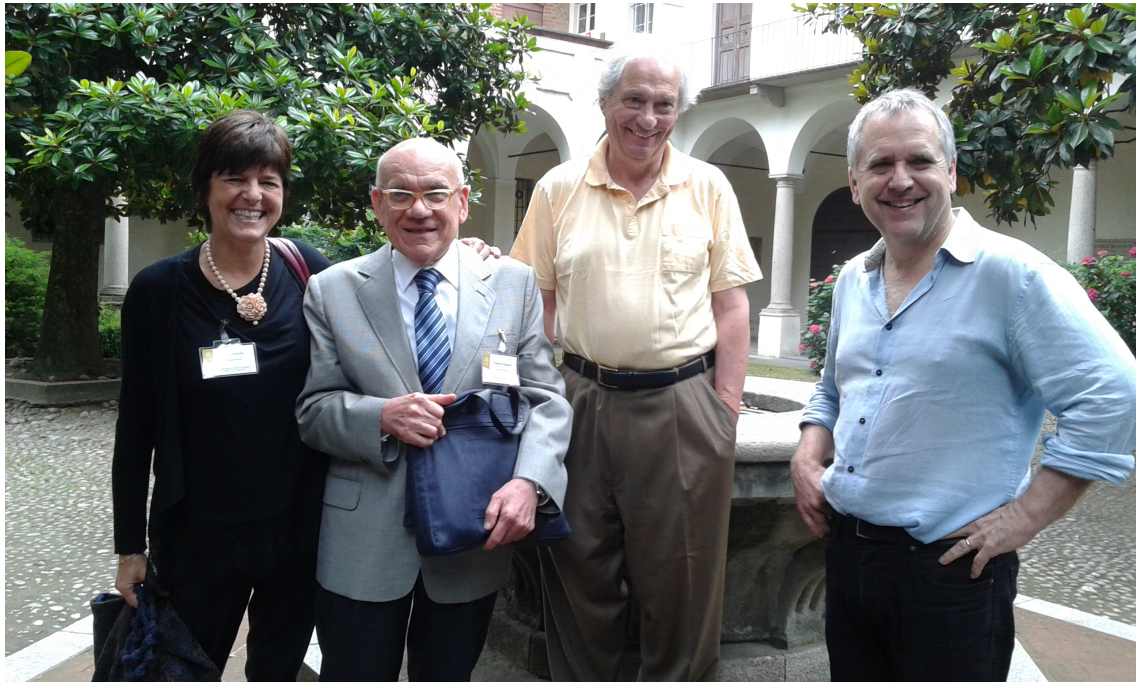
# Bayesian predictive-based uncertainty quantification

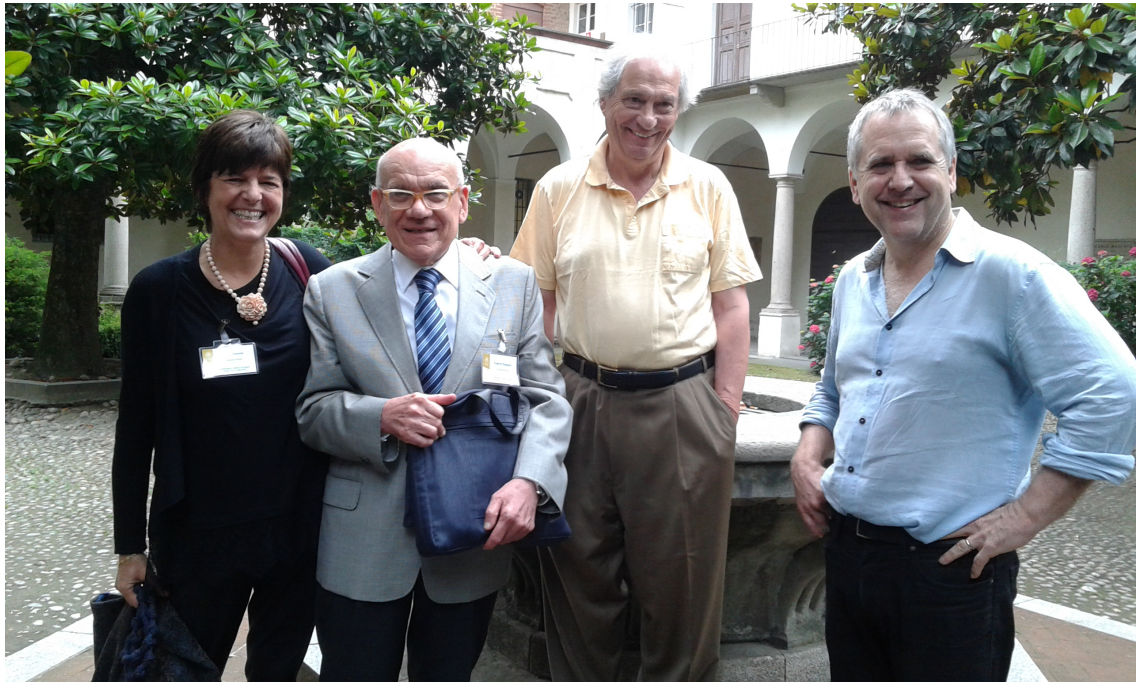
**Sonia Petrone**

Bocconi University, Milan

Second SISBayes workshop, 4-5 September 2025,
University of Padova



tre Maestri: Eugenio Regazzini, Persi Diaconis, Michael Jordan – Conference in honor of Eugenio Regazzini, 10-11 June 2016, Pavia



tre Maestri: Eugenio Regazzini, Persi Diaconis, Michael Jordan – Conference in honor of Eugenio Regazzini, 10-11 June 2016, Pavia

Here, joint works with Sandra Fortini

# Inferential vs predictive approach in Stats

We are all familiar, at least since Leo Breiman's (2001, *Statistical Science*), with the "two cultures" - classic statistical inference versus algorithmic prediction.
And the more so, with Stats and AI...

The Bayesian approach has prediction in its foundations, and can naturally combine both cultures.

# Classic: from inference to prediction

In classic statistics, prediction is guided by the study of the phenomenon of interest and the resulting inferential model.

In the Bayesian approach,

$$(X_1, \ldots, X_n) \mid \theta \sim p(x_1, \ldots, x_n \mid \theta), \quad n \geq 1$$

where $\theta$ is described as random with prior distribution $\pi$,
from which we obtain the **predictive distribution**

$$\mathbf{X_{n+1}} \mid \mathbf{x_{1:n}} \sim \mathbf{p_n}(\mathbf{x_{n+1}} \mid \mathbf{x_{1:n}}) = \int \mathbf{p}(\mathbf{x_{n+1}} \mid \mathbf{x_{1:n}}, \theta)\mathbf{d}\pi(\theta \mid \mathbf{x_{1:n}}),$$

with **full Bayesian uncertainty quantification**.

# Classic: from inference to prediction

In classic statistics, prediction is guided by the study of the phenomenon of interest and the resulting inferential model.

In the Bayesian approach,

$$(X_1, \ldots, X_n) \mid \theta \sim p(x_1, \ldots, x_n \mid \theta), \quad n \geq 1$$

where $\theta$ is described as random with prior distribution $\pi$,
from which we obtain the **predictive distribution**

$$\mathbf{X_{n+1}} \mid \mathbf{x_{1:n}} \sim \mathbf{p_n}(\mathbf{x_{n+1}} \mid \mathbf{x_{1:n}}) = \int \mathbf{p}(\mathbf{x_{n+1}} \mid \mathbf{x_{1:n}}, \theta)\mathbf{d}\pi(\theta \mid \mathbf{x_{1:n}}),$$

with **full Bayesian uncertainty quantification**.

However, specifying the proper model or eliciting the prior may be difficult (e.g., parameters lose interpretation in black-box models). Moreover, **computations** may be overwhelming, especially with streaming data.

# from prediction to inference

On the other hand, we have a wealth of *predictive algorithms* [a strategy to provide predictions, with no explicit likelihood or priors] that perform well.. but often lack clean understanding and uncertainty quantification.

This talk is a review, but a guideline-aim is to show how, by taking a Bayesian predictive approach, we may
- read recursive predictive algorithms as Bayesian predictive learning rules,
- understand the statistical model & prior implicitly used, if any,
- and provide full Bayesian uncertainty quantification.

Examples and potential applications are many, in statistics (e.g., quasi-Bayes approximations of costly Bayesian procedures), and in machine-learning and AI contexts (e.g., understanding if In-Context Learning of LLMs is any Bayesian, or how trained transformers learn)[1]

[1] Susan Wei's talk at *post-Bayes* workshop, UCL, May 2025

1. Bayesian predictive approach.
   - What is a Bayesian predictive rule?
   - Basic desiderata – in random sampling
   - From prediction to inference

2. Asymptotic exchangeability and predictive algorithms
   - Asymptotic exchangeability
   - From the predictive rule (algorithms) to inference
   - Turning algorithms into Bayesian predictive rules
   - Bayesian uncertainty quantification - approximating the posterior distribution

3. Directions and ongoing work
   - Predictive "efficiency"
   - Multivariate extensions
   - Understanding time (order) dependence
   - Beyond random sampling

# 1. Bayesian predictive approach

* In the decision-theoretical foundations of Bayesian statistics, we have agents acting under incomplete information
(not dealing with replicates of experiments)

* Probability is the prescribed way to formalize incomplete information (uncertainty)

* and should be expressed on observable facts.

Thus the modeling effort is to elicit $\mathbf{p}(\mathbf{x_1}, \ldots, \mathbf{x_n})$ (for any $n$).

Models may convey valuable information, but are just a ring of the chain

$$(X_1, \ldots, X_n) \rightarrow \text{ models, parameters } \rightarrow X_{n+1}$$

thus, properties of models andinference should be thought of in their effects on prediction

$\rightarrow$ We can directly reason on what is relevant for prediction and assign

$$\mathbf{p}(\mathbf{x_1}, \ldots, \mathbf{n_n}) = \mathbf{p_0}(\mathbf{x_1})\mathbf{p_1}(\mathbf{x_2} \mid \mathbf{x_1}) \cdots \mathbf{p_{n-1}}(\mathbf{x_n} \mid \mathbf{x_{1:n-1}}).$$

# What is a Bayesian predictive rule?

Because (incomplete) information is expressed through probability, learning is expressed through conditional probability.

* The predictive distribution $\mathbf{P_n}(\cdot) = \mathbb{P}(\mathbf{X_{n+1}} \in \cdot \mid \mathbf{X_1}, \ldots, \mathbf{X_n})$ formalizes how we learn from the data $(X_1, \ldots, X_n)$ on the future observation $X_{n+1}$. (*not* meant as the true mechanism that generates $x_{n+1}$ given $x_{1:n}$)

Any predictive rule in this sense is Bayesian.

* The predictive distributions give the finite-dimensional

$$p(x_1, \ldots, x_n) = p_0(x_1)p_1(x_2 \mid x_1) \cdots p_{n-1}(x_n \mid x_{1:n-1}).$$

The predictive rule $(P_n)_{n \geq 0}$ characterizes the law $\mathbb{P}$ of the process, $(X_n)_{n \geq 1} \sim \mathbb{P}$ (Ionescu-Tulcea theorem).

# Basic desiderata for prediction

There are no formal constraints in assigning the predictive rule... but **natural desiderata!**

$\rightarrow$ **Basic setting: random sampling**

# Basic desiderata for prediction

There are no formal constraints in assigning the predictive rule... but **natural desiderata!**

→ **Basic setting: random sampling**

- exchangeability Not mandatory, but natural to judge that labels do not carry information ('order does not matter')

$$p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

Extending to the infinite sequence $(X_n)_{n \geq 1} \sim \mathbb{P}$, natural to ask invariance under every finite permutation, i.e. $(X_n)_{n \geq 1}$ exchangeable.

# Basic desiderata for prediction

There are no formal constraints in assigning the predictive rule... but **natural desiderata!**

→ **Basic setting: random sampling**

- exchangeability Not mandatory, but natural to judge that labels do not carry information ('order does not matter')

$$p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

Extending to the infinite sequence $(X_n)_{n \geq 1} \sim \mathbb{P}$, natural to ask invariance under every finite permutation, i.e. $(X_n)_{n \geq 1}$ exchangeable.

- Check prediction with facts! Minimal requirement: for $n$ large, $P_n$ should agree with empirical frequences: $\mathbb{P}(d(P_n, \hat{F}_n) \to 0) = 1$.

# Basic desiderata for prediction

There are no formal constraints in assigning the predictive rule... but **natural desiderata!**

→ **Basic setting: random sampling**

- exchangeability Not mandatory, but natural to judge that labels do not carry information ('order does not matter')

$$p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

Extending to the infinite sequence $(X_n)_{n \geq 1} \sim \mathbb{P}$, natural to ask invariance under every finite permutation, i.e. $(X_n)_{n \geq 1}$ exchangeable.

- Check prediction with facts! Minimal requirement: for $n$ large, $P_n$ should agree with empirical frequences: $\mathbb{P}(d(P_n, \hat{F}_n) \to 0) = 1$.

- Link with inference? **Does $P_n$ implicitly use a model and a prior?** This would give the latter a predictive justification, and allow inference.

# Exchangeability

- exchangeability. The predictive rule $(P_n)_{n\geq 0}$ characterizes an exchangeable law $\mathbb{P}$ for $(X_n)_{n\geq 1}$ iff
  $*p(x_{n+1} \mid x_{1:n})$ invariant to permutations of $x_1, \ldots, x_n$;
  $*p(x_{n+1}, \ldots, x_{n+k} \mid x_{1:n})$ invariant to permutations of $x_{n+1}, \ldots, x_{n+k}$

# Exchangeability

- exchangeability. The predictive rule $(P_n)_{n\geq 0}$ characterizes an exchangeable law $\mathbb{P}$ for $(X_n)_{n\geq 1}$ iff
  $*p(x_{n+1} \mid x_{1:n})$ invariant to permutations of $x_1, \ldots, x_n$;
  $*p(x_{n+1}, \ldots, x_{n+k} \mid x_{1:n})$ invariant to permutations of $x_{n+1}, \ldots, x_{n+k}$
- Check with facts. Under exchangeability, the empirical $\hat{F}_n$ and the predictive distributions $P_n$ converge, to the same limit $\tilde{F}$, a random distribution

$$\lim \hat{F}_n = \lim P_n = \tilde{F}.$$

# Exchangeability

- exchangeability. The predictive rule $(P_n)_{n\geq 0}$ characterizes an exchangeable law $\mathbb{P}$ for $(X_n)_{n\geq 1}$ iff
  $*p(x_{n+1} \mid x_{1:n})$ invariant to permutations of $x_1, \ldots, x_n$;
  $*p(x_{n+1}, \ldots, x_{n+k} \mid x_{1:n})$ invariant to permutations of $x_{n+1}, \ldots, x_{n+k}$

- Check with facts. Under exchangeability, the empirical $\hat{F}_n$ and the predictive distributions $P_n$ converge, to the same limit $\tilde{F}$, a random distribution
$$\lim \hat{F}_n = \lim P_n = \tilde{F}.$$

- Link to inference: de Finetti representation theorem. (*informal*). If $(X_n)_{n\geq 1} \sim \mathbb{P}$ is exchangeable, then $\mathbb{P}$ can be represented as
$$X_i \mid \tilde{F} \overset{iid}{\sim} \tilde{F},$$
where $\tilde{F} = \lim \hat{F}_n = \lim P_n$.

# predictive characterizations

Thus, in a predictive approach, we can directly move from the predictive rule $(P_n)_n n \geq 0$; if exchangeable, it characterize the 'model' as $\tilde{F} = \lim P_n$ and the prior as its distribution.

[2] Blackwell & Mac Queen, *Ann. Statist.*, 1973

# predictive characterizations

Thus, in a predictive approach, we can directly move from the predictive rule $(P_n)_n n \geq 0$; if exchangeable, it characterize the 'model' as $\tilde{F} = \lim P_n$ and the prior as its distribution.

Example. Pólya sequences[2] $(X_n)_{n\geq 1}$ such that $X_1 \sim P_0$ and for $n \geq 1$

$$X_{n+1} \mid x_{1:n} \sim P_n(\cdot) = \frac{\alpha}{\alpha + n} P_0(\cdot) + \frac{n}{\alpha + n} \sum_{i=1}^{n} \delta_{x_i}(\cdot)/n.$$

**Thrm** *The sequence* $(X_n)$ *is exchangeable.*
$P_n \to \tilde{F}$ *and* $X_i \mid \tilde{F} \overset{iid}{\sim} \tilde{F}$, *where* $\tilde{F} \sim DP(\alpha, P_0)$.

[2] Blackwell & Mac Queen, *Ann. Statist.*, 1973

# predictive characterizations

Thus, in a predictive approach, we can directly move from the predictive rule $(P_n)_n n \geq 0$; if exchangeable, it characterize the 'model' as $\tilde{F} = \lim P_n$ and the prior as its distribution.

Example. Pólya sequences[2] $(X_n)_{n \geq 1}$ such that $X_1 \sim P_0$ and for $n \geq 1$

$$X_{n+1} \mid x_{1:n} \sim P_n(\cdot) = \frac{\alpha}{\alpha + n} P_0(\cdot) + \frac{n}{\alpha + n} \sum_{i=1}^{n} \delta_{x_i}(\cdot)/n.$$

**Thrm** *The sequence* $(X_n)$ *is exchangeable.*
$P_n \to \tilde{F}$ *and* $X_i \mid \tilde{F} \overset{iid}{\sim} \tilde{F}$, *where* $\tilde{F} \sim DP(\alpha, P_0)$.

Predictive characterizations have a long tradition in Bayesian statistics, and encourage a **tractable** predictive rule. Yet, often not easy to have exchangeability AND a tractable predictive rule!
E..g., beyond trivial cases, smoothing the point masses $\delta_{x_i}(\cdot)$ with kernels $K(\cdot; x_i)$ when dealing with continuous data breaks exchangeability.

[2] Blackwell & Mac Queen, *Ann. Statist.*, 1973

# 2. Asymptotic exchangeability

Our idea is to accept an approximation in terms of asymptotic exchangeability.

# 2. Asymptotic exchangeability

Our idea is to accept an approximation in terms of asymptotic exchangeability.

Exchangeability implies that $P_n \to \tilde{F}$. The reverse is not true, but:
**Thrm**(Aldous, 1983) If $P_n$ converges to a random probability distribution $\tilde{F}$, then $(X_n)_{n\geq 1}$ is asymptotically exchangeable;

# 2. Asymptotic exchangeability

Our idea is to accept an approximation in terms of asymptotic exchangeability.

Exchangeability implies that $P_n \to \tilde{F}$. The reverse is not true, but:
**Thrm**(Aldous, 1983) If $P_n$ converges to a random probability distribution $\tilde{F}$, then $(X_n)_{n \geq 1}$ is asymptotically exchangeable; informally, for $n > N$ large,

$$X_n \mid \tilde{F} \overset{iid}{\approx} \tilde{F} \quad \text{with } \tilde{F} \sim \pi.$$

an approx exchangeable setting where, again, **the model is $\tilde{F} = \lim P_n$ and its probability law is the "implicit" prior.**
Moreover, (Rigo; see Bissiri & Walker, *EJS*, 2025) also $\tilde{F} = \lim \hat{F}_n$.

How can we assign a convergent predictive rule?

How can we get a link to parametric models?

How can we approximate inference, not having the (explicit) prior?

**How can we assign a convergent predictive rule?**

How can we get a link to parametric models?

How can we approximate inference, not having the (explicit) prior?

# Martingale predictive, or c.i.d. sequences

* A sufficient condition for $P_n \to \tilde{F}$ is that $(P_n)_{n\geq 0}$ is a measure-valued martingale; equivalently, $X_{n+k} \mid X_{1:n} \stackrel{d}{=} X_{n+1} \mid X_{1:n} \quad k \geq 1$, i.e. $(X_n)_{n\geq 1}$ **is c.i.d.**(Berti et al, Ann. Prob.2004)

* Thrm.(Kallenberg, 1998) $(X_n)$ *exchangeable if and only if it is stationary and* $(P_n)$ *is a martingale.*
(we break stationarity: time - the order - matters, in the initial stage)

# Turning algorithms into Bayesian predictive rules

There are many *predictive algorithms*, that we can read as Bayesian predictive learning rules!

A 'statistical' example. Consider the popular DP mixture model

$$X_t \mid G \overset{iid}{\sim} f_G(x) = \int k(x \mid \theta) dG(\theta), \quad G \sim\sim DP(\alpha, G_0).$$

Suppose data arrive sequentially, and interest is in estimating the mixing distribution $G$, updating the estimate as a new $x_t$ becomes available. Computations; MCMC no! sequential MC, or sequential VB...?

# Turning algorithms into Bayesian predictive rules

There are many *predictive algorithms*, that we can read as Bayesian predictive learning rules!

A 'statistical' example. Consider the popular DP mixture model

$$X_t \mid G \overset{iid}{\sim} f_G(x) = \int k(x \mid \theta) dG(\theta), \quad G \sim\sim DP(\alpha, G_0).$$

Suppose data arrive sequentially, and interest is in estimating the mixing distribution $G$, updating the estimate as a new $x_t$ becomes available. Computations; MCMC no! sequential MC, or sequential VB...?

$\rightarrow$ "A recursive algorithm": start from $G_0(\theta)$ at $t = 0$ and for $t \geq 1$ recursively update

$$G_t(\theta) = (1 - \alpha_n) G_{t-1}(\theta) + \alpha_n G_{t-1}(\theta \mid x_t).$$

At step $t = n$, $G_n(\theta)$ is the proposed estimate of $G(\theta)$. ("Newton's algorithm" in the BNP literature).

# Turning algorithms into Bayesian predictive rules

There are many *predictive algorithms*, that we can read as Bayesian predictive learning rules!

A 'statistical' example. Consider the popular DP mixture model

$$X_t \mid G \overset{iid}{\sim} f_G(x) = \int k(x \mid \theta) dG(\theta), \quad G \sim\sim DP(\alpha, G_0).$$

Suppose data arrive sequentially, and interest is in estimating the mixing distribution $G$, updating the estimate as a new $x_t$ becomes available. Computations; MCMC no! sequential MC, or sequential VB...?

$\rightarrow$ "A recursive algorithm": start from $G_0(\theta)$ at $t = 0$ and for $t \geq 1$ recursively update

$$G_t(\theta) = (1 - \alpha_n) G_{t-1}(\theta) + \alpha_n G_{t-1}(\theta \mid x_t).$$

At step $t = n$, $G_n(\theta)$ is the proposed estimate of $G(\theta)$. ("Newton's algorithm" in the BNP literature).
But, **uncertainty around $G_n$? In fact, is this algorithm any Bayesian?**

# A Bayesian predictive rule

By taking a predictive approach, we (Fortini & P. (2020), JRSS,B) can read it as a Bayesian predictive learning rule,

$$X_{n+1} \mid x_{1:n} \sim P_n = F_{G_n}(x) = \int K(x \mid \theta) dG_n(\theta)$$

$$= (1 - \alpha_n) P_{n-1} + \alpha_n \int K(\cdot \mid \theta) dG_{n-1}(\theta \mid x_n).$$

# A Bayesian predictive rule

By taking a predictive approach, we (Fortini & P. (2020), JRSS,B) can read it as a Bayesian predictive learning rule,

$$X_{n+1} \mid x_{1:n} \sim P_n = F_{G_n}(x) = \int K(x \mid \theta) dG_n(\theta)$$

$$= (1-\alpha_n) P_{n-1} + \alpha_n \int K(\cdot \mid \theta) dG_{n-1}(\theta \mid x_n).$$

Then we show that $(P_n)_{n \geq 0}$ is a martingale, and $P_n$ converges to $\tilde{F} = F_{\tilde{G}} = \int K(\cdot \mid \theta d\tilde{G}(\theta)$, where $\tilde{G} = \lim G_n$.

**Thus, the algorithm is using an asymptotically exchangeable Bayesian mixture model**; for $n > N$ large

$$X_i \mid G \overset{iid}{\approx} f_G(x) = \int k(x \mid \theta) dG(\theta)$$

with an implicit prior on $\tilde{G}$.

# A Bayesian predictive rule

By taking a predictive approach, we (Fortini & P. (2020), JRSS,B) can read it as a Bayesian predictive learning rule,

$$X_{n+1} \mid x_{1:n} \sim P_n = F_{G_n}(x) = \int K(x \mid \theta) dG_n(\theta)$$

$$= (1-\alpha_n) P_{n-1} + \alpha_n \int K(\cdot \mid \theta) dG_{n-1}(\theta \mid x_n).$$

Then we show that $(P_n)_{n \geq 0}$ is a martingale, and $P_n$ converges to $\tilde{F} = F_{\tilde{G}} = \int K(\cdot \mid \theta d\tilde{G}(\theta)$, where $\tilde{G} = \lim G_n$.

**Thus, the algorithm is using an asymptotically exchangeable Bayesian mixture model**; for $n > N$ large

$$X_i \mid G \overset{iid}{\approx} f_G(x) = \int k(x \mid \theta) dG(\theta)$$

with an implicit prior on $\tilde{G}$.

$\rightarrow$ **The prior, and the posterior distribution of** $\tilde{G}$ are not explicit, but we will approximate them!

# Example: online gradient descent

$\rightarrow$ Classify items that arrive sequentially: $(X_n, Y_n)$, where $X_n$ are features and $Y_n$ is type, $0-1$. One models

$$P(Y_n = 1 \mid x_n) = g(x_n, \beta)$$

for a known, but possibly complex, $g$, e.g. neural network, and an unknown vector $\beta$.

Given a training sample $(x_1, y_1), \ldots, (x_n, y_n)$, it is common to estimate $\beta$ by minimizing a loss function $L(\beta; x_{1:n}, y_{1:n})$ measuring the discrepancy between the actual values $y_1, \ldots, y_n$ and the ones predicted by the model.

A popular recursive algorithm is the online gradient descent that is initialized at $\beta_0$ and for $n = 1, 2, \ldots$

$$\beta_n = \beta_{n-1} - \frac{1}{n} \nabla_\beta L(\beta_{n-1}; x_n, y_n).$$

E.g., with the binary cross entropy loss, and logistic function $g(x, \beta)$,

$$\beta_n = \beta_{n-1} + \frac{1}{n \log 2}(y_n - g(x_n, \beta_{n-1}))x_n.$$

# A Bayesian predictive rule!

We can read it as a Bayesian predictive learning rule for the sequence $((X_n, Y_n))_{n\geq 1}$, by taking $X_i \overset{iid}{\sim} p_x$ and letting

$$(X_{n+1}, Y_{n+1}) \mid x_{1:n}, y_{1:n} \sim P_n(x, y)$$

where $P_n$ is such that

$$P(Y_{n+1} = 1 \mid x_{1:n}, y_{1:n}, x_{n+1}) = g(x_{n+1}, \beta_n).$$

Then under general conditions $\beta_n \to \tilde{\beta}$, random, and $P_n(x, y)$ converges to a random $\tilde{F}(x, y)$, such that, for $n$ large

$$Y_n \mid x_n, \tilde{\beta} \overset{ind}{\approx} Bernoulli(g(x_{n+1}, \tilde{\beta})),$$

with an implicit prior on $\tilde{\beta}$. Moreover

$$\beta_n = E(\tilde{\beta} \mid x_{1:n}, y_{1:n}).$$

# inference on $\tilde{\beta}$

We can then make Bayesian inference on $\tilde{\beta}$ ("without the prior") with an implicit prior on $\tilde{\beta}$.
BUT, how can we obtain the implied posterior distribution?

Not by predictive resampling, as we's need to sample pairs $(x_{n+1}, y_{n+1}), \ldots$, thus need the distribution of the $X_i$!

$\rightarrow$ Provide asymptotic approximation of the implicit posterior of $\tilde{\beta}$

We prove that, under regularity conditions,

$$\tilde{\beta} \mid x_{1:n}, y_{1:n} \approx \mathcal{N}_d(\beta_n, \frac{V_n}{n})$$

where

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

does not depend on $P_X$; and asymptotic credible intervals for $\tilde{\beta}$.

How can we assign a convergent predictive rule?

**How can we get a link to parametric models?**

How can we approximate inference, not having the (explicit) prior?

# Link with parametric models

In the examples, the model $\tilde{F} = \lim P_n$ implied by the algorithm (the predictive rule) is semi-parametric (in the mixture example, $\tilde{F} = F_{\tilde{G}} = \int K(\cdot \mid \theta d\tilde{G}(\theta)$, the parameter is $\tilde{G}$), or parametric, $\tilde{F} = F_{\tilde{\beta}}$.

**In the general case, $P_n \to \tilde{F}$, and the asymptotic model is $\tilde{F}$. Parametric if $P_n \to F_{\tilde{\theta}}$.**

# Link with parametric models

In the examples, the model $\tilde{F} = \lim P_n$ implied by the algorithm (the predictive rule) is semi-parametric (in the mixture example, $\tilde{F} = F_{\tilde{G}} = \int K(\cdot \mid \theta d\tilde{G}(\theta)$, the parameter is $\tilde{G}$), or parametric, $\tilde{F} = F_{\tilde{\beta}}$.

**In the general case, $P_n \to \tilde{F}$, and the asymptotic model is $\tilde{F}$. Parametric if $P_n \to F_{\tilde{\theta}}$.**

In parametric models $F_\theta(x)$, under exchangeability, $\tilde{\theta}$ typically is the limit of a predictive sufficient statistic, which inspires the specification

$$P_n(x_{n+1} \mid X_1; \ldots, X_n) = F(x_{n+1} \mid T_n(X_1, \ldots, X_n)) \equiv F_{T_n}(x_{n+1}),$$

where $T_n$ is computed recursively as a function of $T_{n-1}$ and $x_n$; in particular,

$$T_n = T_{n-1} + \alpha_n h(T_{n-1}, X_n).$$

# Link with parametric models

In the examples, the model $\tilde{F} = \lim P_n$ implied by the algorithm (the predictive rule) is semi-parametric (in the mixture example, $\tilde{F} = F_{\tilde{G}} = \int K(\cdot \mid \theta) d\tilde{G}(\theta)$, the parameter is $\tilde{G}$), or parametric, $\tilde{F} = F_{\tilde{\beta}}$.

**In the general case, $P_n \to \tilde{F}$, and the asymptotic model is $\tilde{F}$. Parametric if $P_n \to F_{\tilde{\theta}}$.**

In parametric models $F_\theta(x)$, under exchangeability, $\tilde{\theta}$ typically is the limit of a predictive sufficient statistic, which inspires the specification

$$P_n(x_{n+1} \mid X_1; \ldots, X_n) = F(x_{n+1} \mid T_n(X_1, \ldots, X_n)) \equiv F_{T_n}(x_{n+1}),$$

where $T_n$ is computed recursively as a function of $T_{n-1}$ and $x_n$; in particular,

$$T_n = T_{n-1} + \alpha_n h(T_{n-1}, X_n).$$

Under conditions, $T_n \to \tilde{\theta}$ and $P_n = F_{T_n} \to F_{\tilde{\theta}}$. Thus, asymptotically,

$$X_n \mid \tilde{\theta} \overset{d}{\approx} F_{\tilde{\theta}} \quad \text{where } \tilde{\theta} = \lim T_n \text{ has an implicit prior law.}$$

# Link with parametric models

In the examples, the model $\tilde{F} = \lim P_n$ implied by the algorithm (the predictive rule) is semi-parametric (in the mixture example, $\tilde{F} = F_{\tilde{G}} = \int K(\cdot \mid \theta) d\tilde{G}(\theta)$, the parameter is $\tilde{G}$), or parametric, $\tilde{F} = F_{\tilde{\beta}}$.

**In the general case, $P_n \to \tilde{F}$, and the asymptotic model is $\tilde{F}$. Parametric if $P_n \to F_{\tilde{\theta}}$.**

In parametric models $F_\theta(x)$, under exchangeability, $\tilde{\theta}$ typically is the limit of a predictive sufficient statistic, which inspires the specification

$$P_n(x_{n+1} \mid X_1; \ldots, X_n) = F(x_{n+1} \mid T_n(X_1, \ldots, X_n)) \equiv F_{T_n}(x_{n+1}),$$

where $T_n$ is computed recursively as a function of $T_{n-1}$ and $x_n$; in particular,

$$T_n = T_{n-1} + \alpha_n h(T_{n-1}, X_n).$$

Under conditions, $T_n \to \tilde{\theta}$ and $P_n = F_{T_n} \to F_{\tilde{\theta}}$. Thus, asymptotically,

$$X_n \mid \tilde{\theta} \overset{d}{\approx} F_{\tilde{\theta}} \quad \text{where } \tilde{\theta} = \lim T_n \text{ has an implicit prior law.}$$

* This is related to the *plug-in predictive* (Walker, 2022; see Fong & Yiu, 2024+) where $T_n = \hat{\theta}_n$, e.g. MLE.

How can we assign a convergent predictive rule?

How can we get a link to parametric models?

**How can we approximate inference, not having the (explicit) prior?**

# posterior approximation

The prior remains implicit. How can we approximate the posterior distribution?

- Sample from it! predictive Monte Carlo, or predictive resampling, (Fortini & P., JRSS, 2020; Fong, Holmes & Walker, JRSS, B, 2023).

- Gaussian asymptotic approximations of the posterior distribution of $\theta$, in parametric cases; or of the posterior distribution of $F(t)$, for a fixed $t$ or on a grid, or of the entire process $\tilde{F}$, in the general case.

The latter are not BvM results; rather refine Doob's theorem ("Doob-BvM").

# Predictive Monte Carlo

**Suppose $P_n \to \tilde{F}$, and interest is in inference on $F(t)$, or in a functional $\tilde{\theta} = \theta(\tilde{F})$, e.g. $\int x d\tilde{F}(x)$.**

We can design a *predictive Monte Carlo*, or *predictive resampling* algorithm to sample from their prior and posterior distributions.[3]

[3] Fortini & P., *JRSS, B*, (2020); Fong, Holmes, Walker, *JRSS, B*, (2023)

# Predictive Monte Carlo

**Suppose $P_n \to \tilde{F}$, and interest is in inference on $F(t)$, or in a functional $\tilde{\theta} = \theta(\tilde{F})$, e.g. $\int x d\tilde{F}(x)$.**

We can design a *predictive Monte Carlo*, or *predictive resampling* algorithm to sample from their prior and posterior distributions.[3]

**Sampling from the posterior**

- Given data $x_{1:n}$, use the predictive rule to generate $(x_{n+1}, x_{n+2}, \ldots)$ truncated at a large $N$: **get** $x_{1:N}$.

[3] Fortini & P., *JRSS, B*, (2020); Fong, Holmes, Walker, *JRSS, B*, (2023)

# Predictive Monte Carlo

**Suppose $P_n \to \tilde{F}$, and interest is in inference on $F(t)$, or in a functional $\tilde{\theta} = \theta(\tilde{F})$, e.g. $\int x d\tilde{F}(x)$.**

We can design a *predictive Monte Carlo*, or *predictive resampling* algorithm to sample from their prior and posterior distributions.[3]

**Sampling from the posterior**

- Given data $x_{1:n}$, use the predictive rule to generate $(x_{n+1}, x_{n+2}, \ldots)$ truncated at a large $N$: **get** $x_{1:N}$.
- Compute the empirical $\hat{F}_N(t) = \sum_{i=1}^{N} \delta_{x_i}(t)$, or the pred $P_N(t \mid x_{1:N})$.
  Because $P_n \to \tilde{F}$, and $N$ is large, $P_N(t)$ approximates $\tilde{F}(t)$,
  a sample from the posterior of $\tilde{F}(t)$.

[3] Fortini & P., *JRSS, B*, (2020); Fong, Holmes, Walker, *JRSS, B*, (2023)

# Predictive Monte Carlo

**Suppose $P_n \to \tilde{F}$, and interest is in inference on $F(t)$, or in a functional $\tilde{\theta} = \theta(\tilde{F})$, e.g. $\int x d\tilde{F}(x)$.**

We can design a *predictive Monte Carlo*, or *predictive resampling* algorithm to sample from their prior and posterior distributions.[3]

**Sampling from the posterior**

- Given data $x_{1:n}$, use the predictive rule to generate $(x_{n+1}, x_{n+2}, \ldots)$ truncated at a large $N$: **get** $x_{1:N}$.
- Compute the empirical $\hat{F}_N(t) = \sum_{i=1}^{N} \delta_{x_i}(t)$, or the pred $P_N(t \mid x_{1:N})$.
  Because $P_n \to \tilde{F}$, and $N$ is large, $P_N(t)$ approximates $\tilde{F}(t)$,
  a sample from the posterior of $\tilde{F}(t)$.
- Repeat $M$ times: (approx) Monte Carlo sample of size $M$ from the posterior.

[3] Fortini & P., *JRSS, B*, (2020); Fong, Holmes, Walker, *JRSS, B*, (2023)

# Predictive Monte Carlo

**Suppose $P_n \to \tilde{F}$, and interest is in inference on $F(t)$, or in a functional $\tilde{\theta} = \theta(\tilde{F})$, e.g. $\int x d\tilde{F}(x)$.**

We can design a *predictive Monte Carlo*, or *predictive resampling* algorithm to sample from their prior and posterior distributions.[3]

**Sampling from the posterior**

- Given data $x_{1:n}$, use the predictive rule to generate $(x_{n+1}, x_{n+2}, \ldots)$ truncated at a large $N$: **get** $x_{1:N}$.
- Compute the empirical $\hat{F}_N(t) = \sum_{i=1}^{N} \delta_{x_i}(t)$, or the pred $P_N(t \mid x_{1:N})$.
  Because $P_n \to \tilde{F}$, and $N$ is large, $P_N(t)$ approximates $\tilde{F}(t)$,
  a sample from the posterior of $\tilde{F}(t)$.
- Repeat $M$ times: (approx) Monte Carlo sample of size $M$ from the posterior.

- If interest in $\tilde{\theta} = \theta(\tilde{F})$: compute the empirical $\theta(\hat{F}_N)$ at each iteration, to get an approx Monte Carlo sample from the posterior dist. of $\tilde{\theta}$.

**If $n = 0$ : sampling from the prior.**

[3] Fortini & P., *JRSS, B*, (2020); Fong, Holmes, Walker, *JRSS, B*, (2023)

# Asymptotic Gaussian approximations

Predictive Monte Carlo does not require MCMC!
But also interesting to have analytic approximations – in some cases, better! e.g., in the logistic regression ex: we should sample pairs $(X_n, Y_n)$ but do not have $P_X$.

# Asymptotic Gaussian approximations

Predictive Monte Carlo does not require MCMC!
But also interesting to have analytic approximations – in some cases, better! e.g., in the logistic regression ex: we should sample pairs $(X_n, Y_n)$ but do not have $P_X$.

Let's consider the general case, $P_n \to \tilde{F}$ thus for $n$ large

$$X_n \mid \tilde{F} \overset{iid}{\approx} \tilde{F}.$$

Here I only consider the case where $(P_n)_{n\geq 0}$ is a martingale.
Interest in the **posterior distribution of $\tilde{F}(t)$ at a fixed $t$.**

# Asymptotic Gaussian approximations

Predictive Monte Carlo does not require MCMC!
But also interesting to have analytic approximations – in some cases, better! e.g., in the logistic regression ex: we should sample pairs $(X_n, Y_n)$ but do not have $P_X$.

Let's consider the general case, $P_n \to \tilde{F}$ thus for $n$ large

$$X_n \mid \tilde{F} \overset{iid}{\approx} \tilde{F}.$$

Here I only consider the case where $(P_n)_{n\geq 0}$ is a martingale.
Interest in the **posterior distribution of $\tilde{F}(t)$ at a fixed $t$.**

Remember that $\tilde{F} = \lim P_n$. Given data $x_{1:n}$, we are uncertain about the limit $\tilde{F}$ of $P_n$. uncertainty formalized in the posterior distribution of $\tilde{F}$ is such an uncertainty, 'due to the missing obs", and depends on the behavior of $P_n$.
Intuitively, if $P_n \to \tilde{F}$ at a fast rate, then for finite $n$ we will be fairly sure about its limit, reflected in a small posterior variance.

# Asymptotic Gaussian approximations

Predictive Monte Carlo does not require MCMC!
But also interesting to have analytic approximations – in some cases, better! e.g., in the logistic regression ex: we should sample pairs $(X_n, Y_n)$ but do not have $P_X$.

Let's consider the general case, $P_n \to \tilde{F}$ thus for $n$ large

$$X_n \mid \tilde{F} \overset{iid}{\approx} \tilde{F}.$$

Here I only consider the case where $(P_n)_{n\geq 0}$ is a martingale.
Interest in the **posterior distribution of $\tilde{F}(t)$ at a fixed $t$.**

Remember that $\tilde{F} = \lim P_n$. Given data $x_{1:n}$, we are uncertain about the limit $\tilde{F}$ of $P_n$. uncertainty formalized in the posterior distribution of $\tilde{F}$ is such an uncertainty, 'due to the missing obs", and depends on the behavior of $P_n$.
Intuitively, if $P_n \to \tilde{F}$ at a fast rate, then for finite $n$ we will be fairly sure about its limit, reflected in a small posterior variance.

# From the predictive to the posterior distr.

Consider the predictive updates

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

how the predictive distribution at $t$ varies in response to the new observation $x_n$. Let

$$V_{t,n} = \frac{1}{n}\sum_{k=1}^{n} k^2 {\Delta_{t,k}}^2.$$

**Thrm**[4] *Under regularity conditions,*

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

*for $\mathbb{P}$-almost all $\omega = (x_1, x_2, \ldots)$.*
The approximation is centered on $P_n(t) = E(\tilde{F}(t) \mid x_{1:n})$.
The asymptotic variance depends on the way the predictive distribution learns from the data.

[4] Fortini & P., *Phil. Trans. Roy. Soc.*, 2023; Fortini & P., *Statistical Science*, 2025

# Predictive efficiency?

We can obtain asymptotic credible intervals for $\tilde{F}(t)$ given $x_{1:n}$

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

The size of the credible interval depends on the **convergence rate** of $P_n$.

# Predictive efficiency?

We can obtain asymptotic credible intervals for $\tilde{F}(t)$ given $x_{1:n}$

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

The size of the credible interval depends on the **convergence rate** of $P_n$.

**Yet, is $P_n$ "learning well?** is it "**efficient**"?

It has to balance a convergence rate with a proper **learning rate** (Ex, if $P_n = P_0$ for any $n$, it converges immediately, but does not learn from the data!)

## example

Example: $\tilde{F} \sim DP(\alpha P_0)$; then

$$P_n(t) = \frac{\alpha}{\alpha + n} P_0(t) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{x_i}(t)$$

## example

Example: $\tilde{F} \sim DP(\alpha P_0)$; then

$$P_n(t) = \frac{\alpha}{\alpha + n} P_0(t) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{x_i}(t) = P_{n-1}(t) + \frac{1}{\alpha + n}(\delta_{x_n}(t) - P_{n-1}(t)),$$

and

$$\Delta_{t,n}(t) = \frac{1}{\alpha + n}[\delta_{x_n}(t) - P_{n-1}(t)]$$

depends on $\alpha$: if $\alpha$ large, $\Delta_{t,n}$ is small: given $x_{1:n-1}$, we do not learn much from the new observation $x_n$ in predicting $x_{n+1}$

# example

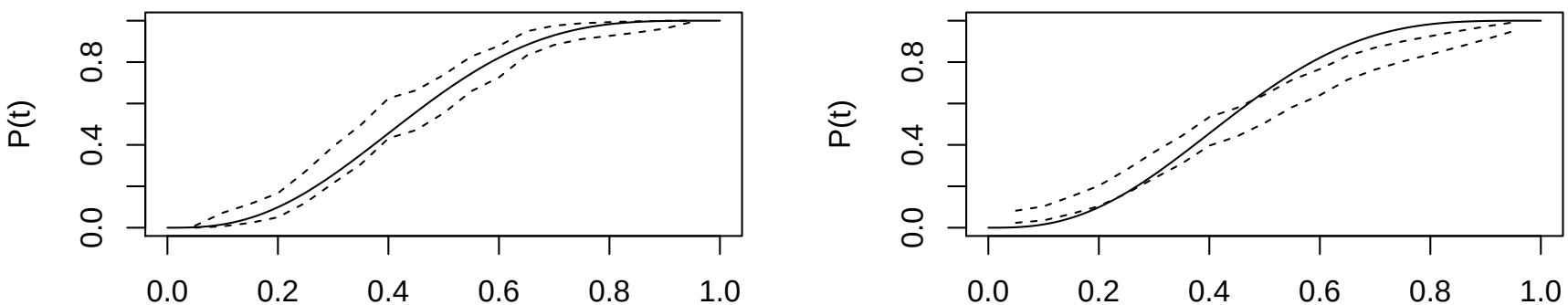
Example: $\tilde{F} \sim DP(\alpha P_0)$; then

$$P_n(t) = \frac{\alpha}{\alpha+n} P_0(t) + \frac{1}{\alpha+n} \sum_{i=1}^{n} \delta_{x_i}(t) = P_{n-1}(t) + \frac{1}{\alpha+n}(\delta_{x_n}(t) - P_{n-1}(t)),$$

and

$$\Delta_{t,n}(t) = \frac{1}{\alpha+n}[\delta_{x_n}(t) - P_{n-1}(t)]$$

depends on $\alpha$: if $\alpha$ large, $\Delta_{t,n}$ is small: given $x_{1:n-1}$, we do not learn much from the new observation $x_n$ in predicting $x_{n+1}$




Marginal 0.95 credible intervals for $\tilde{F}(t)$ for $t$ on a grid: $\alpha = 1$ (left panel) and $\alpha = 100$ (right). Solid curve: True $F$.

# Predictive inferential-efficiency?

This calls for a notion of **predictive efficiency**..

In frequentist stats: inferential efficiency: (asymptotic) variance of unbiased efficient estimators in terms of Fisher information..

Is there a notion of *efficiency* in Bayesian Stats?
- rather, loss function and optimality..
- scoring rules, calibration..

- Here: IF the $X_i$ are indeed iid from $F_{true}$, is the predictive rule able to learn that, "efficiently"? giving good frequentist coverage of the implied credible intervals?

# Predictive inferential-efficiency?

This calls for a notion of **predictive efficiency**..

In frequentist stats: inferential efficiency: (asymptotic) variance of unbiased efficient estimators in terms of Fisher information..

Is there a notion of *efficiency* in Bayesian Stats?
- rather, loss function and optimality..
- scoring rules, calibration..

- Here: IF the $X_i$ are indeed iid from $F_{true}$, is the predictive rule able to learn that, "efficiently"? giving good frequentist coverage of the implied credible intervals?

The latter is usually studied through BvM results. Ours are not BvM results: they hold with prob one, moreover the asymptotic variance is expressed in terms of the predictive updates, not of Fisher information.

Yet, in a predictive approach, they may suggest conditions on the predictive rule – on its *learning rate* – that ensure *inferential efficiency* and good frequentist coverage.

# Some preliminary results

Consider the 'parametric predictive' rule

$$P_n(x_{n+1} \mid X_1; \ldots, X_n) = F(x_{n+1} \mid T_n(1, \ldots, X_n)) \equiv F_{T_n}(x_{n+1},$$

with

$$T_n = T_{n-1} + \alpha_n \, h(T_{n-1}, X_n)$$

Under conditions, $T_n \to \tilde{\theta}$ and $P_n = F_{T_n} \to F_{\tilde{\theta}}$, thus, asymptotically, $X_n \mid \tilde{\theta} \overset{iid}{\approx} F_{\tilde{\theta}}$ where $\tilde{\theta} = \lim T_n$, and has an implicit prior.

# Some preliminary results

Consider the 'parametric predictive' rule

$$P_n(x_{n+1} \mid X_1; \ldots, X_n) = F(x_{n+1} \mid T_n(1, \ldots, X_n)) \equiv F_{T_n}(x_{n+1},$$

with

$$T_n = T_{n-1} + \alpha_n \, h(T_{n-1}, X_n)$$

Under conditions, $T_n \to \tilde{\theta}$ and $P_n = F_{T_n} \to F_{\tilde{\theta}}$, thus, asymptotically, $X_n \mid \tilde{\theta} \overset{iid}{\approx} F_{\tilde{\theta}}$ where $\tilde{\theta} = \lim T_n$, and has an implicit prior.

Conditions on the updating:
* $\alpha_n$ positive decreasing with $\sum_{n=1}^{\infty} \alpha_n = \infty$ and $\sum_{k \geq n} \alpha_k^2 < \infty$, and
* $h$ satisfies $E(h(\theta, X)) = 0$ and $E(h_j(\theta, X)^2) < \infty$ where $X \sim F_\theta$.

Then $(T_n)_{n \geq 0}$ is a uniformly integrable martingale converging to $\tilde{\theta}$.

# posterior distribution of $\tilde{\theta}$

By the a.s. conditional CLT for martingales[5], we can show that, for $\mathbb{P}$-almost all $(x_1, x_2, \ldots)$,

$$\sqrt{b_n}(\tilde{\theta} - T_n) \mid x_{1:n} \to \mathcal{N}(0, U_{\tilde{\theta}(x_{1:\infty})})$$

where $b_n = (\sum_{k \geq n} \alpha_k^2)^{-1}$ and

$$\begin{aligned} U_{\tilde{\theta}} &= \lim E(h(T_n, X_n)h(T_n, X_n)^T \mid X_1, \ldots, X_{n-1}) \\ &= \mathrm{Var}(h(\tilde{\theta}, X) \mid \tilde{\theta}) \end{aligned}$$

Thus, for $n$ large,

$$\tilde{\theta} \mid x_{1:n} \approx \mathcal{N}(T_n, \frac{U_{\tilde{\theta}}}{b_n}),$$

[5] Crimaldi, 2009

Thus, for $n$ large,

$$\tilde{\theta} \mid x_{1:n} \approx \mathcal{N}(T_n, \frac{U_{\tilde{\theta}}}{b_n}).$$

* The rate, $b_n^{-1} = \sum_{k \geq n} \alpha_k^2$, is generally slower than $1/n$, unless $\alpha_n \sim 1/n$.

* The asymptotic posterior variance
$U_\theta = \lim E(h(T_n, X_n)h(T_n, X_n)^T \mid X_1, \ldots, X_{n-1})$, that again depends on the predictive updates, correspond to the inverse of Fisher information $I_n(\theta)$ for the model $p_\theta$ if the 'loss' $h$ is suitably chosen.

Thus, for $n$ large,

$$\tilde{\theta} \mid x_{1:n} \approx \mathcal{N}(T_n, \frac{U_{\tilde{\theta}}}{b_n}).$$

* The rate, $b_n^{-1} = \sum_{k \geq n} \alpha_k^2$, is generally slower than $1/n$, unless $\alpha_n \sim 1/n$.

* The asymptotic posterior variance
$U_\theta = \lim E(h(T_n, X_n)h(T_n, X_n)^T \mid X_1, \ldots, X_{n-1})$, that again depends on the predictive updates, correspond to the inverse of Fisher information $I_n(\theta)$ for the model $p_\theta$ if the 'loss' $h$ is suitably chosen.

$\rightarrow$ An idea for such a "predictive inferential-efficiency" is to use *score-adjusted predictive rules* with

$$T_n = T_{n-1} + \alpha_n \, I(T_{n-1})^{-1} \, s(T_{n-1}, x_n)$$

where $s(\theta, X) = \partial \, log p_\theta(x)/\partial\theta$ is the score function, (elaborating from Walker (2022), Holmes & Walker (2023), Wang & Holmes (2024); Fong & Yiu (2024+).
E.g., this suggests to **improve the online gradient descent updating** in the logistic example by including $I(T_{n-1})^{-1}$.

# Frequentist coverage?

Again, the asymptotic approximations still hold if the random variance $U_{\tilde{\theta}}$ is replaced by 'an estimate' (e.g. $U_{T_n}$) that converges to $U_{\tilde{\theta}}$, allowing to obtain predictive-based asymptotic credible intervals for $\tilde{\theta}$ –
-+ that would have good frequentist coverage (very informally: "for almost all $\theta$")

...BUT not BvM yet; rather "Doob-BvM", with conditions on the predictive updates
(some recent results by Fong & Yiu, 2024+, and by Fortini & P. - ongoing )

# Multivariate extensions

For short, the previous results were shown for the univariate case. For martingale predictive rules $P_n \to \tilde{F}$, so that

$$X_i \mid \tilde{F} \overset{iid}{\approx} \tilde{F} \quad \text{for } n \text{ large.}$$

we can approximate the posterior distribution of $\tilde{F}(t)$ for a fixed $t$

$\to$ The results extend to the posterior distribution of $\tilde{F}(t)$ for $t$ on a grid, and to the entire distribution $\tilde{F}$.
The latter may allow to approximate the posterior distribution of functionals of $\tilde{F}$ (e.g., $\mu = \int x d\tilde{F}(x)$, quantiles $\tilde{F}^{-1}(p)$, ...)

# predictive inference on $[\tilde{F}(t_1), \ldots, \tilde{F}(t_k)]$

Consider a grid $t_{1:k}$ with $\mathbb{P}(X_1 \in \{t_1, \ldots, t_k\}) = 0$, and the column vector of predictive updates $\mathbf{\Delta}_n = [\Delta_{t_1,n}, \ldots, \Delta_{t_k,n}]^T$.

**Proposition**

If there is a sequence $(\alpha)_n$ with $\alpha_n > 0$, $\sum_{n=1}^{\infty} \alpha_n = \infty$ and $\sum_{k \geq n} \alpha_k^2 < \infty$ such that

- $E(\sup_n \alpha_n^{-1/2} \mid \Delta_{t_j,n}|) < \infty$
- $\sum_{n=1}^{\infty} \alpha_n^{-2} E(\Delta_{t_j,n}^4) < \infty$, $j = 1, \ldots, k$
- $E(\alpha_n^{-2} \mathbf{\Delta}_{\mathbf{t},n} \mathbf{\Delta}_{\mathbf{t},n}^T \mid X_1, \ldots, X_{n-1}) \to \tilde{\mathbf{U}}_{t_{1:k}}$, $\mathbb{P}$-a.s., for a positive definite random matrix $\tilde{\mathbf{U}}_{t_{1:k}}$,

then, $\mathbb{P}$-a.s.

$$\sqrt{b_n} \begin{bmatrix} \tilde{F}(t_1) - P_n(t_1) \\ \vdots \\ \tilde{F}(t_k) - P_n(t_k) \end{bmatrix} | X_1, \ldots, X_n \xrightarrow{d} \mathcal{N}_k(0, \tilde{\mathbf{U}}_{t_{1:k}})$$

where $b_n = (\sum_{k \geq n} \alpha_k^2)^{-1}$.

# inference on $\tilde{F}$

## Theorem

*Suppose the same conditions hold for any* $(t_1, \ldots, t_k)$.
*If there exists* $n_0$ *and a non-decreasing random element* $\tilde{H}$ *of* $D$ *(space of cadlag functions, with Skorohod topology) such that for every* $n \geq n_0$,

$$E(\alpha_n^{-2}(\Delta_{t,n} - \Delta_{s,n})^2 \mid X_1, \ldots, X_n) \leq \tilde{H}(t) - \tilde{H}(s) \quad \mathbb{P}\text{-a.s.}$$

*for every* $s < t$, *then,* $\mathbb{P}$*-a.s.,*

$$\sqrt{b_n}\,(P_n - \tilde{F}) \mid X_1, \ldots, X_n \xrightarrow{d} \mathbb{G}(U),$$

*where* $b_n = (\sum_{k \geq n} \alpha_k^2)^{-1}$, *and* $\mathbb{G}(U)$ *is a centered Gaussian process with kernel* $U$ *satisfying* $U(t_i, t_j) = \mathbf{U}_{(t_i,t_j)}[i,j]$. *and* $U(t,t) = \mathbf{U}_t$.

# Time dependence

Breaking exchangeability, order matters. In the DP mixture example, Newton's algorithm – as a predictive rule – is implicitly assuming a time-dependent mixture model

$$X_i \mid \tilde{G}_n \overset{iid}{\sim} \int K(\cdot \mid \theta) d\tilde{G}_n(\theta)$$

with a specific temporal evolution of the random $(G_n)$; in particular, $E(G_n(\cdot) = G_0(\cdot)$ and $\tilde{G}_n$ converges to a random $\tilde{G}$; (details in Fortini & P. *JRSS, B* 2020).

This may be an interesting model when one actually has time. In a static setting, it is an asymptotic approximation of an exchangeable mixture model with $\tilde{G}$ constant – intuitively, the closer to it the 'more stable' the $\tilde{G}_n$ are.

* Again, the *learning coefficients* $\alpha_n$ have a crucial dual role in
– driving the rate of convergence of $\tilde{G}_n$ to the limit $\tilde{G}$ (we'd like it to be fast, to quickly reach exchangeability)
– driving the learning rate (we'd like to be fairly slow, to actually learn from the observations as they become available).
* The role of the *loss function* is unexplored in this semi-parametric setting.

# Beyond random sampling

I have here considered the basic setting, random sampling – exchangeability.

Extensions to more elaborated sampling schemes and settings - e.g., fixed-design regression and partial exchangeability, and time series ...- are only partially developed.

A direction for extensions to partially exchangeable structures and possibly Markov chains is to move from the notion of *partially c.i.d.* arrays (Fortini, P. & Sporysheva, *Bernoulli*, 2016).

# Final remarks & open problems

- We can take a Bayesian predictive approach to deal with (some classes of) recursive predictive algorithms
- Is the predictive rule implicitly using an inferential scheme?
  Then we can provide Bayesian uncertainty quantification (without the (explicit) prior!)
- Inferential properties depends on the capacity of the predictive distribution of learning from the data.
  $\rightarrow$ "predictive efficiency"?
  $\rightarrow$ Frequentist properties?
  $\rightarrow$ More genuinely predictive criteria? Calibration and scoring rules?
- more on extensions beyond random sampling (asymptotic partial exchangeability, ...)
- $\vdots$

**Thank you for attending this lecture!**

# references


- Fortini, S. and Petrone, S. (2020) *JRSS, B.*
- Fortini, S. and Petrone, S. (2023) Prediction-based uncertainty quantification for exchangeable sequences. *Phil. Trans. Royal Soc. A*, 381:20220142.
- **Fortini, S. and Petrone, S. (2025) Exchangeability, prediction and predictive modeling in Bayesian statistics. Statistical Science**