Online Active Model Selection for Pre-trained Classifiers: Supplementary Materials

A Proofs and Supplementary Lemmas **13**

A.1 On the Choice of Query Probability . . . 13

A.2 Mix Loss Properties . . . 13

A.3 Lemmas for Regret Bound . . . 14

A.4 Proof of Theorem 2 . . . 16

A.5 Proof of Theorem 6 . . . 16

A.6 Proof of Theorem 4 . . . 17

A.7 Proof of Theorem 3 . . . 18

A.8 Proof of Theorem 7 . . . 18

A.9 Proof of Theorem 8 . . . 18

B Example for Large Number of Updates **18**

C Experiments **18**

C.1 Details on the Model Collections . . . 18

C.2 Note on the Baselines . . . 19

C.3 Performance of Model Picker on models with low accuracies . . . 20

C.4 Hyperparameters . . . 21

A Proofs and Supplementary Lemmas

A.1 On the Choice of Query Probability

Here we elaborate on the discussion for (1). Let $\boldsymbol{w}_t$ be the current distribution over experts, and $\boldsymbol{w}_{t+1}^{(c)}$ be the hypothetical distribution over the experts having observed the loss if the true label is c.

First, according to (Narayanan and Rakhlin, 2010, Lemma 1), the divergence $\mathrm{KL}(\boldsymbol{w}_t \| \boldsymbol{w}_{t+1}^{(c)})$ accumulates into the regret. This means that if we do *not* update $\boldsymbol{w}_t$ accordingly, we *miss* this amount of information.

Second, the KL divergence between $\boldsymbol{w}_t$ and $\boldsymbol{w}_{t+1}^{(c)}$ computes

$$\mathrm{KL}(\boldsymbol{w}_t \| \boldsymbol{w}_{t+1}^{(c)}) = \eta r + \log(r e^{-\eta} + 1 - r),$$

where $r = \langle \boldsymbol{w}_t, \boldsymbol{\ell}_t^c \rangle$. By Höffdings inequality, one can show that this quantity is between 0 and $\eta^2/8$. As the variance is between 0 and $1/4$, it makes sense to scale the KL divergence to the range $[0, 1/4]$, by multiplying it by $2/\eta^2$. The following lemma completes the comparison promised in Section 4.1. We drop the subscript t for readability.

Lemma 1. *For a distribution over the experts $\boldsymbol{w}$ and a fixed $\boldsymbol{\ell} \in \{0,1\}^k$, define $\boldsymbol{w}_+ \propto \boldsymbol{w}\, e^{-\eta \boldsymbol{\ell}}$. Then*

$$\left| \operatorname*{Var}_{A \sim \boldsymbol{w}} \boldsymbol{\ell}_A - \frac{2}{\eta^2} \mathrm{KL}(\boldsymbol{w} \| \boldsymbol{w}_+) \right| < \frac{1}{18\sqrt{3}} \eta + O(\eta^2).$$

Proof. Define $r = \langle \boldsymbol{w}, \boldsymbol{\ell} \rangle$. We have that

$$\mathrm{KL}(\boldsymbol{w} \| \boldsymbol{w}_+) = \eta r + \log(r e^{-\eta} + 1 - r) = \log \mathbb{E}\, e^{-\eta (X - \mathbb{E}\, X)},$$

where X is a Bernoulli random variable with $\mathbb{E}\, X = r$. Note that the equation above is the cumulant generating function of X and has the Taylor series

$$\log \mathbb{E}\, e^{-\eta (X - \mathbb{E}\, X)} = \frac{\eta^2}{2} \operatorname{Var}(X) + \frac{\eta^3}{6} \kappa_3 + O(\eta^4),$$

where κ_3 is the third cumulant. Note that by the relation between cumulants of a Bernoulli random variable, we have

$$\kappa_3 = r(1-r) \frac{d}{dr} \kappa_2 = r(1-r)(1-2r).$$

Easy algebra finds that for $r \in [0,1]$, we have $\kappa_3 \in [-1/6\sqrt{3}, 1/6\sqrt{3}]$.

Summing all up, we find

$$\frac{2}{\eta^2} \mathrm{KL}(\boldsymbol{w} \| \boldsymbol{w}_+) = \operatorname{Var}(X) + \frac{\kappa_3}{3} \eta + O(\eta^2),$$

and the result of the lemma follows. □

A.2 Mix Loss Properties

Define $\widehat{\boldsymbol{L}}_{T,*} := \min_{i \in [k]} \widehat{\boldsymbol{L}}_{T,i}$.

Lemma 2. *The cumulative mix loss M_T is bounded above by $\widehat{\boldsymbol{L}}_{T,*} + \frac{\log k}{\eta_T}$.*

Before stating the proof, first we bring a standard lemma:

Lemma 3 (de Rooij et al. (2013))**.** *The cumulative mix loss M_T has the following properties for constant learning rates ($\eta_t \equiv \eta$ for all $t \geq 1$):*

(i) $M_T = -\frac{1}{\eta} \log \sum_{i \in [k]} e^{-\eta \widehat{\boldsymbol{L}}_{T,i}} + \frac{1}{\eta} \log k,$

(ii) $M_T \leq \widehat{\boldsymbol{L}}_{T,*} + \frac{1}{\eta} \log k.$

Moreover, for any sequence of decaying learning rates $\{\eta_t\}_{t\geq 1}$, let $M_T(\{\eta_t\})$ be the corresponding cumulative mix loss, and set $M_T(\eta_T)$ be the cumulative mix loss for fixed learning rate η_T. Then, it holds that $M_T(\{\eta_t\}) \leq M_T(\eta_T)$.

Proof. Define $W_t = \sum_{i\in[k]} e^{-\eta \widehat{\boldsymbol{L}}_{t,i}}$. For part (i) observe that

$$\langle \boldsymbol{w}_t, e^{-\eta\hat{\boldsymbol{\ell}}_t}\rangle = \sum_{i\in[k]} \frac{e^{-\eta\widehat{\boldsymbol{L}}_{t-1,i}}}{W_{t-1}} e^{-\eta\hat{\boldsymbol{\ell}}_{t,i}} = \frac{W_t}{W_{t-1}}.$$

Hence, $\log\langle \boldsymbol{w}_t, e^{-\eta\hat{\boldsymbol{\ell}}_t}\rangle = -\eta m_t = \log W_t - \log W_{t-1}$, and $M_T = \frac{1}{\eta}(\log W_0 - \log W_T)$. Observing that $W_0 = k$ gives (i).

Noticing that $W_T \geq e^{-\eta\widehat{\boldsymbol{L}}_{T,*}}$ easily implies (ii).

For the last part of the lemma, first we prove that M_T for constant learning rate η is nonincreasing in η. This is shown by looking at the derivative of M_T with respect to η which is equal to

$$\tfrac{1}{\eta^2}\log\sum_{i\in[k]} e^{-\eta\widehat{\boldsymbol{L}}_{T,i}} - \frac{1}{\eta}\frac{\sum \widehat{\boldsymbol{L}}_{T,i} e^{-\eta\widehat{\boldsymbol{L}}_{T,i}}}{\sum e^{-\eta\widehat{\boldsymbol{L}}_{T,i}}} - \tfrac{1}{\eta^2}\log k \leq 0,$$

as $\log\sum_{i\in[k]} e^{-\eta\widehat{\boldsymbol{L}}_{T,i}} \leq \log k$.

Now we can prove the last part of the lemma.

$$\sum_{t=1}^{T} m_t(\{\eta_t\}) = \sum_{t=1}^{T} M_t(\eta_t) - M_{t-1}(\eta_t) \leq \sum_{t=1}^{T} M_t(\eta_t) - M_{t-1}(\eta_{t-1}) \leq M_T(\eta_T). \qquad \square$$

Proof of Lemma 2. By the lemma above, we see that $M_T \leq M_T(\eta_T) \leq \widehat{\boldsymbol{L}}_{T,*} + \frac{\log k}{\eta_T}$, where we lower bounded the sum over the models by the one that corresponds to $\widehat{\boldsymbol{L}}_{T,*}$. $\square$

A.3 Lemmas for Regret Bound

Lemma 4. *For any $\eta > 0$ it holds that*

$$\mathbb{E}\Big[\log\langle \boldsymbol{w}_t, e^{-\eta\hat{\boldsymbol{\ell}}_t}\rangle + \eta\langle \boldsymbol{w}_t, \hat{\boldsymbol{\ell}}_t\rangle\Big] \leq \eta^2.$$

Proof. If the predictions are all the same, there is nothing to prove, as $\hat{\boldsymbol{\ell}}_t \equiv 0$ and the expectation vanishes.

Let $\boldsymbol{w}_c = \sum_{i:\boldsymbol{p}_t(i)=c} \boldsymbol{w}_{t,i}$ and set c^* to be the true label of this round. We can then rewrite q_t as $q_t = \max_{c\in\mathcal{C}} \boldsymbol{w}_c(1-\boldsymbol{w}_c) \vee \eta$. Observe that

$$\log\langle \boldsymbol{w}_t, e^{-\eta\hat{\boldsymbol{\ell}}_t}\rangle = \log\Big[(1-\boldsymbol{w}_{c^*})\exp\Big\{-\tfrac{\eta}{q_t}Q_t\Big\} + \boldsymbol{w}_{c^*}\Big].$$

It is clear that $\mathbb{E}\,\langle \boldsymbol{w}_t, \hat{\boldsymbol{\ell}}_t\rangle = \langle \boldsymbol{w}_t, \ell_t\rangle = 1 - \boldsymbol{w}_{c^*}$. Hence, the expected value in the lemma is equal to

$$\eta(1-\boldsymbol{w}_{c^*}) + q_t \cdot \log[(1-\boldsymbol{w}_{c^*})\exp\{-\eta/q_t\} + \boldsymbol{w}_{c^*}]. \tag{2}$$

Our desired result follows from Lemma 5 by setting $x = 1 - \boldsymbol{w}_{c^*}$ and noticing that $q_t \geq \boldsymbol{w}_{c^*}(1-\boldsymbol{w}_{c^*}) \vee \eta$. $\square$

Lemma 5. *For all $x \in (0,1)$ and all $\eta \in (0,1]$, defining $u = x(1-x)\vee\eta$, one has*

$$f(x,u,\eta) := \eta x + u\log\Big[x\,e^{-\frac{\eta}{u}} + 1 - x\Big] \leq \eta^2.$$

Moreover, for fixed x and η, $f(x,u,\eta)$ is decreasing in u for $u \geq \eta$.

Proof. Note that the value of the LHS and RHS agree when $\eta = 0$, so we have to prove that for all $\eta \geq 0$, the derivative of the LHS is at most η. Fix some $x \in [0, 1)$. We prove this fact in two cases:

Case where $\eta \leq x(1-x)$. In this case, $u = x(1-x)$. For brevity, define $y := \frac{\eta}{x(1-x)}$. The derivative of f with respect to η becomes

$$\frac{(e^y - 1)x(1-x)}{e^y(1-x) + x},$$

and we are left with proving

$$\frac{(e^y - 1)}{e^y(1-x) + x} \leq 2y$$

As $0 < y \leq 1$, we have that $1 + y + y^2/2 \leq e^y \leq 1 + y + (e-2)y^2$. Replacing these bounds in the equation above leaves us with proving that

$$\frac{1 + (e-2)y}{1 + (y + y^2/2)(1-x)} \leq 2.$$

For a fixed y, the left hand side is increasing in x, and hence, it is enough to prove that

$$1 + (e-2)y \leq 2,$$

but this is true as $y \leq 1$ and $e - 1 < 2$. Thus, we are done with the proof of this case.

Case where $\eta > x(1-x)$. In this case, $u = \eta$, and $f(x, \eta) = \eta x + \eta \log\big[xe^{-1} + 1 - x\big]$. To prove the claim, we have to show that $x + \log\big[xe^{-1} + 1 - x\big] \leq \eta$, or, as the left hand side does not depend on η, we shall prove

$$x + \log\big[xe^{-1} + 1 - x\big] \leq x(1-x),$$

or, equivalently,

$$1 - (1 - e^{-1})x \leq e^{-x^2},$$

which is proven in Lemma 6. Thus, in both cases, we have proved our inequality and we are done with the proof of the first part of lemma.

We now prove the monotonicity of f with respect to u. For that, we show the derivative of f with respect to u is nonpositive. The derivative computes

$$f'(u) = \log\Big[xe^{-\eta/u} + 1 - x\Big] + \frac{xe^{-\eta/u}\eta/u}{xe^{-\eta/u} + 1 - x}.$$

Define $a = \eta/u$. The equation above is zero for $a = 0$. So it suffices to show that the derivative of above is nonpositive for $0 \leq a \leq 1$. Computing the derivative w.r.t. a and setting it less than 0 is equivalent to

$$xe^{-a} + 1 - x - xe^{-a} \geq 0,$$

which is true. Hence, we are done. □

Lemma 6. *For all $x \in [0, 1]$ one has $1 - (1 - e^{-1})x \leq \exp\{-x^2\}$.*

Proof. Note that $\exp(-x^2)$ is concave on $[0, \sqrt{1/2}]$ and convex on $[\sqrt{1/2}, 1]$. Also at $x = 0$ and $x = 1$, both sides are equal. Hence, we just have to show that at $x = \sqrt{1/2}$, the right hand side is bigger than the left hand side, which automatically shows the inequality for $x \in [0, \sqrt{1/2}]$, and we have to show that the derivative of the right hand side is smaller than the left hand side at $x = 1$, which automatically shows the inequality for the other half of the interval, as $\exp(-x^2)$ is convex there. For the first part, evaluate

$$\exp(-1/2) - 1 + (1 - 1/e)\sqrt{1/2} = (1 - 1/\sqrt{e})\left(\frac{1}{\sqrt{2}}(1 + \frac{1}{\sqrt{e}}) - 1\right) \geq 0.$$

For the second part, note that $\frac{d}{dx}\exp(-x^2) = -2x\exp(-x^2)$, and at $x = 1$ it is equal to

$$-2/e \leq -(1 - 1/e),$$

as $3/e > 1$. Hence, we are done. □

A.4 Proof of Theorem 2

Proof. We assume that at all rounds we have $q_t > 0$, as there is no label request on the rounds that all models predict the same. First observe that

$$\mathbb{E}\sum_{t=1}^{T} Q_t \le \mathbb{E}\left\{\sum_{t=1}^{T}\eta_t + \textstyle\sum_{c\in C}\boldsymbol{w}_{t,c}(1-\boldsymbol{w}_{t,c})\right\},$$

as maximum of positive numbers is less than their sum. Next, at round t suppose that the true label is c_t. As $x(1-x)$ is concave and $\sum_{c\neq c_t}\boldsymbol{w}_{t,c} = 1-\boldsymbol{w}_{t,c_t} = \langle \boldsymbol{w}_t, \boldsymbol{\ell}_t\rangle =: r_t$, using Jensen's inequality we have

$$\begin{aligned}
&\sum_{c\in C}\boldsymbol{w}_{t,c}(1-\boldsymbol{w}_{t,c})\\
&\qquad = \boldsymbol{w}_{t,c_t}(1-\boldsymbol{w}_{t,c_t}) + \sum_{c\neq c_t}\boldsymbol{w}_{t,c}(1-\boldsymbol{w}_{t,c})\\
&\qquad \le \boldsymbol{w}_{t,c_t}(1-\boldsymbol{w}_{t,c_t}) + (1-\boldsymbol{w}_{t,c_t})\Big(1-\tfrac{1-\boldsymbol{w}_{t,c_t}}{|C|-1}\Big)\\
&\qquad = r_t\Big(2-\tfrac{|C|}{|C|-1}r_t\Big).
\end{aligned}$$

Using Jensen now for the concave function $x(2-\frac{|C|}{|C|-1}x)$, we get

$$\sum_{t=1}^{T}\sum_{c\in C}\boldsymbol{w}_{t,c}(1-\boldsymbol{w}_{t,c}) \le T\cdot\Big(\tfrac{\sum r_t}{T}\Big)\Big(2-\tfrac{|C|}{|C|-1}\tfrac{\sum r_t}{T}\Big).$$

Now observe that if the expected total loss of the best model is L^*, by our regret bound in Theorem 1 we have

$$\mathbb{E}\sum r_t \le 2\sqrt{2T\log k} + L^*.$$

Also note that for $x \le \frac{|C|-1}{|C|}$, the function $x(2-\frac{|C|}{|C|-1}x)$ is increasing. Hence, for large enough T (as described in the theorem), $\frac{1}{T}\sum r_t \le \frac{|C|-1}{|C|}$, and we have

$$\mathbb{E}\sum_{t=1}^{T}\sum_{c\in C}\boldsymbol{w}_{t,c}(1-\boldsymbol{w}_{t,c}) \le \Big(2\sqrt{2T\log k}+L^*\Big)\cdot\Big(2-\tfrac{|C|}{|C|-1}\Big(2\sqrt{2\log k/T}+L^*/T\Big)\Big).$$

Noting that $\sum\eta_t \le \sqrt{2T\log k}$, one obtains the result. □

A.5 Proof of Theorem 6

Proof. First, we remind the following martingale inequality, which is an improved version of McDiarmid's:

Lemma 7 (Seldin and Lugosi (2017))**.** *Let $\xi_1,\dots,\xi_T$ be a martingale difference sequence with respect to the filteration $\{\mathcal{F}_t\}_{t\le T}$, where each ξ_t is integrable and bounded. Let $M_t := \sum_{s\le t}\xi_s$ be the associated martingale. Define $\nu_T = \sum_{t\le T}\mathbb{E}\{\xi_t^2 \mid \mathcal{F}_{t-1}\}$ and $c_T = \max_{t\le T}\xi_t$. Then for any $\beta,\nu,c>0$,*

$$\Pr\left\{\left(M_T \ge \sqrt{2\nu\beta T} + \frac{1}{3}c\beta T\right)\wedge(\nu_T\le\nu)\wedge(c_T\le c)\right\} \le e^{-\beta T}.$$

Remember that the weight of model i at the end of round t is proportional to $\exp\{-\eta_{t+1}\widehat{\boldsymbol{L}}_t\}$. Hence, identifying the best model i^* after round t reduces to the fact that $\widehat{\boldsymbol{L}}_{t,i^*} = \min_{j\in[k]}\widehat{\boldsymbol{L}}_{t,j}$. The probability of this event not happening can be bounded by a union bound on the models:

$$\Pr\{\exists j\neq i^* : \widehat{\boldsymbol{L}}_{t,i^*} \ge \widehat{\boldsymbol{L}}_{t,j}\} \le \sum_{j\neq i^*}\Pr\{\widehat{\boldsymbol{L}}_{t,i^*} \ge \widehat{\boldsymbol{L}}_{t,j}\} = \sum_{j\neq i^*}\Pr\{\widetilde{D}_{t,j}\le 0\},$$

where we define $\widetilde{D}_{t,j} = \widehat{\boldsymbol{L}}_{t,j} - \widehat{\boldsymbol{L}}_{t,i^*}$. From now on, we focus on a single model j and drop the index j from Δ_j and θ_j. Set $d_t := \boldsymbol{\ell}_{t,j} - \boldsymbol{\ell}_{t,i^*}$ and define

$$\xi_t := \Delta - \frac{d_t Q_t}{q_t}.$$

Note that $\mathbb{E}\{\xi_t \mid \mathcal{F}_{t-1}\} = \mathbb{E}\{\mathbb{E}\{\xi_t \mid \boldsymbol{\ell}_t, \mathcal{F}_{t-1}\} \mid \mathcal{F}_{t-1}\} = 0$. Moreover, the following holds:

$$\xi_t \le \Delta + \frac{1}{q_t} \le \Delta + \eta_t^{-1}, \quad \mathbb{E}\{\xi_t^2 \mid \mathcal{F}_{t-1}\} = \mathbb{E}\Big\{\tfrac{d_t^2}{q_t} \mid \mathcal{F}_{t-1}\Big\} - \Delta^2 \le \theta\eta_t^{-1} - \Delta^2.$$

The sum of the conditional variances up to T satisfies

$$\sum_{t=1}^{T} \mathbb{E}\{\xi_t^2 \mid \mathcal{F}_{t-1}\} \le T\eta_T^{-1}\theta - T\Delta^2 =: \nu$$

Also, set $c = \Delta + \eta_T^{-1}$. By lemma above we have

$$\Pr\left\{\widetilde{D}_{T,j} = \sum_{t=1}^{T} \frac{d_t Q_t}{q_t} \le T\Delta - \sqrt{2T\nu\beta} - \tfrac{1}{3}c\beta T\right\} \le e^{-\beta T}.$$

We will find the largest β such that the right hand side of the inequality above becomes positive. As it is a quadratic polynomial in $\sqrt{\beta}$, we should have that

$$\sqrt{\beta} \le \frac{\sqrt{2\nu/T + \frac{4}{3}c\Delta} - \sqrt{2\nu/T}}{\frac{2}{3}c}.$$

Now we lower bound the right hand side, and write $\gamma := \eta_T^{-1}$ for brevity:

$$\begin{aligned}
\frac{\sqrt{2\nu/T + \frac{4}{3}c\Delta} - \sqrt{2\nu/T}}{\frac{2}{3}c} &\ge \frac{4\Delta\sqrt{2\nu/T}}{8\nu/T + \frac{4}{3}c\Delta} && \text{as } \sqrt{x+a} - \sqrt{x} \ge \frac{2a\sqrt{x}}{4x+a} \\
&= \frac{3\Delta\sqrt{2\nu/T}}{6\nu/T + c\Delta} \\
&= \frac{3\Delta\sqrt{2(\gamma\theta - \Delta^2)}}{6\gamma\theta - 5\Delta^2 + \gamma\Delta} \\
&= \frac{3\sqrt{2}}{\sqrt{\gamma}} \frac{\Delta\sqrt{\theta - \Delta^2/\gamma}}{6\theta + \Delta - 5\Delta^2/\gamma} \\
&\ge \frac{3}{\sqrt{\gamma}} \frac{\Delta\sqrt{\theta}}{7\theta} && \text{as } \theta \ge \Delta \text{ and } \Delta^2/\gamma \le \theta/2.
\end{aligned}$$

Hence, we conclude that setting

$$\beta = 0.18\sqrt{\log k}\frac{1}{\sqrt{2T}}\frac{\Delta^2}{\theta}$$

gives the desired property. The proof follows by plugging in the value of β and taking a union bound over the experts. □

A.6 Proof of Theorem 4

The proof is very much similar to Theorem 6. The difference is that the conditional variance is bounded above by

$$\mathbb{E}\{\xi_t^2 \mid \mathcal{F}_{t-1}\} \le \eta_t^{-1} - \Delta^2,$$

and the rest of the proof follows by setting $\theta = 1$.

A.7 Proof of Theorem 3

The first bound is standard and can be found in (Slivkins, 2019). The argument is completed by noting that the expected accuracy gap will be bounded by

$$\frac{1}{T}\,\mathbb{E}[\boldsymbol{L}_{T,I_\tau} - \boldsymbol{L}_{T,i^*}] \le \sqrt{\frac{8\log k}{T}},$$

and setting the right hand side less than ε.

For the second part, we use Theorem 4. With probability at most $k \cdot e^{-0.18\Delta^2\sqrt{T\log k}}$ the recommended expert is not the best, for which its accuracy gap is at most 1, and otherwise, the best expert is returned, with accuracy gap 0. Combining the two gives the result.

A.8 Proof of Theorem 7

The proof is very similar to Theorem 3, with the difference that here one upper bounds the accuracy gap by $\max_i \Delta_i$ instead of 1.

A.9 Proof of Theorem 8

First we prove a lemma that help us proving the theorem:

Lemma 8. *The expected number of times that the recommendation π_t is not the best model is a constant up to any round and is bounded by $\frac{62k}{\lambda^2 \log k}$.*

Proof. By Theorem 6, we know that the probability of not recommending the best model at round t is upper bounded by $k \cdot e^{-0.18\lambda\sqrt{T\log k}}$. Using integral approximation, one finds that $\sum_{t=1}^{\infty} e^{-a\sqrt{t}} \le 2/a^2$ for all $a > 0$. This gives

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{I}_{\{\pi_t \neq i^*\}}\right] = \sum_{t=1}^{\infty} \Pr\{\pi_t \neq i^*\} \le \sum_{t=1}^{\infty} k \cdot e^{-0.18\lambda\sqrt{T\log k}} \le \frac{62k}{\lambda^2 \log k}. \qquad \square$$

Using the lemma, over T rounds, we make at most $\frac{62k}{\lambda^2 \log k}$ mistakes, for which we get at most $\max_i \Delta_i$ added to the regret, and in other rounds, we make no mistakes, hence no regrets on those rounds. Adding up gives the result.

B Example for Large Number of Updates

Consider a binary classification scenario with two models. Set the loss sequence to be $(1,0),(0,1),(1,0),\ldots$, that is, on the odd rounds the second model is correct and on the even rounds, the first one. One can see that the probability of querying the label is always $\boldsymbol{w}_{t,1}\boldsymbol{w}_{t,2} \vee \eta_t$ for all t. Hence, this probability is always near $1/4$, as the models weights are always around $1/2$. Hence, the total number of queries is linear.

C Experiments

C.1 Details on the Model Collections

- CIFAR-10: As an image classification dataset, we train 80 models on CIFAR-10 dataset varying in machine learning models (ranging from DenseNet, Resnet to VGG), architecture and parameter setting. The ensemble of models have accuracies between 55-92% on a test set consists of 10 000 instances.
- ImageNet: This dataset consists of 102 image classification models (ranging from ResNet, Inception to MobileNet) pre-trained on ImageNet that are available on TensorFlow Hub. The accuracy of models occupy the range in 50-80%. For each model, we obtain the ImageNet validation dataset with 50 000 data examples, and furthermore normalize and resize them according to expected input format for each model, and finally conduct inference on the given model to produce predicted labels.

Table 1: Datasets characteristics

Dataset	#Classes	#Instances	#Models	Accuracy of Models
CIFAR-10	10	10 000	80	55-92%
ImageNet	1 000	50 000	102	50-80%
Drift	6	3 000	9	25-60%
EmoContext	4	5 509	8	88-92%
CIFAR-10 (worse models)	10	10 000	80	40-70%

- Drift: For the Drift dataset, we trained models on the gas sensor drift data that is collected over a course of three years. The dataset has ten batches, each collected in different months. We trained an SVM classifier on each of the batch but the last one, and use the last batch of size 3 000 as test set. Although each model has good training accuracy on the batch it is trained on, namely above 90%, their accuracy on the test set lies in 25-60%. This is due to the drift behaviour of sensor data among different time intervals.

- EmoContext: This dataset consist of pretrained models that are the development history of a participant on EmoContexttask in SemEval 2019. The task aims to detect emotions from text leveraging contextual information which is deemed challenging due to the lack of facial expressions and voice modulations. We treat each development as an individual pretrained model where development stages differ in various word representations including ELMo Peters et al. (2018) and GloVe Pennington et al. (2014). The dataset consists of 8 pre-trained models whose accuracy varies between 88-92% on the test set of size 5 509.

- CIFAR-10 V2: We also train a set of models with relatively lower accuracies on CIFAR-10 and call it CIFAR-10 V2. The sole purpose behind creating such a collection is to investigate the performance of Model Picker on a practical scenario like this. Using similar model architectures to that of CIFAR-10, the pretrained models have accuracy between 40-70% on a test set of size 10 000.

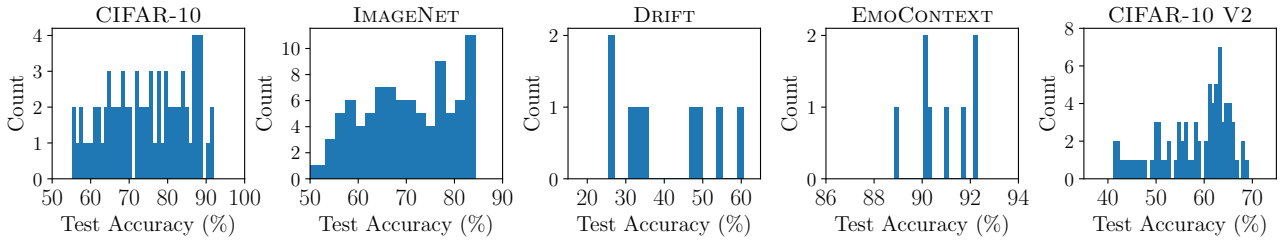


Figure 3: Counts of model accuracies

The properties of model collections for all datasets are depicted in Table 1 on page 19 and Figure 3.

C.2 Note on the Baselines

In this section, we provide further details on some of the adapted baseline methods, namely, query by committee (Entropy), importance weighted active learning (Importance) and efficient active learning (Efal).

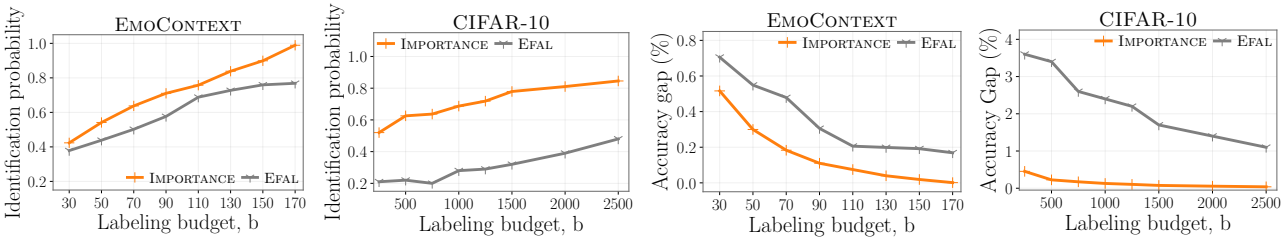
- *Query by Committee*: As indicated in Section 5, we adapt the query-by-committee paradigm proposed in Dagan and Engelson (1995) for model selection in the online setting. The query by committee method consist of two sub-strategies (a) ensemble learning, and (b) determining a maximal disagreement measure. The ensemble learning indicates how the committee is formed from the candidate classifiers. This step is crucial to make the disagreement measure more reliable while aiming to form a set of classifiers with high accuracy. In literature, there exist many ensemble learning methods including Abe and Mamitsuka (1998); Melville and Mooney (2004); Breiman (1996); Freund and Schapire (1995). Most, if not all, of these methods are either designed for pool-based sampling or for cases where observed data is stored. Bagging predictors Breiman (1996) proposes to improve performance of a single predictor by forming a committee from multiple versions of it, where the versions are trained on the bootstrap replicates of training data. This is followed by Abe and Mamitsuka (1998) where diverse ensembles are generated using bagging and boosting techniques previously introduced by Freund and Schapire (1995). These strategies focus on a setting where the observed data is stored as opposed to our setting. Another popular ensemble learning algorithm, Active-Decorate relies on the existence of artificial

training data to form a diverse set of examples. In our setting, however, we assume neither storing of previously seen data nor availability of artificial data. In the online setting, however, one could benefit from the strategy introduced in Freund and Schapire (1995). Upon seeing the label c_t, the authors propose to update the belief on the models such that $\boldsymbol{w}_t \propto \boldsymbol{w}_{t-1}\beta^{\boldsymbol{\ell}_t}$. We note that, this update rule very closely resembles that of the structural query by committee, which we include in our numerical analysis. In fact, it is identical when both of β are tuned to query budget b amount of label in average over many realizations.

As a disagreement measure, popular choices include vote margin, vote entropy and KL divergence between the label distributions of each committee member and the consensus in Settles and Craven (2008). We first note that the latter two are equivalent for 0-1 loss functions ℓ. The former, vote margin is measured by the difference between the votes of most voted and second most voted label. We omit this in our analysis motivated by the preliminary observation on the success of entropy over the vote margin.

- *Importance Weighted Active Learning*: As indicated earlier, we implement the importance weighted active learning algorithm, introduced by Beygelzimer et al. (2008). Formally, upon seeing a new instance x_t, the algorithm computes a rejection threshold θ_t using sample complexity bounds, and update the hypothesis space $\mathcal{H}_t$ to contain only the models whose weighted error is θ_t greater than weighted error of the current best model at time t. The sampling probability q_t is set to $\max_{i,j\in\mathcal{H}_t, c\in[C]} \boldsymbol{\ell}_{t,i}^{(c)} - \boldsymbol{\ell}_{t,j}^{(c)}$. We use 0-1 loss. Therefore, adaptation in our setting becomes making query decision based merely on the disagreement between the *surviving* hypotheses at time t. That is, we query the label c_t if and only if the surviving classifiers at time t disagree on the labeling of x_t.

- *Efficient Active Learning*: We adapt the efficient active learning algorithm presented by Beygelzimer et al. (2010, 2011). In a manner similar to the importance weighted approach, the efficient active learning algorithm also uses the importance weighted framework. Upon receiving a new instance x_t, the algorithm measures the weighted error estimate between two competing models, and specifies a sampling probability based on a threshold that is a function of $\frac{C_0 \log t}{t-1}$ for some parameter $C_0 > 0$. If the gap between the estimated weighted errors of two competing models are below this threshold, then the label c_t is queried. Otherwise, the algorithm computes the sampling probability q_t that is roughly $\min\left\{1, \mathcal{O}(1/G_k^2 + 1/G_k)\frac{C_0 \log k}{k-1}\right\}$ where $G_k = \min_{i\in[k]} \boldsymbol{L}_{t,i} - \min_{j\in[k], j\neq i} \boldsymbol{L}_{t,j}$. We refer to Algorithm 1 of Beygelzimer et al. (2010) for further details. In our implementation, we consider the threshold parameter C_0 as hyperparameter and tune for efficient active learning algorithm to request amount of labels not exceeding the labelling budget b. However, as indicated in Figure 4, it underperforms the importance weighted active learning algorithm. However, it is crucial to emphasize again that these methods are meant to improve supervised training of classifiers instead of ranking of pretrained models. We include them in our comparison for the completeness.



(a) Identification probability (b) Accuracy gap

Figure 4: Comparison of importance weighted methods {Importance, Efal} and on the EmoContext and CIFAR-10 datasets

C.3 Performance of Model Picker on models with low accuracies

As mentioned in Section 5, we conduct another numerical analysis on the performance of Model Picker when pretrained models have relatively lower accuracies. Towards that, we train 80 models on CIFAR-10 varying in machine learning models and parameters. The accuracy of pretrained models line in 40-70% over a test set of of size 10 000. We compare the model selection methods over this new model collection by following the exact same procedure as in the Section 5. We use a stream size of 5 000 and average the results over 500 realizations. Figure 5 summarize the comparison. When the accuracy of pre-trained models are low, the query by committee

algorithm expectedly underperforms as the disagreement measure becomes noisy under the existence of models with low accuracies. Model Picker, on the other hand, noticeably outperforms in returning the true best model as well as the ranking of the models (Figure 5). The regret analysis in Figure 5 suggests that the structural query by committee method maintains a low regret throughout the streaming process as well as for different labeling budgets, and very closely followed by Model Picker.

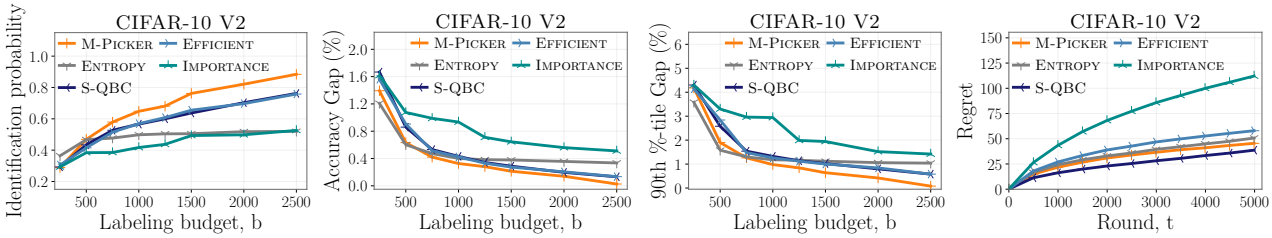


Figure 5: Performance evaluation of model selection methods on CIFAR-10 V2 dataset that consist of pre-trained models with low accuracies.

C.4 Hyperparameters

The hyperparameter tuning is performed via grid search. For each grid point, we run the experiment for 100 realizations and compute the average number of requests. The grid search was performed over the following search space:

- CIFAR-10: Model Picker: `[0, 3000]`, Entropy: `[0, 20]`, S-QBC: `[0, 10]`, Importance: `[0, 0.9]`, Efal: `[0, 1.5e-2]`
- ImageNet: Model Picker: `[0, 135]`, Entropy: `[0, 22]`, S-QBC: `[0, 20]`, Importance: `[0, 1]`
- Drift: Model Picker: `[0, 60]`, Entropy: `[0, 4]`, S-QBC: `[0, 4]`, Importance: `[0, 05]`
- EmoContext: Model Picker: `[0, 60]`, Entropy: `[0, 4]`, S-QBC: `[0, 4]`, Importance: `[0, 05]`, Efal: `[0, 1e-2]`
- CIFAR-10 V2: Model Picker: `[0, 1000]`, Entropy: `[0, 3]`, S-QBC: `[0, 10]`, Importance: `[0, 0.9]`, Efal: `[0, 1e-1]`

with grid size of 250 where grid points are equally spaced. The respective number of requests for each grid point can be found in our publicly available repository[8].

Remark that the amount of requests by Model Picker saturates when Model Picker reaches at a high identification probability. Therefore, the update probability is upscaled with a very high value such that Model Picker queries large number of labels, and thus comparison to other methods for large budget constraints are made possible. Practically, this would not be required as Model Picker itself decides when to stop requesting labels. For example, when the update probability is upscaled by a factor of 11 for CIFAR-10 V2 dataset, the number of requests made by Model Picker is 3 800 labels, whereas an upscaling of 835 is used to enable Model Picker requests nearly 4 800 labels.

[8] https://github.com/DS3Lab/online-active-model-selection