Proceedings of Machine Learning Research 157, 2021 ACML 2021

Lifelong Learning with Branching Experts (Appendix)

Appendix A. Proofs in Section 3

A.1. Proof of Theorem 1

Recall that based on the ε-cover assumption, for any expert i, his true loss $\ell_{\tau,i}$ at step τ differs from the surrogate loss $\bar{\ell}_{\tau,i} = \ell_{\tau,\pi_{\tau+1}(i)}$ by at most ε. As our algorithm computes the probability of playing expert i at a later step t based on such a loss $\bar{\ell}_{\tau,i}$ instead of $\ell_{\tau,i}$, it is simpler to use $\bar{\ell}_{\tau,i}$ as a surrogate loss for expert i at step τ and bound a surrogate regret in terms of such surrogate losses, which we will do in the following. It is straightforward to see that the true regret is only larger by at most $2\varepsilon T$.

Let us divide the T time steps into M consecutive intervals $I_1, I_2, \ldots, I_M$, such that the set of experts grows at the start but remains the same within each interval. For an interval I_m, let $I_m = [a_m, b_m]$, so that $a_1 = 1$, $b_M = T$, and $a_{m+1} = b_m + 1$ for $1 \leq m \leq M - 1$.

Note that the regret in each interval I_m can be bounded using standard approaches for the non-branching case. Here, we follow that in Section 2.5 of (Bubeck, 2011) for Hedge algorithm with time-varying learning rates.

Consider any interval I_m. For any $t \in I_m$, let

$$\Phi_t = \frac{1}{\eta_{t+1}} \log \left(\frac{1}{N_{t+1}} \sum_{i \in [N_{t+1}]} e^{-\eta_{t+1} \bar{L}_{t,i}} \right).$$

As shown in the proof of Theorem 2.4 in (Bubeck, 2011),

$$\langle p_t, \bar{\ell}_t \rangle \leq \Phi_{t-1} - \Phi_t + \eta_t \tag{1}$$

when both steps t and $t+1$ belong to I_m so that $N_t = N_{t+1}$ and $\eta_{t+1} = \sqrt{\frac{m}{t+1} \log N_{t+1}} \leq \eta_t$. For step $t = b_m$ ending interval I_m, the next step $t+1 = a_{m+1}$ belongs to interval I_{m+1} with $N_{t+1} > N_t$ experts, and the inequality (1) may not hold. In this case, with $t = b_m$, we define

$$\Phi'_t = \frac{1}{\eta'_{t+1}} \log \left(\frac{1}{N_t} \sum_{i \in [N_t]} e^{-\eta'_{t+1} \bar{L}_{t,i}} \right),$$

with $\eta'_{t+1} = \sqrt{\frac{m}{t+1} \log N_t} \leq \eta_t$, so that we can similarly have

$$\langle p_t, \bar{\ell}_t \rangle \leq \Phi_{t-1} - \Phi'_t + \eta_t.$$

For convenience later, we let $\Phi'_T = \Phi_T$ at the last step. Then, the following lemma bounds the expected loss of our algorithm in interval I_m, which we prove in Appendix A.1.1.

Lemma 1 $\sum_{t \in I_m} \langle p_t, \bar{\ell}_t \rangle \le \Phi_{b_{m-1}} - \Phi'_{b_m} + \sum_{t \in I_m} \eta_t$,

To bound the total expected loss of our algorithm, we sum the bound above over m to get

$$\sum_{t=1}^{T} \langle p_t, \bar{\ell}_t \rangle \le \sum_{m=1}^{M} \left(\Phi_{b_{m-1}} - \Phi'_{b_m} \right) + \sum_{t=1}^{T} \eta_t.$$

The second term on the righthand side above is at most

$$\sum_{t=1}^{T} \sqrt{\frac{M \log N}{t}} \le 2\sqrt{MT \log N},$$

while the first term can be decomposed as

$$\sum_{m=1}^{M} \left(\Phi_{b_{m-1}} - \Phi_{b_m} \right) + \sum_{m=1}^{M} \left(\Phi_{b_m} - \Phi'_{b_m} \right)$$

with the first sum above becomes $\Phi_0 - \Phi_T \le \bar{L}_{T,i^*} + \frac{\log N}{\eta_{T+1}}$. Finally, the theorem follows from the following lemma, which we prove in Appendix A.1.2.

Lemma 2 $\sum_{m=1}^{M} \left(\Phi_{b_m} - \Phi'_{b_m} \right) \le \mathcal{O}(\sqrt{TM \log N})$.

A.1.1. Proof of Lemma 1

Recall that I_m is the interval $[a_m, b_m]$, and $\langle p_t, \ell_t \rangle$ is at most $(\Phi_{t-1} - \Phi_t) + \eta_t$ for $t \in I_m \setminus \{b_m\}$ and is at most $(\Phi_{t-1} - \Phi'_t) + \eta_t$ for $t = b_m$. Therefore,

$$\begin{aligned} \sum_{t \in I_m} \langle p_t, \ell_t \rangle &\le \left(\sum_{t=a_m}^{b_m - 1} (\Phi_{t-1} - \Phi_t) \right) + \left(\Phi_{b_m - 1} - \Phi'_{b_m} \right) + \sum_{t \in I_m} \eta_t \\ &\le \Phi_{a_m - 1} - \Phi'_{b_m} + \sum_{t \in I_m} \eta_t \\ &= \Phi_{b_{m-1}} - \Phi'_{b_m} + \sum_{t \in I_m} \eta_t. \end{aligned}$$

A.1.2. Proof of Lemma 2

First note that for $m = M$, we have $b_M = T$ and $\Phi_T = \Phi'_T$. Now consider any $1 \le m \le M - 1$, let $t = b_m$, and recall that $\eta'_{t+1} = \sqrt{\frac{m \log N_t}{t+1}} < \sqrt{\frac{(m+1) \log N_{t+1}}{t+1}} = \eta_{t+1}$. Then we have

$$\begin{aligned} \Phi_t &= \frac{1}{\eta_{t+1}} \log \left(\frac{1}{N_{t+1}} \sum_{i \in [N_{t+1}]} e^{-\eta_{t+1} \bar{L}_{t,i}} \right) \\ &\le \frac{1}{\eta_{t+1}} \log \max_{i \in [N_{t+1}]} e^{-\eta_{t+1} \bar{L}_{t,i}} \\ &= \frac{1}{\eta'_{t+1}} \log \max_{i \in [N_{t+1}]} e^{-\eta'_{t+1} \bar{L}_{t,i}} \\ &\le \frac{1}{\eta'_{t+1}} \log \sum_{i' \in [N_t]} e^{-\eta'_{t+1} (\bar{L}_{t,i'} - 1)}, \end{aligned}$$

where the last line follows from the fact that for any expert $i \in [N_{t+1}]$ split from expert $i' \in [N_t]$,

$$\bar{L}_{t,i} = \bar{L}_{t-1,i'} + \ell_{t,i} \geq \bar{L}_{t-1,i'} + \ell_{t,i'} - 1 = \bar{L}_{t,i'} - 1.$$

This implies that for $t = b_m$,

$$\begin{aligned} \Phi_t - \Phi'_t &\leq \frac{1}{\eta'_{t+1}} \log \sum_{i' \in [N_t]} e^{-\eta'_{t+1}(\bar{L}_{t,i'}-1)} - \Phi'_t \\ &= \frac{1}{\eta'_{t+1}} \log \left(N_t e^{\eta'_{t+1}} \right) \\ &= \frac{1}{\eta'_{t+1}} \log N_t + 1 \\ &\leq \mathcal{O}\left(\sqrt{\frac{T \log N_t}{m}} \right). \end{aligned}$$

By summing over m, we have

$$\begin{aligned} \sum_{m=1}^{M-1} \left(\Phi_{b_m} - \Phi'_{b_m} \right) &\leq \sum_{m=1}^{M-1} \mathcal{O}\left(\sqrt{\frac{T \log N}{m}} \right) \\ &\leq \mathcal{O}\left(\sqrt{TM \log N} \right), \end{aligned}$$

which completes the proof of the lemma.

A.2. Proof of Theorem 2

Let us first focus on the case where $\varepsilon \leq \Delta/8$. The following lemma shows that a suboptimal expert is unlikely to have small accumulated loss, which we prove in Appendix A.2.1.

Lemma 3 *For any expert i with gap Δ_i and any time t,*

$$\Pr\left[\bar{L}_{t,i} - \bar{L}_{t,i^*} \leq \frac{1}{4}\Delta_i t \right] \leq e^{-\Omega(\Delta_i^2 t)}.$$

This implies that a suboptimal expert i with gap Δ_i is unlikely to be played by our Hedge-based algorithm after some time step $\tilde{t}_i$, and here we choose

$$\tilde{t}_i = \frac{c}{\Delta_i^2} \log N, \text{ for a large enough constant } c.$$

Recall that in the non-branching case, (Mourtada and Gaïffas, 2019) relied on the regret bound in adversarial setting to bound the regret before the last of these steps $\tilde{t}_i$'s. In our case with branching experts, we cannot afford to do so as the adversarial bound has an extra $\sqrt{M}$ factor which we would like to avoid. Instead, we will take a more careful analysis, by partitioning experts into groups and bounding their regrets separately. More precisely, for any $r \geq 0$, let

$$A_r = \{ i : \Lambda_r \leq \Delta_i \leq 2\Lambda_r \} \text{ where } \Lambda_r = 2^r \Delta,$$

and let $\tilde{r} = \max_{i \in A_r} \tilde{t}_i$. In the following, we will use the notation $q_{t,i}$ to denote the probability that our algorithm plays expert i in time t, with $q_{t,i} = 0$ for $i \notin H_t$.

The next lemma bounds the pseudo-regret before step $\tilde{r}$ for experts in each group A_r, which we prove in Appendix A.2.2.

Lemma 4 *For any r, $\sum_{t \le \tilde{r}} \sum_{i \in A_r} \Delta_i \cdot q_{t,i} \le \mathcal{O}(\frac{\log N}{\Lambda_r})$.*

To bound the pseudo-regret after step $\tilde{r}$, we rely on the following lemma, which we prove in Appendix A.2.3.

Lemma 5 *For any r, $\sum_{t > \tilde{r}} \sum_{i \in A_r} \Delta_i \cdot q_{t,i} \le \mathcal{O}(\frac{1}{\Lambda_r})$.*

By summing the regret bounds in the last two lemmas, we can conclude that the total pseudo-regret is at most

$$\sum_{r \ge 0} \mathcal{O}\left(\frac{\log N}{\Lambda_r}\right) \le \sum_{r \ge 0} \mathcal{O}\left(\frac{\log N}{2^r \Delta}\right) \le \mathcal{O}\left(\frac{\log N}{\Delta}\right).$$

Note that the bound in this case is bad when the gap Δ is small. However, by a standard approach, one can show that the regret is still at most $\mathcal{O}(\sqrt{T \log N})$, by considering some threshold $\Delta' = \sqrt{(\log N)/T}$ for gaps. This is because those experts with gaps at most Δ' contribute a total regret of at most $\Delta' T \le \mathcal{O}(\sqrt{T \log N})$, while those remaining, if any, contribute at most $\mathcal{O}(\frac{\log N}{\Delta'}) \le \mathcal{O}(\sqrt{T \log N})$.

For the second case, we rely on the following, which we prove in Appendix A.2.4. The theorem then follows by combining the bounds in these two cases.

Lemma 6 *If $\varepsilon > \Delta/8$, then the regret of our algorithm is at most $\mathcal{O}(\sqrt{T \log N} + \varepsilon T)$.*

A.2.1. Proof of Lemma 3

Recall that for any expert j, $\bar{L}_{t,j} = \sum_{\tau=1}^{t} \bar{\ell}_{\tau,j}$ where $\bar{\ell}_{\tau,j} = \ell_{\tau,\pi_{\tau+1}(j)}$ with $\pi_\tau(j)$ being the expert representing expert j at step τ, and we have $|L_{t,j} - \bar{L}_{t,j}| \le \varepsilon t$ by the ε-cover assumption. Now consider any expert i with gap Δ_i so that $\mathbb{E}[L_{t,i} - L_{t,i^*}] = \Delta_i t$, where

$$i^* = \arg\min_{i \in H} \mathbb{E}[\ell_{t,i}]$$

denotes the optimal expert in H. By Hoeffding bound, we have

$$\Pr\left[L_{t,i} - L_{t,i^*} \le \frac{1}{2}\Delta_i t\right] \le e^{-\Omega(\Delta_i^2 t)}.$$

Then the lemma follows as $\bar{L}_{t,i} - \bar{L}_{t,i^*}$ and $L_{t,i} - L_{t,i^*}$ differ by at most $2\varepsilon t \le \frac{1}{4}\Delta_i t$, under the assumption that $\varepsilon \le \Delta_i/8$.

A.2.2. Proof of Lemma 4

For any r, we have

$$\sum_{t \le \tilde{r}} \sum_{i \in A_r} \Delta_i \cdot q_{t,i} \le \sum_{t \le \tilde{r}} 2\Lambda_r \le \frac{2c}{\Lambda_r} \log N,$$

as $\tilde{r} \le \frac{c}{\Lambda_r^2} \log N$.

A.2.3. PROOF OF LEMMA 5

Recall the definition that $i^* = \arg\min_{i\in H} \mathbb{E}[\ell_{t,i}]$, $\Delta_i = \mathbb{E}[\ell_{t,i} - \ell_{t,i^*}]$, and $\Delta = \min_{i\neq i^*} \Delta_i$, as well as

$$A_r = \{i \in H : \Lambda_r \le \Delta_i \le 2\Lambda_r\} \text{ where } \Lambda_r = 2^r\Delta.$$

Consider any r and note that any expert $i \in A_r$ has $\Lambda_r \le \Delta_i \le 2\Lambda_r$. Thus for any t, we have

$$\sum_{i\in A_r} \Delta_i \cdot q_{t,i} \le 2\Lambda_r \sum_{i\in A_r} q_{t,i},$$

and we can focus on bounding the sum $\sum_{i\in A_r} q_{t,i}$, which is the probability that our algorithm plays an expert in A_r at step t.

Consider any step $t > \tilde{r}$, and let B_t denote the event that

$$\exists i \in A_r \text{ such that } \bar{L}_{t-1,i} - \bar{L}_{t-1,i^*} \le \frac{1}{4}\Delta_i(t-1),$$

which by Lemma 3 and a union bound is at most

$$\sum_{i\in A_r} e^{-\Omega\left(\Delta_i^2 t\right)} \le N \cdot e^{-\Omega\left(\Lambda_r^2 \tilde{r}\right)} \cdot e^{-\Omega\left(\Lambda_r^2 (t-\tilde{r})\right)}.$$

Then using the fact that $\tilde{r} \ge \frac{c}{(2\Lambda_r)^2}\log N$ for a large enough constant c, we obtain

$$\Pr\left[B_t\right] \le e^{-\Omega\left(\Lambda_r^2(t-\tilde{r})\right)}.$$

Now assume that the event B_t does not happen. Recall that our algorithm plays any active expert i in A_r with probability

$$\frac{e^{-\eta_t \bar{L}_{t-1,i}}}{\sum_j e^{-\eta_t \bar{L}_{t-1,j}}} \le e^{-\eta_t\left(\bar{L}_{t-1,i}-\bar{L}_{t-1,i^*}\right)} \le e^{-\eta_t \frac{1}{4}\Delta_i(t-1)}.$$

When there are N_t active experts, this is at most

$$\begin{aligned} e^{-\Omega\left(\sqrt{\Lambda_r^2 t \log N_t}\right)} &\le e^{-\Omega\left(\sqrt{\Lambda_r^2 \tilde{r} \log N_t}\right)} \cdot e^{-\Omega\left(\sqrt{\Lambda_r^2 t}\right)} \\ &\le \frac{1}{N_t} \cdot e^{-\Omega\left(\sqrt{\Lambda_r^2 t}\right)}, \end{aligned}$$

which implies that the contribution from all the active experts in A_r is at most

$$N_t \cdot \frac{1}{N_t} \cdot e^{-\Omega\left(\sqrt{\Lambda_r^2 t}\right)} \le e^{-\Omega\left(\sqrt{\Lambda_r^2 t}\right)}.$$

Note that this holds for any possible values of N_t, conditioned on the event B_t not happening. Therefore, we can conclude that for any step $t > \tilde{r}$,

$$\begin{aligned} \sum_{i\in A_r} q_{t,i} &\le \Pr\left[B_t\right] + \left(1 - \Pr\left[B_t\right]\right) e^{-\Omega\left(\sqrt{\Lambda_r^2 t}\right)} \\ &\le e^{-\Omega\left(\Lambda_r^2(t-\tilde{r})\right)} + e^{-\Omega\left(\sqrt{\Lambda_r^2 t}\right)}. \end{aligned}$$

Finally, by summing the bound over $t > \tilde{r}$ and using the same analysis from the proof of Theorem 2 in (Mourtada and Gaïffas, 2019), which is based on the inequalities that $\sum_{t\geq 1} 2^{-\alpha t} \leq \mathcal{O}(1/\alpha)$ and $\sum_{t\geq 1} 2^{-\alpha\sqrt{t}} \leq \mathcal{O}(1/\alpha^2)$, for any $\alpha > 0$, we obtain

$$\sum_{t>\tilde{r}} \sum_{i\in A_r} \Delta_i \cdot q_{t,i} \leq 2\Lambda_r \cdot \mathcal{O}\left(\frac{1}{\Lambda_r^2}\right) \leq \mathcal{O}\left(\frac{1}{\Lambda_r}\right).$$

This proves the lemma.

A.2.4. Proof of Lemma 6

Recall that in the proof of Theorem 2, we have shown that in the first case when the smallest gap Δ satisfies the condition $\varepsilon \leq \Delta/8$, then the regret is at most $\mathcal{O}(\frac{\log N}{\Delta})$. Now we bound the regret for the other case, with $\varepsilon > \Delta/8$. Note that those experts with gaps at most 8ε only contribute a total regret of at most $8\varepsilon T$. For those remaining experts with larger gaps, we can apply the bound in the first case to bound their regret by $\mathcal{O}(\frac{\log N}{\varepsilon})$. Thus, the total regret in this case is at most $\mathcal{O}(\frac{\log N}{\varepsilon} + \varepsilon T)$. Then let us consider two subcases, depending on whether or not $\varepsilon > \Delta'/8$.

In the first subcase, when $\varepsilon > \Delta'/8$, the regret upper bound above becomes $\mathcal{O}(\sqrt{T\log N} + \varepsilon T)$. In the second subcase, with $\varepsilon \leq \Delta'/8$, we can apply the upper bound in the first case to bound the regret of those experts, if any, with gaps at least Δ' by $\mathcal{O}(\frac{\log N}{\Delta'}) \leq \mathcal{O}(\sqrt{T\log N})$. Those experts with gaps at most Δ' only contribute a total regret of at most $T \cdot \Delta' = \sqrt{T\log N}$. Thus, in both subcases, the regret is at most $\mathcal{O}(\sqrt{T\log N} + \varepsilon T)$. This proves the lemma.

A.3. Proof of Theorem 3

The key lemma to prove the theorem is the following, which we prove in Appendix A.3.1. As we will also use it later in different settings, we describe it in a slightly more general form, by considering any loss vectors f_t's, instead of just the surrogate loss vectors $\bar{\ell}_t$'s used by Algorithm 2, as well as bounding the regret starting from any step against any expert.

Lemma 7 *Suppose we run Algorithm 2 using the loss vectors f_t's and consider the regret with respect to them against an expert i starting from some time step s. Let p_t be the distribution it plays at step t, let $r_{t,i} = \mathbb{E}_{j\sim p_t}[f_{t,j}] - f_{t,i}$ be the regret at step t, and let $V = \sum_{t=s}^{T} r_{t,i}^2$. Let M_i be the number of branching steps of i's representatives starting from step s. Then*

$$\sum_{t=s}^{T} r_{t,i} \leq \mathcal{O}\left(\sqrt{V(M_i \log N + \log(NT))}\right).$$

Applying this lemma with the surrogate loss vectors $\bar{\ell}_t$'s and noting that $V \leq T$, we obtain a regret bound of $\mathcal{O}(\sqrt{T(M_i \log N + \log(NT))}$ with respect to such losses. By incorporating the approximation errors based on the ε-cover assumption, we can conclude that the regret with respect to the true losses is larger by at most $\mathcal{O}(\varepsilon T)$, which proves the theorem.

A.3.1. Proof of Lemma 7

For any t, let $W_t = \sum_{i,\eta} W_{t,i}^{\eta}$. Recall that

$$\hat{W}_{t+1,i}^{\eta} = W_{t,i}^{\eta}(1 + \eta r_{t,i}) \text{ and } \tilde{W}_{t+1,i}^{\eta} = \alpha_{t+1} \cdot \frac{1}{\hat{N}_{t+1}} + (1 - \alpha_{t+1}) \cdot \hat{W}_{t+1,i}^{\eta}.$$

Then by definition, we have

$$W_{t+1} = \sum_{i,\eta} W^{\eta}_{t+1,i} = \sum_{i,\eta} \tilde{W}^{\eta}_{t+1,i} \leq \alpha_{t+1} + \sum_{i,\eta} W^{\eta}_{t,i} + \sum_{i,\eta} \eta W^{\eta}_{t,i} r_{t,i},$$

with the last sum above being

$$\sum_{i,\eta} \eta W^{\eta}_{t,i} r_{t,i} = \left(\sum_{i',\eta'} \eta' W^{\eta'}_{t,i'} \right) \sum_{i} p_{t,i} \left(\mathop{\mathbb{E}}_{j\sim p_t} [f_{t,j}] - f_{t,i} \right) = 0,$$

since $\mathbb{E}_{j\sim p_t}[f_{t,j}] = \sum_j p_{t,j} f_{t,j}$. Consequently, we have

$$W_{t+1} \leq \alpha_{t+1} + W_t \leq \cdots \leq \sum_{\tau=1}^{t+1} \alpha_\tau + W_0 \leq \mathcal{O}(1). \tag{2}$$

On the other hand, for each new (i, η) branching out from some (i', η'), we have

$$W^{\eta}_{t+1,i} = \tilde{W}^{\eta'}_{t+1,i'} / |C_{i',\eta'}| \geq N^{-1} 2^{-1} \tilde{W}^{\eta'}_{t+1,i'},$$

which implies that

$$W^{\eta}_{t+1,i} \geq N^{-M_i} 2^{-\log_2 t} \left(\prod_{\tau=s}^{t} (1 - \alpha_{\tau+1})(1 + \eta r_{\tau,i}) \right) W^{\eta}_{s,i},$$

and hence

$$\begin{aligned}
\ln W^{\eta}_{t+1,i} &\geq -\mathcal{O}(M_i \log N + \log t) - \sum_{\tau=s}^{t} (\alpha_{\tau+1} + \alpha^2_{\tau+1}) + \sum_{\tau=s}^{t} (\eta r_{\tau,i} - \eta^2 r^2_{\tau,i}) + \ln W^{\eta}_{s,i} \\
&\geq -\mathcal{O}(M_i \log N + \log(Nt)) + \eta \sum_{\tau=s}^{t} (r_{\tau,i} - \eta r^2_{\tau,i}),
\end{aligned}$$

as $\ln W^{\eta}_{s,i} \geq \ln(\frac{1}{2N} \cdot \frac{\alpha_s}{\hat{N}_s}) \geq -\mathcal{O}(\log(Nt))$. Combining this with the bound

$$\ln W^{\eta}_{t+1,i} \leq \ln W_{t+1} \leq \mathcal{O}(1)$$

from Eq. (2), we have

$$\sum_{\tau=s}^{t} r_{\tau,i} \leq \eta \sum_{\tau=s}^{t} r^2_{\tau,i} + \frac{\mathcal{O}(M_i \log N + \log(Nt))}{\eta},$$

for any t, i and η. For $t = T$, this implies the existence of some $\eta \in Q_T$ with

$$\eta = \Theta\left(\sqrt{\frac{M_i \log N + \log(NT)}{V}} \right)$$

such that the regret starting from step s against expert i can be bounded as

$$\sum_{\tau=s}^{T} r_{\tau,i} \leq \mathcal{O}\left(\sqrt{V(M_i \log N + \log(NT))} \right),$$

which proves the lemma.

A.4. Proof of Theorem 4

Let us start from the case that

$$\varepsilon \leq \frac{\Delta}{8} \text{ and } \Delta \geq \sqrt{\frac{\log N}{T}}. \tag{3}$$

We will rely on the following nice property in stochastic setting that the optimal expert i^* appears early and it soon makes no more branching. More precisely, consider the following event C_t, for any step t.

- C_t: step t is the first time step such that the best expert i^* branches out before it and none branches out from i^* since then.

Then we have the following lemma which we prove in Appendix B.2.1.

Lemma 8 *There is some step $s_0 \leq \mathcal{O}(\frac{\log(N/\Delta)}{\Delta})$ such that for any $t \geq s_0$, $\Pr[C_t] \leq \frac{1}{t^3}$.*

Let us first consider the regret with respect to the surrogate losses $\bar{\ell}_t$'s starting from step s_0, based on Lemma 7. The key observation is that when the event C_t happens, there is no branching from the optimal expert i^* after step t, and we can bound the expected regret afterward by by Lemma 7 with $M_{i^*} = 0$ and

$$V = \sum_{t \geq s_0} \left(\mathop{\mathbb{E}}_{i \sim p_t} [\bar{\ell}_{t,i}] - \bar{\ell}_{t,i^*} \right)^2,$$

while we can simply bound the regret before by t. As the event C_t is only determined by the losses before step t and the losses afterwards are independent, we have

$$\begin{aligned}
\sum_{t \geq s_0} \left(\mathop{\mathbb{E}}_{i \sim p_t} [\bar{\ell}_{t,i}] - \bar{\ell}_{t,i^*} \right) &\leq \sum_{t \geq s_0} \Pr[C_t] \cdot \left(t + \mathcal{O}\left(\mathbb{E}\left[\sqrt{V \log(NT)} \right] \right) \right) \\
&\leq \sum_{t \geq s_0} \frac{t}{t^3} + \sum_{t \geq s_0} \Pr[C_t] \cdot \mathcal{O}\left(\sqrt{\mathbb{E}[V] \log(NT)} \right) \\
&\leq \mathcal{O}(1) + \mathcal{O}\left(\sqrt{\mathbb{E}[V] \log(NT)} \right) \\
&\leq \mathcal{O}\left(\sqrt{\mathbb{E}[V] \log(NT)} \right),
\end{aligned}$$

since the events C_t's are disjoint from each other. It remains to bound $\mathbb{E}[V]$, for which we follow the approach in the proof of Theorem 11 in (Gaillard et al., 2014). Let the random variable S be the number of steps after s_0 such that suboptimal experts are played, and we have $\mathbb{E}[V] \leq \mathbb{E}[S]$. To bound $\mathbb{E}[S]$, note that with $\varepsilon \leq \frac{\Delta}{8}$, we have

$$\mathbb{E}\left[\bar{\ell}_{t,i} - \bar{\ell}_{t,i^*} \right] \geq \Delta - 2\varepsilon \geq \frac{6\Delta}{8},$$

for any t and $i \neq i^*$, which implies that the regret above is at least $\mathbb{E}[S] \cdot \Omega(\Delta)$. As a result, we have

$$\mathbb{E}[S] \cdot \Omega(\Delta) \leq \mathcal{O}\left(\sqrt{\mathbb{E}[S] \log(NT)} \right),$$

which implies that

$$\mathbb{E}[S] \leq \mathcal{O}\left(\frac{\log(NT)}{\Delta^2} \right).$$

Thus, we can conclude that

$$\sum_{t \geq s_0} \left(\mathop{\mathbb{E}}_{i \sim p_t} [\bar{\ell}_{t,i}] - \bar{\ell}_{t,i^*} \right) \leq \mathcal{O}\left(\sqrt{\mathbb{E}[S] \log(NT)} \right) \leq \mathcal{O}\left(\frac{\log(NT)}{\Delta} \right).$$

Recall that our goal is to bound the regret with respect to the true losses, instead of the surrogate losses above, and a simple analysis would introduce an additional $\mathcal{O}(\varepsilon T)$ term which we would like to avoid. Note that $|\bar{\ell}_{t,i} - \ell_{t,i}| = 0$ if $i = i^*$ and $|\bar{\ell}_{t,i} - \ell_{t,i}| \leq \varepsilon$ otherwise (by the ε-cover assumption). Therefore, the additional regret for going from the surrogate losses to the true ones after step s_0 has an expected value of at most

$$\mathbb{E}[S] \cdot \varepsilon \leq \mathcal{O}\left(\frac{\varepsilon(\log(NT))}{\Delta^2} \right) \leq \mathcal{O}\left(\frac{\log(NT)}{\Delta} \right).$$

As a result, the total expected pseudo regret is at most

$$s_0 + \mathcal{O}\left(\frac{\log(NT)}{\Delta} \right) + \mathbb{E}[S] \cdot \varepsilon \leq \mathcal{O}\left(\frac{\log(N/\Delta)}{\Delta} \right) + \mathcal{O}\left(\frac{\log(NT)}{\Delta} \right) \leq \mathcal{O}\left(\frac{\log(NT)}{\Delta} \right),$$

since $\Delta \geq 1/T$ according to the condition in (7).

The case when the condition does not hold in (7) can be handled using the same approach as in the proof of Theorem 2. This completes the proof of Theorem 4.

A.4.1. PROOF OF LEMMA 8

For any step t, consider the following event

- B_t: there exists some $i \neq i^*$ such that for any $\tau \leq t-2$, $|\ell_{\tau,i} - \ell_{\tau,i^*}| \leq \varepsilon$.

Then clearly $\Pr[C_t] \leq \Pr[B_t]$. To bound each $\Pr[B_t]$, note that for any $i \neq i^*$ and any step τ,

$$\mathbb{E}[\ell_{\tau,i} - \ell_{\tau,i^*}] \geq \Delta,$$

so that with the condition $\varepsilon \leq \frac{\Delta}{8}$, we have

$$\Pr[|\ell_{\tau,i} - \ell_{\tau,i^*}| \leq \varepsilon] \leq \Pr[\ell_{\tau,i} - \ell_{\tau,i^*} \leq \varepsilon] \leq \frac{1-\Delta}{1-\varepsilon} \leq 1 - (\Delta - \varepsilon) \leq 1 - \frac{7}{8}\Delta.$$

As the loss functions are independent from each other, we have

$$\begin{aligned} \Pr[B_t] &\leq \sum_{i \neq i^*} \Pr[\forall \tau \leq t-2 : |\ell_{\tau,i} - \ell_{\tau,i^*}| \leq \varepsilon] \\ &\leq N\left(1 - \frac{7}{8}\Delta\right)^{t-2} \\ &\leq \frac{1}{t^3} \end{aligned}$$

when $t \geq s_0$ for some $s_0 \leq \mathcal{O}(\frac{\log(N/\Delta)}{\Delta})$. This proves the lemma.

Appendix B. Proofs in Section 4

B.1. Proof of Theorem 5

Note that the expected loss of the algorithm equals $\sum_{k,s} \mathbb{E}_{g\sim\mathcal{G}_{k,s}}[\hat{\ell}_{k,s}(g)]$, where $\mathcal{G}_{k,s}$ denotes the distribution played by $\texttt{alg}_G$ at that time, and the regret can be decomposed into two parts:

$$\sum_k \sum_s \left(\mathop{\mathbb{E}}_{g\sim\mathcal{G}_{k,s}} \left[\hat{\ell}_{k,s}(g)\right] - \hat{\ell}_{k,s}(g^*) \right) \tag{4}$$

$$+ \sum_k \sum_s \left(\hat{\ell}_{k,s}(g^*) - \ell_{k,s}(g^*, h_k^*) \right). \tag{5}$$

Then the theorem follows from the following two lemmas which bound these two parts, respectively. We will prove them in Appendix B.1.1 and B.1.2, respectively.

Lemma 9 *The sum in (4) is* $\mathcal{O}(\sqrt{T(M_{g^*}\log N_G + \log T)} + \varepsilon T)$.

Lemma 10 *The sum in (5) is* $\mathcal{O}(\sqrt{TM_H \log N_H} + \varepsilon T)$.

B.1.1. Proof of Lemma 9

We can see the sum as the regret of $\texttt{alg}_G$ for learning representations with respect to the loss vectors $\hat{\ell}_{k,s}$'s, since we use them to update $\texttt{alg}_G$ based on our Algorithm 2. Then the lemma follows from Theorem 3 once we show that for any step t, G_t forms a good cover with respect to these loss vectors. For this, we claim that any g is covered by $\bar{g} = \pi_t(g) \in G_t$. This is because for any $\tau < t$, $\pi_{\tau+1}(g) = \pi_{\tau+1}(\bar{g})$ (recall the tree-shape branching structure), and

$$\hat{\ell}_\tau(g) - \hat{\ell}_\tau(\bar{g}) = \mathbb{E}\left[\ell_\tau(\pi_{\tau+1}(g), h) - \ell_\tau(\pi_{\tau+1}(\bar{g}), h)\right] = 0,$$

where the expectation is taken over h sampled from $\mathcal{H}_\tau^{\bar{g}_\tau}$. As a result, we can apply Theorem 3 with $\varepsilon = 0$ and obtain a regret bound of $\mathcal{O}(\sqrt{T(M_{g^*}\log N_G + \log T})$ with respect to such loss vectors, which proves the lemma.

B.1.2. Proof of Lemma 10

Fix any task k, and let us bound the inner sum of (5) in our main text, which is

$$\sum_s (\hat{\ell}_{k,s}(g^*) - \ell_{k,s}(g^*, h_k^*)). \tag{6}$$

To ease the notation, let us drop the index k. We also use the notation $\bar{g}_s = \pi_{s+1}(g^*)$, $H_s^* = H_s^{\bar{g}_s}$ and $\mathcal{H}_s^* = \mathcal{H}_s^{\bar{g}_s}$.

Then let us express the sum in (6) as

$$\sum_s (\alpha_s + \beta_s + \gamma_s),$$

where

$$\begin{aligned}
\alpha_s &= \mathop{\mathbb{E}}_{h\sim\mathcal{H}_s^*}[\ell_s(\bar{g}_s,h)] - \bar{\ell}_s^{g^*}(h_k^*),\\
\beta_s &= \bar{\ell}_s^{g^*}(h_k^*) - \ell_s(\bar{g}_s,h_k^*)\\
&= \ell_s(\bar{g}_s,\pi_{s+1}^{\bar{g}_s}(h_k^*)) - \ell_s(\bar{g}_s,h_k^*),\\
\gamma_s &= \ell_s(\bar{g}_s,h_k^*) - \ell_s(g^*,h_k^*)\\
&= \ell_s(\pi_{s+1}(g^*),h_k^*) - \ell_s(g^*,h_k^*).
\end{aligned}$$

Note that $\beta_s \le \varepsilon$ and $\gamma_s \le \varepsilon$ based on Assumptions 2.3 & 2.2 respectively. On the other hand, as $\ell_s(\bar{g}_s,h) = \ell_s(\bar{g}_s,\pi_{s+1}^{\bar{g}_s}(h)) = \bar{\ell}_s^{g^*}(h)$ for any $h \in H_s^*$, we have

$$\sum_s \alpha_s = \sum_s \left(\mathop{\mathbb{E}}_{h\sim\mathcal{H}_s^*}\left[\bar{\ell}_s^{g^*}(h)\right] - \bar{\ell}_s^{g^*}(h_k^*)\right).$$

Then we can see the sum above as the regret of $\mathtt{alg}_H^{g^*}$ in task k with respect to such surrogate loss functions $\bar{\ell}_s^{g^*}$'s. This is because the distribution $\mathcal{H}_s^* = \mathcal{H}_s^{\bar{g}_s}$ is updated according to the loss functions

$$\bar{\ell}_\tau^{\bar{g}_s}(h) = \ell_\tau(\bar{g}_\tau,\bar{h}_\tau) = \bar{\ell}_\tau^{g^*}(h)$$

for $\tau < s$, by noting that $\pi_{\tau+1}(\bar{g}_s) = \pi_{\tau+1}(g^*) = \bar{g}_\tau$ due to the tree-shape branching structure. Of course it is possible that g^* is never split out and our algorithm never actually runs it in the task. In this case, we can still imagine actually running it, and calculate what its regret would be with respect to such surrogate loss functions. To apply Theorem 1, it remains to show that for any s, the active set H_s^* forms a 5ε-cover with respect to such loss functions.

We claim that any h is 5ε-covered by its representative in $H_s^* = H_s^{\bar{g}_s}$, denoted as h_s. To see this, consider any $\tau < s$, and let us use the notation $x \approx_\varepsilon y$ for $|x-y| \le \varepsilon$. Then by Assumptions 2.2 & 2.3, we have

$$\bar{\ell}_\tau^{g^*}(h) = \ell_\tau(\bar{g}_\tau,\bar{h}_\tau) \approx_\varepsilon \ell_\tau(\bar{g}_\tau,h) \approx_\varepsilon \ell_\tau(\bar{g}_s,h),$$

since $\pi_{\tau+1}(\bar{g}_s) = \bar{g}_\tau$ as discussed before. Similarly, we also have

$$\bar{\ell}_\tau^{g^*}(h_s) \approx_\varepsilon \ell_\tau(\bar{g}_\tau,h_s) \approx_\varepsilon \ell_\tau(\bar{g}_s,h_s) \approx_\varepsilon \ell_\tau(\bar{g}_s,h),$$

as $h_s = \pi_s^{\bar{g}_s}(h)$. Consequently, we can conclude that $\bar{\ell}_\tau^{g^*}(h) \approx_{5\varepsilon} \bar{\ell}_\tau^{g^*}(h_s)$ for any $\tau < s$.

Therefore, we can apply Theorem 1 to upper bound $\sum_s \alpha_s$ and hence the sum in (6) by $\mathcal{O}(\sqrt{M_{H,k}T_k \log N_H} + \varepsilon T_k)$, where T_k is the length of task k and $M_{H,k}$ is the number of steps at which g^*'s representatives branch out new predictors. Finally, by summing over k and applying the Cauchy-Schwarz inequality, we obtain an upper bound of

$$\mathcal{O}\left(\sqrt{\sum_k M_{H,k}}\sqrt{\sum_k T_k \log N_H} + \varepsilon T\right),$$

where $\sum_k M_{H,k} = M_H$ and $\sum_k T_k = T$. The lemma then follows.

B.2. Proof of Theorem 6

Before the proof, let us make some remarks. Recall that in the one-task stochastic setting, our Theorem 2 relies heavily on the nice property that the loss distribution of each expert is fixed during the whole time, which is needed for Lemmas 4 & 5. However, in lifelong learning, the loss distribution of a representation can actually change in different tasks as different predictors are allowed. This means that a representation which looks good previously may turn out very bad later, which makes the learning of representations hard. Although we can still apply Algorithm 1 and show that after some time step $\tilde{t}$, each suboptimal representation is unlikely to be chosen, the loss functions $\hat{\ell}_t(\cdot)$'s before step $\tilde{t}$ looks rather like adversarial ones. Therefore, we rely on a different algorithm, our Algorithm 2, for learning representations, which has a better form of the adversarial regret bound, given in Lemma 7. In particular, it can utilize the nice property in the stochastic setting that the optimal g^* appears early and soon makes no more branching.

The proof for the theorem is very similar to that for Theorem 4. Let us consider the following analogous event C_t, for any step t.

- C_t: step t is the first time step such that the best representation g^* branches out before it and none branches out from g^* since then.

Then we have the following analogous lemma which we prove in Appendix B.2.1.

Lemma 11 *There is some step $\tilde{s} \le \mathcal{O}(\frac{\log(N_G/\Delta)}{\Delta})$ such that for any $t \ge \tilde{s}$, $\Pr[C_t] \le \frac{1}{t^3}$.*

Let us start from the case that

$$\varepsilon \le \frac{\Delta}{8} \text{ and } \Delta \ge \sqrt{\frac{\log N_G + K \log N_H}{T}}. \tag{7}$$

Recall from the proof of Theorem 5 that the regret can be decomposed into two parts:

$$\sum_t \left(\mathop{\mathbb{E}}_{g \sim \mathcal{G}_t} \left[\hat{\ell}_t(g) \right] - \hat{\ell}_t(g^*) \right) + \sum_t \left(\hat{\ell}_t(g^*) - \ell^*_t(g^*) \right), \tag{8}$$

with $\ell^*_t(g) = \ell_\tau(g, h^*_k)$ when step t belongs to task k.

Note that the second sum in Eq. (8) corresponds to the learning of predictors for g^*. We bound its expected value using the following lemma, which we prove in Appendix B.2.2.

Lemma 12 *For any steps a and b, with $a \le b$,*

$$\mathbb{E}\left[\sum_{s=a}^{b} \left(\hat{\ell}_s(g^*) - \ell^*_s(g^*) \right) \right] \le \mathcal{O}\left(\frac{\log N_G + K \log N_H}{\Delta} \right).$$

On the other hand, the first sum in Eq. (8) corresponds to the learning of representations. Based on Lemma 11 given above as well as Lemma 7 in Appendix A.3, we have the following, which we prove in Appendix B.2.3.

Lemma 13 *The first sum in Eq. (8) assuming $\Delta \ge 1/T$ is at most*

$$\mathcal{O}\left(\frac{\log(N_G T) + K \log N_H}{\Delta} \right).$$

By combining these two lemmas together, we obtain the regret bound of $\mathcal{O}\left(\frac{\log(N_G T)+K\log N_H}{\Delta}\right)$ when the condition in Eq. (7) holds.

Next, let us consider the case that

$$\Delta < \sqrt{\frac{\log N_G + K \log N_H}{T}}.$$

In this case with a small gap, the situation becomes different in the lifelong learning setting with multiple tasks. More precisely, a representation can have different gaps in different tasks, as different predictors can be used, and the gap of a representation is defined as the smallest among them. This means that even if its smallest gap is small, it can have large gaps in other tasks. Therefore, we can not use the same argument in the proof of Theorems 2 and 4 for the case of small gaps, and in fact our lower bound in Theorem 9 shows that it is impossible to do much better than that in the adversarial setting. Therefore, in this case, we simply apply our adversarial regret bound for learning representations. On the other hand, the argument in the proof of Theorems 2 and 4 still works for learning predictors for g^*, and consequently we obtain a regret bound of

$$\mathcal{O}\left(\sqrt{T(M_{g^*}\log N_G + \log T)} + \sqrt{TK\log N_H} + \varepsilon T\right).$$

Note that when $\varepsilon \le \mathcal{O}(\Delta)$, the term εT in the regret bound above is dominated by other terms. This completes the proof of the theorem.

B.2.1. PROOF OF LEMMA 11

For any step t, consider the following event

- B_t: there is some $g \neq g^*$ such that for any $\tau \le t-2$ and $h \in H$, $|\ell_\tau(g,h) - \ell_\tau(g^*,h)| \le \varepsilon$.

Then clearly we have $\Pr[C_t] \le \Pr[B_t]$. To bound each $\Pr[B_t]$, note that for any $g \neq g^*$ and any step τ in task k,

$$\mathbb{E}\left[\ell_\tau(g,h_k^*) - \ell_\tau(g^*,h_k^*)\right] \ge \Delta,$$

which implies that

$$\begin{aligned}
\Pr\left[\forall h : |\ell_\tau(g,h) - \ell_\tau(g^*,h)| \le \varepsilon\right] &\le \Pr\left[|\ell_\tau(g,h_k^*) - \ell_\tau(g^*,h_k^*)| \le \varepsilon\right] \\
&\le \Pr\left[\ell_\tau(g,h_k^*) - \ell_\tau(g^*,h_k^*) \le \varepsilon\right] \\
&\le 1 - (\Delta - \varepsilon) \\
&\le 1 - \frac{7}{8}\Delta.
\end{aligned}$$

As the loss functions are independent from each other, we have

$$\begin{aligned}
\Pr\left[B_t\right] &\le \sum_{g\neq g^*} \Pr\left[\forall \tau \le t-2, \forall h : |\ell_\tau(g,h) - \ell_\tau(g^*,h)| \le \varepsilon\right] \\
&\le N_G\left(1 - \frac{7}{8}\Delta\right)^{t-2} \\
&\le \frac{1}{t^3}
\end{aligned}$$

when $t \ge \tilde{s}$ for some $\tilde{s} \le \mathcal{O}(\frac{\log(N_G/\Delta)}{\Delta})$, which proves the lemma.

B.2.2. Proof of Lemma 12

As discussed in the proof of Lemma 10 in Appendix B.1.2, the sum

$$\sum_{s=a}^{b} \left(\hat{\ell}_s(g^*) - \ell_s^*(g^*)\right) \tag{9}$$

corresponds to the regret of learning the predictors of g^* with respect to the surrogate loss functions $\bar{\ell}_s^{g^*}(\cdot)$'s. While we could apply the adversarial regret bound derived there, here we aim for a better bound, in terms of pseudo regret, in the stochastic setting. We would like to apply our results in the one-task stochastic setting, but there are some issues which we need to handle. The first issue is that during a task, the distribution of the surrogate loss $\bar{\ell}_s^{g^*}(h)$ may change, because it depends on the representative $\pi_{s+1}(g^*)$ which may change. The second issue is that each surrogate loss $\bar{\ell}_s^{g^*}(h)$ is only an approximation of the true loss $\ell_s(g^*, h)$, and even though each approximation error can be bounded by 2ε, their accumulation may contribute $\Omega(\varepsilon T)$ to the total regret, which we would like to avoid.

Consider any steps $a \le b$, and recall the definition that $\ell_s^*(g) = \ell_s(g, h_k^*)$ when step s belongs to task k. Following the proof of Lemma 10 in Appendix B.1.2, let us express the sum in (9) as

$$\sum_{s=a}^{b} \left(\hat{\ell}_s(g^*) - \ell_s^*(g^*)\right) = \sum_{s=a}^{b} \left(\alpha_s + \beta_s + \gamma_s\right),$$

where for step s belonging to task k,

$$\begin{aligned}
\alpha_s &= \hat{\ell}_s(g^*) - \bar{\ell}_s^{g^*}(h_k^*) \\
&= \mathop{\mathbb{E}}_{h\sim\mathcal{H}_s^*}\left[\bar{\ell}_s^{g^*}(h)\right] - \bar{\ell}_s^{g^*}(h_k^*), \\
\beta_s &= \bar{\ell}_s^{g^*}(h_k^*) - \ell_s(\bar{g}_s, h_k^*) \\
&= \ell_s(\bar{g}_s, \pi_{s+1}^{\bar{g}_s}(h_k^*)) - \ell_s(\bar{g}_s, h_k^*), \\
\gamma_s &= \ell_s(\bar{g}_s, h_k^*) - \ell_s(g^*, h_k^*) \\
&= \ell_s(\pi_{s+1}(g^*), h_k^*) - \ell_s(g^*, h_k^*).
\end{aligned}$$

Note that α_s corresponds to the regret at step s, while β_s and γ_s correspond to the approximation errors. The lemma follows from the following three propositions for bounding their expected sums respectively.

Proposition B.1 $\sum_{s=a}^{b} \mathbb{E}[\alpha_s] \le \mathcal{O}(\frac{K}{\Delta}\log N_H)$.

Proof As discussed previously in the proof of Lemma 10, $\sum_s \alpha_s$ corresponds to the regret of learning predictors for g^* with respect to the loss functions $\bar{\ell}_s^{g^*}(\cdot)$'s. Instead of applying the adversarial regret bound there, we would like to apply our results in the one-task stochastic setting. For this, we need to show that these loss functions have desirable gaps during each task k. Consider any $h \ne h_k^*$ with a gap for ℓ_s (instead of for $\bar{\ell}_s^{g^*}$) defined as

$$\Delta_h = \mathbb{E}\left[\ell_s(g^*, h) - \ell_s(g^*, h_k^*)\right]$$

which is at least $\Delta \geq 8\varepsilon$. Then on one hand, we have

$$\begin{aligned}\mathbb{E}\left[\bar{\ell}_s^{g^*}(h) - \bar{\ell}_s^{g^*}(h_k^*)\right] &\geq \mathbb{E}\left[\ell_s(\pi_{s+1}(g^*), h) - \ell_s(\pi_{s+1}(g^*), h_k^*)\right] - 2\varepsilon \\ &\geq \mathbb{E}\left[\ell_s(g^*, h) - \ell_s(g^*, h_k^*)\right] - 4\varepsilon \\ &= \Delta_h - 4\varepsilon \\ &\geq \frac{1}{2}\Delta_h.\end{aligned}$$

On the other hand, one can similarly show that

$$\mathbb{E}\left[\bar{\ell}_s^{g^*}(h) - \bar{\ell}_s^{g^*}(h_k^*)\right] \leq \Delta_h + 4\varepsilon \leq \frac{3}{2}\Delta_h.$$

Thus, although the expected loss $\mathbb{E}[\bar{\ell}_s^{g^*}(h)]$ may have a changing gap as s varies during each task, it always falls in the small range between $\frac{1}{2}\Delta_h$ and $\frac{3}{2}\Delta_h$. Then it is straightforward to check that for each task, similar bounds as in Lemmas 3 & 4 & 5 still hold by almost identical proofs, and the pseudo regret of each task is at most $\mathcal{O}(\frac{1}{\Delta}\log N_H)$. By summing the bounds over at most K tasks, one can then obtain the bound

$$\sum_s \mathbb{E}[\alpha_s] \leq \mathcal{O}\left(\frac{K}{\Delta}\log N_H\right).$$

■

Proposition B.2 $\sum_{s=a}^{b} \mathbb{E}[\gamma_s] \leq \mathcal{O}(\log(N_G/\Delta))$.

Proof Note that for any step s,

$$\mathbb{E}[\gamma_s] \leq \varepsilon \cdot \Pr[g^* \neq \pi_{s+1}(g^*)] = \varepsilon \cdot \Pr[g^* \notin G_{s+1}].$$

From Lemma 11, we know that $\Pr[g^* \notin G_{s+1}] \leq \frac{1}{s^3}$ when $s \geq \tilde{s}$. As $\tilde{s} \leq \mathcal{O}(\frac{\log(N_G/\Delta)}{\Delta})$, we have

$$\sum_{s=a}^{b} \mathbb{E}[\gamma_s] \leq \sum_{s \leq \tilde{s}} \varepsilon \Pr[g^* \notin G_{s+1}] + \sum_{s > \tilde{s}} \varepsilon \Pr[g^* \notin G_{s+1}] \leq \varepsilon\tilde{s} + \sum_{s > \tilde{s}} \frac{\varepsilon}{s^3} \leq \mathcal{O}(\log(N_G/\Delta)).$$

■

Proposition B.3 $\sum_{s=a}^{b} \mathbb{E}[\beta_s] \leq \mathcal{O}(K\log(N_H/\Delta))$.

Proof Let us partition the time steps according to tasks and consider the sum corresponding to each task separately. For each task k, the corresponding sum can be bounded by $\mathcal{O}(\log(N_H/\Delta))$ using the same analysis as in the proof of Proposition B.2, as the optimal predictor h_k^* is also likely to appear in $s_0 \leq \mathcal{O}(\frac{\log(N_H/\Delta)}{\Delta})$ steps according to Lemma 8. The proposition then follows by summing the bounds over tasks. ■

B.2.3. Proof of Lemma 13

We would like to apply Lemma 7 with starting step $\tilde{s}$ given in Lemma 11 using the loss functions $\hat{\ell}_t$'s, which have the corresponding

$$r_{t,g^*} = \mathop{\mathbb{E}}_{g\sim\mathcal{G}_t}\left[\hat{\ell}_t(g)\right] - \hat{\ell}_t(g^*) \text{ and } V = \sum_{t\geq\tilde{s}} r_{t,g^*}^2.$$

We know from Lemma 11 that for some $\tilde{s} \leq \mathcal{O}(\frac{\log(N_G/\Delta)}{\Delta})$, the event C_t happens with probability at most $1/t^3$. Following the proof of Theorem 4, we decompose the regret in two parts. For that before step $\tilde{s}$, we use the trivial upper bound of $\tilde{s}$. For the remaining steps, we do the following, depending on when the event C_t happens.

As in the proof of Theorem 4, the key observation is that when the event C_t happens, there is no branching from the optimal representation g^* after step t, and we can bound the expected regret afterward by $\mathcal{O}(\sqrt{\mathbb{E}[V]\log(N_GT)})$ according to Lemma 7, while we can simply bound the regret before by t. As the event C_t is only determined by the losses before step t and the losses afterwards are independent, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\geq\tilde{s}} r_{t,g^*}\right] &\leq \sum_{t\geq\tilde{s}} \Pr[C_t]\cdot\left(t+\mathcal{O}\left(\sqrt{\mathbb{E}[V]\log(N_GT)}\right)\right)\\
&\leq \sum_{t\geq\tilde{s}}\frac{t}{t^3} + \sum_{t\geq\tilde{s}}\Pr[C_t]\cdot\mathcal{O}\left(\sqrt{\mathbb{E}[V]\log(N_GT)}\right)\\
&\leq \mathcal{O}(1)+\mathcal{O}\left(\sqrt{\mathbb{E}[V]\log(N_GT)}\right)\\
&\leq \mathcal{O}\left(\sqrt{\mathbb{E}[V]\log(N_GT)}\right),
\end{aligned}$$

as the events C_t's are disjoint from each other.

It remains to bound $\mathbb{E}[V]$. As in the proof of Theorem 4, let S denote the number of steps after $\tilde{s}$ such that suboptimal representations are played, and we know that $\mathbb{E}[V] \leq \mathbb{E}[S]$. Next, we would like to bound $\mathbb{E}[S]$. However, the situation now becomes more difficult as the loss functions $\hat{\ell}_t$'s may not always have a large gap for a suboptimal representation g. To handle this issue, let us decompose the regret as

$$\sum_{t\geq\tilde{s}} r_{t,g^*} = \sum_{t\geq\tilde{s}}\left(\mathop{\mathbb{E}}_{g\sim\mathcal{G}_t}\left[\hat{\ell}_t(g)\right]-\ell_t^*(g^*)\right)+\sum_{t\geq\tilde{s}}\left(\ell_t^*(g^*)-\hat{\ell}_t(g^*)\right),$$

where $\ell_t^*(g^*) = \ell_t(g^*, h_k^*)$ when step t belongs to task k. Then we can again have that the expectation of the first sum above is at least $\mathbb{E}[S]\cdot\Omega(\Delta)$. On the other hand, the expectation of the second sum corresponds to the negation of the expected regret for learning predictors for g^*, which according to Lemma 12 is at least $-\mathcal{O}\left(\frac{\log N_G + K\log N_H}{\Delta}\right)$. As a result, we have

$$\mathbb{E}[S]\cdot\Omega(\Delta) - \mathcal{O}\left(\frac{\log N_G + K\log N_H}{\Delta}\right) \leq \sum_{t\geq\tilde{s}} r_{t,g^*} \leq \mathcal{O}\left(\sqrt{\mathbb{E}[S]\log(N_GT)}\right),$$

which implies that

$$\mathbb{E}[S] \leq \mathcal{O}\left(\frac{\log(N_GT)+K\log N_H}{\Delta^2}\right).$$

Therefore, we can conclude that

$$\mathbb{E}\left[\sum_{t\geq\tilde{s}} r_{t,g^*}\right] \leq \mathcal{O}\left(\sqrt{\mathbb{E}\left[S\right]\log(N_G T)}\right) \leq \mathcal{O}\left(\frac{\log(N_G T)+K\log N_H}{\Delta}\right).$$

Then the lemma follows as $\sum_{t\leq\tilde{s}} r_{t,g^*} \leq \tilde{s} \leq \mathcal{O}(\frac{\log(N_G/\Delta)}{\Delta}) \leq \mathcal{O}(\frac{\log(N_G T)}{\Delta})$, assuming $\Delta > \frac{1}{T}$.

B.3. Proof of Corollary 7

Let us start from the adversarial setting. Recall that our algorithm updates each G_t and H_t^g to satisfy Assumptions 2.2 &2.3, which allows us to apply the regret bound in Theorem 5. It remains to bound the numbers N_G and N_H there.

Our goal is to bound them in terms of their covering numbers, defined as follows.

Definition 14 *The ε-covering number of G is defined as the size of the smallest $G' \subseteq G$ such that any $g \in G$ has some $g' \in G'$ such that for any loss function ℓ and any $h \in H$, $|\ell(g,h)-\ell(g',h)| \leq \varepsilon$. Moreover, for any $g \in G$, the ε-covering number of H^g is defined as the size of the smallest $H' \subseteq H^g$ such that any $h \in H$ has some $h' \in H'$ such for any loss function ℓ, $|\ell(g,h)-\ell(g,h')| \leq \varepsilon$.*

Following (Cohen and Mannor, 2017), we define their empirical versions with respect to a set $\mathcal{L}$ of loss functions as follows.

Definition 15 *The covering number $\mathcal{N}_G(\varepsilon,\mathcal{L})$ is the size of the smallest $\hat{G} \subseteq G$ satisfying the condition that any $g \in G$ is ε-close to some $\hat{g} \in \hat{G}$ in the sense that*

$$\forall \ell \in \mathcal{L}, \forall h \in H, |\ell(g,h)-\ell(\hat{g},h)| \leq \varepsilon.$$

Definition 16 *For any $g \in G$, the covering number $\mathcal{N}_H^g(\varepsilon,\mathcal{L})$ is the size of the smallest $\hat{H} \subseteq H$ satisfying the condition that any $h \in H$ is ε-close to some $\hat{h} \in \hat{H}$ in the sense that*

$$\forall \ell \in \mathcal{L}, |\ell(g,h)-\ell(g,\hat{h})| \leq \varepsilon.$$

We also need the notions of packing numbers defined as follows.

Definition 17 *The packing number $\mathcal{P}_G(\varepsilon,\mathcal{L})$ is the size of the largest $\hat{G} \subseteq G$ which forms an ε-packing in the sense that any distinct $g_1, g_2 \in \hat{G}$ are not ε-close to each other, so that*

$$\exists \ell \in \mathcal{L}, \exists h \in H, \text{ such that } |\ell(g_1,h)-\ell(g_2,h)| > \varepsilon.$$

Definition 18 *For any $g \in G$, the packing number $\mathcal{P}_H^g(\varepsilon,\mathcal{L})$ is the size of the largest $\hat{H} \subseteq H$ which forms an ε-packing for g in the sense that any distinct $h_1, h_2 \in \hat{H}$ are not ε-close to each other, so that*

$$\exists \ell \in \mathcal{L} \text{ such that } |\ell(g,h_1)-\ell(g,h_2)| > \varepsilon.$$

Now to bound N_G, observe that as in (Cohen and Mannor, 2017), the way in which we construct each G_t ensures that G_T forms an ε_G-packing, which implies that $|G_T| \leq \mathcal{P}_G(\varepsilon_G,\mathcal{L}_T)$, with $\mathcal{L}_T$ denoting the set containing all the loss functions in T steps. Furthermore, according to Lemma 2 in (Cohen and Mannor, 2017), we have $\mathcal{P}_G(\varepsilon_G,\mathcal{L}_T) \leq \mathcal{N}_G(\varepsilon_G/2,\mathcal{L}_T)$. Therefore, we can upper bound the size of G_T by the empirical covering number $\mathcal{N}_G(\varepsilon_G/2, \mathcal{L}_T)$, which is clearly upperbounded by the covering number $\mathcal{N}_G(\varepsilon_G/2, \mathcal{L})$, with $\mathcal{L}$ being the set of all possible loss functions.

To bound N_H, let us consider any task k, let $\mathcal{L}_T^k$ denote the set of loss functions appearing during task k, and suppose task k ends at step t. Then we know that $\bar{g}_t = \pi_{t+1}(g^*)$ is ε_G-close to g^* with respect to loss functions in $\mathcal{L}_T^k$. Next, we argue that the set $H_{t+1}^{\bar{g}_t}$ forms an ε'-packing for g^*, with $\varepsilon' = \varepsilon_H - 2\varepsilon_G$.

Consider any $h_a, h_b \in H_{t+1}^{\bar{g}_t}$. Assuming h_b is added after h_a, let $\tau_b \in [t]$ denote the time step when the loss ℓ_{τ_b} forces h_b to be added to the set, with $\bar{g}_{\tau_b} = \pi_{\tau_b+1}(g^*)$. As we add h_b to the set only when it is not ε_H-close to any existing element, including h_b, in the set, we must have

$$|\ell_{\tau_b}(\bar{g}_{\tau_b}, h_a) - \ell_{\tau_b}(\bar{g}_{\tau_b}, h_b)| > \varepsilon_H.$$

Furthermore, since $\bar{g}_{\tau_b} = \pi_{\tau_b+1}(g^*)$, we have

$$\ell_{\tau_b}(\bar{g}_{\tau_b}, h_a) \approx_{\varepsilon_G} \ell_{\tau_b}(g^*, h_a)$$

as well as

$$\ell_{\tau_b}(\bar{g}_{\tau_b}, h_b) \approx_{\varepsilon_G} \ell_{\tau_b}(g^*, h_b).$$

Therefore, we must have

$$|\ell_{\tau_b}(g^*, h_a) - \ell_{\tau_b}(g^*, h_b)| > \varepsilon_H - 2\varepsilon_G = \varepsilon',$$

which implies that $H_{t+1}^{\bar{g}_t}$ forms an ε'-packing for g^* with respect to $\mathcal{L}_T^k$.

As a result, we can conclude that

$$|H_{t+1}^{\bar{g}_t}| \leq \mathcal{P}_H^{g^*}(\varepsilon', \mathcal{L}_T^k) \leq \mathcal{N}_H^{g^*}(\varepsilon'/2, \mathcal{L}_T^k),$$

and therefore we have $N_H \leq \max_k \mathcal{N}_H^{g^*}(\varepsilon'/2, \mathcal{L}_T^k)$, which is clearly bounded by the covering number $\mathcal{N}_H^{g^*}(\varepsilon'/2, \mathcal{L})$. Assuming that $\varepsilon_G \leq \varepsilon_H/4$, we have $\varepsilon'/2 \geq \varepsilon_H/4$, which implies that N_H is upperbounded by the $\varepsilon_H/4$-covering number. This completes the proof of the adversarial setting.

Now let us move on to the stochastic setting, and recall that here we only aim for a better bound on the part of regret corresponding to the learning of predictors for g^*. As we use Algorithm 1 to learn the predictors for each task separately, we can focus on each task k, with T_k steps. We would like to apply our Theorem 2 for the finite case, but some care is needed. The key difference is that, for the event B_t defined there, we can no longer bound its probability by a simple union bound as there are now an infinite number of experts. To deal with this issue, we rely on the assumption that the set H of experts has a finite ε-covering number N, and let S denote such a subset of experts which achieves this ε-covering number.

Given ε, let us choose $\Delta = 32\varepsilon$ here, and consider as before the set

$$A_r = \{i \in H : \Lambda_r \leq \Delta_i \leq 2\Lambda_r\} \text{ where } \Lambda_r = 2^r \Delta,$$

for $r \geq 0$. As the regret contributed by experts not in any such set is at most $\Delta T = \mathcal{O}(\varepsilon T)$, we can focus on experts in these sets. Now fix any $r \geq 0$. Let us define the event B_t:

$$\exists i \in A_r \text{ such that } \bar{L}_{t-1,i} - \bar{L}_{t-1,i^*} \leq \frac{1}{8}\Delta_i(t-1).$$

Moreover, let us define the set

$$S_r = \{i \in S : \Lambda_r - \varepsilon \le \Delta_i \le 2\Lambda_r + \varepsilon\},$$

as well as the event C_t:

$$\exists j \in S_r \text{ such that } \bar{L}_{t-1,j} - \bar{L}_{t-1,i^*} \le \frac{1}{4}\Delta_j(t-1).$$

We claim that the event B_t implies the event C_t. To see this, consider any $i \in A_r$ and let $j \in S$ be its ε-cover, so that $\ell(i) \approx_\varepsilon \ell(j)$ for any loss ℓ, using the notation $x \approx_\varepsilon y$ for $|x-y| \le \varepsilon$. This implies that $\Delta_i \approx_\varepsilon \Delta_j$ and hence $\Delta_i \le \frac{32}{31}\Delta_j$, as well as $j \in S_r$. Then note that

$$\bar{L}_{t-1,i} \approx_{\varepsilon(t-1)} L_{t-1,i} \approx_{\varepsilon(t-1)} L_{t-1,j} \approx_{\varepsilon(t-1)} \bar{L}_{t-1,j},$$

which implies that

$$\bar{L}_{t-1,j} - \bar{L}_{t-1,i^*} \approx_{3\varepsilon(t-1)} \bar{L}_{t-1,i} - \bar{L}_{t-1,i^*}.$$

Therefore, if $\bar{L}_{t-1,i} - \bar{L}_{t-1,i^*} \le \frac{1}{8}\Delta_i(t-1)$, then

$$\bar{L}_{t-1,j} - \bar{L}_{t-1,i^*} \le \left(\frac{1}{8}\Delta_i + 3\varepsilon\right)(t-1)$$

where

$$\frac{1}{8}\Delta_i + 3\varepsilon \le \frac{7}{32}\Delta_i \le \frac{7}{31}\Delta_j \le \frac{1}{4}\Delta_j.$$

This proves our claim, and we have

$$\Pr[B_t] \le \Pr[C_t] \le \sum_{j \in S_r} e^{-\Omega\left(\Delta_j^2 t\right)} \le e^{-\Omega\left(\Lambda_r^2(t-\tilde{r})\right)},$$

for some $\tilde{r} \le \mathcal{O}(\frac{1}{\Delta^2}\log N_H)$, as in the proof of Lemma 5. Then it is straightforward to check that all the remaining proof works and we can achieve the regret bound of $\mathcal{O}(\sqrt{T_k \log N_H} + \varepsilon T_k)$ as in Theorem 2.

Finally, by summing the bound over task k, and combing the adversarial bound for learning representations, we obtain the regret bound for the stochastic setting. This completes the proof of the corollary.

Appendix C. Proofs in Section 6

C.1. Sub-optimality of previous algorithms

First, the algorithm of (Gofer et al., 2013) uses a constant learning rate, and according to Proposition 7 in (Mourtada and Gaïffas, 2019), such an algorithm has a regret lower bound of $\Omega(\sqrt{T\log N})$ in the stochastic setting, even when a large gap Δ exists. Thus, it cannot achieve an upper bound of the form $\mathcal{O}(\frac{1}{\Delta}\log N)$ we are looking for, which motivates us to design a different algorithm.

Next, let us consider the algorithm of (Cohen and Mannor, 2017) which resets and restarts the learning every time a new expert branches out. In the following, we show that its regret depends on the number of branching steps in the stochastic setting, which we would like to avoid.

To make the algorithm of (Cohen and Mannor, 2017) suffer such a regret, our strategy is to make suboptimal experts branch out in appropriate time steps so that the algorithm must restart many times and suffer large enough regret each time. More precisely, we design the loss functions as follows, with any $\Delta \in (0, \frac{1}{4}]$ and $\varepsilon < \frac{\Delta}{N}$.

- We let expert N be the optimal expert with deterministic loss $\ell_{t,N} = 0$ for each step t.
- We let each expert j, for $\frac{N}{2} \leq j < N$, be a suboptimal one, with deterministic loss $\ell_{t,j} = \Delta + 2(N - j - 1)\varepsilon$ for each step t.
- We let each remaining expert i, for $1 \leq i < \frac{N}{2}$, be a suboptimal one, with stochastic loss

$$\ell_{t,i} = \Delta + (1 - \Delta)x_{t,i}$$

for each step t, where each $x_{t,i}$ is an independent Bernoulli random variable with mean p_i, for some p_i to be determined next.

Note that the loss functions has a gap Δ, and we would like to have many branching steps which are at least $s = \frac{1}{\Delta^2}\log N$ steps apart from each other, so that we can have the algorithm suffer a regret of $\Omega(\sqrt{s \log N})$ between two branching steps. Note that for our choice of ε, any expert $j \geq \frac{N}{2}$ must branch out in the beginning. On the other hand, any expert $i < \frac{N}{2}$ can branch out at some step t only if $x_{t,i} = 1$.

Our idea is to choose each p_i appropriately so that each expert i is likely to succeed in branching out in the time interval $I_i = [t_i, r_i]$, with desirable t_i and r_i satisfying $t_{i+1} = r_i + s$ (with $r_0 = 0$ for convenience). For this, we would like each of the following three bad events to happen with probability at most $\delta = \frac{1}{2M}$, for some M to be determined later, with $\hat{t} = \min\{\tau : x_{\tau,i} = 1\}$, which is the first time that $x_{\tau,i} = 1$.

- B_1: $\hat{t} < t_i$.
- B_2: $\hat{t} > r_i$.
- B_3: $x_{\hat{t},j} = 1$ some $j \neq i$.

It is easy to see that expert i succeeds (branching out in the interval I_i) if none of the events happens. Note that we can have $\Pr[B_1] \leq t_i p_i \leq \delta$ with $p_i = \frac{\delta}{t_i}$. We can also have $\Pr[B_2] \leq (1 - p_i)^{r_i} \leq \delta$ with $r_i = \frac{1}{p_i}\log\frac{1}{\delta} = \frac{t_i}{\delta}\log\frac{1}{\delta}$, which with the notation $\alpha = \frac{1}{\delta}\log\frac{1}{\delta}$ implies that

$$r_i = \alpha t_i = \alpha(r_{i-1} + s) = \sum_{j=1}^{i} \alpha^j s \in [\alpha^i s, \alpha^{i+1} s].$$

Moreover, we have $\Pr[B_3] \leq \sum_{j<\frac{N}{2}} p_j$ because after fixing $\hat{t}$ as well as the randomness of expert i, the distribution of $x_{\hat{t},j}$, for any $j \neq i$, is still independent from each other. As $p_j = \frac{\delta}{t_j} = \frac{\delta}{r_{j-1}+s}$, we have

$$\Pr[B_3] \leq \sum_{j<\frac{N}{2}} \frac{\delta}{r_{j-1} + s} \leq \delta.$$

As a result, the probability that some expert fails is at most $M\delta = \frac{1}{2}$. Moreover, as $r_i \leq s\alpha^{i+1}$, we can have $r_i \leq T$ for $i \geq M$, with some $M = \Omega(\frac{\log T}{\log\log T})$.

Next, let us consider the case that all the experts in $[M]$ succeed, so that there are M branching steps, which fall in those M intervals $I_1, \ldots, I_M$, with $I_i = [t_i, r_i]$. As the algorithm restarts at each branching step, it suffices to show a large regret lower bound between two such branching steps, denoted as b_i and b_{i+1}, of experts i and $i+1$, which are at least s steps apart, as $b_{i+1} - b_i \geq t_{i+1} - r_i = s$. Recall that after τ steps of update since a restart, the algorithm with a time-varying learning rate $\eta_\tau = \sqrt{(c \log N_\tau)/\tau}$ plays suboptimal experts with probability at least

$$\left(\frac{N}{2} - 1\right) e^{-\eta_\tau 2\Delta\tau} \geq \left(\frac{N}{2} - 1\right) e^{-\sqrt{4c\Delta^2\tau \log N}} \geq \frac{1}{2}$$

when $\tau \leq \tilde{t}$, for some $\tilde{t} \geq \Omega(\frac{1}{\Delta^2} \log N)$. This implies that between steps b_i and b_{i+1}, the pseudo regret of the algorithm is at least

$$\Delta \cdot \frac{1}{2} \cdot \Omega\left(\frac{1}{\Delta^2} \log N\right) \geq \Omega\left(\frac{1}{\Delta} \log N\right),$$

and the total pseudo regret is at least

$$\Omega\left(\frac{M}{\Delta} \log N\right),$$

which depends on the number of branching steps M.

C.2. Proof of Theorem 8

Our proof is based on the previous approaches for proving lower bounds in the branching settings (Gofer et al., 2013) and the lifelong learning setting (Wu et al., 2019). Both rely on the well-known lower bound of $\Omega(\sqrt{T \log N})$ for the traditional case of experts problem, with N experts in one task of length T (see e.g. Section 3.7 in (Cesa-Bianchi and Lugosi, 2006)).

Following (Wu et al., 2019), we prove our lower bound by considering two special cases. First, let us consider the case when the predictor set for each representation has only one element, and the problem reduces to that of learning representations. Consider the scenario in which the T steps are evenly divided into M_G intervals, where the representations do not branch within each interval. Given any algorithm, our strategy is to make it suffer a large regret by adding N_G/M_G new representations at the start of each interval. More precisely, at the start of the first interval, there are N_G/M_G representations, and we can have the algorithm suffer a regret of at least $\Omega(\sqrt{(T/M_G)\log(N_G/M_G)})$ based on the lower bound for the experts problem. In each later interval $i \geq 2$, we take the best representation g_{i-1} in the previous interval and have it split into N_G/M_G representations, with S_i denoting this set of N_G/M_G representations. We let representations not in S_i remain bad by giving them large losses in this interval, while the new best representation is now hiding in this set S_i. We use representation in S_i to confuse the algorithm by letting them share the same losses as that of g_{i-1} up to interval $i-1$, but in interval i, the algorithm face a new experts problem with this set S_i of experts, and again we can establish the same regret lower bound as in the first interval. Therefore, the total regret the algorithm suffers is M_G times that in each interval, which is

$$\Omega(\sqrt{TM_G \log(N_G/M_G)}).$$

Next, let us consider the case when there is only one representation, and the problem reduces to that of learning predictors. In this case, the tasks become unrelated to each other and we can bound the regret of each task separately, since the offline algorithm is allowed to use a different predictor for a different task. Thus, we divide the time steps and the branching steps evenly for these K tasks, each having $T_k = T/K$ steps and $M_{H,k} = M_H/K$ branching steps. Using the argument as in the first case, we can establish for each task k a regret lower bound of $\Omega(\sqrt{T_k M_{H,k} \log(N_H/M_{H,k})})$. Multiplying this lower bound by K, we obtain a total regret lower bound of

$$\Omega(\sqrt{T M_H \log(N_H K/M_H)}).$$

Finally, since the problem has these two special cases, the regret lower bound is at least the maximum of these two lower bound, which is at least the average of the two. This proves the theorem.

C.3. Proof of Theorem 9

Our approach here follows closely that for Theorem 8, but now we instead rely on the stochastic lower bound for the standard case of experts problem, also of the form $\Omega(\sqrt{T \log N})$, which can be found in Proposition 4 of (Mourtada and Gaïffas, 2019).

Again, we prove the bounds by considering two special cases. In the first case, the predictor set for each representation has only one element, and the problem reduces to that of learning representations. Although the loss distribution of a representation must remain fixed during a task, it can change when a new task starts as that of its predictor can change. Therefore, we now make the branching happen only at the start of a new task, unlike in the adversarial case. That is, we will use the first $K' = \min\{M_G, K\}$ tasks to force a large regret by adding N_G/K' new representations each time, with each of these K' tasks lasting for $\Omega(T/K')$ steps. Following our approach for Theorem 8, but now using instead the stochastic lower bound of (Mourtada and Gaïffas, 2019), we can establish a regret lower bound of

$$\Omega\left(\sqrt{TK' \log(N_G/K')}\right).$$

The second case is when there is only one representation. In this case, the tasks become unrelated to each other, so we can establish a lower bound for each task separately and add these bounds together, just as in our proof of Theorem 8. However, as allowing branching in a single task does not make the problem harder in the stochastic setting, as shown by our Theorem 2, we simply apply (Mourtada and Gaïffas, 2019) to obtain a lower bound of $\Omega(\sqrt{(T/K) \log N_H})$ for each task, with each lasting for T/K steps. By multiplying this by K, we obtain a total regret lower bound of

$$\Omega\left(\sqrt{TK \log N_H}\right).$$

Finally, by combining these two lower bounds for these two special cases, we obtain the claimed lower bound of the theorem.

References

Sébastien Bubeck. Introduction to online optimization. *Lecture Notes*, 2011.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Alon Cohen and Shie Mannor. Online learning with many experts. *arXiv preprint arXiv:1702.07870*, 2017.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 176–196, 2014.

Eyal Gofer, Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for branching experts. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 618–638, 2013.

Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research (JMLR)*, 20(83):1–28, 2019.

Yi-Shan Wu, Po-An Wang, and Chi-Jen Lu. Lifelong optimization with low regret. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 448–456, 2019.