VERTEBRAL. In each realization, we randomly selected stream-size aligned data from testing-set and make it as online streaming data which is the input of each algorithm. Thus, we got independent result for each realization.

A small realization number would increase the variance of the results due to the randomness of stream order. A large realization number would make the result be more stable but at the cost of increasing computational cost (time, memory, etc.). We chose the realization number by balancing both aspects.

D Proofs for the Stochastic Setting

In this section, we focus on the stochastic setting. We first prove the regret bound presented in Theorem 1 and then prove the query complexity presented in Theorem 2 for Algorithm 1.

D.1 Proof of Theorem 1

Before providing the proof of Theorem 1, we first introduce the following lemma.

Lemma 8. *Fix $\tau \in (0,1)$. Let q_{t,i^*} be the probability of the optimal policy i^* maintained by Algorithm 1 at t, and let $b = p_{\min} \log_c(1/p_{\min})$, where $p_{\min} = \min_{s,i} \pi(\mathbf{x}_s)$ denotes the minimal model selection probability by any policy[6]. When $t \geq \left(\frac{\ln \frac{(|\Pi^*|-1)\tau}{1-\tau}}{\sqrt{\ln|\Pi^*|}\left(\Delta - \sqrt{\frac{2b^2}{t}\ln\frac{2}{\delta}}\right)} \right)^2$, with probability at least $1-\delta$, it holds that $q_{t,i^*} \geq \tau$.*

Proof of Lemma 8. W.l.o.g, we assume $\mu_1 \leq \mu_2 \leq \dots \mu_{n+k}$. Recall that we define $\Delta = \min_{i \neq i^*} \Delta_i = \mu_2 - \mu_1 = \frac{\mathbb{E}[\widetilde{L}_{t,2} - \widetilde{L}_{t,1}]}{t}$, and π_1 is the policy with the minimal expected loss.

Define

$$\delta_t \triangleq \tilde{\ell}_{t-1,i'} - \tilde{\ell}_{t-1,1}. \tag{7}$$

where $i' \triangleq \arg\min_{i \neq 1} \widetilde{L}_{t-1,i}$ denotes the index of the best empirical policy up to $t-1$ other than π_1. Therefore for $i \geq 2$, it holds that

$$\widetilde{L}_{t-1,i'} - \widetilde{L}_{t-1,i} = \sum_{s=1}^{t-1} \delta_s \leq 0.$$

We have $q_{t,i^*} = q_{t,1} = \frac{\exp(-\eta_t \widetilde{L}_{t-1,1})}{\sum_{i=1}^{|\Pi^*|} \exp(-\eta_t \widetilde{L}_{t-1,i})}$ as the weight of optimal expert at round t. Therefore

$$\begin{aligned} q_{t,i^*} = q_{t,1} &= \frac{\exp\left(-\eta_t \widetilde{L}_{t-1,1}\right)}{\sum_{i=1}^{|\Pi^*|} \exp\left(-\eta_t \widetilde{L}_{t-1,i}\right)} \\ &\overset{(a)}{=} \frac{\exp\left(-\eta_t \widetilde{L}_{t-1,1} + \eta_t \widetilde{L}_{t-1,i'}\right)}{\sum_{i=1}^{|\Pi^*|} \exp\left(-\eta_t \widetilde{L}_{t-1,i} + \eta_t \widetilde{L}_{t-1,i'}\right)} \\ &\overset{(b)}{=} \frac{\exp\left(\eta_t \sum_{s=1}^{t} \delta_s\right)}{\exp\left(\eta_t \sum_{s=1}^{t} \delta_s\right) + \sum_{i=2}^{|\Pi^*|} \exp\left(-\eta_t \widetilde{L}_{t-1,i} + \eta_t \widetilde{L}_{t-1,i'}\right)} \\ &\geq \frac{\exp\left(\eta_t \sum_{s=1}^{t} \delta_s\right)}{\exp\left(\eta_t \sum_{s=1}^{t} \delta_s\right) + |\Pi^*| - 1} \end{aligned} \tag{8}$$

where step (a) is by dividing the cumulative loss of sub-optimal policy $\pi_{i'}$ and step (b) is by the definition of δ_t in Equation (7).

[6]We assume $p_{\min} > 0$ per the policy regularization criterion in Appendix C.3. (cf. Algorithm 1 on "Regularized policy $\bar{\pi}(\mathbf{x}_t)$").

Let $\tau \in (0,1)$, such that $q_{t,i^*} \geq \frac{\exp\left(\eta_t \sum_{s=1}^{t} \delta_s\right)}{\exp\left(\eta_t \sum_{s=1}^{t} \delta_s\right)+|\Pi^*|-1} \geq \tau$. Plugging in $\eta_t = \sqrt{\frac{\ln |\Pi^*|}{t}}$ and define $\overline{\delta_t} = \frac{1}{t}\sum_{s=1}^{t} \delta_s$, we get

$$\frac{\exp\left(\sqrt{\ln |\Pi^*|}\sqrt{t}\cdot \overline{\delta_t}\right)}{\exp\left(\sqrt{\ln |\Pi^*|}\sqrt{t}\cdot \overline{\delta_t}\right)+|\Pi^*|-1} \geq \tau$$

Therefore, we obtain $\exp\left(\sqrt{\ln |\Pi^*|}\sqrt{t}\cdot \overline{\delta_t}\right) \geq \frac{(|\Pi^*|-1)\tau}{1-\tau}$. Rearranging the terms, we get

$$t \geq \left(\frac{\ln \frac{(|\Pi^*|-1)\tau}{1-\tau}}{\sqrt{\ln |\Pi^*|}\cdot \overline{\delta_t}}\right)^2$$

Next, we seek a high probability upper bound on $\overline{\delta}_t$. Denote $\Delta_i \triangleq \mu_i - \mu_1$ for $i \in 1, \ldots, |\Pi^*|$. We know

$$P(\overline{\delta}_t \leq \Delta_2 - \epsilon) \overset{\text{(a)}}{\leq} P(\overline{\delta}_t \leq \Delta_{i'} - \epsilon) = P\left(\frac{1}{t}\sum_{s=1}^{t} \delta_s - \Delta_{i'} \leq -\epsilon\right) \overset{\text{(b)}}{\leq} e^{-\frac{t\epsilon^2}{2b^2}} \tag{9}$$

Here, step (9a) is by the fact that $\Delta_2 = \min_{i\neq 1}\Delta_i \leq \Delta_{i'}$, and step (9b) is by Hoeffding's inequality where b denotes the upper bound on $|\delta_s|$. Further note that

$$\begin{aligned}\delta_{s+1} = \tilde{\ell}_{s,i'} - \tilde{\ell}_{s,1} = \frac{U_s}{z_s}\langle \pi_{i'}(\mathbf{x}_s) - \pi_1(\mathbf{x}_s), \mathbb{I}\{\hat{\mathbf{y}}_s \neq y_s\}\rangle &\leq \frac{\langle \pi_{i'}(\mathbf{x}_s), \mathbb{I}\{\hat{\mathbf{y}}_s \neq y_s\}\rangle}{z_s} \\ &\overset{\text{Eq. (4)}}{\leq} U_s \frac{\langle \pi_{i'}(\mathbf{x}_s), \mathbb{I}\{\hat{\mathbf{y}}_s \neq y_s\}\rangle}{\frac{1}{c}\sum_{y\in\mathcal{Y}}\langle \boldsymbol{w}_s, \mathbb{I}\{\hat{\mathbf{y}}_s \neq y\}\rangle \log_c \frac{1}{\langle \boldsymbol{w}_s, \mathbb{I}\{\hat{\mathbf{y}}_s \neq y\}\rangle}}\end{aligned}$$

Given $p_{\min} = \min_{s,i}\pi(\mathbf{x}_s)$, we obtain $\delta_{s+1} \leq \frac{1}{p_{\min}\log_c(1/p_{\min})}$ and similarly, $\delta_{s+1} \geq -\frac{\langle \pi_1(\mathbf{x}_s), \mathbb{I}\{\hat{\mathbf{y}}_s \neq y_s\}\rangle}{z_s} \geq -\frac{1}{p_{\min}\log_c(1/p_{\min})}$. We hence conclude that $|\delta_{s+1}| \leq b$.

Let $2e^{-\frac{t\epsilon^2}{2b^2}} = \delta$. Therefore, when $t \geq \left(\frac{\ln \frac{(|\Pi^*|-1)\tau}{1-\tau}}{\sqrt{\ln |\Pi^*|}(\Delta-\epsilon)}\right)^2 = \left(\frac{\ln \frac{(|\Pi^*|-1)\tau}{1-\tau}}{\sqrt{\ln |\Pi^*|}\left(\Delta - \sqrt{\frac{2b^2}{t}\ln\frac{2}{\delta}}\right)}\right)^2$, it holds that $q_{t,i^*} \geq \tau$ with probability at least $1-\delta$.

□

Lemma 9. *At round t, when* $t \geq \left(\frac{\ln\frac{|\Pi^*|-1}{\gamma} + \sqrt{\ln |\Pi^*|\cdot 2b^2 \ln\frac{2}{\delta}}}{\sqrt{\ln |\Pi^*|}\Delta}\right)^2$, *it holds that the arm chosen by the best policy i^* will be the arm chosen by Algorithm 1 with probability at least $1-\delta$. That is,* $\arg\max\left\{\sum_{i\in[|\Pi^*|]} q_{t,i}\pi_i(\mathbf{x}_t)\right\} = \arg\max\{\pi_{i^*}(\mathbf{x}_t)\}$.

Proof of Lemma 9. At round t, for Algorithm 1, we have loss $\sum_{j=1}^{k}\mathbb{I}\left\{j = \arg\max\left\{\sum_{i\in[|\Pi^*|]} q_{t,i}\pi_i(\boldsymbol{x}_t)\right\}\right\}\widehat{\ell}_{t,j}$. Let $q_{t,i^*} \geq \tau$. At round t, the best policy i^*'s top weight arm j_{t,i^*}'s probability $\max\{\pi_{i^*}(\boldsymbol{x}_t)\}$ is at least $\frac{1}{k}$. The second rank probability of $\pi_{i^*}(\boldsymbol{x}_t)$ is $\max_j\,[\pi_{i^*}(\boldsymbol{x}_t)]_{j\neq \mathrm{maxind}(\pi_{i^*}(\boldsymbol{x}_t))}$. Let us define

$$\begin{aligned}\gamma &:= \min_{\boldsymbol{x}_t}\left\{\max_{w_j\in \boldsymbol{w}^t_{i^*}} w_j - \max_{w_j\in \boldsymbol{w}^t_{i^*}, j\neq \mathrm{maxind}(\boldsymbol{w}^t_{i^*})} w_j\right\} \\ &= \max\{\pi_{i^*}(\boldsymbol{x}_t)\} - \max_j\left\{[\pi_{i^*}(\boldsymbol{x}_t)]_{j\neq \mathrm{maxind}(\pi_{i^*}(\boldsymbol{x}_t))}\right\},\end{aligned} \tag{10}$$

as the minimal gap in model distribution space of best policy. The arm recommended by the best policy i^* of CAMS will dominate CAMS's selection, when we have

$$q_{t,i^*}\cdot\max\left\{\pi_{i^*}(\boldsymbol{x}_t)\right\}\geq(1-q_{t,i^*})+q_{t,i^*}\left(\max_j\left[\pi_{i^*}(\boldsymbol{x}_t)\right]_{j\neq\text{maxind}(\pi_{i^*}(\boldsymbol{x}_t))}\right)\tag{11}$$

Rearranging the terms, and by

$$q_{t,i^*}\cdot\gamma\overset{\text{Eq. (10)}}{=}q_{t,i^*}\left(\max\left\{\pi_{i^*}(\boldsymbol{x}_t)\right\}-\max_j\left[\pi_{i^*}(\boldsymbol{x}_t)\right]_{j\neq\text{maxind}(\pi_{i^*}(\boldsymbol{x}_t))}\right)\geq(1-q_{t,i^*})$$

Therefore, we get $\tau\cdot(\gamma)\geq(1-\tau)$, and thus $\tau\geq\frac{1}{\gamma+1}$.

Set $\tau\geq\frac{1}{\gamma+1}$. By Lemma 8, we get

$$\begin{aligned}
t&\geq\left(\frac{\ln\frac{|\Pi^*-1|\tau}{1-\tau}}{\sqrt{\ln|\Pi^*|}(\Delta-\epsilon)}\right)^2\\
&\geq\left(\frac{\ln\left(\frac{|\Pi^*|-1}{\gamma}\right)}{\sqrt{\ln|\Pi^*|}(\Delta-\epsilon)}\right)^2\\
&\overset{(c)}{\geq}\left(\frac{\ln\frac{|\Pi^*|-1}{\gamma}}{\sqrt{\ln|\Pi^*|}\Delta-\sqrt{\ln|\Pi^*|\cdot\frac{2b^2}{t}\ln\frac{2}{\delta}}}\right)^2
\end{aligned}$$

where the last step is by applying $2e^{-\frac{t\epsilon^2}{2b^2}}=\delta$, thus, $\epsilon=\sqrt{\frac{2b^2}{t}\ln\frac{2}{\delta}}$. Dividing both sides by t

$$\begin{aligned}
1&\overset{(d)}{\geq}\left(\frac{\ln\frac{|\Pi^*|-1}{\gamma}}{\sqrt{\ln|\Pi^*|\cdot t}\Delta-\sqrt{\ln|\Pi^*|\cdot 2b^2\ln\frac{2}{\delta}}}\right)^2\\
\ln\frac{|\Pi^*|-1}{\gamma}&\leq\sqrt{t}\sqrt{\ln(|\Pi^*|)}\Delta-\sqrt{\ln(|\Pi^*|)\cdot 2b^2\ln\frac{2}{\delta}}\\
t&\geq\left(\frac{\ln\frac{|\Pi^*|-1}{\gamma}+\sqrt{\ln|\Pi^*|\cdot 2b^2\ln\frac{2}{\delta}}}{\sqrt{\ln|\Pi^*|}\Delta}\right)^2.
\end{aligned}$$

So, when $t\geq\left(\frac{\ln\frac{|\Pi^*|-1}{\gamma}+\sqrt{\ln|\Pi^*|\cdot 2b^2\ln\frac{2}{\delta}}}{\sqrt{\ln|\Pi^*|}\Delta}\right)^2$, it holds that $\arg\max\left\{\sum_{i\in[|\Pi^*|]}q_{t,i}\pi_i(\boldsymbol{x}_t)\right\}=\arg\max\left\{\pi_{i^*}(\boldsymbol{x}_t)\right\}$. □

Proof of Theorem 1. Therefore, with probability at least $1-\delta$, we get constant regret $\left(\frac{\ln\frac{|\Pi^*|-1}{\gamma}+\sqrt{\ln|\Pi^*|\cdot 2b^2\ln\frac{2}{\delta}}}{\sqrt{\ln|\Pi^*|}\Delta}\right)^2$.

Furthermore, with probability at most δ, the regret is upper bounded by T. Thus, we have

$$\begin{aligned}
\overline{\mathcal{R}}(T)&\leq(1-\delta)\left(\frac{\ln\frac{|\Pi^*|-1}{\gamma}+\sqrt{\ln|\Pi^*|\cdot 2b^2\ln\frac{2}{\delta}}}{\sqrt{\ln|\Pi^*|}\Delta}\right)^2+\delta T\\
&\overset{(a)}{\leq}\left(1-\frac{1}{T}\right)\left(\frac{\ln\frac{|\Pi^*|-1}{\gamma}+b\sqrt{\ln|\Pi^*|\cdot(2\ln T+2\ln 2)}}{\sqrt{\ln|\Pi^*|}\Delta}\right)^2+1\\
&=O\left(\frac{b\ln T}{\Delta^2}+\left(\frac{\ln\frac{|\Pi^*|-1}{\gamma}}{\sqrt{\ln|\Pi^*|}\Delta}\right)^2\right),
\end{aligned}$$

where step (a) by setting $\delta=\frac{1}{T}$, and where γ in Eq. (10) is the min gap. □