A Experiments continued

In this section, we discuss our experimental setup more thoroughly and present more results. Each plot depicts ≥ 50 independent simulations, and the error bands depict 68% bootstrap confidence intervals. For the NDBAL query selection algorithm, we used the heuristic suggested in Section 6: we sampled $m = 500$ candidate atoms from $\mathcal{D}$ and $n = 300$ pairs of structures from π_t and chose the atom that empirically minimized equation (5).

A.1 Models, sampling, and evaluation

In our experiments, we used the posterior update in equation (4) with $\ell(z, y)$ as the logistic loss, i.e.

$$\ell(z, y) = \log\left(1 + e^{-zy}\right).$$

In this setting, it is not possible to express π_t in closed form. However, we can still approximately sample from π_t using the Metropolis-adjusted Langevin Algorithm (MALA) (Dwivedi et al., 2018). If we let

$$f(w) \;=\; -\sum_{i=1}^{t} \beta\ell(\langle w, x_i\rangle, y_i) - \frac{1}{2\sigma^2}\|w\|^2$$

then MALA is a Markov chain in which we maintain a vector $W_t \in \mathbb{R}^d$ and transition to W_{t+1} according to the following process.

(i) Sample $V \sim \mathcal{N}(W_t - \eta\nabla f(W_t), 2\eta I_d)$.

(ii) Calculate $\alpha = \min\left\{1, \exp\left(f(W_t) - f(V) + \frac{1}{4\eta}\left(\|V - W_T + \eta\nabla f(W_t)\|^2 - \|W_t - V + \eta\nabla f(V)\|^2\right)\right)\right\}$.

(iii) With probability α, $W_{t+1} = V$. Otherwise, set $W_{t+1} = W_t$.

The only hyper-parameter that needs to be set is $\eta > 0$. This parameter should be carefully chosen: if η is too large then the walk may never accept the proposed state, and if η is too small then the walk may not move far enough to get to a large probability region. The best choice of η ultimately depends on the distribution we are sampling from, and unfortunately for us, our distributions are changing. Our fix is to adjust η on the fly so that the average number of times that step (iii) rejects is not too close to 0 or to 1. A reasonable rejection rate is about 0.4 (Roberts and Rosenthal, 1998).

Finally, in all of our evaluations we recorded an approximation of the *average* error of the posterior distribution π_t. This consists of sampling structures $g_1, \ldots, g_n \sim \pi_t$ and calculating

$$\widehat{\text{error}}(\pi_t) \;=\; \frac{1}{n}\sum_{i=1}^{n} d(g_i, g^*)$$

where $d(\cdot, \cdot)$ is the distance function for the task at hand. In our experiments, this distance takes the following forms.

- Classification error: $d(w, w') = \Pr_{x\sim\text{unif}(\mathcal{S}^{d-1})}(\text{sign}(\langle w, x\rangle) \neq \text{sign}(\langle w^*, x\rangle)) = \frac{1}{\pi}\arccos\left(\frac{\langle w, w'\rangle}{\|w\|\|w'\|}\right)$.
- Best item identification: $d(w, w') = \mathbb{1}[i_w \neq i_{w'}]$.
- Approximate best item identification: $d(w, w') = \|x_{i_w} - x_{i_{w'}}\|$.

In the above, $i_w = \arg\max_i \langle w, x_i\rangle$ is the top item under w in the choice model setting. We used $n = 300$ in our experiments.

A.2 Classification experiments

In Figure 2, we have classification experiments under logistic noise across different dimensions d and standard deviations σ. In all of the experiments, we used the logistic loss update on the posterior with $\beta = 1$ and a prior distribution of $\mathcal{N}(0, \sigma^2 I_d)$.

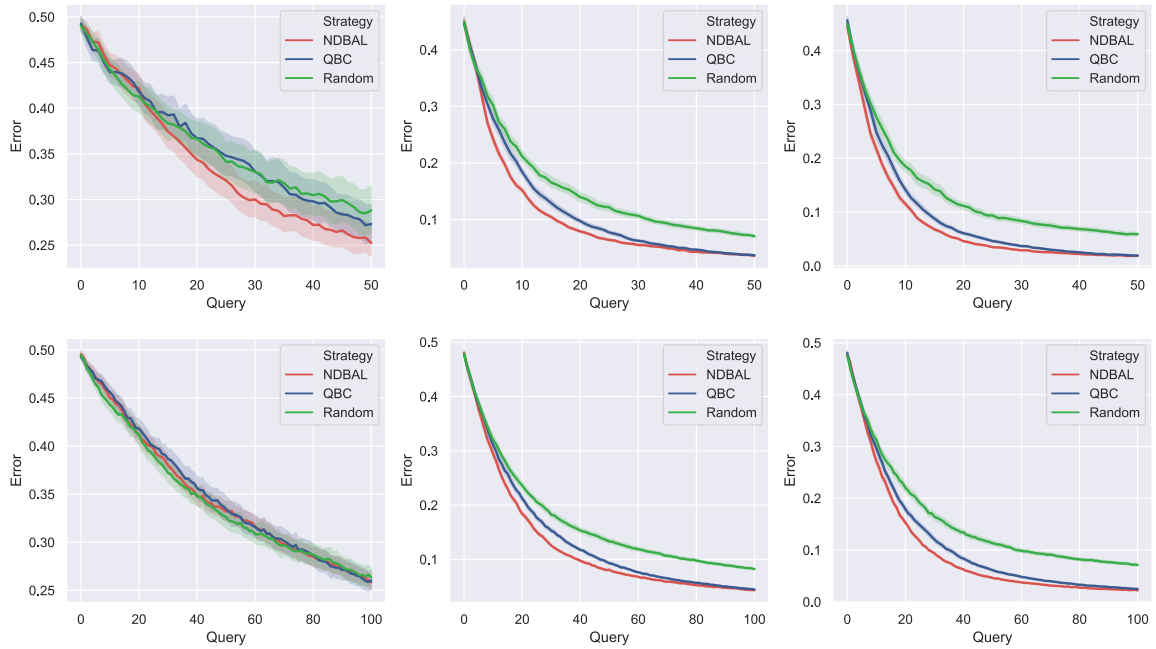

Figure 2: Logistic noise experiments. *Top to bottom*: $d = 5, 10$. *Left to right*: $\sigma = 1$, 5, 10.

A.3 Logit choice model experiments

In Figure 3, we have logit choice model experiments across different dimensions d, numbers of items n, and standard deviations σ. In all of the experiments, we used the logistic loss update on the posterior with $\beta = 1$ and a prior distribution of $\mathcal{N}(0, \sigma^2 I_d)$.

B Dasgupta's splitting index

We will make use of the original splitting index of Dasgupta (2005) and its multiclass extension in Balcan and Hanneke (2012). Let $E = ((g_1, g_1'), \ldots, (g_n, g_n'))$ be a sequence of structure pairs. We say that an atom a ρ-splits E if

$$\max_y |E_a^y| \;\leq\; (1-\rho)|E|.$$

$\mathcal{G}$ has splitting index (ρ, ϵ, τ) if for any edge sequence E such that $d(g, g') > \epsilon$ for all $(g, g') \in E$, we have

$$\Pr_{a\sim\mathcal{D}}(a \text{ } \rho\text{-splits } E) \geq \tau.$$

The following theorem, which we will use heavily, demonstrates that the average splitting index can be bounded by the splitting index. It is analogous to Lemma 3 of Tosh and Dasgupta (2017).

Theorem 14. *Fix $\mathcal{G}$, $\mathcal{D}$, and π. If $\mathcal{G}$ has splitting index index (ρ, ϵ, τ) then it has average splitting index* $(\frac{\rho}{4\lceil \log_2 1/\epsilon \rceil}, 2\epsilon, \tau)$.

From the proof of Lemma 3 by Tosh and Dasgupta (2017), it is easy to see that so long as $d(\cdot, \cdot)$ is symmetric and takes values in $[0, 1]$, the same arguments imply Theorem 14.

C Proofs from Section 3

C.1 Proof of Lemma 2

To prove Lemma 2, we will appeal to the following multiplicative Chernoff-Hoeffding bound (Angluin and Valiant, 1977).

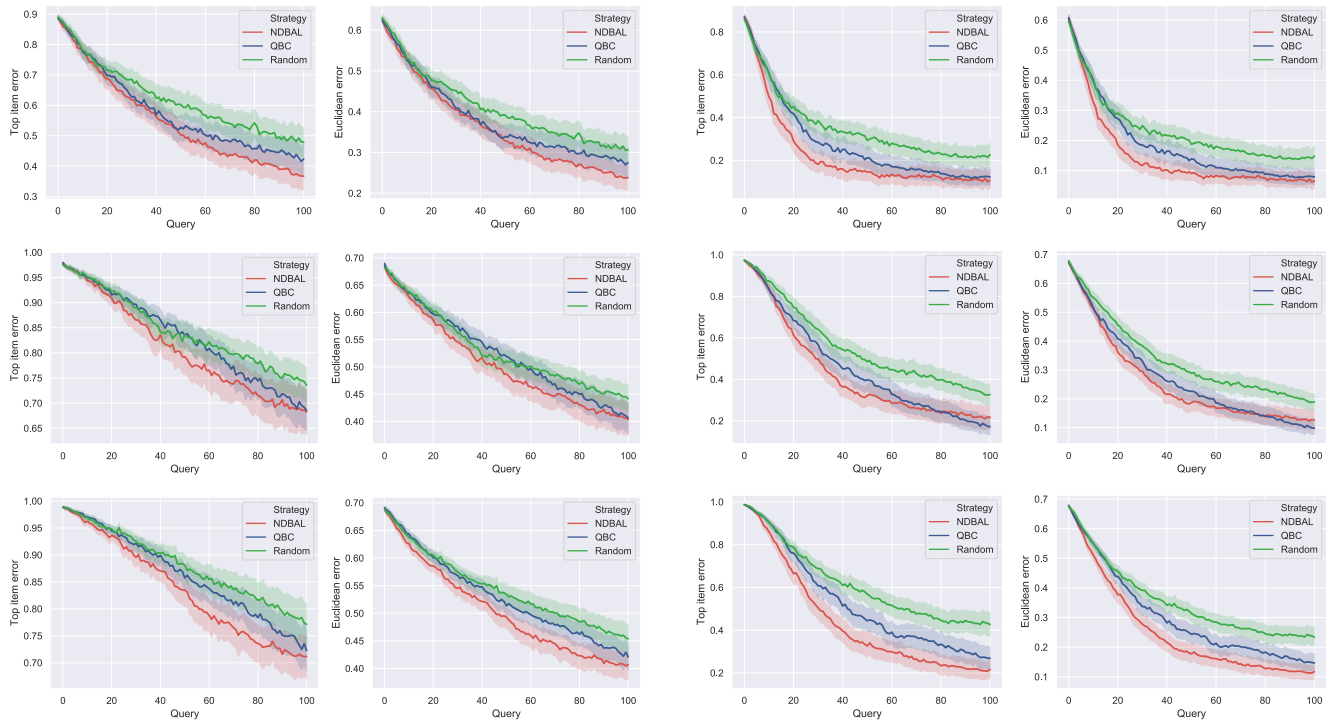

Figure 3: Logit choice model experiments with $d = 10$. *Top to bottom*: $n = 10, 50, 100$. *Left to right*: $\sigma = 1, 5$.

Lemma 15. *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $[0, 1]$ and let $X = \sum X_i$ and $\mu = \mathbb{E}[X]$. Then for $0 < \beta < 1$,*

(i) $\Pr(X \leq (1-\beta)\mu) \leq \exp\left(-\frac{\beta^2\mu}{2}\right)$ *and*

(ii) $\Pr(X \geq (1+\beta)\mu) \leq \exp\left(-\frac{\beta^2\mu}{3}\right)$.

The key observation in proving Lemma 2 is that if a ρ-average splits π, then for all $y \in \mathcal{Y}$ we have

$$\text{avg-diam}(\pi) - \pi(\mathcal{G}_a^y)^2 \text{avg-diam}(\pi|_{\mathcal{G}_a^y}) \;\geq\; \rho \,\text{avg-diam}(\pi).$$

On the other hand, if a does not ρ-average split π, then there is some $y \in \mathcal{Y}$ such that

$$\text{avg-diam}(\pi) - \pi(\mathcal{G}_a^y)^2 \text{avg-diam}(\pi|_{\mathcal{G}_a^y}) \;<\; \rho \,\text{avg-diam}(\pi).$$

Moreover, if $g, g' \sim \pi$, then

$$\mathbb{E}[d(g, g')(1 - \mathbb{1}[g(a) = y = h'(a)])] \;=\; \text{avg-diam}(\pi) - \pi(\mathcal{G}_a^y)^2 \text{avg-diam}(\pi|_{\mathcal{G}_a^y}).$$

Using these facts, along with Lemma 15, we have the following result.

Lemma 2. *Pick $\alpha, \delta > 0$. If* SELECT *is run with atoms $a_1, \ldots, a_m$, one of which ρ-average splits π, then with probability $1 - \delta$,* SELECT *returns a data point that $(1-\alpha)\rho$-average splits π while sampling no more than*

$$\frac{12}{\alpha^2(1-\alpha)\rho \,\text{avg-diam}(\pi)} \log \frac{m + |\mathcal{Y}|}{\delta}$$

pairs of structures in total.

Proof. Define $K_N^{a,y} = \inf\{K : S_K^{a,y} \geq N\}$. Recalling that $S_k^{a,y} = \sum_{i=1}^k d(g_i, g_i')(1 - \mathbb{1}[g_i(a) = y = g_i'(a)])$, we have the following relationship between $K_N^{a,y}$ and $S_k^{a,y}$.

$$\Pr(K_N^{a,y} \leq k) \;=\; \Pr(S_{k_o}^{a,y} \geq N \text{ for some } k_o \leq k) \;\leq\; \Pr(S_k^{a,y} \geq N)$$
$$\Pr(K_N^{a,y} > k) \;=\; \Pr(S_{k_o}^{a,y} < N \text{ for all } k_o \leq k) \;=\; \Pr(S_k^{a,y} < N)$$

Now let a^* be the atom that ρ-average splits π. Then for all $y \in \mathcal{Y}$, we have

$$\Pr\left(K_N^{a^*,y} > \frac{N}{(1+\epsilon/2)(1-\epsilon)\rho\,\text{avg-diam}(\pi)}\right) \;\le\; \exp\left(-\frac{N\epsilon^2(1+\epsilon)^2}{8(1-\epsilon(1+\epsilon)/2)}\right).$$

On the other hand we know for any data point a that does not $(1-\epsilon)\rho$-average split π, there is some $y \in \mathcal{Y}$ such that

$$\Pr\left(K_N^{a,y} \le \frac{N}{(1+\epsilon/2)(1-\epsilon)\rho\,\text{avg-diam}(\pi)}\right) \;\le\; \exp\left(-\frac{N\epsilon^2}{12(1-\epsilon/2)}\right).$$

Taking a union bound over $\mathcal{Y}$ and all the a's, we have

$$\Pr\left(\text{we choose } a_i \text{ that does not } (1-\epsilon)\rho\text{-average split } \pi\right) \;\le\; |\mathcal{Y}|\exp\left(-\frac{N\epsilon^2}{4(2-\epsilon)}\right) + m\exp\left(-\frac{N\epsilon^2}{6(2+\epsilon)}\right).$$

By our choice of N, this is less than δ. □

D Proofs from Section 4

D.1 Proof of Lemma 3

Lemma 3. *Pick $k \ge 2$. Suppose Assumption 2 holds and $\beta \le \lambda/(2+2k^2)$. If we query an atom a_t that ρ-average splits π_{t-1}, then in expectation over the randomness of the response y_t, we have*

$$\mathbb{E}\left[\frac{\text{avg-diam}(\pi_t)}{\pi_t(g^*)^k}\,\middle|\,\mathcal{F}_{t-1}, a_t\right] = (1-\Delta)\,\frac{\text{avg-diam}(\pi_{t-1})}{\pi_{t-1}(g^*)^k}$$

where $\Delta \ge \rho\lambda\beta/2$.

Proof. To simplify notation, take $\pi = \pi_{t-1}$. Suppose that we query $a \in \mathcal{A}$. Enumerate the potential responses as $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$. The definition of average splitting implies that there exists a symmetric matrix $R \in [0,1]^{m\times m}$ satisfying

- $R_{ii} \le 1-\rho$ for all i,
- $\sum_{i,j} R_{ij} = 1$, and
- $R_{ij}\,\text{avg-diam}(\pi) = \sum_{g\in\mathcal{G}_a^{y_i}, g'\in\mathcal{G}_a^{y_j}} \pi(g)\pi(g')d(g,g')$.

Let us assume w.l.o.g. that $g^*(a) = y_1$. Define the quantity

$$Q_a^i \;:=\; \pi(G_a^{y_i}) + e^{-\beta}\sum_{j\ne i}\pi(G_a^{y_j}) \;=\; \pi(G_a^{y_i}) + e^{-\beta}(1-\pi(G_a^{y_i})) \;\le\; 1.$$

We now derive the form of avg-diam(π_t). In the event that $y_t = i$, we have

$$\begin{aligned}
\text{avg-diam}(\pi_t) &= \sum_{h,h'\in\mathcal{H}} \pi_t(h)\pi_t(h')d(h,h') \\
&= \left(\frac{1}{Q_a^i}\right)^2\left(\sum_{g,g'\in\mathcal{G}_a^{y_i}}\pi(g)\pi(g')d(g,g') + 2e^{-\beta}\sum_{j\ne i}\sum_{g\in\mathcal{G}_a^{y_1},g'\in\mathcal{G}_a^{y_j}}\pi(g)\pi(g')d(g,g')\right. \\
&\qquad\left. + e^{-2\beta}\sum_{j\ne i,k\ne i}\sum_{g\in\mathcal{G}_a^{y_j},g'\in\mathcal{G}_a^{y_k}}\pi(g)\pi(g')d(g,g')\right) \\
&= \left(\frac{1}{Q_a^i}\right)^2\left(R_{ii} + 2e^{-\beta}\sum_{j\ne i}R_{ij} + e^{-2\beta}\sum_{j\ne i,k\ne i}R_{jk}\right)\text{avg-diam}(\pi)
\end{aligned}$$

$$= \left(\frac{1}{Q_a^i}\right)^2 \left(R_{ii} + 2e^{-\beta}\sum_{j\neq i} R_{ij} + e^{-2\beta}\left(1 - R_{ii} - 2\sum_{j\neq i} R_{ij}\right)\right) \text{avg-diam}(\pi)$$
$$= \left(\frac{1}{Q_a^i}\right)^2 \left(e^{-2\beta} + (1-e^{-2\beta})R_{ii} + 2(e^{-\beta} - e^{-2\beta})\sum_{j\neq i} R_{ij}\right) \text{avg-diam}(\pi).$$

We can also derive the form of $\frac{1}{\pi_t(g^*)^k}$:

$$\frac{1}{\pi_t(g^*)^k} \;=\; \begin{cases} \left(\frac{Q_a^1}{\pi(g^*)}\right)^k & \text{if } y_t = y_1 \\ \left(\frac{Q_a^i}{e^{-\beta}\pi(g^*)}\right)^k & \text{if } y_t = y_i \neq y_1 \end{cases}$$

Define

$$\Delta_t \;:=\; \frac{\pi(g^*)^k}{\text{avg-diam}(\pi)} \cdot \mathbb{E}\left[\frac{\text{avg-diam}(\pi_t)}{\pi_t(g^*)^k}\right].$$

If we take $\eta(y_i|a) = \gamma_i$ and assume w.l.o.g. that $\gamma_1 > \gamma_2 \geq \gamma_3 \geq \cdots$, then

$$\begin{aligned}
\Delta_t &= \gamma_1 (Q_a^1)^{k-2}\left(e^{-2\beta} + (1-e^{-2\beta})R_{11} + 2(e^{-\beta} - e^{-2\beta})\sum_{j\neq 1} R_{1j}\right) \\
&\quad + \sum_{i\geq 2} \gamma_i (Q_a^1)^{k-2} e^{k\beta}\left(e^{-2\beta} + (1-e^{-2\beta})R_{ii} + 2(e^{-\beta} - e^{-2\beta})\sum_{j\neq i} R_{ij}\right) \\
&\leq (1-\gamma_1)e^{(k-2)\beta} + \gamma_1\left(e^{-2\beta} + (1-e^{-2\beta})R_{11} + 2(e^{-\beta} - e^{-2\beta})\sum_{j\neq 1} R_{1j}\right) \\
&\quad + \gamma_2\left((e^{k\beta} - e^{(k-2)\beta})\sum_{i\geq 2} R_{ii} + 2(e^{(k-1)\beta} - e^{(k-2)\beta})\sum_{i\geq 2}\sum_{j\neq i} R_{ij}\right) \\
&\leq (1-\gamma_1)e^{(k-2)\beta} + \gamma_1(1-e^{-2\beta})R_{11} + \gamma_2(e^{k\beta} - e^{(k-2)\beta})\sum_{i\geq 2} R_{ii} \\
&\quad + \left(\gamma_1(e^{-\beta} - e^{-2\beta}) + \gamma_2(e^{(k-1)\beta} - e^{(k-2)\beta})\right)\left(1 - \sum_{i\geq 1} R_{ii}\right)
\end{aligned}$$

Using the inequalities $1 + x \leq e^x \leq 1 + x + x^2$ for $|x| \leq 1$ and Assumption 2, we can verify that the following inequalities hold for our choice of β:

$$\begin{gathered}
\gamma_2(e^{k\beta} - e^{(k-2)\beta}) \;\leq\; \gamma_1(e^{-\beta} - e^{-2\beta}) + \gamma_2(e^{(k-1)\beta} - e^{(k-2)\beta}) \;\leq\; \gamma_1(1-e^{-2\beta}) \\
(1-\gamma_1)e^{(k-2)\beta} + \gamma_1(1-e^{-2\beta}) \;\leq\; 1 \\
\gamma_1(1-e^{-\beta}) + \gamma_2(e^{(k-1)\beta} - e^{(k-2)\beta}) \;\leq\; -\beta\lambda/2
\end{gathered}$$

Using our restrictions on the structure of R, the above inequalities imply

$$\begin{aligned}
\Delta_t &\leq (1-\gamma_1)e^{(k-2)\beta} + (1-\rho)\gamma_1(1-e^{-2\beta}) + \rho\left(\gamma_1(e^{-\beta} - e^{-2\beta}) + \gamma_2(e^{(k-1)\beta} - e^{(k-2)\beta})\right) \\
&= (1-\gamma_1)e^{(k-2)\beta} + \gamma_1(1-e^{-2\beta}) + \rho\left(\gamma_1(1-e^{-\beta}) + \gamma_2(e^{(k-1)\beta} - e^{(k-2)\beta})\right) \\
&\leq 1 + \rho\left(\gamma_1(1-e^{-\beta}) + \gamma_2(e^{(k-1)\beta} - e^{(k-2)\beta})\right) \\
&\leq 1 - \rho\lambda\beta/2.
\end{aligned}$$

□

D.2 Proof of Lemma 4

Lemma 4. *Pick $k \geq 1$. Suppose Assumption 2 holds and $\beta \leq \lambda/k$. Then for any query a_t, we have $\mathbb{E}\left[1/\pi_t(g^*)^k \,|\, \mathcal{F}_{t-1}, a_t\right] \leq 1/\pi_{t-1}(g^*)^k$.*

Proof. Suppose we query a at step t. Denote by $\gamma_i = \eta(y_i \,|\, a)$ and $\pi_i = \pi_{t-1}(\mathcal{G}_a^{y_i})$, and assume w.l.o.g that $g^*(a) = y_1$ and $\gamma_1 > \gamma_2 \geq \gamma_3 \geq \cdots$. Then we have

$$\mathbb{E}\left[\frac{1}{\pi_t(g^*)^k} \mid \pi_{t-1}(g^*)\right] = \frac{\gamma_1(\pi_1 + e^{-\beta}(1-\pi_1))^k}{\pi_{t-1}(g^*)^k} + \sum_{i\geq 2}\frac{\gamma_i(e^{\beta}\pi_i + 1 - \pi_i)^k}{\pi_{t-1}(g^*)^k}$$
$$= \frac{1}{\pi_{t-1}(g^*)^k}\left(\gamma_1(\pi_1 + e^{-\beta}(1-\pi_1))^k + \sum_{i\geq 2}\gamma_i(e^{\beta}\pi_i + 1 - \pi_i)^k\right)$$

Denote the term in parenthesis by Δ_t. Using the inequalities $1 + x \leq e^x \leq 1 + x + x^2$ for $|x| \leq 1$, for our choice of β we have

$$\begin{aligned}
\Delta_t &\leq \gamma_1(\pi_1 + (1 - \beta + \beta^2)(1-\pi_1))^k + \sum_{i\geq 2}\gamma_i((1+\beta+\beta^2)\pi_i + 1 - \pi_i)^k \\
&= \gamma_1(1 - \beta(1-\beta)(1-\pi_1))^k + \sum_{i\geq 2}\gamma_i(1 + \pi_i\beta(1+\beta))^k \\
&\leq \gamma_1 \exp(-k\beta(1-\beta)(1-\pi_1)) + \sum_{i\geq 2}\gamma_i \exp(k\pi_i\beta(1+\beta)) \\
&\leq \gamma_1(1 - k\beta(1-\beta)(1-\pi_1) + (k\beta(1-\beta)(1-\pi_1))^2) + \sum_{i\geq 2}\gamma_i(1 + k\pi_i\beta(1+\beta) + (k\pi_i\beta(1+\beta))^2) \\
&= 1 + k\beta\left((1+\beta)\sum_{i\geq 2}\gamma_i\pi_i - \gamma_1(1-\beta)(1-\pi_1)\right) + k^2\beta^2\left((1+\beta)^2\sum_{i\geq 2}\gamma_i\pi_i^2 + \gamma_1(1-\beta)^2(1-\pi_1)^2\right) \\
&\leq 1 + k\beta(1-\pi_1)\left(\gamma_2(1+\beta) - \gamma_1(1-\beta)\right) + k^2\beta^2(1-\pi_1)^2\left(\gamma_2(1+\beta)^2 + \gamma_1(1-\beta)^2\right) \\
&= 1 + k\beta(1-\pi_1)\left(\beta(\gamma_1+\gamma_2)\left(1 + k(1-\pi_1) + \beta^2 k(1-\pi_1)\right) - (\gamma_1 - \gamma_2)(1 + 2\beta^2 k(1-\pi_1)\right) \\
&\leq 1 + k\beta(1-\pi_1)(\beta k - \lambda) \leq 1.
\end{aligned}$$

□

D.3 Proof of Lemma 5

Recall our definitions of the splitting index. Let $E = ((g_1, g_1'), \ldots, (g_n, g_n'))$ be a sequence of structure pairs. We say that an atom a ρ-splits E if

$$\max_y |E_a^y| \leq (1-\rho)|E|.$$

$\mathcal{G}$ has splitting index (ρ, ϵ, τ) if for any edge sequence E such that $d(g, g') > \epsilon$ for all $(g, g') \in E$, we have

$$\Pr_{a\sim\mathcal{D}}(a \text{ } \rho\text{-splits } E) \geq \tau.$$

Lemma 16. *Pick $\gamma, \epsilon > 0$. If $\mathcal{G}$ is finite and Assumption 1 holds, then there exists a constant $p > 0$ such that $\mathcal{G}$ has splitting index $((1-\gamma)p, \epsilon, \gamma p)$*

Proof. Given Assumption 1 and the finiteness of $\mathcal{G}$, we know that there is some $p > 0$ such that for any $g, g' \in \mathcal{G}$ satisfying $d(g, g') > 0$, we have $\Pr_{a\sim\mathcal{D}}(g(a) \neq g'(a)) \geq p$. Now suppose that we have a collection of edges $E \subset \binom{\mathcal{G}}{2}$ such that $d(g, g') > \epsilon$ for all $(g, g') \in E$. A random atom $a \sim \mathcal{D}$ will split some random number Z of these edges. Note that $\mathbb{E}Z \geq p|E|$. Moreover, by Markov's inequality, we have

$$\Pr(Z \geq (1-\gamma)p|E|)|E| \geq \mathbb{E}Z - (1-\gamma)p|E| \geq p|E| - (1-\gamma)p|E| = \gamma p|E|.$$

Simplifying the above, and substituting our definition of splitting gives us

$$\Pr_{a\sim\mathcal{D}}(a \text{ } (1-\gamma)p\text{-splits } E) \geq \gamma p.$$

□

Lemma 16 and Theorem 14 together imply the following corollary.

Corollary 17. *If $\mathcal{G}$ is finite and Assumption 1 holds, then there exists a constant $p > 0$ such that $\mathcal{G}$ has average splitting index $\left(\frac{p}{8(\log_2(1/\epsilon)+2)}, \epsilon, p/2\right)$.*

Given this result, we can now prove the following claim.

Lemma 5. *If Assumption 1 holds and* NDBAL *is run with constants $\alpha, \delta \in (0,1)$, then there is a constant $c > 0$, depending on $\alpha, \delta, d(\cdot,\cdot), \mathcal{G}$ and $\mathcal{D}$, such that for every round t,* NDBAL *queries a point that ρ_t-average split π_t satisfying $\mathbb{E}[\rho_t \,|\, \mathcal{F}_{t-1}] \;\geq\; \frac{c}{1-\log(\text{avg-diam}(\pi_t))}$.*

Proof. By Corollary 17, there is some constant $p > 0$ such that every distribution π_t is (ρ, τ)-average splittable with

$$\rho \;:=\; \frac{p}{8\left(\log_2 \frac{1}{\text{avg-diam}(\pi_t)} + 2\right)} \quad \text{and} \quad \tau \;:=\; p/2.$$

Suppose that NDBAL draws $m_t \geq 1$ candidate queries at round t. By the definition of average splittability, we have

$$\Pr(\text{at least one of } m_t \text{ draws } \rho\text{-average splits } \pi_{t-1}) \;\geq\; 1 - (1-\tau)^{m_t} \;\geq\; \tau \;\geq\; p/2.$$

Conditioned on both of this happening, Lemma 2 tells us that SELECT will choose a point that $(1-\alpha)\rho$-average splits π_t with probability $1-\delta$. Putting these together, along with the fact that $\rho_t \geq 0$ always, gives us the lemma. □

D.4 Proof of Theorem 6

Theorem 6. *If Assumptions 1 and 2 hold, $\beta \leq \lambda/10$, and $\pi_o(g^*) > 0$, then $\mathbb{E}_{g\sim\pi_t}[d(g, g^*)] \to 0$ a.s.*

Proof. Let $X_t = \text{avg-diam}(\pi_t)$ and $Y_t = 1/\pi_t(g^*)^2$. Since $\beta \leq \lambda/10$, Lemmas 3 and 5, together with the inequality $x/(1+\log(1/x)) \geq x^2$ for $x \in (0,1)$, imply

$$\mathbb{E}[X_t Y_t \,|\, \mathcal{F}_{t-1}] \;\leq\; X_{t-1}Y_{t-1} - cX_{t-1}^2 Y_{t-1} \tag{6}$$

for some constant $c > 0$. Since $X_t Y_t$ and Y_t are positive supermartingales, we have that $X_t Y_t \to Z$ and $Y_t \to Y$ for some random variables Z, Y almost surely. Moreover, since $Y_t, Y \geq 1$ almost surely, we have $X_t^2 Y_t \to W$ for some random variable W almost surely.

Iterating expectations in equation (6) and using the fact that $X_t Y_t \geq 0$, we have

$$0 \;\leq\; \mathbb{E}[X_t Y_t] \;\leq\; \frac{\text{avg-diam}(\pi_o)}{\pi_o(g^*)^2} - c\sum_{i=1}^{t-1} \mathbb{E}[X_i^2 Y_i].$$

In particular, we know $\lim_{t\to\infty} \mathbb{E}[X_t^2 Y_t] = 0$. By Fatou's lemma, this implies

$$0 \;\leq\; \mathbb{E}\left[\lim_{t\to\infty} X_t^2 Y_t\right] \;\leq\; \lim_{t\to\infty} \mathbb{E}[X_t^2 Y_t] \;=\; 0.$$

Thus, we have

$$\lim_{t\to\infty} \frac{\text{avg-diam}(\pi_t)^2}{\pi_t(g^*)^2} \;=\; \lim_{t\to\infty} X_t^2 Y_t \;=\; 0$$

almost surely. By the Continuous Mapping Theorem, this implies $\frac{\text{avg-diam}(\pi_t)}{\pi_t(g^*)} \to 0$ almost surely. The inequality

$$0 \leq \mathbb{E}_{g\sim\pi_t}[d(g, g^*)] \leq \frac{\text{avg-diam}(\pi_t)}{\pi_t(g^*)}$$

finishes the proof. □

D.5 Proof of Theorem 7

Theorem 7. *Let $\epsilon, \delta > 0$ and $\epsilon_o = \epsilon\delta\pi(g^*)/4$. If Assumption 2 holds, $\mathcal{G}$ has average splitting index (ρ, ϵ_o, τ) and* NDBAL *is run with $\beta \le \lambda/10$ and $\alpha = 1/2$, then with probability $1-\delta$,* NDBAL *encounters a distribution π_t satisfying $\mathbb{E}_{g\sim\pi_t}[d(g, g^*)] \le \epsilon$ while the resources used satisfy:*

(a) $T \le \frac{2}{\rho\lambda\beta(1-\beta)} \max\left(\ln \frac{1}{\epsilon\pi(g^*)^2}, \frac{2e^{2\beta}}{\rho\lambda\beta(1-\beta)} \ln\frac{1}{\delta}\right)$ *rounds, with one query per round,*

(b) $m_t \le \frac{1}{\tau}\log\frac{4t(t+1)}{\delta}$ *atoms drawn per round, and*

(c) $n_t \le O\left(\frac{1}{\rho\epsilon_o}\log\frac{(m_t+|\mathcal{Y}|)t(t+1)}{\delta}\right)$ *structures sampled per round.*

Proof. We will show that for some round t, NDBAL must encounter a posterior distribution π_t satisfying $\text{avg-diam}(\pi_t)/\pi(g^*)^2 \le \epsilon$ while using the resources described in the theorem statement. By Lemma 1, this will imply that $\mathbb{E}_{g\sim\pi_t}[d(g, g^*)] \le \epsilon$ for the same round t.

Lemma 4 implies that $1/\pi_t(g^*)^2$ is a positive supermartingale for our choice of β. From standard martingale theory (Resnick, 2013), we have $\pi_t(g^*)^2 \ge \delta\pi(g^*)^2/4$ for $t = 1, \ldots, T$ with probability at least $1 - \delta/4$.

Conditioned on this event, we have by a union bound that if we sample $m_t = \frac{1}{\tau}\log\frac{4t(t+1)}{\delta}$ data points at every round t, then with probability $1-\delta/4$, one of those data points will ρ-average split π_t for every round in which $\text{avg-diam}(\pi_t)/\pi_t(g^*)^2 > \epsilon$. Conditioned on drawing such points, Lemma 2 tells us that for all rounds t, SELECT terminates with a data point that $\rho/2$-average splits π_t with probability $1-\delta/4$ after drawing n_t hypotheses, for the value of n_t given in the statement.

Let us condition on all of these events happening. For round t define the random variable

$$\Delta_t = 1 - \frac{\text{avg-diam}(\pi_t)}{\pi_t(g^*)^2} \cdot \frac{\pi_{t-1}(g^*)^2}{\text{avg-diam}(\pi_{t-1})}.$$

If π_{t-1} satisfies $\text{avg-diam}(\pi_t)/\pi_t(g^*)^2 > \epsilon$, then the query x_t $\rho/2$-average splits π_{t-1}. By Lemma 3,

$$\mathbb{E}[\Delta_t \,|\, \mathcal{F}_{t-1}] \;\ge\; \frac{1}{2}\rho\lambda\beta(1-\beta).$$

Now suppose by contradiction that $\text{avg-diam}(\pi_t)/\pi_t(g^*)^2 > \epsilon$ for $t = 1, \ldots, T$. Then we have $\mathbb{E}[\Delta_1 + \ldots + \Delta_T] \ge \frac{T}{2}\rho\lambda\beta(1-\beta)$. To see that this sum is concentrated about its expectation, we notice that $\Delta_t \in [1 - e^{2\beta}, 1]$ since

$$e^{-\beta}\pi_{t-1}(g) \;\le\; \pi_t(g) \;\le\; e^{\beta}\pi_{t-1}(g)$$

for all $g \in \mathcal{G}$ which implies

$$e^{-2\beta} \;\le\; \frac{\text{avg-diam}(\pi_t)}{\pi_t(g^*)^2} \cdot \frac{\pi_{t-1}(g^*)^2}{\text{avg-diam}(\pi_{t-1})} \;\le\; e^{2\beta}.$$

By the Azuma-Hoeffding inequality (Azuma, 1967; Hoeffding, 1963), if T achieves the value in the theorem statement, then with probability $1-\delta$,

$$\Delta_1 + \cdots + \Delta_T \;>\; \frac{1}{2}\mathbb{E}[\Delta_1 + \cdots + \Delta_T] \;\ge\; \frac{T}{8}\rho\lambda\beta(1-\beta) \;\ge\; \ln\frac{1}{\epsilon\pi(g^*)^2}.$$

However, this is a contradiction since

$$\epsilon \;<\; \frac{\text{avg-diam}(\pi_T)}{\pi_T(g^*)^2} \;=\; (1-\Delta_1)\cdots(1-\Delta_T)\frac{\text{avg-diam}(\pi)}{\pi(g^*)^2} \;\le\; \exp\left(-(\Delta_1 + \cdots + \Delta_T)\right)\frac{1}{\pi(g^*)^2}.$$

Thus, with probability $1-\delta$, we must have encountered a distribution π_t in some round $t = 1, \ldots, T$ satisfying $\text{avg-diam}(\pi_t)/\pi_t(g^*)^2 \le \epsilon$. □

D.6 Proof of Theorem 9

To begin, we will utilize the following result on our stopping criterion.

Lemma 18. *Pick $\epsilon, \delta > 0$ and let $n_t = \frac{48}{\epsilon} \log \frac{t(t+1)}{\delta}$. If at the beginning of each round t, we draw $E = (\{g_1, g_1'\}, \ldots, \{g_{n_t}, g_{n_t}'\}) \sim \pi_t$, then with probability $1 - \delta$*

$$\frac{1}{n_t} \sum_{i=1}^{n_t} d(g_i, g_i') > \frac{3\epsilon}{4} \quad \textit{if} \quad \text{avg-diam}(\pi_t) > \epsilon$$
$$\frac{1}{n_t} \sum_{i=1}^{n_t} d(g_i, g_i') \leq \frac{3\epsilon}{4} \quad \textit{if} \quad \text{avg-diam}(\pi_t) \leq \epsilon/2$$

for all rounds $t \geq 1$.

The proof of Lemma 18 follows from applying a union bound to Lemma 7 of Tosh and Dasgupta (2017).

For a round t, let V_t denote the version space, i.e. the set of structures consistent with the responses seen so far. Then we may write

$$\pi_t(g) \;=\; \frac{\pi(g)\mathbb{1}[g \in V_t]}{\pi(V_t)} \quad \text{and} \quad \nu_t(g) \;=\; \frac{\nu(g)\mathbb{1}[g \in V_t]}{\nu(V_t)}.$$

Assumption 3 tells us that we have the following upper bound.

$$D(\pi_t, \nu_t) \;\leq\; \lambda^2 \text{avg-diam}(\pi_t).$$

Thus, the average diameter of avg-diam(π_t) is a meaningful surrogate for the objective $D(\pi_t, \nu_t)$ in this setting.

Recalling the definition of average splitting, we know that if we always query points that ρ-average the current posterior, then after t rounds we will have

$$\pi(V_t)^2 \text{avg-diam}(\pi_t) \;\leq\; (1-\rho)^t \pi(V_0)^2 \text{avg-diam}(\pi) \;\leq\; e^{-\rho t}.$$

While this demonstrates that the potential function $\pi(V_t)^2$avg-diam(π_t) is decreasing exponentially quickly, it does not by itself guarantee that avg-diam(π_t) is itself decreasing. What is needed is a lower bound on the factor $\pi(V_t)$. The following lemma, which is a generalization of a result due to Freund et al. (1997), provides us with just that, provided that $\mathcal{G}$ has bounded graph dimension.

Lemma 19. *Suppose $g^* \sim \nu$ where ν is a prior distribution over a hypothesis class $\mathcal{G}$ with graph dimension d_G, and say $|\mathcal{Y}| \leq k$. Let $c > 0$ and $a_1, \ldots, a_m$ be any atomic questions, and let $V^* = \{g \in \mathcal{G} : g(a_i) = g^*(a_i)$ for all $i\}$, then*

$$\Pr\left(\log\left(\frac{1}{\nu(V^*)}\right) \geq c + d_G \log \frac{em(k+1)}{d_G}\right) \;\leq\; e^{-c}.$$

To prove this, we need the following generalization of Sauer's lemma.

Lemma 20 (Corollary 3 of Haussler and Long (1995))**.** *Let d, m, k be s.t. $d \leq m$. Let $F \subset \{1, \ldots, k\}^m$ s.t. F has graph dimension less than d. Then,*

$$|F| \leq \sum_{i=0}^{d} \binom{m}{i} (k+1)^i \leq \left(\frac{em(k+1)}{d}\right)^d.$$

Proof of Lemma 19. Let $V_1, \ldots, V_N \subset \mathcal{G}$ denote the partition of $\mathcal{G}$ induced by our atomic questions. Note that if $g^* \sim \nu$, then the probability $V^* = V_i$ is exactly $\nu(V_i)$. Let $S \subset \{1, \ldots N\}$ consist of all indices i satisfying $\log \frac{1}{\nu(V_i)} \geq c + \log N$. Rearranging, we have

$$\sum_{i \in S} \nu(V_i) \;\leq\; e^{-c} \cdot \frac{|S|}{N} \;\leq\; e^{-c}.$$

From Lemma 20, we have $\log N \leq d_G \log \frac{em(k+1)}{d_G}$, which finishes the proof. □

Given the above, we are now ready to prove Theorem 9.

Theorem 9. *Suppose $\mathcal{G}$ has average splitting index $(\rho, \epsilon/(2\lambda^2), \tau)$ and graph dimension d_G. If Assumptions 3 and 4 hold, then with probability $1-\delta$, modified* NDBAL *terminates with a distribution π_t satisfying $D(\pi_t, \nu_t) \le \epsilon$ while using the following resources:*

(a) $T \le O\left(\frac{d_G}{\rho}\left(\log\frac{|\mathcal{Y}|\lambda}{\epsilon\tau\delta} + \log^2\frac{d_G}{\rho}\right)\right)$ *rounds with one query per round,*

(b) $m_t \le O\left(\frac{1}{\tau}\log\frac{t}{\delta}\right)$ *atoms drawn per round, and*

(c) $n_t \le O\left(\left(\frac{\lambda^2}{\epsilon\rho}\right)\log\frac{(m_t+|\mathcal{Y}|)t}{\delta}\right)$ *structures sampled per round.*

Proof. If we use the stopping criterion from Lemma 18 with the threshold $3\epsilon/4\lambda^2$, then at the expense of drawing an extra $\frac{48\lambda^2}{\epsilon}\log\frac{t(t+1)}{\delta}$ hypotheses for each round t, we are guaranteed that with probability $1-\delta$ if we ever encounter a round t in which avg-diam$(\pi_t) \le \epsilon/(2\lambda^2)$ then we terminate and we also never terminate whenever avg-diam$(\pi_K) > \epsilon$. Thus if we do ever terminate at some round t, then with high probability

$$D(\pi_t, \nu_t) \;\le\; \lambda^2 \text{avg-diam}(\pi_t) \;\le\; \epsilon.$$

It remains to be shown that we will encounter such a posterior. Note that if we draw $m_t \ge \frac{1}{\tau}\log\frac{t(t+1)}{\delta}$ atoms per round, then with probability $1-\delta$ one of them will ρ-average split π_t if avg-diam$(\pi_t) > \epsilon/(2\lambda^2)$. Conditioned on this happening, Lemma 2 guarantees that that with probability $1-\delta$ SELECT finds a point that $\rho/2$-average splits π_t while drawing at most $O\left(\frac{\lambda^2}{\epsilon\rho}\log\frac{(m_t+|\mathcal{Y}|)t(t+1)}{\delta}\right)$.

If after T rounds we still have not terminated, then avg-diam$(\pi_T) > \epsilon/(2\lambda^2)$. However, we also know

$$\pi(V_T)^2 \text{avg-diam}(\pi_T) \;\le\; e^{-\rho T/2}.$$

Now suppose that in each round t, we have seen m_t atoms $x_1^{(t)}, \ldots, x_{m_t}^{(t)}$, and define

$$V_{T^*} = \{h \in \mathcal{H} \,:\, h(x_i^{(t)}) = h^*(x_i^{(t)}) \text{ for } t = 1, \ldots, T, i = 1, \ldots, m_t\}.$$

Clearly, $V_{T^*} \subset V_T$. By Lemma 19, we have with probability $1-\delta$,

$$\pi(V_T) \;\ge\; \pi(V_{T^*}) \;\ge\; \frac{1}{\lambda}\nu(V_{T^*}) \;\ge\; \frac{1}{\lambda}\cdot\frac{\delta}{T(T+1)}\left(\frac{d_G}{em^{(T)}(|\mathcal{Y}|+1)}\right)^{d_G}$$

for all rounds $T \ge 1$, where $m^{(T)} = \sum_{t=1}^{T} m_t$.

Plugging this in with the above, we have

$$\text{avg-diam}(\pi_T) \;\le\; \frac{e^{-\rho T/2}}{\pi(V_T)^2} \;\le\; \lambda^2 \exp\left(2d_G\log\frac{em^{(T)}(|\mathcal{Y}|+1)}{d_G} + 2\log\frac{T(T+1)}{\delta} - \frac{\rho T}{2}\right).$$

Suppose $m_t = \frac{1}{\tau}\log\frac{t(t+1)}{\delta}$. Then we can upper bound $m^{(T)}$ as

$$m^{(T)} \;=\; \sum_{t=1}^{T} m_t \;\le\; \frac{T}{\tau}\log\frac{T(T+1)}{\delta}.$$

Putting everything together, we have

$$\frac{\epsilon}{2\lambda^2} \;\le\; \text{avg-diam}(\pi_T) \;\le\; \lambda^2\exp\left(2\log\frac{T(T+1)}{\delta} + 2d_G\log\left(\frac{e(|\mathcal{Y}|+1)}{d_G}\cdot\frac{T}{\tau}\log\frac{T(T+1)}{\delta}\right) - \frac{\rho T}{2}\right).$$

Letting $C = 2d_G\log\frac{e(|\mathcal{Y}|+1)}{d_G\tau}$ and $b = \frac{1}{\delta}$, the right-hand side is less than $\epsilon/(2\lambda^2)$, whenever

$$T \;\ge\; \frac{2}{\rho}\max\left\{C + \log\frac{2\lambda^4}{\epsilon} + 6(d_G+1)\log T, C + \log\frac{2\lambda^4}{\epsilon} + \log b + 2d_G\log\left(3b\log(b)\right)\right\}.$$

Additionally, note that $T \ge \frac{2}{\rho}\left(C + \log\frac{1}{\epsilon} + 6(d_G+1)\log T\right)$, whenever

$$T \;\ge\; \frac{4}{\rho}\max\left\{C + \log\frac{2\lambda^4}{\epsilon}, 24(d_G+1)\log^2\left(\frac{96(d_G+1)}{\rho}\right)\right\}.$$

The value of T provided in the theorem statement, satisfies all of these inequalities. Thus, with probability $1-4\delta$, we must have encountered a round in which avg-diam$(\pi_t) < \epsilon/(2\lambda^2)$ and terminated. □

D.7 Proof of Theorem 10

The following result is analogous to Theorem 2 of Dasgupta (2005).

Theorem 21. *Fix $\mathcal{G}$ and $\mathcal{D}$. Suppose that $\mathcal{G}$ does not have splitting index (ρ, ϵ, τ) for some $\rho, \epsilon \in (0,1)$ and $\tau \in (0, 1/2)$. Then any interactive learning strategy which with probability $> 3/4$ over the random sampling from $\mathcal{D}$ finds a structure $g \in \mathcal{G}$ within distance $\epsilon/2$ of any target in $\mathcal{G}$ must draw at least $1/\tau$ atoms from $\mathcal{D}$ or must make at least $1/\rho$ queries.*

From the proof of Theorem 2 of Dasgupta (2005), it is easy to see that so long as $d(\cdot, \cdot)$ is symmetric, the same arguments imply Theorem 21. For completeness, we include its proof here.

Proof. Since $\mathcal{G}$ does not have splitting index (ρ, ϵ, τ), there is some set of edges $E \subset \binom{\mathcal{G}}{2}$ such that $d(g, g') > \epsilon$ for all $(g, g') \in E$ and

$$\Pr_{a \sim \mathcal{D}}(a \ \rho\text{-splits } E) < \tau.$$

Let V denote the vertices of E. Then distinguishing between structures in V requires at least $1/\rho$ queries or at least $1/\tau$ atoms.

To see this, suppose we draw less than $1/\tau$ atoms. Then with probability at least $(1-\tau)^{1/\tau} \geq 1/4$ none of these atoms ρ-splits E, i.e. for each of these atoms there is some response $y \in \mathcal{Y}$ such that less than $\rho|E|$ edges are eliminated. Thus, there is some $g^* \in V$ such that requires us to query at least $1/\rho$ atoms to distinguish it from the rest of the structures in V. □

Combining the above with Theorem 14, we have the following corollary.

Theorem 10. *Fix $\mathcal{G}$, $\mathcal{D}$ and $d(\cdot, \cdot)$. If $\mathcal{G}$ does not have average splitting index $(\frac{\rho}{4\lceil \log 1/\epsilon \rceil}, 2\epsilon, \tau)$ for some $\rho, \epsilon \in (0,1)$ and $\tau \in (0, 1/2)$, then any interactive learning strategy which with probability $> 3/4$ over the random sampling from $\mathcal{D}$ finds a structure $g \in \mathcal{G}$ within distance $\epsilon/2$ of any target in $\mathcal{G}$ must draw at least $1/\tau$ atoms from $\mathcal{D}$ or must make at least $1/\rho$ queries.*

E Proofs from Section 5

E.1 Proof of Theorem 11

We will utilize the following result from Dasgupta (2005).

Lemma 22 (Lemma 11 from Dasgupta (2005))**.** *For any $d \geq 2$, let x, y be vectors in $\mathbb{R}^d$ separated by an angle of $\theta \in [0, \pi]$. Let $\tilde{x}, \tilde{y}$ be their projections into a randomly chosen two-dimensional subspace. There is an absolute constant $c_o > 0$ (which does not depend on d) such that with probability at least 3/4 over the choice of subspace, the angle between $\tilde{x}$ and $\tilde{y}$ is at least $c_o\theta$.*

Given the above, we prove Theorem 11.

Theorem 11. *Suppose μ is spherically symmetric. Under distance $d_r(\cdot, \cdot)$, $\mathcal{G}$ has average splitting index $(\frac{1}{16\lceil \log(2/\epsilon) \rceil}, \epsilon, c\epsilon)$ for some absolute constant $c > 0$.*

The proof of Theorem 11 closely mirrors that of Theorem 10 Dasgupta (2005). For completeness, we produce its proof here.

Proof. We make two key observations here.

- A weight vector $w \in \mathcal{G}$ ranks x over y if and only if $\langle w, x - y \rangle > 0$.
- If x, y are drawn from a spherically symmetric distribution, then $z = x - y$ also follows a spherically symmetric distribution.

From these two observations, we know that if $w, w' \in \mathcal{G}$, then $d(w, w') = \theta/\pi$ where θ is the angle lying between w and w'.

Suppose $w_1, w_1', \ldots, w_n, w_n'$ are a sequence of edges such that $d(w_i, w_i') \geq \epsilon$, which implies their corresponding angles satisfy $\theta_i \geq \epsilon\pi$. Suppose we project the pairs onto a randomly drawn 2-d subspace, to get $\tilde{w}_1, \tilde{w}_1', \ldots, \tilde{w}_n, \tilde{w}_n'$. Let c_o be the absolute constant from Lemma 22. Call an edge $\tilde{w}_i, \tilde{w}_i'$ good if the resulting angle satisfies $\tilde{\theta}_i \geq c_o\epsilon\pi$.

By Lemma 22, the expected number of good edges for a randomly chosen 2-d subspace is $n/2$. By Markov's inequality, with probability $1/2$, at least $n/2$ edges are good.

Let us suppose that we have chosen a 2-d subspace/plane that results in at least $n/2$ good edges. Call these projected edges $\tilde{w}_1, \tilde{w}_1', \ldots, \tilde{w}_m, \tilde{w}_m'$. Without loss of generality, assume that the clockwise angle $\tilde{\theta}_i$ from $\tilde{w}_i$ to $\tilde{w}_i'$ satisfies $c_o\epsilon\pi \geq \tilde{\theta}_i \leq \pi$. Notice that if z_o is in our plane and satisfies $\langle \tilde{w}_i, z_o \rangle \geq 0$ for at least $n/2$ edges and $\langle \tilde{w}_i', z_o \rangle \leq 0$ for at least $n/2$ edges, then querying any points x_o, y_o such that $x_o - y_o = z_o$ will eliminate at least half of the $\tilde{w}_i$. Moreover, it is enough to query any pair x, y such that $x - y = z$ satisfies that x's counterclockwise angle is in the range $[0, c_o\epsilon\pi]$ or $[\pi, \pi + c_o\epsilon\pi]$, since such a pair will eliminate either $\tilde{w}_i$ or $\tilde{w}_i'$. Thus, querying such an x, y pair will result in eliminating at least $1/2$ of the good edges, which is at least $1/4$ of all the edges.

Since $z = x - y$ follows a spherically symmetric distribution, the probability of drawing such a pair is at least $c_o\epsilon\pi/2$. Thus, the splitting index here is $(1/4, \epsilon, c_o\epsilon\pi/2)$, and Theorem 11 follows by applying Theorem 14. □

E.2 Proof of Lemma 13

Lemma 13. *Let $\mu(\mathcal{I}) = \alpha$. Under distance $d_{\mathcal{I}}(\cdot, \cdot)$, $\mathcal{G}_{k,\mathcal{I}}$ has average splitting index $(\frac{1}{16\lceil \log(2/\epsilon) \rceil}, \epsilon, \frac{\epsilon\alpha}{2})$.*

Proof. We will first bound the splitting index and then invoke Theorem 14. Suppose that $g_1, g_1', \ldots, g_n, g_n' \in \mathcal{G}_{k,\alpha}$ are a sequence of edges satisfying $d_{\mathcal{I}}(g_i, g_i') \geq \epsilon$ for all $i = 1, \ldots, n$. Note that for each g_i, g_i' there are associated reals $\ell_i < u_i$ and $\ell_i' < u_i'$ such that

$$\ell_i, \ell_i' \ \leq \mathcal{I} \ \leq \ u_i, u_i'.$$

From the definition of $d_{\mathcal{I}}(g_i, g_i')$, we have

$$\epsilon \ \leq \ d_{\mathcal{I}}(g_i, g_i') \ = \ \mu(\ell_i, \ell_i') + \mu(u_i, u_i')$$

where $\mu(a, b)$ is the probability mass of the interval bounded by a and b. Call an edge *left-leaning* if $\mu(\ell_i, \ell_i') \geq \epsilon/2$ and *right-leaning* if $\mu(u_i, u_i') \geq \epsilon/2$.

Suppose without loss of generality that at least half of the edges are right-leaning (the case where half are left-leaning can be handled symmetrically), and order them as $g_1, g_1', \ldots, g_m, g_m'$ such that $u_1 \leq u_2 \leq \cdots \leq u_m$. Moreover, let us also assume without loss of generality that $u_i < u_i'$. Let r denote the point $u_i < r \leq u_i'$ such that $\mu(u_i, r) = \epsilon/2$. Suppose we query a pair x, y where $x \in \mathcal{I}$ and $y \in (u_{m/2}, r)$, notice that such a pair satisfies.

$$x < u_1 \leq \cdots \leq u_{m/2} < y < u_{m/2}' \leq \cdots \leq u_m'.$$

If we query this pair and the result is that they should belong to the same cluster, then we may eliminate at least one endpoint of edges $g_1, g_1', \ldots, g_{m/2}, g_{m/2}'$. On the other hand, if the result is that they should belong to different clusters, then we may eliminate at least one endpoint of edges $g_{m/2}, g_{m/2}', \ldots, g_m, g_m'$. In either case, we eliminate at least half of these m edges. Since this is only the right-leaning edges, at least one quarter of the original edges are eliminated. Finally, the probability of drawing such a pair x, y is $\alpha \cdot \epsilon$.

Thus, $\mathcal{G}_{k,\mathcal{I}}$ has splitting index $(1/4, \epsilon, \alpha\epsilon)$. Theorem 14 finishes the proof. □

E.3 Proof of Theorem 12

We will make use of the following result from Dasgupta (2005).

Lemma 23 (Corollary 3 from Dasgupta (2005))**.** *Suppose there are structures $g_o, g_1, \ldots, g_N \in \mathcal{G}$ such that*

1. *$d(g_o, g_i) > \epsilon$ for all $i = 1, \ldots, N$ and*
2. *the sets $\{a : g_o(a) \neq g_i(a)\}$ are disjoint for all $i = 1, \ldots, N$.*

Then for any $\tau > 0$ and any $\rho > 1/N$, $\mathcal{G}$ is not (ρ, ϵ, τ)-splittable. Thus, any active learning scheme that finds $g \in \mathcal{G}$ satisfying $d(g, g^*) < \epsilon/2$ for any $g^* \in \mathcal{G}$ must use at least N labels in the worst case.

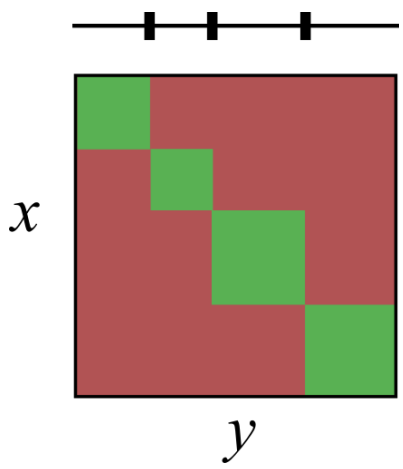

Figure 4: Viewing an interval-based clustering as a classifier over $\mathbb{R}^2$. The green regions correspond to 'must-link' constraints, and the red regions correspond to 'cannot-link' constraints.

Given this, we have the following lemma lower bounding the query complexity of a particular subset of $\mathcal{G}_{k,\mathcal{I}}$.

Lemma 24. *Say $\mu(\mathcal{I}) \leq 1/2$. There is a subset $\mathcal{G}_o \subset \mathcal{G}_{k+2,\mathcal{I}}$ of $N = \min\{k, \frac{1}{\sqrt{8\epsilon}}\} + 1$ clusterings such that learning $\mathcal{G}_o$ under distance $d_c(\cdot,\cdot)$ requires at least $N-1$ queries, no matter how many unlabeled data points are drawn.*

Proof. For ease of exposition, say that μ is uniform over the interval $[0,1]$ and that $\mathcal{I} = [0,\alpha]$ for some $\alpha \leq 1/2$. We will consider the case where $k \leq \frac{1}{\sqrt{8\epsilon}}$, the other case can be proven symmetrically.

Define g_o as the clustering with dividing points

$$a_1 = \alpha,\ a_2 = \alpha + \frac{1-\alpha}{k},\ a_3 = \alpha + \frac{2(1-\alpha)}{k},\ \ldots,\ a_k = \alpha + \frac{(k-1)(1-\alpha)}{k}.$$

We also define g_i as the clustering with the same dividing points except it has an additional dividing point at $b_i = \frac{a_i + a_{i+1}}{2} = \alpha + \frac{(2i-1)(1-\alpha)}{2k}$ for $i = 1, \ldots k$, where we take $a_{k+1} = 1$. Then it can be seen that

$$d(g_o, g_i) \ = \ 2 \cdot \Pr_{x\sim\mu}(x \in (a_i, b_i)) \cdot \Pr_{y\sim\mu}(y \in (b_i, a_{i+1})) \ = \ \frac{1}{2}\left(\frac{1-\alpha}{k}\right)^2 \ \geq \ \epsilon.$$

Moreover, we also have that the sets $\{(x,y) : g_o(x,y) \neq g_i(x,y)\}$ are disjoint for all $i = 1, \ldots, N$. This is readily observed after making the transformation from an interval-based clustering to binary classifier over $[0,1]^2$. Applying Lemma 23 finishes the proof. □

Given Lemmas 13 and 24, we can now prove Theorem 12.

Theorem 12 *(Formal statement) Let $\epsilon > 0$. There is a setting of $k = \Theta(1/\sqrt{\epsilon})$ and a subset $\mathcal{G} \subseteq \mathcal{G}_{k+2,\mathcal{I}}$ that is polynomially-sized in k such that any active learning algorithm that is guaranteed to find any target in $\mathcal{G}$ up to distance ϵ in distance $d_c(\cdot,\cdot)$ must make at least $\Omega(k)$ queries, but* NDBAL *with distance $d_{\mathcal{I}}(\cdot,\cdot)$ and prior π uniform over $\mathcal{G}$ requires $O(\log^2(k/\epsilon\delta))$ queries.*

Proof. Take $k = \Theta(1/\sqrt{\epsilon})$ and let $\mathcal{G}_o \subset \mathcal{G}_{k+2,\mathcal{I}}$ be the subset from Lemma 24. Take $\mathcal{G}$ to be any subset of $\mathcal{G}_{k+2,\mathcal{I}}$ such that (a) $\mathcal{G}$ has size polynomial in k and (b) $\mathcal{G}_o \subseteq \mathcal{G}$. By Lemma 24, we know that learning under distance $d_c(\cdot,\cdot)$ requires at least $|\mathcal{G}_o| = \Theta(k)$ queries.

On the other hand, consider running NDBAL with distance $d_{\mathcal{I}}(\cdot,\cdot)$ and prior π uniform over $\mathcal{G}$. The results in Theorem 7 and Lemma 13 tell us that NDBAL requires $O(\log^2(k/\epsilon))$ queries to find a posterior π_t over $\mathcal{G}$ such that $\mathbb{E}_{g\sim\pi_t}[d_{\mathcal{I}}(g, g^*)] \leq \epsilon$. To turn this into a high probability result, simply apply Markov's inequality to get that NDBAL requires $O(\log^2(k/\epsilon\delta))$ queries in order to find a posterior π_t such that with probability $1-\delta$ if $g \sim \pi_t$ then $d_{\mathcal{I}}(g, g^*) \leq \epsilon$. □

F Noisy fast convergence

In this section, we give rates of convergence in the Bayesian setting under noise. We start by defining the quantity

$$Z_t = \sum_{g \in \mathcal{G}} \pi(g) \exp\left(-\beta \sum_{i=1}^{t} \mathbb{1}[g(x_i) \neq y_i]\right).$$

The following lemma is analogous to Lemma 3.

Lemma 25. *Pick $\beta, \rho > 0$. If at step t, our query ρ-average splits π_{t-1}, then*

$$Z_t^2 \Phi(\pi_t) \;\leq\; \left[1 - \rho(1 - e^{-\beta})\right] Z_{t-1}^2 \Phi(\pi_{t-1}).$$

Proof. Suppose that we query atom a_t and receive label y_t. Enumerate the potential responses as $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$. The definition of average splitting implies that there exists a symmetric matrix $R \in [0,1]^{m \times m}$ satisfying

- $R_{ii} \leq 1 - \rho$ for all i,
- $\sum_{i,j} R_{ij} = 1$, and
- $R_{ij}\,\text{avg-diam}(\pi) = \sum_{g \in \mathcal{G}_a^{y_i}, g' \in \mathcal{G}_a^{y_j}} \pi(g)\pi(g')d(g,g')$.

Define the quantity

$$Q_a^i \;:=\; \pi(G_a^{y_i}) + e^{-\beta} \sum_{j \neq i} \pi(G_a^{y_j}) \;=\; \pi(G_a^{y_i}) + e^{-\beta}(1 - \pi(G_a^{y_i})) \;\leq\; 1.$$

Note that if $y_t = y_i$, we have

$$Q_a^i \;=\; \sum_g \pi_{t-1}(g) \exp\left(-\beta \mathbb{1}[g(a_t) \neq y_t]\right) \;=\; \sum_g \frac{1}{Z_{t-1}} \pi(g) \exp\left(-\beta \sum_{j=1}^{t} \mathbb{1}[g(a_j) \neq y_j]\right) \;=\; \frac{Z_t}{Z_{t-1}}$$

Thus, if we observe $y_t = y_i$, then

$$\begin{aligned} Z_t^2 \text{avg-diam}(\pi_t) &= (Q_a^i Z_{t-1})^2 \sum_{g,g'} \frac{1}{(Q_a^i)^2} \pi_{t-1}(g)\pi_{t-1}(g')d(g,g') \exp\left(-\beta(\mathbb{1}[g(a_t) \neq y_i] + \mathbb{1}[g(a_t) \neq y_t])\right) \\ &= \left(R_{ii} + e^{-2\beta} \sum_{j,k \neq i} R_{jk} + e^{-\beta} \cdot 2 \sum_{j \neq i} R_{ij}\right) Z_{t-1}^2 \text{avg-diam}(\pi_{t-1}) \\ &\leq \left((1-\rho) + e^{-\beta}\rho\right) Z_{t-1}^2 \text{avg-diam}(\pi_{t-1}) \;=\; \left(1 - \rho(1 - e^{-\beta})\right) Z_{t-1}^2 \text{avg-diam}(\pi_{t-1}). \end{aligned}$$

□

Suppose we receive query/label pairs $(a_1, y_1), \ldots, (a_t, y_t)$ where the noise level at a_i is q_i, then the true posterior distribution under Assumption 3 is

$$\nu_t(g) \;=\; \frac{1}{\widehat{Z}_t} \nu(g) \exp\left(-\sum_{i=1}^{t} \mathbb{1}[g(a_i) \neq y_i] \ln \frac{1 - q_i}{q_i}\right)$$

where $\widehat{Z}_t$ is the normalizing constant

$$\widehat{Z}_t \;=\; \sum_g \nu(g) \exp\left(-\sum_{i=1}^{t} \mathbb{1}[g(a_i) \neq y_i)] \ln \frac{1 - q_i}{q_i}\right).$$

The following lemma will be useful in bounding this quantity.

Lemma 26. *Suppose $Y_1, \dots, Y_t$ are independent random variables such that*

$$Y_i \;=\; \begin{cases} \ln \frac{1-q_i}{q_i} & \text{with probability } q_i \\ 0 & \text{with probability } 1-q_i \end{cases}$$

With probability $1-\delta$, we have

$$\sum_{i=1}^t Y_i \;\le\; \sum_{i=1}^t q_i \ln \frac{1-q_i}{q_i} + \sqrt{t \ln \frac{2}{\delta}} \left(\ln \frac{2t}{\delta} \right).$$

Proof. We begin by partitioning the random variables Y_i into two groups. We say Y_i is 'small' if $q_i \le \frac{\delta}{2t}$ and 'big' otherwise. Then with probability at least $1-\delta/2$, all small Y_i satisfy $Y_i = 0$. Let us condition on this happening.

Now each big Y_i takes values in $[0, \ln \frac{2t}{\delta}]$. By Hoeffding's inequality, we have that with probability at least $1-\delta/2$

$$\sum_{i=1}^t Y_i \;\le\; \sum_{i=1}^t \mathbb{E}[Y_i] + \sqrt{t \ln \frac{2}{\delta}} \left(\ln \frac{2t}{\delta} \right) \;\le\; \sum_{i=1}^t q_i \ln \frac{1-q_i}{q_i} + \sqrt{t \ln \frac{2}{\delta}} \left(\ln \frac{2t}{\delta} \right). \qquad \square$$

Given the above, we can lower bound $\widehat{Z}_t$ under Assumption 3.

Lemma 27. *Let $\delta \in (0,1)$ and let $\mathcal{G}$ have graph dimension d_G. Suppose Assumption 3 holds. If in the course of running* NDBAL *we observe m atoms, of which we query $a_1, \dots, a_t$ where the noise level at a_i is q_i, then with probability $1-\delta$ over the randomness of the responses we observe,*

$$\log \frac{1}{\widehat{Z}_t} \;\le\; \log \frac{2}{\delta} + d_G \log \frac{em(|\mathcal{Y}|+1)}{d_G} + \sum_{i=1}^t q_i \ln \frac{1-q_i}{q_t} + \sqrt{t \log \frac{3}{\delta}} \left(\log \frac{3t}{\delta} \right)$$

Proof. By Assumption 3, we know $g^* \sim \nu$. Let U be the set of m atoms observed in running NDBAL and let $V^* = \{g \in \mathcal{G} : g(a) = g^*(a) \text{ for } a \in U\}$. By Lemma 19, we have with probability $1-\delta/2$

$$\log \frac{1}{\nu(V^*)} \;\le\; \log \frac{2}{\delta} + d_G \log \frac{em(|\mathcal{Y}|+1)}{d_G}.$$

Now let $g \in V^*$ and say the responses on atoms $a_1, \dots, a_t$ are $y_1, \dots, y_t$, respectively. By Lemma 26, we have with probability $1-\delta/2$

$$\sum_{i=1}^t \mathbb{1}[g(a_i) \ne y_i] \ln \frac{1-q_i}{q_i} \;\le\; \sum_{i=1}^t q_i \ln \frac{1-q_i}{q_t} + \sqrt{t \log \frac{6}{\delta}} \left(\log \frac{6t}{\delta} \right).$$

Combining the above concentration results with the inequality

$$\widehat{Z}_t \;\ge\; \sum_{g \in V^*} \nu(g) \exp\left(-\sum_{i=1}^t \mathbb{1}[g(a_i) \ne y_i] \ln \frac{1-q_i}{q_i} \right)$$

gives us the lemma. $\square$

We will assume that the noise distribution is restricted to classification noise.

Assumption 5. *There exists a $q \in (0,1)$ and $g^* \in \mathcal{G}$ such that $\eta(g^*(a) \,|\, a) = 1-q$.*

If we know the noise level, then the appropriate setting of β is $\ln \frac{1-q}{q}$, in which case we recover the bound

$$\mathcal{D}(\pi_t, \nu_t) \;\le\; \lambda^2 \text{avg-diam}(\pi_t). \tag{7}$$

Given the above, we can now prove the following theorem.

Theorem 28. *Suppose $\mathcal{G}$ has average splitting index $(\rho, \epsilon/(2\lambda^2), \tau)$ and graph dimension d_G. If Assumptions 3 and 5 hold, $\gamma = \frac{\rho}{2} \cdot \frac{1-2q}{1-q} - q \ln \frac{1-q}{q} > 0$, and $\beta = \ln \frac{1-q}{q}$, then with probability $1-\delta$ modified* NDBAL *terminates with a distribution π_t satisfying $D(\pi_t, \nu_t) \le \epsilon$ while using the following resources:*

(a) less than $T = O\left(\frac{1}{\gamma}\log^3\frac{1}{\gamma\delta} + \frac{d_G}{\gamma}\log\left(\frac{d_G\lambda|\mathcal{Y}|}{\epsilon\tau\delta}\log\left(\frac{d_G\lambda|\mathcal{Y}|}{\epsilon\tau\delta}\right)\right)\right)$ rounds with one query per round,

(b) $m_t \le O\left(\frac{1}{\tau}\log\frac{t}{\delta}\right)$ atoms drawn per round, and

(c) $n_t \le O\left(\left(\frac{\lambda^2}{\epsilon\rho}\right)\log\frac{(m_t+|\mathcal{Y}|)t}{\delta}\right)$ structures sampled per round.

Proof. If we use the stopping criterion from Lemma 18 with the threshold $3\epsilon/4\lambda^2$, then at the expense of drawing an extra $\frac{48\lambda^2}{\epsilon}\log\frac{t(t+1)}{\delta}$ hypotheses for each round t, we are guaranteed that with probability $1-\delta$ if we ever encounter a round t in which avg-diam$(\pi_t) \le \epsilon/(2\lambda^2)$ then we terminate and we also never terminate whenever avg-diam$(\pi_K) > \epsilon$. Thus if we do ever terminate at some round t, equation (7) guarantees

$$D(\pi_t, \nu_t) \;\le\; \epsilon.$$

Note that if we draw $m_t \ge \frac{1}{\tau}\log\frac{t(t+1)}{\delta}$ atoms per round, then with probability $1-\delta$ one of them will ρ-average split π_t if avg-diam$(\pi_t) > \epsilon/(2\lambda^2)$. Conditioned on this happening, Lemma 2 guarantees that that with probability $1-\delta$ SELECT finds a point that $\rho/2$-average splits π_t while drawing at most $O\left(\frac{\lambda^2}{\epsilon\rho}\log\frac{(m_t+|\mathcal{Y}|)t(t+1)}{\delta}\right)$.

If after T rounds we still have not terminated, then avg-diam$(\pi_T) > \epsilon/(2\lambda^2)$. By Lemma 25 we also know

$$Z_T^2\,\text{avg-diam}(\pi_T) \;\le\; \exp\left(-\rho(1-e^{-\beta})T/2\right) \;=\; \exp\left(-\frac{\rho T}{2}\cdot\frac{1-2q}{1-q}\right).$$

By Lemma 27, we have that for all rounds $t \ge 1$, with probability $1-\delta$,

$$\log\frac{1}{Z_t} \;\le\; \log\frac{2t(t+1)}{\delta} + d_G\log\frac{em^{(t)}(|\mathcal{Y}|+1)}{d_G} + tq\ln\frac{1-q}{q} + \sqrt{t\log\frac{4t(t+1)}{\delta}}\left(\log\frac{4t^2(t+1)}{\delta}\right).$$

Where $m^{(t)}$ is the number of atoms sampled up to time t, which can be bounded as

$$m^{(t)} \;\le\; \frac{t}{\tau}\log\frac{t(t+1)}{\delta}.$$

Putting this together, we can conclude that avg-diam$(\pi_T) \le \epsilon/(2\lambda^2)$ whenever

$$T \;\ge\; \max\frac{2}{\gamma}\left\{\sqrt{T\log\frac{4T(T+1)}{\delta}}\left(\log\frac{4T^2(T+1)}{\delta}\right),\right.$$
$$\left.\log\frac{2T(T+1)}{\delta} + d_G\log\left(\frac{e(|\mathcal{Y}|+1)}{d_G}\cdot\frac{T}{\tau}\log\frac{T(T+1)}{\delta}\right) + \log\frac{2\lambda^2}{\epsilon}\right\}.$$

Note that $T \ge \frac{2}{\gamma}\sqrt{T\log\frac{4T(T+1)}{\delta}}\left(\log\frac{4T^2(T+1)}{\delta}\right)$ whenever $T \ge \frac{4}{\gamma^2}\log^3\left(\frac{4T^2(T+1)}{\delta}\right)$ and this is satisfied for

$$T \ge \frac{4c_1}{\gamma^2}\left(\log^3\frac{4}{\gamma^2} + \log^3\frac{4}{\delta}\right)$$

where $c_1 = 2^{22}$ suffices.

Further, we have $T \ge \frac{2}{\gamma}\left(\log\frac{2T(T+1)}{\delta} + d_G\log\left(\frac{e(|\mathcal{Y}|+1)}{d_G}\cdot\frac{T}{\tau}\log\frac{T(T+1)}{\delta}\right) + \log\frac{2\lambda^2}{\epsilon}\right)$ is satisfied whenever we have $T \ge \frac{2}{\gamma}\left((1+d_G)\log\frac{2T(T+1)}{\delta} + d_G\log\left(\frac{e(|\mathcal{Y}|+1)}{\tau d_G}\right) + \log\frac{2\lambda^2}{\epsilon}\right)$. We can achieve this with

$$T \ge \frac{2c_2}{\gamma}\left(d_G\log\frac{e(|\mathcal{Y}|+1)}{\tau d_G} + \log\frac{2\lambda^2}{\epsilon} + c_2(1+d_G)\log\left(\frac{4(1+d_G)}{\gamma\delta}\left(d_G\log\frac{e(|\mathcal{Y}|+1)}{\tau d_G} + \log\frac{2\lambda^2}{\epsilon}\right)\right)\right)$$

where $c_2 = 50$ suffices. □