Active Online Learning with Hidden Shifting Domains: Supplementary Materials

A Proofs for upper bounds

A.1 Proof of Theorem 1

We provide the proof of Theorem 1 in this section. We focus on regret and query complexity bounds on one domain I_u, and sum over domain u to obtain Theorem 1. Recall that we define the interaction history between the learner and the environment up to time t be $H_t = \{x_{1:t}, f_{1:t}, \xi_{1:t}\}$; we abbreviate $\mathbb{E}[\cdot|x_t, H_{t-1}]$ as $\mathbb{E}_{t-1}[\cdot]$.

The following lemma upper bounds the regret with sum of uncertainty estimates, $\Delta_t = \tilde{\eta}^2 \min\left(1, \|x_t\|^2_{M_t^{-1}}\right)$. A similar lemma has appeared in Cesa-Bianchi et al. (2009, Lemma 1).

Lemma 1. *In the setting of Theorem 1, with probability* $1 - \frac{\delta}{2}$*, for all* $t \in [T]$*,* $(\hat{y}_t - \langle \theta^*, x_t \rangle)^2 = \tilde{O}(\Delta_t)$.

Proof of Lemma 1. Denote the value of $M, \mathcal{Q}$ at the beginning of round t as $M_t, \mathcal{Q}_t$. Let $\lambda = 1/C^2$, $V_t = M_t - \lambda I = \sum_{s\in\mathcal{Q}_t} x_s x_s^\top$. Therefore, $\hat{\theta}_t = M_t^{-1}(\sum_{s\in\mathcal{Q}_t} x_s y_s) = M_t^{-1}(V_t\theta^* + \sum_{s\in\mathcal{Q}_t} \xi_s x_s)$, and

$$\langle x_t, \hat{\theta}_t - \theta^* \rangle = \sum_{s\in\mathcal{Q}_t} \xi_s (x_t^\top M_t^{-1} x_s) - \lambda x_t^\top M_t^{-1} \theta^*. \tag{2}$$

The first term is a sum over a set of independent sub-Gaussian random variables, so it is $(\eta\sigma)^2$-sub-Gaussian with $\sigma^2 = \sum_{s\in\mathcal{Q}_t} x_t^\top M_t^{-1} x_s x_s^\top M_t^{-1} x_t \le x_t^\top M_t^{-1} x_t$. Define event

$$E_t = \left\{ \left| \sum_{s\in\mathcal{Q}_t} \xi_s (x_t^\top M_t^{-1} x_s) \right| \le \eta \sqrt{2\ln\frac{4T}{\delta}} \|x_t\|_{M_t^{-1}} \right\}.$$

By standard concentration of subgaussian random variables, we have $\mathbb{P}(E_t) \ge 1 - \frac{\delta}{2T}$. Define $E = \cap_{t=1}^T E_t$. By union bound, we have $\mathbb{P}(E) \ge 1 - \frac{\delta}{2}$. We henceforth condition on E happening, in which case the first term of Equation (2) is bounded by $\eta\sqrt{2\ln(4T/\delta)}\|x_t\|_{M_t^{-1}}$ at every time step t.

Meanwhile, the second term of Equation (2) can be bounded by Cauchy-Schwarz:

$$\left|\lambda x_t^\top M_t^{-1}\theta^*\right| = \lambda \left|\langle M_t^{-1/2} x_t, M_t^{-1/2}\theta^* \rangle\right| \le \lambda \|x_t\|_{M_t^{-1}} \|\theta^*\|_{M_t^{-1}} \le \sqrt{\lambda}\|\theta^*\|_2 \|x_t\|_{M_t^{-1}},$$

which is at most $\|x_t\|_{M_t^{-1}}$, since $\|\theta^*\|_2 \le C$ and $\lambda = 1/C^2$. Using the basic fact that $(A+B)^2 \le 2A^2 + 2B^2$,

$$(\langle x_t, \hat{\theta}_t \rangle - \langle x_t, \theta^* \rangle)^2 \le (4\eta^2 \ln(4T/\delta) + 2)\|x_t\|^2_{M_t^{-1}}.$$

Since $\hat{y}_t = \mathrm{clip}(\langle x_t, \hat{\theta}_t \rangle) \in [-1, 1]$ and $\left|\langle x_t, \theta^* \rangle\right| \le 1$, we also trivially have $(\hat{y}_t - \langle \theta^*, x_t \rangle)^2 \le 4$. Therefore,

$$\begin{aligned}
(\hat{y}_t - \langle \theta^*, x_t \rangle)^2 &\le \min\left(4, (4\eta^2 \ln(4T/\delta') + 2)\|x_t\|^2_{M_t^{-1}}\right) \\
&\le (4\eta^2 \ln(2T/\delta') + 4) \cdot \min\left(1, \|x_t\|^2_{M_t^{-1}}\right) \\
&\le \tilde{O}\left(\tilde{\eta}^2 \min\left(1, \|x_t\|^2_{M_t^{-1}}\right)\right) = \tilde{O}(\Delta_t).
\end{aligned}$$

□

The following lemma bounds the sum of uncertainty estimates for k *queried* examples in a domain:

Lemma 2. *Let $a_1, \ldots, a_k$ be k vectors in $\mathbb{R}^d$. For $i \in [k]$, define $N_i = \lambda I + \sum_{j=1}^{i-1} a_j a_j^\top$. Then, for any $S \subseteq [k]$, $\sum_{i\in S} \min\left(1, \|a_i\|^2_{N_i^{-1}}\right) \le \ln(\det(\lambda I + \sum_{i\in S} a_i a_i^\top)/\det(\lambda I))$.*

Proof of Lemma 2. We denote by $N_{i,S} = \lambda I + \sum_{j\in S: j\le i-1} a_j a_j^\top$. As S is a subset of $[k]$, we have that $N_{i,S} \preceq N_i$. Consequently, $\|a_i\|_{N_i^{-1}} \le \|a_i\|_{N_{i,S}^{-1}}$. Therefore,

$$\sum_{i\in S} \min\left(1, \|a_i\|^2_{N_i^{-1}}\right) \le \sum_{i\in S} \min\left(1, \|a_i\|^2_{N_{i,S}^{-1}}\right) \le \ln\left(\frac{\det(\lambda I + \sum_{i\in S} a_i a_i^\top)}{\det(\lambda I)}\right),$$

where the second inequality is well-known (see e.g. Lattimore and Szepesvári, 2018, Lemma 19.4). □

Proof of Theorem 1. Let $p_t = \min(1, \alpha\Delta_t)$ be the learner's query probability at time t; it is easy to see that $\mathbb{E}_{t-1}[q_t] = p_t$. Let random variable $Z_t = q_t\Delta_t$. We have the following simple facts:

1. $Z_t \le \tilde{\eta}^2$,
2. $\mathbb{E}_{t-1} Z_t = p_t\Delta_t$,
3. $\mathbb{E}_{t-1} Z_t^2 \le \tilde{\eta}^2 \cdot \mathbb{E}_{t-1} Z_t \le \tilde{\eta}^2 p_t \Delta_t$.

For every $u \in [m]$, define event

$$F_u = \left\{ \left| \sum_{t\in I_u} p_t\Delta_t - \sum_{t\in I_u} q_t\Delta_t \right| \le O\left(\tilde{\eta}\sqrt{\sum_{t\in I_u} p_t\Delta_t \ln\frac{T}{\delta}} + \tilde{\eta}^2 \ln\frac{T}{\delta}\right) \right\}. \tag{3}$$

Applying Freedman's inequality to $\{Z_t\}_{t\in I_u}$ (see e.g. Bartlett et al., 2008, Lemma 2), we have that $\mathbb{P}(F_u) \ge 1 - \frac{\delta}{4m}$.

Similarly, define

$$G = \left\{ \left| \sum_{t=1}^{T} p_t - \sum_{t=1}^{T} q_t \right| \le O\left(\sqrt{\sum_{t=1}^{T} p_t \ln\frac{T}{\delta}} + \ln\frac{T}{\delta}\right) \right\}. \tag{4}$$

Applying Freedman's inequality to $\{q_t\}_{t\in I_u}$, we have that $\mathbb{P}(G) \ge 1 - \frac{\delta}{4}$.

Furthermore, define $H = E \cap (\cap_{u=1}^m F_u) \cap G$, where E is the event defined in the proof of Lemma 1. By union bound, $\mathbb{P}(H) \ge 1 - \delta$. We henceforth condition on H happening.

By the definition of F_u, Solving for $\sum_{t\in I_u} p_t\Delta_t$ in Equation (3), we get that

$$\sum_{t\in I_u} p_t\Delta_t = \tilde{O}\left(\sum_{t\in I_u} q_t\Delta_t + \tilde{\eta}^2\right). \tag{5}$$

Using Lemma 2 with $\{a_i\}_{i=1}^k = \{x_t\}_{t\in\mathcal{Q}_T}$, and $S = I_u \cap \mathcal{Q}_T$, we get that

$$\begin{aligned}\sum_{t\in I_u} q_t\Delta_t &\le \tilde{\eta}^2 \cdot \ln\det\left(I + C^2 \sum_{t\in I_u\cap\mathcal{Q}_T} x_t x_t^\top\right) \\ &\le 2\tilde{\eta}^2 d_u \ln\left(1 + C^2 T_u/d_u\right) = \tilde{O}(\tilde{\eta}^2 d_u).\end{aligned}$$

In combination with Equation (5), we have $\sum_{t\in I_u} p_t\Delta_t = \tilde{O}(\tilde{\eta}^2 d_u)$.

We divide the examples in domain u into high and low risk subsets with index sets $I_{u,+}$ and $I_{u,-}$ (abbrev. I_+ and I_- hereafter). Formally,

$$I_+ = \{t \in I_u : \alpha\Delta_t > 1\}, \quad I_- = I - I_+.$$

We consider bounding the regrets and the query complexities in these two sets respectively:

1. For every t in I_+, as $p_t = 1$, label y_t is queried, so

$$\sum_{t\in I_+} \Delta_t = \sum_{t\in I_+} q_t\Delta_t \le \sum_{t\in I_u} q_t\Delta_t = \tilde{O}(\tilde{\eta}^2 d_u).$$

   Since for every t in I_-, $\Delta_t > 1/\alpha$, we have $\sum_{t\in I_+}\Delta_t > |I_+|/\alpha$. This implies that $\sum_{t\in I_+} p_t = |I_+| = \tilde{O}(\alpha\tilde{\eta}^2 d_u)$.

2. For every t in I_-, $p_t = \alpha\Delta_t$. Therefore, $\sum_{t\in I_-}\alpha\Delta_t^2 = \sum_{t\in I_-} p_t\Delta_t \le \sum_{t\in I_u} p_t\Delta_t = \tilde{O}(\tilde{\eta}^2 d_u)$. By Cauchy-Schwarz, and the fact that $|I_-| \le T_u$, we get that $\sum_{t\in I_-}\Delta_t \le \sqrt{|I_-|\cdot(\sum_{t\in I_-}\Delta_t^2)} = \tilde{O}(\tilde{\eta}\sqrt{d_u T_u/\alpha})$.

   Consequently, $\sum_{t\in I_-} p_t = \sum_{t\in I_-}\alpha\Delta_t \le \tilde{O}(\tilde{\eta}\sqrt{\alpha d_u T_u})$.

Summing over the two cases, we have

$$\sum_{t\in I_u} p_t \le \tilde{O}\left(\alpha\tilde{\eta}^2 d_u + \tilde{\eta}\sqrt{\alpha d_u T_u}\right), \quad \sum_{t\in I_u}\Delta_t \le \tilde{O}\left(\tilde{\eta}^2 d_u + \tilde{\eta}\sqrt{d_u T_u/\alpha}\right),$$

If $\alpha \le \frac{1}{\tilde{\eta}^2}\frac{T_u}{d_u}$, we have $\alpha\tilde{\eta}^2 d_u \le \tilde{\eta}\sqrt{\alpha d_u T_u}$, otherwise we use the trivial bound $\sum_{t\in I_u} p_t \le T_u$. Therefore, the above bounds can be simplified to

$$\sum_{t\in I_u} p_t \le \tilde{O}\left(\min\{T_u, \tilde{\eta}\sqrt{\alpha d_u T_u}\}\right), \quad \sum_{t\in I_u}\Delta_t \le \tilde{O}\left(\max\{\tilde{\eta}^2 d_u, \tilde{\eta}\sqrt{d_u T_u/\alpha}\}\right). \tag{6}$$

For the query complexity, from the definition of event G, applying AM-GM inequality on Equation (4), we also have

$$Q = \sum_{t=1}^{T} q_t = \tilde{O}\left(\sum_{t=1}^{T} p_t + 1\right) = \tilde{O}\left(\sum_{u=1}^{m}\min\{T_u, \tilde{\eta}\sqrt{\alpha d_u T_u}\} + 1\right).$$

For the regret guarantee, we have by the definition of event E and Lemma 1 that

$$\sum_{t=1}^{T}(\hat{y}_t - \langle\theta^*, x_t\rangle)^2 \le \tilde{O}\left(\sum_{t=1}^{T}\Delta_t^2\right) = \tilde{O}\left(\sum_{u=1}^{m}\left(\sum_{t\in I_u}\Delta_t^2\right)\right).$$

Using the second inequality of Equation (6), we get

$$\sum_{t=1}^{T}(\hat{y}_t - \langle\theta^*, x_t\rangle)^2 \le \tilde{O}\left(\sum_{u=1}^{m}\max\{\tilde{\eta}^2 d_u, \tilde{\eta}\sqrt{d_u T_u/\alpha}\}\right).$$

The theorem follows. □

A.2 Proof of Theorem 2

Before going into the proof, we set up some useful notations. Define $I = \{0, 1, \ldots, k\}$ as the index set of the α_i's of interest. Recall the number of copies $k = 1 + \lceil 3\log T\rceil \le 2 + 3\log T$. Recall also that $B' = \lfloor B/k\rfloor$ is the label budget for each copy.

Let $p_t^i = \min(1, \alpha_i\Delta_t)$ be the intended query probability of copy i at time step t; let $r_t^i \sim \text{Bernoulli}(p_t^i)$ be the attempted query decision of copy i at time step t; let $A_t^i = \mathbb{1}\left[\sum_{j=1}^{t-1} r_j^i < B'\right]$, i.e. the indicator that copy i has not reached its budget limit at time step t. Using this notation, the actual query decision of copy i, q_t^i, can be written as $r_t^i A_t^i$.

We have the following useful observation that gives a sufficient condition for copy i to be within its label budget:

Lemma 3. *Given $i \in [k]$, if $\sum_{t=1}^{T} A_t^i r_t^i < B'$, the following items hold:*

1. $\sum_{t=1}^{T} r_t^i < B'$.
2. *For all $t \in [T]$, $A_t^i = 1$, i.e. copy i does not run of label budget throughout.*

Proof. Suppose for the sake of contradiction that $\sum_{t=1}^{T} r_t^i \geq B'$. Consider the first B' occurrences of $r_j^i = 1$; call them $J = \{j_1, \ldots, j_{B'}\}$. It can be seen that for all $j \in J$, $A_j^i = 1$. Therefore,

$$\sum_{t=1}^{T} A_t^i r_t^i \geq \sum_{j \in J} A_j^i r_j^i \geq |J| = B',$$

which contradicts with the premise that $\sum_{t=1}^{T} A_t^i r_t^i < B'$.

The second item immediately follows from the first item, as $\sum_{j=1}^{T} r_j^i < B'$ implies that $\sum_{j=1}^{t-1} r_j^i < B'$ for every $t \in [T]$. □

Complementary to the above lemma, we can also see that for every $i \in [k]$, $\sum_{t=1}^{T} A_t^i r_t^i = \sum_{t=1}^{T} q_t^i \leq B'$ is trivially true. We next give a key lemma that generalizes Theorem 1, and upper bounds $\sum_{t=1}^{T} A_t^i r_t^i$ for all i's beyond the above trivial B' bound.

Lemma 4. *There exists $C = \mathrm{polylog}(T, \frac{1}{\delta}) \geq 1$, such that with probability $1 - \delta/2$,*

$$\sum_{t=1}^{T} A_t^i \Delta_t \leq C \cdot \tilde{\eta} \sum_u \sqrt{d_u T_u}/\sqrt{\alpha_i}, \text{ and } \sum_{t=1}^{T} A_t^i r_t^i \leq C \cdot \tilde{\eta}\sqrt{\alpha_i} \sum_u \sqrt{d_u T_u},$$

for every $i \in I$ such that $\alpha_i \in \left[\frac{1}{\tilde{\eta}^2}\left(\frac{1}{\sum_u \sqrt{d_u T_u}}\right)^2, \frac{1}{\tilde{\eta}^2} \min_{u \in [m]} \frac{T_u}{d_u}\right]$.

Proof. Applying Freedman's inequality to the martingale difference sequence $\{A_t^i(r_t^i - p_t^i)\}_{t=1}^{T}$, we get that with probability $1 - \delta/4$,

$$\sum_{t=1}^{T} A_t^i r_t^i = \tilde{O}\left(\sum_{t=1}^{T} A_t^i p_t^i + 1\right). \tag{7}$$

Applying Freedman's inequality to $\{A_t^i(r_t^i - p_t^i)\Delta_t \mathbb{1}[t \in I_u]\}_{t=1}^{T}$, and take a union bound over all $u \in [m]$, we get that with probability $1 - \delta/4$,

$$\sum_{t \in I_u} A_t^i p_t^i \Delta_t = \tilde{O}\left(\sum_{t \in I_u} A_t^i r_t^i \Delta_t + \tilde{\eta}^2\right).$$

Using Lemma 2 with $\{a_i\}_{i=1}^{k} = \{x_t\}_{t \in \mathcal{Q}_T}$, and $S = I_u \cap \mathcal{Q}_T$ we get that, deterministically, $\sum_{t \in I_u} A_t^i r_t^i \Delta_t \leq \sum_{t \in I_u} q_t \Delta_t = \tilde{O}(\tilde{\eta}^2 d_u)$. So with probability $1 - \delta/4$,

$$\sum_{t \in I_u} A_t^i p_t^i \Delta_t = \tilde{O}(\tilde{\eta}^2 d_I). \tag{8}$$

We henceforth condition on Equations (7) and (8) occuring, which happens with probability $1 - \delta/2$ by union bound. Let $I_+ = \{t \in I_u : \alpha_i \Delta_j > 1\}$, and $I_- = I_u - I_+$.

1. For I_+, by Equation (8), $\sum_{t \in I_+} A_j^i \Delta_j = \tilde{O}(\tilde{\eta}^2 d_u) \implies \sum_{j \in I_+} A_j^i p_j^i = \tilde{O}(\alpha_i \tilde{\eta}^2 d_u)$.
2. For I_-, by Equation (8), $\sum_{j \in I_-} A_j^i \alpha_i \Delta_j^2 = \sum_{j \in I_-} A_j^i p_j \Delta_j = \tilde{O}(\tilde{\eta}^2 d_u)$; this implies that $\sum_{j \in I_-} A_j^i \Delta_j = \tilde{O}(\tilde{\eta}\sqrt{d_u T_u/\alpha_i})$. In this event, we also have $\sum_{j \in I_-} A_j^i p_j^i = \sum_{j \in I_-} A_j^i \alpha_i \Delta_j = \tilde{O}(\tilde{\eta}\sqrt{d_u T_u \alpha_i})$.

Summing over the two cases, we have

$$\sum_{t\in I_u} A_t^i p_t^i \le \tilde{O}(\alpha_i\tilde{\eta}^2 d_u + \tilde{\eta}\sqrt{\alpha_i d_u T_u}), \quad \sum_{t\in I_u} A_t^i \Delta_t \le \tilde{O}(\tilde{\eta}^2 d_u + \tilde{\eta}\sqrt{d_u T_u/\alpha_i}),$$

By the assumption that $\alpha_i \le \frac{1}{\tilde{\eta}^2}\min_u \frac{T_u}{d_u}$, for every u, we have, $\alpha_i\tilde{\eta}^2 d_u \le \tilde{\eta}\sqrt{\alpha_i d_u T_u}$. This implies that

$$\sum_{t\in I_u} A_t^i p_t^i \le \tilde{O}(\tilde{\eta}\sqrt{\alpha_i d_u T_u}), \quad \sum_{t\in I_u} A_t^i \Delta_t \le \tilde{O}(\tilde{\eta}\sqrt{d_u T_u/\alpha_i}). \tag{9}$$

Summing over $u \in [m]$, we get

$$\sum_{t=1}^{T} A_t^i p_t^i \le \tilde{O}(\tilde{\eta}\sum_{u=1}^{m}\sqrt{\alpha_i d_u T_u}), \quad \sum_{t=1}^{T} A_t^i \Delta_t \le \tilde{O}(\tilde{\eta}\sum_{u=1}^{m}\sqrt{d_u T_u/\alpha_i}).$$

Therefore, using Equation (7), we have

$$\sum_{t=1}^{T} A_t^i r_t^i \le \tilde{O}\left(\sum_{t=1}^{T} A_t^i p_t^i + 1\right) \le \tilde{O}\left(\tilde{\eta}\sum_{u=1}^{m}\sqrt{\alpha_i d_u T_u} + 1\right) \le \tilde{O}\left(\tilde{\eta}\sum_{u=1}^{m}\sqrt{\alpha_i d_u T_u}\right),$$

where the last inequality uses the assumption that $\alpha_i \ge \frac{1}{\tilde{\eta}^2}\left(\frac{1}{\sum_u \sqrt{d_u T_u}}\right)^2$. The lemma follows. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. First, the query complexity of Fixed-Budget QuFUR is B by construction, as the algorithm maintains k copies of QuFUR, and each copy consumes at most $B' = \lfloor B/k \rfloor$ labels.

We now bound the regret of Fixed-Budget QuFUR. We consider $\overline{B} = Ck(\sum_u \sqrt{d_u T_u}) \cdot \min_{u\in[m]} \sqrt{T_u/d_u} = \tilde{O}\left((\sum_u \sqrt{d_u T_u}) \cdot \min_{u\in[m]} \sqrt{T_u/d_u}\right)$, where $C = \mathrm{polylog}(T, \frac{1}{\delta}) \ge 1$ is defined in Lemma 4. We will show that if $B \in (0, \overline{B}]$, with probability $1-\delta$, the regret of Fixed-Budget QuFUR is at most $\tilde{O}\left(\frac{\tilde{\eta}^2(\sum_u \sqrt{d_u T_u})^2}{B}\right)$.

If $B < 2C\tilde{\eta}^2 k$, the regret of the algorithm is trivially upper bounded by $4T$, which is clearly $\tilde{O}\left(\frac{\tilde{\eta}^2(\sum_u \sqrt{d_u T_u})^2}{B}\right)$. Therefore, throughout the rest of the proof, we consider $B \in [2C\tilde{\eta}^2 k, \overline{B}]$.

Recall that $I = \left\{\frac{2^i}{T^2} : i \in \{0, 1, \ldots, k\}\right\}$. We denote by $\alpha_{\min} = \frac{1}{T^2}$ the minimum element of I, and $\alpha_{\max} = \frac{2^k}{T^2} \ge T$ the maximum element of I.

Denote by

$$i_B = \max\left\{i \in I : C\tilde{\eta}\sqrt{\alpha_i}\sum_{u=1}^{m}\sqrt{d_u T_u} < B'\right\} = \max\left\{i \in I : \alpha_i < \left(\frac{B'}{C\tilde{\eta}\sum_u \sqrt{d_u T_u}}\right)^2\right\}.$$

As $B \in [2C\tilde{\eta}^2 k, \overline{B}]$, we have $\left(\frac{B'}{C\tilde{\eta}\sum_u \sqrt{d_u T_u}}\right)^2 \in (\alpha_{\min}, \alpha_{\max}]$. Indeed, $\left(\frac{B'}{C\tilde{\eta}\sum_u \sqrt{d_u T_u}}\right)^2 \le \left(\frac{\overline{B}}{Ck\tilde{\eta}\sum_u \sqrt{d_u T_u}}\right)^2 \le T \le \alpha_{\max}$, $\left(\frac{B'}{C\tilde{\eta}\sum_u \sqrt{d_u T_u}}\right)^2 \ge \left(\frac{\tilde{\eta}}{\sum_u \sqrt{d_u T_u}}\right)^2 > \alpha_{\min}$, as $\sum_u \sqrt{d_u T_u} \le \sum_u T_u = T$.

Therefore, by the definition of i_B, we have

$$\alpha_{i_B} \in \left[\frac{1}{2}\left(\frac{B'}{C\tilde{\eta}\sum_u \sqrt{d_u T_u}}\right)^2, \left(\frac{B'}{C\tilde{\eta}\sum_u \sqrt{d_u T_u}}\right)^2\right) \tag{10}$$

Again by our assumption on B, $\frac{1}{2}\left(\frac{B'}{C\tilde{\eta}\sum_u\sqrt{d_uT_u}}\right)^2 \geq \tilde{\eta}^2\left(\frac{1}{\sum_u\sqrt{d_uT_u}}\right)^2 \geq \frac{1}{\tilde{\eta}^2}\left(\frac{1}{\sum_u\sqrt{d_uT_u}}\right)^2$, $\left(\frac{B'}{C\tilde{\eta}\sum_u\sqrt{d_uT_u}}\right)^2 \leq \left(\frac{\overline{B}}{Ck\tilde{\eta}\sum_u\sqrt{d_uT_u}}\right)^2 \leq \frac{1}{\tilde{\eta}^2}\min_{u\in[m]}\frac{T_u}{d_u}$. Therefore,

$$\alpha_{i_B} \in \left[\frac{1}{\tilde{\eta}^2}\left(\frac{1}{\sum_u\sqrt{d_uT_u}}\right)^2, \frac{1}{\tilde{\eta}^2}\min_{u\in[m]}\frac{T_u}{d_u}\right].$$

Hence, the premises of Lemma 4 is satisfied for $i = i_B$; this gives that with probability $1-\delta/2$,

$$\sum_{t=1}^{T} A_t^{i_B}\Delta_t \leq C\cdot\tilde{\eta}\sum_u\sqrt{d_uT_u}/\sqrt{\alpha_{i_B}}, \tag{11}$$

and

$$\sum_{t=1}^{T} A_t^{i_B} r_t^{i_B} \leq C\cdot\tilde{\eta}\sqrt{\alpha_{i_B}}\sum_u\sqrt{d_uT_u}. \tag{12}$$

Now from Equation (12) and the definition of i_B, we have

$$\sum_{t=1}^{T} A_t^{i_B} r_t^{i_B} \leq C\cdot\tilde{\eta}\sqrt{\alpha_{i_B}}\sum_u\sqrt{d_uT_u} < B'.$$

Applying Lemma 3, we deduce that for all t in $[T]$, $A_t^{i_B} = 1$. Plugging this back to Equation (11), we have

$$\begin{aligned}\sum_{t=1}^{T}\Delta_t =& \sum_{t=1}^{T} A_t^{i_B}\Delta_t\\ \leq& C\cdot\tilde{\eta}\sum_u\sqrt{d_uT_u}/\sqrt{\alpha_{i_B}}\\ \leq& \tilde{O}\left(\frac{\tilde{\eta}^2(\sum_u\sqrt{d_uT_u})^2}{B}\right).\end{aligned}$$

where the second inequality is from the lower bound of α_{i_B} in Equation (10).

Combining the above observation with Lemma 1, along with the union bound, we get that with probability $1-\delta$,

$$R = \sum_{t=1}^{T}(\hat{y}_t - \langle\theta^*, x_t\rangle)^2 = \tilde{O}\left(\sum_{t=1}^{T}\Delta_t\right) = \tilde{O}\left(\frac{\tilde{\eta}^2(\sum_u\sqrt{d_uT_u})^2}{B}\right). \qquad \square$$

A.3 Proof of Theorem 4

Define

$$\beta_k = \beta_k(\mathcal{F},\delta) := 8\eta^2\log\left(4\mathcal{N}(\mathcal{F},1/T^2,\|\cdot\|_\infty)/\delta\right) + 2k/T^2(16+\sqrt{2\eta^2\ln\left(16k^2/\delta\right)}), \tag{13}$$

and

$$R_u := \frac{T_u}{T^2} + 4\min(d'_u, T_u) + 4d'_u\beta_T\ln T_u = \tilde{O}\left(\eta^2 d'_u\log\mathcal{N}(\mathcal{F},T^{-2},\|\cdot\|_\infty)\right).$$

Analogous to Theorem 1, the following theorem provides the query and regret guarantees of of Algorithm 3.

Theorem 5. *Suppose the example sequence $\{x_t\}_{t=1}^T$ has the following structure: $[T]$ has an admissible partition $\{I_u : u\in[m]\}$, where for each u, $|I_u| = T_u$, and the eluder dimension of $\mathcal{F}$ w.r.t. $\{x_t\}_{t\in I_u}$ is d'_u. Suppose $\alpha \leq \frac{1}{\tilde{\eta}^2}\min_{u\in[m]}\frac{T_u}{R_u}$. With probability $1-\delta$, Algorithm 3 satisfies:*
1. Its query complexity $Q = \tilde{O}(\tilde{\eta}\cdot\sqrt{\alpha}\sum_u\sqrt{R_uT_u})$.
2. Its regret $R = \tilde{O}(\tilde{\eta}\cdot\sum_u\sqrt{R_uT_u})/\sqrt{\alpha}$.

We shall prove Theorem 4 directly below; the proof of Theorem 5 follows as a corollary, using the same argument in the proof of Theorem 2; we note that the admissibility condition on domain partition $\{I_u\}_{u=1}^m$ ensures that $\{A_t^i(r_t^i - p_t^i)\mathbb{1}[t \in I]\}_{t=1}^T$ and $\{A_t^i(r_t^i - p_t^i)\Delta_t\mathbb{1}[t \in I]\}_{t=1}^T$ are still martingale difference sequences in our proof.

Proof of Theorem 4. We focus on proving the analogues of Lemma 1 and Lemma 2; the rest of the proof follows the same argument as the proof of Theorem 2 and is therefore omitted.

Lemma 5 (Analogue of Lemma 1)**.** *With probability* $1 - \delta/2$, $R \le \sum_{t=1}^T \Delta_t$.

Proof. Recall that the confidence set at time t is $\mathcal{F}_t = \{f \in \mathcal{F} : \sum_{i \in \mathcal{Q}_t} (f(x_i) - \hat{f}_t(x_i))^2 \le \beta_{|\mathcal{Q}_t|}(\mathcal{F}, \delta)\}$. By Russo and Van Roy (2013, Proposition 2), we have that with probability $1 - \delta/2$, $f^* \in \mathcal{F}_t$, for all $t \in [T]$.

Meanwhile, if $f^* \in \mathcal{F}_t$, for all $t \in [T]$, $(\hat{f}_t(x_t) - f^*(x_t))^2 \le \sup_{f_1, f_2 \in \mathcal{F}_t}(f_1(x_t) - f_2(x_t))^2 = \Delta_t$. This implies that the regret is bounded by $R \le \sum_{t=1}^T \Delta_t$. □

Lemma 6 (Analogue of Lemma 2)**.** $\sum_{t \in I_u} q_t \Delta_t \le R_u$.

Proof. Let $k = |I_u \cap \mathcal{Q}_T|$ and write $d = d'_u$ as a shorthand. Let $(D_1, \dots, D_k)$ be $\{\Delta_t : t \in I_u \cap \mathcal{Q}_T\}$ sorted in non-increasing order. We have

$$\sum_{t \in I_u \cap \mathcal{Q}_T} \Delta_t = \sum_{j=1}^k D_j = \sum_{j=1}^k D_j \mathbb{1}[D_j \le 1/T^4] + \sum_{j=1}^k D_j \mathbb{1}[D_j > 1/T^4].$$

Clearly, $\sum_{j=1}^k D_j \mathbb{1}[D_j \le 1/T^4] \le \frac{T_u}{T^2}$.

We know for all $j \in [k]$, $D_j \le 4$. In addition, $D_j > \epsilon^2 \iff \sum_{t \in I_u \cap \mathcal{Q}_T} \mathbb{1}[\Delta_t > \epsilon^2] \ge j$. By Lemma 7 below, this can only occur if $j < (4\beta_T/\epsilon^2 + 1)d$. Thus, when $D_j > \epsilon^2$, $j < (4\beta_T/\epsilon^2 + 1)d$, which implies $\epsilon^2 < \frac{4\beta_T d}{j-d}$. This shows that if $D_j > 1/T^4$, $D_j \le \min\left\{4, \frac{4\beta_T d}{j-d}\right\}$. Therefore $\sum_j D_j \mathbb{1}[D_j > 1/T^4] \le 4d + \sum_{j=d+1}^k \frac{4\beta_T d}{j-d} \le 4d + 4d\beta_T \log T_u$.

Consequently,

$$\sum_{t \in I_u} q_t \Delta_t = \sum_{t \in I_u \cap \mathcal{Q}_T} \Delta_t \le \min\left\{4T_u, \frac{T_u}{T^2} + 4d'_u + 4d'_u \beta_T \log T_u\right\} \le R_u. \quad \square$$

□

The following lemma generalizes Russo and Van Roy (2013, Proposition 3), in that it considers a subsequence of examples coming from a subdomain of $\mathcal{X}$. We define $\dim_I^E$ as the eluder dimension of $\mathcal{F}$ with respect to support $\{x_t : t \in I\}$. It can be easily seen that $\dim_{I_u}^E \le \dim_u^E$.

Lemma 7. *Fix $I \subseteq [T]$. If $\{\beta_t \ge 0\}_{t=1}^T$ is a nondecreasing sequence and $\mathcal{F}_t := \{f \in \mathcal{F} : \sum_{i \in \mathcal{Q}_t} (f(x_i) - \hat{f}_t(x_i))^2 \le \beta_{|\mathcal{Q}_t|}(\mathcal{F}, \delta)\}$, then*

$$\forall \epsilon > 0, \sum_{t \in I \cap \mathcal{Q}_T} \mathbb{1}[\Delta_t > \epsilon^2] < \left(\frac{4\beta_T}{\epsilon^2} + 1\right) dim_I^E(\mathcal{F}, \epsilon).$$

Proof. Let $k = |I \cap \mathcal{Q}_T|$, $(a_1, \dots, a_k) = (x_t : t \in I \cap \mathcal{Q}_T)$, and $(b_1, \dots, b_k) = (\Delta_t : t \in I \cap \mathcal{Q}_T)$. First, we show that if $b_j > \epsilon^2$ then a_j is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(a_1, \dots, a_{j-1})$, for $j \le k$, in other words, if there exist K disjoint subsequences of $(a_1, \dots, a_{j-1})$ such that a_j is ϵ-dependent on all of them, then $K < \frac{4\beta_T}{\epsilon^2}$.

Indeed, suppose $b_j > \epsilon^2$ and $a_j = x_t$, there are $f_1, f_2 \in \mathcal{F}_t$ such that $f_1(a_j) - f_2(a_j) > \epsilon$. By definition, if a_j is ϵ-dependent on a subsequence $(a_{i_1}, \dots, a_{i_p})$ of $(a_1, \dots, a_{j-1})$, then $\sum_{l=1}^p (f_1(a_{i_l}) - f_2(a_{i_l}))^2 > \epsilon^2$. Thus, if $a_j = x_t$ is ϵ-dependent on K subsequences of $(a_1, \dots, a_{j-1})$, then $\sum_{i \in \mathcal{Q}_t} (f_1(x_i) - f_2(x_i))^2 > K\epsilon^2$. By the triangle inequality,

$$\sqrt{\sum_{i \in \mathcal{Q}_t} (f_1(x_i) - f_2(x_i))^2} \le \sqrt{\sum_{i \in \mathcal{Q}_t} (f_1(x_i) - f^*(x_i))^2} + \sqrt{\sum_{i \in \mathcal{Q}_t} (f_2(x_i) - f^*(x_i))^2} \le 2\sqrt{\beta_T}.$$

Algorithm 4 Fixed-budget QuFUR for general function class

Require: Hypotheses set $\mathcal{F}$, time horizon T, label budget B, parameter δ, noise level η.
Labeled dataset $\mathcal{Q} \leftarrow \emptyset$.
$k \leftarrow 3\lceil \log_2 T \rceil$.
for $i = 0$ to k **do**
    Parameter $\alpha_i \leftarrow 2^i/T^2$.
for $t = 1$ to T **do**
    Predict $\hat{f}_t \leftarrow \operatorname{argmin}_{f\in\mathcal{F}} \sum_{i\in\mathcal{Q}} (f(x_i) - y_i)^2$.
    Confidence set $\mathcal{F}_t \leftarrow \{f \in \mathcal{F} : \sum_{i\in\mathcal{Q}} (f(x_i) - \hat{f}(x_i))^2 \le \beta_{|\mathcal{Q}|}(\mathcal{F}, \delta)\}$,
where $\beta_k := 8\eta^2 \log(4\mathcal{N}(\mathcal{F}, 1/T^2, \|\cdot\|_\infty)/\delta) + 2k/T^2(16 + \sqrt{2\eta^2 \ln(16k^2/\delta)})$.
    Uncertainty estimate $\Delta_t = \sup_{f_1, f_2 \in \mathcal{F}_t} |f_1(x_t) - f_2(x_t)|^2$.
    for $i = 0$ to k **do**
        if $\sum_{j=1}^{t-1} q_j^i < \lfloor B/k \rfloor$ **then**
            With probability $\min\{1, \alpha_i \Delta_t\}$, set $q_t^i = 1$.
    if $\sum_i q_t^i > 0$ **then**
        Query y_t. $\mathcal{Q} \leftarrow \mathcal{Q} \bigcup \{t\}$.

Thus, $K < 4\beta_T/\epsilon^2$.

Next, we show that in any sequence of elements in I, $(c_1, \ldots, c_\tau)$, there is some c_j that is ϵ-dependent on at least $\tau/d - 1$ disjoint subsequences of $(c_1, \ldots, c_{j-1})$, where $d := \dim_I^E(\mathcal{F}, \epsilon)$. For any integer K satisfying $Kd + 1 \le \tau \le Kd + d$, we will construct K disjoint subsequences $C_1, \ldots, C_K$. First let $C_i = (c_i)$ for $i \in [K]$. If c_{K+1} is ϵ-dependent on $C_1, \ldots, C_K$, our claim is established. Otherwise, select a C_i such that c_{K+1} is ϵ-independent and append c_{K+1} to C_i. Repeat for all $j > K + 1$ until c_j is ϵ-dependent on each subsequence or $j = \tau$. In the latter case $\sum |C_i| \ge Kd$, and $|C_i| = d$. In this case, c_τ must be ϵ-dependent on each subsequence, by the definition of $\dim_I^E$.

Now take $(c_1, \ldots, c_\tau)$ to be the subsequence $(a_{t_1}, \ldots, a_{t_\tau})$ of $(a_1, \ldots, a_k)$ consisting of elements a_j for which $b_j > \epsilon^2$. We proved that each a_{t_j} is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(a_1, \ldots, a_{t_j - 1})$. Thus, each c_j is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(c_1, \ldots, c_{j-1})$.[5] Combining this with the fact that there is some c_j that is ϵ-dependent on at least $\tau/d - 1$ disjoint subsequences of $(c_1, \ldots, c_{j-1})$, we have $\tau/d - 1 < 4\beta_T/\epsilon^2$. Thus, $\tau < (4\beta_T/\epsilon^2 + 1)d$. □

A.4 Analysis of uniform query strategy for online active linear regression with oblivious adversary

Theorem 6. *With probability $1 - \delta$, the uniformly querying strategy with probability μ achieves $\mathbb{E}[R] = \tilde{O}\left(\frac{\tilde{\eta}^2 d}{\mu}\right)$ and $\mathbb{E}[Q] = \mu T$.*

Proof sketch. As $Q = \sum_{t=1}^T q_t$ is a sum of T iid Bernoulli random variables with means μ, $\mathbb{E}[Q] = \mu T$.

We now bound the regret of the algorithm. We still define $\Delta_t = \tilde{\eta}^2 \min\{1, \|x_t\|^2_{M_t^{-1}}\}$.

Using Lemma 2 with $\{a_i\}_{i=1}^k = \{x_t\}_{t=1}^T$, and $S = \mathcal{Q}_T$, $\sum_t q_t \Delta_t = \tilde{O}(\tilde{\eta}^2 d)$. Let $Z_t = q_t \Delta_t$. We have $Z_t \le \Delta_t \le \tilde{\eta}^2$, $\mathbb{E}_{t-1} Z_t = \mu \Delta_t$, and $\mathbb{E}_{t-1} Z_t^2 \le \tilde{\eta}^2 \mu \Delta_t$. Applying Freedman's inequality, with probability $1 - \delta/2$,

$$\sum_{t=1}^T \mu\Delta_t - \sum_{t=1}^T q_t \Delta_t = O\left(\tilde{\eta}\sqrt{\sum_{t=1}^T \mu\Delta_t \ln(\ln T/\delta)} + \tilde{\eta}^2 \ln(\ln T/\delta))\right).$$

The above inequality implies that $\sum_{t=1}^T \Delta_t = \tilde{O}\left(\frac{\tilde{\eta}^2 d}{\mu}\right)$. Now, applying Lemma 1 and take the union bound, we have that

[5] To see this, observe that if c is ϵ-dependent on a sequence S, then c must also be ϵ-dependent on any supersequence of S.

with probability $1-\delta$,

$$R = \tilde{O}\left(\sum_{t=1}^{T}\Delta_t\right) = \tilde{O}\left(\frac{\tilde{\eta}^2 d}{\mu}\right).$$

Use the basic relationship between the expectation and tail probability $\mathbb{E}[R] = \int_0^\infty \mathbb{P}(R \geq a)da$, we conclude that $\mathbb{E}[R] = \tilde{O}\left(\frac{\tilde{\eta}^2 d}{\mu}\right)$. □

B Proofs for lower bounds

B.1 Proof of Theorem 3

Theorem 3. *For any $\eta \geq 1$, any set of positive integers $\{(d_u, T_u)\}_{u=1}^m$ and integer B that satisfy*

$$d_u \leq T_u, \forall u \in [m], \quad \sum_{u=1}^m d_u \leq d, \quad B \geq \sum_{u=1}^m d_u,$$

there exists an oblivious adversary such that:
1. It uses a ground truth linear predictor $\theta^\star \in \mathbb{R}^d$ such that $\|\theta^*\|_2 \leq \sqrt{d}$, and $\left|\langle\theta^*, x_t\rangle\right| \leq 1$; in addition, the noises $\{\xi_t\}_{t=1}^T$ are sub-Gaussian with variance proxy η^2.
2. It shows example sequence $\{x_t\}_{t=1}^T$ such that $[T]$ can be partitioned into m disjoint nonempty subsets $\{I_u\}_{u=1}^m$, where for each u, $|I_u| = T_u$, and $\{x_t\}_{t\in I_u}$ lie in a subspace of dimension d_u.
3. Any online active learning algorithm $\mathcal{A}$ with label budget B has regret $\Omega((\sum_{u=1}^m \sqrt{d_u T_u})^2/B)$.

Proof. Our proof is inspired by Vovk (2001, Theorem 2). For $u \in [m]$ and $i \in [d_u]$, define $c_{u,i} = e_{\sum_{v=1}^{u-1} d_v + i}$, where e_j denotes the j-th standard basis of $\mathbb{R}^d$. It can be easily seen that all $c_{u,i}$'s are orthonormal. In addition, for a vector $\theta \in \mathbb{R}^d$, denote by $\theta_{u,i} = \theta_{\sum_{v=1}^{u-1} d_v + i}$.

For task u, we construct domain $\mathcal{X}_u = \mathrm{span}(c_{u,i} : i \in [d_u])$. The sequence of examples shown by the adversary is the following: it is divided to m blocks, where the u-th block occupies a time interval $I_u = [\sum_{v=1}^{u-1} T_v + 1, \sum_{v=1}^{u} T_v]$; Each block is further divided to d_u subblocks, where for $i \in [d_u - 1]$, subblock (u,i) spans time interval $I_{u,i} = [\sum_{v=1}^{u-1} T_v + (i-1)\lfloor T_u/d_u \rfloor + 1, \sum_{v=1}^{u-1} T_v + i\lfloor T_u/d_u \rfloor]$, and subblock (u, d_u) spans time interval $I_{u,d_u} = [\sum_{v=1}^{u-1} T_v + (d_u - 1)\lfloor T_u/d_u \rfloor + 1, \sum_{v=1}^{u-1} T_v + T_u]$. At block u, examples from domain $\mathcal{X}_u$ are shown; furthermore, for every t in $I_{u,i}$, i.e. in the (u,i)-th subblock, example $c_{u,i}$ is repeatedly shown to the learner. Observe that (u,i)-th subblock contains at least $\lfloor \frac{T_u}{d_u} \rfloor \geq \frac{T_u}{2d_u}$ examples, as $T_u \geq d_u$.

We first choose θ^* from distribution D_θ, such that for every coordinate $j \in [d]$, $\theta^*_i \sim \mathrm{Beta}(1,1)$, which is also the uniform distribution over $[0,1]$. Given θ^*, the adversary reveals labels using the following mechanism: given x_t, it draws $y_t \sim \mathrm{Bernoulli}(\langle\theta^*, x_t\rangle)$ independently and optionally reveals it to the learner upon learner's query. Specifically, given θ^*, if $t \in I_{u,i}$, $y_t \sim \mathrm{Bernoulli}(\theta^*_{u,i})$. By Hoeffding's Lemma, $\xi_t = y_t - \theta^*_{u,i}$ is zero mean subgaussian with variance proxy $\frac{1}{4} \leq \eta^2$.

Denote by $N_{u,i}(t) = \sum_{s\in I_{u,i}: s\leq t} q_s$ the number of label queries of the learner in domain (u,i) up to time t. Because the learner satisfies a budget constraint of B under all environments, we have

$$\mathbb{E}\left[\sum_{u=1}^m \sum_{i=1}^{d_u} N_{u,i}(T) \mid \theta^*\right] \leq B.$$

Adding $2\sum_{u=1}^m d_u$ on both sides and by linearity of expectation, we get

$$\sum_{u=1}^m \sum_{i=1}^{d_u} \mathbb{E}\left[(N_{u,i}(T) + 2) \mid \theta^*\right] \leq B + 2\sum_{u=1}^m d_u \leq 3B. \tag{14}$$

On the other hand, we observe that the expected regret of the algorithm can be written as follows:

$$\mathbb{E}\left[R\right] = \mathbb{E}\left[\sum_{u=1}^m \sum_{i=1}^{d_u} \sum_{t\in I_{u,i}} (\hat{y}_t - \theta^*_{u,i})^2\right],$$

where the expectation is with respect to both the choice of θ^* and the random choices of $\mathcal{A}$.

We define a filtration $\{\mathcal{F}_t\}_{t=1}^T$, where $\mathcal{F}_t$ is the σ-algebra generated by $\left\{(x_s, q_s, y_s q_s)\right\}_{s=1}^t$, which encodes the informative available to the *learner* up to time step t.[6] We note that $\hat{y}_t$ is $\mathcal{F}_{t-1}$-measurable. Denote by $N_{u,i}^+(t) = \sum_{s \in I_{u,i}: s \le t} q_s \cdot \mathbb{1}(y_s = 1)$, which is the number of 1 labels seen on example $c_{u,i}$ by the learner up to round $t-1$. Observe that both $N_{u,i}^+(t-1)$ and $N_{u,i}(t-1)$ are $\mathcal{F}_{t-1}$-measurable.

Observe that conditioned on the interaction logs $(x_s, q_s, y_s q_s)_{s=1}^{t-1}$, the posterior distribution of $\theta_{u,i}^*$ is $\mathrm{Beta}(1 + N_{u,i}^+(t-1), 1 + N_{u,i}(t-1) - N_{u,i}^+(t-1))$. Therefore, define random variable $\hat{y}_t^* = \mathbb{E}\left[\theta_{u,i}^* \mid \mathcal{F}_{t-1}\right] = \frac{1+N_{u,i}^+}{2+N_{u,i}}$, we have by bias-variance decomposition,

$$\begin{aligned}\mathbb{E}\left[(\hat{y}_t - \theta_{u,i})^2 \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[(\hat{y}_t^* - \theta_{u,i}^*)^2 \mid \mathcal{F}_{t-1}\right] + (\hat{y}_t - \hat{y}_t^*)^2 \\ &\ge \mathbb{E}\left[(\hat{y}_t^* - \theta_{u,i}^*)^2 \mid \mathcal{F}_{t-1}\right]\end{aligned}$$

Summing over all time steps, we have

$$\mathbb{E}[R] \ge \mathbb{E}\left[\sum_{u=1}^m \sum_{i=1}^{d_u} \sum_{t \in I_{u,i}} (\hat{y}_t^* - \theta_{u,i}^*)^2\right].$$

On the other hand, from Lemma 8, we have for all $t \in I_{u,i}$,

$$\mathbb{E}\left[(\hat{y}_t - \theta_{u,i}^*)^2 \mid N_{u,i}(T), \theta^*\right] \ge \frac{f(\theta_{u,i}^*)}{2(N_{u,i}(T)+2)},$$

where $f(\gamma) = \min(\gamma \cdot (1-\gamma), (2\gamma-1)^2)$.

By the tower property of conditional expectation and conditional Jensen's inequality, we have

$$\mathbb{E}\left[(\hat{y}_t - \theta_{u,i})^2 \mid \theta^*\right] \ge \mathbb{E}\left[\frac{f(\theta_{u,i}^*)}{N_{u,i}(T)+2} \mid \theta^*\right] \ge \frac{f(\theta_{u,i}^*)}{2(\mathbb{E}\left[N_{u,i}(T) \mid \theta^*\right]+2)}.$$

Summing over all t in $I_{u,i}$, and then summing over all subblocks $(u,i) : u \in [m], i \in [d_u]$, and using the aforementioned fact that the (u,i) subblock has at least $\frac{T_u}{2d_u}$ examples, we have

$$\begin{aligned}\mathbb{E}\left[R \mid \theta^*\right] &= \sum_{u=1}^m \sum_{i=1}^{d_u} \sum_{t \in I_{u,i}} \mathbb{E}\left[(\hat{y}_t - \theta_{u,i})^2 \mid \theta^*\right] \\ &\ge \sum_{u=1}^m \sum_{i=1}^{d_u} \frac{T_u/d_u \cdot f(\theta_{u,i}^*)}{4(\mathbb{E}\left[N_{u,i}(T) \mid \theta^*\right]+2)}. \end{aligned} \tag{15}$$

Combining the above inequality with Equation (14), we have:

$$\begin{aligned}3B \cdot \mathbb{E}\left[R \mid \theta^*\right] &\ge \left(\sum_{u=1}^m \sum_{i=1}^{d_u} \frac{T_u/d_u \cdot f(\theta_{u,i}^*)}{4(\mathbb{E}\left[N_{u,i}(T) \mid \theta^*\right]+2)}\right) \cdot \left(\sum_{u=1}^m \sum_{i=1}^{d_u} \mathbb{E}\left[(N_{u,i}(T) \mid \theta^*\right]+2)\right) \\ &\ge \frac{1}{4}\left(\sum_{u=1}^m \sum_{i=1}^{d_u} \left(\sqrt{T_u/d_u} \cdot \sqrt{f(\theta_{u,i}^*)}\right)\right)^2.\end{aligned}$$

[6] This notion should be distinguished from the history notion H_t defined before, in that it does not include the labels not queried by the learner up to time step t. For s in $[t]$, we use $y_s q_s$ to indicate the labeled data information acquired at time step s; if $q_s = 1$, $y_s q_s = y_s$, encoding the fact that the learner has access to label y_s; otherwise $q_s = 0$, $y_s q_s$ is always 0, meaning that the learner does not have label y_s available.

where the second inequality is from Cauchy-Schwarz. Now taking expectation over θ, using Jensen's inequality and Lemma 9 that $\mathbb{E}\sqrt{f(\theta^*_{u,i})} \geq \frac{1}{25}$, and some algebra yields

$$3B \cdot \mathbb{E}[R] \geq \frac{1}{2}\left(\sum_{u=1}^{m}\sum_{i=1}^{d_u}\left(\sqrt{T_u/d_u}\cdot\mathbb{E}\left[\sqrt{f(\theta^*_{u,i})}\right]\right)\right)^2 \geq \frac{1}{2500}\left(\sum_{u=1}^{m}\sqrt{d_u T_u}\right)^2.$$

In conclusion, we have

$$\mathbb{E}[R] \geq \frac{\left(\sum_{u=1}^{m}\sum_{i=1}^{d_u}\sqrt{T_u/d_u}\right)^2}{7500\cdot B}.$$

As the above expectation is over θ^* chosen randomly from D_θ, there must exists an θ^* from $\mathrm{supp}(D_\theta) = [0,1]^d$ such that

$$\mathbb{E}\left[R \mid \theta^*\right] \geq \frac{\left(\sum_{u=1}^{m}\sum_{i=1}^{d_u}\sqrt{T_u/d_u}\right)^2}{7500\cdot B}$$

holds. This θ^* has ℓ_2 norm at most $\sqrt{\sum_{j=1}^{d}(\theta^*_j)^2} \leq \sqrt{d}$. □

Lemma 8. *If t is in $I_{u,i}$, then*

$$\mathbb{E}\left[(\hat{y}^*_t - \theta^*_{u,i})^2 \mid N_{u,i}(T), \theta^*\right] \geq \frac{f(\theta^*_{u,i})}{2(N_{u,i}(T)+2)},$$

where $f(\gamma) = \min\left(\gamma(1-\gamma), (2\gamma-1)^2\right)$.

Proof. We condition on $N_{u,i}(T) = m$, and a value of θ^*. Recall that $\hat{y}^*_t = \frac{1+N^+_{u,i}}{2+N_{u,i}} = \frac{1+N^+_{u,i}}{2+m}$, where $N^+_{u,i}$ can be seen as drawn from the binomial distribution $\mathrm{Bin}(m, \theta^*_{u,i})$.

$$\begin{aligned}
&\mathbb{E}\left[(\hat{y}^*_t - \theta^*_{u,i})^2 \mid N_{u,i}(T) = m, \theta^*\right] \\
=&\mathbb{E}\left[\left(\frac{1+N^+_{u,i}}{2+m} - \theta^*_{u,i}\right)^2 \mid N_{u,i}(T) = m, \theta^*\right] \\
=&\frac{m\theta^*_{u,i}(1-\theta^*_{u,i})}{(m+2)^2} + \frac{(2\theta^*_{u,i}-1)^2}{(m+2)^2} \\
\geq&\frac{m+1}{(m+2)^2} f(\theta^*_{u,i}) \geq \frac{f(\theta^*_{u,i})}{2(m+2)}.
\end{aligned}$$

□

Lemma 9. *Suppose $Z \sim \mathrm{Beta}(1,1)$. Then $\mathbb{E}\left[\sqrt{f(Z)}\right] \geq \frac{1}{25}$.*

Proof. We observe that

$$\mathbb{E}\left[\sqrt{f(Z)}\right] = \int_{[0,1]}\sqrt{f(z)}dz \geq \int_{[\frac{1}{5},\frac{2}{5}]}\sqrt{f(z)}dz,$$

Now, for all $z \in [\frac{1}{5}, \frac{2}{5}]$, $\sqrt{f(z)} \geq \sqrt{\frac{1}{25}} = \frac{1}{5}$, which implies that the above integral is at least $\frac{1}{25}$. □

B.2 Lower bound for unstructured domains

We have the following lower bound in the case when there is no domain structure.

Theorem 7. *For any set of positive integers d, T, B such that $d \leq T$ and $d \leq B$, there exists an oblivious adversary such that:*

1. *it uses a ground truth linear predictor $\theta^\star \in \mathbb{R}^d$ such that $\|\theta^*\|_2 \leq \sqrt{d}$, and $\left|\langle\theta^*, x_t\rangle\right| \leq 1$.*

2. *any online active learning algorithm $\mathcal{A}$ with label budget B has regret at least $\Omega\left(\frac{dT}{B}\right)$.*

Proof. This is an immediate consequence of Theorem 3, by setting $m = 1$, $d_1 = d$, $T_1 = T$, and the label budget equal to B. □

C The c-cost model for online active learning

We consider the following variant of our learning model, which models settings where the cost ratio between a unit of square loss regret and a label query is c to 1. In this setting, the interaction protocol between the learner and the environment remains the same, with the goal of the learner modified to minimizing the total cost, formally $W = cR + Q$. We call the above model the *c-cost model*. We will show that Algorithm 1 achieves optimal cost up to constant factors, for a wide range of values of η and c.

Theorem 8. *For any $\eta \geq 1$, set of positive integers $\{(d_u, T_u)\}_{u=1}^m$ such that $d_u \leq T_u, \forall u \in [m]$, $\sum_{u=1}^m d_u \leq d$, cost ratio $c \geq \max_u \frac{d_u}{T_u}$, there exists an oblivious adversary such that:*

1. *it uses a ground truth linear predictor $\theta^\star \in \mathbb{R}^d$ such that $\|\theta^*\|_2 \leq \sqrt{d}$, and $\left|\langle \theta^*, x_t \rangle\right| \leq 1$; in addition, the subgaussian variance proxy of noise is η^2.*
2. *it shows example sequence $\{x_t\}_{t=1}^T$ such that $[T]$ can be partitioned into m disjoint nonempty subsets $\{I_u\}_{u=1}^m$, where for each u, $|I_u| = T_u$, and $\{x_t\}_{t \in I_u}$ lie in a subspace of dimension d_u.*
3. *any online active learning algorithm $\mathcal{A}$ has total cost $\Omega\left(\sqrt{c} \cdot (\sum_{u=1}^m \sqrt{d_u T_u})\right)$.*

Proof. Consider any algorithm $\mathcal{A}$. Same as in the proof of Theorem 3, we will choose θ^* randomly where each of its coordinates is drawn independently from the $\mathrm{Beta}(1,1)$ distribution, and show the exact same sequence of instances $\{x_t\}_{t=1}^T$ and reveals the labels the same say as in that proof. It can be seen that the η_t's are subgaussian with variance proxy 1, which is also subgaussian with variance proxy η^2.

As $\mathcal{A}$ can behave differently under different environments, we define $\mathbb{E}\left[Q \mid \theta^*\right]$ as $\mathcal{A}$'s query complexity conditioned on the adversary choosing ground truth linear predictor θ^*.

We conduct a case analysis on the random variable $\mathbb{E}\left[Q \mid \theta^*\right]$:

1. If there exists some $\theta^* \in [0,1]^d$, $\mathbb{E}\left[Q \mid \theta^*\right] \geq \sqrt{c}\left(\sum_{u=1}^m \sqrt{d_u T_u}\right)$, then we are done: under the environment where the ground truth linear predictor is θ^*, the total cost of $\mathcal{A}$, $\mathbb{E}\left[W \mid \theta^*\right]$, is clearly at least $\mathbb{E}\left[Q \mid \theta^*\right] \geq \Omega\left(\sqrt{c}\left(\sum_{u=1}^m \sqrt{d_u T_u}\right)\right)$.
2. If for every $\theta^* \in [0,1]^d$, $\mathbb{E}\left[Q \mid \theta^*\right] \leq \sqrt{c}\left(\sum_{u=1}^m \sqrt{d_u T_u}\right)$, $\mathcal{A}$ can be viewed as an algorithm with label budget $B = \sqrt{c}\left(\sum_{u=1}^m \sqrt{d_u T_u}\right)$. By the premise that $c \geq \max_u \frac{d_u}{T_u}$, we get that $B \geq \sum_{u=1}^m \sqrt{d_u T_u} \cdot \sqrt{\frac{d_u}{T_u}} = \sum_{u=1}^m d_u$. Therefore, from the proof of Theorem 3, we get that there exists a θ^* in $[0,1]^d$, such that

$$\mathbb{E}\left[R \mid \theta^*\right] \geq \frac{(\sum_u \sqrt{d_u T_u})^2}{B} \geq \Omega\left(\frac{1}{\sqrt{c}}\left(\sum_u \sqrt{d_u T_u}\right)\right),$$

   which implies that the total cost of $\mathcal{A}$, under the environment where the ground truth linear predictor is θ^*, $\mathbb{E}\left[W \mid \theta^*\right]$, is at least $c \cdot \mathbb{E}\left[R \mid \theta^*\right] \geq \Omega\left(\sqrt{c}\left(\sum_u \sqrt{d_u T_u}\right)\right)$.

In summary, in both cases, there is an oblivious adversary that uses θ^* in $[0,1]^d$, under which $\mathcal{A}$ has a expected cost of $\Omega\left(\sqrt{c}\left(\sum_u \sqrt{d_u T_u}\right)\right)$. □

In the theorem below, we discuss the optimality of Algorithm 1 in the c-cost model for a range of problem parameters.

Theorem 9. *Suppose $\eta \in [1, O(1)]$; in addition, consider a set of $\{(T_u, d_u)\}_{u=1}^m$, such that $\min_u T_u/d_u \geq \eta$. Fix $c \in [\max_u \frac{d_u}{T_u}, \frac{1}{\eta^2} \min_u \frac{T_u}{d_u}]$. We have*

1. *Under all environments with domain dimension and duration $\{(T_u, d_u)\}_{u=1}^m$, such that $\|\theta^*\| \leq C$ and $\max_{t\in[T]} |\langle \theta^*, x_t \rangle| \leq 1$, QuFUR$(c)$ (with the knowledge of norm bound C) has the guarantee that*

$$W \leq \tilde{O}\left(\sqrt{c} \cdot \sum_u \sqrt{T_u d_u}\right),$$

2. *For any algorithm, there exists an environment with domain dimension and duration $\{(T_u, d_u)\}_{u=1}^m$ such that $\|\theta^*\| \leq \sqrt{d}$ and $\max_{t\in[T]} |\langle \theta^*, x_t \rangle| \leq 1$, under which the algorithm must have the following cost lower bound:*

$$W \geq \Omega\left(\sqrt{c} \cdot \sum_u \sqrt{T_u d_u}\right),$$

Proof. We show the two items respectively:

1. As $c \leq \tilde{\eta}^2 \min_u \frac{T_u}{d_u}$, and $c \geq \max_u \frac{d_u}{T_u} \geq \frac{1}{\tilde{\eta}^2}(\frac{1}{(\sum_u \sqrt{d_u T_u})^2})$, applying Theorem 1, we have that QuFUR(c) achieves the following regret and query complexity guarantees:

$$Q \leq O\left(\tilde{\eta}\sqrt{c} \sum_u \sqrt{T_u d_u}\right), \quad R \leq O\left(\tilde{\eta} \sum_u \sqrt{T_u d_u}/\sqrt{c}\right).$$

   This implies that

$$W = cQ + R \leq O\left(\tilde{\eta} \sum_u \sqrt{T_u d_u} \cdot \sqrt{c}\right) = O\left(\sqrt{c} \cdot \sum_u \sqrt{T_u d_u}\right).$$

2. By the condition that $c \geq \max_u \frac{d_u}{T_u}$, applying Theorem 8, we get the item. □

D The regret definition

Recall that in the main text, we define the regret of an algorithm as $R = \sum_{t=1}^T (\hat{y}_t - f^*(x_t))^2$. This is different from the usual definition of regret in online learning, which measures the difference between the loss of the learner and that of the predictor f^*: $\mathrm{Reg} = \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \sum_{t=1}^T (f^*(x_t) - y_t)^2$.

We show a standard result in this section that the expectation of these two notions coincide.

Theorem 10. $\mathbb{E}[R] = \mathbb{E}[\mathrm{Reg}]$.

Proof. Denote by $\mathcal{F}_{t-1}$ be the σ-algebra generated by all observations up to time $t-1$, and x_t. As a shorthand, denote by $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$.

Let $Z_t = (\hat{y}_t - y_t)^2 - (f^*(x_t) - y_t)^2$; we have

$$\begin{aligned}
\mathbb{E}_{t-1} Z_t &= \mathbb{E}_{t-1}\left[(\hat{y}_t - f^*(x_t) + f^*(x_t) - y_t)^2 - (f^*(x_t) - y_t)^2\right] \\
&= \mathbb{E}_{t-1}\left[(f^*(x_t) - \hat{y}_t)^2 + 2(\hat{y}_t - f^*(x_t))(f^*(x_t) - y_t)\right] \\
&= (f^*(x_t) - \hat{y}_t)^2
\end{aligned}$$

where the last inequality uses the fact that $\mathbb{E}_{t-1}(f^*(x_t) - y_t) = 0$ and $\hat{y}_t - f^*(x_t)$ is $\mathcal{F}_{t-1}$-measurable. Consequently, $\mathbb{E} Z_t = \mathbb{E}(f^*(x_t) - \hat{y}_t)^2$. The theorem is concluded by summing over all time steps t from 1 to T. □

E Online to batch conversion

In this section we show that by an standard application of online to batch conversion Cesa-Bianchi et al. (2004a) on QuFUR, we obtain new results on active linear regression under the batch learning setting.

First we recall a standard result on online to batch conversion; for completeness we provide its proof here.

Theorem 11. *Suppose online active learning algorithm $\mathcal{A}$ sequentially receives a set of iid examples $(x_t, y_t)_{t=1}^T$ drawn from D, and at every time step t, it outputs predictor $\hat{f}_t : \mathcal{X} \to \mathcal{Y}$. In addition, suppose $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function. Define regret* $\mathrm{Reg} = \sum_{t=1}^T \ell(\hat{f}_t(x_t), y_t) - \sum_{t=1}^T \ell(f^*(x_t), y_t)$*, and define $\ell_D(f) = \mathbb{E}_{(x,y)\sim D}\ell(f(x), y)$. If $\mathbb{E}[\mathrm{Reg}] \le R_0$, then,*

$$\mathbb{E}\left[\mathbb{E}_{f\sim\mathrm{uniform}(\hat{f}_1,\dots,\hat{f}_T)}\ell_D(f)\right] - \ell_D(f^*) \le \frac{R_0}{T}.$$

Proof. As $\mathrm{Reg} = \sum_{t=1}^T \ell(\hat{f}_t(x_t), y_t) - \sum_{t=1}^T \ell(f^*(x_t), y_t)$, We have

$$\begin{aligned} R_0 \ge \mathbb{E}[\mathrm{Reg}] &= \sum_{t=1}^T \mathbb{E}\left[\ell_D(\hat{f}_t)\right] - \mathbb{E}\left[\sum_{t=1}^T \ell(f^*(x_t), y_t)\right] \\ &= T \cdot \left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\ell_D(\hat{f}_t)\right] - \mathbb{E}_{(x,y\sim D}\ell(f^*(x), y).\right) \end{aligned}$$

The theorem is proved by dividing both sides by T and recognizing that

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\ell_D(\hat{f}_t)\right] = \mathbb{E}_{f\sim\mathrm{uniform}(\hat{f}_1,\dots,\hat{f}_T)}\ell_D(f). \qquad \square$$

Combining Theorem 11 with Theorem 2, we have the following adaptive excess loss guarantee of Fixed-Budget QuFUR (Algorithm 2) when run on iid data with hidden domain structure.

Theorem 12. *Suppose the unlabeled data distribution D_X is a mixture distribution: $D_X = \sum_{u=1}^m p_u D_u$, where D_u is a distribution supported on a subspace of $\mathbb{R}^d$ of dimension d_u and is a subset of $\left\{x : \|x\|_2 \le 1, \left|\langle\theta^*, x\rangle\right| \le 1\right\}$. The conditional distribution of y given x is $y = \langle\theta^*, x\rangle + \xi$ where ξ is a subgaussian with variance proxy η^2. In addition, suppose we are given integer B, T_0 such that $T_0 \ge \Omega\left(\max\left(\frac{B}{\sum_u \sqrt{d_u p_u}\cdot\min_u \sqrt{\frac{p_u}{d_u}}}, \frac{\ln m}{\min_u p_u}\right)\right)$. If Algorithm 2 is given dimension d, time horizon $T \ge T_0$, label budget B, norm bound C, noise level η as input, then:*
1. It uses T unlabeled examples.
2. Its query complexity Q is at most B.
3. Denote by $\ell(\hat{y}, y) = (\hat{y} - y)^2$ the square loss. We have,

$$\mathbb{E}\left[\mathbb{E}_{f\sim\mathrm{uniform}(\hat{f}_1,\dots,\hat{f}_T)}\ell_D(f)\right] - \ell_D(f^*) \le O\left(\frac{\tilde{\eta}^2(\sum_u \sqrt{d_u p_u})^2}{B}\right).$$

Proof sketch. From Theorem 11 it suffices to show that

$$\mathbb{E}[\mathrm{Reg}] \le O\left(\frac{\tilde{\eta}^2 T \cdot (\sum_u \sqrt{d_u p_u})^2}{B}\right).$$

By Theorem 10, $\mathbb{E}[\mathrm{Reg}] = \mathbb{E}[R]$, it therefore suffices to show that

$$\mathbb{E}[R] \le O\left(\frac{\tilde{\eta}^2 T \cdot (\sum_u \sqrt{d_u p_u})^2}{B}\right).$$

We first show a high probability upper bound of R. Given a sequence of unlabeled examples $\{x_t\}_{t=1}^T$, we denote by S_u the subset of examples drawn from component D_u, and denote by T_u the size of S_u. From the assumption of D_u, we know that S_u all lies in a subspace of dimension d_u.

Define event E as follows:

$$E = \left\{ \forall u \in [m] \centerdot T_u \in \left[\frac{Tp_u}{2}, 2Tp_u \right] \right\}.$$

From the assumption that $T \geq T_0 \geq \Omega(\frac{\ln m}{\min_u p_u})$, we have that by Chernoff bound and union bound, $\mathbb{P}(E) \geq 1 - \frac{1}{T^2}$.

Conditioned on event E happening, we have that by the assumption that $T \geq T_0 \geq \frac{B}{\sum_u \sqrt{d_u p_u} \cdot \min_u \sqrt{\frac{p_u}{d_u}}}$,

$$B \leq \tilde{O}\left(T \cdot \sum_u \sqrt{d_u p_u} \min_u \sqrt{\frac{p_u}{d_u}} \right) \leq \tilde{O}\left(\sum_u \sqrt{d_u T_u} \min_u \sqrt{\frac{T_u}{d_u}} \right).$$

Therefore, applying Theorem 2, we have that conditioned on event E happening, with probability $1 - \frac{1}{T^2}$ over the draw of $\{y_t\}_{t=1}^T$,

$$R \leq O\left(\frac{\tilde{\eta}^2 \cdot (\sum_u \sqrt{d_u T_u})^2}{B} \right) \leq O\left(\frac{\tilde{\eta}^2 T \cdot (\sum_u \sqrt{d_u p_u})^2}{B} \right).$$

Combining the above two equations and using union bound, we conclude that with probability $1 - \frac{2}{T^2}$,

$$R \leq O\left(\frac{\tilde{\eta}^2 T \cdot (\sum_u \sqrt{d_u p_u})^2}{B} \right).$$

Observe that with probability 1, $\hat{y}_t \in [-1, 1]$ and $\langle \theta^*, x_t \rangle \in [-1, 1]$. Therefore, $R = \sum_{t=1}^T (\hat{y}_t - \langle \theta^*, x_t \rangle)^2 \in [0, 4T]$. Hence,

$$\mathbb{E}[R] \leq \left(1 - \frac{2}{T^2}\right) \cdot O\left(\frac{\tilde{\eta}^2 T \cdot (\sum_u \sqrt{d_u p_u})^2}{B} \right) + \frac{2}{T^2} \cdot 4T = O\left(\frac{\tilde{\eta}^2 T \cdot (\sum_u \sqrt{d_u p_u})^2}{B} \right).$$

The theorem follows. □

F Kernelisation of QuFUR

We extend QuFUR (α) to kernel regression, following an approach similar to Valko et al. (2013). Assume mapping $\phi : \mathrm{R}^d \to \mathcal{H}$ maps the data to a reproducing kernel Hilbert space. Assume $\|\theta^*\| \leq C = \tilde{O}(1)$, and $\|\phi(x)\| \leq 1$, $\langle \phi(x), \theta^* \rangle^2 \leq 1$, for all x. Define the kernel function $k(x, x') = \phi(x)^\top \phi(x'), \forall x, x' \in \mathrm{R}^d$. Assume the ground-truth label is generated via $y_t = \phi(x_t)^\top \theta^* + \xi_t$.

The kernelised QuFUR algorithm is as follows: Let $\mathcal{Q}_t$ denote the set of indices of the queried examples up to round $t-1$. Denote $M_t = \lambda I + K_t$ where K_t is the kernel matrix $[k(x, x')]_{x,x' \in \mathcal{Q}_t}$, and $\lambda = 1/C^2$. Define column vector $k_t = [k(x_t, x)]_{x \in \mathcal{Q}_t}^\top$. We predict $\hat{y}_t = \mathrm{clip}(k_t^\top M_t^{-1} Y_{\mathcal{Q}_t})$. Uncertainty estimate $\Delta_t = \tilde{\eta}^2 \min\{1, \|k_t\|^2_{M_t^{-1}})\}$, where $\|k_t\|^2_{M_t^{-1}} = \frac{1}{\lambda}(k(x_t, x_t) - k_t^\top M_t^{-1} k_t)$. We still query with probability $\min\{1, \alpha \Delta_t\}$.

A trivial regret and query guarantee is similar to Theorem 1, with d_u replaced by the dimension of the support of $\phi(x)$ for x in domain u, which is possibly infinite. Below we obtain a trade-off dependent on the *effective dimension* $\tilde{d}_u$ of $\mathcal{X}_u$ defined in equation (16). For example, $\tilde{d}_u = \tilde{O}((\log T_u)^{d_u+1})$ for the RBF kernel (Srinivas et al., 2009).

Theorem 13. *Suppose the inputs $\{x_t\}_{t=1}^T$ have the following structure: $[T]$ can be partitioned into m disjoint nonempty subsets $\{I_u\}_{u=1}^m$, where for each u, $|I_u| = T_u$, and the effective dimension of $\{x_t\}_{t \in I_u}$ is $\tilde{d}_u$. If kernelised QuFUR receives inputs dimension d, time horizon T, norm bound $C = \tilde{O}(1)$, noise level η, parameter α, then, with probability $1 - \delta$:*

1. Its query complexity $Q = \tilde{O}\left(\sum_{u=1}^m \min\{T_u, \tilde{\eta}\sqrt{\alpha \tilde{d}_u T_u}\} + 1 \right)$.

2. Its regret $R = \tilde{O}\left(\sum_{u=1}^m \max\{\tilde{\eta}^2 \tilde{d}_u, \tilde{\eta}\sqrt{\tilde{d}_u T_u / \alpha}\} \right)$.

Let S denote the set of indices for queried examples in domain u. Suppose $|S| = s$. If the i-th queried example in domain u happens at time t, we define $a_{i,u} = \phi(x_t)$, $\Phi_i = [\phi(x)^\top]_{x \in \mathcal{Q}_t}^\top$, $\Phi_{i,u} = [\phi(x)^\top]_{x \in \mathcal{Q}_t \cap I_u}^\top$, $N_{i,u} = {\Phi_{i,u}}^\top \Phi_{i,u} + \lambda I$, for all $i \in [s]$. Note that $\|k_t\|^2_{M_t^{-1}} = a_{i,u}^\top (\Phi_i^\top \Phi_i + \lambda I) a_{i,u} \le \|a_{i,u}\|^2_{N_{i,u}^{-1}}$. We still have that

$$\sum_{t \in I_u} q_t \Delta_t = \sum_{i \in S} \tilde{\eta}^2 \min\left(1, \|k_t\|^2_{M_t^{-1}}\right) \le \tilde{\eta}^2 \sum_{i \in S} \min\left(1, \|a_{i,u}\|^2_{N_{i,u}^{-1}}\right).$$

We now focus on bounding $\|a_{i,u}\|^2_{N_{i,u}^{-1}}$. We use the following lemma:

Lemma 10 (Lemma 3 of Valko et al. (2013)). *For all $i \in [s]$, the eigenvalues of $N_{i,u}$ can be arranged so that $\lambda_{j,i-1} \le \lambda_{j,i}$ for all $j \ge 1$; $\lambda_{j,i} \le \lambda_{j-1,i}$ for all $j \ge 2$; $\lambda_{j,0} = \lambda$ for all j, and*

$$\|a_{i,u}\|^2_{N_{i,u}^{-1}} \le \left(4 + \frac{6}{\lambda}\right) \sum_{j=1}^{i} \frac{\lambda_{j,i} - \lambda_{j,i-1}}{\lambda_{j,i-1}}.$$

Let $\Lambda_{s,j} = \sum_{i>j} (\lambda_{i,s} - \lambda)$. The effective dimension of domain u is defined as follows:

$$\tilde{d}_u = \min\{j : j\lambda \ln s > \Lambda_{s,j}\}. \tag{16}$$

The effective dimension is a proxy for the number of principle directions over which the projection of x_t's in domain u in the RKHS is spread. If they fall in a subspace of $\mathcal{H}$ of dimension $\tilde{d}'$, then $\tilde{d}'_u \le \tilde{d}'$. More generally it captures how quickly the eigenvalues of ${\Phi_{i,u}}^\top \Phi_{i,u}$ decrease.

We prove Lemma 10 below for completeness. We use the following lemma as a black box:

Lemma 11 (Lemma 19 of Auer (2002)). *Let $\lambda_1 \ge \cdots \ge \lambda_d \ge 0$. The eigenvalues $\nu_1, \ldots, \nu_d$ of the matrix $diag(\lambda_1, \ldots, \lambda_d) + zz^\top$ with $\|z\| \le 1$ can be arranged such that there are $y_{h,j} \ge 0$, $1 \le h < j \le d$, and the following holds:*

$$\nu_j \ge \lambda_j \tag{17}$$

$$\nu_j = \lambda_j + z_j^2 - \sum_{h=1}^{j-1} y_{h,j} + \sum_{k=j+1}^{d} y_{j,k} \tag{18}$$

$$\sum_{h=1}^{j-1} y_{h,j} \le z_j^2 \tag{19}$$

$$\sum_{j=h+1}^{d} y_{h,j} \le \nu_h - \lambda_h \tag{20}$$

$$\sum_{j=1}^{d} \nu_j = \sum_{j=1}^{d} \lambda_j + \|z\|^2 \tag{21}$$

and if $\lambda_h > \lambda_j + 1$ then

$$y_{h,j} < \frac{z_j^2 z_h^2}{\lambda_h - \lambda_j - 1}. \tag{22}$$

Proof of Lemma 10. We omit the domain index subscript u for clarity. Assume $\phi = \phi_{\mathcal{E}}$ where $\mathcal{E}$ is some basis for $\mathcal{H}$. Let $\mathcal{B}$ be any basis of $\mathcal{H}$ extended from a maximal linearly independent subset of $\{a_j\}_{j \le i}$. If $Q_{\mathcal{B}\mathcal{E}}$ denotes the change of basis matrix from $\mathcal{B}$ to $\mathcal{E}$ then $\Phi_{\mathcal{E},i} = \Phi_{\mathcal{B},i} Q_{\mathcal{B}\mathcal{E}}$ and

$$\Phi_{\mathcal{E},i}^\top \Phi_{\mathcal{E},i} = Q_{\mathcal{B}\mathcal{E}}^\top \Phi_{\mathcal{B},i}^\top \Phi_{\mathcal{B},i} Q_{\mathcal{B}\mathcal{E}}$$

where $\Phi_{\mathcal{B},i}$, $\Phi_{\mathcal{E},i}$ denote Φ_i with respect to the basis $\mathcal{B}$, $\mathcal{E}$. Thus the eigenvalues of $\Phi_{\mathcal{E},i}^\top \Phi_{\mathcal{E},i}$ do not depend on the basis, and we can focus on $\Phi_{\mathcal{B},i}^\top \Phi_{\mathcal{B},i}$, which has zeros everywhere outside its top-left $i \times i$-submatrix. Denote this submatrix as C_i.

We apply Lemma 11 by setting $d = i$, $\lambda_1 \geq \cdots \geq \lambda_d \geq \lambda$ as the eigenvalues of $C_i + \lambda I_i$, and z as the first i entries of the vector ${Q_{\mathcal{BE}}^{\top}}^{-1} a_i$. Our target turns into

$$\|a_i\|^2_{N_i^{-1}} = \sum_{j=1}^{d} \frac{z_j^2}{\lambda_j}$$

For any $1 \leq h < j \leq d$, if $\lambda_h > \lambda_j + 3$, by inequality (22), we have

$$y_{h,j} \leq \frac{1}{2} z_j^2 z_h^2,$$

and since $\|z\| \leq 1$,

$$\sum_{h:h<j,\lambda_h>\lambda_j+3} y_{h,j} \leq \frac{z_j^2}{2} \sum_{h:h<j,\lambda_h>\lambda_j+3} z_h^2 \leq \frac{z_j^2}{2}.$$

If $\lambda_h \leq \lambda_j + 3$, since $\lambda_j \geq \lambda$, $\lambda_j \geq \frac{\lambda}{\lambda+3}\lambda_h$, so

$$\begin{aligned}
\sum_{j=1}^{d} \sum_{h<j:\lambda_h\leq\lambda_j+3} \frac{y_{h,j}}{\lambda_j} &\leq \frac{\lambda+3}{\lambda} \sum_{j=1}^{d} \sum_{h<j:\lambda_h\leq\lambda_j+3} \frac{y_{h,j}}{\lambda_h} \\
&\leq \frac{\lambda+3}{\lambda} \sum_{h=1}^{d} \sum_{j=h+1}^{d} \frac{y_{h,j}}{\lambda_h} \\
&\leq \frac{\lambda+3}{\lambda} \sum_{j=1}^{d} \frac{\nu_j - \lambda_j}{\lambda_j}
\end{aligned}$$

where the last step is due to inequality (20).

By Equation (18),

$$\begin{aligned}
z_j^2 &\leq \nu_j - \lambda_j + \sum_{h=1}^{j-1} y_{h,j} = \nu_j - \lambda_j + \sum_{h<j,\lambda_h>\lambda_j+3} y_{h,j} + \sum_{h<j,\lambda_h\leq\lambda_j+3} y_{h,j} \\
&\leq \nu_j - \lambda_j + \frac{z_j^2}{2} + \sum_{h<j,\lambda_h\leq\lambda_j+3} y_{h,j},
\end{aligned}$$

so

$$\begin{aligned}
\sum_{j=1}^{d} \frac{z_j^2}{\lambda_j} &\leq 2 \sum_{j=1}^{d} \frac{\nu_j - \lambda_j}{\lambda_j} + 2 \sum_{j=1}^{d} \sum_{h<j:\lambda_h\leq\lambda_j+3} \frac{y_{h,j}}{\lambda_j} \\
&\leq \left(2 + 2 \cdot \frac{\lambda+3}{\lambda}\right) \sum_{j=1}^{d} \frac{\nu_j - \lambda_j}{\lambda_j} \\
&= \left(4 + \frac{6}{\lambda}\right) \sum_{j=1}^{d} \frac{\nu_j - \lambda_j}{\lambda_j},
\end{aligned}$$

or equivalently,

$$\|a_i\|^2_{N_i^{-1}} \leq \left(4 + \frac{6}{\lambda}\right) \sum_{j=1}^{i} \frac{\lambda_{j,i} - \lambda_{j,i-1}}{\lambda_{j,i-1}}.$$

□

The proof of Theorem 13 is similar to that of Theorem 1. We only prove the following analogue to Lemma 2.

Lemma 12. $\sum_{i\in S}\min\left(1, \|a_{i,u}\|^2_{N_{i,u}^{-1}}\right) \le \tilde{O}(\tilde{d}_u)$.

Proof. By Equation (10),

$$\begin{aligned}\sum_{i\in S}\|a_{i,u}\|^2_{N_{i,u}^{-1}} &\le \left(4+\frac{6}{\lambda}\right)\sum_{i=1}^{s}\sum_{j=1}^{i}\frac{\lambda_{j,i}-\lambda_{j,i-1}}{\lambda_{j,i-1}}\\ &\le \left(4+\frac{6}{\lambda}\right)\sum_{i=1}^{s}\left[\sum_{j=1}^{\tilde{d}_u}\frac{\lambda_{j,i}-\lambda_{j,i-1}}{\lambda_{j,i-1}}+\sum_{j=\tilde{d}_u+1}^{s}\frac{\lambda_{j,i}-\lambda_{j,i-1}}{\lambda_{j,i-1}}\right]\end{aligned}$$

Since we assume $C=\tilde{O}(1)$, we have $4+\frac{6}{\lambda}=\tilde{O}(1)$. To bound the second term, since the denominators are at least λ,

$$\begin{aligned}\sum_{i=1}^{s}\sum_{j=\tilde{d}_u+1}^{s}\frac{\lambda_{j,i}-\lambda_{j,i-1}}{\lambda_{j,i-1}} &\le \frac{1}{\lambda}\sum_{i=1}^{s}\sum_{j=\tilde{d}+1}^{s}(\lambda_{j,i}-\lambda_{j,i-1})\\ &= \frac{1}{\lambda}\sum_{j=\tilde{d}_u+1}^{s}(\lambda_{j,s}-\lambda)\\ &\le \tilde{d}_u\ln s\end{aligned}$$

where the last inequality follows from Definition 16.

To bound the first term, define $\alpha_{j,i}=\lambda_{j,i}-\lambda_{j,i-1}$, so the first term becomes

$$\sum_{i=1}^{s}\sum_{j=1}^{\tilde{d}_u}\frac{\alpha_{j,i}}{\sum_{p=1}^{i-1}\alpha_{j,p}+\lambda}.$$

To upper bound this term, we solve the following relaxed optimization program

$$\begin{aligned}&\max\left\{\sum_{i=1}^{s}\sum_{j=1}^{\tilde{d}_u}\frac{\alpha_{j,i}}{\sum_{p=1}^{i-1}\epsilon_{j,p}+\lambda}\right\}\\ &s.t.\forall i\in[s],\sum_{j=1}^{\tilde{d}_u}\alpha_{j,i}=\sum_{j=1}^{\tilde{d}_u}\epsilon_{j,i}\le 1.\end{aligned}$$

The optimal solution is $\alpha_{j,i}=\epsilon_{j,i}=1/\tilde{d}_u$, for all j,i. We verify this via the KKT conditions below. Write the Lagrangian

$$\begin{aligned}L(\alpha,\epsilon,\mu,g) &= \sum_{i=1}^{s}\sum_{j=1}^{\tilde{d}_u}\frac{\alpha_{j,i}}{\sum_{p=1}^{i-1}\epsilon_{j,p}+\lambda}-\sum_{i=1}^{s}(\mu_i(\sum_j\alpha_{j,i}-\sum_j\epsilon_{j,i}))-\sum_{i=1}^{s}(g_i(\sum_j\alpha_{j,i}-1))\\ \frac{\partial L}{\partial\alpha_{j,i}} &= \frac{1}{\sum_{p=1}^{i-1}\epsilon_{j,p}+\lambda}-\mu_i-g_i\\ \frac{\partial L}{\partial\epsilon_{j,i}} &= -\sum_{q=i+1}^{s}\frac{\alpha_{j,q}}{(\sum_{p=1}^{q-1}\epsilon_{j,p}+\lambda)^2}+\mu_i\end{aligned}$$

Plugging in $\alpha_{j,i}=\epsilon_{j,i}=1/\tilde{d}_u$, for all j,i,

$$\begin{aligned}\mu_i &= \sum_{q=i+1}^{s}\frac{\tilde{d}_u}{(q-1+\lambda\tilde{d}_u)^2}\ge 0\\ g_i &= \frac{\tilde{d}_u}{i-1+\lambda\tilde{d}_u}-\mu_i\ge 0\end{aligned}$$

Therefore the maximum objective value is $\tilde{d_u} \sum_{i=1}^{s} \frac{1}{i-1+\lambda \tilde{d_u}} = \tilde{O}(\tilde{d_u} \log(\frac{s}{\lambda \tilde{d_u}} + 1))$.

Summing up both terms completes the proof. □

G Comparison of oracle baseline and QuFUR in large budget settings

Consider the optimization program

$$\min_{\mu} \sum_{u=1}^{m} d_u/\mu_u, \text{ s.t. } \sum_{u=1}^{m} \mu_u T_u \leq B, \mu_u \in [0,1], \forall u \in [m]. \tag{23}$$

Theorem 14. *The solution to 23, $\{\mu_u\}_{u=1}^m$, has the following structure: there exists a constant C, such that*

$$\mu_u = \min\left(1, C\sqrt{\frac{d_u}{T_u}}\right).$$

Proof. Since the constraints are linear, define the Lagrangian $L(\mu, \lambda, \gamma) = \sum_u \frac{d_u}{\mu_u} + \lambda(\sum_u T_u \mu_u - B) + \gamma^\top(\mu - 1)$, where $\lambda \in \mathbb{R}$, $\gamma \in \mathbb{R}^m$. By the complementary slackness condition,

1. If $\gamma_u > 0$, $\mu_u = 1$. In this case $\gamma_u = d_u - \lambda T_u$.
2. If $\mu < 1$, $\gamma_u = 0$. In this case $\mu_u = \sqrt{\frac{d_u}{\lambda T_u}}$.

The proof is complete by taking $C = 1/\sqrt{\lambda}$. □

Theorem 14 implies that for $B > \sum_u \sqrt{d_u T_u} \min \sqrt{T_u/d_u}$, if we always query each domain with a fixed probability, the optimal solution is to query all T_u examples from domain u when $\sqrt{d_u/T_u} > \tau$, and query with probability proportional to $\sqrt{d_u/T_u}$ for the rest of the domains. With this setting of μ, in domain u, the total query complexity is $T_u\mu_u = \min\left(T_u, C\sqrt{d_u T_u}\right)$; the regret is $\tilde{\eta}^2 \frac{d_u}{\mu_u} = \max\left(\tilde{\eta}^2 d_u, \frac{\tilde{\eta}^2}{C}\sqrt{d_u T_u}\right)$.

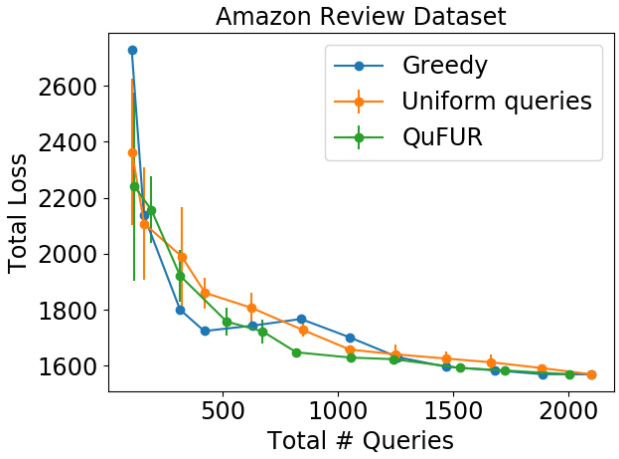
We observe that QuFUR(α) (query w.p. $\min\{1, \alpha\Delta_t\}$) achieves the same upper bound. Specifically, for every setting of C, consider $\alpha = (\frac{C}{\tilde{\eta}})^2$. Define $U_1 = \left\{u : C^2 d_u > T_u\right\}$, and $U_2 = \left\{u : C^2 d_u \leq T_u\right\}$. In other words, U_1 (resp. U_2) is the collection of domains where the domain-aware uniform sampling baseline uses query probability μ_u is $= 1$ (resp. < 1). Observe that U_1 and U_2 constitutes a partition of $[m]$.

1. For $u \in U_1$, the domain-aware uniform querying baseline sets $\mu_u = 1$ and has query complexity T_u and regret $\tilde{\eta}^2 d_u$. On the other hand, QuFur(α) has the same query complexity bound of T_u trivially, and has a regret of $\tilde{\eta}^2\left(d_u + \frac{1}{C}\sqrt{d_u T_u}\right) = O(\tilde{\eta}^2 d_u)$, matching the baseline performance.
2. For $u \in U_2$, the baseline sets $\mu_u = C\sqrt{\frac{d_u}{T_u}}$, and has query complexity $C\sqrt{d_u T_u}$ and regret $\tilde{\eta}^2 \frac{1}{C}\sqrt{d_u T_u}$. On the other hand, QuFur(α) has the query complexity bound of $C^2 d_u + C\sqrt{d_u T_u} = C\sqrt{d_u T_u}$, and has a regret of $\tilde{\eta}^2\left(d_u + \frac{1}{C}\sqrt{d_u T_u}\right) \leq \tilde{\eta}^2 \frac{1}{C}\sqrt{d_u T_u}$, matching the baseline performance.

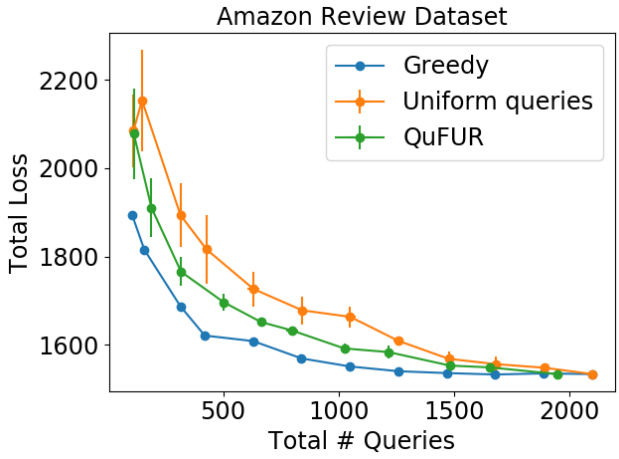
H Additional experimental details

For linear classification experiments, we use the same query strategy as Algorithm 1, i.e. querying with probability $\min\{1, \alpha\Delta_t\}$. As to prediction strategy, we train a linear model with NLL loss and Adam optimizer (learning rate 0.003, weight decay 0.001). After each new query, we train the model on all queried data for 3 additional epochs, with batch size 64.

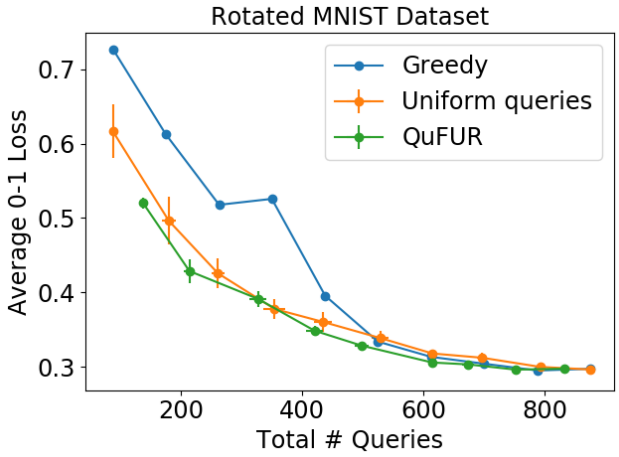
Figure 4 shows the tradeoff curves for alternative domain setups on different datasets. QuFUR maintains competitive performance when we reverse the order of domains (Figures 4a 4c 4e), interleave domains (Figure 4d), and make the domains homogeneous in duration (Figure 4f), with the exception of randomly shuffled inputs from all domains (Figure 4b). In this case, since the inputs are iid, greedy queries are the optimal strategy. However, greedy is unlikely to perform well whenever there is domain shift.



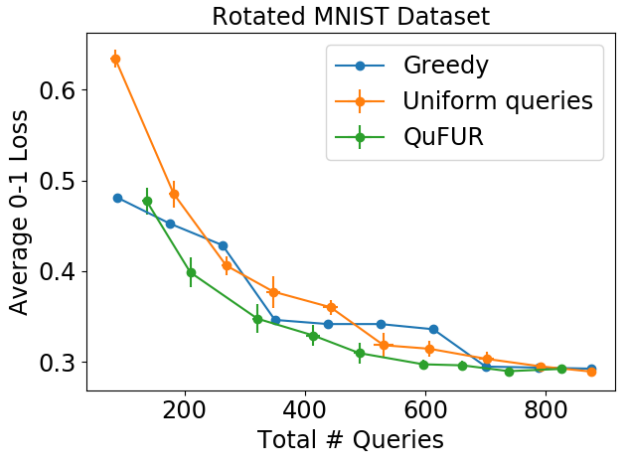
(a) Amazon reviews dataset with video games topic duration 1200 + grocery topic duration 600 + automobile topic duration 300.



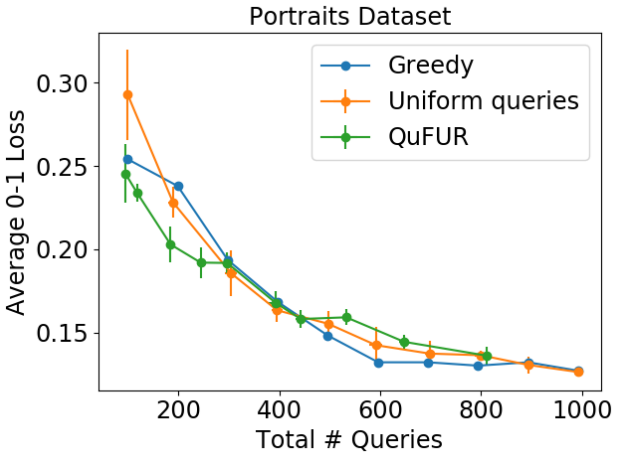
(b) Amazon reviews dataset with randomly shuffled inputs from all 3 topics.



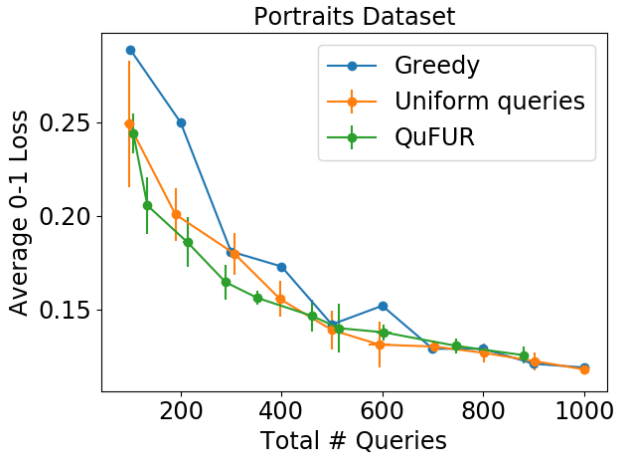
(c) Rotated MNIST dataset with 60°-rotation duration 125 + 30°-rotation duration 250 + no-rotation duration 500.



(d) Rotated MNIST dataset with 60°-rotation duration 250 + 30°-rotation duration 250 + 60°-rotation duration 125 + no-rotation duration 125 + 60°-rotation duration 125.



(e) Portraits dataset when we use the first 32, 64, 128, 256, 512 images from each time period.



(f) Portraits dataset with the first 200 images from all domains.

Figure 4: Tradeoff curves for alternative domain setups.