Appendices

Appendix Contents

A THE MOE-F ALGORITHM

This section records a detailed version of Algorithm 1, which can also be rolled-forward online as the target process $Y_\cdot$ is dynamically observed.

Algorithm 2: The MoE-F Algorithm

```
Input: A time-horizon $T \in \mathbb{N}_+$, $N$ (pre-trained) experts $f^{(1)}, \dots, f^{(N)}$, hyperparameters $\lambda > 0$,
        $\alpha \in (0,1)$ and $k \in \mathbb{N}_+$, target $(Y_t)_{t=0}^{T-1}$, and input signal $x_{[0:T-1]}$.
Output: A posterior Mixture Weights $w_t$
/* Initialize                                                    */
Initialize $\pi \overset{\text{def.}}{=} (\pi^{(n:i)})_{n,i=1}^N \leftarrow (1/N)_{n,i=1}^N$
$Q \leftarrow (1/(N-1) I_{i \neq j} - 1\, I_{i=j})_{i,j=1}^N$
$(L_-^{(n)})_{n=1}^N \leftarrow 0$
for $t = 0, \dots, T-1$ do
    For $n = 1, \dots, N$ in parallel
        $\bar{A} \leftarrow \bar{A}_t^{(n)}(\pi^{(n)}, Y_t)$
        $B \leftarrow B_t^{(n)}(Y_t)$ $\tilde{\pi} \leftarrow \pi$
        $L^{(n)} \leftarrow \ell\big(f^{(n)}(x_{[0:t]})\big)$
        $\Delta L \leftarrow L^{(n)} - L_-^{(n)}$
        $\Delta \overline{W} \leftarrow \frac{\Delta L - \bar{A}}{B}$
        $L_-^{(n)} \leftarrow L^{(n)}$
        /* Update components of $n^{th}$ expert's posterior $(\pi^{(n)})$          */
        for $i = 1, \dots, N$ do
            $A \leftarrow A_t^{(n)}(e_i, Y_t)$
            drift $\leftarrow Q_i^\top \tilde{\pi}^{(n)}$
            diffusion $\leftarrow \tilde{\pi}^{(n:i)}\,(A - \bar{A})/B$
            $\pi^{(n:i)} \leftarrow \tilde{\pi}^{(n:i)} +$ drift $+$ diffusion$\Delta\overline{W}$
        end for
        $\pi^{(n)} \leftarrow \pi^{(n)} / \sum_i \pi^{(n:i)}$
        $s_n \leftarrow \ell\big(Y_t, (\pi^{(n)})^\top F(x_{[0:t]})\big)$
        $\hat{Y}_t^{(n)} \leftarrow (\pi^{(n)})^\top F(x_{[0:t]})$          // Calculate expert scores
    end
    $\bar{\pi} \leftarrow \big(e^{-\lambda\, s_n} / \big(\sum_{i=1}^N e^{-\lambda\, s_i}\big)\big)_{n=1}^N$          // Get Expert Scores
    $\hat{Y}_t \overset{\text{def.}}{=} \bar{\pi}^\top\, \big(\hat{Y}_t^{(n)}\big)_{n=1}^N$          // Get time $t$ prediction
    /* Update $Q$                                                  */
    $\tilde{Q} \leftarrow Q$
    for $n = 1, \dots, N$ do
        $P^{(n)} \leftarrow \bar{\pi}$
    end for
    $P \leftarrow (1-\alpha)\, P + \alpha I_N$
    $Q \leftarrow \operatorname{ReLU}(\log(P)) - \operatorname{diag}(\bar{1}_N^\top \operatorname{ReLU}(\log(P)))$
end for
return Sequence of Mixture Predictions $(\hat{Y}_t)_{t=0}^{T-1}$
```

B Proofs

This section contains the proofs of our main result, generalizations thereof, and variants which apply to the quadratic (squared) loss. In the latter case, the necessary modifications to the algorithm and the overall proof structure are relatively similar but with key technical differences.

Mild Generalizations and Further Discussion. We will consider the slightly more general case where the target process $Y_{\cdot}$ follows the generalized dynamics

$$Y_t = Y_0 + \underbrace{\int_0^t w_t^\top F(x_s)\, ds}_{\text{Best Expert Estimate}} + \underbrace{\int_0^t \sigma_s\, dW_s}_{\text{(Generalized) Idiosyncratic Residual}} \quad , \tag{7}$$

where, there are constants $\alpha, C \geq 0$ with $C \leq 1$ such that: for each $t \geq 0$ one has $\sigma_t = C\, e^{-\alpha\, t}$. By the Itô-isometry, see [Cohen & Elliott, 2015, Lemma 12.1.4], we have that the variance of $\int_0^t \sigma_s\, dW_s$ is given by

$$\varsigma_t^2 \overset{\text{def.}}{=} \mathbb{E}\left[\left(\int_0^t \sigma_s\, dW_s\right)^2\right] = \begin{cases} C^2\,(1-e^{-\alpha 2t})/(2\alpha) & \text{if } \alpha > 0 \\ t & \text{if } \alpha = 0 \end{cases} \tag{8}$$

Observe that, if $\alpha > 0$ then the variance of $\int_0^t \sigma_s\, dW_s$ asymptotically stabilizes at 1, as t becomes arbitrarily large. In contrast, the variance of $\int_0^t \sigma_s\, dW_s$ diverges in the case where $\alpha = 0$ (which is the case considered in the main body of our paper).

Intuition behind the choice of assumed fluctuations/diffusion. The intuition behind this modelling choice for the diffusion coefficient $\sigma_{\cdot}$ is based on ideas behind concentration of measure. Consider the case where $\alpha > 0$ in equation (8). Since we will be considering classification applications, then we will not want the idiosyncratic residual $\int_0^t \sigma_s\, W_s$ to push fluctuate outside the unit interval $[0,1]$, or rather the probability that any fluctuation of $\int_0^t \sigma_s\, W_s$ is "large" should be small. Since $\int_0^t \sigma_s\, W_s$ has a Gaussian distribution, then note, by standard Gaussian concentration inequalities, we have that

$$\mathbb{P}\left(\left|\int_0^t \sigma_s\, W_s\right| \geq 1/2\right) \leq e^{(1/2)^2/(2\varsigma_t^2)} = e^{-\alpha/\left(4C^2(1-e^{-\alpha 2t})\right)} \leq e^{-\alpha/(C^2 4)} \leq e^{-\alpha/4}. \tag{9}$$

We can control the probability that any fluctuation is "large", meaning larger than $1/2$, by setting the right-hand side of equation (9) to be a prespecified "small" value $\delta \in (0,1]$ and solving for the required $\alpha > 0$ parameter in terms of δ yields the specification $\alpha = \ln(1/\delta^4)$. If $d \geq 2$, then we may set $C = \frac{\sqrt{2}}{d}$ purely for convenience in simplifying expressions below.

In this case, for any hyperparameter $0 < \delta \leq 1$, the quantities in Theorem 1 become

$$\pi_t^{(n:i)} = w_0^i + \int_0^t (Q_s)_i^\top \pi_s^{(n)}\, ds + \int_0^t \frac{\pi_s^{(n:i)}\,\left(A_s(e_i, Y_{[0:s]}) - \bar{A}_s(Y_{[0:s]})^\top \pi_s^{(n)}\right)}{B_s(Y_{[0:s]})} d\overline{W}_u^{(n)}, \tag{10}$$

where $(Q_t)_i$ denotes the i^{th} row of the transitions matrix Q_t at time $t \geq 0$, $w_0^i \overset{\text{def.}}{=} \mathbb{P}(w_0 = e_i)$ and where the "innovations process" $\overline{W}_{\cdot}^{(n)} \overset{\text{def.}}{=} (\overline{W}_t^{(n)})_{t\geq 0}$ is the following $(\mathbb{P}, \mathcal{F}_{\cdot}^n)$-Brownian motion

$$\overline{W}_s^{(n)} \overset{\text{def.}}{=} \int_0^s \frac{dL_{[0:u]}^{(n)} - \bar{A}_u(Y_{[0:u]})}{B_u(Y_{[0:u]})} du,$$

where the "stochastic differential" $dL_{[0:u]}^{(n)}$ is given by

$$\begin{aligned} dL_{[0:u]}^{(n)} =& \left(2\big(Y_u - f^{(n)}(x_u)\big)[\nabla_u f^{(n)}(x_u) + w_u^\top F(x_u)] - e^{-2t\ln(1/\delta^4)}\right) du \\ &+ \frac{2^{3/2}}{d}\,(Y_u - f^{(n)}(x_u)) e^{-t\ln(1/\delta^4)}\, dW_u. \end{aligned}$$

B.1 Proof of Theorem 1

We are now ready to state and prove two versions, one of which generalizes, our first main result (Theorem 1). We consider two cases. We obtain our main result by customizing Theorem 1 to our classification problem, where $D = 1$ and the range of each expert is in $\{0, 1\} \subset \mathbb{R}$, and setting σ to be a specific constant in $(0, \infty)$. For convenience, if we postulate that $\sigma_t = 1$; i.e. it is a constant function of the path $y_{[0:t]}$ and of time $t \geq 0$. Note that, by the Itô-isometry, see [Cohen & Elliott, 2015, Lemma 12.1.4] the *idiosyncratic residual* term $\int_0^t \sigma_s(Y_{[0:s]})\, dW_s$ in equation (1) has a centred normal random distribution with variance $\int_0^t \sigma_s^2\, ds = t$.

B.1.1 Case I: Binary Cross-Entropy Case

We now state and prove a mild generalization of Theorem 1.

Theorem 3 (Optimal Optimistic Prior for n^{th} Expert - Squared Loss Case)**.** *Consider the binary cross-entropy loss*

$$\ell(\hat{y}, y) \overset{\text{def.}}{=} y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})$$

and fix a continuously differentiable path $x_\cdot \in C^1(\mathbb{R})$.
Under Assumptions 4.1, the best a posteriori estimate of the n^{th} expert, $\pi_t^{(n)}$, satisfies the following stochastic differential equation

$$\pi_t^{(n:i)} = w_0^i + \int_0^t (Q_t)_i^\top \pi_s^{(n)}\, ds + \int_0^t \frac{\pi_s^{(n:i)} \left(A_s(e_i, Y_{[0:s]}) - \bar{A}_s(\pi^{(n)}, Y_{[0:s]})\right)}{B_s(Y_{[0:s]})} d\overline{W}_u^{(n)}, \tag{11}$$

where $(Q_t)_i$ denotes the i^{th} row of the transitions matrix Q_t at time $t \geq 0$, $w_0^i \overset{\text{def.}}{=} \mathbb{P}(w_0 = e_i)$,

$$A_t^{(n)}(w, y_{[0:t]}) \overset{\text{def.}}{=} -\frac{\left(Y_s - f^{(n)}(x_{[0:t]})\right) \Delta f^{(n)}(x_{[0:t]})}{\left(1 - f^{(n)}(x_{[0:t]})\right) f^{(n)}(x_{[0:t]})} - \log\left(\frac{f^{(n)}(x_{[0:t]})}{1 - f^{(n)}(x_{[0:t]})}\right) [w_s^\top F(x_{[0:s]})]$$

$$\bar{A}_t^{(n)}(\pi^{(n)}, y_{[0:t]}) \overset{\text{def.}}{=} \sum_{i=1}^{d} A_t(e_i, Y_{[0:t]})\, \pi^{(n:i)}$$

$$B_t^{(n)}(y_{[0:t]}) \overset{\text{def.}}{=} -\log\left(\frac{f^{(n)}(x_{[0:t]})}{1 - f^{(n)}(x_{[0:t]})}\right) e^{s \ln(\delta^4)}$$

$$F(x_{[0:t]}) \overset{\text{def.}}{=} \left(f^{(1)}(x_{[0:t]}), \dots, f^{(N)}(x_{[0:t]})\right)$$

and the "innovations process" $\overline{W}_\cdot^{(n)} \overset{\text{def.}}{=} (\overline{W}_t^{(n)})_{t \geq 0}$ is the following $(\mathbb{P}, \mathcal{F}_\cdot^n)$-Brownian motion

$$\overline{W}_s^{(n)} \overset{\text{def.}}{=} \int_0^s \frac{dL_{[0:u]}^{(n)} - \bar{A}_u(Y_{[0:u]})}{B_u(Y_{[0:u]})} du,$$

where

$$dL_{[0:u]}^{(n)} = d\ell(Y_t, \hat{f}^{(n)}(x_{[0:t]}) = \frac{\left(Y_s - f^{(n)}(x_{[0:t]})\right) \Delta f^{(n)}(x_{[0:t]})}{\left(1 - f^{(n)}(x_{[0:t]})\right) f^{(n)}(x_{[0:t]})} + \log\left(\frac{f^{(n)}(x_{[0:t]})}{1 - f^{(n)}(x_{[0:t]})}\right) [w_s^\top F(x_{[0:s]})]$$
$$+ \log\left(\frac{f^{(n)}(x_{[0:t]})}{1 - f^{(n)}(x_{[0:t]})}\right) e^{s \ln(\delta^4)}\, dW_s.$$

Proof. Fix a continuously differentiable path $x_\cdot \in C^1(\mathbb{R})$. Define $\Delta f(x_{[0:t]}) \overset{\text{def.}}{=} \partial_t f(x_{[0:t]})$ and set $\ell : \mathbb{R} \ni y \mapsto (y - f(x_{[0:t]}))^2$.

First, observe that: for all $t \geq 0$, all $x_\cdot \in C^1(\mathbb{R})$ and all $y \in \mathbb{R}$ one has

$$\begin{aligned}
\ell_t(y) &= -\big(y\,\log(f^{(n)}(x_{[0:t]})) + (1-y)\,\log(1-f^{(n)}(x_{[0:t]}))\big)\\
\frac{\partial \ell_t}{\partial t}(y) &= -\frac{\big(y - f^{(n)}(x_{[0:t]})\big)\,\Delta f^{(n)}(x_{[0:t]})}{\big(1-f^{(n)}(x_{[0:t]})\big)\,f^{(n)}(x_{[0:t]})},\\
\frac{\partial \ell_t}{\partial y}(y) &= -\log\bigg(\frac{f^{(n)}(x_{[0:t]})}{1-f^{(n)}(x_{[0:t]})}\bigg),\\
\frac{\partial^2 \ell_t}{\partial y^2}(y) &= 0.
\end{aligned} \tag{12}$$

Since $\ell \in C^\infty(\mathbb{R}^d \times \mathbb{R}^d)$, we assumed that the path $x_\cdot \in C^1([0:\infty),\mathbb{R}^d)$ then this, together with the postulated dynamics on $Y_\cdot^{(n)}$ imply that Itô's Lemma/Formula, see [Cohen & Elliott, 2015, Theorem 14.2.4], used on the map $\ell_t^{(n)} : [0,\infty) \times \mathbb{R}^D \ni (t,y) \to (y - f^{(n)}(x_t))^2 \in \mathbb{R}$ pre-composed with $Y_\cdot$ applies. Whence, our and the assumed dynamics on $Y_\cdot$, postulated in equation (1), imply that the process $L_t^{(n)} \overset{\text{def.}}{=} \ell_t^{(n)}(Y_t)$ satisfies the following stochastic differential equation

$$\begin{aligned}
L_t^{(n)} = L_0^{(n)} + \int_0^t & \frac{\partial \ell_s^{(n)}}{\partial s}(Y_s)\\
&+ \frac{\partial \ell_s^{(n)}}{\partial y}(Y_s)\,[w_s^\top F(x_{[0:s]})]\\
&+ \frac{1}{2}\,\frac{\partial^2 \ell_t^{(n)}}{\partial y^2}(Y_s)\,2e^{s\ln(\delta^8)}ds\\
&+ \int_0^t \frac{\partial \ell_t^{(n)}}{\partial y}(Y_s)e^{s\ln(\delta^4)}\,dW_s\\
= L_0^{(n)} + \int_0^t & \frac{(-1)\big(Y_s - f^{(n)}(x_{[0:t]})\big)\,\Delta f^{(n)}(x_{[0:t]})}{\big(1-f^{(n)}(x_{[0:t]})\big)\,f^{(n)}(x_{[0:t]})} \qquad (13)\\
&+ (-1)\log\bigg(\frac{f^{(n)}(x_{[0:t]})}{1-f^{(n)}(x_{[0:t]})}\bigg)\,[w_s^\top F(x_{[0:s]})]\\
&+ \frac{1}{2}0\;ds\\
&+ \int_0^t (-1)\log\bigg(\frac{f^{(n)}(x_{[0:t]})}{1-f^{(n)}(x_{[0:t]})}\bigg)\,e^{s\ln(\delta^4)}\,dW_s\\
= L_0^{(n)} + \int_0^t & -\frac{\big(Y_s - f^{(n)}(x_{[0:t]})\big)\,\Delta f^{(n)}(x_{[0:t]})}{\big(1-f^{(n)}(x_{[0:t]})\big)\,f^{(n)}(x_{[0:t]})} - \log\bigg(\frac{f^{(n)}(x_{[0:t]})}{1-f^{(n)}(x_{[0:t]})}\bigg)\,[w_s^\top F(x_{[0:s]})]\;\,ds\\
&- \log\bigg(\frac{f^{(n)}(x_{[0:t]})}{1-f^{(n)}(x_{[0:t]})}\bigg)\int_0^t e^{s\ln(\delta^4)}\,dW_s
\end{aligned} \tag{14}$$

Synchronizing our notation with [Liptser & Shiryaev, 2001a, Equation (9.1)], we write

$$\begin{aligned}
A_t(w, y_{[0:t]}) &\overset{\text{def.}}{=} -\frac{\big(Y_s - f^{(n)}(x_{[0:t]})\big)\,\Delta f^{(n)}(x_{[0:t]})}{\big(1-f^{(n)}(x_{[0:t]})\big)\,f^{(n)}(x_{[0:t]})} - \log\bigg(\frac{f^{(n)}(x_{[0:t]})}{1-f^{(n)}(x_{[0:t]})}\bigg)\,[w_s^\top F(x_{[0:s]})]\\
B_t(y_{[0:t]}) &\overset{\text{def.}}{=} -\log\bigg(\frac{f^{(n)}(x_{[0:t]})}{1-f^{(n)}(x_{[0:t]})}\bigg)\,e^{s\ln(\delta^4)}
\end{aligned} \tag{15}$$

Under Assumptions 4.1, we may apply [Liptser & Shiryaev, 2001a, Theorem 9.1] to deduce that Then the a posteriori probability $\pi_t^{n:0} \overset{\text{def.}}{=} (\pi_t)_0$, satisfies a system of equations

$$\pi_t^{(n:i)} = w_0^i + \int_0^t (Q_t)_i^\top \pi_s^{(n)}\,ds + \int_0^t \frac{\pi_s^{(n:i)}\,\big(A_s(e_i, Y_{[0:s]}) - \bar{A}_s(\pi^{(n)}, Y_{[0:s]})\big)}{B_s(Y_{[0:s]})}d\overline{W}_u, \tag{16}$$

where $(Q_t)_i$ denotes the i^{th} row of the Q_t/transitions matrix Q_t at time $t \geq 0$, $w_0^i \overset{\text{def.}}{=} \mathbb{P}(w_0 = e_i)$, and the "innovations process" is the $(\mathbb{P}, \mathcal{F}^n_\cdot)$-Brownian motion given by

$$\overline{W}^{(n)}_s \overset{\text{def.}}{=} \int_0^s \frac{dL^{(n)}_{[0:u]} - \bar{A}_u(\pi^{(n)}, Y_{[0:u]})}{B_u(Y_{[0:u]})} du \tag{17}$$

and where

$$\bar{A}_t(\pi^{(n)}, y_{[0:t]}) \overset{\text{def.}}{=} \sum_{i=1}^d A_s(e_i, Y_{[0:t]})\, \pi^{(n:i)}.$$

This completes the proof. □

Remark 1. Setting $\delta = 1$ in the previous derivation yields the formulation of Theorem 1 found in the main body of the paper.

B.1.2 Case II: Squared Loss

Theorem 4 (Optimal Optimistic Prior for n^{th} Expert - Squared Loss Case)**.** *Let $\ell(\hat{y}, y) \overset{\text{def.}}{=} (y - \hat{y})^2$ and fix a continuously differentiable path $x_\cdot \in C^1(\mathbb{R})$.*

Under Assumptions 4.1, the best a posteriori estimate of the n^{th} expert, $\pi^{(n)}_t$, satisfies the following stochastic differential equation

$$\pi^{(n:i)}_t = w_0^i + \int_0^t (Q_t)_i^\top \pi^{(n)}_s \, ds + \int_0^t \frac{\pi^{(n:i)}_s \big(A_s(e_i, Y_{[0:s]}) - \bar{A}_s(\pi^{(n)}, Y_{[0:s]})\big)}{B_s(Y_{[0:s]})} d\overline{W}^{(n)}_u, \tag{18}$$

where $(Q_t)_i$ denotes the i^{th} row of the transitions matrix Q_t at time $t \geq 0$, $w_0^i \overset{\text{def.}}{=} \mathbb{P}(w_0 = e_i)$,

$$\begin{aligned}
A^{(n)}_t(w, y_{[0:t]}) &\overset{\text{def.}}{=} 2\big(y_t - f^{(n)}(x_{[0:t]})\big)\big(w_t^\top F(x_{[0:t]}) - \Delta f^{(n)}(x_{[0:t]}) + e^{s\ln(\delta^8)}\big)\\
\bar{A}^{(n)}_t(\pi^{(n)}, y_{[0:t]}) &\overset{\text{def.}}{=} \sum_{i=1}^d A_t(e_i, Y_{[0:t]})\, \pi^{(n:i)}\\
B^{(n)}_t(y_{[0:t]}) &\overset{\text{def.}}{=} 2^{3/2}(y_t - f^{(n)}(x_{[0:t]}))e^{t\ln(\delta^4)}\\
F(x_{[0:t]}) &\overset{\text{def.}}{=} \big(f^{(1)}(x_{[0:t]}), \dots, f^{(N)}(x_{[0:t]})\big)
\end{aligned}$$

and the "innovations process" $\overline{W}^{(n)}_\cdot \overset{\text{def.}}{=} (\overline{W}^{(n)}_t)_{t\geq 0}$ is the following $(\mathbb{P}, \mathcal{F}^n_\cdot)$-Brownian motion

$$\overline{W}^{(n)}_s \overset{\text{def.}}{=} \int_0^s \frac{dL^{(n)}_{[0:u]} - \bar{A}_u(Y_{[0:u]})}{B_u(Y_{[0:u]})} du,$$

where

$$\begin{aligned}
dL^{(n)}_{[0:u]} = d\ell(Y_t, \hat{f}^{(n)}(x_{[0:t]}) =& 2\big(Y_t - f^{(n)}(x_{[0:t]})\big)\big([w_s^\top F(x_{[0:t]})] - \Delta f^{(n)}(x_{[0:t]}) + e^{t\ln(\delta^8)}\big)\\
&+ 2\big(Y_t - f^{(n)}(x_{[0:t]})\big)e^{t\ln(\delta^4)}\, dW_t.
\end{aligned}$$

Proof. Fix a continuously differentiable path $x_\cdot \in C^1(\mathbb{R})$. Define $\Delta f(x_{[0:t]}) \overset{\text{def.}}{=} \partial_t f(x_{[0:t]})$ and set $\ell : \mathbb{R} \ni y \mapsto (y - f(x_{[0:t]}))^2$.

First, observe that: for all $t \geq 0$, all $x_\cdot \in C^1(\mathbb{R})$ and all $y \in \mathbb{R}$ one has

$$\begin{aligned}
\ell^{(n)}_t(y) &= \big(y - f^{(n)}(x_{[0:t]})\big)^2\\
\frac{\partial \ell_t}{\partial t}(y) &= 2\big(y - f^{(n)}(x_{[0:t]})\big)\big(-\Delta f^{(n)}(x_{[0:t]})\big),\\
\frac{\partial \ell_t}{\partial y}(y) &= 2\big(y - f^{(n)}(x_{[0:t]})\big),\\
\frac{\partial^2 \ell_t}{\partial y^2}(y) &= 2.
\end{aligned} \tag{19}$$

Since $\ell \in C^\infty(\mathbb{R}^d \times \mathbb{R}^d)$, we assumed that the path $x_\cdot \in C^1([0:\infty), \mathbb{R}^d)$ then this, together with the postulated dynamics on $Y_\cdot^{(n)}$ imply that Itô's Lemma/Formula, see [Cohen & Elliott, 2015, Theorem 14.2.4], used on the map $\ell_t^{(n)} : [0,\infty) \times \mathbb{R}^D \ni (t,y) \to (y - f^{(n)}(x_t))^2 \in \mathbb{R}$ pre-composed with $Y_\cdot$ applies. Whence, the computations in equation (19) and the assumed dynamics on $Y_\cdot$, postulated in equation (1), imply that the process $L_t^{(n)} \overset{\text{def.}}{=} \ell_t^{(n)}(Y_t)$ satisfies the following stochastic differential equation

$$\begin{aligned} L_t^{(n)} &= L_0^{(n)} + \int_0^t \frac{\partial \ell_s^{(n)}}{\partial s}(Y_s) + \frac{\partial \ell_s^{(n)}}{\partial y}(Y_s)\,[w_s^\top F(x_{[0:s]})] + \frac{1}{2}\frac{\partial^2 \ell_t^{(n)}}{\partial y^2}(Y_s)\,2e^{s\ln(\delta^8)}ds \\ &\quad + \int_0^t \frac{\partial \ell_t^{(n)}}{\partial y}(Y_s)e^{s\ln(\delta^4)}\,dW_s \\ &= L_0^{(n)} + \int_0^t 2\big(Y_s - f^{(n)}(x_{[0:s]})\big)\big([w_s^\top F(x_{[0:s]})] - \Delta f^{(n)}(x_{[0:s]}) + e^{s\ln(\delta^8)}\big)ds \\ &\quad + \int_0^t 2\big(Y_s - f^{(n)}(x_{[0:s]})\big)\sqrt{2}e^{s\ln(\delta^4)}\,dW_s \end{aligned} \tag{20}$$

Synchronizing our notation with [Liptser & Shiryaev, 2001a, Equation (9.1)], we write

$$\begin{aligned} A_t(w, y_{[0:t]}) &\overset{\text{def.}}{=} 2\big(y_t - f^{(n)}(x_{[0:t]})\big)\big(w_t^\top F(x_{[0:t]}) - \Delta f^{(n)}(x_{[0:t]}) + e^{s\ln(\delta^8)}\big) \\ B_t(y_{[0:t]}) &\overset{\text{def.}}{=} 2^{3/2}(y_t - f^{(n)}(x_{[0:t]}))e^{t\ln(\delta^4)} \end{aligned} \tag{21}$$

Under Assumptions 4.1, we may apply [Liptser & Shiryaev, 2001a, Theorem 9.1] to deduce that Then the a posteriori probability $\pi_t^{n:0} \overset{\text{def.}}{=} (\pi_t)_0$, satisfies a system of equations

$$\pi_t^{(n:i)} = w_0^i + \int_0^t (Q_t)_i^\top \pi_s^{(n)}\,ds + \int_0^t \frac{\pi_s^{(n:i)}\,\big(A_s(e_i, Y_{[0:s]}) - \bar{A}_s(\pi^{(n)}, Y_{[0:s]})\big)}{B_s(Y_{[0:s]})} d\overline{W}_u, \tag{22}$$

where $(Q_t)_i$ denotes the i^{th} row of the Q_t/transitions matrix Q_t at time $t \geq 0$, $w_0^i \overset{\text{def.}}{=} \mathbb{P}(w_0 = e_i)$, and the "innovations process" is the $(\mathbb{P}, \mathcal{F}_\cdot^n)$-Brownian motion given by

$$\overline{W}_s^{(n)} \overset{\text{def.}}{=} \int_0^s \frac{dL_{[0:u]}^{(n)} - \bar{A}_u(\pi^{(n)}, Y_{[0:u]})}{B_u(Y_{[0:u]})} du \tag{23}$$

and where

$$\bar{A}_t(\pi^{(n)}, y_{[0:t]}) \overset{\text{def.}}{=} \sum_{i=1}^d A_s(e_i, Y_{[0:t]})\,\pi^{(n:i)}.$$

This completes the proof. □

Remark 2. Setting $\delta = 1$ in the previous derivation yields the formulation of Theorem 1 found in the main body of the paper.

B.2 Proof of Theorem 2

The proof of Theorem 2 relies on the following result. Briefly, this result guarantees for the validity of the perturbation to the transition probability defined by

$$P_t^\alpha \overset{\text{def.}}{=} (1-\alpha)P_t + \alpha I_N \tag{24}$$

for arbitrary $N \in \mathbb{N}_+$, $P_t \in \mathcal{P}_N^U$, $\alpha \in (0,1)$, and where I_N is the $N \times N$ identity matrix.

In what follows, we will use $\Delta_N \overset{\text{def.}}{=} \{w \in [0,1]^N : \sum_{n=1}^N w_n\}$ to denote the probability N-simplex; which corresponds to the probability (measures) distributions supported on N points. Here, these N points are the experts themselves, and the probability of selecting any expert is interpreted as the relative credibility we ascribe to its historical predictive power.

Proposition 2 (Regularity of Perturbations). *Let $N \in \mathbb{N}_+$, $\lambda > 0$, $s_1, \ldots, s_N \in \mathbb{R}$, $\bar{\pi} \in \Delta_N$ be given by*

$$\bar{\pi} \overset{\text{def.}}{=} \operatorname{Softmin}\left(\lambda\,(s_n)_{n=1}^N\right) \text{ and } P_t \overset{\text{def.}}{=} [(\bar{\pi})_{n=1}^N]^N.$$

For every $\alpha \in (0,1)$, the matrix P_t^α in equation (24), is invertible and (row) stochastic. If, moreover, all its real eigenvalues are non-negative, then $\log(P_t^\alpha)$ is well-defined and its rows sum to 0. In particular, setting $\alpha \geq 1 - 1/N$ guarantees that $\log(P_t^\alpha)$ exists, if P_t^α has real eigenvalues.

We will now show our second main result, and the intermediate lemmata leading up to it. The next lemma states that if a (row) stochastic matrix is constructed by filling each of its rows with an element of the probability simplex, then shining it by an arbitrarily small abound and growing its diagonal proportionally yields a (row) stochastic matrix, which is necessarily invertible.

Lemma 1 (Invertible Perturbations). *Let $N \in \mathbb{N}_+$ let $\pi \in \Delta_N$. If P is a (row) stochastic matrix, then, for any $\alpha \in (0,1)$, the matrix $(1-\alpha)P + \alpha I_N$ is an invertible (row) stochastic matrix.*

Proof of Lemma 1. Let $\mathbf{1}_N \in \mathbb{R}^N$ be such that: for each $i = 1, \ldots, N$ we have $(\mathbf{1}_N)_i = 1$ *(i.e. $\mathbf{1}_N$ is a matrix of ones)*. By construction $P = (\pi, \ldots, \pi)^\top$. Therefore, P can be written as an outer product via

$$P = \mathbf{1}_N\,\pi^\top \tag{25}$$

Therefore, for any $\alpha \in (0,1)$, the perturbed matrix $(1-\alpha)P + \alpha I_N$ can be expressed as

$$(1-\alpha)\,P = \big((1-\alpha)\cdot \mathbf{1}_N\big)\,\pi^\top, \tag{26}$$

i.e. $(1-\alpha)\,P$ can be expressed as an outer product of vectors in $\mathbb{R}^N$; namely of $\big((1-\alpha)\cdot\mathbf{1}_N\big)$ and $\pi^\top$. Consequentially, our matrix of interest can be written as

$$(1-\alpha)\,P + \alpha I_N = \big((1-\alpha)\cdot \mathbf{1}_N\big)\,\pi^\top + \alpha I_N. \tag{27}$$

Note that, $\alpha\, I_N$ is invertible since $\det(\alpha I_N) = \alpha^N > 0$. Thus, the main result of Bartlett [1951] can be applied, which yields the condition: if αI_N is invertible (which it is) and if

$$1 + \mathbf{1}_N^\top \big(\alpha\, I_N\big)^{-1} \pi \neq 0 \tag{28}$$

then $\alpha I_N + \big((1-\alpha)\cdot\mathbf{1}_N\big)\,\pi^\top$ is invertible. Thus, we only need to verify that the condition holds in our case. Simplifying equation (28) yields

$$\begin{aligned}
-1 \neq\, & \mathbf{1}_N^\top\big(\alpha\, I_N\big)^{-1}\pi && (29)\\
=& \frac{1}{\alpha}\,\mathbf{1}_N^\top \pi && (30)\\
=& \frac{1}{\alpha}\sum_{n=1}^N \pi_n && (31)\\
=& \frac{1}{\alpha}\,1 = \frac{1}{\alpha}, && (32)
\end{aligned}$$

where equation (32) held since $\pi \in \Delta_N$. Consequentially, the identity in equation (27) and the computation in equation (29)-equation (32) imply that $(1-\alpha)\,P + \alpha I_N$ is invertible if $\alpha \neq -1$.

Finally, since P is (row) stochastic and so is I_N then, for each $i = 1, \ldots, N$, we have that

$$\sum_{j=1}^N \big((1-\alpha)\,P_i + \alpha I_N\big)_j = (1-\alpha)\sum_{j=1}^N P_{i,j} + \alpha 1 = (1-\alpha)\,1 + \alpha = 1.$$

Whence, $(1-\alpha)P + \alpha I_N$ is (row) stochastic also. □

We now provide a set of *sufficient* condition on α, guaranteeing that the principal logarithm of P_t^α is well-defined. Furthermore, this lemme also shows that for $\alpha \in (0,1)$ large enough, as a function of N, the matrix $\log(P_t^\alpha)$ is necessarily a valid candidate for a Markov transition matrix (i.e. each of its rows sum to 0).

Lemma 2 (Sufficient Condition for Existence). *If $\alpha \geq 1 - \frac{1}{N^2}$ then if either of the following holds:*

(i) ***Real Case:*** *P_t^α has no complex eigenvalues,*

(ii) ***Complex Case:*** *P is doubly stochastic (i.e. row and column stochastic),*

then $\log(P_t^\alpha)$ *exists and its rows sum to* 0.

Proof of Lemma 2. Since P_t^α is an $N \times N$ real (thus complex) matrix with real eigenvalues, then define

$$m \overset{\text{def.}}{=} \mathrm{tr}(P_t^\alpha)/N \text{ and } s^2 \overset{\text{def.}}{=} \mathrm{tr}((P_t^\alpha)^2)/N - m^2. \tag{33}$$

First, observe that since the entries of P_t (in particular its diagonal elements) are all positive then

$$\begin{aligned} m &= \mathrm{tr}(P_t^\alpha)/N & (34)\\ &= \frac{1}{N} \sum_{i=1}^{N} \big((1-\alpha)\,\pi_i + \alpha\big) & (35)\\ &= \frac{1}{N} \Big((1-\alpha) \sum_{i=1}^{N} \pi_i + \alpha \sum_{i=1}^{N} 1\Big) & (36)\\ &= \frac{1}{N} \Big((1-\alpha) + \alpha N\Big) & (37)\\ &= \frac{1}{N} \Big(1 + \alpha(N-1)\Big). & (38) \end{aligned}$$

Next, we compute s^2. By Lemma 1, we have that P_t^α is a stochastic matrix and, therefore, so is its square (as the product of stochastic matrices is stochastic). Note that

$$\mathrm{tr}((P_t^\alpha)^2) \leq \max_{S \in \mathrm{Stoch}(N)} \mathrm{tr}(S) = \mathrm{tr}(I_N) = N \tag{39}$$

where $\mathrm{Stoch}(N)$ is the set of $N \times N$ stochastic matrices. Therefore, we bound s^2, defined in equation (33) using the “extremal trace bound” in equation (39) via

$$\begin{aligned} s^2 &= \mathrm{tr}(P_t^\alpha)/N - m^2 & (40)\\ &\leq N/N - m^2 & (41)\\ &= 1 - \frac{1}{N^2}\left(1 + \alpha(N-1)\right)^2. & (42) \end{aligned}$$

That is

$$-s \geq -\big(1 - \frac{1}{N^2}\left(1 + \alpha(N-1)\right)^2\big)^{1/2}.$$

Now, using lower-bound on the minimal eigenvalue of a square complex matrix with real eigenvalues using m and s in [Wolkowicz & Styan, 1980, Theorem 2.1] we have that

$$\begin{aligned} \lambda_{\text{min}}(P_t^\alpha) \geq& m - s(N-1)^{1/2} & (43)\\ \geq& \frac{1}{N}\Big(1 + \alpha(N-1)\Big) - \Big(1 - \frac{1}{N^2}\left(1 + \alpha(N-1)\right)^2\Big)^{1/2} (N-1)^{1/2} & (44)\\ =& \frac{1}{N}\Big(1 + \alpha(N-1)\Big) - \Big(N^2 - \left(1 + \alpha(N-1)\right)^2\Big)^{1/2} \frac{(N-1)^{1/2}}{N} & (45)\\ =& \frac{1}{N}\Bigg(\Big(1 + \alpha(N-1)\Big) & (46)\\ & - \Big[(N-1)\Big((1-\alpha^2)N^2 + 2\alpha(\alpha-1)N - (\alpha-1)^2\Big)\Big]^{1/2}\Bigg). & (47) \end{aligned}$$

If $\alpha \geq 1 - \frac{1}{N^2}$ and $N > 1$ (which is always the case) then

$$\Bigg(\Big(1 + \alpha(N-1)\Big) - \Big[(N-1)\Big((1-\alpha^2)N^2 + 2\alpha(\alpha-1)N - (\alpha-1)^2\Big)\Big]^{1/2}\Bigg) > 0. \tag{48}$$

Therefore, equation (48) together with equation (46) imply that

$$\lambda_{\min}(P_t^\alpha) > 0$$

whenever $\alpha \geq 1 - \frac{1}{N^2}$.

Therefore, [Dunford & Schwartz, 1958, Theorem VII.1.10] implies that $\log(P_t^\alpha)$ exists, since P_t^α is a matrix whose spectrum does not contain $(-\infty, 0]$. Moreover, [Davies, 2010, Lemma 1] guarantees that the rows of $\log(P_t^\alpha)$ sum to 0.

Finally, we note that since I_N is doubly stochastic then so is $\bar{P_t^\alpha}$ provided that P is. Therefore, $\bar{P_t^\alpha}$ is also a stochastic matrix and so is the product $\bar{P_t^\alpha} P_t^\alpha$ (as the product of (row) stochastic matrices is again a (row) stochastic matrix). Whence

$$s_a^2 \overset{\text{def.}}{=} \operatorname{tr}(\overline{P_t^\alpha} P_t^\alpha)/N - m^2 \leq 1 - \frac{1}{N^2}\big(1 + \alpha(N-1)\big)^2 \tag{49}$$

and the same argument may be applied with s_a^2 in place of s^2 upon using [Wolkowicz & Styan, 1980, Theorem 3.1] in place of [Wolkowicz & Styan, 1980, Theorem 2.1]; however, in this case we do not need to assume that the eigenvalues of P_t^α are real. In either case, this concludes our proof. □

B.2.1 Completion of The Proof of Theorem 2

Proof of Theorem 2. **Step 1 - Minimizer of Inner Problem equation (Inner):**
Since the elements of the set of $N \times N$ uniform stochastic matrices $\mathcal{P}_N^U$ all have identical rows then, P is an optimizer of equation (Inner) if and only if its first row is a minimizer of

$$\min_{P \in \mathcal{P}_N^U} \sum_{n=1}^N P_{1,n}\, \ell(Y_t^{(n)}, Y_t) + \frac{1}{\lambda} \sum_{n=1}^N P_{1,n}\, \log(w_n/N). \tag{50}$$

Since the matrices in $\mathcal{P}_N^U$ are row-stochastic, then all their rows belong to the N simplex Δ_N. Therefore, P is a minimizer of equation (50) if and only if its first row, which we denote by $\pi \overset{\text{def.}}{=} (P_{1,1}, \dots, P_{1,N}) \in \Delta_N$ is a minimizer of

$$\min_{\pi \in \Delta_N} \sum_{n=1}^N \pi_j\, \ell(Y_t^{(n)}, Y_t) + \frac{1}{\lambda} \sum_{n=1}^N \pi_n\, \log(w_n/N). \tag{51}$$

Since the elements of Δ_N are in bijection with the set of probability measures on the N-point set $\{1, \dots, N\}$, $\lambda > 0$, and $\sum_{n=1}^N \pi_n\, \log(w_n/N)$ is the KL-divergence (relative entropy) between the probability measure $\sum_{n=1}^N \pi_n \delta_n$ and the uniform measure $\sum_{n=1}^N \frac{1}{N} \delta_N$ (both on $\{1, \dots, N\}$) then [Wang et al., 2020, Proposition 1] the unique minimizer of equation (51) is given by

$$\bar{\pi} \overset{\text{def.}}{=} \operatorname{Softmin}\big(\lambda\, (\ell(Y^{(n)}, Y))_{n=1}^N\big).$$

Consequentially, the matrix $P \in \mathcal{P}_N^U$ whose rows are π is a minimizer of equation (Inner).

Step 2 - Minimizer of Outer Problem equation (Outer):
By Proposition 2, for every $\alpha(0,1)$ the matrix

$$P^\alpha \overset{\text{def.}}{=} (1-\alpha)P + \alpha I_N$$

is row-stochastic and for α "large enough"; meaning for $\alpha \in (1 - 1/N, 1)$, the matrix P^α has all its eigenvalues in $(0, \infty)$. Therefore, by [Davies, 2010, Theorem 12] there is a minimizer of equation (Outer) and it is given in closed-form by

$$Q \overset{\text{def.}}{=} \operatorname{ReLU}\big(\log(P_t^\alpha)\big).$$

This concludes our proof. □

Proof of Proposition 2. By Lemma 1, the matrix P_t^α is (row) stochastic and invertible. Thus, it has no zero-eigenvalues. If, moreover, the assumption holds that P_t^α has no negative eigenvalues then [Dunford & Schwartz, 1958, Theorem VII.1.10] guarantees that the (principle branch) of the matrix logarithm of P_t^α exists. Consequentially, [Davies, 2010, Lemma 1] applies from which we deduce that the rows of $\log(P_t^\alpha)$ sum to 0. The last claim follows directly from Lemma 2 (i). □

B.3 Proof of the Stability Guarantee in Proposition 1

The following generalizes, thus implies, Proposition 1.

Lemma 3 (Maximal KL Divergence for Perturbation in Lemma 1)**.** *Let* $\pi \in \Delta_N$, $\alpha \in [0,1)$, $i = 1, \ldots, N$, *and let for each* $i = 1, \ldots, N$ *let* $\pi^{\alpha,i} = (1-\alpha)\pi + \alpha\, e_i$ *where* $\{e_i\}_{i=1}^N$ *is the standard basis of* $\mathbb{R}^N$. *If* $p_{\min} \overset{\text{def.}}{=} \min_{i=1,\ldots,N}\, p_i > 0$ *then*

$$\max_{i=1,\ldots,N} \operatorname{KL}(\pi|\pi^{\alpha,i}) \le 2\alpha\biggl(-\frac{\log(p_{\min})}{1/p_{\min}-1} - \frac{\log((1-\alpha)\,p_{\min})}{1/((1-\alpha)\,p_{\min})-1}\biggr).$$

Proof of Lemma 3. By the (sharp) reverse Pinsker inequality in [Binette, 2019, Theorem 1], as formulated in [Binette, 2019, Example A], yields the bound

$$\operatorname{KL}(\pi|\pi^{\alpha,i}) \le \operatorname{TV}(\pi|\pi^{\alpha,i})\biggl(\frac{\log(1/p_{\min})}{1/p_{\min}-1} + \frac{\log(1/p_{\min}^{\alpha,i})}{1/p_{\min}^{\alpha,i}-1}\biggr) \tag{52}$$

where TV is the total variation distance between π and $\pi^{\alpha,i}$, and $p_{\min}^{\alpha,i} = \min_{i=1,\ldots,N}\, \pi^{\alpha,i}$. By construction

$$(1-\alpha)p_{\min} \le p_{\min}^{\alpha,i} \le (1-\alpha)p_{\min} + \alpha.$$

Whence,

$$\frac{1}{p_{\min}^{\alpha,i}} \le \frac{1}{(1-\alpha)p_{\min}} \text{ and } \frac{1}{1/p_{\min}^{\alpha,i}-1} \le \frac{1}{1/((1-\alpha)p_{\min}+\alpha)-1}. \tag{53}$$

Incorporating equation (53) into equation (52) yields

$$\begin{aligned}
\operatorname{KL}(\pi|\pi^{\alpha,i}) \le& \operatorname{TV}(\pi,\pi^{\alpha,i})\biggl(-\frac{\log(p_{\min})}{1/p_{\min}-1} - \frac{\log((1-\alpha)\,p_{\min})}{1/((1-\alpha)\,p_{\min})-1}\biggr)\\
=& \sum_{j=1}^N |\pi_j - \pi_j^{\alpha,i}|\biggl(-\frac{\log(p_{\min})}{1/p_{\min}-1} - \frac{\log((1-\alpha)\,p_{\min})}{1/((1-\alpha)\,p_{\min})-1}\biggr)\\
=& \sum_{j=1}^N |\alpha\pi_j + \alpha I_{j=i}|\biggl(-\frac{\log(p_{\min})}{1/p_{\min}-1} - \frac{\log((1-\alpha)\,p_{\min})}{1/((1-\alpha)\,p_{\min})-1}\biggr)\\
\le& \biggl(\sum_{j=1}^N |\alpha\pi_j| + \alpha\biggr)\biggl(-\frac{\log(p_{\min})}{1/p_{\min}-1} - \frac{\log((1-\alpha)\,p_{\min})}{1/((1-\alpha)\,p_{\min})-1}\biggr)\\
\le& \biggl(\alpha\sum_{j=1}^N |\pi_j| + \alpha\biggr)\biggl(-\frac{\log(p_{\min})}{1/p_{\min}-1} - \frac{\log((1-\alpha)\,p_{\min})}{1/((1-\alpha)\,p_{\min})-1}\biggr)\\
\le& \biggl(\alpha\sum_{j=1}^N \pi_j + \alpha\biggr)\biggl(-\frac{\log(p_{\min})}{1/p_{\min}-1} - \frac{\log((1-\alpha)\,p_{\min})}{1/((1-\alpha)\,p_{\min})-1}\biggr) && (54)\\
\le& (\alpha+\alpha)\biggl(-\frac{\log(p_{\min})}{1/p_{\min}-1} - \frac{\log((1-\alpha)\,p_{\min})}{1/((1-\alpha)\,p_{\min})-1}\biggr) && (55)\\
=& 2\alpha\biggl(-\frac{\log(p_{\min})}{1/p_{\min}-1} - \frac{\log((1-\alpha)\,p_{\min})}{1/((1-\alpha)\,p_{\min})-1}\biggr).
\end{aligned}$$

where equation (54) held since $\pi \in \Delta_N$ and therefore, $\pi_i \ge 0$ for each $i = 1, \ldots, N$, and equation (55) held since $\sum_{i=1}^N \pi_i = 1$ again due to the fact that $\pi \in \Delta_N$. □

C Datasets and Benchmarks

C.1 NIFTY Dataset

The **N**ews-**I**nformed **F**inancial **T**rend **Y**ield (NIFTY) dataset Saqur et al. [2024] is a processed and curated daily news headlines dataset for the stock (US Equities) market price movement prediction task. NIFTY is comprised of two related datasets, NIFTY-LM and NIFTY-RL. In this section we outline the composition of the two datasets, and comment on additional details.

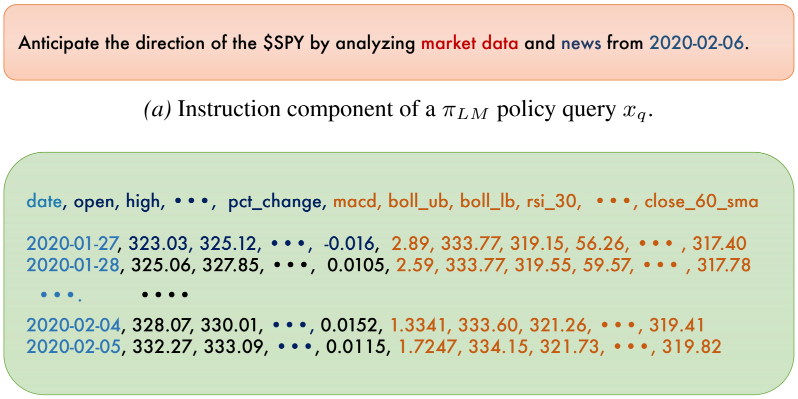
Dataset statistics . Table 6 and Table 7 present pertinent statistics related to the dataset.

Table 6: Statistics and breakdown of splits sizes

Category	Statistics
Number of data points	2111
Number of Rise/Fall/Neutral label	558 / 433 / 1122
Train/Test/Evaluation split	1477 / 317 / 317

Table 7: Date Ranges of news headlines in splits

Split	Num. Samples	Date range
Train	1477	2010-01-06 to 2017-06-27
Valid	317	2017-06-28 to 2019-02-12
Test	317	2019-02-13 to 2020-09-21



Anticipate the direction of the $SPY by analyzing market data and news from 2020-02-06.

(a) Instruction component of a π_{LM} policy query x_q.

```
date, open, high, • • •,  pct_change, macd, boll_ub, boll_lb, rsi_30,  • • •, close_60_sma

2020-01-27, 323.03, 325.12, • • •,  -0.016,  2.89, 333.77, 319.15, 56.26, • • • , 317.40
2020-01-28, 325.06, 327.85, • • •,  0.0105, 2.59, 333.77, 319.55, 59.57, • • • , 317.78
  • • •.          • • • •
2020-02-04, 328.07, 330.01, • • •, 0.0152, 1.3341, 333.60, 321.26, • • •, 319.41
2020-02-05, 332.27, 333.09, • • •, 0.0115, 1.7247, 334.15, 321.73, • • •, 319.82
```

(b) The market's **history** is provided as the past t days of numerical statistics like the (OHLCV) price (in blue) and common technical indicators (in orange) (e.g. moving averages) data.

Figure 6: Breaking down the instruction or prompt prefix, and market context components of a prompt, x_p.

C.1.1 NIFTY-LM: SFT Fine-tuning Dataset

The NIFTY-LM prompt dataset was created to finetune and evaluate LLMs on predicting future stock movement given previous market data and news headlines. The dataset was assembled by aggregating information from three distinct sources from January 6, 2010, to September 21, 2020. The compilation includes headlines from The **Wall Street Journal** and **Reuters News**, as well as market data of the $SPY index from **Yahoo Finance**. The NIFTY-LM dataset consists of:

- **Meta data**: Dates and data ID.
- **Prompt** (x_p): LLM question ($x_{question}$), market data from previous days ($x_{context}$), and news headlines (x_{news}).
- **Response**: Qualitative movement label (x_r) $\in \{Rise, Fall, Neutral\}$, and percentage change of the closing price of the $SPY index.

To generate LLM questions, ($\boldsymbol{x_{question}}$), the authors used the self-instruct Wang et al. [2023] framework and OpenAI GPT4 to create 20 synthetic variations of the instruction below:

> Create 20 variations of the instruction below.
> Examine the given market information and news headlines data on DATE to forecast whether the $SPY index will rise, fall, or remain unchanged. If you think the movement will be less than 0.5%, then return 'Neutral'. Respond with Rise, Fall, or Neutral and your reasoning in a new paragraph.

Where DATE would be substituted later, during the training phase with a corresponding date.

Context. The key 'context' ($\boldsymbol{x_{context}}$) was constructed to have newline delimited market metrics over the past T ($\approx$ 10) days (N.B. Not all market data for the past days for were available and therefore prompts might have less than 10 days of market metrics.).

Table 8 show the details of financial context provided in each day's sample.

Table 8: Summary of the dataset columns with their respective descriptions.

Column Name	**Description**
Date	Date of the trading session
Opening Price	Stock's opening market price
Daily High	Highest trading price of the day
Daily Low	Lowest trading price of the day
Closing Price	Stock's closing market price
Adjusted Closing Price	Closing price adjusted for splits and dividends
Volume	Total shares traded during the day
Percentage Change	Day-over-day percentage change in closing price
MACD	Momentum indicator showing the relationship between two moving averages
Bollinger Upper Band	Upper boundary of the Bollinger Bands, set at two standard deviations above the average
Bollinger Lower Band	Lower boundary, set at two standard deviations below the average
30-Day RSI	Momentum oscillator measuring speed and change of price movements
30-Day CCI	Indicator identifying cyclical trends over 30 days
30-Day DX	Indicates the strength of price trends over 30 days
30-Day SMA	Average closing price over the past 30 days
60-Day SMA	Average closing price over the past 60 days

News Headlines. $(\boldsymbol{x_{news}})$: Final list of filtered headlines from the aggregation pipeline. The non-finance related headlines were filtered out by performing a similarity search with SBERT model, "all-MiniLM-L6-v2" Reimers & Gurevych [2019]. Each headline was compared to a set of artificially generated financial headlines generated by GPT-4, with the prompt *"Generate 20 financial news headlines"*. Headlines with a similarity score below 0.2, were excluded from the dataset. To respect the prompting 'context length' of LLMs, in instances where the prompt exceeded a length of 3000 words, a further refinement process was employed. This process involved the elimination of words with a tf-idf Sammut & Webb [2010] score below 0.2 and truncating the prompt to a maximum of 3000 words.

It is also important to note that the dataset does not encompass all calendar dates within the specified time range. This limitation emanates from the trading calendar days, and absence of relevant financial news headlines for certain dates.

Label. $(\boldsymbol{x_r})$: The label is determined by the percentage change in closing prices from one day to the next, as defined in equation 56. This percentage change is categorized into three labels: {Rise, Fall, Neutral}, based on the thresholds specified in equation 57.

$$PCT_{\text{change}} = \left(\frac{\text{Closing Price}_t - \text{Closing Price}_{t-1}}{\text{Closing Price}_{t-1}} \right) \times 100\% \tag{56}$$

$$x_r = \begin{cases} \text{Fall} & \text{if } PCT_{\text{change}} < -0.5\% \\ \text{Neutral} & \text{if } -0.5\% \leq PCT_{\text{change}} \leq 0.5\% \\ \text{Rise} & \text{if } PCT_{\text{change}} > 0.5\% \end{cases} \tag{57}$$

C.2 NIFTY-RL: Preferences Dataset

The preference dataset is a variation of the fine-tuning dataset and it is designed for alignment training of LLMs using reward model. In NIFTY-RL, labels are omitted and replaced with chosen and rejected results. The chosen result is a label corresponding to a rise, a fall or neutral movement in the stock market and is equivalent to the response in NIFTY-LM. The rejected result is a random label not equal to the chosen label.

- **Metadata**: Includes dates and data identifiers.

- **Prompt** (x_p): Includes an LLM instruction ($x_{question}$), preceding market data ($x_{context}$), and relevant news headlines (x_{news}).
- **Chosen Result**: A qualitative movement label (x_r) from $\{Rise, Fall, Neutral\}$ indicating the predicted market trend.
- **Rejected Result**: A label ($\overline{x}_r$) randomly selected from $\{Rise, Fall, Neutral, Surrender\} \setminus \{x_r\}$, representing an incorrect market prediction.

C.3 FLARE BENCHMARK DATASETS

Stock Movement Prediction Datasets and Tasks: Flare-SM tasks. **FLARE** proposed by Xie et al. [2023], extends to include one financial prediction task – the **CIKM** dataset Wu et al. [2018] as an evaluation task among (four) other general financial NLP tasks. Under the hood, this benchmark is a fork of the '*lm-eval*' harness Gao et al. [2021] with addendums. Other stock price movement prediction from social dataset include what is referred to as *ACL18* (or, 'acl18') in this paper is essentially the **StockNet** Xu & Cohen [2018] dataset which comprises of stock tweets of 88 stock tickers from 9 financial market industries from Twitter over two years (from 2014-2015) aligned with their corresponding historical price data. **BigData22** [Soun et al., 2022] is another more recent tweets dataset comprising of tweets about 50 stock tickers during the period 2019-07-05 to 2020-06-30.

Table 9: Summary of Flare stock price movement datasets. The 'Stocks' column indicates the total number of different stock tickers referenced. The 'Tweets' and 'Days' columns represent the number of tweets and days respectively in each dataset.

Data	Stocks	Tweets	Days	Start Date	End Date
ACL18	87	106,271	696	2014-01-02	2015-12-30
BigData22	50	272,762	362	2019-07-05	2020-06-30
CIKM18	38	955,788	352	2017-01-03	2017-12-28

D ADDITIONAL BACKGROUND MATERIAL

In an effort to keep our paper as self-contained as possible, this section contains additional background material used in our derivations and in the formulations of our technical results.

D.1 MATRIX LOGARITHMS

The (principal) logarithm of an $N \times N$ matrix A whose spectrum does not contain $(-\infty, 0]$ in $\mathbb{C}$ is defined by

$$\log(A) \overset{\text{def.}}{=} \frac{1}{2\pi i} \int_\gamma \log(z)\,(zI_N - A)^{-1}\,dz$$

where $\log(z)$ is the principal logarithm of z (in the complex plane) and γ is a closed curve in $\mathbb{C} \setminus (-\infty, 0]$ containing the eigenspectrum of A.

D.2 THE SHIRAYEV-WONHAM FILTER

To keep our paper as self-contained as possible, we included some brief background on *stochastic filtering*. Namely, this appendix contains background material on the Shirayev-Wonham (stochastic) filter, studied in [Liptser & Shiryaev, 2001b, Chapter 9].

Consider a complete probability space $(\Omega, \mathcal{F}, P)$ equipped with a non-decreasing sequence of right-continuous sub-σ-algebras $\mathcal{F}_t, 0 \le t \le T$. Let $\theta = (\theta_t, \mathcal{F}_t)$, $0 \le t \le T$, denote a real right-continuous Markov process taking values in the countable set $E = \{\alpha, \beta, \gamma, \dots\}$. Additionally, let $W = (W_t, \mathcal{F}_t)$, $0 \le t \le T$, be a standard Wiener process independent of θ, and let ξ_0 be a $\mathcal{F}_0$-measurable random variable independent of θ. We assume the existence of nonanticipative functionals $A_t(\epsilon, x)$ and $B_t(x)$ that define

$$d\xi_t = A_t(\theta_t, \xi)dt + B_t(\xi)dW_t \tag{58}$$

and satisfy the following conditions.

$$A_t^2(\epsilon_t, x) \le L_1 \int_0^t (1 + x_s^2) dK(s) + L_2(1 + \epsilon_t^2 + x_t^2), \tag{59}$$

$$0 < C \le B_t^2(x) \le L_1 \int_0^t (1 + x_s^2) dK(s) + L_2(1 + x_t^2), \tag{60}$$

$$|A_t(\epsilon_t, x) - A_t(\epsilon_t, y)|^2 + |B_t(x) - B_t(y)|^2 \le L_1 \int_0^t (x_s - y_s)^2 dK(s) + L_2(x_t - y_t)^2, \tag{61}$$

where C, L_1, L_2 are certain constants, $K(s)$ is a non-decreasing right continuous function, $0 \le K(s) \le 1, x \in C_T, y \in C_T, \epsilon_t \in E, 0 \le t \le T$.

Along with Equations (59) to (61) it will also be assumed that

$$M\xi_0^2 < \infty, \tag{62}$$

and

$$M \int_0^T \theta_t^2 dt < \infty. \tag{63}$$

Define

$$p_\beta(t) \overset{\text{def.}}{=} P(\theta_t = \beta),$$
$$p_{\beta\alpha}(t, s) \overset{\text{def.}}{=} P(\theta_t = \beta | \theta_s = \alpha), \quad 0 \le s < t \le T, \quad \beta, \alpha \in E,$$

and assume there exist a function $\lambda_{\alpha\beta}(t), 0 \le t \le T, \alpha, \beta \in E$, that is

$$\text{continuous over } t, \text{ (uniformly over } \alpha, \beta) \tag{64}$$
$$|\lambda_{\alpha\beta}(t)| \le K \tag{65}$$
$$|p_{\beta\alpha}(t + \Delta, t) - \delta(\beta, \alpha) - \lambda_{\alpha\beta}(t) \cdot \Delta| \le o(\Delta), \tag{66}$$

where $\delta(\beta, \alpha)$ is a Kronecker's symbol and the value $o(\Delta)/\Delta \to 0$ as $\Delta \to 0$ (uniformly over α, β).

Let the Equations (59) to (66) be fulfilled. Then the a posteriori probability $\pi_\beta(t)$, $\beta \in \mathcal{E}$, satisfies a system of equations

$$\pi_\beta(t) = p_\beta(0) + \int_0^t \mathcal{L}^* \pi_\beta(u) du + \int_0^t \pi_\beta(u) \frac{A_u(\beta, \xi) - \bar{A}_u(\xi)}{B_u(\xi)} d\overline{W}_u,$$

where

$$\mathcal{L}^* \pi_\beta(u) = \sum_{\gamma \in \mathcal{E}} \lambda_{\gamma\beta}(u) \pi_\gamma(u) \quad \text{and} \quad \bar{A}_u(\xi) = \sum_{\gamma \in \mathcal{E}} A_u(\gamma, \xi) \pi_\gamma(u),$$

and $\overline{W} = (\overline{W}_t, \mathcal{F}_t)$ is a Wiener process with

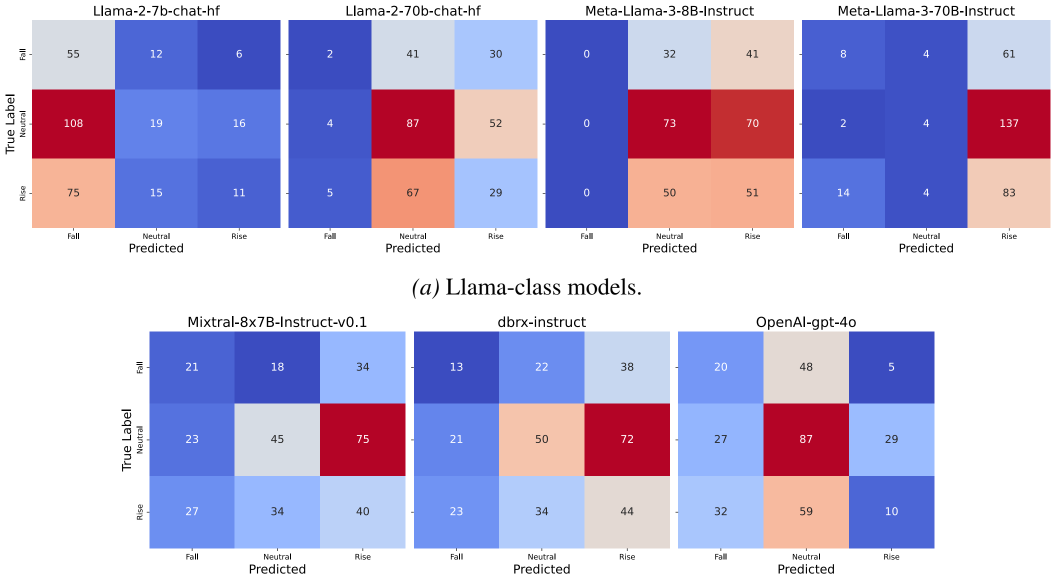
$$\overline{W}_t = \int_0^t \frac{d\xi_u - \bar{A}_u(\xi)}{B_u(\xi)} du.$$

E Additional Discussions

We have updated the anonymous git repo with all the experiment results and added script to easily replicate the main results of the paper in Table 2.

E.1 F1-Measure in FMM Task

For evaluation of model performances on the financial market movement (FMM) task using ternary class labels, we use the scikit-learn library methods. For multi-class, n, labels (with $n > 2$), the choice of *averaging* is important. The tabulated results were evaluated with default averaging set to "*weighted*" – where metrics for each label were calculated, then average weighted by their corresponding support (the number of true instances for each label). This alters the global averaging 'macro' (equal weight) for label imbalance. Imbalanced label support can result in an F-score that is lower or not in between precision and recall.



(a) Llama-class models.

(b) Mixture-of-experts class models and the state-of-the-art GPT-4 model.

Figure 7: Confusion matrices for Table 2. The first row highlights the Llama-class models, and the second row focuses on mixture-of-experts and GPT-4 models.

E.2 Time-Series Forecasting Experiments: Additional Details

This section provides additional discussion in support of the long-horizon time-series forecasting (LTSF) experiment covered in §5.2.

E.2.1 Forecasting granularity in Time-Series Forcasting: IMS vs. DMS

In LTSF works, the decoding granularity is dichotomized in the following two categories:

I) Iterated/Incremental Multi-step (IMS). : This is auto-regressive language-modeling or generative style prediction decoding where each step in prediction horizon H is iteratively predicted: $\hat{x}_{t+1}$, and is used for the prediction for the subsequent time-step. The common limitations of such forecasting granularity is 'error accumulation' over time as the decoder builds on the error from previous steps while iteratively making subsequent predictions. Additionally, the run-time complexity is to the order of the length of the horizon H.

II) Direct Multi-step (DMS). : As the name implies, in this decoding or forecasting approach, the entire forecasting horizon H is predicted at one go: $\hat{x}_{t+1:t+H}$. While computationally more attractive (much faster than IMS), this approach may not incorporate seasonality/periodicity in the time-series. Modern TSF specialist models, especially the transformer-based TSF architectures, tend to follow this scheme to avoid the quadratic cost (to the length of input) associated with attention.

E.2.2 Channel Independent Strategy

Recent advancements in Long-Term Series Forecasting (LTSF) have increasingly embraced a **Channel Independent (CI) approach** for handling multivariate time series data Han et al. [2024]. The CI strategy simplifies forecasting by isolating each (channel or feature as) univariate time series within the dataset, allowing the model to focus on predicting individual channels independently. Unlike traditional methods that leverage the entire multivariate historical data to make forecasts, the CI approach seeks a shared function $f : x^{(i)}_{t-L+1:t} \in \mathbb{R}^L \rightarrow \hat{x}^{(i)}_{t+1:t+H} \in \mathbb{R}^H$ for each univariate series, providing a streamlined model for each channel and reducing the need to account for inter-channel dependencies.

E.2.3 Our Setup

For our experiments, the historical observation window (*aka. look-back window* or *lag period*), L, is kept constant at 720 time-steps to be consistent and comparable with the literature. We follow the channel-independent strategy similar to the three expert models used for filtering - giving us C number of distinct features or channels of an agent's observation. The (MSE) loss is then measured as the discrepancy between the predicted values $\bar{x}^{(i)}_{t+1:t+H}$ and the ground truth $y^{(i)}_{t+1:t+H}$ as

$$\mathcal{L} = \frac{1}{C} \sum_{i=1}^{C} \left\| y^{(i)}_{t+1:t+H} - \bar{x}^{(i)}_{t+1:t+H} \right\|_2^2. \tag{67}$$