

826 A Proofs for Section 3

827 In this section, we analyze the regret of Algorithm 3 in the policy class setting and prove Theorem 3.1.

Algorithm 3 EXP4 with Delay-Adapted Loss Estimators (EXP4-DALE)

1: **inputs:**

- Finite policy class $\Pi \subseteq \mathcal{X} \rightarrow \mathcal{A}$ with $|\Pi| = N$,
- Upper bound on the sum of delays, D .
- Step size $\eta > 0$.

2: Initialize $p_1 \in \Delta_N$ as the uniform distribution over Π .

3: **for** round $t = 1, \dots, T$ **do**

4: Receive context $x_t \in \mathcal{X}$.

5: Sample $\pi \sim p_t$ and play $a_t = \pi(x_t)$.

6: Observe feedback $(s, L(x_s, a_s))$ for all $s \leq t$ with $s + d_s = t$ and construct loss estimators

$$\hat{c}_{s,i} = \frac{L(x_s, a_s) \mathbb{I}[\pi_i(x_s) = a_s]}{\max\{Q_{s,a_s}, \tilde{Q}_{s,a_s}^t\}} \quad \forall i \in [N], \quad (2)$$

where we define $Q_{s,a} = \sum_{i=1}^N p_{s,i} \mathbb{I}[\pi_i(x_s) = a]$ and $\tilde{Q}_{s,a}^t = \sum_{i=1}^N p_{t,i} \mathbb{I}[\pi_i(x_s) = a]$.

7: Update

$$p_{t+1,i} \propto p_{t,i} \exp\left(-\eta \sum_{s:s+d_s=t} \hat{c}_{s,i}\right). \quad (3)$$

828 Throughout this section, we use the notation $\mathbb{E}_t[\cdot]$ to denote an expectation conditioned on the entire
829 history up to round t . We define the standard (unbiased) importance-weighted loss estimators by

$$\tilde{c}_{t,i} = \frac{L(x_t, a_t) \mathbb{I}[\pi_i(x_t) = a_t]}{Q_{t,a_t}} \quad \forall i \in [N], \quad (4)$$

830 **Theorem A.1.** Algorithm 3 attains the following expected regret bound:

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{\log N}{\eta} + \frac{\eta}{2} \mathbb{E} \left[\sum_{t=1}^T \sum_{i=1}^N p_{t+d_t,i} \tilde{c}_{t,i}^2 \right] + 2 \mathbb{E} \left[\sum_{t=1}^T \|p_{t+d_t} - p_t\|_1 \right].$$

831 *Proof.* The regret may be decomposed as follows:

$$\begin{aligned} \mathcal{R}_T &= \sum_{t=1}^T c_t \cdot (p_t - p^*) \\ &= \underbrace{\sum_{t=1}^T p_t \cdot (c_t - \hat{c}_t)}_{Bias_1} + \underbrace{\sum_{t=1}^T p^* \cdot (\hat{c}_t - c_t)}_{Bias_2} + \underbrace{\sum_{t=1}^T (p_t - p_{t+d_t}) \cdot \hat{c}_t}_{Drift} + \underbrace{\sum_{t=1}^T (p_{t+d_t} - p^*) \cdot \hat{c}_t}_{OMD}, \end{aligned} \quad (5)$$

832 where $c_{t,i} = L(x_t, \pi_i(x_t))$ for $i \in [N]$. The *OMD* term can be bounded by referring to Lemma 9 of
833 [35] which asserts that

$$\sum_{t=1}^T (p_{t+d_t} - p^*) \cdot \hat{c}_t \leq \frac{\log N}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \sum_{i=1}^{|\Pi|} p_{t+d_t,i} \tilde{c}_{t,i}^2. \quad (6)$$

834 while noting that this lemma does not require a specific form of loss estimators, only that they
835 are nonnegative, as is the case for our delay-adapted estimators defined in Equation (2). We also
836 note that the *Bias*₂ term is non-positive in expectation, since the delay-adapted estimators satisfy
837 $\mathbb{E}_t[\hat{c}_{t,i}] \leq c_{t,i}$ for $i \in [N]$. Thus, to conclude the proof we are left with bounding the *Drift* and
838 *Bias*₁ terms, whose bounds are given in Lemma A.2 and Lemma A.3 that follow. \square

839 *Proof of Theorem 3.1* First, we show that

$$\mathbb{E} \left[\sum_t \sum_i p_{t+d_t, i} \hat{c}_{t, i}^2 \right] \leq KT.$$

840 Indeed, using the definition of the delay-adapted loss estimators \hat{c}_t , it holds that

$$\begin{aligned} \mathbb{E} \left[\sum_t \sum_i p_{t+d_t, i} \hat{c}_{t, i}^2 \right] &= \mathbb{E} \left[\sum_t \sum_i p_{t+d_t, i} \left(\frac{L(x_t, a_t) \mathbb{I}[\pi_i(x_t) = a_t]}{\max\{Q_{t, a_t}, \tilde{Q}_{t, a_t}^{t+d_t}\}} \right)^2 \right] \\ &\leq \mathbb{E} \left[\sum_t \frac{1}{\tilde{Q}_{t, a_t}^{t+d_t}} \sum_i \frac{p_{t+d_t, i} \mathbb{I}[\pi_i(x_t) = a_t]}{Q_{t, a_t}} \right] \\ &= \mathbb{E} \left[\sum_t \frac{1}{Q_{t, a_t}} \right] = \mathbb{E} \left[\sum_t \sum_a \frac{Q_{t, a}}{Q_{t, a}} \right] = KT. \end{aligned}$$

841 Thus, using Theorem A.1 together with Lemma A.4 gives the bound claimed in Theorem 3.1 \square

842 **Lemma A.2** (Bounding the Drift term). *The Drift term given in Equation (5) is bounded in*
843 *expectation as follows:*

$$\mathbb{E} \left[\sum_{t=1}^T (p_t - p_{t+d_t}) \cdot \hat{c}_t \right] \leq \mathbb{E} \left[\sum_{t=1}^T \|p_t - p_{t+d_t}\|_1 \right].$$

844 *Proof.* First, we note that the delay-adapted loss estimators \hat{c}_t are upper-bounded by the standard,
845 conditionally unbiased importance-weighted estimators \tilde{c}_t defined in Equation (4). Therefore, we can
846 bound the *Drift* term as follows:

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^T (p_t - p_{t+d_t}) \cdot \hat{c}_t \right] &\leq \mathbb{E} \left[\sum_{t=1}^T \sum_{i=1}^N |p_{t, i} - p_{t+d_t, i}| \hat{c}_{t, i} \right] \\ &\leq \mathbb{E} \left[\sum_{t=1}^T \sum_{i=1}^N |p_{t, i} - p_{t+d_t, i}| \tilde{c}_{t, i} \right] \\ &= \mathbb{E} \left[\sum_{t=1}^T \sum_{i=1}^N |p_{t, i} - p_{t+d_t, i}| \cdot c_{t, i} \right] \\ &\leq \mathbb{E} \left[\sum_{t=1}^T \|p_t - p_{t+d_t}\|_1 \right], \end{aligned}$$

847 where the last step follows from Hölder's inequality and the fact that $\|c_t\|_\infty \leq 1$. \square

848 **Lemma A.3.** *The Bias₁ term given in Equation (5) is bounded in expectation as follows:*

$$\mathbb{E} \left[\sum_t p_t \cdot (c_t - \hat{c}_t) \right] \leq \mathbb{E} \left[\sum_{t=1}^T \|p_t - p_{t+d_t}\|_1 \right].$$

849 *Proof.* We note losses and loss estimators can be indexed by actions rather than policies and use the
850 the notation $c_{t, a} = L(x_t, a)$ and $\hat{c}_{t, a} = \frac{c_{t, a} \mathbb{I}[a_t = a]}{M_{t, a}}$ where $M_{t, a} = \max\{Q_{t, a}, \tilde{Q}_{t, a}^{t+d_t}\}$. Therefore,
851 using the fact that $\mathbb{E}_t[\hat{c}_{t, a}] = c_{t, a} \frac{Q_{t, a}}{M_{t, a}}$, the Bias₁ term can be bounded as follows:

$$\begin{aligned}
\mathbb{E} \left[\sum_{t=1}^T p_t \cdot (c_t - \hat{c}_t) \right] &= \mathbb{E} \left[\sum_{t=1}^T \sum_{a=1}^K Q_{t,a} (L(x_t, a) - \hat{c}_{t,a}) \right] \\
&= \mathbb{E} \left[\sum_{t=1}^T \sum_{a=1}^K Q_{t,a} L(x_t, a) \left(1 - \frac{Q_{t,a}}{M_{t,a}} \right) \right] \\
&\leq \mathbb{E} \left[\sum_{t=1}^T \sum_{a=1}^K \frac{Q_{t,a}}{M_{t,a}} (M_{t,a} - Q_{t,a}) \right] \\
&\leq \mathbb{E} \left[\sum_{t=1}^T \sum_{k=1}^K \left(\max \{ Q_{t,a}, \tilde{Q}_{t,a}^{t+d_t} \} - Q_{t,a} \right) \right] \\
&\leq \mathbb{E} \left[\sum_{t=1}^T \sum_{a=1}^K \left| \tilde{Q}_{t,a}^{t+d_t} - Q_{t,a} \right| \right].
\end{aligned}$$

852 Now, by the definition of $Q_{t,a}$, $\tilde{Q}_{t,a}^{t+d_t}$ and the triangle inequality, we have

$$\mathbb{E} \left[\sum_{t=1}^T \sum_{a=1}^K \left| \tilde{Q}_{t,a}^{t+d_t} - Q_{t,a} \right| \right] \leq \mathbb{E} \left[\sum_{t=1}^T \sum_{a=1}^K \sum_{i: \pi_i(x_t)=a} |p_{t+d_t, i} - p_{t, i}| \right] = \mathbb{E} \left[\sum_{t=1}^T \|p_t - p_{t+d_t}\|_1 \right],$$

853 concluding the proof. \square

854 **Lemma A.4** (Distribution drift). *The following holds for the iterates $\{p_t\}_{t=1}^T$ of Algorithm 3:*

$$\mathbb{E} \left[\sum_{t=1}^T \|p_{t+d_t} - p_t\|_1 \right] \leq \eta(D + T).$$

855 *Proof.* Define

$$F_t(p) = p \cdot \sum_{s: s+d_s < t} \hat{c}_s + \frac{1}{\eta} R(p),$$

856 where $R(p) = \sum_{i=1}^N p_i \log p_i$, so that $p_t = \arg \min_{p \in \Delta_\Pi} F_t(p)$. Note that $R(\cdot)$ is 1-strongly convex
857 with respect to $\|\cdot\|_1$, and therefore $F_t(\cdot)$ are $1/\eta$ -strongly convex. Thus, using first-order optimality
858 conditions for p_t and p_{t+1} , we have:

$$\begin{aligned}
F_t(p_{t+1}) &\geq F_t(p_t) + \nabla F_t(p_t) \cdot (p_{t+1} - p_t) + \frac{1}{2\eta} \|p_{t+1} - p_t\|_1^2 \geq F_t(p_t) + \frac{1}{2\eta} \|p_{t+1} - p_t\|_1^2, \\
F_{t+1}(p_t) &\geq F_{t+1}(p_{t+1}) + \nabla F_{t+1}(p_{t+1}) \cdot (p_t - p_{t+1}) + \frac{1}{2\eta} \|p_{t+1} - p_t\|_1^2 \geq F_{t+1}(p_{t+1}) + \frac{1}{2\eta} \|p_{t+1} - p_t\|_1^2.
\end{aligned}$$

859 Summing the two inequalities, we obtain

$$\begin{aligned}
\frac{1}{\eta} \|p_{t+1} - p_t\|_1^2 &\leq F_{t+1}(p_t) - F_t(p_t) + F_t(p_{t+1}) - F_{t+1}(p_{t+1}) \\
&= \left(\sum_{s: s+d_s=t} \hat{c}_s \right) \cdot (p_t - p_{t+1}) \\
&\leq \sum_i \left(\sum_{s: s+d_s=t} \hat{c}_{s,i} \right) |p_{t,i} - p_{t+1,i}| \\
&\leq \sum_i \left(\sum_{s: s+d_s=t} \tilde{c}_{s,i} \right) |p_{t,i} - p_{t+1,i}|,
\end{aligned}$$

860 where $\tilde{c}_{s,i}$ are the standard (unbiased) importance-weighted loss estimators. Taking expectations
 861 while using $\mathbb{E}[(\cdot)^2] \geq (\mathbb{E}[\cdot])^2$ and Hölder's inequality, we obtain

$$\begin{aligned} \frac{1}{\eta}(\mathbb{E}\|p_{t+1} - p_t\|_1)^2 &\leq \frac{1}{\eta}\mathbb{E}\left[\|p_{t+1} - p_t\|_1^2\right] \\ &\leq \mathbb{E}\left[\sum_i \left(\sum_{s:s+d_s=t} c_{s,i}\right) \cdot |p_{t+1,i} - p_{t,i}|\right] \\ &\leq m_t \mathbb{E}\|p_{t+1} - p_t\|_1, \end{aligned}$$

862 where $m_t = |\{s : s + d_s = t\}|$ is the number of observations that arrive on round t . Dividing through
 863 by the right-hand side of the inequality above, we obtain

$$\mathbb{E}\|p_{t+1} - p_t\|_1 \leq \eta m_t,$$

864 and using the triangle inequality we have

$$\mathbb{E}\|p_{t+d_t} - p_t\|_1 \leq \sum_{s=1}^{d_t} \mathbb{E}\|p_{t+s} - p_{t+s-1}\|_1 \leq \eta \sum_{s=1}^{d_t} m_{t+s-1} = \eta M_{t,d_t},$$

865 where M_{t,d_t} is the number of observations that arrive between rounds t and $t + d_t - 1$. Using Lemma
 866 C.7 in [20], we conclude the proof via

$$\mathbb{E}\left[\sum_{t=1}^T \|p_{t+d_t} - p_t\|_1\right] \leq \eta \sum_{t=1}^T M_{t,d_t} \leq \eta(D + T).$$

867

□

868 B Proofs for Section 4.2

869 B.1 Proof of Theorem 4.6

870 In this subsection, we provide the proofs of the lemmas required to derive regret guarantees for
871 algorithm DA-FA (Algorithm 1), proving Theorem 4.6

872 Consider the following regret decomposition,

$$\begin{aligned} \mathcal{R}_T &= \sum_{t=1}^{d_{\max}} (p_t - p_*(\cdot | x_t)) \cdot \ell(x_t, \cdot) + \sum_{t=d_{\max}+1}^T (p_t - p_*(\cdot | x_t)) \cdot \hat{f}_{\tau^t}(x_t, \cdot) \\ &\quad + \sum_{t=d_{\max}+1}^T p_t \cdot (\ell(x_t, \cdot) - \hat{f}_t(x_t, \cdot)) + \sum_{t=d_{\max}+1}^T p_*(\cdot | x_t) \cdot (\hat{f}_t(x_t, \cdot) - \ell(x_t, \cdot)) \\ &\quad + \sum_{t=d_{\max}+1}^T (p_t - p_*(\cdot | x_t)) \cdot (\hat{f}_t(x_t, \cdot) - \hat{f}_{\tau^t}(x_t, \cdot)). \end{aligned}$$

873 We bound each term individually in the following lemmas and claims, and then we combine all the
874 bounds to derive Theorem 4.6

875 **Claim B.1.** *With probability 1, it holds that*

$$\sum_{t=1}^{d_{\max}} (p_t - p_*(\cdot | x_t)) \cdot \ell(x_t, \cdot) \leq d_{\max}.$$

876 *Proof.* This follows immediately by the fact that $\ell(\cdot)$ is bounded in $[0, 1]$. □

877 **Lemma B.2** (Restatement of Lemma 4.7). *With probability 1, it holds that*

$$\sum_{t=d_{\max}+1}^T (p_t(\cdot) - p_*(\cdot | x_t)) \cdot \hat{f}_{\tau^t}(x_t, \cdot) \leq \frac{KT}{\gamma} - \sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} \frac{p_*(a | x_t)}{\gamma p_t(a)}.$$

878 *Proof.* For $t \in \{d_{\max} + 1, d_{\max} + 2, \dots, T\}$, let $R_t(p)$ denote the objective of the convex minimiza-
879 tion problem in Equation (1), i.e.,

$$R_t(p) = \sum_{a \in \mathcal{A}} p(a) \cdot \hat{f}_{\tau^t}(x_t, a) - \frac{1}{\gamma} \sum_{a \in \mathcal{A}} \log(p(a)).$$

880 Hence,

$$(\nabla R_t(p))_a = \hat{f}_{\tau^t}(x_t, a) - \frac{1}{\gamma p(a)}.$$

881 Since $p_*(\cdot | x_t)$ is a feasible solution and p_t is the optimal solution, by first-order optimality conditions
882 we have

$$\sum_{a \in \mathcal{A}} p_*(a | x_t) \left(\hat{f}_{\tau^t}(x_t, a) - \frac{1}{\gamma p_t(a)} \right) - \sum_{a \in \mathcal{A}} p_t(a) \left(\hat{f}_{\tau^t}(x_t, a) - \frac{1}{\gamma p_t(a)} \right) \geq 0,$$

883 Thus,

$$\sum_{a \in \mathcal{A}} (p_*(a | x_t) - p_t(a)) \hat{f}_{\tau^t}(x_t, a) \geq \sum_{a \in \mathcal{A}} \frac{p_*(a | x_t)}{\gamma p_t(a)} - \frac{K}{\gamma}.$$

884 Which implies that

$$\sum_{a \in \mathcal{A}} (p_t(a) - p_*(a | x_t)) \hat{f}_{\tau^t}(x_t, a) \leq \frac{K}{\gamma} - \sum_{a \in \mathcal{A}} \frac{p_*(a | x_t)}{\gamma p_t(a)}.$$

885 We conclude that

$$\sum_{t=d_{\max}+1}^T (p_t - p_*(\cdot | x_t)) \cdot \hat{f}_{\tau^t}(x_t, \cdot) \leq \frac{KT}{\gamma} - \sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} \frac{p_*(a | x_t)}{\gamma p_t(a)}.$$

886 □

887 **Lemma B.3** (Restatement of Lemma 4.8). *It holds true that*

$$\mathbb{E} \left[\sum_{t=d_{\max}+1}^T p_t \cdot \left(\ell(x_t, \cdot) - \hat{f}_t(x_t, \cdot) \right) \right] \leq \frac{KT}{\gamma} + \gamma \mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F}, \eta}).$$

888 *Proof.* For this term, we apply the oracle expected regret bound for the non-delayed function
889 approximation. By Assumption 4.2 the following holds.

$$\begin{aligned} & \mathbb{E} \left[\sum_{t=d_{\max}+1}^T p_t \cdot \left(\ell(x_t, \cdot) - \hat{f}_t(x_t, \cdot) \right) \right] \\ & \leq \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} p_t(a) \left(\ell(x_t, a) - \hat{f}_t(x_t, a) \right) \right] \\ & = \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} \sqrt{\frac{\gamma}{\gamma}} p_t(a) \left(\ell(x_t, a) - \hat{f}_t(x_t, a) \right) \right] \\ & \leq \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} \frac{p_t(a)}{\gamma} \right] + \gamma \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} p_t(a) \left(\ell(x_t, a) - \hat{f}_t(x_t, a) \right)^2 \right] \quad (\text{AM-GM}) \\ & = \frac{(T - d_{\max})K}{\gamma} + \gamma \sum_{t=d_{\max}+1}^T \mathbb{E} \left[\left(\hat{f}_t(x_t, a_t) - \ell(x_t, a_t) \right)^2 \right] \\ & \leq \frac{(T - d_{\max})K}{\gamma} + \gamma \sum_{t=1}^T \mathbb{E} \left[\left(\hat{f}_t(x_t, a_t) - \ell(x_t, a_t) \right)^2 \right] \\ & \leq \frac{KT}{\gamma} + \gamma \mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F}, \eta}), \end{aligned}$$

890 where in the final transition we used Assumption 4.5 which implies that the observations are given to
891 the oracle in the same order that they arrive to the CMAB algorithm, which allows us to invoke the
892 regret guarantee of the non-delayed oracle. \square

893 **Lemma B.4** (Restatement of Lemma 4.9). *It holds true that*

$$\mathbb{E} \left[\sum_{t=d_{\max}+1}^T p_{\star}(\cdot | x_t) \cdot \left(\hat{f}_t(x_t, \cdot) - \ell(x_t, \cdot) \right) \right] \leq \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} \frac{p_{\star}(a | x_t)}{\gamma p_t(a)} \right] + \gamma \mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F}, \eta}).$$

894 *Proof.* For this term, we would like to use a change-of-measure technique using AM-GM to be able
895 to apply the oracle's expected regret bound for the non-delayed function approximation. Again,
896 by Assumption 4.2 the following holds.

$$\begin{aligned} & \mathbb{E} \left[\sum_{t=d_{\max}+1}^T p_{\star}(\cdot | x_t) \cdot \left(\hat{f}_t(x_t, \cdot) - \ell(x_t, \cdot) \right) \right] \\ & \leq \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} p_{\star}(a | x_t) \cdot \left(\hat{f}_t(x_t, a) - \ell(x_t, a) \right) \right] \\ & = \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} p_{\star}(a | x_t) \sqrt{\frac{\gamma p_t(a)}{\gamma p_t(a)}} \cdot \left(\hat{f}_t(x_t, a) - \ell(x_t, a) \right) \right] \\ & \leq \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} \frac{p_{\star}^2(a | x_t)}{\gamma p_t(a)} \right] + \gamma \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} p_t(a) \left(\hat{f}_t(x_t, a) - \ell(x_t, a) \right)^2 \right] \quad (\text{AM-GM}) \\ & \leq \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} \frac{p_{\star}(a | x_t)}{\gamma p_t(a)} \right] + \gamma \sum_{t=d_{\max}+1}^T \mathbb{E} \left[\left(\hat{f}_t(x_t, a_t) - \ell(x_t, a_t) \right)^2 \right] \end{aligned}$$

$$\begin{aligned}
&\leq \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} \frac{p_{\star}(a|x_t)}{\gamma p_t(a)} \right] + \gamma \sum_{t=1}^T \mathbb{E} \left[\left(\hat{f}_t(x_t, a_t) - \ell(x_t, a_t) \right)^2 \right] \\
&\leq \mathbb{E} \left[\sum_{t=d_{\max}+1}^T \sum_{a \in \mathcal{A}} \frac{p_{\star}(a|x_t)}{\gamma p_t(a)} \right] + \gamma \mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F}, \eta}),
\end{aligned}$$

897 where in the final transition we used Assumption 4.5 as in Lemma 4.8 □

898 We now proceed to prove our final lemma.

899 **Lemma B.5** (Restatement of Lemma 4.10). *Under Assumption 4.4 it holds true that*

$$\mathbb{E} \left[\sum_{t=d_{\max}+1}^T (p_t - p_{\star}(\cdot | x_t)) \cdot \left(\hat{f}_t(x_t, \cdot) - \hat{f}_{\tau^t}(x_t, \cdot) \right) \right] \leq 2\sqrt{d_{\max} D \beta}.$$

900 *Proof.* Using Hölder's inequality and the triangle inequality, we have

$$\begin{aligned}
&\sum_{t=d_{\max}+1}^T (p_t - p_{\star}(\cdot | x_t)) \cdot \left(\hat{f}_t(x_t, \cdot) - \hat{f}_{\tau^t}(x_t, \cdot) \right) \\
&\leq \sum_{t=d_{\max}+1}^T \|p_t - p_{\star}(\cdot | x_t)\|_1 \cdot \|\hat{f}_t(x_t, \cdot) - \hat{f}_{\tau^t}(x_t, \cdot)\|_{\infty} \\
&\leq 2 \sum_{t=d_{\max}+1}^T \sum_{i=1}^{d_{\tau^t}} \|\hat{f}_{t-i}(x_t, \cdot) - \hat{f}_{t-(i-1)}(x_t, \cdot)\|_{\infty} \quad (\tau^t = t - d_{\tau^t}) \\
&\leq 2 \sum_{t=d_{\max}+1}^T \sum_{i=1}^{d_{\tau^t}} \|\hat{f}_{t-i} - \hat{f}_{t-(i-1)}\|_{\infty} \\
&\leq 2 \sum_{t=d_{\max}+1}^T \sigma_t \|\hat{f}_t - \hat{f}_{t+1}\|_{\infty},
\end{aligned}$$

901 where σ_t is the number of pending observations (that is, which have not yet arrived) as of round t .
902 Now, using the Cauchy-Schwarz inequality, the fact that $\mathbb{E}[\sqrt{\cdot}] \leq \sqrt{\mathbb{E}[\cdot]}$ and Assumption 4.4, we
903 finally obtain

$$\begin{aligned}
\mathbb{E} \left[\sum_{t=d_{\max}+1}^T (p_t - p_{\star}(\cdot | x_t)) \cdot \left(\hat{f}_t(x_t, \cdot) - \hat{f}_{\tau^t}(x_t, \cdot) \right) \right] &\leq 2 \sqrt{\left(\sum_{t=d_{\max}+1}^T \sigma_t^2 \right) \cdot \left(\sum_{t=d_{\max}+1}^T \mathbb{E}[\|\hat{f}_t - \hat{f}_{t+1}\|_{\infty}^2] \right)} \\
&\leq 2\sqrt{d_{\max} D \beta},
\end{aligned}$$

904 where we used the fact that $\sigma_t \leq d_{\max}$ for all t (since at every round σ_t can increase at most by one
905 and an observation can remain pending for at most d_{\max} rounds), and the fact that $\sum_t \sigma_t = D$, which
906 follows since when summing the delays, each delay d_t contributes once to exactly d_t rounds with
907 pending observations, and all pending observations are covered in this manner. □

908 We can now prove Theorem 4.6

909 **Theorem B.6** (Restatement of Theorem 4.6). *Let $\gamma = \sqrt{\frac{KT}{\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F}, \eta})}}$. Then the following expected*
910 *regret bound holds for Algorithm 1*

$$\mathbb{E}[\mathcal{R}_T] \leq O \left(\sqrt{KT \left(\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F}, \eta}) \right)} + \sqrt{d_{\max} D \beta} \right).$$

911 *Proof of Theorem 4.6* Putting the results of Claim B.1 (taking expectation on both sides) and Lem-
 912 mas 4.7 to 4.10 all together, the expected regret is bounded as follows.

$$\mathbb{E}[\mathcal{R}_T] \leq d_{\max} + 2\frac{KT}{\gamma} + 2\gamma\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta}) + 2\sqrt{d_{\max}D\beta}.$$

913 Choosing $\gamma = \sqrt{\frac{KT}{\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta})}}$ yields the desired bound. \square

914 B.2 Regret and Stability analysis of Vovk's aggregating forecaster for the square-loss

915 We consider a hedge-based version of Vovk's aggregating forecaster [39], presented in Algorithm 2
 916 for the square loss under the realizability assumption (Assumption 4.1) and a finite function class \mathcal{F} .

917 We denote by $z_t = (x_t, a_t) \in \mathcal{X} \times \mathcal{A}$ the input of each function $f \in \mathcal{F} \subseteq \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ at
 918 time step $t \in [T]$, where $x_1, \dots, x_T \in \mathcal{X}$ is a sequence of contexts generated throughout, and
 919 $a_1, \dots, a_T \in \mathcal{A}$ is the sequence of actions, where a_i was chosen for the context x_i , for all $i \in [T]$.
 920 Also, let $y_1, \dots, y_T \in [0, 1]$ are such that $\mathbb{E}[y_t | z_t] = f_\star(z_t)$, and $f(z_t) \in [0, 1]$ for all $f \in \mathcal{F}$. We
 921 consider the square loss, and prove the following guarantee for the iterates of Algorithm 2

922 **Theorem B.7** (Restatement of Theorem 4.11). *For $t \in [T]$ denote by $q_t \in \Delta(\mathcal{F})$ the probability*
 923 *measure over functions in \mathcal{F} computed by Algorithm 2 at time step t , for the $t-1$ -length prefix of the*
 924 *sequence $\{(z_\tau, y_\tau)\}_{\tau=1}^T$.*

925 *Then, the sequence of measures $\{q_t\}_{t=1}^T$ satisfies the followings for any $\eta \leq 1/18$:*

926 1. *Expected regret:*

$$\sum_{t=1}^T \mathbb{E}[(f_t(z_t) - f_\star(z_t))^2] \leq \frac{2\log|\mathcal{F}|}{\eta}.$$

927 2. *Stability:*

$$\sum_{t=1}^T \mathbb{E}[KL(q_t \| q_{t+1})] \leq 9\eta^2 \cdot \frac{2\log|\mathcal{F}|}{\eta} = 18\eta \log|\mathcal{F}|.$$

928 *Proof.* WLOG, since Hedge is invariant under adding a constant loss in each round we can subtract
 929 $(f_\star(z_t) - y_t)^2$ from the loss of all functions. In particular, after the subtraction, f_\star has a cumulative
 930 loss of 0. Therefore $q_t(f) \propto w_t(f)$, and $w_{t+1}(f) = w_t(f)e^{-\eta((f(z_t) - y_t)^2 - (f_\star(z_t) - y_t)^2)}$. Denote
 931 $W_t = \sum_f w_t(f)$.

932 We have, as $W_1 = |\mathcal{F}|$ and $W_{T+1} \geq 1$ (since $w_t(f_\star) = w_1(f_\star) = 1$ for all t), that

$$\log \frac{W_{T+1}}{W_1} \geq -\log|\mathcal{F}|.$$

933 On the other hand, for small enough η (smaller than a constant),

$$\begin{aligned} \log \frac{W_{T+1}}{W_1} &= \sum_{t=1}^T \log \frac{W_{t+1}}{W_t} \\ &= \sum_{t=1}^T \log \mathbb{E}_{f \sim q_t} \left[e^{-\eta((f(z_t) - y_t)^2 - (f_\star(z_t) - y_t)^2)} \right] \\ &\leq \sum_{t=1}^T \log \mathbb{E}_{f \sim q_t} \left[(1 - \eta((f(z_t) - y_t)^2 - (f_\star(z_t) - y_t)^2) + \eta^2((f(z_t) - y_t)^2 - (f_\star(z_t) - y_t)^2)^2) \right] \\ &\quad (e^x \leq 1 + x + x^2 \text{ for } x < 1) \\ &\leq \sum_{t=1}^T -\eta \mathbb{E}_{f \sim q_t} [(f(z_t) - y_t)^2 - (f_\star(z_t) - y_t)^2] + \eta^2 \mathbb{E}_{f \sim q_t} [(f(z_t) - y_t)^2 - (f_\star(z_t) - y_t)^2]^2 \\ &\quad (\log(1 + x) \leq x) \end{aligned}$$

$$\begin{aligned}
&= \sum_{t=1}^T -\eta \mathbb{E}_{f \sim q_t} [(f(z_t) - f_*(z_t))^2 + 2(f(z_t) - f_*(z_t))(f_*(z_t) - y_t)] \\
&\quad + \eta^2 \mathbb{E}_{f \sim q_t} [((f(z_t) - f_*(z_t))^2 + 2(f(z_t) - f_*(z_t))(f_*(z_t) - y_t))^2] \\
&\leq \sum_{t=1}^T -\eta \mathbb{E}_{f \sim q_t} [(f(z_t) - f_*(z_t))^2 + 2(f(z_t) - f_*(z_t))(f_*(z_t) - y_t)] \\
&\quad + 9\eta^2 \mathbb{E}_{f \sim q_t} (f(z_t) - f_*(z_t))^2.
\end{aligned}$$

934 Rearranging, we obtain that

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}_{f \sim q_t} [(f(z_t) - f_*(z_t))^2 + 2(f(z_t) - f_*(z_t))(f_*(z_t) - y_t)] \\
&\leq \frac{\log |\mathcal{F}|}{\eta} + 9\eta \sum_{t=1}^T \mathbb{E}_{f \sim q_t} [(f(z_t) - f_*(z_t))^2].
\end{aligned}$$

935 Taking expectation over y_1, \dots, y_T :

$$\mathbb{E}_{y_1, \dots, y_T} \left[\sum_{t=1}^T \mathbb{E}_{f \sim q_t} [(f(z_t) - f_*(z_t))^2] \right] \leq \frac{\log |\mathcal{F}|}{\eta} + 9\eta \mathbb{E}_{y_1, \dots, y_T} \left[\sum_{t=1}^T \mathbb{E}_{f \sim q_t} [(f(z_t) - f_*(z_t))^2] \right].$$

936 If, suppose $\eta \leq 1/18$ then we immediately obtain

$$\mathbb{E}_{y_1, \dots, y_T} \left[\sum_{t=1}^T \mathbb{E}_{f \sim q_t} [(f_t(z_t) - f_*(z_t))^2] \right] \leq \frac{2 \log |\mathcal{F}|}{\eta},$$

937 and the expected regret of Algorithm 2 can now be bounded by

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E} [(f_t(z_t) - f_*(z_t))^2] &= \sum_{t=1}^T \mathbb{E} \left[\left(\sum_{f \in \mathcal{F}} q_t(f) (f(z_t) - f_*(z_t)) \right)^2 \right] \\
&\leq \sum_{t=1}^T \mathbb{E} \left[\sum_{f \in \mathcal{F}} q_t(f) (f(z_t) - f_*(z_t))^2 \right] \\
&= \mathbb{E}_{y_1, \dots, y_T} \left[\sum_{t=1}^T \mathbb{E}_{f \sim q_t} [(f_t(z_t) - f_*(z_t))^2] \right] \\
&\leq \frac{2 \log |\mathcal{F}|}{\eta},
\end{aligned}$$

938 where we have used Jensen's inequality. This concludes the proof of part 1. of the theorem.

939 For the second part, we observe that

$$\begin{aligned}
KL(q_t \| q_{t+1}) &= \mathbb{E}_{f \sim q_t} \left[\log \frac{q_t(f)}{q_{t+1}(f)} \right] \\
&= \eta \mathbb{E}_{f \sim q_t} [(f(z_t) - y_t)^2] + \log \mathbb{E}_{f \sim q_t} e^{-\eta(f(z_t) - y_t)^2} \\
&= \eta \mathbb{E}_{f \sim q_t} [f(z_t) - y_t)^2 - (f_*(z_t) - y_t)^2] + \log \mathbb{E}_{f \sim q_t} [e^{-\eta((f(z_t) - y_t)^2 - (f_*(z_t) - y_t)^2)}] \\
&\leq \eta \mathbb{E}_{f \sim q_t} [(f(z_t) - y_t)^2 - (f_*(z_t) - y_t)^2] \\
&\quad + \log \mathbb{E}_{f \sim q_t} [1 - \eta((f(z_t) - y_t)^2 - (f_*(z_t) - y_t)^2) + \eta^2((f(z_t) - y_t)^2 - (f_*(z_t) - y_t)^2)^2] \\
&\quad \quad \quad (e^x \leq 1 + x + x^2 \text{ for } x < 1) \\
&\leq \eta^2 \mathbb{E}_{f \sim q_t} [(f(z_t) - y_t)^2 - (f_*(z_t) - y_t)^2]^2 \quad (\log(1 + x) \leq x) \\
&= \eta^2 \mathbb{E}_{f \sim q_t} [((f(z_t) - f_*(z_t))^2 + 2(f(z_t) - f_*(z_t))(f_*(z_t) - y_t))^2]
\end{aligned}$$

$$\leq 9\eta^2 \mathbb{E}_{f \sim q_t} [(f(z_t) - f_\star(z_t))^2].$$

Therefore, taking expectation over y_1, \dots, y_T and using part 1. we obtain

$$\sum_{t=1}^T \mathbb{E}[KL(q_t \| q_{t+1})] \leq 9\eta^2 \cdot \frac{2 \log |\mathcal{F}|}{\eta} = 18\eta \log |\mathcal{F}|,$$

yields the second part of the theorem. \square

C Lower Bounds

C.1 Proof of Theorem 4.3

In this subsection, we present a lower bound indicating that an additional assumption on the oracle is necessary in order to obtain sub-linear regret in the general function approximation setting. The lower bound shows that with no additional assumption on the least squares oracle, any algorithm incurs linear regret in the presence of delayed feedback, even for a constant delay of $d = 1$.

Theorem C.1 (Restatement of Theorem 4.3). *For any CMAB algorithm ALG in the function approximation setting (that is, ALG can only access \mathcal{F} via the oracle) there exists a CMAB instance with constant delay $d = 1$ over a realizable loss function class \mathcal{F} with $|\mathcal{F}| = T + 1$ with an online oracle $\mathcal{O}_{sq}^{\mathcal{F}}$ satisfying $\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F}}) = 0$, on which ALG attains regret $\mathcal{R}_T = \Omega(T)$.*

Proof. Consider a CMAB instance over $\mathcal{X} = \{x_1, x_2, \dots, x_T\}$, $\mathcal{A} = \{a_1, a_2\}$ and $\mathcal{F} = \{f_1, f_2, \dots, f_T, f_\star\}$, where f_\star is the true loss function. The functions in \mathcal{F} are sampled randomly as follows:

$$f_\star(x_i, a_1) = \text{Ber}\left(\frac{1}{2}\right), \quad f_\star(x_i, a_2) = 1 - f_\star(x_i, a_1), \quad i \in \{T\},$$

$$f_i(x, a) = \begin{cases} f_\star(x, a), & x = x_i, \\ \text{Ber}\left(\frac{1}{2}\right), & \text{otherwise}, \end{cases} \quad i \in [T].$$

The online sequence of contexts is defined by $x_1, x_2, x_3, \dots, x_T$ in order. We consider an oracle which at round t outputs the function $\hat{f}_t = f_t$. It is easy to see that the least-squares regret of this oracle is zero because $\hat{f}_t(x_t, \cdot) = f_\star(x_t, \cdot)$. Now, at round t , due to the delay, ALG only has relevant information on x_t given by $\{f_1(x_t, \cdot), \dots, f_{t-1}(x_t, \cdot)\}$, all of which are random i.i.d. $\text{Ber}(\frac{1}{2})$ random variables, with the true loss $f_\star(x_t, \cdot)$ being either $(1, 0)$ or $(0, 1)$ with equal probability and independently of the previous observations of ALG. Therefore, however ALG chooses the next action $a^{(t)}$, with probability $\frac{1}{2}$ it will incur a loss of 1 while simultaneously the other action will have a loss of zero. This means that in expectation over the random construction of \mathcal{F} , the algorithm will incur $\Omega(\frac{T}{2})$ regret. By the probabilistic method, we know that there exists a fixture of \mathcal{F} depending on ALG on which ALG suffers linear regret, as claimed. \square

C.2 Lower Bounds for Contextual MAB with Delayed Feedback

In this subsection, we establish lower bounds on the expected regret for CMAB with delayed feedback. Our construction is based on the approach of [10] via a reduction from the full information variant with non-delayed feedback using a blocking argument.

We begin with a lower bound for the policy class setting with a finite policy class Π , which relies on a reduction from the problem of (agnostic) prediction with expert advice, for which known lower bounds exist in the literature (see e.g. [9]).

We then present a lower bound for the realizable function approximation setting with a finite loss function class \mathcal{F} , and for that we construct an explicit hard instance for the full-information non-delayed variant, with a regret lower bound of $\Omega(\sqrt{T \log |\mathcal{F}|})$.

We remark that while a regret lower bound of $\Omega(\sqrt{KT} + \sqrt{D})$ can be immediately inferred from the results of [9] who consider the special case of multi-armed bandits, our goal is to show that the

dependence on $\log |\mathcal{F}|$ where \mathcal{F} appears jointly with the delay dependence. While the dependence of $\sqrt{KT \log |\mathcal{F}|}$ is known to be tight for CMAB with general function approximation, it is nontrivial that the delay dependent term also contains a dependence on $\log |\mathcal{F}|$, which we prove in the construction that follows.

In our construction we consider the full feedback setting, where for each round t and observed context $x_t \in \mathcal{X}$, the learner observes the entire loss vector $(\ell(x_t, a))_{a \in \mathcal{A}}$ after choosing an action $a_t \in \mathcal{A}$.

Theorem C.2. *There exists a finite policy class $\Pi \subseteq \mathcal{A}^{\mathcal{X}}$ mapping contexts to actions, a delay sequence (d_1, \dots, d_T) with maximal delay d and sum of delays $D = \Theta(dT)$ such that for any CMAB algorithm there is instance of the CMAB problem for which the algorithm incurs expected regret*

$$\mathbb{E}[\mathcal{R}_T] \geq \Omega(\sqrt{D \log |\Pi|}).$$

Proof. We observe that CMAB with a policy class can be viewed as a special case of the prediction with expert advice framework [9], where each policy corresponds to an expert, provides a prediction for each context. Hence, the classical lower bound of $\Omega(\sqrt{T \log |\Pi|})$ for the full-information expert setting (see Cesa-Bianchi and Lugosi [9], chapter 2) applies in the absence of delays.

Returning to the delayed CMAB problem, construct a delay sequence (d_1, \dots, d_T) in which d is the maximal delay and $D = \sum_{t=1}^T d_t = \Theta(dT)$ as follows:

Divide the time horizon into $T/(d+1)$ blocks, each containing $d+1$ consecutive rounds. For each block $b \in \{0, 1, \dots, T/(d+1)-1\}$ and each round $\tau \in \{b(d+1), b(d+1)+1, \dots, (b+1)(d+1)-1\}$, define the delay as $d_\tau = d - (\tau - b(d+1))$. That is, within each block, the delays decrease from d to 0, in this corresponding order. This also implies that $D = \frac{T}{d+1} \sum_{i=0}^d i = \frac{T}{d+1} \cdot \frac{(d+1)(d+0)}{2} = \frac{Td}{2}$. This construction ensures that feedback from all rounds within a block is revealed simultaneously at the end of the block.

The loss sequence is constructed as follows: Consider the loss sequence $(\ell_1, \dots, \ell_{T/(d+1)})$ given by a lower bound construction for prediction with expert advice over $T/(d+1)$ rounds. The loss of the first round of each block b is defined to be ℓ_b , and remains the same throughout the block. Now, note that given this construction, the algorithm essentially faces a prediction with expert advice problem over $T/(d+1)$ rounds (the rounds on which information is obtained), with loss values in the range $[0, d+1]$. We remark that we can assume without loss of generality that the algorithm fixes a policy π_b at the start of block b and uses it to play actions throughout the entire block, as it does not learn new information within the block.

Thus, we can aggregate each block into a single “super-round” of a reduced expert problem. Specifically, for block b , define the aggregate loss of each expert π as $\ell_b(\pi) = \sum_{\tau=b(d+1)}^{(b+1)(d+1)-1} \ell(x_\tau, \pi(x_\tau))$. Even if we allow the algorithm to observe full feedback, it essentially observes the full aggregated loss vector over actions in each block, so this construction corresponds to a well-defined instance of prediction with expert advice over the $T/(d+1)$ rounds which are the initial rounds of the blocks.

The resulting reduced problem has $T/(d+1)$ rounds with losses in $[0, d+1]$. Applying the lower bound from Cesa-Bianchi and Lugosi [9] to this reduced problem yields:

$$\mathbb{E}[\mathcal{R}_T] \geq \Omega \left((d+1) \cdot \sqrt{\frac{T}{d+1} \log |\Pi|} \right) = \Omega \left(\sqrt{(d+1)T \log |\Pi|} \right) = \Omega \left(\sqrt{D \log |\Pi|} \right),$$

which completes the proof. \square

We now combine this result with the classical lower bound of $\Omega(\sqrt{KT \log |\Pi|})$ for CMAB with bandit feedback and a finite policy class, which is based on reductions from prediction with expert advice (see, e.g., [29, 6, 7, 9]). This yields the following lower bound for CMAB with delayed feedback in the policy class setting:

Corollary C.3. *For CMAB with delayed bandit feedback and a finite policy class Π , the expected regret satisfies*

$$\mathbb{E}[\mathcal{R}_T] \geq \Omega \left(\sqrt{TK \log |\Pi|} + \sqrt{D \log |\Pi|} \right).$$

1020 To prove a corresponding regret lower bound for the realizable function approximation setting, we
 1021 similarly require a regret lower bound of $\Omega\left(\sqrt{T \log |\mathcal{F}|}\right)$ for the full-information non-delayed variant
 1022 of the problem. Such a lower bound, however, does not exist in the literature as far as we are aware,
 1023 so we exhibit an explicit construction in the following lemma.

1024 **Lemma C.4.** *Let $\mathcal{A} = \{a_1, a_2\}$ be action set, and let $\mathcal{X} = \{x_1, \dots, x_n\}$ be a set of n contexts where
 1025 $n \leq T$. Then for any CMAB algorithm there exists a finite loss function class $\mathcal{F} \subseteq \{\mathcal{X} \times \mathcal{A} \rightarrow [0, 1]\}$
 1026 of size $|\mathcal{F}| = 2^n$ and a CMAB instance which is realizable with respect to \mathcal{F} , on which the expected
 1027 regret of the CMAB algorithm is lower bounded by*

$$\mathbb{E}[\mathcal{R}_T] \geq \Omega\left(\sqrt{nT}\right) = \Omega\left(\sqrt{T \log |\mathcal{F}|}\right).$$

1028 *Proof.* Across all of the instances which we construct, the context is chosen uniformly at random
 1029 from \mathcal{X} . We define the function class \mathcal{F} as the set of 2^n functions f which, for each $x \in \mathcal{X}$, are
 1030 defined via $f(x, a_i) = \frac{1}{2} - \varepsilon$ and $f(x, a_j) = \frac{1}{2}$ for the other action $a_j \neq a_i$ (that is, each function in
 1031 \mathcal{F} has a distinct choice of optimal actions across all n contexts), and we choose $\varepsilon = \sqrt{n/100T}$.

1032 Prior to the interaction, a function $f_\star \in \mathcal{F}$ is selected uniformly at random and the losses are defined
 1033 to be Bernoulli random variables according to f_\star , ensuring realizability holds. More specifically,
 1034 $\ell(x, a)$ will be a Bernoulli random variable with parameter $f_\star(x, a)$ for all $x \in \mathcal{X}, a \in \mathcal{A}$.

1035 Now, by standard arguments of statistical estimation, since the true loss function f_\star was sampled at
 1036 random, as long as a given context $x \in \mathcal{X}$ has not appeared more than $\Omega(1/\varepsilon^2)$ times, the CMAB
 1037 algorithm must incur instantaneous regret of ε conditioned on this context. Since the contexts are
 1038 sampled uniformly at random and the loss values for one context reveal no information about the
 1039 loss for different contexts, the algorithm must incur expected regret of at least $\Omega(\varepsilon t)$ on the first t
 1040 rounds as long as each context has been sampled $o(1/\varepsilon^2)$ times. With high probability, all contexts
 1041 are sampled sufficiently many times only after $t = \Omega(n/\varepsilon^2)$ rounds, implying that the expected
 1042 regret of the algorithm over T rounds is lower bounded by

$$\mathbb{E}[\mathcal{R}_T] \geq \Omega\left(\varepsilon \cdot \frac{n}{\varepsilon^2}\right) = \Omega\left(\frac{n}{\varepsilon}\right) = \Omega\left(\sqrt{nT}\right),$$

1043 which concludes the proof. □

1044 We remark that the construction in the above proof is similar to the lower bound given by [3] for the
 1045 bandit case, but here the proof is considerably simpler as the algorithm is not required to perform
 1046 exploration in order to obtain sufficient feedback.

1047 Thus, by combining the lower bound for contextual bandits with function approximation under bandit
 1048 feedback [3] with the delayed feedback result above using the same reduction as we described in the
 1049 proof of Theorem C.2, we obtain:

1050 **Corollary C.5.** *For any CMAB algorithm in the realizable function approximation setting over finite
 1051 loss function classes, there exists a finite function class \mathcal{F} and a distribution over losses which is
 1052 realizable by \mathcal{F} , for which the expected regret satisfies*

$$\mathbb{E}[\mathcal{R}_T] \geq \Omega\left(\sqrt{TK \log |\mathcal{F}|} + \sqrt{D \log |\mathcal{F}|}\right).$$