A. Experimental Details

A.1. Finding Cost Ranges with Online Approximation

Consider the maximum and minimum costs for a fixed label y at round i, both of which may be suppressed. First, define

$$\underline{g}_w \triangleq \arg\min_{g\in\mathcal{G}} \widehat{R}(g) + w(g(x) - 0)^2$$
$$\overline{g}_w \triangleq \arg\min_{g\in\mathcal{G}} \widehat{R}(g) + w(g(x) - 1)^2$$

and recall the definition of $g_{i,y} \triangleq \mathrm{argmin}_{g\in\mathcal{G}} \widehat{R}(g)$ given in Algorithm 1. Owing to the monotonicity property of $\hat{R}(g, w, c; y)$ (Lemma 1), an alternative to MINCOST and MAXCOST is to find

$$\underline{w} := \max\{w \mid \widehat{R}(\underline{g}_w) - \widehat{R}(g_{i,y}) \le \Delta_i\} \tag{9}$$
$$\overline{w} := \max\{w \mid \widehat{R}(\overline{g}_w) - \widehat{R}(g_{i,y}) \le \Delta_i\} \tag{10}$$

and return $\underline{g}_{\underline{w}}(x)$ and $\overline{g}_{\overline{w}}(x)$ as the minimum and maximum costs. We use two steps of approximation here. Using the definition of $\overline{g}_w$ and $\underline{g}_w$ we have:

$$\widehat{R}(\underline{g}_w) - \widehat{R}(g_{i,y}) \le w \cdot g_{i,y}(x)^2 - w \cdot \underline{g}_w(x)^2$$
$$\widehat{R}(\overline{g}_w) - \widehat{R}(g_{i,y}) \le w \cdot (g_{i,y}(x) - 1)^2 - w \cdot (\overline{g}_w(x) - 1)^2.$$

We use this upper bound in place of $\widehat{R}(g_w) - \widehat{R}(g_{i,y})$ in Eqs. (9) and (10). Second, we replace $g_{i,y}$, $\underline{g}_w$, and $\overline{g}_w$ with approximations obtained by online updates. More specifically, we replace $g_{i,y}$ with $g^o_{i,y}$, the current regressor produced by all online updates so far, and approximate the others by

$$\underline{g}_w(x) \approx g^o_{i,y}(x) - w \cdot s(x, 0, g^o_{i,y})$$
$$\overline{g}_w(x) \approx g^o_{i,y}(x) + w \cdot s(x, 1, g^o_{i,y}),$$

where $s(x, y, g^o_{i,y}) \ge 0$ is a *sensitivity* value that approximates the change in prediction on x resulting from an online update to $g^o_{i,y}$ with features x and label y. The computation of this sensitivity value is governed by the actual online update where we compute the derivative of the change in the prediction as a function of the importance weight w for a hypothetical example with cost 0 or cost 1 and the same features. This is possible for essentially all online update rules on importance weighted examples and it corresponds to taking the limit as $w \to 0$ of the change in prediction due to an update divided by w. By inspection this requires only $\mathcal{O}(d)$ time per example, where d is the average number of non-zero features. With these two steps, we obtain approximate minimum and maximum costs using

$$g^o_{i,y}(x) - \underline{w}^o \cdot s(x, 0, g^o_{i,y})$$
$$g^o_{i,y}(x) + \overline{w}^o \cdot s(x, 1, g^o_{i,y}),$$

where

$$\underline{w}^o \;\triangleq\; \max\{w \mid w\left(g^o_{i,y,}(x)^2 - (g^o_{i,y}(x) - w \cdot s(x, 0, g^o_{i,y}))^2\right) \le \Delta_i\}$$
$$\overline{w}^o \;\triangleq\; \max\{w \mid w\left((g^o_{i,y,}(x) - 1)^2 - (g^o_{i,y}(x) + w \cdot s(x, 1, g^o_{i,y}) - 1)^2\right) \le \Delta_i\}.$$

The online update guarantees that $g^o_{i,y}(x) \in [0, 1]$. Since the minimum cost is lower bounded by 0, we have $\underline{w}^o \in \left(0, \frac{g^o_{i,y}(x)}{s(x,0,g^o_{i,y})}\right]$. Finally, because the objective $w(g^o_{i,y}(x))^2 - w(g^o_{i,y}(x) - w \cdot s(x, 0, g^o_{i,y}))^2$ is increasing in w within this range (which can be seen by inspecting the derivative), we can find $\underline{w}^o$ with binary search. Using the same techniques, we also obtain an approximate maximum cost.

It is worth noting that the approximate cost ranges (without the sensitivity trick) are contained in the exact cost ranges because we approximate the difference in squared error by an *upper bound*. Hence, the query rule in this online algorithm should be more aggressive than the query rule in Algorithm 1.

Table 2. Best learning rates

	ImageNet 20	ImageNet 40	RCV1-v2	POS	NER	NER-wiki
passive	1	1	0.5	1.0	0.5	0.5
active (10^{-1})	0.05	0.1	0.5	1.0	0.1	0.5
active (10^{-2})	0.05	0.5	0.5	1.0	0.5	0.5
active (10^{-3})	1	10	0.5	10	0.5	0.5

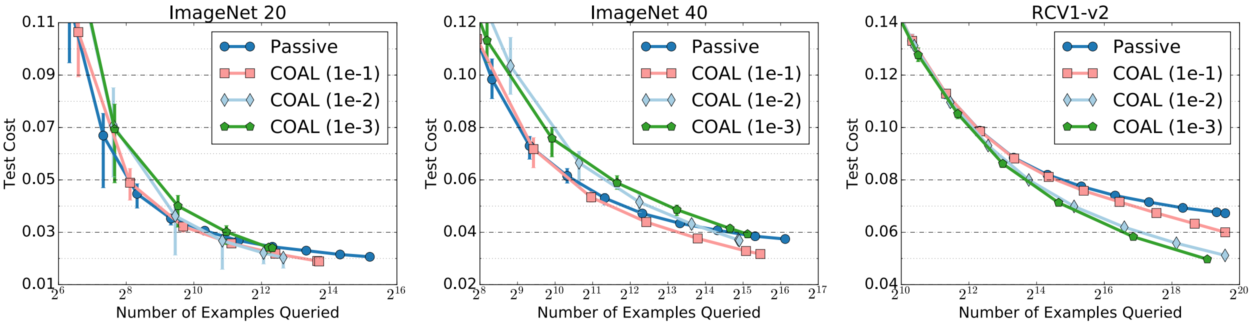

Figure 3. Additional figures for simulated active learning experiments. The plots show the test cost as a function of the number of examples where even a single query was issued.

A.2. Choosing the Learning Rate

For all experiments, we show the results obtained by the best learning rate for each mellowness on each dataset. We choose the best learning rate as follows. For each dataset let $\mathrm{perf}(m, l, q, t)$ denote the test performance of the algorithm using mellowness m and learning rate l on the t^{th} permutation of the training data under a query budget of $2^{(q-1)} \cdot 10 \cdot K, q \geq 1$. Let $\mathrm{query}(m, l, q, t)$ denote the number of queries actually made. Note that $\mathrm{query}(m, l, q, t) < 2^{(q-1)} \cdot 10 \cdot K$ if the algorithm runs out of the training data before reaching the q^{th} query budget[8]. To evaluate the trade-off between test performance and number of queries, we define the following performance measure:

$$\mathrm{AUC}(m, l, t) = \frac{1}{2} \sum_{q=1}^{q_{\max}} \Big(\mathrm{perf}(m, l, q+1, t) + \mathrm{perf}(m, l, q, t)\Big) \cdot \left(\log_2 \frac{\mathrm{query}(m, l, q+1, t)}{\mathrm{query}(m, l, q, t)}\right), \tag{11}$$

where $q_{\max}$ is the minimum q such that $2^{(q-1)} \cdot 10$ is larger than the size of the training data. This performance measure is the area under the curve of test performance against numbers of queries in $\log_2$ scale. A large value means the test performance quickly improves with the number of queries. The best learning rate for mellowness m is then chosen as

$$l^\star(m) \triangleq \arg\max_{l} \mathrm{median}_{1 \leq t \leq 100} \quad \mathrm{AUC}(m, l, t).$$

The best learning rates for different datasets and mellowness settings are in Table 2.

A.3. Additional Figures for Simulated Active Learning

In Figure 3, we plot the test error as a function of the number of examples for which at least one query was requested, for each dataset and mellowness parameter. This experimentally corresponds to the L_1 term in our label complexity analysis.

In comparison with Figure 2 involving the total number of queries, the improvements offered by active learning are slightly less dramatic here. This suggests that our algorithm queries just a few labels for each example, but does end up issuing at least one query on most of the examples. Nevertheless, one can still achieve test cost competitive with passive learning using a factor of 2-16 less labeling effort, as measured by L_1.

In Figure 4, we compare COAL with the two active learning baselines, ALLORNONE and NODOM described in Section 6, along with passive learning, on the RCV1-v2 dataset. As in the ImageNet 40 results, here COAL substantially outperforms both baselines and passive learning. However, here ALLORNONE offers marginal improvement over passive, while

[8] In fact, we check the test performance only in between examples, so $\mathrm{query}(m, l, q, t)$ may be larger than $2^{(q-1)} \cdot 10 \cdot K$ by an additive factor of K, which is negligibly small.

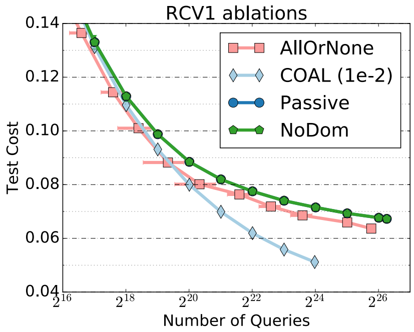

Figure 4. Test cost versus number of queries for COAL, in comparison with active and passive baselines on the RCV1-v2 dataset. Passive learning and NoDom are nearly identical.

NoDom improves over passive on ImageNet 40. Thus, depending on the task, the baselines can improve performance, but COAL is reliably better.

B. Running time analysis

Throughout this section, fix an x, y pair, an iteration i, as well a radius Δ and an accuracy ϵ. We focus on approximating $c_+(x, \mathcal{G}(\Delta; y))$ (See Eqs. (2) and (7)), approximating the minimum cost is very similar. To simplify notation, we drop dependence on x and y. We recall our earlier notation $\widehat{R}_i(g; y)$ (Eq. (5)), except we drop the dependence on both y and i which are fixed in this section. We also recall some other important notation which is accordingly simplified for brevity:

$$\widehat{R}(g) = \hat{\mathbb{E}}[(g(x) - c(y))^2 \mathbb{1}\,(y \text{ queried on } x)], \qquad \widetilde{R}(g, w, c) = \widehat{R}(g) + w(g(x) - c)^2$$

$$g_{\min} = \operatorname{argmin}_{g \in \mathcal{G}} \widehat{R}(g), \qquad \mathcal{G}(\Delta) = \{g \in \mathcal{G} : \widehat{R}(g) - \min_g \widehat{R}(g) \leq \Delta\}$$

$$c_+(\alpha\Delta) = \max_{g \in \mathcal{G}(\alpha\Delta)} g(x), \qquad c_\star = c_+(\Delta).$$

$\widehat{R}(g)$ is the empirical square loss used to define the set of good regressors $\mathcal{G}(\Delta)$ in the algorithm. The precise form of $\widehat{R}(g)$ does not matter in this section. $\widetilde{R}(g, w, c)$ is the empirical square loss with one additional example, with features x, target c, and weight w. $g_{\min}$ is the empirical square loss minimizer, which is the center of the ball $\mathcal{G}(\Delta)$. This functional is used to define new square loss problems in our algorithm. Our goal is to find a number $\hat{c}$ such that,

$$c_\star \leq \hat{c} \leq c_+(4\Delta) + \sqrt{3}\epsilon.$$

Finally, let $g_\star$ be any function such that $g_\star(x) = c_\star$ and $\widehat{R}(g_\star) - \widehat{R}(g_{\min}) \leq \Delta$. In other words $g_\star$ realizes the maximum cost on example x. Note that $g_\star$ is *not* the same regressor that satisfies the realizability condition.

We start the running time analysis with several lemmas characterizing the behavior of various components of the algorithm.

An important structure to the square loss problem is a monotonicity property of both the risk functional and the predictions.

Lemma 1. *For any c and for $w' \geq w \geq 0$, let $g = \operatorname{argmin}_g \widetilde{R}(g, w, c)$ and $g' = \operatorname{argmin}_g \widetilde{R}(g, w', c)$. Then*

$$\widehat{R}(g') \geq \widehat{R}(g) \text{ and } (g'(x) - c)^2 \leq (g(x) - c)^2.$$

Proof. By the definitions,

$$\begin{aligned}
\widehat{R}(g') + w'(g'(x) - c)^2 &= \widetilde{R}(g', w', c) \leq \widehat{R}(g) + w'(g(x) - c)^2 \\
&= \widehat{R}(g) + w(g(x) - c)^2 + (w' - w)(g(x) - c)^2 \\
&\leq \widehat{R}(g') + w(g'(x) - c)^2 + (w' - w)(g(x) - c)^2.
\end{aligned}$$

Rearranging shows that

$$(w' - w)(g'(x) - c)^2 \leq (w' - w)(g(x) - c)^2.$$

Since $w' \geq w$, we have $(g'(x) - c)^2 \leq (g(x) - c)^2$, which is the second claim. For the first claim, the definition of g gives

$$\widehat{R}(g) + w(g(x) - c)^2 \leq \widehat{R}(g') + w(g'(x) - c)^2$$

Rearranging this inequality gives,

$$\widehat{R}(g') - \widehat{R}(g) \geq w((g(x) - c)^2 - (g'(x) - c)^2) \geq 0. \qquad \square$$

The next critical lemma shows that the termination condition in Line 6 of MAXCOST meets the accuracy guarantee.

Lemma 2. *If $c \geq c_\star, w \geq \Delta/\epsilon^2$ and $g = \operatorname{argmin}_g \widetilde{R}(g, w, c)$ then $g(x) \geq c_\star - \epsilon$. Further, if $g \in \mathcal{G}(\Delta)$, then $g(x) \leq c_\star$.*

Proof. The second claim is straightforward by the definition of $c_\star$.

For the first claim, we work to establish a contradiction. Suppose that $g(x) < c_\star - \epsilon$. By the facts that g is the minimizer of $\widetilde{R}(g, w, c)$, $g_{\min}$ is the minimizer of $\widehat{R}(g)$, and $c \geq c_\star$, we have

$$w(c - (c_\star - \epsilon))^2 < w(c - g(x))^2 \leq \widetilde{R}(g, w, c) - \widehat{R}(g_{\min}) \leq \widetilde{R}(g_\star, w, c) - \widehat{R}(g_{\min}) \leq \Delta + w(c - c_\star)^2.$$

We may further lower bound (again using $c \geq c_\star$),

$$(c - c_\star + \epsilon)^2 = (c - c_\star)^2 + 2(c - c_\star)\epsilon + \epsilon^2 \geq (c - c_\star)^2 + \epsilon^2.$$

Rearranging proves that $w < \Delta/\epsilon^2$. The contrapositive is that if $w \geq \Delta/\epsilon^2$, then we must have $g(x) \geq c_\star - \epsilon$, which is the desired claim. $\square$

The next lemma is the main result for the BINARYSEARCH subroutine.

Lemma 3. *Suppose we invoke the subroutine* BINARYSEARCH *with parameters ϵ and Δ. Then it terminates in polynomial time with $O(\log_2(1/(\epsilon^2)))$ oracle calls. The algorithm outputs two regressors (g_ℓ, g_h) and if $c \geq c_\star$ is passed as input then $c_\star \in [g_\ell(x), g_h(x)]$. If additionally, $g_h \notin \mathcal{G}(4\Delta)$ then $c_\star \leq (g_\ell(x) + g_h(x))/2$.*

Proof. The logarithmic running time is fairly straightforward since in each iteration the algorithm halves the interval, has initial interval of size Δ/ϵ^2 and terminates when the interval is smaller than 2Δ. Thus for $T \geq \log_2(1/(2\epsilon^2))$ the interval has size at most

$$2^{-T}(\Delta/\epsilon^2) \leq 2^{-\log_2(1/(2\epsilon^2))}(\Delta/\epsilon^2) = 2\Delta$$

Hence the number of iterations is upper bounded by $\lceil \log_2(1/(2\epsilon^2)) \rceil$.

For the first termination claim, the invariant that we maintain is that for all $t \geq 1$, $g_{t,h} = \operatorname{argmin}_g \widetilde{R}(g, w_{t,h}, c)$ satisfies $g_{t,h}(x) \geq c_\star$ while $g_{t,\ell} = \operatorname{argmin}_g \widetilde{R}(g, w_{t,\ell}, c)$ satisfies $g_{t,\ell}(x) \leq c_\star$.

For $g_{t,h}$, we first establish the base case. Observe that $g_{1,h} = g_c$ (computed in MAXCOST just before the invocation of BINARYSEARCH) and $\widehat{R}(g_c) \geq \widehat{R}(g_{\min}) + \Delta$ by the termination check in Line 6. By construction, in this iteration and in all others, we have that $g_{t,h} \notin \mathcal{G}(\Delta)$, since this is the requirement for updating $w_{t,h}$. But since $g_{t,h}$ minimizes the risk function $\widetilde{R}(g, w_{t,h}, c)$ we get,

$$\widehat{R}(g_{\min}) + \Delta + w_{t,h}(g_{t,h}(x) - c)^2 \leq \widetilde{R}(g_{t,h}, w_{t,h}, c) \leq \widetilde{R}(g_\star, w_{t,h}, c) \leq \widehat{R}(g_{\min}) + \Delta + w_{t,h}(c_\star - c)^2.$$

Since $c \geq c_\star$, this implies that $g_{t,h}(x) \geq c_\star$.

The proof for $g_{t,\ell}$ is simpler, since we only shrink the interval up if we find something in $\mathcal{G}(\Delta)$. By definition of $c_\star$ this check guarantees that $g_\ell(x) \leq c_\star$.

For the second termination claim we must use the fact that $|w_{t,h} - w_{t,\ell}| \leq 2\Delta$ by the termination condition and $g_h \notin \mathcal{G}(4\Delta)$. Let t be the terminal iteration, so $g_\ell = \operatorname{argmin}_g \widetilde{R}(g, w_{t,\ell}, c)$ and analogously for g_h. Assume for the sake of contradiction that $c_\star \geq (g_h(x) + g_\ell(x))/2$. Since $c \geq c_\star$, this implies that

$$\begin{aligned}\widehat{R}(g_{\min}) + 4\Delta + w_{t,h}(g_h(x) - c)^2 \leq \widetilde{R}(g_h, w_{t,h}, c) \leq \widetilde{R}(g_\star, w_{t,h}, c) &\leq \widehat{R}(g_{\min}) + \Delta + w_{t,h}(c - c_\star)^2 \\ &\leq \widehat{R}(g_{\min}) + \Delta + w_{t,h}\left(c - \frac{g_h(x) + g_\ell(x)}{2}\right)^2.\end{aligned}$$

Similarly we have

$$\widehat{R}(g_{\min}) + w_{t,\ell}(g_\ell(x) - c)^2 \le \widetilde{R}(g_\ell, w_{t,\ell}, c) \le \widetilde{R}(g_\star, w_{t,\ell}, c) \le \widehat{R}(g_{\min}) + \Delta + w_{t,\ell}\left(c - \frac{g_h(x) + g_\ell(x)}{2}\right)^2.$$

Adding the two equations gives

$$\begin{aligned}
&2\Delta + w_{t,\ell}(c - g_\ell(x))^2 + w_{t,h}(c - g_h(x))^2 \le (w_{t,h} + w_{t,\ell})\left(c - \frac{g_h(x) + g_\ell(x)}{2}\right)^2 \\
\Rightarrow\ &2\Delta + w_{t,h}\left[(c - g_\ell(x))^2 + (c - g_h(x))^2\right] \le 2w_{t,h}\left(c - \frac{g_h(x) + g_\ell(x)}{2}\right)^2 + (w_{t,h} - w_{t,\ell})(c - g_\ell(x))^2 \\
\Rightarrow\ &2\Delta + w_{t,h}\left[(c - g_\ell(x))^2 + (c - g_h(x))^2\right] \le 2w_{t,h}\left(c - \frac{g_h(x) + g_\ell(x)}{2}\right)^2 + 2\Delta \quad \text{since } c, g_\ell(x) \in [0,1] \\
\Rightarrow\ &\frac{1}{2}(c - g_\ell(x))^2 + \frac{1}{2}(c - g_h(x))^2 \le \left(c - \frac{g_h(x) + g_\ell(x)}{2}\right)^2.
\end{aligned}$$

The last line is a contradiction since $\mathbb{E}[f(Z)] \ge f(\mathbb{E}[Z])$ for convex f, which can be applied by taking $Z = \text{Unif}(\{g_\ell(x), g_h(x)\})$ and $f(y) = (c - y)^2$. □

The last lemma ensures sufficient progress in the case when $g_h \notin \mathcal{G}(4\Delta)$, which is crucial for the oracle complexity bound.

Lemma 4. *Suppose $c \ge c_\star$ and that there exists $g \in \mathcal{G}(\Delta)$ such that $c - g(x) = \delta$ with $\delta \in [\sqrt{3}\epsilon, 1]$. Then if the output (g_ℓ, g_h) of* BINARYSEARCH *satisfies $g_h \notin \mathcal{G}(4\Delta)$, then $g_h(x) \le c + \delta - \epsilon^2$.*

Proof. We never use a weight larger than Δ/ϵ^2 by the initialization of $w_{1,\ell}, w_{1,h}$. Now suppose that we output g_h such that $\widehat{R}(g_h) - \widehat{R}(g_{\min}) > 4\Delta$, which by construction is the minimizer of $\widetilde{R}(\cdot, w, c)$ for some $w \le \Delta/\epsilon^2$. Then

$$\widehat{R}(g_{\min}) + 4\Delta + w(g_h(x) - c)^2 \le \widetilde{R}(g_h, w, c) \le \widetilde{R}(g, w, c) \le \widehat{R}(g_{\min}) + \Delta + w(g(x) - c)^2 = \widehat{R}(g_{\min}) + \Delta + w\delta^2.$$

Rearranging, using the fact that $\Delta \ge w\epsilon^2$, and dividing through by $w > 0$ gives

$$(g_h(x) - c)^2 \le \delta^2 - 3\epsilon^2.$$

The condition on δ ensures that the right hand side is non-negative. It is easy to see that $\delta^2 - 3\epsilon^2 \le (\delta - \epsilon^2/\delta)^2$ simply by expanding the square. Hence we get that

$$|g_h(x) - c| \le |\delta - \epsilon^2/\delta| \le \delta - \epsilon^2.$$

We can safely remove the absolute value on the right hand side since we have the condition that $\delta \ge \sqrt{3}\epsilon$, which ensures that $\delta - \epsilon^2/\delta$ is non-negative. The absolute value on the left hand side can also be removed, since if $g_h(x) \le c$ we have already proved what is required. Specifically, since we must have $\epsilon \in (0, 1)$ for the preconditions of the lemma to be satisfied and $\delta \ge \sqrt{3}\epsilon$, it follows that $c \le c + \delta - \epsilon^2$. Since g_h is the result of an oracle call with weight $w \le \Delta/\epsilon^2$, either it has $\widehat{R}(g_h) - \widehat{R}(g_{\min}) \le 4\Delta$, or it must have $g_h(x) \le c + \delta - \epsilon^2$. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. The first step of the proof is to inductively verify that $c \ge c_\star, h \ge c_\star, \ell \le c_\star$ at all steps in the algorithm execution. These invariants are clearly maintained at the onset of the algorithm. Now suppose they are maintained at the onset of some iteration. If g_c satisfies $\widehat{R}(g_c) \le \widehat{R}(g_{\min}) + \Delta$, then by Lemma 2 we are done. Otherwise, we obtain two regressors (g_ℓ, g_h) from BINARYSEARCH. For the lower bound, we always have $g_\ell(x) \le c_\star$ by Lemma 3, which verifies the inductive step for ℓ. For the upper bound to $c_\star$, if $\widehat{R}(g_h) - \widehat{R}(g_{\min}) \le 4\Delta$ then by Lemma 3, we know that $c_\star \le g_h(x)$, but we also know that $g_h(x) \le c_+(4\Delta)$ by the definition, so we are done. The last case is when $\widehat{R}(g_h) > 4\Delta$, but here we may apply the second statement of Lemma 3, which asserts that $c_\star \le (g_h(x) + g_l(x))/2$. The settings of ℓ, h, c now verify the inductive claim, since $\ell \ge g_l$ implies that $(h + l)/2 \ge (g_h(x) + g_l(x))/2 \ge c_\star$.

This immediately proves the correctness of the algorithm, since the loop stopping condition, along with the invariant, guarantees that $c \geq c_\star \geq \ell$ which means that

$$\hat{c} - c_\star \leq \hat{c} - \ell = \frac{h - \ell}{2} \leq \sqrt{3}\epsilon.$$

For the iteration complexity, we must apply Lemma 4. In particular, we use the width of the interval $[\ell, h]$ which contains $c_\star$ as a potential function and show that it decreases with every step. Let δ_t denote $h - \ell$, which is the width of the interval before the t^{th} iteration (so $\delta_1 = 1$). Every non-terminal iteration satisfies $c \geq c_\star$. Moreover, for any $t > 1$, we use as the regressor g, the one that achieved the value ℓ used to define c. This ensures that $g(x) = \ell$. Furthermore, in application of Lemma 4, we set $\delta \triangleq c - g(x) = c - \ell$, which conveniently gives $2\delta = \delta_t = h - \ell$. Recall that we entered the loop at t^{th} iteration, meaning that $\delta_t \geq 2\sqrt{3}\epsilon$ and hence $\delta \in [\sqrt{3}\epsilon, 1]$. Lemma 4 states that either we terminate successfully, or we are guaranteed that $g_h(x) \leq c + \delta - \epsilon^2$. This means that

$$\delta_{t+1} = g_h(x) - \max\{\ell, g_\ell(x)\} \leq c + \delta - \epsilon^2 - \ell = \delta + \delta - \epsilon^2 = \delta_t - \epsilon^2,$$

where the first equality used $c - \ell = \delta$ which is true by definition. Since we terminate at the first T such that $\delta_T \leq 2\sqrt{3}\epsilon$, we require at most $O(1/\epsilon^2)$ iterations. By Lemma 3, each iteration takes $O(\log(1/\epsilon))$ oracle calls. □

C. Generalization analysis

To bound the generalization error of Algorithm 1, we start by defining the central random variable in the analysis. At round i, recall our notation $Q_i(y) = \mathbb{1}$ (query y on example x_i) which indicates the query rule. The central random variable is,

$$M_i(g; y) \quad \triangleq \left((g(x_i) - c_i(y))^2 - (f^\star(x_i; y) - c_i(y))^2\right) Q_i(y). \tag{12}$$

Here (x_i, c_i) is the i^{th} example and cost presented to the algorithm. For simplicity, we write M_i when the dependence on g and y is clear from context. For a vector regressor f, we write

$$M_i(f; y) \triangleq M_i(f(\cdot; y); y).$$

We also recall some of the constants and notation defined in Algorithm 1 which are heavily used throughout this appendix.

$$\Delta_i = \frac{\kappa\epsilon_{i-1}}{i-1}, \quad \epsilon_i = \left(\frac{n}{i}\right)^\beta \log\left(\frac{2n^2|\mathcal{G}|K}{\delta}\right), \quad \kappa = 80.$$

$$\widehat{R}_i(g; y) = \frac{1}{i-1}\sum_{j=1}^{i-1}\left[(g(x_j) - c_j(y))^2 Q_j(y)\right].$$

$$g_{i,y} = \operatorname*{argmin}_{g \in \mathcal{G}} \widehat{R}_i(g; y), \quad \text{and} \quad f_i = \{g_{i,y}\}_{y=1}^K.$$

$$\mathcal{G}_i(y) = \{g \in \mathcal{G} \mid \forall y,\ \widehat{R}_i(g; y) \leq \widehat{R}_i(g_{i,y}; y) + \Delta_i\},$$

$$\mathcal{F}_i = \{f \in \mathcal{G}^K \mid \forall y,\ \widehat{R}_i(f(\cdot; y); y) \leq \widehat{R}_i(g_{i,y}; y) + \Delta_i\}.$$

For Δ_1 we use the convention that $1/0 = \infty$ so the initial radius is infinite. Let $\mathbb{E}_i[\cdot]$ and $\mathrm{Var}_i[\cdot]$ denote the expectation and variance conditioned on all randomness up to and including round $i - 1$. With these definitions, we turn to several supporting claims.

C.1. Supporting Lemmata

Theorem 7 (Freedman-type inequality (Beygelzimer et al., 2011; Agarwal et al., 2014)). *Let $X_1, \ldots, X_T$ be a sequence of real-valued random variables. Assume for all $t \in \{1, \ldots, T\}$ that $|X_t| \leq R$ and $\mathbb{E}[X_t | X_1, \ldots, X_{t-1}] = 0$. Define $S = \sum_{t=1}^T X_t$ and $V = \sum_{t=1}^T \mathbb{E}[X_t^2 | X_1, \ldots, X_{t-1}]$. For any $\delta \in (0, 1)$ and $\lambda \in [0, 1/R]$, with probability at least $1 - \delta$,*

$$S \leq (e - 2)\lambda V + \frac{\ln(1/\delta)}{\lambda}.$$

Lemma 5 (Concentration of squared loss)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $g \in \mathcal{G}, y \in Y, i \in [n], t \in [n]$:*

$$\left|\sum_{j=i}^{i+t-1} \mathbb{E}_j[M_j] - \sum_{j=i}^{i+t-1} M_j\right| \;\le\; 2\sqrt{\sum_{j=i}^{i+t-1} \underset{j}{\mathrm{Var}}[M_j]\nu_n} + 2\nu_n,$$

where $\nu_n \triangleq \log\left(\frac{2n^2|\mathcal{G}|K}{\delta}\right)$.

Note that $\epsilon_i = \frac{n}{i}^{\beta}\nu_n$, is a scaled version of the confidence bound here, where the scaling shrinks polynomially with i.

Proof. First observe that by the rescaling of the failure parameter, we can apply Freedman's inequality for each i, t, y, g and for each tail and a union bound proves the result.

We now apply the Freedman-type inequality in Theorem 7. For a fixed $g \in \mathcal{G}, y \in Y$, the random variable M_i is measurable with respect to the σ-field $\sigma(\{(x_j, \{(c(x_j;y), Q_j(y))\}_{y \in Y})\}_{j=1}^{i})$, so $M_i - \mathbb{E}_i[M_i]$ forms a martingale difference sequence adapted to this filtration. Moreover $M_i - \mathbb{E}_i[M_i]$ and $\mathbb{E}_i[M_i] - M_i$ are both conditionally centered and clearly at most 2. Thus Freedman's inequality gives,

$$\sum_{j=i}^{i+t-1} M_j - \mathbb{E}_j[M_j] \le 2\sqrt{\sum_{j=i}^{i+t-1} \underset{j}{\mathrm{Var}}[M_j]\nu_n} + 2\nu_n,$$

except with probability $\frac{\delta}{2n^2|\mathcal{G}|K}$. This follows by observing that $(e-2) \le 1$ and setting $\lambda = \sqrt{\nu_n/V}$, provided it meets the constraint $\lambda \le 1/R$. Otherwise we set $\lambda = 1/R$ and use the fact that $1/R \le \sqrt{\nu_n/V}$.

The bound on the right hand side also holds for the lower tail, again except with same probability. Thus a union bound over both tails, all $g \in \mathcal{G}, y \in Y$ and pairs i, t gives the result. □

Lemma 6 (Bounding variance of regression regret)**.** *We have for all $(g, y) \in \mathcal{G} \times Y$,*

$$\begin{aligned}
\mathbb{E}_i[M_i] &= \mathbb{E}_i\left[Q_i(y)(g(x_i) - f^\star(x_i;y))^2\right],\\
\underset{i}{\mathrm{Var}}[M_i] &\le 4\mathbb{E}_i[M_i].
\end{aligned}$$

Proof. We take expectation of M_i over the cost conditioned on a fixed example $x_i = x$ and a fixed query outcome $Q_i(y)$:

$$\begin{aligned}
\mathbb{E}[M_i \mid x_i = x, Q_i(y)] &= Q_i(y) \times \mathbb{E}_c[g(x)^2 - f^\star(x;y)^2 - 2c(y)(g(x) - f^\star(x;y)) \mid x_i = x]\\
&= Q_i(y)\big(g(x)^2 - f^\star(x;y)^2 - 2f^\star(x;y)(g(x) - f^\star(x;y))\big)\\
&= Q_i(y)(g(x) - f^\star(x;y))^2.
\end{aligned}$$

The second equality is by Assumption 1, which implies $\mathbb{E}[c(y) \mid x_i = x] = f^\star(x;y)$. Taking expectation over x_i and $Q_i(y)$, we have

$$\mathbb{E}_i[M_i] \;=\; \mathbb{E}_i\left[Q_i(y)(g(x_i) - f^\star(x_i;y))^2\right].$$

For the variance:

$$\begin{aligned}
\underset{i}{\mathrm{Var}}[M_i] &\le \mathbb{E}_i[M_i^2]\\
&= \mathbb{E}_i\left[Q_i(y)(g(x_i) - f^\star(x_i;y))^2(g(x_i) + f^\star(x_i;y) - 2c(y))^2\right]\\
&\le 4 \cdot \mathbb{E}_i\left[Q_i(y)(g(x_i) - f^\star(x_i;y))^2\right]\\
&= 4\mathbb{E}_i[M_i].
\end{aligned}$$

□

Lemma 7 (Sharp cost-sensitive bound)**.** *For all $i > 0$, if $f^\star \in \mathcal{F}_i$, then for all $f \in \mathcal{F}_i$*

$$\mathbb{E}_{x,c}[c(h_f(x)) - c(h_{f^\star}(x))] \le \min_{\zeta > 0} \left\{ \zeta P_\zeta + \mathbb{1}(\zeta \le 2\eta_i) 2\eta_i + \frac{4\eta_i^2}{\zeta} + \frac{6}{\zeta}\sum_y \mathbb{E}_i[M_i(f;y)] \right\},$$

where $P_\zeta = \Pr_{x\sim\mathcal{D}}[\min_{y\neq h_{f^\star}(x)} f^\star(x,y) \le f^\star(x,h_{f^\star}(x)) + \zeta]$ is the probability that the expected cost of the second best and best label are within ζ of each other.

Proof. Fix some $f \in \mathcal{F}_i$, and let $y(x) = h_f(x)$ and $y^\star(x) = h_{f^\star}(x)$ for shorthand. Define $S_\zeta(x) = \mathbb{1}(f^\star(x,y(x)) \le f^\star(x,y^\star(x)) + \zeta)$ and $S'_\zeta(x) = \mathbb{1}\left(\min_{y\neq y^\star(x)} f^\star(x,y) \le f^\star(x,y^\star(x)) + \zeta\right)$. Observe that for fixed ζ, $S_\zeta(x)\mathbb{1}(y(x) \neq y^\star(x)) \le S'_\zeta(x)$ for all x. We can also majorize the complementary indicator to obtain the inequality

$$S_\zeta^C(x) \le \frac{(f^\star(x,y(x)) - f^\star(x,y^\star(x)))}{\zeta}.$$

We begin with the definition of realizability, which gives

$$\begin{aligned}
\mathbb{E}_{x,c}[c(h_f(x)) - c(h_{f^\star}(x)] &= \mathbb{E}_x\left[(f^\star(x,y(x)) - f^\star(x,y^\star(x)))\,\mathbb{1}(y(x)\neq y^\star(x))\right] \\
&= \mathbb{E}_x\left[\left(S_\zeta(x) + S_\zeta^C(x)\right)(f^\star(x,y(x)) - f^\star(x,y^\star(x)))\,\mathbb{1}(y(x)\neq y^\star(x))\right] \\
&\le \zeta\mathbb{E}_x S'_\zeta(x) + \mathbb{E}_i\left[S_\zeta^C(x)\mathbb{1}(y(x)\neq y^\star(x))(f^\star(x,y(x)) - f^\star(x,y^\star(x)))\right].
\end{aligned}$$

The first term here is exactly the ζP_ζ term in the bound. We now focus on the second term, which depends on our query rule. For this we must consider three cases.

Case 1. If both $y(x)$ and $y^\star(x)$ are not queried, then it must be the case that both have small cost ranges. This follows since $f \in \mathcal{F}_i$ and $h_f(x) = y(x)$ so $y^\star(x)$ does not dominate $y(x)$. Moreover, since the cost ranges are small on both $y(x)$ and $y^\star(x)$, since we know that $f^\star$ is well separated under event $S_\zeta^C(x)$, the relationship between ζ and η_i governs whether we make a mistake or not. Specifically, we get that $S_\zeta^C(x)\mathbb{1}(y(x)\neq y^\star(x))\,\mathbb{1}(\text{no query}) \le \mathbb{1}(\zeta \le 2\eta_i)$ at round i. In other words, if we do not query and the separation is big but we make a mistake, then it must mean that the cost range threshold η_i is also big.

Using this argument, we can bound the second term as,

$$\begin{aligned}
&\mathbb{E}_i\left[S_\zeta^C(x)\mathbb{1}(y(x)\neq y^\star(x))\,\mathbb{1}(y(x),y^\star(x)\text{ not queried})(f^\star(x,y(x)) - f^\star(x,y^\star(x)))\right] \\
&\le \mathbb{E}_i\left[S_\zeta^C(x)\mathbb{1}(y(x)\neq y^\star(x))\,\mathbb{1}(y(x),y^\star(x)\text{ not queried})(f^\star(x,y(x)) - f(x,y(x)) + f(x,y^\star(x)) - f^\star(x,y^\star(x)))\right] \\
&\le \mathbb{E}_i\left[S_\zeta^C(x)\mathbb{1}(y(x)\neq y^\star(x))\,\mathbb{1}(y(x),y^\star(x)\text{ not queried})\,2\eta_i\right] \\
&\le \mathbb{E}_i\left[\mathbb{1}(\zeta\le 2\eta_i)\,2\eta_i\right] = \mathbb{1}(\zeta \le 2\eta_i)\,2\eta_i.
\end{aligned}$$

Case 2. If both $y(x)$ and $y^\star(x)$ are queried, we can easily relate the second term to the square loss,

$$\begin{aligned}
&\mathbb{E}_i\left[S_\zeta^C(x)\mathbb{1}(y(x),y^\star(x)\text{ both queried})(f^\star(x,y(x)) - f^\star(x,y^\star(x)))\right] \\
&\le \frac{1}{\zeta}\mathbb{E}_i\left[\mathbb{1}(y(x),y^\star(x)\text{ both queried})(f^\star(x,y(x)) - f^\star(x,y^\star(x)))^2\right] \\
&\le \frac{1}{\zeta}\mathbb{E}_i\left[\mathbb{1}(y(x),y^\star(x)\text{ both queried})(f^\star(x,y(x)) - f(x,y(x)) + f(x,y^\star(x)) - f^\star(x,y^\star(x)))^2\right] \\
&\le \frac{2}{\zeta}\mathbb{E}_i\left[\mathbb{1}(y(x)\text{ queried})(f^\star(x,y(x)) - f(x,y(x)))^2 + \mathbb{1}(y^\star(x)\text{ queried})(f(x,y^\star(x)) - f^\star(x,y^\star(x)))^2\right] \\
&\le \frac{2}{\zeta}\sum_y \mathbb{E}_i\left[Q_i(y)(f^\star(x,y) - f(x,y))^2\right] = \frac{2}{\zeta}\sum_y \mathbb{E}_i\left[M_i(f;y)\right].
\end{aligned}$$

Passing from the second to third line here is justified by the fact that $f^\star(x,y(x)) \ge f^\star(x,y^\star(x))$ and $f(x,y(x)) \le f(x,y^\star(x))$ so we added two non-negative quantities together. The last step uses Lemma 6. While not written, we also use the event $\mathbb{1}(y(x)\neq y^\star(x))$ to avoid losing a factor of 2.

Case 3. The last case is if one label is queried and the other is not. Both cases here are analogous, so we do the derivation for when $y(x)$ is queried but $y^\star(x)$ is not. Since in this case, $y^\star(x)$ is not dominated ($h_f(x)$ is never dominated provided $f \in \mathcal{F}_i$), we know that the cost range for $y^\star(x)$ must be small. Using this fact, and essentially the same argument as in case 2, we get

$$\begin{aligned}
&\mathbb{E}_i\left[S_\zeta^C(x)\mathbb{1}\left(y(x) \text{ queried}, y^\star(x) \text{ not}\right)\left(f^\star(x,y(x)) - f^\star(x,y^\star(x))\right)\right]\\
&\frac{1}{\zeta}\mathbb{E}_i\left[\mathbb{1}\left(y(x) \text{ queried}, y^\star(x) \text{ not}\right)\left(f^\star(x,y(x)) - f^\star(x,y^\star(x))\right)^2\right]\\
&\le \frac{2}{\zeta}\mathbb{E}_i\left[\mathbb{1}\left(y(x) \text{ queried}, y^\star(x) \text{ not}\right)\left(f^\star(x,y(x)) - f(x,y(x))\right)^2 + \left(f(x,y^\star(x)) - f^\star(x,y^\star(x))\right)^2\right]\\
&\le \frac{2\eta_i^2}{\zeta} + \frac{2}{\zeta}\mathbb{E}_i\left[\mathbb{1}\left(y(x) \text{ queried}\right)\left(f^\star(x,y(x)) - f(x,y(x))\right)^2\right]\\
&\le \frac{2\eta_i^2}{\zeta} + \frac{2}{\zeta}\sum_y \mathbb{E}_i\left[M_i(f;y)\right].
\end{aligned}$$

We also obtain this term for the other case where $y^\star(x)$ is queried by $y(x)$ is not.

To summarize, adding up the contributions from these cases (which is an over-estimate since at most one case can occur and all are non-negative), we get

$$\mathbb{E}_{x,c}[c(h_f(x)) - c(h_{f^\star}(x)] \le \zeta P_\zeta + \mathbb{1}\left(\zeta \le 2\eta_i\right)2\eta_i + \frac{4\eta_i^2}{\zeta} + \frac{6}{\zeta}\sum_y \mathbb{E}_i\left[M_i(f;y)\right].$$

This bound holds for any ζ, so it holds for the minimum. □

C.2. Proof of Theorem 3

Conditioning on the high-probability event in Lemma 5, we prove the theorem by induction. Define

$$\Delta_i' = \min\{1, \frac{\kappa\nu_n}{i-1}\}, \quad \nu_n = \log\left(\frac{2n^2|\mathcal{G}|K}{\delta}\right).$$

We will make use of the following simple fact, which applies since $i \le n$, so the premultiplier on ϵ_i is at least 1.

Fact 1. *For all $i \in [n]$, we have $\nu_n \le \epsilon_i$.*

Concretely we consider the inductive hypothesis:

$$\forall y, \forall i \ge 1, \quad \widehat{R}_i(f^\star(\cdot;y);y) \le \min_{g\in\mathcal{G}} \widehat{R}_i(g;y) + \frac{c_0\nu_n}{i-1} \quad \text{and} \quad \mathbb{E}[c(h_{f_i}(x)) - c(h_{f^\star}(x))] \le \min_{\zeta>0}\left\{\zeta P_\zeta + \frac{2K\Delta_i'}{\zeta}\right\} \quad (13)$$

where $c_0 = 10$. The first claim in particular implies that $f^\star(\cdot;y) \in \mathcal{F}_i$ since we chose $\Delta_{i+1} = \kappa\epsilon_i/i$ and using Fact 1. For the base case $i = 1$, observe that the right hand side of the first inequality is infinity but the empirical squared loss is 0 for all regressors. Hence the first claim is trivially satisfied. Moreover, because the excess cost-sensitive classification risk is always upper-bounded by 1, it is trivially bounded by $\frac{2K\Delta_1'}{\zeta}$ for any $\zeta \in [0,1]$. For $\zeta > 1$, we have $\zeta P_\zeta = \zeta$ so again the bound is trivial.

Now assume the inductive hypothesis holds for the first i rounds, $i \ge 1$. We want to analyze the set $\mathcal{F}_{i+1}$, which is computed at the end of the i^{th} iteration of Algorithm 1 based on i examples (technically the beginning of the $(i+1)^{\text{st}}$

iteration). Invoking Lemma 5, with parameters 1 and i, and Lemma 6, we have for all $(g, y) \in \mathcal{G} \times Y$,

$$\begin{aligned}\sum_{j=1}^{i} \mathbb{E}_j[M_j(g;y)] - \sum_{j=1}^{i} M_j(g;y) &\leq 2\sqrt{4\nu_n \sum_{j=1}^{i} \mathbb{E}_j[M_j(g;y)]} + 2\nu_n \\ &\leq 2\left(4\nu_n + \frac{1}{4}\sum_{j=1}^{i} \mathbb{E}_j[M_j(g;y)]\right) + 2\nu_n \\ &= 10\nu_n + \frac{1}{2}\sum_{j=1}^{i} \mathbb{E}_j[M_j].\end{aligned}$$

This bound implies that

$$-\sum_{j=1}^{i} M_j \;\leq\; 10\nu_n, \quad \text{(since } \mathbb{E}_j[M_j(g;y)] \geq 0 \text{ by Lemma 6)}$$

and therefore

$$\widehat{R}_{i+1}(f^\star(\cdot;y);y) \;\leq\; \widehat{R}_{i+1}(g;y) + \frac{c_0\nu_n}{i}. \tag{14}$$

Since this bound applies for all $g \in \mathcal{G}$, it proves the first part of the inductive claim.

Next we prove that the empirical squared loss minimizer f_{i+1} after iteration i has small excess risk. Fix some label y. To simplify notation, we drop the dependence on y and define for any j:

$$g_j \triangleq f_j(\cdot;y), \quad g^\star \triangleq f^\star(\cdot;y), \quad \mathcal{G}_j \triangleq \mathcal{G}_j(y), \quad \widehat{R}_j(g) \triangleq \widehat{R}_j(g;y).$$

Let M_j be defined for g_{i+1} and y according to Eq. (12). We first prove that since g_{i+1} is the empirical loss minimizer at round i for label y, it must have been in the version space $\mathcal{G}_j(y)$ for all $j \in \{1, \ldots, i+1\}$.

Because g_{i+1} is the loss minimizer for label y after round i, we have

$$\sum_{j=1}^{i} M_j = \sum_{j=1}^{i} M_j(g_{i+1};y) \;\leq\; 0.$$

Now suppose $g_{i+1} \notin \mathcal{G}_{t+1}$ for some $t \in \{0, \ldots, i\}$. We have

$$\begin{aligned}\sum_{j=1}^{t} M_j &= t\left(\widehat{R}_{t+1}(g_{i+1}) - \widehat{R}_{t+1}(g^\star)\right) \\ &= t\left(\widehat{R}_{t+1}(g_{i+1}) - \widehat{R}_{t+1}(g_{t+1}) + \widehat{R}_{t+1}(g_{t+1}) - \widehat{R}_{t+1}(g^\star)\right) \\ &\geq \kappa\epsilon_t - c_0\nu_n. \end{aligned} \tag{15}$$

The last inequality here follows since $g_{i+1} \notin \mathcal{G}_{t+1}$ so it must have $\widehat{R}_{t+1}(g_{i+1}) - \widehat{R}_{t+1}(g_{t+1}) \geq \kappa\epsilon_t/t$ by the elimination rule. Simultaneously, we use Eq. (14) which lower bounds the second term. Combining this inequality with the fact that $\sum_{j=1}^{i} M_j \leq 0$ gives

$$\sum_{j=t+1}^{i} M_j \leq c_0\nu_n - \kappa\epsilon_t. \tag{16}$$

Applying Lemmas 5 and 6 along with the inequality $\sqrt{4ab} \leq a/\alpha + \alpha b$ for all $\alpha > 0$, gives

$$\sum_{j=t+1}^{i} \mathbb{E}_j[M_j] - \sum_{j=t+1}^{i} M_j \;\leq\; 2\sqrt{4\nu_n \sum_{j=t+1}^{i} \mathbb{E}_j[M_j]} + 2\nu_n \leq \frac{1}{2}\sum_{j=t+1}^{i} \mathbb{E}_j[M_j] + 10\nu_n. \tag{17}$$

Combining the last inequality and Eq. (16), we get

$$\sum_{j=t+1}^{i} \mathbb{E}_j[M_j] \le 20\nu_n + 2c_0\nu_n - 2\kappa\epsilon_t \le (40 - 2\kappa)\epsilon_t < 0.$$

The strict inequality here is based on Fact 1 and the parameter setting $\kappa = 80$. This is a contradiction since $\mathbb{E}_j[M_j]$ is a quadratic form and hence non-negative by Lemma 6. The same analysis applies to every y. Therefore, we know that the empirical square loss vector regressor f_{i+1} is in $\mathcal{F}_j$ for all $j \in \{1, \ldots, i+1\}$, and hence we can apply Lemma 7 for all of these rounds, to obtain

$$\begin{aligned} & i\big(\mathbb{E}_{x,c}[c(h_{f_{i+1}}(x)) - c(h_{f^\star}(x))]\big) \\ & \le \min_{\zeta > 0} \left\{ \sum_{j=1}^{i} \left(\zeta P_\zeta + \mathbb{1}\left(\zeta \le 2\eta_j\right) 2\eta_j + \frac{4\eta_j^2}{\zeta} + \frac{6}{\zeta} \sum_y \mathbb{E}_j\left[M_j(f_{i+1}; y)\right] \right) \right\} \\ & \le \min_{\zeta > 0} \left\{ i\zeta P_\zeta + \sum_{j=1}^{i} \left(\mathbb{1}\left(\zeta \le 2\eta_j\right) 2\eta_j + \frac{4\eta_j^2}{\zeta} + \frac{6}{\zeta} \sum_y \mathbb{E}_j\left[M_j(f_{i+1}; y)\right] \right) \right\}. \end{aligned}$$

We study the four terms separately. The first one is straightforward and contributes ζP_ζ to the instantaneous cost sensitive regret. Using our definition of $\eta_j = 1/\sqrt{j}$ the second term can be bounded as

$$\sum_{j=1}^{i} \mathbb{1}\left(\zeta < 2\eta_j\right) 2\eta_j = \sum_{j=1}^{\lceil 4/\zeta^2 \rceil} \frac{2}{\sqrt{j}} \le 4\sqrt{\lceil 4/\zeta^2 \rceil} \le \frac{12}{\zeta}.$$

The inequality above, $\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \le 2\sqrt{n}$, is well known. For the third term, using our definition of η_j gives

$$\sum_{j=1}^{i} \frac{4\eta_j^2}{\zeta} = \frac{4}{\zeta} \sum_{j=1}^{i} \frac{1}{j} \le \frac{4}{\zeta}(1 + \log(i)).$$

Finally, the fourth term can be bounded using Lemma 5 (Eq. (17) with $t = 0$), which reveals

$$\sum_{j=1}^{i} \mathbb{E}_j[M_j] \le 2 \sum_{j=1}^{i} M_j + 20\nu_n$$

Since for each y, $\sum_{j=1}^{i} M_j(f_i; y) \le 0$ for the empirical square loss minimizer (which is what we are considering now), we get

$$\frac{6}{\zeta} \sum_y \sum_{j=1}^{i} \mathbb{E}_j[M_j(f_{i+1}; y)] \le \frac{120}{\zeta} K\nu_n.$$

And hence, we obtain

$$\begin{aligned} \mathbb{E}_{x,c}[c(x; h_{f_{i+1}}(x)) - c(x; h_{f^\star}(x))] & \le \min_{\zeta > 0} \left\{ \zeta P_\zeta + \frac{1}{\zeta i}\left(4\log(i) + 16 + 120K\nu_n\right) \right\} \\ & \le \min_{\zeta > 0} \left\{ \zeta P_\zeta + \frac{140K\nu_n}{\zeta i} \right\} \le \min_{\zeta > 0} \left\{ \zeta P_\zeta + \frac{2\kappa K\nu_n}{\zeta i} \right\} \end{aligned}$$

To obtain this last bound, we observe that $1 \le \log(i) \le \nu_n$ under our assumption that $\delta < 1/e$ so the coefficient in the numerator is at most 140. The inductive claim follows by the definition of Δ'_{i+1}. Or more precisely, if $\Delta'_{i+1} = 1$ then the inductive claim is trivial and otherwise we have proved what is required.

D. Label complexity analysis

D.1. Supporting Lemmata

Our label complexity analysis builds on the following lemma, which uses the sets $\mathcal{G}_i^\star$ and $\mathcal{G}_i$:

$$\mathcal{G}_i(\Delta; y) \triangleq \{g \mid \widehat{R}_i(g; y) - \min_{g' \in \mathcal{G}} \widehat{R}_i(g'; y) \le \Delta\}, \tag{18}$$

$$\mathcal{G}_i^\star(\Delta; y) \triangleq \Big\{g \;\Big|\; \frac{1}{i-1}\sum_{j=1}^{i-1} Q_j(y)(g(x_j) - f^\star(x_j; y))^2 \le \Delta\Big\}. \tag{19}$$

Throughout we use the definitions.

$$\Delta_i \triangleq \kappa\epsilon_{i-1}/(i-1), \kappa \triangleq 80, c_0 \triangleq 10, c_1 \triangleq 25/3, c_2 \triangleq 1/3, \eta_i \triangleq 1/\sqrt{i}, \nu_n \triangleq \log\left(\frac{2n^2|\mathcal{G}|K}{\delta}\right)$$

These are the constants defined in Algorithm 1 with some additional numerical constants that we use in the analysis. We also require a new definition:

$$I_\beta(i) = \max\{t \in \mathbb{N} | (t-1) \le (c_2/c_1)^{1/\beta}(i-1)\}. \tag{20}$$

Note that $I_\beta(i)$ is well defined for $i \ge 1$ since the right hand side is non-negative. However $I_\beta(i)$ could be as small as 1. We first study the I_β functional.

Fact 2. *Define $i_\beta \triangleq 2(c_1/c_2)^{1/\beta} + 1$. Then for $i \ge i_\beta$, we have*

$$I_\beta(i) - 1 \ge \max\{(c_2/c_1)^{1/\beta}(i-1)/2, 2\}.$$

Proof. The proof is by direct calculation.

$$I_\beta(i) - 1 = \lfloor (c_2/c_1)^{1/\beta}(i-1)\rfloor \ge \lfloor (c_2/c_1)^{1/\beta}(i_\beta - 1)\rfloor = 2$$
$$I_\beta(i) - 1 \ge (c_2/c_1)^{1/\beta}(i-1) - 1 = (c_2/c_1)^{1/\beta}(i-1) - \frac{(c_2/c_1)^{1/\beta}(i_\beta - 1)}{2} \ge \frac{(c_2/c_1)^{1/\beta}(i-1)}{2}. \qquad \square$$

We now turn to the more intricate lemmas.

Lemma 8. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $i \ge 1$ and all y,*

$$\mathcal{G}_i^\star(c_2\Delta_i; y) \subset \mathcal{G}_i(\Delta_i; y) \subset \mathcal{G}_i(4\Delta_i; y) \subset \mathcal{G}_i^\star(c_1\Delta_i; y) \subset \mathcal{G}_{I_\beta(i)}^\star(c_2\Delta_{I_\beta(i)}; y),$$

where $I_\beta(i)$ is in Eq. (20).

Proof. The second containment is trivial.

Recall our earlier definition that for a fixed $g \in \mathcal{G}$ and $y \in Y$,

$$M_j \triangleq \left((g(x_j) - c_j(y))^2 - (f^\star(x_j; y) - c_j(y))^2\right) Q_j(y).$$

Let $\mathbb{E}_c[M_j]$ and $\mathrm{Var}_c[M_j]$ denote the expectation and variance taken with respect to the cost c at round j, conditioned on all randomness up to round $j-1$ and on x_j. Following the same proof for Lemma 6, we have that

$$\mathbb{E}_c[M_j] = Q_j(y)(g(x_j) - f^\star(x_j; y))^2, \qquad \text{and} \qquad \operatorname*{Var}_c[M_j] \le 4\mathbb{E}_c[M_j(g; y)].$$

It is also easy to prove a concentration result similar to Lemma 5 where $\mathbb{E}_j[M_j]$ and $\mathrm{Var}_j[M_j]$ are replaced by $\mathbb{E}_c[M_j]$ and $\mathrm{Var}_c[M_j]$, respectively. Thus we have for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $(g, y) \in \mathcal{G} \times Y$ and all $i, t \in [n]$:

$$\left|\sum_{j=i}^{i+t-1} \mathbb{E}_c[M_j] - \sum_{j=i}^{i+t-1} M_j\right| \;\le\; 2\sqrt{4\nu_n \sum_{j=i}^{i+t-1} \mathbb{E}_c[M_j]} + 2\nu_n, \tag{21}$$

where $\nu_n = \log\left(\frac{2n^2|\mathcal{G}|K}{\delta}\right)$ as in Lemma 5. This bound, via the inequality $\sqrt{4ab} \le \alpha a + b/\alpha$ implies

$$\sum_{j=i}^{i+t-1} \mathbb{E}_c[M_j] \le 2\sum_{j=i}^{i+t-1} M_j + 20\nu_n \tag{22}$$

$$\sum_{j=i}^{i+t-1} M_j \le \frac{3}{2}\sum_{j=i}^{i+t-1} \mathbb{E}_c[M_j] + 10\nu_n \tag{23}$$

We start with proving the first containment. Fix some round i, some label y, and some $g \in \mathcal{G}_i^\star(c_2\Delta_i; y)$. Conditioning on the above high-probability event and starting with Eq. (23), we have

$$\begin{aligned}\sum_{j=1}^{i-1} M_j &\le \frac{3}{2}\cdot\left(\sum_{j=1}^{i-1}\mathbb{E}_c[M_j]\right) + 10\nu_n \le \frac{3}{2}\cdot(i-1)\cdot c_2\Delta_i + 10\nu_n \\ &= \frac{3}{2}c_2\kappa\epsilon_{i-1} + 10\nu_n \le \left(\frac{\kappa}{2}+c_0\right)\epsilon_{i-1}.\end{aligned}$$

Above, the second inequality is by

$$\sum_{j=1}^{i-1}\mathbb{E}_c[M_j] \;=\; \sum_{j=1}^{i-1} Q_j(y)(g(x_j) - f^\star(x_j;y))^2 \;\le\; c_2\Delta_i \times (i-1)$$

since $g \in \mathcal{G}_i^\star(c_2\Delta_i; y)$, and the final inequality uses $\nu_n \le \epsilon_{i-1}$ (Fact 1) and our choices of κ, c_0 and c_2. Using the above bound and with $g_i = \mathrm{argmin}_{g\in\mathcal{G}} \widehat{R}_i(g;y)$, we have

$$\begin{aligned}(i-1)\cdot\left(\widehat{R}_i(g;y) - \widehat{R}_i(g_i;y)\right) &= \sum_{j=1}^{i-1} M_j + (i-1)\left(\widehat{R}_i(f^\star;y) - \widehat{R}_i(g_i;y)\right) \\ &\le (\kappa/2 + c_0)\epsilon_{i-1} + c_0\nu_n \le \kappa\epsilon_{i-1},\end{aligned}$$

where the first inequality is by the above upper bound on $\sum_{j=1}^{i-1} M_j$ and Eq. (14), which upper bounds the excess empirical square loss of $f^\star$. Thus, $g \in \mathcal{G}_i(\Delta_i; y) \subset \mathcal{G}_i(4\Delta_i; y)$.

To prove the third containment, we fix some i, y, and $g \in \mathcal{G}_i(4\Delta_i; y)$. Starting from (22) we have

$$\begin{aligned}\sum_{j=1}^{i-1}\mathbb{E}_c[M_j] &\le 2\sum_{j=1}^{i-1} M_j + 20\nu_n \\ &= 2(i-1)\cdot(\widehat{R}_i(g;y) - \widehat{R}_i(f^\star;y)) + 20\nu_n \\ &\le 2(i-1)\cdot(\widehat{R}_i(g;y) - \widehat{R}_i(g_i;y)) + 20\nu_n \\ &\le 8\kappa\epsilon_{i-1} + 20\nu_n \\ &\le c_1\kappa\epsilon_{i-1},\end{aligned}$$

where the second inequality is by the fact that g_i is the square loss minimizer at round i for label y, the third inequality is by $g \in \mathcal{G}_i(4\Delta_i; y)$, and the last inequality is by $\nu_n \le \epsilon_{i-1}$ (Fact 1) and our choices of c_1 and κ. Thus, $g \in \mathcal{G}_i^\star(c_1\Delta_i; y)$.

For the final containment, observe that

$$(i-1)c_1\Delta_i = c_1\kappa\epsilon_{i-1} = c_1\kappa\left(\left(\frac{n}{i-1}\right)^\beta \nu_n\right) = c_2\kappa\left(\left[\left(\frac{c_1}{c_2}\right)^{1/\beta}\frac{n}{i-1}\right]^\beta \nu_n\right).$$

Using the definition of $I_\beta(i)$ in Eq. (20), we get that $(i-1)c_1\Delta_i \le (I_\beta(i)-1)c_2\Delta_{I_\beta(i)}$. Of course we always have $I_\beta(i) - 1 \le i-1$ since $c_2 \le c_1$. Hence,

$$\sum_{j=1}^{I_\beta(i)-1}\mathbb{E}_j Q_j(y)(g(x_j) - f^\star(x_j;y))^2 \le \sum_{j=1}^{i-1}\mathbb{E}_j Q_j(y)(g(x_j) - f^\star(x_j;y))^2 \le (i-1)c_1\Delta_i \le (I_\beta(i)-1)c_2\Delta_{I_\beta(i)}.$$

Thus we get that $\mathcal{G}_i^\star(c_1\Delta_i) \subset \mathcal{G}^\star_{I_\beta(i)}(c_2\Delta_{I_\beta(i)})$. □

Before bounding the label complexity, we first prove the following regret bound:

Lemma 9. *For any $\delta \le 1/e$, with probability at least $1-\delta$, for all $i \ge 1$ and for all vector regressors $f \in \mathcal{F}_i^\star(c_2\Delta_i) \triangleq \prod_y \mathcal{G}_i^\star(c_2\Delta_i; y)$,*

$$\mathbb{E}_{x,c}\left[c(x,h_f(x)) - c(x,h_{f^\star}(x))\right] \le \min_{\zeta>0}\left\{\zeta P_\zeta + \frac{14K\Delta_i}{\zeta}\right\}.$$

Note that this cost-sensitive regret bound is polynomially worse than the one in Theorem 3 that we prove just for the empirical risk minimizer f_i. This is because we set the confidence radius Δ_i using a polynomial function of n/i, which will be important for our label complexity analysis.

Proof. The proof follows a similar argument to that of Lemma 7 in that we must argue that each $g \in \mathcal{G}_i^\star(c_2\Delta_i; y)$ is involved in driving the query rule for a large fraction of the rounds. First observe that $f^\star \in \mathcal{F}_i^\star(c_2\Delta_i)$ for $i \ge 1$ by the definition of $\mathcal{F}_i^\star$.

Next, fix a label y and a function $g \in \mathcal{G}^\star_{i+1}(c_2\Delta_{i+1}; y)$ for $i \ge 0$. We prove that $g \in \mathcal{G}_{t+1}(\Delta_t)$ for all $t \in \{0,\ldots,i\}$. In search of a contradiction, suppose that $g \notin \mathcal{G}_{t+1}(\Delta_{t+1})$ for some $t \in \{0,\ldots,i\}$. First, since $g \in \mathcal{G}^\star_{i+1}(c_2\Delta_{i+1}; y)$, using the Freedman-style deviation bound in Eq. (23), we have

$$\sum_{j=1}^i M_j \le \frac{3}{2}\sum_{j=1}^i \mathbb{E}_c[M_j] + 10\nu_n \le \left(\frac{3}{2}c_2\kappa + c_0\right)\epsilon_i.$$

Here we also use the definition of $\Delta_{i+1} = \kappa\epsilon_i/i$, $c_0 = 10$, and Fact 1.

At the same time, since $g \notin \mathcal{G}_{t+1}(\Delta_{t+1}; y)$, we know that

$$\Delta_{t+1} < \hat{R}_{t+1}(g) - \hat{R}_{t+1}(g_{t+1}) \le \hat{R}_{t+1}(g) - \hat{R}_{t+1}(g^\star) + \frac{c_0\nu_n}{t}.$$

The last inequality uses Eq. (14). Together with the above, this implies that

$$\sum_{j=t+1}^i M_j \le \left(\frac{3}{2}c_2\kappa + c_0\right)\epsilon_i - \kappa\epsilon_t + c_0\nu_n.$$

Now, since $i \ge t$ and $\beta \in (0,1)$, we get that $\epsilon_i < \epsilon_t$. Using Eq. (22) as before, we get

$$\sum_{j=t+1}^i \mathbb{E}_c[M_j] \le 2\sum_{j=t+1}^i M_j + 20\nu_n \le 2\left(\frac{3}{2}c_2\kappa + c_0\right)\epsilon_i - 2\kappa\epsilon_t + 4c_0\nu_n \le (-\kappa + 6c_0)\epsilon_t < 0.$$

The last non-strict inequality follows from the fact that $\epsilon_t \ge \epsilon_i \ge \nu_n$ since $i \ge t$, and then the strict inequality is by our choices for the constants. This is a contradiction since the left hand side is a quadratic form and so, $g \in \mathcal{G}_{t+1}(\Delta_{t+1})$ for all $t \in \{0,\ldots,i\}$.

This argument applies for all y, and hence, we may apply Lemma 7, so that for all regressors $f \in \mathcal{F}_i^\star(c_2\Delta_{i+1})$,

$$\begin{aligned} i\cdot\left(\mathbb{E}_{x,c}[c(x,h_f(x)) - c(x;h_{f^\star}(x))]\right) &\le \min_{\zeta>0}\left\{i\zeta P_\zeta + \sum_{j=1}^i\left(\mathbb{1}\left(\zeta \le 2\eta_j\right)2\eta_j + \frac{4\eta_j^2}{\zeta} + \frac{6}{\zeta}\sum_y \mathbb{E}_j\left[M_j(f;y)\right]\right)\right\} \\ &\le \min_{\zeta>0}\left\{i\zeta P_\zeta + \frac{16+4\log(i)}{\zeta} + \frac{6}{\zeta}\sum_y\sum_{j=1}^i \mathbb{E}_j\left[M_j(f;y)\right]\right\}. \end{aligned}$$

The last inequality here uses identical bounds as the proof of Theorem 3.

In a similar way to (17), we use Lemma 5 to obtain

$$\begin{aligned}\sum_{j=1}^{i}\mathbb{E}_j[M_j(f;y)] &\le 2\sum_{j=1}^{i}M_j(f;y) + 20\nu_n = 2i\cdot\left(\widehat{R}_{i+1}(f;y) - \widehat{R}_{i+1}(f^\star;y)\right) + 20\nu_n\\ &\le 2i\cdot\left(\widehat{R}_{i+1}(f;y) - \widehat{R}_{i+1}(f_{i+1};y)\right) + 20\nu_n\\ &\le (2\kappa+20)\epsilon_i.\end{aligned}$$

The last bound uses the definition of Δ_{i+1} and Fact 1, along with the fact that $\mathcal{G}^\star_{i+1}(c_2\Delta_{i+1};y) \subset \mathcal{G}_{i+1}(\Delta_{i+1};y)$ so we know the empirical risk to f_{i+1} is controlled. Finally, we collect the latter three terms and the constant $6(2\kappa+20)+20$ (which requires $\delta < 1/e$). This gives,

$$\mathbb{E}_{x,c}\left[c(x,h_f(x)) - c(x,h_{f^\star}(x))\right] \le \min_{\zeta>0}\left\{\zeta P_\zeta + \frac{14\kappa K\epsilon_i}{i\zeta}\right\}.$$

This proves the statement since we are considering $f \in \mathcal{F}^\star_{i+1}(c_2\Delta_{i+1})$ and $\kappa\epsilon_i/i = \Delta_{i+1}$. □

For the rest of the analysis, it will be convenient to introduce the shorthand $\widehat{\gamma}(x_i,y) = \widehat{c}_+(x_i,y) - \widehat{c}_-(x_i,y)$, where $\widehat{c}_+(x_i,y)$ and $\widehat{c}_-(x_i,y)$ are the approximate maximum and minimum costs computed in Algorithm 1 at round i. Moreover, let Y_i be the set of non-dominated labels at round i of the algorithm, which in the pseudocode we call Y'. Formally, $Y_i = \{y \mid \widehat{c}_-(x_i,y) \le \min_{y'}\widehat{c}_+(x_i,y')\}$.

Lemma 10 (Cost Range Translation)**.** *Fix i and suppose that the conclusions of Lemmas 8 and 9 hold. Then for any x,y pair, we have*

$$\widehat{\gamma}(x_i,y) \le \gamma(x_i,y,\mathcal{F}_{csr}(r_{I_\beta(i)})) + \eta_i/2,$$

where $r_i = \min_{\zeta>0}\left\{\zeta P_\zeta + \frac{14K\Delta_i}{\zeta}\right\}$ and $I_\beta(i)$ is in Eq. (20).

Proof. We have

$$\begin{aligned}\widehat{\gamma}(x_i,y) &\le \gamma(x_i,y,\mathcal{G}^\star_i(c_1\Delta_i;y)) + \frac{\eta_i}{2} && \text{(By Theorem 1, setting of } \epsilon \text{ in Algorithm 1 and Lemma 8)}\\ &\le \gamma(x_i,y,\mathcal{F}_{\text{csr}}(r_{I_\beta(i)})) + \frac{\eta_i}{2} && \text{(By Lemmas 8 and 9)}\end{aligned}$$

□

Lemma 11. *Fix i and suppose that the conclusions of Lemmas 8 and 9 hold. Define $y^\star_i = \operatorname{argmin}_y f^\star(x_i;y), \bar{y}_i = \operatorname{argmin}_y \widehat{c_+}(x_i,\mathcal{G}_i(y)), \tilde{y}_i = \operatorname{argmin}_{y\neq y^\star_i}\widehat{c_-}(x_i,\mathcal{G}_i(y))$. Then for $y \neq y^\star_i$, we have*

$$y \in Y_i \Rightarrow f^\star(x_i;y) - f^\star(x_i;y^\star_i) - \frac{\eta_i}{2} \le \left(\gamma(x_i,y) + \gamma(x_i,y^\star_i)\right),$$

and for $y^\star_i$:

$$|Y_i| > 1 \ \wedge\ y^\star_i \in Y_i \Rightarrow f^\star(x_i;\tilde{y}_i) - f^\star(x_i;y^\star_i) - \frac{\eta_i}{2} \le \left(\gamma(x_i,\tilde{y}_i) + \gamma(x_i,y^\star_i)\right).$$

In both bounds, all the cost ranges are computed using $\mathcal{F}_{csr}(r_{I_\beta(i)})$.

Proof. Suppose $y \neq y^\star_i$

$$\begin{aligned}y \in Y_i &\Rightarrow \widehat{c_-}(x_i,\mathcal{G}_i(y)) \le \widehat{c_+}(x_i,\mathcal{G}_i(\bar{y}_i))\\ &\Rightarrow \widehat{c_-}(x_i,\mathcal{G}_i(y)) \le \widehat{c_+}(x_i,\mathcal{G}_i(y^\star_i))\\ &\Rightarrow c_-(x_i,\mathcal{G}^\star_i(c_1\Delta_i;y)) \le c_+(x_i,\mathcal{G}^\star_i(c_1\Delta_i;y^\star_i)) + \frac{\eta_i}{2}\\ &\Rightarrow f^\star(x_i;y) - \gamma(x_i,\mathcal{G}^\star_i(c_1\Delta_i;y)) \le f^\star(x_i;y^\star_i) + \gamma(x_i,\mathcal{G}^\star_i(c_1\Delta_i;y^\star_i)) + \frac{\eta_i}{2}\\ &\Rightarrow f^\star(x_i;y) - f^\star(x_i;y^\star_i) - \frac{\eta_i}{2} \le \gamma(x_i,\mathcal{G}^\star_i(c_1\Delta_i;y)) + \gamma(x_i,\mathcal{G}^\star_i(c_1\Delta_i;y^\star_i))\\ &\Rightarrow f^\star(x_i;y) - f^\star(x_i;y^\star_i) - \frac{\eta_i}{2} \le \left(\gamma(x_i,y,\mathcal{F}_{\text{csr}}(r_{I_\beta(i)})) + \gamma(x_i,y^\star_i,\mathcal{F}_{\text{csr}}(r_{I_\beta(i)}))\right).\end{aligned}$$

For $y_i^\star$ we need to consider two cases. First assume $y_i^\star = \bar{y}_i$. Then

$$\begin{aligned} |Y_i| > 1 \wedge y_i^\star \in Y_i \wedge y_i^\star = \bar{y}_i &\Rightarrow \widehat{c_-}(x_i, \mathcal{G}_i(\tilde{y}_i)) \le \widehat{c_+}(x_i, \mathcal{G}_i(y_i^\star)) \\ &\Rightarrow f^\star(x_i, \tilde{y}_i) - f^\star(x_i, y_i^\star) - \frac{\eta_i}{2} \le \gamma(x_i, \tilde{y}_i) + \gamma(x_i, y_i^\star). \end{aligned}$$

This is true since if $|Y_i| > 1$ then it must be the case that $\tilde{y}_i$ is confused, since it has the minimal lower cost estimate. On the other hand if $y_i^\star \neq \bar{y}_i$ then

$$\begin{aligned} |Y_i| > 1 \wedge y_i^\star \in Y_i \wedge y_i^\star \neq \bar{y}_i &\Rightarrow \widehat{c_+}(x_i, \mathcal{G}_i(\bar{y}_i)) \le \widehat{c_+}(x_i, \mathcal{G}_i(y_i^\star)) \\ &\Rightarrow \widehat{c_-}(x_i, \mathcal{G}_i(\tilde{y}_i)) \le \widehat{c_+}(x_i, \mathcal{G}_i(y_i^\star)) \\ &\Rightarrow f^\star(x_i, \tilde{y}_i) - f^\star(x_i, y_i^\star) - \frac{\eta_i}{2} \le \gamma(x_i, \tilde{y}_i) + \gamma(x_i, y_i^\star). \end{aligned}$$

The second step here is because the search for $\tilde{y}_i$ includes $\bar{y}_i$, since the latter is not $y_i^\star$. Thus we obtain

$$\begin{aligned} |Y_i| > 1 \ \wedge \ y_i^\star \in Y_i &\Rightarrow \widehat{c_-}(x_i, \mathcal{G}_i(\tilde{y}_i)) \le \widehat{c_+}(x_i, \mathcal{G}_i(y_i^\star)) \\ &\Rightarrow f^\star(x_i; \tilde{y}_i) - f^\star(x_i; y_i^\star) - \frac{\eta_i}{2} \le \left(\gamma(x_i, \tilde{y}_i) + \gamma(x_i, y_i^\star)\right), \end{aligned}$$

as desired. □

D.2. Low Noise (Massart) Case (Theorem 6)

Fix some round i. Let $\mathcal{F}_i$ be the set of vector regressors used at round i of COAL and let $\mathcal{G}_i(y)$ be the corresponding regressors for label y. Let $\bar{y}_i \triangleq \mathrm{argmin}_y\, \widehat{c_+}(x_i, \mathcal{G}_i(y))$, $y_i^\star = \mathrm{argmin}_y\, f^\star(x_i; y)$, and $\tilde{y}_i \triangleq \mathrm{argmin}_{y \neq y_i^\star}\, \widehat{c_-}(x_i, \mathcal{G}_i(y))$. Assume Lemmas 8 and 9 hold. The label complexity L_2 for round i is

$$\sum_y Q_i(y) = \sum_y \mathbb{1}\left(|Y_i| > 1 \wedge y \in Y_i\right) \mathbb{1}\left(\widehat{\gamma}(x_i, y, \mathcal{F}_i) > \eta_i\right) = \sum_y \mathbb{1}\left(|Y_i| > 1 \wedge y \in Y_i\right) Q_i(y).$$

We need to do two things with $Q_i(y)$, so we have duplicated it here. First, observe that $y \in Y_i$ implies that there exists a vector regressor $f \in \mathcal{F}_i$ such that $h_f(x_i) = y$. This follows since the domination condition means that there exists $g \in \mathcal{G}_i(y)$ such that $g(x_i) \le \min_{y' \neq y} \max_{g' \in \mathcal{G}_i(y')} g'(x_i)$. Since we are using a factored representation, we can take f to use g on the y^{th} coordinate and use the maximizers for all the other coordinates. Moreover, $|Y_i| > 1$ implies there exists a regressor that *does not* predict y. Of course, through Lemmas 8 and 9, we know that $\mathcal{F}_i \subset \mathcal{F}_{\text{csr}}(r_{I_\beta(i)})$, and so we get the bound:

$$\mathbb{1}\left(|Y_i| > 1 \wedge y \in Y_i\right) \le \mathbb{1}\left(\exists f, f' \in \mathcal{F}_{\text{csr}}(r_{I_\beta(i)}) \mid h_f(x_i) = y \wedge h_{f'}(x_i) \neq y\right).$$

For $y \neq y_i^\star$, we take f' to be $f^\star$ which is always in the cost-sensitive regret ball. For $y_i^\star$, we take f' to be any regressor such that $h_{f'}(x_i) = \tilde{y}_i$, which must exist in the ball if $|Y_i| > 1$. We will use these as an upper bound on $Q_i(y)$ momentarily.

Secondly, we apply Lemma 11 along with the Massart noise assumption. For $y \neq y_i^\star$

$$\begin{aligned} \mathbb{1}\left(|Y_i| > 1 \wedge y \in Y_i\right) &\le \mathbb{1}\left(f^\star(x_i; y) - f^\star(x_i; y_i^\star) - \frac{\eta_i}{2} \le \gamma(x_i, y) + \gamma(x_i, y_i^\star)\right) \\ &\le \mathbb{1}\left(\tau - \frac{\eta_i}{2} \le \gamma(x_i, y) + \gamma(x_i, y_i^\star)\right). \end{aligned}$$

Recall that we use the convention that all quantities without an explicit regressor ball use $\mathcal{F}_{\text{csr}}(r_{I_\beta(i)})$. For $y_i^\star$ we obtain the same inequality but using $\tilde{y}_i$ via Lemma 11. Together this gives the bound:

$$L_2 \le \sum_{y \neq y_i^\star} \mathbb{1}\left(\tau - \eta_i/2 \le \gamma(x_i, y) + \gamma(x_i, y_i^\star)\right) \times Q_i(y) + \mathbb{1}\left(\tau - \eta_i/2 \le \gamma(x_i, \tilde{y}_i) + \gamma(x_i, y_i^\star)\right) \times Q_i(y_i^\star).$$

Let us focus on just one of these terms (say where $y \neq y_i^\star$) and consider any round i where $\tau \ge 2\eta_i$.

$$\begin{aligned} \mathbb{1}\left(\tau - \eta_i/2 \le \gamma(x_i, y) + \gamma(x_i, y_i^\star)\right) Q_i(y) &\le \mathbb{1}\left(\tau/2 \le \gamma(x_i, y) + \gamma(x_i, y_i^\star)\right) Q_i(y) \\ &\le \mathbb{1}\left(\tau/4 \le \gamma(x_i, y)\right) Q_i(y) + \mathbb{1}\left(\tau/4 \le \gamma(x_i, y_i^\star)\right) Q_i(y). \end{aligned}$$

Using the upper bound on $Q_i(y)$, the first term here is clearly bounded by

$$\mathbb{1}\left(\tau/4 \leq \gamma(x_i, y)\right) \mathbb{1}\left(\exists f, f' \in \mathcal{F}_{\text{csr}}(r_{I_\beta(i)}) \mid h_f(x_i) = y \wedge h_{f'}(x_i) \neq y\right) \triangleq D_i(y).$$

Fortunately, the second term is bounded in the same way, since we know that $h_{f^\star} \in \mathcal{F}_{\text{csr}}(r_{I_\beta(i)})$, the fact that some f with $h_f(x_i) = y \neq y_i^\star$ exists implies that the second term is at most $D_i(y_i^\star)$.

The last term, which involves $Q_i(y_i^\star)$ is bounded in essentially the same way, since we know that when $|Y_i| > 1$ (which is all we are considering), there exists two functions $f, f' \in \mathcal{F}_i$ such that $h_f(x_i) = \tilde{y}_i$ and $h_{f'}(x_i) = y_i^\star$. Thus we can bound the label complexity at round i by

$$D_i(\tilde{y}_i) + D_i(y_i^\star) + \sum_{y \neq y_i^\star} \left(D_i(y) + D_i(y_i^\star)\right) \leq K D_i(y_i^\star) + 2\sum_y D_i(y).$$

For the rounds i where $\tau < 2\eta_i$ we simply upper bound the label complexity by K.

The last step in the proof is to apply Freedman's inequality to the sequence of indicators. The conditional mean of each term is at most (for rounds i where $\tau > 2\eta_i$),

$$\mathbb{E}_i\left[K D_i(y_i^\star) + 2\sum_y D_i(y)\right] \leq \frac{4 r_{I_\beta(i)}}{\tau}\left[K\theta_1 + 2\theta_2\right].$$

The part involving θ_2 is straightforward and the premultiplier follows since we are measuring the probability of querying with a cost range parameter of $\tau/4$ and over a cost-sensitive regret ball of radius $r_{I_\beta(i)}$ in $D_i(y)$. To obtain θ_1 we use the fact that if $D_i(y_i^\star) = 1$, then certainly there exists some confused label, namely $y_i^\star$, and hence the indicator in θ_1 is also 1.

The range is $3K$ since $D_i(y) \in \{0, 1\}$ and since the terms are non-negative, the variance is at most the range times the mean. In such cases, Freedman's inequality gives

$$X \leq \mathbb{E}X + 2\sqrt{R\mathbb{E}X\log(1/\delta)} + 2R\log(1/\delta) \leq 2\mathbb{E}X + 3R\log(1/\delta),$$

with probability at least $1 - \delta$ where X is the non-negative random variable with range R and expectation $\mathbb{E}X$. The last step is by the fact that $2\sqrt{ab} \leq a + b$.

In our case, we get that with probability at least $1 - \delta$,

$$\sum_{i=i^\star}^{n} K D_i(y_i^\star) + 2\sum_y D_i(y) \leq \sum_{i=i^\star}^{n} \frac{8 r_{I_\beta(i)}}{\tau}\left[K\theta_1 + 2\theta_2\right] + 9K\log(1/\delta).$$

Here we only consider rounds $i \geq i^\star$ where $i^\star$ is the smallest index such that $\tau < 2\eta_{i^\star}$ and $i^\star \geq i_\beta$ (Recall Fact 2). For the first $i^\star$ rounds, we will upper bound the per-round label complexity by K, so that the overall label complexity is at most

$$\begin{aligned} & K i^\star + \sum_{i=i^\star}^{n} \frac{8 r_{I_\beta(i)}}{\tau}\left[K\theta_1 + 2\theta_2\right] + 9K\log(1/\delta) \\ & \leq K\sum_{i=1}^{n} \mathbb{1}\left(\tau \leq 2\eta_i\right) + K i_\beta + + \sum_{i=i_\beta}^{n} \frac{8 r_{I_\beta(i)}}{\tau}\left[K\theta_1 + 2\theta_2\right] + 9K\log(1/\delta) \end{aligned}$$

Using our choice of $\eta_i = 1/\sqrt{i}$, the first term is at most $K\lceil 4/\tau^2\rceil$. The second term is bounded by Fact 2. The last step is to use the definition of $r_{I_\beta(i)}$ to simplify the sum. Since we are in the Massart noise case, we will set $\zeta = \tau$ in the definition of r_i in Lemma 10. Since $P_\tau = 0$ by the definition of the noise condition, this yields $r_i = 14K\Delta_i/\tau$. Substituting this choice, along with our definition of Δ_i yields

$$\begin{aligned}
\sum_{i=i_\beta}^{n} r_{I_\beta(i)} &= \frac{14\kappa n^\beta K\nu_n}{\tau}\sum_{i=i_\beta}^{n}(I_\beta(i)-1)^{-1-\beta}\\
&\le \frac{14\kappa n^\beta K\nu_n}{\tau}\times\left(2^{(1+\beta)}\times\left(\frac{c_1}{c_2}\right)^{\frac{1+\beta}{\beta}}\sum_{i=i_\beta}^{n}(i-1)^{-1-\beta}\right)\\
&\le \frac{56(c_1/c_2)\kappa n^\beta K\nu_n}{\tau}\left[\left(\frac{c_1}{c_2}\right)^{\frac{1}{\beta}}\sum_{i=2}^{n}(i-1)^{-1}\right]\\
&\le \frac{56(c_1/c_2)\kappa n^\beta K\nu_n}{\tau}\left(\frac{c_1}{c_2}\right)^{\frac{1}{\beta}}(2\times\log(n)).
\end{aligned}$$

Including the extra $O(K)$ term, the overall bound is

$$\begin{aligned}
&K\left(\lceil\frac{4}{\tau^2}\rceil+2(c_1/c_2)^{1/\beta}+1\right)+\frac{8\times 56\times 25\times 2\kappa n^\beta K\nu_n}{\tau^2}\left(\frac{c_1}{c_2}\right)^{\frac{1}{\beta}}\log(n)[K\theta_1+2\theta_2]+9K\log(1/\delta)\\
&\le a_0 25^{1/\beta}\left(\frac{n^\beta K\log(n)\nu_n}{\tau^2}[K\theta_1+2\theta_2]+\frac{K\log(1/\delta)}{\tau^2}\right),
\end{aligned}$$

where a_0 is a universal constant.

For L_1 we can use a very similar argument. First,

$$L_1=\sum_i \mathbb{1}\left(|Y_i|>1\wedge\exists y\in Y_i,\widehat{\gamma}(x_i,y,\mathcal{F}_i)>\eta_i\right)\ \le\sum_i \mathbb{1}\left(|Y_i|>1\wedge\exists y\in Y_i,\gamma(x_i,y,\mathcal{F}_{\text{csr}}(r_{I_\beta(i)}))>\eta_i/2\right).$$

This inequality is an application of Lemma 10. Now as above, we know that,

$$|Y_i|>1\wedge y\in Y_i\Rightarrow\exists f,f'\in\mathcal{F}_{\text{csr}}(r_{I_\beta(i)}),h_f(x_i)=y\wedge h_{f'}(x_i)\neq y,$$

since if $y\in Y_i$ then some classifier must select it, and since $|Y_i|>1$, something else must also be selected. We also know that we can always take f' to be $f^\star$ when $y\neq y_i^\star$. For $y_i^\star$ we can always take the classifier to be the one that predicts $\tilde{y}_i$.

Moreover we also have that when $\tau\ge 2\eta_i$,

$$\begin{aligned}
|Y_i|>1\wedge y\in Y_i&\Rightarrow f^\star(x_i;y)-f^\star(x_i;y_i^\star)-\eta_i/2\le\gamma(x_i,y)+\gamma(x_i,y_i^\star)\\
&\Rightarrow\tau/4\le\gamma(x_i,y)\vee\tau/4\le\gamma(x_i,y_i^\star)
\end{aligned}$$

Thus, putting things together, and considering only rounds where $\tau\ge 2\eta_i$ we get

$$L_1\le\sum_{i=1}^{n}\mathbb{1}\left(\tau<2\eta_i\right)+\sum_{i=1}^{n}\mathbb{1}\left(\exists y\mid\exists f,f'\in\mathcal{F}_{\text{csr}}(r_{I_\beta(i)}),h_f(x_i)=y\wedge h_{f'}(x_i)\neq y\wedge\gamma(x,y)\ge\tau/4\right).$$

Here we dropped the $\gamma(x;y_i^\star)\ge\tau/4$ term from consideration since the term gets included in the existential quantifier when the chosen label $y=y_i^\star$. Now we may apply Freedman's inequality to upper bound L_1 by

$$L_1\le i_\beta+\lceil 4/\tau^2\rceil+2\sum_{i=i_\beta}^{n}\frac{4r_{I_\beta(i)}}{\tau}\theta_1+2\log(1/\delta)\le a_0 25^{1/\beta}\left(\frac{n^\beta K\log(n)\nu_n}{\tau^2}\theta_1+\frac{\log(1/\delta)}{\tau^2}\right),$$

where a_0 is a universal constant.

D.3. High noise case (Theorem 5)

Fix some round i. Let $\mathcal{F}_i$ be the set of vector regressors used at round i of COAL and let $\mathcal{G}_i(y)$ be the corresponding regressors for label y. Let $\bar{y}_i \triangleq \operatorname{argmin}_y \widehat{c_+}(x_i, \mathcal{G}_i(y))$, $y_i^\star = \operatorname{argmin}_y f^\star(x_i; y)$, and $\tilde{y}_i \triangleq \operatorname{argmin}_{y \neq y_i^\star} \widehat{c_-}(x_i, \mathcal{G}_i(y))$. Assume Lemmas 8 and 9 hold. The label complexity L_2 for round i is

$$\sum_y Q_i(y) = \sum_y \mathbb{1}\left(|Y_i| > 1 \wedge y \in Y_i\right) \mathbb{1}\left(\widehat{\gamma}(x_i, y, \mathcal{F}_i) > \eta_i\right).$$

First we apply Lemma 10 on the latter indicator to get

$$\sum_y \mathbb{1}\left(|Y_i| > 1 \wedge y \in Y_i\right) \mathbb{1}\left(\gamma(x_i, y, \mathcal{F}_{\text{csr}}(r_{I_\beta(i)})) \geq \eta_i/2\right).$$

For the former indicator, observe that $y \in Y_i$ implies that there exists a vector regressor $f \in \mathcal{F}_i$ such that $h_f(x_i) = y$. This follows since the domination condition means that there exists $g \in \mathcal{G}_i(y)$ such that $g(x_i) \leq \min_{y'} \max_{g' \in \mathcal{G}_i(y')} g'(x_i)$. Since we are using a factored representation, we can take f to use g on the y^{th} coordinate and use the maximizers for all the other coordinates.

Since $y \in Y_i$ implies there exists $f \in \mathcal{F}_i$ such that $h_f(x_i) = y$, and by Lemmas 8 and 9, we get that $f \in \mathcal{F}_{\text{csr}}(r_{I_\beta(i)})$. Similarly there exists $f' \in \mathcal{F}_i$ such that $h_{f'}(x_i) \neq y$. Thus we can bound the the label complexity for round i as,

$$\begin{aligned}
&\sum_y \mathbb{1}\left(\exists f, f' \in \mathcal{F}_{\text{csr}}(r_{I_\beta(i)}) \mid h_f(x_i) = y \neq h_{f'}(x_i)\right) \mathbb{1}\left(\gamma(x_i, y, \mathcal{F}_{\text{csr}}(r_{I_\beta(i)})) \geq \eta_i/2\right) \\
&= \sum_y \mathbb{1}\left(x \in \text{DIS}(r_{I_\beta(i)}, y) \wedge \gamma(x_i, y, \mathcal{F}_{\text{csr}}(r_{I_\beta(i)})) \geq \eta_i/2\right).
\end{aligned}$$

Now we can apply Freedman's inequality on the sequence here to find that with probability at least $1 - \delta$,

$$L_2 \leq K i_\beta + \sum_{i=i_\beta}^{n} \frac{4 r_{I_\beta(i)}}{\eta_i} \theta_2 + 3K \log(1/\delta)$$

Again $i_\beta = 2(c_1/c_2)^{1/\beta} + 1$ is from Fact 2. We just need to upper bound the sequence:

$$\begin{aligned}
\sum_{i=i_\beta}^{n} \frac{r_{I_\beta(i)}}{\eta_i} &= 2 \sum_{i=i_\beta}^{n} \sqrt{i} \sqrt{\frac{14 K \kappa n^\beta \nu_n}{(I_\beta(i) - 1)^{1+\beta}}} \\
&\leq 2\sqrt{14 K \kappa n^\beta \nu_n} \times \sum_{i=i_\beta}^{n} \sqrt{\frac{2^{1+\beta} i}{(c_2/c_1)^{\frac{1+\beta}{\beta}} (i-1)^{1+\beta}}} \\
&\leq 2\sqrt{14 K \kappa n^\beta \nu_n} \times \sum_{i=i_\beta}^{n} \sqrt{\frac{2^{2+\beta}}{(c_2/c_1)^{\frac{1+\beta}{\beta}} (i-1)^{\beta}}} \\
&\leq \sqrt{448 (c_1/c_2)^{\frac{1+\beta}{\beta}} K \kappa n^\beta \nu_n} \times \sum_{i=1}^{n-1} i^{-\beta/2} \\
&\leq 2\sqrt{448 (c_1/c_2)^{\frac{1+\beta}{\beta}} K \kappa n^\beta \nu_n} \times n^{1-\beta/2} \\
&\leq 2\sqrt{448 (c_1/c_2)^{\frac{1+\beta}{\beta}} K \kappa \nu_n} \times n.
\end{aligned}$$

The first line follows by the definition of η_i and by optimizing the bound in Lemma 9 using the definition of Δ_i. The second line uses Fact 2. The remaining steps are simple calculations using $\beta \in (0, 1)$ and an integral bound.

Thus in total we get a label complexity of

$$L_2 \leq a_0 (25)^{1/\beta} \left(n \theta_2 \sqrt{K \nu_n} + K \log(1/\delta)\right).$$

Similarly for L_1 we can derive the bound

$$L_1 \leq \sum_i \mathbb{1}\left(\exists y \mid \gamma(x_i, y, \mathcal{F}_{\text{csr}}(r_{I_\beta(i)})) \geq \eta_i/2 \wedge x \in \text{DIS}(r_{I_\beta(i)}, y)\right).$$

and then apply Freedman's inequality to this sequence to obtain that with probability at least $1-\delta$

$$L_1 \leq i_\beta + 2\sum_{i=i_\beta}^{n} \frac{2r_{I_\beta}}{\eta_i}\theta_1 + 3\log(1/\delta) \leq a_0(25)^{1/\beta}\left(n\theta_1\sqrt{K\nu_n} + \log(1/\delta)\right).$$