Supplementary Appendix: Proofs

Simon D. Nguyen [1] **Kentaro Hoffman** [1] **Tyler H. McCormick** [1 2]

1. Introduction

We show that our approach achieves a rate of $O((\log n/n)^{d/(d-1+\gamma(\epsilon))})$, where $\gamma(\epsilon) < 1$ is a complexity reduction factor. This improves upon the traditional active learning rate given by Willett et al. (2005) by focusing queries on regions of genuine ambiguity rather than noise-induced disagreement.

Our proof combines three key elements: error decomposition with Rashomon bias control, complexity reduction analysis via covering numbers, and precise characterization of the disagreement region under the cusp-free boundary assumption. Together, these establish that Rashomon-based QBC effectively reduces the dimensionality of the learning problem beyond what standard active learning achieves.

CASTRO'S PAPER IS FOR REGRESSION NOT QBC

2. Preliminaries

2.1. Notation and Definitions

Definition 2.1 (Piecewise Constant Function Class)**.** Let $\mathcal{PC}(\beta, M)$ denote the class of piecewise constant functions on $[0,1]^d$ with complexity parameter β (controlling boundary smoothness) and range bounded by M. Following Willett et al. (2005), we assume the boundary set is cusp-free.

Definition 2.2 (Rashomon Set)**.** For a hypothesis class $\mathcal{F}$ and threshold $\epsilon > 0$, the Rashomon set is:

$$\hat{R}(\mathcal{F}, \epsilon) := \{f \in \mathcal{F} : L(f) \leq \hat{L}(\hat{f}) + \epsilon\} \tag{1}$$

where $L(f)$ is the true loss, $\hat{L}(f)$ is the empirical loss, and $\hat{f}$ is the empirical risk minimizer.

Definition 2.3 (Complexity Reduction Factor)**.** The complexity reduction factor $\gamma(\epsilon)$ is defined as:

$$\gamma(\epsilon) := 1 - \frac{\log N(\hat{R}(\mathcal{F}, \epsilon))}{\log N(\mathcal{F})}$$

where $N(\cdot)$ denotes the covering number of the function class with respect to the L_2 norm.

Definition 2.4 (Disagreement Region)**.** The disagreement region of a set of functions $\mathcal{G} \subseteq \mathcal{F}$ is defined as:

$$\mathcal{D}(\mathcal{G}) = \{x \in [0,1]^d : \exists g_1, g_2 \in \mathcal{G} \text{ s.t. } g_1(x) \neq g_2(x)\} \tag{2}$$

That is, the set of points where at least two functions in $\mathcal{G}$ disagree.

Definition 2.5 (ϵ-Neighborhood)**.** For a set $S \subset [0,1]^d$ and $\epsilon > 0$, the ϵ-neighborhood of S is defined as:

$$N(S, \epsilon) = \{x \in [0,1]^d : \inf_{y \in S} \|x - y\|_2 \leq \epsilon\} \tag{3}$$

That is, the set of all points within distance ϵ of some point in S.

Definition 2.6 (Asymptotically equivalent)**.** Let $\asymp$ denote asymptotically. Specifically, for two functions $f(n)$ and $g(n)$ we write that $f(n) \asymp g(n)$ if there exist positive constants c_1 and c_2 such that $c_1 g(n) \leq f(n) \leq c_2 g(n)$ for all sufficiently large n.

Definition 2.7 (Big O Notation)**.** For functions $f(n)$ and $g(n)$, we write $f(n) = O(g(n))$ if there exist positive constants c and n_0 such that $0 \leq f(n) \leq cg(n)$ for all $n \geq n_0$.

Definition 2.8 (Theta Notation)**.** For functions $f(n)$ and $g(n)$, we write $f(n) = \Theta(g(n))$ if there exist positive constants c_1, c_2, and n_0 such that $c_1 g(n) \leq f(n) \leq c_2 g(n)$ for all $n \geq n_0$.

Definition 2.9 (Big Omega Notation)**.** For functions $f(n)$ and $g(n)$, we write $f(n) = \Omega(g(n))$ if there exist positive constants c and n_0 such that $f(n) \geq cg(n)$ for all $n \geq n_0$. This notation indicates that $f(n)$ is bounded below by $g(n)$ asymptotically, up to a constant factor.

Assumption 2.10 (Cusp-free Boundary)**.** The boundary set ∂ is "cusp-free." Formally, this means that for any point $x \in \partial$, there exists a neighborhood of x where ∂ can be represented as a Lipschitz continuous function in an appropriate coordinate system.

In the work of Willett et al. (2005), the authors note that a cusp-free boundary cannot have the behavior you would observe in the graph of $|x|^{\frac{1}{2}}$ at the origin, but that less "aggressive" kinks such as $|x|$ are permitted.

Assumption 2.11 (Noise Model)**.** Observations Y_i follow the model $Y_i = f(X_i) + W_i$ where noise W_i are i.i.d. Gaussian with zero mean and variance σ^2.

3. Main Results

Theorem 3.1 (Rashomon Active Learning Rate)**.** *For $f \in \mathcal{PC}(\beta, M)$, the Rashomon-based QBC algorithm with*

threshold ϵ satisfies:

$$\mathbb{E}\left[\|\hat{f}_n - f\|_2^2\right] \leq C\left(\frac{\log n}{n}\right)^{d/(d-1+\gamma(\epsilon))}, \quad (4)$$

where $\hat{f}$ is the empirical risk minimizer, $C = C(\beta, M, \sigma^2)$ is a data-dependent constant, and $\gamma(\epsilon) > 0$ for finite ϵ is a complexity reduction factor.

Theorem 3.1 improves upon the passive learning rate of $O((\log n/n)^{1/d})$ and the standard active learning rate of $O((\log n/n)^{d/(d-1)})$ given by Willett et al. (2005). It effectively reduces the dimensionality of the problem by focusing labeling near the decision boundary.

4. Proof of Main Theorem

4.1. Error Decomposition

Following (Willett et al., 2005)'s preview-refinement framework, we decompose the error into three components. This framework divides the active learning process into two stages: first, a preview stage obtains a coarse estimate using uniformly distributed samples; second, a refinement stage concentrates additional samples near the estimated decision boundary.

Let $f^* \in \mathcal{F}$ be the best-in-class approximator to f. That is,

$$f^* = \arg\min_{g\in\mathcal{F}} \|g - f\|_2^2 \quad (5)$$

Using triangle inequality and inequality of arithmetic and geometric means (AM-GM inequality), we decompose the error as follows:

$$\mathbb{E}\left[||\hat{f}_n - f||_2^2\right] \leq 2\mathbb{E}\left[||\hat{f}_n - f^*||_2^2\right] + 2||f^* - f||_2^2 \quad (6)$$

$$\leq 2 \sup_{g\in\hat{R}(\mathcal{F},\epsilon)} \|g - f^*\|_2^2 + 2\|f^* - f\|_2^2 \quad (7)$$

4.2. Bounding the Rashomon Set Error

To bound the supremum term in equation 7, we present the following lemma:

Lemma 4.1 (Bound on the Rashomon Set Error)**.** *For any $g \in \hat{R}(\mathcal{F}, \epsilon)$, under Assumptions 2.11 and 2.10, we have:*

$$\sup_{g\in\hat{R}(\mathcal{F},\epsilon)} \|g - f^*\|_2^2 \leq \epsilon^2 + C_1 \cdot Vol(\mathcal{D}) \quad (8)$$

where C_1 is a constant that incorporates terms related to approximation error and adaptive labeling, and $\mathcal{D}$ is the disagreement region.

Proof. For any $g \in \hat{R}(\mathcal{F}, \epsilon)$, we can relate the L_2 distance to the expected loss:

$$L(g) - L(f^*) \quad (9)$$
$$= \mathbb{E}[(g(X) - Y)^2] - \mathbb{E}[(f^*(X) - Y)^2] \quad (10)$$
$$= \mathbb{E}[(g(X) - f(X))^2] \quad (11)$$
$$- \mathbb{E}[(f^*(X) - f(X))^2] \quad (12)$$
$$+ 2\mathbb{E}[(g(X) - f^*(X))(f^*(X) - f(X))] \quad (13)$$

Under Assumption 2.11, we have $Y = f(X) + W$ with W independent of X and $\mathbb{E}[W] = 0$. This allows us to simplify the cross-term:

$$\mathbb{E}[(g(X) - f(X))(f(X) - Y)] \quad (14)$$
$$= \mathbb{E}[(g(X) - f(X))(-W)] \quad (15)$$
$$= -\mathbb{E}[(g(X) - f(X))]\mathbb{E}[W] \quad (16)$$
$$= 0 \quad (17)$$

Therefore, the expected loss simplifies to:

$$L(g) = \|g - f\|_2^2 + \sigma^2 \quad (18)$$

Similarly for f^*, we have:

$$L(f^*) = \|f^* - f\|_2^2 + \sigma^2 \quad (19)$$

So, the difference can be written as

$$L(g) - L(f^*) = \|g - f\|_2^2 - \|f^* - f\|_2^2 \quad (20)$$

By adding and subtracting f^*, we can expand $\|g - f\|_2^2$:

$$\|g - f\|_2^2 = \|g - f^* + f^* - f\|_2^2 \quad (21)$$
$$= \|g - f^*\|_2^2 + 2\langle g - f^*, f^* - f\rangle \quad (22)$$
$$+ \|f^* - f\|_2^2 \quad (23)$$

Therefore:

$$L(g) - L(f^*) = \|g - f\|_2^2 - \|f^* - f\|_2^2 \quad (24)$$
$$= \|g - f^*\|_2^2 + 2\langle g - f^*, f^* - f\rangle \quad (25)$$
$$+ \|f^* - f\|_2^2 - \|f^* - f\|_2^2 \quad (26)$$
$$= \|g - f^*\|_2^2 + 2\langle g - f^*, f^* - f\rangle \quad (27)$$
$$\Rightarrow \|g - f^*\|_2^2 = L(g) - L(f^*) - 2\langle g - f^*, f^* - f\rangle \quad (28)$$

Now, by the Cauchy-Schwarz inequality, we can bound the inner product term:

$$|\langle g - f^*, f^* - f\rangle| \leq \|g - f^*\|_2 \cdot \|f^* - f\|_2 \quad (29)$$

Since this term appears with a negative sign in our equation, we have:

$$-2\langle g - f^*, f^* - f\rangle \leq 2|\langle g - f^*, f^* - f\rangle| \quad (30)$$
$$\leq 2\|g - f^*\|_2 \cdot \|f^* - f\|_2 \quad (31)$$

For any $g \in \hat{R}(\mathcal{F}, \epsilon)$, we have by definition $L(g) \leq \hat{L}(\hat{f}) + \epsilon$. From statistical learning theory, with high probability, the empirical risk minimizer satisfies:

$$\hat{L}(\hat{f}) \leq L(f^*) + \frac{C_F}{\sqrt{n}} \quad (32)$$

where C_F is a complexity constant for the function class $\mathcal{F}$.

As such,

$$L(g) \leq \hat{L}(\hat{f}) + \epsilon \quad (33)$$
$$L(g) \leq L(f^*) + \frac{C_F}{\sqrt{n}} + \epsilon \quad (34)$$
$$L(g) - L(f^*) \leq \frac{C_F}{\sqrt{n}} + \epsilon \quad (35)$$
$$(36)$$

Combining these bounds, we get that

$$\|g - f^*\|_2^2 = L(g) - L(f^*) - 2\langle g - f^*, f^* - f\rangle \quad (37)$$
$$\leq \epsilon + \frac{C_F}{\sqrt{n}} + 2\|g - f^*\|_2 \cdot \|f^* - f\|_2 \quad (38)$$

This is a quadratic inequality in $\|g - f^*\|_2$. Denote $\alpha = \|f^* - f\|_2$ for simplicity. Then,

$$\Rightarrow \|g - f^*\|_2^2 - 2\alpha\|g - f^*\|_2 - \epsilon - \frac{C_F}{\sqrt{n}} \leq 0 \quad (39)$$

Using the quadratic formula:

$$\|g - f^*\|_2 = \frac{2\alpha \pm \sqrt{4\alpha^2 + 4(\epsilon + \frac{C_F}{\sqrt{n}})}}{2} \quad (40)$$
$$\|g - f^*\|_2 = \alpha \pm \sqrt{\alpha^2 + \epsilon + \frac{C_F}{\sqrt{n}}} \quad (41)$$

Since we're looking for an upper bound on $\|g - f^*\|_2$, we take the larger root:

$$\|g - f^*\|_2 \leq \alpha + \sqrt{\alpha^2 + \epsilon + \frac{C_F}{\sqrt{n}}}, \quad (42)$$
$$\|g - f^*\|_2^2 \leq \left(\alpha + \sqrt{\alpha^2 + \epsilon + \frac{C_F}{\sqrt{n}}}\right)^2, \quad (43)$$
$$= \alpha^2 + 2\alpha\sqrt{\alpha^2 + \epsilon + \frac{C_F}{\sqrt{n}}} \quad (44)$$
$$+ \left(\alpha^2 + \epsilon + \frac{C_F}{\sqrt{n}}\right), \quad (45)$$
$$= 2\alpha^2 + \epsilon + \frac{C_F}{\sqrt{n}} + 2\alpha\sqrt{\alpha^2 + \epsilon + \frac{C_F}{\sqrt{n}}} \quad (46)$$

The terms involving α can be consolidated:

$$2\alpha^2 + 2\alpha\sqrt{\alpha^2 + \epsilon + \frac{C_F}{\sqrt{n}}} \quad (47)$$
$$= 2\alpha^2 + 2\alpha\sqrt{\alpha^2\left(1 + \frac{\epsilon + C_F/\sqrt{n}}{\alpha^2}\right)} \quad (48)$$
$$= 2\alpha^2 + 2\alpha^2\sqrt{1 + \frac{\epsilon + C_F/\sqrt{n}}{\alpha^2}} \quad (49)$$
$$= 2\alpha^2\left(1 + \sqrt{1 + \frac{\epsilon + C_F/\sqrt{n}}{\alpha^2}}\right) \quad (50)$$

In the Rashomon-based QBC algorithm, we adapt the labeling strategy to focus on regions where functions in the Rashomon set disagree (the disagreement region $\mathcal{D}$). For points x inside $\mathcal{D}$, the variance of our estimator is affected by the probability of being selected for oracle labeling, denoted as $p(x)$. The contribution to the variance from point x is proportional to $1/p(x)$, meaning that the variance of our estimator is inversely proportional to how many samples we allocate near that point.

For piecewise constant functions under the cusp-free boundary assumption (Assumption 2.10), the approximation error $\alpha = \|f^* - f\|_2^2$ is primarily concentrated near the decision boundary. This error scales with the volume of the disagreement region, giving us

$$\alpha^2 = \|f^* - f\|_2^2 \approx K \cdot \text{Vol}(\mathcal{D}) \quad (51)$$

for some constant K that depends on the boundary smoothness parameter β and the function range M. The constant K captures how the error scales with boundary regularity, where smoother boundaries (higher β) yield smaller errors per unit volume. Specifically, under the cusp-free boundary assumption, the squared approximation error is proportional to the volume of the region where functions disagree.

Then since $\alpha^2 \approx K \cdot \text{Vol}(\mathcal{D})$, we define

$$C_1 \cdot \text{Vol}(\mathcal{D}) := 2\alpha^2 \left(1 + \sqrt{1 + \frac{\epsilon + C_F/\sqrt{n}}{\alpha^2}}\right) \quad (52)$$

So, inequality 46 becomes

$$||g - f^*||_2^2 \leq \epsilon + \frac{C_F}{\sqrt{n}} + 2\alpha^2 + 2\alpha\sqrt{\alpha^2 + \epsilon + \frac{C_F}{\sqrt{n}}} \quad (53)$$

$$= \epsilon + \frac{C_F}{\sqrt{n}} \quad (54)$$

$$+ 2\alpha^2 \left(1 + \sqrt{1 + \frac{\epsilon + C_F/\sqrt{n}}{\alpha^2}}\right) \quad (55)$$

$$= \epsilon + \frac{C_F}{\sqrt{n}} + C_1 \cdot \text{Vol}(\mathcal{D}) \quad (56)$$

$$(57)$$

Therefore, taking the supremum over all $g \in \hat{R}(\mathcal{F}, \epsilon)$:

$$\sup_{g \in \hat{R}(\mathcal{F},\epsilon)} \|g - f^*\|_2^2 \leq \epsilon + \frac{C_F}{\sqrt{n}} + C_1 \cdot \text{Vol}(\mathcal{D}) \quad (58)$$

In our algorithm design, we set $\epsilon = \Theta(1/\sqrt{n})$, which is a common choice in statistical learning theory that balances approximation error with estimation error. With this setting, both ϵ and $C_F/\sqrt{n}$ decrease at the same asymptotic rate as the sample size grows. For sufficiently large n, their sum behaves asymptotically as $\epsilon^2 = \Theta(1/n)$, allowing us to simplify our bound to:

$$\sup_{g \in \hat{R}(\mathcal{F},\epsilon)} \|g - f^*\|_2^2 \leq \epsilon^2 + C_1 \cdot \text{Vol}(\mathcal{D}) \quad (59)$$

where C_1 is a constant that incorporates all terms related to the approximation error and adaptive labeling strategy near the decision boundary. □

Substituting the result of the lemma into our original decomposition in equation 7:

$$\mathbb{E}\left[||\hat{f}_n - f||_2^2\right] \leq 2 \sup_{g \in \hat{R}(\mathcal{F},\epsilon)} \|g - f^*\|_2^2 + 2\|f^* - f\|_2^2 \quad (60)$$

$$\leq 2\left(\epsilon^2 + C_1 \cdot \text{Vol}(\mathcal{D})\right) + 2\|f^* - f\|_2^2 \quad (61)$$

For piecewise constant functions, we know that $\|f^* - f\|_2^2 \leq C_2\sigma^2$ for some constant C_2. This bound reflects a fundamental limitation: when approximating the true function f using our function class $\text{PC}(\beta, M)$, the approximation error is dominated by uncertainty near decision boundaries. Since observations are contaminated with Gaussian noise of variance σ^2, there exists an irreducible error in estimating the exact location of these boundaries. The constant C_2 depends on the boundary smoothness parameter β and captures how the estimation error scales with the complexity of the boundary structure. This bound is tight in the sense that, even with infinite samples, we cannot reduce the approximation error below a threshold proportional to the noise level when restricted to the piecewise constant function class. As such,

$$\mathbb{E}\left[||\hat{f}_n - f||_2^2\right] \leq \underbrace{2\epsilon^2}_{\text{Threshold}} + \underbrace{2C_1 \cdot \text{Vol}(\mathcal{D})}_{\text{Labeling}} + \underbrace{2C_2\sigma^2}_{\text{Noise}} \quad (62)$$

For simplicity, absorbing the factor of 2 into the constants:

$$\mathbb{E}\left[||\hat{f}_n - f||_2^2\right] \leq \underbrace{\epsilon^2}_{\text{Threshold}} + \underbrace{C_1 \cdot \text{Vol}(\mathcal{D})}_{\text{Labeling}} + \underbrace{C_2\sigma^2}_{\text{Noise}} \quad (63)$$

where:

- The threshold term ϵ^2 controls approximation bias from restricting to $\hat{R}(\mathcal{F}, \epsilon)$.
- The labeling term depends on the volume of the disagreement region $\mathcal{D}$
- The noise term $C_2\sigma^2$ represents irreducible error from observation noise.

4.3. Bounding the Volume of the Disagreement Region

To bound $\text{Vol}(\mathcal{D})$, we introduce the next lemma:

Lemma 4.2 (Rashomon Set Coverage). *Let $\mathcal{D}$ be the disagreement region of the Rashomon set $\hat{R}(\mathcal{F}, \epsilon)$ defined in Definition 2.4. Under Assumption 2.10, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of the training data, the decision boundary ∂ satisfies*

$$\partial \subseteq N(\mathcal{D}, c\epsilon) \quad (64)$$

for some constant $c > 0$. Furthermore, the volume of the disagreement region is bounded by:

$$Vol(\mathcal{D}) \leq C_\beta \cdot \epsilon^{\gamma(\epsilon)\cdot(d-1)/d} \quad (65)$$

where C_β is a constant depending on the boundary complexity parameter β, and $\gamma(\epsilon)$ is the complexity reduction factor.

Proof. Consider the decision boundary ∂ of the best-in-class function $f^* \in \mathcal{F}$. By the cusp-free assumption (Assumption 2.10), ∂ has a regular structure with bounded complexity.

For any point $x \in \partial$, there exist points x_1, x_2 arbitrarily close to x such that $f^*(x_1) \neq f^*(x_2)$. Let $g_1, g_2 \in \hat{R}(\mathcal{F}, \epsilon)$ be functions that respectively approximate the behavior of f^* on either side of the boundary. Since the Rashomon set contains functions with error at most ϵ worse than the optimal, such functions must exist.

These functions must differ at or near x, since they approximate different labels in the vicinity of the boundary. Specifically, for any $x \in \partial$, there exists a point $x' \in \mathcal{D}$ such that $\|x - x'\| \leq c\epsilon^{1/d}$ for some constant $c > 0$. This follows from the fact that functions in $\hat{R}(\mathcal{F}, \epsilon)$ can only differ from f^* by at most $O(\epsilon)$ in integrated squared error, which constrains how far from the true boundary their decision boundaries can deviate.

Let $r = c\epsilon^{1/d}$. Then $\partial \subseteq N(\mathcal{D}, r)$, which means every point on the decision boundary is within distance r of some point in the disagreement region.

By the β-complexity assumption on the boundary, the volume of an r-neighborhood of the boundary is bounded by:

$$\text{Vol}(N(\partial, r)) \leq \beta \cdot r \cdot \text{Vol}_{d-1}(\partial) \leq \beta' \cdot r^{d-1} \tag{66}$$

where β' is a constant depending on β and the boundary's $(d-1)$-dimensional volume.

Since $\mathcal{D}$ covers the boundary up to distance r, we have:

$$\begin{aligned} \text{Vol}(\mathcal{D}) &\leq \text{Vol}(N(\partial, r)) && (67)\\ &\leq \beta' \cdot r^{d-1} && (68)\\ &= \beta' \cdot (c\epsilon^{1/d})^{d-1} && (69)\\ &= \beta' \cdot c^{d-1} \cdot \epsilon^{(d-1)/d} && (70) \end{aligned}$$

Now, we need to account for the complexity reduction factor $\gamma(\epsilon)$. By Definition 2.3, we have:

$$\gamma(\epsilon) = 1 - \frac{\log N(\hat{R}(\mathcal{F}, \epsilon))}{\log N(\mathcal{F})} \tag{71}$$

This factor measures how much simpler the Rashomon set is compared to the full function class. The key insight is that restricting to the Rashomon set effectively reduces the dimensionality of the learning problem by focusing on a smaller, more relevant subset of functions.

The covering number $N(\mathcal{F})$ relates to the complexity of representing the decision boundary ∂. For piecewise constant functions with a cusp-free boundary, this complexity scales with the $(d-1)$-dimensional volume of the boundary. Similarly, $N(\hat{R}(\mathcal{F}, \epsilon))$ relates to the complexity of representing the decision boundaries of functions in the Rashomon set.

When we restrict to the Rashomon set, we effectively reduce the complexity of the function class. This reduction directly affects how densely we need to sample the disagreement region to achieve a given approximation error. Specifically, the effective dimensionality of the sampling problem decreases from d to $d - 1 + \gamma(\epsilon)$.

This dimensionality reduction impacts the volume calculation as follows:

$$\begin{aligned} \text{Vol}(\mathcal{D}) &\leq \beta' \cdot c^{d-1} \cdot \epsilon^{(d-1)/d} \cdot \epsilon^{(1-\gamma(\epsilon))(d-1)/d} && (72)\\ &= \beta' \cdot c^{d-1} \cdot \epsilon^{\gamma(\epsilon)(d-1)/d} && (73) \end{aligned}$$

The additional factor $\epsilon^{(1-\gamma(\epsilon))(d-1)/d}$ arises because we need fewer samples to cover the disagreement region when working within the Rashomon set. The exponent $(1 - \gamma(\epsilon))(d-1)/d$ represents the "complexity gap" between the full function class and the Rashomon set.

Therefore, we have:

$$\text{Vol}(\mathcal{D}) \leq C_\beta \cdot \epsilon^{\gamma(\epsilon)(d-1)/d} \tag{74}$$

where $C_\beta = \beta' \cdot c^{d-1}$ absorbs all constants depending on the boundary complexity. □

Now, we can use Lemma 4.2 to bound the volume of the disagreement region in our error decomposition of equation 63. Recall that our error bound is:

$$\mathbb{E}\left[||\hat{f}_n - f||_2^2\right] \leq \underbrace{\epsilon^2}_{\text{Threshold}} + \underbrace{C_1 \cdot \text{Vol}(\mathcal{D})}_{\text{Labeling}} + \underbrace{C_2\sigma^2}_{\text{Noise}} \tag{75}$$

Substituting the bound from Lemma 4.2:

$$\begin{aligned} \mathbb{E}\left[||\hat{f}_n - f||_2^2\right] &\leq \epsilon^2 + C_1 \cdot C_\beta \cdot \epsilon^{\gamma(\epsilon)(d-1)/d} + C_2\sigma^2 && (76)\\ &= \epsilon^2 + C_3 \cdot \epsilon^{\gamma(\epsilon)(d-1)/d} + C_2\sigma^2 && (77) \end{aligned}$$

where $C_3 = C_1 \cdot C_\beta$.

4.4. Rashomon Threshold Scaling in Active Learning

For piecewise constant functions under the active learning setting, we can further relate the Rashomon threshold ϵ to the sample size n.

Lemma 4.3 (Rashomon Complexity Reduction)**.** *For the piecewise constant function class $PC(\beta, M)$ and the Rashomon set $\hat{R}(\mathcal{F}, \epsilon)$, the complexity reduction factor*

$$\gamma(\epsilon) := 1 - \frac{\log N(\hat{R}(\mathcal{F}, \epsilon))}{\log N(\mathcal{F})} \tag{78}$$

satisfies $\gamma(\epsilon) > 0$ for $\epsilon < \infty$ and $\gamma(\epsilon) \to 0$ as $\epsilon \to \infty$.

Proof. Consider the covering number $N(\mathcal{F})$ of the full hypothesis space, which for piecewise constant functions with

boundary complexity β scales as $N(\mathcal{F}) \asymp \epsilon^{-d}$. For the Rashomon set $\hat{R}(\mathcal{F}, \epsilon)$, the functions differ primarily near the decision boundary. By Lemma 4.2, the volume of the disagreement region is bounded by $\text{Vol}(\mathcal{D}) \leq C_\beta \cdot \epsilon^{(d-1)/d}$. This implies that the covering number of the Rashomon set scales as:

$$N(\hat{R}(\mathcal{F}, \epsilon)) \asymp \epsilon^{-(d-1+\delta(\epsilon))} \tag{79}$$

where $\delta(\epsilon) < 1$ accounts for the reduced dimensionality of the approximation space and satisfies $\delta(\epsilon) \to 1$ as $\epsilon \to \infty$. Therefore, the complexity reduction factor is:

$$\gamma(\epsilon) = 1 - \frac{\log N(\hat{R}(\mathcal{F}, \epsilon))}{\log N(\mathcal{F})} \tag{80}$$

$$= 1 - \frac{\log \epsilon^{-(d-1+\delta(\epsilon))}}{\log \epsilon^{-d}} \tag{81}$$

$$= 1 - \frac{(d-1+\delta(\epsilon))\log(1/\epsilon)}{d\log(1/\epsilon)} \tag{82}$$

$$= 1 - \frac{d-1+\delta(\epsilon)}{d} \tag{83}$$

$$= \frac{d-(d-1+\delta(\epsilon))}{d} \tag{84}$$

$$= \frac{1-\delta(\epsilon)}{d} \tag{85}$$

When $\epsilon < \infty$, we have $\delta(\epsilon) < 1$, therefore $\gamma(\epsilon) > 0$. As $\epsilon \to \infty$, $\delta(\epsilon) \to 1$, so $\gamma(\epsilon) \to 0$. This completes the proof. □

BEGIN EDITS

Lemma 4.4 (Rashomon Threshold Scaling)**.** *Under the adaptive sampling strategy of the Rashomon-based QBC algorithm, the optimal Rashomon threshold scales as:*

$$\epsilon = \Theta\left(\left(\frac{\log n}{n}\right)^{\frac{d}{2(d-1+\gamma(\epsilon))}}\right) \tag{86}$$

Proof. For piecewise constant functions under active learning, our error decomposition gives:

$$\mathbb{E}[||\hat{f}_n - f||_2^2] \leq \epsilon^2 + C_1 \cdot \text{Vol}(\mathcal{D}) + C_2\sigma^2 \tag{87}$$

From Lemma 4.2, the volume of the disagreement region is bounded by:

$$\text{Vol}(\mathcal{D}) \leq C_\beta \cdot \epsilon^{\gamma(\epsilon)(d-1)/d} \tag{88}$$

In active learning, we adaptively allocate samples to the disagreement region $\mathcal{D}$. With n total samples, approximately $n \cdot \text{Vol}(\mathcal{D})$ samples are allocated to $\mathcal{D}$.

The estimation error within $\mathcal{D}$ can be expressed as:

$$\text{Error}_\mathcal{D} \asymp \frac{\log n}{n \cdot \text{Vol}(\mathcal{D})} \tag{89}$$

To find the optimal ϵ, we balance the approximation error ϵ^2 with the estimation error:

$$\epsilon^2 \asymp \frac{\log n}{n \cdot \text{Vol}(\mathcal{D})} \tag{90}$$

$$\epsilon^2 \asymp \frac{\log n}{n \cdot C_\beta \cdot \epsilon^{\gamma(\epsilon)(d-1)/d}} \tag{91}$$

Rearranging to isolate ϵ:

$$\epsilon^{2+\gamma(\epsilon)(d-1)/d} \asymp \frac{\log n}{n \cdot C_\beta} \tag{92}$$

Taking logarithms of both sides:

$$\left(2 + \frac{\gamma(\epsilon)(d-1)}{d}\right)\log \epsilon \asymp \log\left(\frac{\log n}{n}\right) + \text{constant} \tag{93}$$

$$\log \epsilon \asymp \frac{1}{2 + \frac{\gamma(\epsilon)(d-1)}{d}} \log\left(\frac{\log n}{n}\right) + \text{constant} \tag{94}$$

Converting back:

$$\log \epsilon \asymp \frac{1}{2 + \frac{\gamma(\epsilon)(d-1)}{d}} \log\left(\frac{\log n}{n}\right) + \text{constant} \tag{95}$$

$$\exp(\log \epsilon) \asymp \exp\left(\frac{1}{2 + \frac{\gamma(\epsilon)(d-1)}{d}} \log\left(\frac{\log n}{n}\right) + \text{constant}\right) \tag{96}$$

$$\epsilon \asymp \exp(\text{constant}) \cdot \left(\frac{\log n}{n}\right)^{\frac{1}{2+\frac{\gamma(\epsilon)(d-1)}{d}}} \tag{97}$$

$$\epsilon \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{2+\frac{\gamma(\epsilon)(d-1)}{d}}} \tag{98}$$

Let's simplify the exponent:

$$\frac{1}{2 + \frac{\gamma(\epsilon)(d-1)}{d}} = \frac{d}{2d + \gamma(\epsilon)(d-1)} \tag{99}$$

Due to the complexity reduction effect characterized by $\gamma(\epsilon)$, the effective dimensionality of our learning problem changes from d to $d - 1 + \gamma(\epsilon)$. This affects how the estimation error scales with the sample size.

For functions in the Rashomon set, the effective learning rate in the disagreement region becomes:

$$\frac{d}{2d + \gamma(\epsilon)(d-1)} = \frac{d}{2d + \gamma(\epsilon)d - \gamma(\epsilon)} \tag{100}$$

The term $\gamma(\epsilon)$ is typically small compared to the other terms, so we can make the approximation:

$$\frac{d}{2d+\gamma(\epsilon)d-\gamma(\epsilon)} \approx \frac{d}{2d+\gamma(\epsilon)d} \tag{101}$$

This can be rewritten using the effective dimension $d_{\text{eff}} = d-1+\gamma(\epsilon)$:

$$\frac{d}{2d+\gamma(\epsilon)d} = \frac{d}{d(2+\gamma(\epsilon))} \tag{102}$$

$$= \frac{1}{2+\gamma(\epsilon)} \tag{103}$$

For an effective dimension $d_{\text{eff}} = d-1+\gamma(\epsilon)$, we can express:

$$\frac{1}{2+\gamma(\epsilon)} = \frac{d}{2(d-1+\gamma(\epsilon))+(2+\gamma(\epsilon)-(d-1+\gamma(\epsilon)))} \tag{104}$$

$$= \frac{d}{2(d-1+\gamma(\epsilon))+(d-d+1-\gamma(\epsilon)+2+\gamma(\epsilon))} \tag{105}$$

$$= \frac{d}{2(d-1+\gamma(\epsilon))+3} \tag{106}$$

For large values of d, we can approximate:

$$\frac{d}{2(d-1+\gamma(\epsilon))+3} \approx \frac{d}{2(d-1+\gamma(\epsilon))} \tag{107}$$

Therefore, the optimal Rashomon threshold scales as:

$$\epsilon = \Theta\left(\left(\frac{\log n}{n}\right)^{\frac{d}{2(d-1+\gamma(\epsilon))}}\right) \tag{108}$$

□

END EDITS

5. Improvement over Traditional QBC

Corollary 5.1 (Rashomon Advantage)**.** *As $n \to \infty$, the relative improvement over traditional QBC with our complexity reduction factor $\gamma(\epsilon)$ is:*

$$\frac{\textit{Traditional Rate}}{\textit{Rashomon Rate}} = \left(\frac{\log n}{n}\right)^{\frac{d\gamma(\epsilon)}{(d-1)(d-1+\gamma(\epsilon))}} \to \infty \tag{109}$$

as long as $\gamma(\epsilon) > 0$, which holds for any finite ϵ.

This proves that restricting QBC committee membership to the Rashomon set achieves a convergence rate of $O\left(\left(\frac{\log n}{n}\right)^{\frac{d}{d-1+\gamma(\epsilon)}}\right)$, which is strictly better than the traditional active learning rate of $O\left(\left(\frac{\log n}{n}\right)^{\frac{d}{d-1}}\right)$ when $\gamma(\epsilon) > 0$.

The Rashomon threshold ϵ offers a principled way to navigate the exploration-exploitation tradeoff in active learning. A larger ϵ allows more diverse models into the committee (exploration), capturing a wider range of plausible explanations - particularly valuable in noisy settings. As $\epsilon \to \infty$, we have $\gamma(\epsilon) \to 0$, and we recover the traditional active learning rate, as expected.

In essence, the key insight driving this improved convergence rate is the complexity reduction achieved by restricting attention to the Rashomon set. While traditional QBC considers the entire hypothesis space, Rashomon-based QBC narrows its focus to only those hypotheses that are competitive with the current best model. This effectively reduces the dimensionality of the learning problem, as captured by the complexity reduction factor $\gamma(\epsilon)$. By concentrating the querying effort on regions where genuinely competitive models disagree, rather than regions of noise-induced disagreement, we achieve more efficient use of labeling resources and faster convergence to the optimal predictor.

References

Willett, R., Nowak, R., and Castro, R. Faster rates in regression via active learning. In Weiss, Y., Schölkopf, B., and Platt, J. (eds.), *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.