Probabilistic morphisms, stochastic processes, and Bayesian supervised learning

Hông Vân Lê
Institute of Mathematics, Czech Academy of Sciences

Prague, December 17, 2025

OUTLINE

1. Supervised (machine) learning: frequentist vs Bayesian.

2. Markov kernel, probabilistic morphisms, and Bayesian supervised learning.

3. Bayesian batch learning vs online learning.

4. Gaussian process regression and Kalman filter.

5. Mathematical machine learning.

1 Supervised (machine) learning: frequentist vs Bayesian.

Supervised Bayesian Inference (SBI) Problem]
$\mathcal{X}$ - input space, $\mathcal{Y}$ - measurable label space.
$S_n := \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (\mathcal{X} \times \mathcal{Y})^n$ - training data, $T_m := (t_1, \ldots, t_m) \in \mathcal{X}^m$ - test data.

Problem: estimate the predictive probability measure $\mathcal{P}_{T_m|S_n} \in \mathcal{P}(\mathcal{Y}^m)$ that governs the joint distribution $\big(y'_1, \ldots, y'_m\big) \in \mathcal{Y}^m$ where y'_i is the label of t_i.

- Frequentist statistics assumes the i.i.d. condition, hence it suffices to consider $m = 1$ and $\mathcal{X}$ is measurable.

- Frequentist regards $\{\mathcal{P}(\cdot|x), x \in \mathcal{X}\}$ is an element in $\mathbf{Meas}\big(\mathcal{X}, \mathcal{P}(\mathcal{Y})\big)$. Bayesians regards $\{\mathcal{P}(\cdot|x), x \in \mathcal{X}\}$ as an element in $\mathcal{P}(\mathcal{Y})^{\mathcal{X}}$ with a certainty encoded in a prior probability measure $\mu \in \mathcal{P}(\mathcal{P}(\mathcal{Y})^{\mathcal{X}})$, which can be seen as the law of a stochastic process $(\Omega, \mu) \times \mathcal{X} \to \mathcal{P}(\mathcal{Y})$.

- $(\mathcal{P}(\mathcal{Y}), \Sigma_w)$ is a measurable space s.t. $\forall A \in \Sigma_{\mathcal{Y}}$, the map $\mathcal{P}(\mathcal{Y}) \to \mathbf{R}, \mu \mapsto \mu(A)$, is measurable. $(\mathcal{P}(\mathcal{Y})^{\mathcal{X}}, \Sigma_{cyl})$ is measurable.

- Frequentists approximate the "true" $h : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ by $h_{\mathcal{H}} \in \mathcal{H} \subset \mathbf{Meas}\,(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$ via a loss function and empirical data $S_n \in (\mathcal{X} \times \mathcal{Y})^n$. Bayesians predict $\mathcal{P}_{T_m|S_n} \in \mathcal{P}(\mathcal{Y}^m)$ using prior measure μ and Bayesian inversion.

- Bayesian inversions are best understood best via the concept of probabilistic morphisms, which is a categorical approach to Markov kernels and their calculations.

2. Markov kernel, probabilistic morphism and Bayesian supervised learning

- A Markov kernel : $T \in \mathbf{Meas}\big(\mathcal{X}, \mathcal{P}(\mathcal{Y})\big)$.

- For $T_1 \in \mathbf{Meas}\big(\mathcal{X}, \mathcal{P}(\mathcal{Y})\big)$, $T_2 \in \mathbf{Meas}\big(\mathcal{Y}, \mathcal{P}(\mathcal{Z})\big)$ $T_2 \circ T_1 \in \mathbf{Meas}\big(\mathcal{X}, \mathcal{P}(\mathcal{Z})\big)$ is defined by

$$T_2 \circ T_2(B|x) := \int_{\mathcal{Y}} T_2(B|y) dT_2(y|x),\ B \in \Sigma_{\mathcal{Z}}$$

This composition is associative.

- For $T \in \mathbf{Meas}\big(\mathcal{X}, \mathcal{P}(\mathcal{Y})\big)$, $\underline{T} : \mathcal{X} \rightsquigarrow \mathcal{Y}$ is the probabilistic morphism generated by T.

- For $\mu \in \mathcal{P}(\mathcal{X})$, $(\underline{T})_*\mu \in \mathcal{P}(\mathcal{Y})$:

$$(\underline{T})_*\mu(B) := \int_{\mathcal{X}} T(B|x)d\mu(x).$$

This defines a faithful functor from Category of Markov kernels to Category of measurable spaces.

- For $\mathbf{p} \in \mathbf{Meas}\Big(\mathcal{X}, \mathcal{P}(\mathcal{Y})\Big)$, the reduced graph $\Gamma^\bullet_{\mathbf{p}} : \mathcal{X} \to \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is defined by:

$$\Gamma^\bullet_{\mathbf{p}}(x) := \delta_x \otimes \mathbf{p}(x).$$

Its generated probabilistic morphism is denoted by $\Gamma_{\underline{\mathbf{p}}} : \mathcal{X} \rightsquigarrow \mathcal{X} \times \mathcal{Y}$, the graph of $\underline{\mathbf{p}}$.

$$(\Gamma_{\underline{\mathbf{p}}})_*\mu(A \times B) = \int_{\mathcal{A}} T(B|x)d\mu(x). \qquad (1)$$

- $\forall \mathcal{X}$ Evaluation $E_{X_m} : \mathcal{P}(\mathcal{Y})^{\mathcal{X}} \to \mathcal{P}(\mathcal{Y})^m$ and Inclusion $\mathfrak{m}^m : \mathcal{P}(\mathcal{Y})^m \to \mathcal{P}(\mathcal{Y}^m)$ are measurable.

- A Bayesian statistical model $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{X})$: (Θ, μ_Θ) - probability space, $\mathbf{p} \in \mathbf{Meas}\big(\Theta, \mathcal{P}(\mathcal{X})\big)$. The predictive distribution $\mu_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$ is defined by $\mu_{\mathcal{X}} := (\Pi_{\mathcal{X}})_*(\Gamma_{\underline{\mathbf{p}}})_*\mu_\Theta = (\underline{\mathbf{p}})_*\mu_\Theta$:

$$\mu_{\mathcal{X}}(A) = \int_\Theta \mathbf{p}(A|\theta)d\mu_\Theta(\theta),\ A \in \Sigma_{\mathcal{X}}.$$

- We solve Problem SBI for $\mathcal{X} = \{pt\}$ with help of Bayesian inversion.

- A Markov kernel $\mathbf{q} : \mathcal{X} \to \mathcal{P}(\Theta)$ is called a Bayesian inversion of a Markov kernel $\mathbf{p} : \Theta \to \mathcal{P}(\mathcal{X})$ relative to $\mu_\Theta \in \mathcal{P}(\Theta)$ if

$$(\sigma_{\mathcal{X},\Theta})_*(\Gamma_{\underline{\mathbf{q}}})_*(\underline{\mathbf{p}})_*\mu_\Theta = (\Gamma_{\underline{\mathbf{p}}})_*\mu_\Theta.$$

- For a Bayesian inversion $\mathbf{q}^{(n)} : \mathcal{X}^n \to \mathcal{P}(\Theta)$ of $\mathbf{p}^n : \Theta \to \mathcal{P}(\mathcal{X}^n)$ relative to μ_Θ, the posterior distribution of μ_Θ after seeing $S_n \in \mathcal{X}^n$ is $\mathbf{q}^{(n)}(S_n) \in \mathcal{P}(\Theta)$. The posterior predictive distribution $\mathcal{P}_{T_m|S_n} := (\underline{\mathbf{p}}^n)_*(\mu_{\Theta|S_n}) \in \mathcal{P}(\mathcal{X}^n)$.

- A Bayesian model for the supervised inference problem SBI (Le2025) is $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{P}(\mathcal{Y})^{\mathcal{X}})$, where $\mu_\Theta \in \mathcal{P}(\Theta)$, $\mathbf{p} \in \mathbf{Meas}\,(\Theta, \mathcal{P}(\mathcal{Y})^{\mathcal{X}})$.

(1) For $X_m = (x_1, \ldots, x_m) \in \mathcal{X}^m$, $\mathbf{p}_{X_m} := \mathfrak{m}^m \circ E_{X_m} \circ \mathbf{p} : \Theta \to \mathcal{P}(\mathcal{Y}) \times \ldots \times \mathcal{P}(\mathcal{Y}) \hookrightarrow \mathcal{P}(\mathcal{Y}^m)$ parameterizes sampling distributions of $Y_m = (y_1, \ldots, y_m) \in \mathcal{Y}^m$, where y_i is a label of x_i, with certainty encoded in μ_Θ.

(2) For $S_n \in (\mathcal{X} \times \mathcal{Y})^n$, $\mu_{\Theta|S_n} := \mathbf{q}_{\Pi_{\mathcal{X}}(S_n)}\big(\Pi_{\mathcal{Y}}(S_n)\big)$, where $\mathbf{q}_{\Pi_{\mathcal{X}}(S_n)} : \mathcal{Y}^n \to \mathcal{P}(\Theta)$ is a Bayesian inversion of $\mathbf{p}_{\Pi_{\mathcal{X}}(S_n)} : \Theta \to \mathcal{P}(\mathcal{Y}^n)$ relative to μ_Θ.

(3) The posterior predictive distribution $\mathcal{P}_{T_m|S_n,\mu_\Theta} \in \mathcal{P}(\mathcal{Y}^m)$ is defined as the predictive distribution of the Bayesian statistical model $(\Theta, \mu_{\Theta|S_n}, \mathbf{p}_{T_m}, \mathcal{Y}^m)$, i.e.,

$$\mathcal{P}_{T_m|S_n,\mu_\Theta} := (\underline{\mathbf{p}_{T_m}})_* \mu_{\Theta|S_n} \in \mathcal{P}(\mathcal{Y}^m).$$

(4) The aim of a learner is to estimate and approximate $\mathcal{P}_{T_m|S_n,\mu_\Theta}$.

- $(\mathcal{P}(\mathcal{Y})^{\mathcal{X}}, \mu, \mathrm{Id}_{\mathcal{P}(\mathcal{Y})^{\mathcal{X}}}, \mathcal{P}(\mathcal{Y})^{\mathcal{X}})$ is a universal model for Bayesian supervised learning in the following sense. Assume that $\mathbf{q}^{(m)} : \mathcal{Y}^m \to \mathcal{P}(\Theta)$ is a Bayesian inversion of $\mathfrak{m}^m \circ E_{X_m} \circ \mathbf{p} : \Theta \to \mathcal{P}(\mathcal{Y}^m)$ for $X_m \in \mathcal{X}^m$. Then $\mathbf{p}_* \circ \mathbf{q}^{(m)} : \mathcal{Y}^m \to \mathcal{P}\big(\mathcal{P}(\mathcal{Y})^{\mathcal{X}}\big)$ is a Bayesian inversion of $\mathbf{p}_{X_m} : \mathcal{P}(\mathcal{Y})^{\mathcal{X}} \to \mathcal{P}(\mathcal{Y}^m)$. Consequently, for any $T_m \in \mathcal{X}^m$ and $X_n \in \mathcal{X}^n$ and w.$(\underline{\mathbf{p}_{X_n}})_*\mu_{\Theta}$-a.e. $Y_n \in \mathcal{Y}^n$ we have

$$\mathcal{P}_{T_m|S_n(X_n,Y_n),\mu_\Theta} = \mathcal{P}_{T_m|S_n(X_n,Y_n),\mathbf{p}_*\mu_\Theta} \qquad (2)$$

where the RHS of (2) is the posterior predictive distribution of $\big(\mathcal{P}(\mathcal{Y})^{\mathcal{X}}, \mathbf{p}_*\mu_\Theta, \mathrm{Id}_{\mathcal{P}(\mathcal{Y})^{\mathcal{X}}}, \mathcal{P}(\mathcal{Y})^{\mathcal{X}}\big)$.

3. Bayesian batch learning vs online learning.

- A batch learning algorithm is a map $A : \cup_{n=1}^{\infty} \mathcal{X}^n \times \mathcal{H} \to \mathcal{H}$ where $\mathcal{X}$ is a sample space and $\mathcal{H}$ is a hypothesis space.

- Online learning: $(x_1, \ldots, x_n,)$ is a time-series. We regard $\mathcal{X}^n \times \mathcal{H} \to \mathcal{H}$ as a discrete dynamical system $\mathcal{X} \times (\ldots \times (\mathcal{X} \times \mathcal{H}) \to \mathcal{H}$. Online learning advantages: higher computational efficiency and adaptability.

Theorem 1. Online formula for Bayesian inversion (Le2025)

- $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{P}(\mathcal{Y})^{\mathcal{X}})$ - Bayesian model.
- $S_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (\mathcal{X} \times \mathcal{Y})^n$,
- $S_{n-1} = \big((x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\big)$. Then

$$\mathbf{q}_{\Pi_{\mathcal{X}}(S_n)}(y_n, \ldots, y_1 \| \mu_\Theta) :=$$

$$\mathbf{q}_{x_n}(y_n \| \mathbf{q}_{\Pi_{\mathcal{X}}(S_{n-1})}(y_{n-1}, \ldots, y_n) \| \mu_\Theta).$$

defines a Bayesian inversion

$\mathbf{q}_{\Pi_{\mathcal{X}}(S_n)}(\cdot \| \mu_\Theta) : \mathcal{Y}^n \to \mathcal{P}(\Theta)$ of $\mathbf{p}_{\Pi_{\mathcal{X}}(S_n)} : \Theta \to \mathcal{P}(\mathcal{Y}^n)$ relative to μ_Θ.

Theorem 2. Posterior predictive distribution (Le2025)

- $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{P}(\mathcal{Y})^{\mathcal{X}})$,
- $X_n = (x_1, \ldots, x_n) \in \mathcal{X}^n$,
- $T_m = (t_1, \ldots, t_m) \in \mathcal{X}^m$.
- For $Y_n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ let $S_n := S_n(X_n, Y_n)$ $:= \Big((x_1, y_1), \ldots, (x_n, y_n)\Big) \in (\mathcal{X} \times \mathcal{Y})^n$.

Then $\mathcal{P}^n_m : \mathcal{Y}^n \to \mathcal{P}(\mathcal{Y}^m)$, defined by

$$(y_1, \ldots, y_n) \mapsto \mathcal{P}_{T_m|\Big((x_1,y_1),\ldots(x_n,y_n)\Big),\mu_\Theta},$$

is a reg. conditional probability measure for

$$\mu^0_{T_m,S_n,\mu_\Theta} := \underline{\mathfrak{m}^2(\mathbf{p}_{T_m}, \mathbf{p}_{X_n})}_* \mu_\Theta \in \mathcal{P}(\mathcal{Y}^m \times \mathcal{Y}^n)$$

with respect ot the projection $\Pi_{\mathcal{Y}^n} : \mathcal{Y}^m \times \mathcal{Y}^n \to \mathcal{Y}^n$.

$$\begin{tikzcd}
(\Theta, \mu_{\Theta}) \arrow[rr, squiggly, "\mathfrak{m}^2(\mathbf{p}_{T_m}, \mathbf{p}_{X_n})"] \arrow[d, squiggly, "\underline{\mathbf{p}_{T_m}}"'] \arrow[rrd, squiggly, "\underline{\mathbf{p}_{X_n}}"'] & & \mathcal{Y}^m \times \mathcal{Y}^n \arrow[d, "\Pi_{\mathcal{Y}^n}"] \\
\mathcal{Y}^m & & \mathcal{Y}^n . \arrow[llu, squiggly, "\underline{\mathbf{q}_{X_n}}"']
\end{tikzcd}$$

2) The posterior predictive distribution $\mathcal{P}_{T_m | S_n, \mu_{\Theta}} \in \mathcal{P}(\mathcal{Y}^m)$ can be computed recursively.

4. Gaussian regression and Kalman filter

In Bayesian regression learning, we learn a function $f \in (V)^{\mathcal{X}}$ where: $V = \mathbf{R}^n$

$y = f(x) + \varepsilon(x) : \ f \in ((V)^{\mathcal{X}}, \mu), \ \varepsilon(x) \in (V, \nu_\varepsilon(x))$.

We model probability of y given x, denoted by $\mu_{V|\mathcal{X}}(x) \in \mathcal{P}(V)$, as generated from

$$\text{Ad} : V \times V \to V, (x, y) \mapsto x + y.$$

For $\mu_1, \mu_2 \in \mathcal{P}(V)$ their convolution is defined as follows:

$$\mu_1 * \mu_2 := (\text{Ad})_*(\mu_1 \otimes \mu_2) \in \mathcal{P}(V).$$

Then

$$\mu_{V|\mathcal{X}}(x) = \delta_{f(x)} * \nu_{\varepsilon(x)}.$$

Let $\mathbf{p}^{\varepsilon}, \mathbf{p}^{0} \in \mathbf{Meas}\left((V)^{\mathcal{X}}, \mathcal{P}(V)^{\mathcal{X}}\right)$ be

$$\begin{aligned} \mathbf{p}^{\varepsilon}(f) &:= \delta_f * \nu_{\varepsilon} \in \mathcal{P}(V)^{\mathcal{X}}, \\ (\delta_f * \nu_{\varepsilon})(x) &:= \delta_{f(x)} * \nu_{\varepsilon}(x) \text{ for } x \in \mathcal{X}, \\ \mathbf{p}^{0}(f) &:= \delta_f \in \mathcal{P}(V)^{\mathcal{X}}, \, \delta_f(x) := \delta_{f(x)} \end{aligned}$$

Then we can learn $f \in (V^{\mathcal{X}}, \mu)$ using a Bayesian regression model $\Theta, \mu_{\Theta}, h, (V)^{\mathcal{X}})$ that generates two Bayesian supervised learning models $(\Theta, \mu, \mathbf{p}^{0}, \mathcal{P}(V)^{\mathcal{X}})$ and $(\Theta, \mu, \mathbf{p}^{\varepsilon}, \mathcal{P}(V)^{\mathcal{X}})$.

For $X_m = (x_1, \ldots, x_m) \in \mathcal{X}^m$, $f \in V^{\mathcal{X}}$ $E^V_{X_m}(f) := \big(f(x_1), \ldots, f(x_m)\big)$. $(\Theta, \mu_\Theta, h^\varepsilon_{X_m} := C_{\otimes_{i=1}^m \nu_\varepsilon(x_i)} \circ \delta \circ E^V_{X_m} \circ h,^m)$ is a Bayesian model for learning the distribution of $(y_1 = f(x_1) + \varepsilon(x_1), \ldots, y_m = f(x_m) + \varepsilon(x_m)) \in V^m$.

For $S_n \in (\mathcal{X} \times V)^n$, the posterior distribution $\mu_{\Theta|S_n} \in \mathcal{P}(\Theta)$ is $\mathbf{q}^\varepsilon_{\Pi_{\mathcal{X}}(S_n)}(\Pi_{\mathcal{Y}}(S_n))$ where $\mathbf{q}^\varepsilon_{\Pi_{\mathcal{X}}(S_n)} : V^n \to \mathcal{P}(\Theta)$ is a Bayesian inversion of $h^\varepsilon_{\Pi_{\mathcal{X}}(S_n)} : \Theta \to \mathcal{P}(V^n)$ relative to μ_Θ.

For $T_m = (t_1, \ldots, t_m) \in \mathcal{X}^m$, the posterior predictive distribution $\mathcal{P}_{T_m|S_n,\mu_\Theta} \in \mathcal{P}(V^m)$ of the tuple $\left(y'_1 = f(t_1), \ldots, y'_m = f(t_m)\right)$ after seeing S_n is defined as the predictive distribution of the Bayesian statistical model $(\Theta, \mu_{\Theta|S_n}, h^0_{T_m} := \mathfrak{m}^m \circ E_{T_m} \circ \mathbf{p}^0 \circ h, \mathcal{P}(V^m))$

- $(V^{\mathcal{X}}, \mu, \mathrm{Id}_{V^{\mathcal{X}}}, V^{\mathcal{X}})$ is a universal Bayesian regression model.

- Let $\mathcal{X}$ be an arbitrary set. A Gaussian process regression model is a Bayesian regression model $(\mathbf{R}^{\mathcal{X}}, \mu = \mathcal{GP}(m, K), \mathbf{Id}_{\mathbf{R}^{\mathcal{X}}}, \mathbf{R}^{\mathcal{X}})$ where $\varepsilon(x_i) \in (\mathbf{R}, \mathcal{N}(0, \varepsilon^2(x_i)))$.

- $\varepsilon(x) > 0 \ \forall x \in \mathcal{X} \implies \mathbf{p}^{\varepsilon}_{X_k} : V^{\mathcal{X}} \to \mathcal{P}(V^k)$ is a dominated Markov kernel : $\mathbf{p}^{\varepsilon}_{X_k}(g) \ll \lambda_k$. Hence a Bayesian inversion $\mathbf{q}^{\varepsilon}_{X_k} : \mathbf{R}^k \to \mathcal{P}(\mathbf{R}^{\mathcal{X}})$ of $\mathbf{p}^{\varepsilon}_{X_k} : \mathbf{R}^{\mathcal{X}} \to \mathcal{P}(\mathbf{R}^k)$ relative to $\mu = \mathcal{GP}(m, K)$ can be found by the classical Bayesian inversion formula.

- Theorem 2 in this case. Denote training points by

$$S_n = \Big((x_1, y_1), \ldots, (x_n, y_n)\Big) \in (\mathcal{X} \times \mathbf{R})^n$$

. Let $T_m = (t_1, \ldots, t_m) \in \mathcal{X}^m$ be test points and $X_n := \Pi_{\mathcal{X}}(S_n) = (x_1, \ldots, x_n)$.

$$\begin{tikzcd}
& \Big(\mathbf{R}^{\mathcal{X}}, \mu\Big) \arrow[ddl, squiggly, "\underline{\mathbf{p}^0_{T_m}}"'] \arrow[d, squiggly] & \\
& (\mathbf{R}^m \times \mathbf{R}^n, \mu^0_{(T_m, S_n, \mu)}) \arrow[dl, "\Pi_{\mathbf{R}^m}"] \arrow[dr, "\Pi_{\mathbf{R}^n}"'] & \\
\Big(\mathbf{R}^m, (\underline{\mathbf{p}^0_{T_m}})_* \mu\Big) & & \Big(\mathbf{R}^n, (\underline{\mathbf{p}^\varepsilon_{X_n}})_* \mu \arrow[uul, squiggly, "\underline{\mathbf{q}^\varepsilon_{X_n}}"'] \arrow[ll, squiggly, "\underline{\mathbf{p}^0_{T_m}} \circ \underline{\mathbf{q}^\varepsilon_{X_n}}"']
\end{tikzcd}$$

$$\begin{tikzcd}
 & \Big(\mathbf{R}^{\mathcal{X}}, \mu\Big) \arrow[d, rightsquigarrow] \arrow[ddl, rightsquigarrow, "\underline{\mathbf{p}^0_{T_m}}"'] & \\
 & \left(\mathbf{R}^m \times \mathbf{R}^n, \mu^0_{(T_m,S_n,\mu)}\right) \arrow[dl, "\Pi_{\mathbf{R}^m}"'] \arrow[dr, "\Pi_{\mathbf{R}^n}"] & \\
\Big(\mathbf{R}^m, (\underline{\mathbf{p}^0_{T_m}})_*\mu\Big) & & \left(\mathbf{R}^n, (\underline{\mathbf{p}^\varepsilon_{X_n}})_*\mu\right. \arrow[uul, rightsquigarrow, "\underline{\mathbf{q}^\varepsilon_{X_n}}"'] \arrow[ll, rightsquigarrow, "\underline{\mathbf{p}^0_{T_m}}\circ\underline{\mathbf{q}^\varepsilon_{X_n}}"']
\end{tikzcd}$$

$\mu = \mathcal{GP}(m, K)$, $\mu^0_{(T_m,S_n,\mu)}$ is also a Gaussian measure on $\mathbf{R}^{m+n}$.

$$\mu^0_{(T_m,S_n,\mu)} = \mathcal{N}\Big(m(T_m, S_n, \mu), \Sigma(T_m, S_n, \mu)\Big) .$$

- $\Sigma(T_m, S_n, \mu)$ is the quadratic form whose matrix expression in coordinates $\{e_1, \ldots, e_m\} \in \mathbf{R}^m$ and $\{e_{m+1}, \ldots, e_{m+m}\} \in \mathbf{R}^n$ has the following form:

$$\Sigma(T_m, S_n, \mu) = \left(\begin{array}{cc} \Sigma_{m,m} & \Sigma_{m,n} \\ \Sigma_{n,m} & \Sigma_{n^\varepsilon,n^\varepsilon} \end{array} \right),$$

- $\Sigma_{n^\varepsilon,n^\varepsilon}$ is the covariance quadratic form of the Gaussian measure $(\underline{\mathbf{p}^\varepsilon})_*\mu$ on $\mathbf{R}^n$,

- $\Sigma_{m,m}$ is the covariance quadratic form of the Gaussian measure $(\underline{\mathbf{p}^0_{T_m}})_*\mu$ on $\mathbf{R}^m$,

- $\Sigma_{n,m}$ are components of the covariance matrix of the Gaussian measure $\mu^0_{(T_m,S_n,\mu)}$ on $\mathbf{R}^{m+n}$ in the same coordinates.

Recall that $Y_n = \Pi_{\mathbf{R}}(S_n) \in \mathbf{R}^n$. By Theorem 2, the posterior predictive distribution $\mathcal{P}_{T_m|S_n,\mu} \in \mathcal{P}(\mathbf{R}^m)$ is a Gaussian measure $\mathcal{N}(m_{T_m|S_n,\mu}, \Sigma_{T_m|S_n,\mu})$,

$$m_{T_m|S_n,\mu} = \Sigma_{m,n}(\Sigma_{n^\varepsilon,n^\varepsilon})^{-1}(Y_n) \in \mathbf{R}^m$$

$\Sigma_{T_m|S_n,\mu} = \Sigma_{m,m} - \Sigma_{m,n}\Sigma^{-1}_{n^\varepsilon,n^\varepsilon}\Sigma_{n,m} \in S^2_+(\mathbf{R}^m)$.

where $\Sigma^{-1}_{n^\varepsilon,n^\varepsilon}$ is pseudo inverse of $\Sigma_{n^\varepsilon,n^\varepsilon}$.

This sequential update procedure is known to be equivalent to the celebrated Kalman filter update equations. Kalman filtering is much applied in time series analysis, signal processing, robotic motion planning, trajectory optimization.

5. Final Remarks

- Unified categorical framework for Bayesian supervised learning + $\exists$ universal spaces.

- Equivalence of batch and online learning in Bayesian settings.

- Connection between Gaussian process regression and Kalman filtering.

- Priors on universal parameter space $\mathcal{P}(\mathcal{Y})^{\mathcal{X}}$ (and $\mathcal{Y}^{\mathcal{X}}$) are constructed using projective systems. (in arXiv:2510.16892).

Open Problems.

1. Consistency of posterior predictive distributions $\mathcal{P}_{T_m|S_n,\mu}$ for more general class of Bayesian supervised learning than Gaussian process regression, e.g., for the class of Dependent Dirichlet Process.

2. Efficiency of Bayesian neural networks and neural processes. (BNNs extend traditional neural networks by incorporating Bayesian inference, treating network weights as probability distributions rather than fixed point estimates).

References

- W.F. Lawvere. The category of probabilistic mappings. (1962) https://ncatlab.org/nlab/files/lawvereprobability1962

- N. Chentsov. Statistical decision rules and optimal inference, Nauka: Moscow, Russia (1972). English translation in: Translations of Mathematical Monograph vol. 53, Amer. Math. Soc.: Providence, RI, USA (1982)

- M. Giry. A categorical approach to probability theory. In: B. Banaschewski, editor, Categorical Aspects of Topology and Analysis, Lecture Notes in Mathematics, vol. 915, 68-85, Springer (1982)

-J. Jost, H. V. Lê, and T. D. Tran. Probabilistic morphisms and Bayesian nonparametrics. Eur. Phys. J. Plus 136, 441 (2021).

- H. V. Lê. Probabilistic morphism and Bayesian supervised learning, Mat. Sbornik 216, Nr 5, pp. 161-180 (2025), English translation in Sbornik: Mathematics 216:5 723-741.

- H. V. Lê. Batch learning equals online learning in Bayesian supervised learning, arXiv:2510.16892.

- H. V. Lê, H.Q. Minh, F.Protin, W.Tuschmann: Mathematical Foundations of Machine Learning, Springer, 2026.

THANK YOU FOR YOUR ATTENTION!