussion

A General Framework for Defining Learning Rules

A broad class of recursive predictive rules can be specified by a recursive scheme:

Recursive Predictive Scheme, Fortini and Petrone (2025)

$$X_1 \sim P_0, \quad \text{and for every } n \geq 1, \quad X_{n+1} \mid X_1, \ldots, X_n \sim P_n,$$

with

$$\begin{cases} P_n = q_n(T_n), \\ T_n = h_n(T_{n-1}, X_n), \end{cases}$$

where q_n and h_n are given functions, and T_n is a predictive sufficient statistic.

This formulation allows storing only T_n, which can be easily updated from T_{n-1} and X_n.

Appropriate choices of q_n and h_n can ensure desirable properties of the sequence $(X_n)_{n\geq 1}$, such as asymptotic exchangeability.

Bayesian rules as sequential updating procedures

In an exchangeable setting, many common Bayesian models admit this representation.

Example (Two color Pólya urn)

$$P_n(1) = \frac{\alpha_1 + \sum_{i=1}^n X_i}{\alpha + n} \rightsquigarrow \begin{cases} P_n(1) = \dfrac{\alpha_1 + nT_n}{\alpha + n}, \\ T_n = T_{n-1} + \dfrac{1}{n}(X_n - T_{n-1}) \end{cases}$$

with $T_0 = 0$

Example (Dirichlet process prior)

$$P_n = \frac{\alpha + \sum_{i=1}^n \delta_{X_i}}{\alpha(\mathbb{X}) + n} \rightsquigarrow \begin{cases} P_n = \dfrac{\alpha + nT_n}{\alpha(\mathbb{X}) + n}, \\ T_n = T_{n-1} + \dfrac{1}{n}(\delta_{X_n} - T_{n-1}). \end{cases}$$

with $T_0 = 0$.

Beyond Exchangeability

Measure-Valued Pólya Sequences (Sariev and Savov, 2024)

Let R be a finite, non-null transition kernel on the sample space $\mathbb{X}$, $\gamma > 0$ a constant, and P_0 a probability measure. Define:

$$P_n = \frac{\gamma P_0 + \sum_{i=1}^{n} R_{X_i}}{\gamma + \sum_{i=1}^{n} R_{X_i}(\mathbb{X})}, \quad n \geq 1.$$

Set $T_0 = \gamma P_0$, and define:

$$\begin{cases} P_n = \dfrac{T_n}{T_n(\mathbb{X})}, \\ T_n = T_{n-1} + R_{X_n}. \end{cases}$$

Urn interpretation:
Any measurable set $B \subset \mathbb{X}$ starts with mass $T_0(B)$. At each step n, B receives an additional mass $R_{X_n}(B)$, which depends on the observed value X_n.

A measure-valued Pólya sequence $(X_n)_{n \geq 1}$ is exchangeable if and only if:

$$R_{X_n}(\cdot) = P_0(\cdot \mid \mathcal{G})(X_n)$$

for some σ-algebra $\mathcal{G}$.

A learning process for mixtures

An example of learning rule satisfying the recursive scheme is the learning rule for mixtures introduced in (Fortini and Petrone, 2020) and based on Newton's algorithm (Newton, 2002).

Let λ be a measure on $\mathbb{X}$, G_0 be a measure on Θ, $\alpha_n > 0$, $\sum_n \alpha_n = +\infty$ and $\sum_n \alpha_n^2 < +\infty$. Set $P_0(dx) = \int_\Theta k(x \mid \theta) G_0(d\theta)\lambda(dx)$ and, for $n \geq 1$,

$$\begin{cases} P_n(dx) = \int_\Theta k(x \mid \theta) G_n(d\theta)\lambda(dx) \\ G_n(d\theta) = G_{n-1}(d\theta) + \alpha_n G_{n-1}(d\theta)\left(\dfrac{k(X_n \mid \theta)}{\int_\Theta k(X_n \mid \theta) G_{n-1}(d\theta)} - 1\right) \end{cases}$$

G_n is a weighted average of G_{n-1} and of the **posterior, given X_n, with prior** G_{n-1}.

- (G_n) is a martingale measure and converges to a random probability measure $\tilde{G}$
- (X_n) is CID and therefore asymptotically exchangeable and its asymptotic directing random measure is $\tilde{F}(dx) = \int k(x \mid \theta)\tilde{G}(d\theta)\lambda(dx)$
- (P_n) defines implicitly a novel prior which is absolutely continuous with respect to λ
- although the posterior distribution of $\tilde{G}$ remains unknown, it is possible to sample from it and to give asymptotic Gaussian approximations.

Learning via Loss Functions: Binary Classification

A further example of a recursive learning rule is provided by a process for binary classification introduced in Fortini and Petrone (2025).

At each time step n, a pair (X_n, Y_n) is observed, where X_n denotes a feature vector and $Y_n \in \{0,1\}$ is a binary label.

Sequential learning for classification

Let β_0 be a fixed parameter vector. Define:
$P_0(dx, y) = P_X(dx) \cdot g(x, \beta_0)^y (1 - g(x, \beta_0))^{1-y}$ and

$$\begin{cases} P_n(dx, y) = P_X(dx) \cdot g(x, \beta_n)^y (1 - g(x, \beta_n))^{1-y}, & y \in \{0,1\}, \\ \beta_n = \beta_{n-1} - \alpha_n \nabla_\beta L(\beta_{n-1}; X_n, Y_n), \end{cases}$$

where: - $g(x, \beta)$ is a known function (e.g., logistic function), - L is a loss function (e.g., cross-entropy), - (α_n) is a learning rate sequence.

Asymptotic properties (under mild assumptions):

- $\beta_n \xrightarrow{\text{a.s.}} \tilde{\beta}$, a random limit.
- The sequence (β_n) is a martingale: $\beta_n = \mathbb{E}(\tilde{\beta} \mid X_{1:n}, Y_{1:n})$.
- The sequence $((X_n, Y_n))$ is asymptotically exchangeable.
- The directing random measure $\tilde{F}_{X,Y}$ satisfies: $\tilde{F}_{Y|X} = \text{Bernoulli}(g(X, \tilde{\beta}))$.

Asymptotics for the Posterior Distribution of $\tilde{\beta}$

Although the posterior distribution of $\tilde{\beta}$ is not available in closed form, it is possible to approximate it using asymptotic Gaussian results (Fortini and Petrone, 2025).

Asymptotic Gaussian approximation

Under mild regularity conditions and with learning rate $\alpha_n = \frac{1}{\alpha+n}$:

$$\tilde{\beta} \mid X_{1:n}, Y_{1:n} \approx \mathcal{N}_d \left(\beta_n, \frac{V_n}{n} \right),$$

where

$$V_n = \frac{1}{n} \sum_{k=1}^{n} k^2 (\beta_k - \beta_{k-1})(\beta_k - \beta_{k-1})^T.$$

This allows us to construct **asymptotic credible regions for the random vector $\tilde{\beta}$.**

Importantly, **no knowledge of the distribution P_X is required.**

1. Bayesian statistics as a learning process
2. Exchangeability and prediction
3. Conditionally identically distributed random variables
4. Predictive-based approximations of the posterior distribution.
5. A general framework for defining learning rules
6. Discussion

Discussion

- The Bayesian paradigm provides a powerful framework for learning from data, and performs well in many contexts.

- An alternative perspective is to directly model the **learning process**, rather than starting from a prior.

- This approach offers greater **flexibility** and often leads to more **interpretable** models.

- Specifying simple predictive distributions can greatly simplify computation—**without sacrificing** key properties such as asymptotic exchangeability.

  Many open questions remain, including:

  - the study of **consistency** and **contraction rates**;
  - extensions to more general **symmetries** and **complex models**.

Thank you!

Aldous, D. J. (1985). Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII 1983*, 1117:1–198.

Berti, P., Dreassi, E., Leisen, F., Pratelli, L., and Rigo, P. (2023). Kernel based dirichlet sequences. *Bernoulli*, 29:1321–1342.

Berti, P., Pratelli, L., and Rigo, P. (2004). Limit theorems for a class of identically distributed random variables. *Ann. Probab.*, 32(3):2029–2052.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.*, 1(2):353–355.

de Finetti, B. (1931). Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17(1):298–329.

Dubins, L. and Savage, L. (1965). *How to Gamble if You Must. Inequalities for Stochastic Processes*. McGraw-Hill, New York.

Eggenberger, F. and Pólya, G. (1923). Über die statistik verketteter vorgänge. *Z. Angew. Math. Mech. (Appl. Math. Mech.)*, 3:279–289.

Fortini, S., Ladelli, L., and Regazzini, E. (2000). Exchangeability, predictive distributions and parametric models. *Sankya, Series A*, 62:86–109.

Fortini, S. and Petrone, S. (2020). Quasi-Bayes properties of a procedure for sequential learning in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 82(4):1087–1114.

Fortini, S. and Petrone, S. (2023). Prediction-based uncertainty quantification for exchangeable sequences. *Phil. Trans. R. Soc. A*, 381:20220142.

Fortini, S. and Petrone, S. (2025). Exchangeability, Prediction and Predictive Modeling in Bayesian Statistics. *Statistical Science*, 40(1):40 – 67.

Fortini, S., Petrone, S., and Sporysheva, P. (2018). On a notion of partially conditionally identically distributed sequences. *Stoch. Process. Appl.*, 128(3):819–846.

Hoppe, F. M. (1984). Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.*, 20(1):91–94.

Hoppe, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.*, 25(2):123–159.

Kallenberg, O. (1988). Spreading and predictable sampling in exchangeable sequences and processes. *Ann. Probab.*, 16(2):508–534.

Newton, M. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankyā*, 64:306–322.

Pemantle, R. (2007). A survey on random processes with reinforcement. *Probab. Surv.*, 4:1–79.

Pólya, G. (1931). Sur quelques points de la théorie des probabilités. *Ann. Inst. H. Poincaré*, 1:117–161.

Ramsey, F. (1926). Truth and Probability. In Eagle, A., editor, *Philosophy of Probability: Contemporary Readings.*, pages 52–94. Routledge.

Sariev, H. and Savov, M. (2024). Characterization of exchangeable measure-valued Pólya urn sequences. *Electron. J. Probab.*, 29(none):1 – 23.

Savage, L. (1954). *The Foundations of Statistics*. Wiley, New York.