Learning to Defer For Time Series via Switching State-Space Models

November 24, 2025

Abstract

We present a theoretical framework for Learning-to-Defer (L2D) in non-stationary time series environments where querying experts incurs an explicit consultation cost. We formalize the problem as a sequential decision process under partial observation with a time-varying set of experts. We model the time-varying reliability of black-box experts (e.g., neural networks) as latent states in a Switching Linear Dynamical System (SLDS). This structure captures abrupt environmental shifts (regimes) and gradual performance drift. We derive a tractable inference algorithm using the Interacting Multiple Model (IMM) filter, providing full derivations for the variance spread in Gaussian mixtures, and propose a cost-sensitive myopic selection policy that balances predictive risk, epistemic uncertainty, and operational costs.

Note: might be interested to predict when the expert will be or not available?

1 Problem Formulation

1.1 Preliminaries

We consider a sequential forecasting problem over a discrete time horizon $t = 1, \dots, T$. The system is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t \geq 0}$, representing the information available to the router up to (and including) time t.

- **Context and Target (Regression Setting).** At each time t, nature reveals a context vector $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$, which is $\mathcal{F}_t$-measurable (known to the router at decision time). The associated ground-truth target is a real-valued quantity

$$y_t \in \mathcal{Y} \subseteq \mathbb{R},$$

  which is latent at decision time and becomes observable only through partial feedback after the decision is made; in particular, y_t is not $\mathcal{F}_t$-measurable.

- **Dynamic Expert Universe.** Let $\mathcal{U} = \{1, \dots, N_{\max}\}$ denote the universe of all potential experts (models, human specialists, or external systems). At each time t, only a subset $\mathcal{K}_t \subseteq \mathcal{U}$ is actually available for querying. The set $\mathcal{K}_t$ is allowed to vary over time, capturing phenomena such as system downtime, onboarding or removal of models, or doctor shifts.

- **Expert Predictions.** Each expert $j \in \mathcal{U}$ is associated with a predictive function $f_j : \mathcal{X} \to \mathcal{Y}$. At time t, any available expert $j \in \mathcal{K}_t$ produces a prediction

$$\hat{y}_t^{(j)} \coloneqq f_j(\mathbf{x}_t),$$

  based on the current context $\mathbf{x}_t$.

- **Loss Process and Partial Feedback.** Predictive performance is evaluated via a loss function
$$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty),$$
e.g., squared error $\mathcal{L}(\hat{y}, y) = (\hat{y} - y)^2$ or absolute error $\mathcal{L}(\hat{y}, y) = |\hat{y} - y|$. The instantaneous loss of expert j at time t is the random variable
$$\ell_{j,t} \coloneqq \mathcal{L}(\hat{y}_t^{(j)}, y_t) \in [0, +\infty).$$
We assume that the family $\{\ell_{j,t} : j \in \mathcal{U},\, t \geq 1\}$ forms a well-defined stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$. Because only the queried expert is evaluated, after choosing an expert index $r_t \in \mathcal{K}_t$ at time t the router only observes the loss
$$\ell_{r_t,t},$$
while the losses $\ell_{j,t}$ for $j \neq r_t$ remain unobserved. In particular, $\ell_{r_t,t}$ is not contained in $\mathcal{F}_t$, but becomes available before the next decision epoch and is included in $\mathcal{F}_{t+1}$.

- **Information History and Admissible Policies.** For each $t \geq 1$, define the feedback random element
$$F_t \coloneqq (r_t, \ell_{r_t,t}) \in \mathcal{U} \times [0, +\infty).$$
The interaction history up to, but not including, time t is the finite sequence
$$\mathcal{I}_{t-1} \coloneqq \big((\mathbf{x}_\tau, \mathcal{K}_\tau, F_\tau)\big)_{1 \leq \tau \leq t-1},$$
with the convention that $\mathcal{I}_0$ is the empty sequence. The information available to the router at decision time t is summarized by the random element
$$H_t \coloneqq (\mathcal{I}_{t-1},\, \mathbf{x}_t,\, \mathcal{K}_t).$$
We assume that the observation filtration is generated by this information, i.e.
$$\mathcal{F}_t = \sigma(H_t), \qquad t = 1, \ldots, T.$$
A (deterministic) routing policy is a sequence of measurable mappings
$$\pi_t : \operatorname{supp}(H_t) \to \mathcal{U}, \qquad t = 1, \ldots, T,$$
such that $\pi_t(H_t) \in \mathcal{K}_t$ almost surely. The chosen expert at time t is then
$$r_t \coloneqq \pi_t(H_t),$$
so in particular r_t is $\mathcal{F}_t$-measurable with values in $\mathcal{K}_t$.

Running Example (Medical Triage). To ground these definitions, consider an automated triage system in a hospital Emergency Room.

- The **Context** $\mathbf{x}_t$ represents a patient's initial vitals, laboratory values, and presenting symptoms.
- The **Target** y_t is a real-valued outcome of interest, e.g., a risk score or a biomarker value observed after further testing.
- The **Experts** in $\mathcal{K}_t$ are the diagnostic models or human specialists currently available (e.g., a chest-pain model, a sepsis model, an on-call cardiologist). As doctors start or end their shifts, or as models are deployed/retired, the set $\mathcal{K}_t$ changes over time.

- For a given patient t assigned to specialist j, the **Loss** $\ell_{j,t}$ quantifies the discrepancy between the prediction $f_j(\mathbf{x}_t)$ and the realized outcome y_t (e.g., squared error). Only the loss of the chosen specialist $j = r_t$ is observed and enters subsequent decisions.
- **Non-Stationarity** arises from environmental shifts such as the onset of a flu epidemic, changes in hospital protocols, or gradual degradation of a sensor's calibration, which can all alter the relative performance of experts over time.

The router must therefore select, at each time t, which expert $r_t \in \mathcal{K}_t$ to query based solely on the current patient's features $\mathbf{x}_t$ and the historical feedback $\mathcal{I}_{t-1}$, without access to the full outcome y_t at decision time.

1.2 The Decision Process

The system operates as a *cost-sensitive router*. At each decision epoch $t \in \{1, \dots, T\}$, after observing the information H_t (recall that H_t collects the past interaction history together with the current context $\mathbf{x}_t$ and availability set $\mathcal{K}_t$), the router must select a single expert index $r_t \in \mathcal{K}_t$ to query for the current instance.

- **Action Space.** At time t, the admissible actions form the finite set $\mathcal{K}_t \subseteq \mathcal{U}$. The router chooses an index
$$r_t \in \mathcal{K}_t.$$
- **Consultation Cost.** Each expert $j \in \mathcal{U}$ is associated with a deterministic, non-negative consultation cost
$$\beta_j \geq 0,$$
modeling, for instance, API fees, latency constraints, or computational, energy, and environmental costs. The vector $(\beta_j)_{j \in \mathcal{U}}$ is assumed known to the router.

Realized Cost. At time t, the *total realized cost* incurred when selecting expert $r_t \in \mathcal{K}_t$ is the sum of its prediction loss and its consultation cost:

$$C_t(r_t) \coloneqq \underbrace{\ell_{r_t,t}}_{\text{prediction loss}} + \underbrace{\beta_{r_t}}_{\text{consultation cost}}. \tag{1}$$

The quantity $\ell_{r_t,t}$ is a random variable at decision time t (its distribution is determined by the underlying data-generating process and the chosen expert), whereas β_{r_t} is deterministic and known.

Objective Function. Let H_t denote the state space in which H_t takes values. A (deterministic, possibly non-stationary) routing policy is a sequence of measurable mappings

$$\pi = (\pi_t)_{t=1}^{T}, \qquad \pi_t : \mathsf{H}_t \to \mathcal{U},$$

such that $\pi_t(H_t) \in \mathcal{K}_t$ almost surely. Given a policy π, the chosen expert at time t is

$$r_t \coloneqq \pi_t(H_t),$$

so r_t is $\mathcal{F}_t$-measurable with values in $\mathcal{K}_t$.

The goal is to find a causal policy π that minimizes the expected cumulative cost over the horizon $\{1, \dots, T\}$:

$$\inf_{\pi} \ \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T} C_t(\pi_t(H_t))\right] \ = \ \inf_{\pi} \ \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T} \big(\ell_{\pi_t(H_t),t} + \beta_{\pi_t(H_t)}\big)\right], \tag{2}$$

where the expectation is taken with respect to the joint law of the stochastic process $(\mathbf{x}_t, y_t)_{t=1}^{T}$ and the induced loss process $(\ell_{j,t})_{j,t}$. Any policy π^* attaining the infimum in (2) (if such a policy exists) is called optimal.

Intuition (The "Economy of Diagnosis"). Revisiting the medical triage example:

- **Expert A (Nurse Algorithm):** modest accuracy on complex cases, but very cheap (e.g., $\beta_A \approx 0$);
- **Expert B (Specialist):** high accuracy, but expensive or slow (e.g., $\beta_B \gg 0$).

A purely accuracy-driven scheme (ignoring the consultation costs β_j) might always select Expert B whenever it yields a smaller expected loss. The cost-sensitive objective (2) instead forces the router to trade off accuracy against resource usage, effectively asking: *Is the expected reduction in prediction loss from consulting the specialist worth the additional cost β_B?* For a patient whose context $\mathbf{x}_t$ indicates a simple, low-risk case, the router may select the cheaper Nurse Algorithm; for an ambiguous or high-risk case, it is preferable to pay the higher cost and defer to the specialist.

Remark 1 (Bayes action under squared loss)**.** Assume that at time t the router knows the conditional distribution of y_t given H_t, and that the instantaneous cost for choosing expert $j \in \mathcal{K}_t$ is

$$C_t(j) \coloneqq \big(f_j(\mathbf{x}_t) - y_t\big)^2 + \beta_j.$$

Let $m_t \coloneqq \mathbb{E}[y_t \mid H_t]$ denote the conditional mean. Then any Bayes action at time t is given by

$$j_t^\star \in \arg\min_{j \in \mathcal{K}_t} \Big\{ \big(f_j(\mathbf{x}_t) - m_t\big)^2 + \beta_j \Big\},$$

i.e., the expert whose prediction is closest (in squared distance) to m_t, up to the consultation cost β_j.

2 Generative Model: Switching Linear Dynamical System

In order to construct a cost-sensitive routing policy, we need, at each decision time t, the conditional distribution of the loss of each available expert,

$$\mathbb{P}(\ell_{j,t} \mid H_t), \qquad j \in \mathcal{K}_t.$$

A static regression model of the form $\ell_{j,t} = g_j(\mathbf{x}_t) + \varepsilon_{j,t}$ would be inadequate here, because (i) expert performance is typically non-stationary (concept drift over time), and (ii) we only observe $\ell_{j,t}$ for the selected expert $j = r_t$ (bandit-style partial feedback), which breaks standard i.i.d. assumptions.

We therefore posit that the losses are generated by a latent *Switching Linear Dynamical System* (SLDS), in which a discrete regime process controls the parameters of a continuous latent state describing expert reliability.

2.1 Latent State Dynamics

The core generative assumption is that the observed losses are emissions from a latent process that exhibits both discrete structural changes and continuous evolution. We model this via a Switching Linear Dynamical System specified by the tuple

$$(\mathcal{Z}, (A_k)_{k=1}^M, (Q_k)_{k=1}^M, \Pi).$$

1. Regime Process (Discrete). Let

$$z_t \in \mathcal{Z} \coloneqq \{1, \dots, M\}$$

be a discrete latent variable representing the global environmental *regime* at time t (for instance, "normal operations" versus "outbreak" in the medical example). The regime evolves according to a time-homogeneous Markov chain with transition matrix $\Pi \in [0,1]^{M \times M}$ and some initial distribution μ^0 on $\mathcal{Z}$:

$$\mathbb{P}(z_t = k \mid z_{t-1} = i) \;=\; \Pi_{ik}, \qquad \sum_{k=1}^{M} \Pi_{ik} = 1, \quad i = 1, \dots, M, \tag{3}$$

and

$$\mathbb{P}(z_1 = k) = \mu_k^0, \qquad k = 1, \dots, M.$$

The regime z_t will govern the parameters of the continuous dynamics, allowing the system to switch, for example, between low-variance and high-variance modes.

2. Reliability Process (Continuous). For each expert $j \in \mathcal{U}$, we introduce a continuous latent state

$$\boldsymbol{\alpha}_{j,t} \in \mathbb{R}^{d_\alpha},$$

which is intended to capture the instantaneous *reliability* of expert j at time t (e.g., bias and context-dependent error coefficients; the precise observation model will be specified in the next subsection). The state is assumed to evolve even when the expert is not selected or not currently available, reflecting the idea that an expert's performance can drift over time irrespective of whether we query it.

Conditional on the current regime $z_t = k$, we posit linear-Gaussian dynamics with *shared* regime-specific dynamics across experts:

$$\boldsymbol{\alpha}_{j,t} \;=\; A_k \, \boldsymbol{\alpha}_{j,t-1} \;+\; \mathbf{w}_{j,t}, \qquad \mathbf{w}_{j,t} \sim \mathcal{N}(\mathbf{0},\, Q_k), \tag{4}$$

for all $t \geq 2$ and all $j \in \mathcal{U}$, where

- $A_k \in \mathbb{R}^{d_\alpha \times d_\alpha}$ is the state transition matrix in regime k (e.g., encoding mean-reversion or persistence);
- $Q_k \in \mathbb{S}_{++}^{d_\alpha}$ is the corresponding process noise covariance, governing the rate at which the state diffuses over time in regime k.

One may additionally specify a prior distribution for the initial states, e.g. $\boldsymbol{\alpha}_{j,1} \sim \mathcal{N}(m_{0,j}, P_{0,j})$, independent across j and independent of $(z_t)_{t \geq 1}$.

For clarity, we stress that, conditional on the regime sequence $(z_t)_{t \geq 1}$, the processes $(\boldsymbol{\alpha}_{j,t})_{t \geq 1}$ are assumed independent across experts j and each follows the linear-Gaussian dynamics (4).

Intuition: Drift versus Shift. This hybrid (discrete/continuous) structure separates two qualitatively different forms of non-stationarity:

- **Gradual Drift (via $\boldsymbol{\alpha}_{j,t}$).** The continuous state captures slow, continuous changes in expert reliability, such as a sensor gradually losing calibration or a clinician's performance changing over a shift. In a single-regime linear-Gaussian model, this is handled by the process noise covariance Q_k, as in a standard Kalman filter.
- **Abrupt Shifts (via z_t).** The discrete regime process accounts for sudden, discontinuous changes, such as the emergence of a new virus strain, a sudden change in patient mix, or a hardware failure. These events are modeled by switches in the underlying parameters (A_k, Q_k) and in the prior over $\boldsymbol{\alpha}_{j,t}$.

A standard (single-regime) Kalman filter tends to be *sluggish* in the presence of abrupt shifts: it interprets a large shock as an unusually large noise realization and may require many observations to adapt its state estimate. By contrast, an SLDS can assign high posterior probability to a "volatile" regime k when confronted with an apparent shock, thereby switching to dynamics with, say, larger Q_k and different A_k. This allows rapid adaptation to new conditions while still retaining information accumulated during previous, more stable regimes.

2.2 Dynamic Expert Availability

Our model must accommodate changes in the available expert set $\mathcal{K}_t$ over time. Recall that the latent state process $(\boldsymbol{\alpha}_{j,t})_{t\geq 1}$ is defined for every expert $j \in \mathcal{U}$, independently of whether j is currently available. Dynamic availability therefore affects only the *observation* updates (i.e., when we do or do not receive a loss for expert j), not the underlying state dynamics (4).

At each time t, for each regime $k \in \mathcal{Z}$ and each expert $j \in \mathcal{U}$, the SLDS filter maintains a Gaussian approximation

$$\boldsymbol{\alpha}_{j,t} \,|\, \{z_t = k, H_t\} \;\approx\; \mathcal{N}(m_{j,t}^{(k)}, P_{j,t}^{(k)}),$$

where $m_{j,t}^{(k)} \in \mathbb{R}^{d_\alpha}$ and $P_{j,t}^{(k)} \in \mathbb{S}_{++}^{d_\alpha}$ are the regime-conditional mean and covariance. The effect of dynamic availability can then be described as follows.

- **Expert Removal (Downtime).** Fix $t \geq 2$ and an expert $j \in \mathcal{U}$. Suppose j is available at time $t-1$ but not at time t, i.e.

  $$j \in \mathcal{K}_{t-1} \quad \text{and} \quad j \notin \mathcal{K}_t.$$

  The latent state $\boldsymbol{\alpha}_{j,t}$ is still defined and evolves according to the regime-dependent dynamics (4). In the Kalman/IMM recursion, this means that for expert j at time t we perform *only* the prediction (time-update) step

  $$(m_{j,t-1}^{(k)}, P_{j,t-1}^{(k)}) \;\longmapsto\; (m_{j,t|t-1}^{(k)}, P_{j,t|t-1}^{(k)}),$$

  for each $k \in \mathcal{Z}$, and we skip the measurement (correction) step, since no loss $\ell_{j,t}$ is observed when $j \notin \mathcal{K}_t$. Thus the belief on $\boldsymbol{\alpha}_{j,t}$ is updated purely by passive drift. The expert remains in the global universe $\mathcal{U}$ and can be seamlessly re-integrated when it reappears in $\mathcal{K}_{t'}$ at some later time $t' > t$.

- **Expert Addition (New Expert Joins).** Now suppose $j \in \mathcal{U}$ satisfies

  $$j \notin \mathcal{K}_{t-1} \quad \text{and} \quad j \in \mathcal{K}_t,$$

  i.e., t is the first time at which expert j becomes available (a new model is deployed, or a previously absent server comes online). Since no past loss information is available for j, its state must be initialized from a *population prior*, common across experts with no individual history.

  Formally, for each regime $k \in \mathcal{Z}$ we set

  $$\boldsymbol{\alpha}_{j,t} \,|\, \{z_t = k\} \;\sim\; \mathcal{N}(\mu_{\text{pop}}, \Sigma_{\text{pop}}), \tag{5}$$

  independently of H_t and of the other experts, where $\mu_{\text{pop}} \in \mathbb{R}^{d_\alpha}$ and $\Sigma_{\text{pop}} \in \mathbb{S}_{++}^{d_\alpha}$ are fixed hyperparameters. Equivalently, in terms of the filter parameters, we set

  $$m_{j,t}^{(k)} = \mu_{\text{pop}}, \qquad P_{j,t}^{(k)} = \Sigma_{\text{pop}}, \qquad k \in \mathcal{Z}.$$

  Mathematical role of Σ_{pop}. The choice of Σ_{pop} encodes our epistemic uncertainty about the performance of a newly-arrived expert:

(a) *High Epistemic Uncertainty.* Taking Σ_{pop} to be a large (typically diagonal) positive definite matrix corresponds to a highly diffuse prior on $\boldsymbol{\alpha}_{j,t}$, representing substantial prior uncertainty about the expert's reliability.

(b) *Influence on the Routing Policy.* Under any routing policy that makes use of the *predictive variance* of the loss (for example, UCB/LCB-type rules or mean–variance scores based on $\widehat{\sigma}^2_{j,t} \coloneqq \mathrm{Var}(\ell_{j,t} \mid H_t)$), a large Σ_{pop} induces a large initial value of $\widehat{\sigma}^2_{j,t}$ for the new expert. This, in turn, has the following qualitative effects:

- For a risk-averse, variance-penalizing policy, high initial uncertainty $\widehat{\sigma}^2_{j,t}$ tends to discourage the selection of j until informative feedback has been gathered.
- For an exploration-oriented policy (e.g., one that is optimistic in the face of uncertainty, as in UCB, or that uses Thompson sampling), a large Σ_{pop} increases the probability that j will be selected early, thereby accelerating the reduction of uncertainty about its performance.

(c) *Regime Alignment.* Initializing the expert from the same population prior in each regime k ensures that the SLDS can consistently track the expert's performance as the environment switches between regimes. If desired, one may also specify regime-dependent population priors $(\mu^{(k)}_{\text{pop}}, \Sigma^{(k)}_{\text{pop}})$ to reflect prior knowledge that, say, certain experts are expected to perform differently across regimes.

2.3 Observation Model (Partial Feedback)

To relate the latent reliability states $(\boldsymbol{\alpha}_{j,t})_{j,t}$ to the observed losses, we introduce a fixed feature map

$$\phi : \mathcal{X} \to \mathbb{R}^{d_\alpha}, \qquad \phi_t \coloneqq \phi(\mathbf{x}_t).$$

Conditionally on the regime $z_t = k$ and the state $\boldsymbol{\alpha}_{j,t}$, we model the loss of expert j at time t as a linear-Gaussian emission:

$$\ell_{j,t} \;=\; \phi_t^\top \boldsymbol{\alpha}_{j,t} \;+\; v_{j,t}, \qquad v_{j,t} \sim \mathcal{N}(0,\, R_{k,j}), \tag{6}$$

where $R_{k,j} > 0$ denotes the observation noise variance for expert j in regime k. In other words, for fixed (t, k, j) the emission is obtained by applying the row vector $\phi_t^\top \in \mathbb{R}^{1\times d_\alpha}$ to the latent state $\boldsymbol{\alpha}_{j,t}$ and adding Gaussian noise.

Partial Feedback and Censoring. At decision time t, the router chooses a single expert index $r_t \in \mathcal{K}_t$ based on the information H_t. Subsequently, the system observes the corresponding loss

$$\ell_{r_t,t},$$

while the losses $\ell_{j,t}$ for $j \neq r_t$ remain unobserved. Thus, although the generative model (6) specifies a (latent) loss for every expert $j \in \mathcal{U}$ at each time t, the data available at time t consist only of the single scalar observation $\ell_{r_t,t}$.

From the filtering viewpoint, for each regime $k \in \mathcal{Z}$ and expert $j \in \mathcal{U}$ we maintain Gaussian approximations of the form

$$\boldsymbol{\alpha}_{j,t} \,\big|\, \{z_t = k, H_t\} \;\approx\; \mathcal{N}\big(m^{(k)}_{j,t|t}, P^{(k)}_{j,t|t}\big),$$

with $(m^{(k)}_{j,t|t}, P^{(k)}_{j,t|t})$ obtained recursively from the dynamics (4). Given the *prediction* (time-update) step

$$(m^{(k)}_{j,t-1|t-1}, P^{(k)}_{j,t-1|t-1}) \;\longmapsto\; (m^{(k)}_{j,t|t-1}, P^{(k)}_{j,t|t-1})$$

induced by (4), the observation model (6) yields the standard linear-Gaussian *measurement update* for the selected expert $j = r_t$:

$$(m^{(k)}_{r_t,t|t-1}, P^{(k)}_{r_t,t|t-1}) \longmapsto (m^{(k)}_{r_t,t|t}, P^{(k)}_{r_t,t|t})$$

using the scalar observation $\ell_{r_t,t}$. For all unselected experts $j \neq r_t$, no measurement is available at time t, so the posterior equals the prior,

$$m^{(k)}_{j,t|t} = m^{(k)}_{j,t|t-1}, \qquad P^{(k)}_{j,t|t} = P^{(k)}_{j,t|t-1}.$$

Consequently, for experts that are rarely queried, the uncertainty encoded in $P^{(k)}_{j,t|t}$ increases over time according to the process noise covariance Q_k in the state dynamics (4), reflecting growing epistemic uncertainty about their current reliability.

3 Inference: Interacting Multiple Model (IMM) Filter

Exact Bayesian filtering in a switching linear dynamical system is intractable in general, since the number of regime histories $(z_1, \dots, z_t)$ grows exponentially in t (on the order of M^t for M regimes). We therefore resort to the *Interacting Multiple Model* (IMM) approximation, which maintains, at each time t, a finite set of M Gaussian filters, one per regime $k \in \mathcal{Z}$, and performs a moment-matching "interaction" between them at every step.

For filtering it is convenient to condition on the full interaction history up to time t, including past decisions and observed losses:

$$\mathcal{I}_t \coloneqq \big((\mathbf{x}_\tau, \mathcal{K}_\tau, F_\tau)\big)_{1 \le \tau \le t}, \qquad F_\tau = (r_\tau, \ell_{r_\tau,\tau}).$$

3.1 Belief State Representation

At the end of time t (after having observed $\ell_{r_t,t}$), the IMM filter represents the joint posterior over $(z_t, \boldsymbol{\alpha}_{j,t})$ in the following approximate form:

1. **Regime Probabilities.** The posterior probability of being in regime k:

$$b_t(k) \;\coloneqq\; \mathbb{P}(z_t = k \mid \mathcal{I}_t), \qquad k \in \mathcal{Z}.$$

2. **Regime-Conditional Expert States.** For each expert $j \in \mathcal{U}$ and each regime $k \in \mathcal{Z}$, a Gaussian approximation to the conditional distribution of $\boldsymbol{\alpha}_{j,t}$:

$$\boldsymbol{\alpha}_{j,t} \,\big|\, \{z_t = k, \mathcal{I}_t\} \;\approx\; \mathcal{N}(m^{(k)}_{j,t|t},\, P^{(k)}_{j,t|t}),$$

   where $m^{(k)}_{j,t|t} \in \mathbb{R}^{d_\alpha}$ and $P^{(k)}_{j,t|t} \in \mathbb{S}^{d_\alpha}_{++}$ denote, respectively, the regime-conditional mean and covariance at time t.

Thus the overall posterior at time t is approximated as a finite mixture

$$\mathbb{P}(\boldsymbol{\alpha}_{j,t}, z_t \mid \mathcal{I}_t) \;\approx\; \sum_{k=1}^{M} b_t(k)\, \mathcal{N}(\boldsymbol{\alpha}_{j,t}; m^{(k)}_{j,t|t}, P^{(k)}_{j,t|t}) \otimes \delta_k(z_t),$$

for each $j \in \mathcal{U}$.

3.2 Step 1: Interaction (Mixing of Regime-Conditional Estimates)

The IMM "interaction" step constructs, for each candidate regime k at time $t+1$, a *mixed* Gaussian initial condition by combining the regime-conditional estimates at time t using suitable mixing weights.

Mixing Weights. For a fixed $k \in \mathcal{Z}$, define the conditional mixing probability

$$\mu_{i|k} \coloneqq \mathbb{P}(z_t = i \mid z_{t+1} = k, \mathcal{I}_t), \qquad i \in \mathcal{Z}.$$

Proposition 1 (Mixing weights)**.** For each $k \in \mathcal{Z}$ and $i \in \mathcal{Z}$,

$$\mu_{i|k} = \frac{\Pi_{ik}\, b_t(i)}{\bar{c}_k}, \qquad \bar{c}_k \coloneqq \sum_{l=1}^{M} \Pi_{lk} b_t(l), \tag{7}$$

where Π is the regime transition matrix and $\bar{c}_k = \mathbb{P}(z_{t+1} = k \mid \mathcal{I}_t)$ is the predicted probability of regime k at time $t+1$.

Proof. By Bayes' rule and the Markov property of $(z_t)_{t\geq 1}$,

$$\begin{aligned}
\mu_{i|k} &= \frac{\mathbb{P}(z_{t+1} = k \mid z_t = i, \mathcal{I}_t)\, \mathbb{P}(z_t = i \mid \mathcal{I}_t)}{\mathbb{P}(z_{t+1} = k \mid \mathcal{I}_t)} \\
&= \frac{\Pi_{ik}\, b_t(i)}{\sum_{l=1}^{M} \mathbb{P}(z_{t+1} = k \mid z_t = l)\, \mathbb{P}(z_t = l \mid \mathcal{I}_t)} \\
&= \frac{\Pi_{ik}\, b_t(i)}{\sum_{l=1}^{M} \Pi_{lk}\, b_t(l)},
\end{aligned}$$

which is precisely (7). □

Moment Matching. Fix an expert $j \in \mathcal{U}$ and a candidate regime $k \in \mathcal{Z}$ at time $t+1$. The exact conditional distribution of $\boldsymbol{\alpha}_{j,t}$ given $\{z_{t+1} = k, \mathcal{I}_t\}$ is a mixture of the M Gaussians indexed by $i \in \mathcal{Z}$, with weights $\mu_{i|k}$. The IMM approximation replaces this mixture by a single Gaussian whose first two moments match those of the mixture.

1. **Mixed mean.** Define the mixed mean for expert j and next regime k by

$$m_{j,t|t}^{0,(k)} \coloneqq \sum_{i=1}^{M} \mu_{i|k}\, m_{j,t|t}^{(i)}. \tag{8}$$

2. **Mixed covariance.** The corresponding mixed covariance incorporates both the within-regime variances and the dispersion of the regime-specific means:

$$P_{j,t|t}^{0,(k)} \coloneqq \sum_{i=1}^{M} \mu_{i|k} \Big[P_{j,t|t}^{(i)} + \big(m_{j,t|t}^{(i)} - m_{j,t|t}^{0,(k)}\big)\big(m_{j,t|t}^{(i)} - m_{j,t|t}^{0,(k)}\big)^{\top} \Big]. \tag{9}$$

By construction, $m_{j,t|t}^{0,(k)}$ and $P_{j,t|t}^{0,(k)}$ are, respectively, the first and second conditional moments of the exact mixture $\sum_i \mu_{i|k}\, \mathcal{N}(\cdot; m_{j,t|t}^{(i)}, P_{j,t|t}^{(i)})$.

Intuition. If the regime-specific means $m_{j,t|t}^{(i)}$ are widely spread, the "spread-of-means" term in (9) is large, so the mixed covariance $P_{j,t|t}^{0,(k)}$ reflects substantial uncertainty due to ambiguity about the current regime.

3.3 Step 2: Time Update (Prediction)

Given the mixed initial condition $(m_{j,t|t}^{0,(k)}, P_{j,t|t}^{0,(k)})$ for each $k \in \mathcal{Z}$, the regime-matched Kalman filter performs the linear-Gaussian prediction from t to $t+1$ using the dynamics (4). For each expert $j \in \mathcal{U}$ and regime $k \in \mathcal{Z}$,

$$m_{j,t+1|t}^{(k)} = A_k\, m_{j,t|t}^{0,(k)}, \tag{10}$$

$$P_{j,t+1|t}^{(k)} = A_k\, P_{j,t|t}^{0,(k)} A_k^{\top} + Q_k. \tag{11}$$

This prediction step is performed for all experts $j \in \mathcal{U}$, independently of whether expert j was queried at time t or is available at time $t+1$.

3.4 Step 3: Measurement Update (Correction under Partial Feedback)

At time $t+1$, the router selects a single expert index $r_{t+1} \in \mathcal{K}_{t+1}$ and subsequently observes the corresponding loss $\ell_{r_{t+1},t+1}$. The observation model (6) gives

$$\ell_{r_{t+1},t+1} \;=\; \phi_{t+1}^{\top}\boldsymbol{\alpha}_{r_{t+1},t+1} + v_{r_{t+1},t+1}, \qquad v_{r_{t+1},t+1} \sim \mathcal{N}(0, R_{k,r_{t+1}}),$$

conditional on $z_{t+1} = k$.

For each regime $k \in \mathcal{Z}$ we define the predicted observation mean and innovation variance:

$$\hat{\ell}_{t+1}^{(k)} \coloneqq \phi_{t+1}^{\top} m_{r_{t+1},t+1|t}^{(k)},$$
$$S_{t+1}^{(k)} \coloneqq \phi_{t+1}^{\top} P_{r_{t+1},t+1|t}^{(k)} \phi_{t+1} + R_{k,r_{t+1}}.$$

The innovation (residual) is

$$e_{t+1}^{(k)} \coloneqq \ell_{r_{t+1},t+1} - \hat{\ell}_{t+1}^{(k)}.$$

The scalar-output Kalman update for expert r_{t+1} and regime k is then

$$\text{Kalman gain:} \quad K_{t+1}^{(k)} = P_{r_{t+1},t+1|t}^{(k)} \phi_{t+1} (S_{t+1}^{(k)})^{-1}, \tag{12}$$

$$\text{Updated mean:} \quad m_{r_{t+1},t+1|t+1}^{(k)} = m_{r_{t+1},t+1|t}^{(k)} + K_{t+1}^{(k)} e_{t+1}^{(k)}, \tag{13}$$

$$\text{Updated covariance:} \quad P_{r_{t+1},t+1|t+1}^{(k)} = (I - K_{t+1}^{(k)} \phi_{t+1}^{\top}) P_{r_{t+1},t+1|t}^{(k)}. \tag{14}$$

For all unselected experts $j \neq r_{t+1}$, no observation is available at time $t+1$, so the measurement update is skipped:

$$m_{j,t+1|t+1}^{(k)} = m_{j,t+1|t}^{(k)}, \qquad P_{j,t+1|t+1}^{(k)} = P_{j,t+1|t}^{(k)}.$$

3.5 Step 4: Regime Probability Update

Finally, we update the regime probabilities using the likelihood of the observed loss under each regime hypothesis. For $k \in \mathcal{Z}$, define the (scalar) Gaussian likelihood

$$\Lambda_{t+1}^{(k)} \;\coloneqq\; \mathcal{N}(\ell_{r_{t+1},t+1}; \hat{\ell}_{t+1}^{(k)}, S_{t+1}^{(k)}), \tag{15}$$

i.e., the density of a normal distribution with mean $\hat{\ell}_{t+1}^{(k)}$ and variance $S_{t+1}^{(k)}$ evaluated at $\ell_{r_{t+1},t+1}$.

We have already defined the predicted regime probabilities

$$\bar{c}_k = \mathbb{P}(z_{t+1} = k \mid \mathcal{I}_t) = \sum_{l=1}^{M} \Pi_{lk} b_t(l).$$

Bayes' rule then yields the updated regime probabilities at time $t+1$:

$$b_{t+1}(k) \;\coloneqq\; \mathbb{P}(z_{t+1} = k \mid \mathcal{I}_{t+1}) \;=\; \frac{\Lambda_{t+1}^{(k)}\, \bar{c}_k}{\sum_{l=1}^{M} \Lambda_{t+1}^{(l)}\, \bar{c}_l}, \qquad k \in \mathcal{Z}, \tag{16}$$

where $\mathcal{I}_{t+1}$ is the σ-field generated by the extended history $\mathcal{I}_{t+1}$.

Intuition. If the observed loss $\ell_{r_{t+1},t+1}$ is highly improbable under the regime-k filter (i.e., $\Lambda_{t+1}^{(k)}$ is small), then $b_{t+1}(k)$ is decreased relative to its prediction $\bar{c}_k$. Conversely, regimes under which the observation is well explained (large $\Lambda_{t+1}^{(k)}$) gain posterior probability.

4 Selection Policy

At the end of time t, after running the IMM filter and updating the belief state with the observed loss $\ell_{r_t,t}$, the router must choose an expert $r_{t+1} \in \mathcal{K}_{t+1}$ for the next instance at time $t+1$. Recall that the one-step total cost for expert j is

$$C_{j,t+1} \coloneqq \ell_{j,t+1} + \beta_j.$$

A purely risk-neutral, myopic objective would be to choose r_{t+1} so as to minimize the conditional expectation

$$\mathbb{E}[C_{r_{t+1},t+1} \mid \mathcal{I}_t].$$

In practice, we will consider a one-step *risk-sensitive* criterion that also depends on the conditional variance of $C_{j,t+1}$ given $\mathcal{I}_t$, defined below.

4.1 MMSE Forecast (Predictive Mean)

Fix an expert $j \in \mathcal{U}$. Conditioned on $\mathcal{I}_t$, the latent regime z_{t+1} is distributed according to the one-step-ahead regime probabilities

$$b_{t+1|t}(k) \;\coloneqq\; \mathbb{P}(z_{t+1} = k \mid \mathcal{I}_t), \qquad k \in \mathcal{Z},$$

which, in the IMM recursion, coincide with the predicted regime probabilities $\bar{c}_k$ from Step 4 of the filter.

For each regime $k \in \mathcal{Z}$, the IMM prediction step provides

$$\boldsymbol{\alpha}_{j,t+1} \,|\, \{z_{t+1} = k, \mathcal{I}_t\} \;\approx\; \mathcal{N}(m^{(k)}_{j,t+1|t},\, P^{(k)}_{j,t+1|t}),$$

and the observation model (6) yields

$$\ell_{j,t+1} \,|\, \{z_{t+1} = k, \mathcal{I}_t\} \;\sim\; \mathcal{N}(\mu^{(k)}_{j,t+1|t},\, S^{(k)}_{j,t+1|t}),$$

with

$$\begin{aligned} \mu^{(k)}_{j,t+1|t} &\coloneqq \phi_{t+1}^\top m^{(k)}_{j,t+1|t}, \\ S^{(k)}_{j,t+1|t} &\coloneqq \phi_{t+1}^\top P^{(k)}_{j,t+1|t} \phi_{t+1} + R_{k,j}. \end{aligned}$$

The (approximate) predictive distribution of $\ell_{j,t+1}$ given $\mathcal{I}_t$ is therefore a finite mixture of Gaussians:

$$\mathbb{P}(\ell_{j,t+1} \mid \mathcal{I}_t) \;\approx\; \sum_{k=1}^{M} b_{t+1|t}(k)\, \mathcal{N}(\cdot;\, \mu^{(k)}_{j,t+1|t},\, S^{(k)}_{j,t+1|t}).$$

Under squared loss for predicting $\ell_{j,t+1}$, the optimal point predictor is the conditional expectation (MMSE forecast)

$$\widehat{\mathcal{R}}_{j,t+1|t} \;\coloneqq\; \mathbb{E}[\ell_{j,t+1} \mid \mathcal{I}_t] \;=\; \sum_{k=1}^{M} b_{t+1|t}(k)\, \mu^{(k)}_{j,t+1|t}.$$

Interpretation. The quantity $\widehat{\mathcal{R}}_{j,t+1|t}$ is the regime-averaged predictive loss of expert j at time $t+1$: it combines the predicted reliability state (the regime-conditional means $m^{(k)}_{j,t+1|t}$) with the probability $b_{t+1|t}(k)$ of each regime k being active at time $t+1$.

4.2 Predictive Variance and Epistemic Uncertainty

We now quantify the overall uncertainty associated with expert j's predicted loss via the conditional variance

$$\hat{\sigma}^2_{j,t+1|t} \;\coloneqq\; \mathrm{Var}(\ell_{j,t+1} \mid \mathcal{I}_t).$$

This variance aggregates both *aleatoric* uncertainty (observation noise and within-regime state uncertainty) and *epistemic* uncertainty due to ambiguity over the regime. Since β_j is deterministic, we have

$$\mathrm{Var}(C_{j,t+1} \mid \mathcal{I}_t) = \mathrm{Var}(\ell_{j,t+1} \mid \mathcal{I}_t) = \hat{\sigma}^2_{j,t+1|t}.$$

Proposition 2 (Predictive variance decomposition)**.** For each expert $j \in \mathcal{U}$ and time t,

$$\hat{\sigma}^2_{j,t+1|t} \;=\; \sum_{k=1}^{M} b_{t+1|t}(k)\, S^{(k)}_{j,t+1|t} \;+\; \sum_{k=1}^{M} b_{t+1|t}(k)\, \big(\mu^{(k)}_{j,t+1|t} - \widehat{\mathcal{R}}_{j,t+1|t}\big)^2. \tag{17}$$

Proof. Apply the law of total variance with $Y = \ell_{j,t+1}$ and $Z = z_{t+1}$:

$$\mathrm{Var}(Y \mid \mathcal{I}_t) = \mathbb{E}\big[\mathrm{Var}(Y \mid Z, \mathcal{I}_t) \mid \mathcal{I}_t\big] + \mathrm{Var}\big(\mathbb{E}[Y \mid Z, \mathcal{I}_t] \mid \mathcal{I}_t\big).$$

Conditionally on $\{z_{t+1} = k, \mathcal{I}_t\}$ we have

$$\mathbb{E}[\ell_{j,t+1} \mid z_{t+1} = k, \mathcal{I}_t] = \mu^{(k)}_{j,t+1|t}, \qquad \mathrm{Var}(\ell_{j,t+1} \mid z_{t+1} = k, \mathcal{I}_t) = S^{(k)}_{j,t+1|t},$$

and $\mathbb{P}(z_{t+1} = k \mid \mathcal{I}_t) = b_{t+1|t}(k)$. Therefore,

$$\mathbb{E}\big[\mathrm{Var}(Y \mid Z, \mathcal{I}_t) \mid \mathcal{I}_t\big] = \sum_{k=1}^{M} b_{t+1|t}(k)\, S^{(k)}_{j,t+1|t},$$

and

$$\mathrm{Var}\big(\mathbb{E}[Y \mid Z, \mathcal{I}_t] \mid \mathcal{I}_t\big) = \sum_{k=1}^{M} b_{t+1|t}(k)\, \big(\mu^{(k)}_{j,t+1|t} - \widehat{\mathcal{R}}_{j,t+1|t}\big)^2,$$

since $\widehat{\mathcal{R}}_{j,t+1|t} = \sum_k b_{t+1|t}(k)\, \mu^{(k)}_{j,t+1|t}$. Summing the two terms yields (17). □

Interpretation.

- The first term,
$$\sum_{k=1}^{M} b_{t+1|t}(k)\, S^{(k)}_{j,t+1|t},$$
averages the *within-regime* variances. Each $S^{(k)}_{j,t+1|t}$ captures both the uncertainty in the latent state (through $P^{(k)}_{j,t+1|t}$) and the irreducible observation noise $R_{k,j}$.
- The second term,
$$\sum_{k=1}^{M} b_{t+1|t}(k)\, \big(\mu^{(k)}_{j,t+1|t} - \widehat{\mathcal{R}}_{j,t+1|t}\big)^2,$$
measures the dispersion of the regime-specific predictive means around the overall mean. It quantifies *epistemic* uncertainty due to regime ambiguity: if different regimes predict very different losses for expert j, this term is large.

4.3 Cost-Sensitive Selection Rule

Using the predictive mean and variance, we define a myopic, risk-adjusted selection rule for the next decision. For each available expert $j \in \mathcal{K}_{t+1}$, consider the score

$$J_{j,t+1} \;\coloneqq\; \widehat{\mathcal{R}}_{j,t+1|t} + \beta_j + \lambda\sqrt{\hat{\sigma}^2_{j,t+1|t}},$$

where $\lambda \in \mathbb{R}$ is a user-specified risk parameter. The router then selects

$$r^*_{t+1} \;\in\; \arg\min_{j \in \mathcal{K}_{t+1}} J_{j,t+1}. \tag{18}$$

Operational interpretation.

(i) $\widehat{\mathcal{R}}_{j,t+1|t}$ (**expected loss**) captures pure exploitation: lower values correspond to experts that are predicted to incur smaller errors under the current belief.

(ii) β_j (**consultation cost**) accounts for economic or operational constraints (e.g., latency, API fees, computational cost). An expert with a large β_j must offer a correspondingly lower expected loss to be competitive.

(iii) $\lambda\sqrt{\hat{\sigma}^2_{j,t+1|t}}$ (**risk/uncertainty term**) controls how the policy reacts to predictive uncertainty:

- If $\lambda > 0$ (risk-averse regime), the policy penalizes experts with high predictive variance, favouring stable, well-understood experts.
- If $\lambda < 0$ (exploratory or "optimistic" regime), the policy is encouraged to select experts with high uncertainty, in line with exploration principles: querying such experts yields information that reduces $\hat{\sigma}^2_{j,t+1|t}$ in subsequent steps.

The rule (18) is myopic: it minimizes a one-step, risk-adjusted proxy for the cumulative objective (2). It is not, in general, globally Bayes-optimal over the horizon $\{1, \ldots, T\}$, but it is computationally tractable and directly exploits the IMM-based predictive mean and variance.

Remark 2 (One-step risk-adjusted Bayes action)**.** Fix t and condition on the history $\mathcal{I}_t$. For each expert $j \in \mathcal{K}_{t+1}$,

$$C_{j,t+1} = \ell_{j,t+1} + \beta_j$$

denotes the total next-step cost, with

$$\mathbb{E}[C_{j,t+1} \mid \mathcal{I}_t] = \widehat{\mathcal{R}}_{j,t+1|t} + \beta_j, \qquad \mathrm{Var}(C_{j,t+1} \mid \mathcal{I}_t) = \hat{\sigma}^2_{j,t+1|t}.$$

Define the local risk functional

$$\rho_t(Y) \coloneqq \mathbb{E}[Y \mid \mathcal{I}_t] + \lambda\sqrt{\mathrm{Var}(Y \mid \mathcal{I}_t)},$$

for some parameter $\lambda \in \mathbb{R}$. Then the selection rule (18) is exactly

$$r^*_{t+1} \in \arg\min_{j \in \mathcal{K}_{t+1}} \rho_t(C_{j,t+1}),$$

i.e., a Bayes action for the one-step risk functional ρ_t.

5 Parameter Optimization

The parameters of the SLDS model are collected in

$$\Theta \;\coloneqq\; \Big(\Pi,\; (A_k, Q_k)_{k=1}^{M},\; (R_{k,j})_{k=1,\dots,M;\; j\in\mathcal{U}},\; \text{initial priors}\Big),$$

where:

- Π is the $M \times M$ regime transition matrix;
- $A_k \in \mathbb{R}^{d_\alpha \times d_\alpha}$ and $Q_k \in \mathbb{S}_{++}^{d_\alpha}$ are the regime-dependent state transition and process noise matrices;
- $R_{k,j} > 0$ is the observation noise variance for expert j in regime k;
- the initial priors comprise the initial regime distribution and the Gaussian priors on the latent states (e.g. the population prior $(\mu_{\text{pop}}, \Sigma_{\text{pop}})$).

The feature map $\phi : \mathcal{X} \to \mathbb{R}^{d_\alpha}$ is treated as fixed and not learned.

We observe a single scalar loss per time step, corresponding to the selected expert. The data are

$$\mathbf{D} \;\coloneqq\; \{(\mathbf{x}_t, r_t, \ell_{r_t,t}, \mathcal{K}_t)\}_{t=1}^{T},$$

and, for fixed Θ, the latent variables are the regime sequence $(z_t)_{t=1}^{T}$ and the latent reliability states $(\boldsymbol{\alpha}_{j,t})_{j\in\mathcal{U},\, 1\le t\le T}$.

5.1 Maximum Likelihood Objective

Given that the routing decisions (r_t) and availability sets $(\mathcal{K}_t)$ are treated as exogenous, we seek a maximum likelihood estimate of Θ based on the conditional marginal likelihood of the observed losses given the contexts and actions. Writing

$$\mathcal{I}_t = \big((\mathbf{x}_\tau, \mathcal{K}_\tau, F_\tau)\big)_{1\le\tau\le t}, \qquad F_\tau = (r_\tau, \ell_{r_\tau,\tau}),$$

the incremental likelihood at time t is

$$p_\Theta(\ell_{r_t,t} \mid \mathcal{I}_{t-1}) \;=\; \mathbb{P}_\Theta(\ell_{r_t,t} \mid \mathcal{I}_{t-1}),$$

and the (conditional) log-likelihood over the horizon $\{1, \dots, T\}$ is

$$\mathcal{L}(\Theta) \;\coloneqq\; \sum_{t=1}^{T} \log p_\Theta(\ell_{r_t,t} \mid \mathcal{I}_{t-1}). \tag{19}$$

Maximizing (19) over Θ yields a maximum likelihood estimate conditional on the observed contexts, actions, and availability patterns.

Likelihood term. For each t, the quantity $p_\Theta(\ell_{r_t,t} \mid \mathcal{I}_{t-1})$ can be expressed by marginalizing over the latent regime z_t:

$$p_\Theta(\ell_{r_t,t} \mid \mathcal{I}_{t-1}) = \sum_{k=1}^{M} \underbrace{p_\Theta(\ell_{r_t,t} \mid z_t = k, \mathcal{I}_{t-1})}_{\Lambda_t^{(k)}} \underbrace{\mathbb{P}_\Theta(z_t = k \mid \mathcal{I}_{t-1})}_{\bar{c}_t(k)}. \tag{20}$$

Here:

- $\bar{c}_t(k) \coloneqq \mathbb{P}_\Theta(z_t = k \mid \mathcal{I}_{t-1}) = \sum_{i=1}^{M} \Pi_{ik} b_{t-1}(i)$ is the *predicted* probability of regime k at time t, as in the IMM recursion;

- $\Lambda_t^{(k)}$ is the scalar Gaussian likelihood of the observation $\ell_{r_t,t}$ under the regime-k Kalman filter at time t:

$$\Lambda_t^{(k)} = \mathcal{N}(\ell_{r_t,t}; \mu_{r_t,t|t-1}^{(k)}, S_{r_t,t|t-1}^{(k)}),$$

  where $\mu_{r_t,t|t-1}^{(k)}$ and $S_{r_t,t|t-1}^{(k)}$ are the predicted mean and variance for expert r_t in regime k at time t (as in the IMM filter).

5.2 EM Algorithm for SLDS

Direct maximization of (19) is intractable due to the latent regime sequence and continuous states. We therefore employ an Expectation–Maximization (EM) procedure, using the IMM filter and an associated smoother to approximate the posterior over $(z_t, \boldsymbol{\alpha}_{j,t})_{j,t}$.

Let $Z_{1:T}$ denote the regime sequence and $\boldsymbol{\alpha}_{1:T}$ the collection of all latent states. The EM algorithm iteratively maximizes the expectation of the complete-data log-likelihood

$$\log p_\Theta(\mathbf{D}, Z_{1:T}, \boldsymbol{\alpha}_{1:T})$$

with respect to a posterior approximation $q(Z_{1:T}, \boldsymbol{\alpha}_{1:T})$ induced by the current parameter iterate. Because the IMM posterior is itself an approximation, the resulting scheme is a generalized EM algorithm (it monotonically increases an approximate likelihood surrogate).

Algorithm 1 Cost-Sensitive SLDS Training (Generalized EM with IMM)

Input: data $\mathbf{D} = \{(\mathbf{x}_t, \ell_{r_t,t}, r_t, \mathcal{K}_t)\}_{t=1}^{T}$.

Initialize parameters $\Theta^{(0)}$ (e.g., Π nearly uniform, A_k close to identity, Q_k small, $R_{k,j}$ moderate, and reasonable initial priors).

repeat

E-step (IMM smoothing). Given $\Theta^{(m)}$, run the IMM forward filter and a backward smoother to obtain approximate posterior quantities:

- regime marginals:

$$\gamma_t^{(k)} \;\approx\; \mathbb{P}_{\Theta^{(m)}}(z_t = k \mid \mathcal{I}_T), \qquad 1 \le t \le T,\ k \in \mathcal{Z};$$

- pairwise regime marginals (for transitions):

$$\xi_t^{(i,k)} \;\approx\; \mathbb{P}_{\Theta^{(m)}}(z_{t-1} = i, z_t = k \mid \mathcal{I}_T), \qquad 2 \le t \le T,\ i,k \in \mathcal{Z};$$

- regime-conditional smoothed state statistics for each expert j:

$$\boldsymbol{\alpha}_{j,t} \mid \{z_t = k, \mathcal{I}_T\} \approx \mathcal{N}\big(m_{j,t|T}^{(k)}, P_{j,t|T}^{(k)}\big),$$

together with cross-covariances $P_{j,t,t-1|T}^{(k)} \approx \mathbb{E}[\boldsymbol{\alpha}_{j,t}\boldsymbol{\alpha}_{j,t-1}^{\top} \mid z_t = k, \mathcal{I}_T] - m_{j,t|T}^{(k)} m_{j,t-1|T}^{(k)\top}$ when needed.

M-step (parameter update). Update Θ by maximizing the expected complete-data log-likelihood under the approximate posterior:

$$\Theta^{(m+1)} \;\in\; \arg\max_{\Theta} \mathbb{E}_{q^{(m)}}\big[\log p_\Theta(\mathbf{D}, Z_{1:T}, \boldsymbol{\alpha}_{1:T})\big],$$

where $q^{(m)}$ denotes the IMM-based posterior approximation from the E-step. This yields closed-form updates of the following form:

- **Regime transitions** Π**.** For each $i, k \in \mathcal{Z}$,

$$\Pi_{ik}^{(m+1)} = \frac{\sum_{t=2}^{T} \xi_t^{(i,k)}}{\sum_{t=2}^{T} \sum_{k'=1}^{M} \xi_t^{(i,k')}}.$$

- **State dynamics** (A_k, Q_k)**.** For each regime $k \in \mathcal{Z}$, estimate A_k and Q_k by (approximately) solving a weighted least-squares problem over the latent transitions $\boldsymbol{\alpha}_{j,t-1} \mapsto \boldsymbol{\alpha}_{j,t}$, pooling all experts $j \in \mathcal{U}$ and times t with weights $\gamma_t^{(k)}$:

$$A_k^{(m+1)} \approx \arg\min_{A} \sum_{j \in \mathcal{U}} \sum_{t=2}^{T} \gamma_t^{(k)}\, \mathbb{E}\big[\|\boldsymbol{\alpha}_{j,t} - A\boldsymbol{\alpha}_{j,t-1}\|_2^2 \,\big|\, z_t = k, \mathcal{I}_T\big],$$

and $Q_k^{(m+1)}$ as the corresponding weighted empirical covariance of the residuals $\boldsymbol{\alpha}_{j,t} - A_k^{(m+1)}\boldsymbol{\alpha}_{j,t-1}$.

- **Observation noise** $R_{k,j}$**.** For each (k, j), update $R_{k,j}$ using the observed losses $\ell_{r_t,t}$ whenever expert j was selected ($r_t = j$), weighted by the regime posteriors $\gamma_t^{(k)}$:

$$R_{k,j}^{(m+1)} = \frac{\sum_{t: r_t = j} \gamma_t^{(k)}\, \mathbb{E}\big[\big(\ell_{j,t} - \phi_t^{\top}\boldsymbol{\alpha}_{j,t}\big)^2 \,\big|\, z_t = k, \mathcal{I}_T\big]}{\sum_{t: r_t = j} \gamma_t^{(k)}},$$

where, under the Gaussian approximation,

$$\mathbb{E}\big[(\ell_{j,t} - \phi_t^{\top}\boldsymbol{\alpha}_{j,t})^2 \mid z_t = k, \mathcal{I}_T\big] = (\ell_{r_t,t})^2 - 2\ell_{r_t,t}\,\phi_t^{\top} m_{j,t|T}^{(k)} + \phi_t^{\top}\big(P_{j,t|T}^{(k)} + m_{j,t|T}^{(k)} m_{j,t|T}^{(k)\top}\big)\phi_t.$$

(For t with $r_t \neq j$, no contribution appears because $\ell_{j,t}$ is unobserved.)

until convergence of $\mathcal{L}(\Theta)$ or of the parameter iterates.

Intuition.

- **E-step (attribution).** The E-step approximates, for each time t, how much of the observed loss $\ell_{r_t,t}$ and the latent state evolution should be attributed to each regime k. This produces soft counts for regime occupancy ($\gamma_t^{(k)}$) and transitions ($\xi_t^{(i,k)}$), as well as smoothed trajectories of the latent states under each regime.
- **M-step (recalibration).** The M-step then recalibrates the parameters $(\Pi, A_k, Q_k, R_{k,j})$ so that, under the current posterior attribution, the linear dynamics and noise statistics best explain the observed sequence of losses. For example, periods where $\gamma_t^{(k)}$ is large exert more influence on (A_k, Q_k), so a "volatile" regime will learn larger process noise Q_k if the data exhibit rapid changes during those times.

6 Extension: Joint Subset Selection (Top-K)

In some applications, the router may wish to consult not just a single expert, but a subset

$$S_t \subset \mathcal{K}_t, \qquad |S_t| = K,$$

for the current instance at time t (for instance, to form an ensemble prediction). A naive strategy would select the K experts with smallest individual scores (e.g. lowest predicted loss plus cost), treating them as independent. This ignores *correlation*: if two experts have highly correlated reliability states, querying both may incur redundant cost without commensurate information gain.

We describe an extension of the SLDS framework that models such correlations and a greedy Top-K selection rule that penalizes redundant queries.

6.1 Joint state-space formulation

To capture correlated dynamics, we collect the latent reliability states of all experts into a single vector. Let d_α denote the dimension of the per-expert state, and define

$$\boldsymbol{\alpha}_t \coloneqq \begin{bmatrix} \boldsymbol{\alpha}_{1,t}^\top & \cdots & \boldsymbol{\alpha}_{N,t}^\top \end{bmatrix}^\top \in \mathbb{R}^{Nd_\alpha}.$$

For each regime $k \in \mathcal{Z}$, we define regime-dependent joint dynamics

$$\boldsymbol{\alpha}_{t+1} = A_k^{(\text{joint})} \boldsymbol{\alpha}_t + \mathbf{w}_t^{(k)}, \qquad \mathbf{w}_t^{(k)} \sim \mathcal{N}(0, Q_k^{(\text{joint})}), \tag{21}$$

where:

- $A_k^{(\text{joint})} \in \mathbb{R}^{Nd_\alpha \times Nd_\alpha}$ is a joint transition matrix. A natural choice consistent with the single-expert model is
$$A_k^{(\text{joint})} = I_N \otimes A_k,$$
i.e. the same A_k acts independently on each expert's state, but more general couplings are allowed.
- $Q_k^{(\text{joint})} \in \mathbb{S}_{++}^{Nd_\alpha}$ is a full joint process noise covariance. Its block structure encodes cross-expert correlations: the (i,j) block $(Q_k^{(\text{joint})})_{ij} \in \mathbb{R}^{d_\alpha \times d_\alpha}$ models how shocks affecting expert i co-vary with shocks affecting expert j under regime k. The independent case corresponds to $Q_k^{(\text{joint})}$ block-diagonal.

The regime process (z_t) and its transition matrix Π remain unchanged. In the full SLDS, we maintain regime-conditional joint Gaussians for $\boldsymbol{\alpha}_t$ and combine them via IMM as in the single-expert case.

6.2 Subset observation model

At a fixed time t, the (hypothetical) loss of each expert $j \in \mathcal{U}$ is

$$\ell_{j,t} = \phi_t^\top \boldsymbol{\alpha}_{j,t} + v_{j,t}, \qquad v_{j,t} \sim \mathcal{N}(0, R_{k,j})$$

conditional on regime $z_t = k$, where $\phi_t = \phi(\mathbf{x}_t)$ is the feature vector. In joint form, define the full loss vector

$$\boldsymbol{\ell}_t \coloneqq \begin{bmatrix} \ell_{1,t} & \cdots & \ell_{N,t} \end{bmatrix}^\top \in \mathbb{R}^N.$$

Then, conditional on $z_t = k$ and $\mathcal{I}_{t-1}$,

$$\boldsymbol{\ell}_t = H_t \boldsymbol{\alpha}_t + \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}(0, R_k), \tag{22}$$

where

- $H_t \in \mathbb{R}^{N \times N d_\alpha}$ is block-diagonal with row j equal to

  $$[0, \ldots, 0,\ \phi_t^\top,\ 0, \ldots, 0],$$

  i.e. $\phi_t^\top$ in the block corresponding to expert j and zeros elsewhere;
- $R_k = \mathrm{diag}(R_{k,1}, \ldots, R_{k,N})$ is the diagonal observation noise covariance.

If, at time t, we select a subset $S_t \subset \mathcal{K}_t$ of size K, we observe only the corresponding entries of $\boldsymbol{\ell}_t$. Let $P_{S_t} \in \{0,1\}^{K \times N}$ denote the row selection matrix that picks coordinates in S_t, and define

$$\mathbf{y}_{S_t,t} \coloneqq P_{S_t} \boldsymbol{\ell}_t \in \mathbb{R}^K.$$

Then

$$\mathbf{y}_{S_t,t} = H_{S_t,t} \boldsymbol{\alpha}_t + \mathbf{v}_{S_t,t},$$

where $H_{S_t,t} \coloneqq P_{S_t} H_t$ and $\mathbf{v}_{S_t,t} \sim \mathcal{N}(0, R_{S_t,k})$ with $R_{S_t,k} \coloneqq P_{S_t} R_k P_{S_t}^\top$.

6.3 Greedy sequential Top-K selection

At time t, before observing any losses, the IMM filter (with moment-matching over regimes) yields an approximate Gaussian predictive distribution for the full loss vector

$$\boldsymbol{\ell}_t \mid \mathcal{I}_{t-1} \ \approx\ \mathcal{N}(\widehat{\boldsymbol{\mathcal{R}}}_{t|t-1},\, \Sigma_{t|t-1}),$$

where:

- $\widehat{\boldsymbol{\mathcal{R}}}_{t|t-1} \in \mathbb{R}^N$ collects the one-step-ahead MMSE forecasts

  $$\left(\widehat{\boldsymbol{\mathcal{R}}}_{t|t-1}\right)_j = \widehat{\mathcal{R}}_{j,t|t-1} = \mathbb{E}[\ell_{j,t} \mid \mathcal{I}_{t-1}],$$

  obtained as in the single-expert case;
- $\Sigma_{t|t-1} \in \mathbb{S}_{++}^N$ is the predictive covariance matrix

  $$\Sigma_{t|t-1} = \mathrm{Cov}(\boldsymbol{\ell}_t \mid \mathcal{I}_{t-1}),$$

  whose diagonal entries are the predictive variances of each expert's loss and whose off-diagonal entries encode cross-expert correlations.

Selecting the globally optimal subset S_t of size K with respect to a risk-adjusted objective would require evaluating all $\binom{N}{K}$ subsets, which is combinatorially infeasible for large N. We therefore adopt a forward greedy strategy that, at each step, selects the expert that yields the best marginal improvement when accounting for conditional variances under Gaussian conditioning.

Let $S_t^{(0)} \coloneqq \emptyset$, and let $\Sigma_t^{(0)} \coloneqq \Sigma_{t|t-1}$. For $k = 1, \ldots, K$, perform:

1. **Scoring step.** For each candidate $j \in \mathcal{K}_t \setminus S_t^{(k-1)}$, define the risk-adjusted score

$$\mathrm{Score}_t(j \mid S_t^{(k-1)}) \;\coloneqq\; \widehat{\mathcal{R}}_{j,t|t-1} + \beta_j + \lambda\sqrt{\sigma^2_{j,t}(S_t^{(k-1)})}, \tag{23}$$

where

$$\sigma^2_{j,t}(S_t^{(k-1)}) \;\coloneqq\; \mathrm{Var}\big(\ell_{j,t} \mid \mathcal{I}_{t-1}, \{\ell_{i,t} : i \in S_t^{(k-1)}\}\big)$$

is the conditional predictive variance of expert j's loss given hypothetical observation of the losses of the experts already in $S_t^{(k-1)}$. Under the Gaussian approximation,

$$\sigma^2_{j,t}(S_t^{(k-1)}) = \big(\Sigma_t^{(k-1)}\big)_{jj},$$

i.e. the (j,j) entry of the current conditional covariance matrix.

2. **Selection step.** Choose

$$j^* \in \arg\min_{j \in \mathcal{K}_t \setminus S_t^{(k-1)}} \mathrm{Score}_t(j \mid S_t^{(k-1)}),$$

and update the selected set

$$S_t^{(k)} \;\coloneqq\; S_t^{(k-1)} \cup \{j^*\}.$$

3. **Conditional covariance update (diversity mechanism).** We now update the conditional covariance of the remaining experts' losses given the hypothetical observation of $\ell_{j^*,t}$. Since $\boldsymbol{\ell}_t \mid \mathcal{I}_{t-1} \sim \mathcal{N}(\widehat{\boldsymbol{\mathcal{R}}}_{t|t-1}, \Sigma_t^{(k-1)})$, and we condition on observing $\ell_{j^*,t}$, the conditional covariance of $\boldsymbol{\ell}_t$ (before removing the j^*-coordinate) is given by the standard Gaussian conditioning formula

$$\Sigma_t^{(k)} = \Sigma_t^{(k-1)} - \frac{\Sigma_t^{(k-1)} e_{j^*} e_{j^*}^\top \Sigma_t^{(k-1)}}{\big(\Sigma_t^{(k-1)}\big)_{j^*j^*}}, \tag{24}$$

where $e_{j^*} \in \mathbb{R}^N$ is the j^*-th canonical basis vector. Only the submatrix corresponding to the remaining indices $\mathcal{K}_t \setminus S_t^{(k)}$ is relevant for subsequent iterations, and can be extracted from $\Sigma_t^{(k)}$.

At the end of the K iterations, we output $S_t^{(K)}$ as the selected subset for time t.

Theoretical insight. The update (24) is the conditional covariance of a jointly Gaussian vector under observation of one component; it is equivalent to a Kalman update performed directly in the observation (loss) space. If the loss of a remaining candidate i is highly correlated with that of j^*, then the off-diagonal covariance $\big(\Sigma_t^{(k-1)}\big)_{ij^*}$ is large in magnitude, and the update will significantly reduce the conditional variance $\big(\Sigma_t^{(k)}\big)_{ii}$ of $\ell_{i,t}$, even though i was not selected.

Under an exploratory choice of $\lambda < 0$ in the score (23), this reduction in $\sigma^2_{i,t}(S_t^{(k)})$ makes expert i less attractive in subsequent greedy steps compared to experts whose losses remain weakly correlated with j^* and therefore retain higher conditional variance. The greedy procedure thus tends to produce subsets that are *diverse* in the sense of carrying complementary information rather than duplicating highly correlated experts.

7 Extension: Full-Information Expert Feedback

In the main development, the router observes at each time t only the loss of the selected expert r_t, leading to a censored (bandit-style) feedback model. In some applications—for instance when all experts are implemented as neural networks that can be evaluated offline—it is natural to assume a *full-information* setting in which, once the outcome y_t is revealed, the loss of *every* available expert can be computed, regardless of which expert was actually queried for decision-making.

In this section we formalize this full-information feedback model and describe the corresponding changes to the SLDS–IMM inference and to the routing problem.

7.1 Feedback model with full expert losses

We retain the generative model of Sections 2–3: a latent regime process $(z_t)_{t\geq 1}$, regime-dependent linear dynamics for the expert-specific states $(\boldsymbol{\alpha}_{j,t})_{t\geq 1,\, j\in\mathcal{U}}$, and a linear-Gaussian emission model for the loss. Recall that for each $t \geq 1$ we write

$$\phi_t \coloneqq \phi(\mathbf{x}_t) \in \mathbb{R}^{d_\alpha},$$

and, conditionally on $z_t = k$,

$$\ell_{j,t} = \phi_t^\top \boldsymbol{\alpha}_{j,t} + v_{j,t}, \qquad v_{j,t} \sim \mathcal{N}(0, R_{k,j}),$$

for every expert $j \in \mathcal{U}$.

The change concerns only the *feedback* observed after each decision. In the full-information setting we assume that, once y_t is revealed, the loss

$$\ell_{j,t} = \mathcal{L}(f_j(\mathbf{x}_t), y_t)$$

is available for all experts $j \in \mathcal{K}_t$, not just for the chosen expert r_t.

Formally, define the feedback random element at time t by

$$G_t \;\coloneqq\; (r_t,\; (\ell_{j,t})_{j\in\mathcal{K}_t}) \in \mathcal{U} \times [0,+\infty)^{|\mathcal{K}_t|}.$$

The interaction history up to time t is then

$$\mathcal{I}_t^{\text{full}} \;\coloneqq\; \big((\mathbf{x}_\tau, \mathcal{K}_\tau, G_\tau)\big)_{1\leq\tau\leq t},$$

and the information available at decision time $t+1$ is summarized by

$$H_{t+1}^{\text{full}} \;\coloneqq\; (\mathcal{I}_t^{\text{full}},\, \mathbf{x}_{t+1},\, \mathcal{K}_{t+1}).$$

In this extension, the filtration $(\mathcal{F}_t)_{t\geq 0}$ is redefined as

$$\mathcal{F}_t \;\coloneqq\; \sigma\big(H_t^{\text{full}}\big), \qquad t \geq 0,$$

so that, at each time t, $\mathcal{F}_t$ contains the full panel of losses

$$(\ell_{j,\tau})_{j\in\mathcal{K}_\tau,\, 1\leq\tau\leq t}$$

observed up to time t.

The latent process

$$\Big(z_t,\; (\boldsymbol{\alpha}_{j,t})_{j\in\mathcal{U}},\; \mathbf{x}_t\Big)_{t\geq 1}$$

is unchanged and is assumed to evolve independently of the routing decisions $(r_t)_{t\geq 1}$ (as in the base model): the router affects only the incurred cost, not the environment.

7.2 IMM belief update under full feedback

Under the full-information model, the IMM filter can exploit at each time t the observed losses of *all* experts $j \in \mathcal{K}_t$ rather than only that of the selected expert. For clarity, we keep the standing assumptions:

- conditional on the regime sequence (z_t), the processes $(\boldsymbol{\alpha}_{j,t})_{t\geq 1}$ are independent across $j \in \mathcal{U}$ and follow the regime-dependent linear dynamics;
- conditional on (z_t) and $(\boldsymbol{\alpha}_{j,t})$, the observation noises $(v_{j,t})$ are independent across experts and time.

Fix $t \geq 1$. For each regime $k \in \mathcal{Z}$ and expert $j \in \mathcal{U}$, the prediction step of the IMM filter yields the approximate Gaussian law

$$\boldsymbol{\alpha}_{j,t} \,|\, \{z_t = k,\ \mathcal{I}_{t-1}^{\text{full}}\} \ \approx\ \mathcal{N}(m_{j,t|t-1}^{(k)},\, P_{j,t|t-1}^{(k)}),$$

and the emission model gives

$$\ell_{j,t} \,|\, \{z_t = k,\ \mathcal{I}_{t-1}^{\text{full}}\} \ \sim\ \mathcal{N}(\mu_{j,t|t-1}^{(k)},\, S_{j,t|t-1}^{(k)}),$$

with

$$\begin{aligned}
\mu_{j,t|t-1}^{(k)} &\coloneqq \phi_t^\top m_{j,t|t-1}^{(k)}, \\
S_{j,t|t-1}^{(k)} &\coloneqq \phi_t^\top P_{j,t|t-1}^{(k)} \phi_t + R_{k,j}.
\end{aligned}$$

Measurement update for all experts. Given the observed loss $\ell_{j,t}$, the regime-conditional Kalman update for expert j in regime k is

$$\begin{aligned}
K_{j,t}^{(k)} &\coloneqq P_{j,t|t-1}^{(k)}\, \phi_t\, (S_{j,t|t-1}^{(k)})^{-1}, \\
m_{j,t|t}^{(k)} &= m_{j,t|t-1}^{(k)} + K_{j,t}^{(k)}\, (\ell_{j,t} - \mu_{j,t|t-1}^{(k)}), \\
P_{j,t|t}^{(k)} &= (I - K_{j,t}^{(k)}\, \phi_t^\top) P_{j,t|t-1}^{(k)}.
\end{aligned}$$

In the full-information setting, this update is performed for *every* expert $j \in \mathcal{K}_t$ (and may be skipped for $j \notin \mathcal{K}_t$ if such experts are considered dormant). Thus all experts benefit from each newly revealed outcome y_t, irrespective of whether they were selected for routing at time t.

Regime probability update. Let $b_{t-1}(i) = \mathbb{P}(z_{t-1} = i \mid \mathcal{I}_{t-1}^{\text{full}})$ denote the regime beliefs at time $t-1$, and set

$$\bar{c}_t(k) \ \coloneqq\ \mathbb{P}(z_t = k \mid \mathcal{I}_{t-1}^{\text{full}}) = \sum_{i=1}^{M} \Pi_{ik}\, b_{t-1}(i), \qquad k \in \mathcal{Z},$$

for the one-step regime predictions. Under the Gaussian approximations above and the conditional independence across experts, the joint predictive density of the loss vector

$$(\ell_{j,t})_{j\in\mathcal{K}_t} \,|\, \{z_t = k,\ \mathcal{I}_{t-1}^{\text{full}}\}$$

factorizes as

$$\prod_{j\in\mathcal{K}_t} \mathcal{N}(\ell_{j,t}; \mu_{j,t|t-1}^{(k)}, S_{j,t|t-1}^{(k)}).$$

Define the regime-specific likelihood

$$\Lambda_t^{(k)} \;:=\; \prod_{j\in\mathcal{K}_t} \mathcal{N}(\ell_{j,t};\mu_{j,t|t-1}^{(k)}, S_{j,t|t-1}^{(k)}).$$

Bayes' rule then yields the updated regime beliefs

$$b_t(k) = \mathbb{P}(z_t = k \mid \mathcal{I}_t^{\text{full}}) = \frac{\Lambda_t^{(k)}\,\bar{c}_t(k)}{\sum_{\ell=1}^{M}\Lambda_t^{(\ell)}\,\bar{c}_t(\ell)}, \qquad k\in\mathcal{Z}.$$

Interpretation. Compared to the censored-feedback case, each time step now contributes information about the relative plausibility of the regimes through the entire vector of expert losses $(\ell_{j,t})_{j\in\mathcal{K}_t}$. Regimes under which the observed pattern of losses is unlikely (for many experts simultaneously) receive smaller posterior weight, while regimes consistent with the observed multi-expert performance are reinforced.

7.3 Consequences for the routing problem

Under full-information feedback, the IMM-based predictive mean and variance for each expert j at time $t+1$,

$$\widehat{\mathcal{R}}_{j,t+1|t} = \mathbb{E}[\ell_{j,t+1} \mid \mathcal{I}_t^{\text{full}}], \qquad \hat{\sigma}^2_{j,t+1|t} = \text{Var}(\ell_{j,t+1} \mid \mathcal{I}_t^{\text{full}}),$$

are defined exactly as in Section 4, replacing $\mathcal{I}_t$ by $\mathcal{I}_t^{\text{full}}$. The IMM machinery and the variance decomposition remain unchanged; only the underlying filter now exploits all past losses, not just those of the selected experts, yielding more informative and less selection-biased estimates of the latent dynamics.

Crucially, in this full-information regime the router's decisions $(r_t)_{t\geq 1}$ no longer affect the quality of future information: at each time t, the full panel of losses $(\ell_{j,t})_{j\in\mathcal{K}_t}$ is observed regardless of which expert was used for prediction. Combined with the assumption that the latent process

$$\Big(z_t,\ (\boldsymbol{\alpha}_{j,t})_{j\in\mathcal{U}},\ \mathbf{x}_t\Big)_{t\geq 1}$$

evolves independently of $(r_t)_{t\geq 1}$, this implies that the dynamic optimization problem (2) decomposes into a sequence of one-step problems.

Proposition 3 (Optimality of the myopic Bayes rule under full feedback)**.** Assume the full-information feedback model described above and suppose that the latent process and context sequence

$$\Big(z_t,\ (\boldsymbol{\alpha}_{j,t})_{j\in\mathcal{U}},\ \mathbf{x}_t\Big)_{t\geq 1}$$

do not depend on the routing decisions $(r_t)_{t\geq 1}$. Then any policy π that, at each time t, chooses

$$r_t^\star \in \arg\min_{j\in\mathcal{K}_t} \mathbb{E}\big[C_t(j) \mid \mathcal{I}_{t-1}^{\text{full}}\big] = \arg\min_{j\in\mathcal{K}_t}\{\widehat{\mathcal{R}}_{j,t|t-1} + \beta_j\}$$

is globally optimal for the cumulative objective (2). In particular, the one-step Bayes rule with squared loss (i.e., $\lambda = 0$ in the risk-adjusted score) minimizes

$$\mathbb{E}\Big[\sum_{t=1}^{T} C_t(r_t)\Big]$$

among all admissible policies.

Proof. Fix an admissible policy π. By assumption, the joint law of the latent process and contexts

$$\Big(z_t,\ (\boldsymbol{\alpha}_{j,t})_{j\in\mathcal{U}},\ \mathbf{x}_t\Big)_{t=1}^{T}$$

is independent of π. In the full-information model, for each t the vector of losses $(\ell_{j,t})_{j\in\mathcal{K}_t}$ is observed and added to the history, regardless of the choice r_t. Thus the conditional distribution of the next-step costs

$$\big(C_t(j)\big)_{j\in\mathcal{K}_t} = \big(\ell_{j,t}+\beta_j\big)_{j\in\mathcal{K}_t}$$

given $\mathcal{I}_{t-1}^{\text{full}}$ does not depend on the past actions $(r_1,\dots,r_{t-1})$.

For any such policy,

$$\mathbb{E}_\pi\Big[\sum_{t=1}^{T} C_t(r_t)\Big] = \sum_{t=1}^{T}\mathbb{E}_\pi\big[C_t(r_t)\big] = \sum_{t=1}^{T}\mathbb{E}\Big[\mathbb{E}\big[C_t(r_t)\mid\mathcal{I}_{t-1}^{\text{full}}\big]\Big],$$

where $\mathbb{E}_\pi$ denotes expectation under the process induced by π and the fixed environment. At each time t, conditional on $\mathcal{I}_{t-1}^{\text{full}}$, the inner term

$$\mathbb{E}\big[C_t(r_t)\mid\mathcal{I}_{t-1}^{\text{full}}\big]$$

is minimized by choosing

$$r_t^\star \in \arg\min_{j\in\mathcal{K}_t}\mathbb{E}\big[C_t(j)\mid\mathcal{I}_{t-1}^{\text{full}}\big].$$

Since the sum over t is a sum of such conditional expectations, selecting $r_t^\star$ at each time t yields a policy that minimizes the total expected cost. Using

$$\mathbb{E}\big[C_t(j)\mid\mathcal{I}_{t-1}^{\text{full}}\big] = \widehat{\mathcal{R}}_{j,t|t-1}+\beta_j$$

gives the stated form. □

Interpretation. In the censored (bandit) setting, the router faces an exploration–exploitation trade-off: selecting an expert both incurs cost and determines which loss is observed, influencing what is learned about future performance. Under full expert feedback, the learning process is entirely decoupled from the routing decisions: the SLDS–IMM model is updated using a complete panel of expert losses at each time step, independently of which experts were used for prediction. As a consequence, the optimal routing strategy becomes purely *cost-sensitive*: it suffices to select, at each time t, the expert minimizing the conditional expected cost $\widehat{\mathcal{R}}_{j,t|t-1}+\beta_j$. The variance term $\hat{\sigma}^2_{j,t+1|t}$ can still be used to encode risk preferences (via a parameter λ as in Section 4), but it no longer plays a role in driving active exploration.

8 Extension: Model-Based Horizon Planning via Expert-Driven Context Updates

In the core formulation, the context process $(\mathbf{x}_t)_{t\geq 1}$ is exogenous: at each time t the router observes $\mathbf{x}_t$, selects an expert $r_t\in\mathcal{K}_t$, incurs cost $C_t(r_t)$, and the environment then produces the next context $\mathbf{x}_{t+1}$. The SLDS/IMM machinery models the evolution of the experts' *losses*, not of the contexts themselves.

For horizon-H *planning* and workload forecasting, it is often useful to construct a *surrogate* future in which the router's decisions and the experts' forecasts shape an internal notion of "future context". In this section we describe such a model-based extension. It is important to stress that this extension defines an *internal planning model*: it does not alter the exogenous generative assumptions of Sections **??**–**??**.

8.1 Expert-Driven Context Update Map

We recall that the context space is $\mathcal{X} \subseteq \mathbb{R}^d$, and that at each time t the router observes a context vector $\mathbf{x}_t \in \mathcal{X}$ and a feature vector

$$\phi_t \coloneqq \phi(\mathbf{x}_t) \in \mathbb{R}^m.$$

Each expert $j \in \mathcal{U}$ implements a forecasting function

$$f_j : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R},$$

so that $f_j(\mathbf{x}_t)$ can be interpreted as the expert's prediction for the next value of the underlying time series (or some derived quantity) based on the current context.

To propagate surrogate contexts over a planning horizon, we introduce a deterministic *context update map*

$$\Psi : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}.$$

Intuitively, $\Psi(\mathbf{x}, \hat{y})$ specifies how we would construct the *next* context if we were to treat the forecast $\hat{y}$ as the next observation of the underlying series. For example, in a standard autoregressive setting where the context aggregates the last p lags,

$$\mathbf{x}_t = (y_t, y_{t-1}, \ldots, y_{t-p+1}),$$

a natural choice is

$$\Psi(\mathbf{x}_t, \hat{y}_{t+1}) \coloneqq (\hat{y}_{t+1}, y_t, \ldots, y_{t-p+2}),$$

i.e., shift the lag window and append the forecast in place of the unknown y_{t+1}.

8.2 Surrogate Context Trajectories under a Planned Schedule

Fix a current decision time t and a planning horizon $H \in \mathbb{N}$. For a given starting context $\mathbf{x}_t$ and a *candidate schedule*

$$s \;\coloneqq\; (j_1, \ldots, j_H) \in \mathcal{U}^H,$$

we define a deterministic *surrogate context trajectory*

$$(\tilde{\mathbf{x}}^{(s)}_{t+h})_{h=0}^{H} \subset \mathcal{X}$$

by the recursion

$$\tilde{\mathbf{x}}^{(s)}_{t} \coloneqq \mathbf{x}_t, \tag{25}$$

$$\hat{y}^{(s)}_{t+h} \coloneqq f_{j_h}(\tilde{\mathbf{x}}^{(s)}_{t+h-1}), \qquad h = 1, \ldots, H, \tag{26}$$

$$\tilde{\mathbf{x}}^{(s)}_{t+h} \coloneqq \Psi\big(\tilde{\mathbf{x}}^{(s)}_{t+h-1}, \hat{y}^{(s)}_{t+h}\big), \qquad h = 1, \ldots, H. \tag{27}$$

Thus $\tilde{\mathbf{x}}^{(s)}_{t+h}$ is the context we would obtain at (pseudo-)time $t+h$ if, starting from the current $\mathbf{x}_t$, we were to route to expert j_1 at $t+1$, treat $f_{j_1}(\mathbf{x}_t)$ as the next observation when constructing the context, then route to j_2, and so on.

For each h, we also define the associated feature vector

$$\tilde{\phi}^{(s)}_{t+h} \;\coloneqq\; \phi(\tilde{\mathbf{x}}^{(s)}_{t+h}) \in \mathbb{R}^m.$$

Remark 3 (Internal planning model)**.** The trajectory $(\tilde{\mathbf{x}}^{(s)}_{t+h})_{h=0}^{H}$ is a deterministic function of $(\mathbf{x}_t, s)$ and the expert maps $(f_j)_{j \in \mathcal{U}}$. It is *not* the actual future context trajectory $(\mathbf{x}_{t+h})_{h \geq 1}$ under the true data-generating process, but an internal surrogate used for planning and scheduling. The SLDS/IMM model for expert losses remains exogenous and unchanged.

8.3 Predictive Loss along a Planned Schedule

Let $(b_t(k))_{k\in\mathcal{Z}}$ and $(m^{(k)}_{j,t|t}, P^{(k)}_{j,t|t})$ denote the IMM posterior at time t:

$$b_t(k) = \mathbb{P}(z_t = k \mid \mathcal{I}_t), \qquad \boldsymbol{\alpha}_{j,t} \mid \{z_t = k, \mathcal{I}_t\} \approx \mathcal{N}(m^{(k)}_{j,t|t}, P^{(k)}_{j,t|t}).$$

For each $h \geq 1$, we can propagate the SLDS dynamics forward *without* incorporating future observations to obtain the h-step-ahead predicted reliability states and regime probabilities:

$$\begin{aligned} m^{(k)}_{j,t+h|t},\ P^{(k)}_{j,t+h|t} \quad &\text{for each } j \in \mathcal{U},\ k \in \mathcal{Z}, \\ b_{t+h|t}(k) \ &\coloneqq\ \mathbb{P}(z_{t+h} = k \mid \mathcal{I}_t), \end{aligned}$$

by iterating the IMM time-update recursions h times. These quantities depend only on the dynamics $(\Pi, \mathbf{A}_{k,j}, \mathbf{Q}_{k,j})$ and the current belief at t; they do not depend on the planned schedule s.

Given a schedule s and its surrogate contexts (27), the predictive distribution of the loss incurred at pseudo-time $t+h$ by selecting expert j_h is approximated by

$$\mathbb{P}(\ell_{j_h,t+h} \mid \mathcal{I}_t, s) \ \approx\ \sum_{k=1}^{M} b_{t+h|t}(k)\, \mathcal{N}(\mu^{(k,s)}_{j_h,t+h|t}, S^{(k,s)}_{j_h,t+h|t}),$$

where, under regime k,

$$\begin{aligned} \mu^{(k,s)}_{j_h,t+h|t} &\coloneqq \tilde{\phi}^{(s)\top}_{t+h} m^{(k)}_{j_h,t+h|t}, \\ S^{(k,s)}_{j_h,t+h|t} &\coloneqq \tilde{\phi}^{(s)\top}_{t+h} P^{(k)}_{j_h,t+h|t} \tilde{\phi}^{(s)}_{t+h} + R_{k,j_h}. \end{aligned}$$

The corresponding predictive mean and variance of $\ell_{j_h,t+h}$ under schedule s are

$$\begin{aligned} \widehat{\mathcal{R}}^{(s)}_{j_h,t+h|t} &\coloneqq \mathbb{E}[\ell_{j_h,t+h} \mid \mathcal{I}_t, s] = \sum_{k=1}^{M} b_{t+h|t}(k)\, \mu^{(k,s)}_{j_h,t+h|t}, \\ \hat{\sigma}^{2,(s)}_{j_h,t+h|t} &\coloneqq \mathrm{Var}(\ell_{j_h,t+h} \mid \mathcal{I}_t, s) \\ &= \sum_{k=1}^{M} b_{t+h|t}(k)\, S^{(k,s)}_{j_h,t+h|t} + \sum_{k=1}^{M} b_{t+h|t}(k) (\mu^{(k,s)}_{j_h,t+h|t} - \widehat{\mathcal{R}}^{(s)}_{j_h,t+h|t})^2, \end{aligned}$$

by the same variance decomposition as in Proposition **??**.

We define the *risk-adjusted one-step planning cost* for the h-th element of the schedule s by

$$\widetilde{C}^{(s)}_{t+h} \ \coloneqq\ \widehat{\mathcal{R}}^{(s)}_{j_h,t+h|t} + \beta_{j_h} + \lambda \sqrt{\hat{\sigma}^{2,(s)}_{j_h,t+h|t}}, \tag{28}$$

where $\lambda \in \mathbb{R}$ is the same risk parameter as in the one-step selection rule.

8.4 Horizon-H Planning Objective and Scheduling Forecast

For a fixed schedule $s = (j_1, \ldots, j_H)$ we can aggregate the risk-adjusted planning costs over the horizon, e.g. without discounting,

$$J_{\mathrm{plan}}(s) \ \coloneqq\ \sum_{h=1}^{H} \widetilde{C}^{(s)}_{t+h}, \tag{29}$$

or with a discount factor $\gamma \in (0, 1]$,

$$J^{\gamma}_{\mathrm{plan}}(s) \ \coloneqq\ \sum_{h=1}^{H} \gamma^{h-1} \widetilde{C}^{(s)}_{t+h}.$$

In principle, one could define an *open-loop* horizon-H planning problem by minimizing (29) over $s \in \mathcal{U}^H$. This is combinatorial in H and $|\mathcal{U}|$ and primarily of conceptual interest.

A more practical use of the surrogate model is *scheduling forecast*: instead of optimizing over schedules, we fix a router policy (typically the myopic one-step rule of Section **??**) and *simulate* its behaviour on the surrogate environment defined by (27). Concretely:

- At simulation step $h = 1$, starting from $\tilde{\mathbf{x}}_t^{(s)} = \mathbf{x}_t$ and the belief at time t, we construct candidate surrogate contexts

$$\tilde{\mathbf{x}}_{t+1}^{(j)} \coloneqq \Psi(\mathbf{x}_t, f_j(\mathbf{x}_t)), \qquad j \in \mathcal{K}_{t+1},$$

  compute the corresponding predictive means and variances $(\widehat{\mathcal{R}}_{j,t+1|t}^{(j)}, \hat{\sigma}_{j,t+1|t}^{2,(j)})$, and select

$$\tilde{r}_{t+1} \in \arg\min_{j \in \mathcal{K}_{t+1}} \Big\{ \widehat{\mathcal{R}}_{j,t+1|t}^{(j)} + \beta_j + \lambda \sqrt{\hat{\sigma}_{j,t+1|t}^{2,(j)}} \Big\}.$$

  We then set $\tilde{\mathbf{x}}_{t+1} = \tilde{\mathbf{x}}_{t+1}^{(\tilde{r}_{t+1})}$.

- At simulation step $h = 2$, we repeat the construction starting from $\tilde{\mathbf{x}}_{t+1}$, obtain $\tilde{r}_{t+2}$ and $\tilde{\mathbf{x}}_{t+2}$, and so on up to horizon H.

This yields a simulated sequence $(\tilde{r}_{t+1}, \ldots, \tilde{r}_{t+H})$ of expert indices which approximates the future routing decisions that would be induced by the one-step policy under the internal, expert-driven context dynamics. Repeating this procedure under different modelling choices for Ψ or different initial contexts $\mathbf{x}_t$ provides:

- *Workload forecasts* for each expert over the next H steps (e.g., expected number of routed queries), useful for capacity planning and SLA management;
- A way to compare alternative routing policies (different λ, different cost vectors (β_j)) in terms of their long-run impact on expert usage, without modifying the core SLDS-L2D problem.

Remark 4 (Interpretation and limitations)**.** The horizon-H planning framework described in this section is explicitly *model-based*. The surrogate contexts $\tilde{\mathbf{x}}_{t+h}^{(s)}$ and the induced routing sequence $(\tilde{r}_{t+1}, \ldots, \tilde{r}_{t+H})$ live in an internal closed-loop model where expert forecasts are treated as future observations via the update map Ψ. They need not coincide with the true future contexts and routing decisions under the exogenous data-generating process. As such, this extension is best viewed as a principled scheduling and capacity-planning heuristic built on top of the SLDS/IMM L2D framework, rather than as a modification of the underlying probabilistic model.

9 Conclusion

We presented a theoretically grounded framework for Time-Series L2D. By leveraging Switching State-Space models, we explicitly account for non-stationarity and cost. The formulation supports dynamic expert sets and handles partial feedback via rigorous Bayesian filtering, with extensions for diversity-aware ensemble selection.