The appendix is organized as follows: in Appendix A, we provide proofs of the claims made in the main text. Next, in Appendix B, we provide algorithms and time complexity analyses for the algorithms referenced in the main text. Appendix C contains additional experimental results. In Appendix D, we provide additional details about our experimental setup. Then in Appendix E, we discuss the broader impacts and limitations of LOKI. Finally, in Appendix F, we provide more details and examples of the locus visualizations shown in Figure 1.

A Deferred Proofs

A.1 Proof of Theorem 4.1 (LOKI sample complexity)

In this section, we provide a formal statement and proof of our sample complexity result for LOKI, including additional required definitions and assumptions.

First, we define the required tools to prove the sample complexity bound for LOKI. For the purposes of this proof, we define the Fréchet variance as the function over which the Fréchet mean returns the minimizer.

Definition A.1 (Fréchet variance)**.** The Fréchet variance is defined as

$$\Psi_{\mathbf{w}}(V) := \sum_{i \in [K]} \mathbf{w}_i d^2(V, V_i).$$

Additionally, we will require a technical assumption related to the sensitivity of the Fréchet variance to different node choices.

Assumption A.2 (Fréchet variance is $\frac{1}{\alpha}$-bi-Lipschitz)**.** For a metric space defined by a graph $G = (\mathcal{V}, \mathcal{E})$, the Fréchet variance is K-bi-Lipschitz if there exists a $K \geq 1$ such that

$$\frac{1}{K} d^2(V, \tilde{V}) \leq |\Psi_{\mathbf{w}}(V) - \Psi_{\mathbf{w}}(\tilde{V})| \leq K d^2(V, \tilde{V})$$

for all $V, \tilde{V} \in \mathcal{V}$, and a fixed $\mathbf{w} \in \Delta^{K-1}$. For our purposes, such a K always exists: consider setting $K = \text{diam}(G)^2 \max_{V_1, V_2 \in \mathcal{V}} |\Psi_{\mathbf{w}}(V_1) - \Psi_{\mathbf{w}}(V_2)|$. However, this is a very conservative bound that holds for all graphs that we consider. Instead, we assume access to the largest $\alpha = \frac{1}{K} \leq 1$ such that

$$\alpha d^2(V, \tilde{V}) \leq |\Psi_{\mathbf{w}}(V) - \Psi_{\mathbf{w}}(\tilde{V})|,$$

which may be problem dependent.

Theorem A.3 (LOKI sample complexity)**.** *Let $\mathcal{Y}$ be a set of points on the d dimensional 2-norm ball of radius R, and let $\Lambda \subseteq \mathcal{Y}$ be the set of K observed classes. Assume that Λ forms a $2R/(\sqrt{d}K - 1)$-net under the Euclidean distance. Assume that training examples are generated by drawing n samples from the following process: draw $x \sim \mathcal{N}(y, I)$ where $y \sim$ Unif(Λ), and at test time, draw $x \sim \mathcal{N}(y, I)$ where $y \sim$ Unif$(\mathcal{Y})$. Assume that we estimate a Gaussian mixture model with K components (each having identity covariance) on the training set and obtain probability estimates $\hat{\mathbb{P}}(y_i|x)$ for $i \in [K]$ for a sample (x, y_*) from the test distribution. Then with high probability, under the following model,*

$$\hat{y}_* \in m_\Lambda([\hat{\mathbb{P}}(y_i|x)]_{i \in [K]}) = \arg\min_{y \in \mathcal{Y}} \sum_{i \in [K]} \hat{\mathbb{P}}(y_i|x) d^2(y, y_i)$$

the sample complexity of estimating target y_* from the test distribution $\mathcal{D}_{test}$ with prediction $\hat{y}_*$ is:

$$\mathbb{E}_{(x,y_*) \sim \mathcal{D}_{test}}[d^2(y_*, \hat{y}_*)] \leq O\left(\frac{d}{\alpha}\sqrt{\frac{\log K/\delta}{n}}\left(\frac{1}{\left(R^{1-\frac{2}{d}} - \frac{\log R}{R}\right)} + \sqrt{d}\right)\right)$$

where d is the dimensionality of the input and α is our parameter under Assumption A.2.

Proof. We begin by detailing the data-generating process.

At training time, our underlying data-generating process, $\mathcal{D}_{\text{train}}$, is as follows:

- We begin with a $\frac{2R}{\sqrt{dK-1}}$-net, $\mathcal{Y}$, on the d dimensional 2-norm ball of radius R under the Euclidean distance, and let $\Lambda \subseteq \mathcal{Y}$
- Draw $y \sim$ Unif$(\mathcal{Y})$.
- Discard draws if $y \notin \Lambda$.
- Draw $x \sim \mathcal{N}(y, I)$.

At test time, we do not discard draws and allow for classes not in Λ and use the following data generating process, $\mathcal{D}_{\text{test}}$:

- We begin with a $\frac{2R}{\sqrt{dK-1}}$-net, $\mathcal{Y}$, on the d dimensional 2-norm ball of radius R under the Euclidean distance, and let $\Lambda \subseteq \mathcal{Y}$
- Draw $y \sim$ Unif$(\mathcal{Y})$.
- Draw $x \sim \mathcal{N}(y, I)$.

Given a labeled training dataset $D = \{(x_i, y_i)\}_{i=1}^n$ containing n points drawn from $\mathcal{D}_{\text{train}}$ with $|\Lambda| = K$ distinct classes, we would like to fit a K-component Gaussian mixture model with identity covariance. We first perform mean estimation of each of the classes separately using the median-of-means estimator [15, 27, 24]. Using the estimator yields the following parameter estimation bound [27, 4]:

$$||y_i - \hat{y}_i||_2 \leq O\left(\sqrt{\frac{d \log K/\delta}{n}}\right)$$

with probability $1 - \delta$.

Next, we consider the relationships between four quantities: $\Psi_{\mathbb{P}}(y_*)$, $\Psi_{\mathbb{P}}(\hat{y}_*)$, $\Psi_{\widehat{\mathbb{P}}}(y_*)$, $\Psi_{\widehat{\mathbb{P}}}(\hat{y}_*)$, where $\mathbb{P}$ is the vector of probabilities from the true Gaussian mixture, $\widehat{\mathbb{P}}$ is the vector of probabilities from the estimated model, $y_* \in m_\Lambda(\mathbb{P})$ is the target class, and $\hat{y}_* \in m_\Lambda(\widehat{\mathbb{P}})$ is the predicted class. While it is problem-dependent as to whether $\Psi_{\mathbb{P}}(y_*) \leq \Psi_{\widehat{\mathbb{P}}}(\hat{y}_*)$ or $\Psi_{\mathbb{P}}(y_*) > \Psi_{\widehat{\mathbb{P}}}(\hat{y}_*)$, a similar argument holds for both cases. So without loss of generality, we assume that $\Psi_{\mathbb{P}}(y_*) \leq \Psi_{\widehat{\mathbb{P}}}(\hat{y}_*)$. Then by the definition of the Fréchet mean, the following inequalities hold:

$$\Psi_{\mathbb{P}}(y_*) \leq \Psi_{\widehat{\mathbb{P}}}(\hat{y}_*) \leq \Psi_{\widehat{\mathbb{P}}}(y_*),$$

and consequently,

$$\Psi_{\widehat{\mathbb{P}}}(\hat{y}_*) - \Psi_{\mathbb{P}}(y_*) \leq \Psi_{\widehat{\mathbb{P}}}(y_*) - \Psi_{\mathbb{P}}(y_*). \tag{2}$$

We proceed by obtaining upper and lower bounds of the existing bounds in Equation 2. First, we will obtain an upper bound on $\Psi_{\widehat{\mathbb{P}}}(y_*) - \Psi_{\mathbb{P}}(y_*)$.

$$\begin{aligned}|\Psi_{\widehat{\mathbb{P}}}(y_*) - \Psi_{\mathbb{P}}(y_*)| &= \left|\sum_{i \in [K]} (\widehat{\mathbb{P}} - \mathbb{P}) d^2(y_*, V_i)\right| \\ &= \left|\sum_{i \in [K]} \left(\widehat{\mathbb{P}}(y_i|x) - \mathbb{P}(y_i|x)\right) ||y_* - y_i||_2^2\right| \\ &= \left|\sum_{i \in [K]} \left(\frac{\widehat{\mathbb{P}}(x|y_i)\mathbb{P}(y_i)}{\sum_{j \in [K]} \widehat{\mathbb{P}}(x|y_j)\mathbb{P}(y_j)} - \frac{\mathbb{P}(x|y_i)\mathbb{P}(y_i)}{\sum_{j \in [K]} \mathbb{P}(x|y_j)\mathbb{P}(y_j)}\right) ||y_* - y_i||_2^2\right|\end{aligned} \tag{3}$$

$$\begin{aligned}
&= \left| \sum_{i\in[K]} \left(\frac{\exp\{-\frac{1}{2}||x-\hat{y}_i||_2^2\}\frac{1}{K}}{\sum_{j\in[K]}\exp\{-\frac{1}{2}||x-\hat{y}_j||_2^2\}\frac{1}{K}} \right.\right. \\
&\qquad\qquad \left.\left. - \frac{\exp\{-\frac{1}{2}||x-y_i||_2^2\}\frac{1}{K}}{\sum_{j\in[K]}\exp\{-\frac{1}{2}||x-y_j||_2^2\}\frac{1}{K}} \right) ||y_* - y_i||_2^2 \right|. \\
&= \left| \sum_{i\in[K]} \left(\frac{\exp\{-\frac{1}{2}||x-\hat{y}_i||_2^2\}}{\sum_{j\in[K]}\exp\{-\frac{1}{2}||x-\hat{y}_j||_2^2\}} \right.\right. \\
&\qquad\qquad \left.\left. - \frac{\exp\{-\frac{1}{2}||x-y_i||_2^2\}}{\sum_{j\in[K]}\exp\{-\frac{1}{2}||x-y_j||_2^2\}} \right) ||y_* - y_i||_2^2 \right|. \qquad (4)
\end{aligned}$$

For notational convenience, we define the following:
$a_i := \exp\{-\frac{1}{2}||x-y_i||_2^2\}$,
$\hat{a}_i := \exp\{-\frac{1}{2}||x-\hat{y}_i||_2^2\}$,
$b := \sum_{j\in[K]}\exp\{-\frac{1}{2}||x-y_j||_2^2\}$,
$\hat{b} := \sum_{j\in[K]}\exp\{-\frac{1}{2}||x-\hat{y}_j||_2^2\}$, and
$c_i := ||\hat{y}_* - y_i||_2^2$.

Then (4) becomes:

$$\begin{aligned}
\left| \sum_{i\in[K]} \left(\frac{\hat{a}_i}{\hat{b}} - \frac{a_i}{b} \right) c_i \right| &= \left| \sum_{i\in[K]} \left(\frac{\hat{a}_i}{\hat{b}} - \frac{\hat{a}_i}{b} + \frac{\hat{a}_i}{b} - \frac{a_i}{b} \right) c_i \right| \\
&= \left| \sum_{i\in[K]} \left(\frac{\hat{a}_i - a_i}{b} + \hat{a}_i \left(\frac{1}{\hat{b}} - \frac{1}{b} \right) \right) c_i \right| \\
&\le \left| \sum_{i\in[K]} \left(\frac{\hat{a}_i - a_i}{b} \right) c_i \right| + \left| \sum_{i\in[K]} \left(\hat{a}_i \left(\frac{1}{\hat{b}} - \frac{1}{b} \right) \right) c_i \right| \\
&= \left| \sum_{i\in[K]} \frac{a_i}{b} \left(\frac{\hat{a}_i}{a_i} - 1 \right) c_i \right| + \left| \frac{b-\hat{b}}{b} \sum_{i\in[K]} \frac{\hat{a}_i}{\hat{b}} c_i \right| \\
&= \left| \sum_{i\in[K]} \frac{a_i}{b} \left(\frac{\hat{a}_i}{a_i} - 1 \right) c_i \right| + \left| \left(\sum_{i\in[K]} \frac{a_i}{b} \left(\frac{\hat{a}_i}{a_i} - 1 \right) \right) \left(\sum_{i\in[K]} \frac{\hat{a}_i}{\hat{b}} c_i \right) \right|. \qquad (5)
\end{aligned}$$

Now we define the following: $L := ||x-y_z||_2^2$ with $z \in \arg\min_j ||x-y_j||_2^2$ is the smallest distance to a class mean, and $L + E_i := ||x-y_i||_2^2$ with $E_i > 0$. Similarly, define $\hat{L} := ||x-\hat{y}_z||_2^2$ with $z \in \arg\min_j ||x-\hat{y}_j||_2^2$ is the smallest distance to an estimated class mean, and $\hat{L}+\hat{E}_i := ||x-\hat{y}_i||_2^2$ with $\hat{E}_i > 0$. Finally, we define $T := \sqrt{\hat{L}} + \sqrt{\hat{L}+\hat{E}_i} + \sqrt{L} + \sqrt{L+E_i}$.

Then we can bound the parts separately:

For $i \neq z$, we have

$$\begin{aligned}
\frac{a_i}{b} &= \left(\frac{\exp\{-\frac{1}{2}||x-y_i||_2^2\}}{\sum_{j\in[K]}\exp\{-\frac{1}{2}||x-y_j||_2^2\}} \right) \\
&\le \left(\frac{\exp\{-\frac{1}{2}(L+E_i)\}}{\exp\{-\frac{1}{2}L\} + \exp\{-\frac{1}{2}(L+E_i)\}} \right) \\
&= \frac{1}{\exp\{\frac{1}{2}E_i\}} \qquad (6)
\end{aligned}$$

and in the case of $i = z$, we bound $\frac{a_z}{b} \le 1$. So overall, we have $\frac{a_i}{b} \le \frac{1}{\exp\{\mathbf{1}_{i\neq z}\frac{1}{2}E_i\}}$ for all $i \in [K]$.

$$\begin{aligned} c_i &= ||y_* - y_i||_2^2 \\ &\leq 2||x - y_i||_2^2 + 2||x - y_*||_2^2 \\ &= 2(L + E_i + ||x - y_*||_2^2). \end{aligned} \tag{7}$$

$$\begin{aligned} \frac{\hat{a}_i}{a_i} - 1 &= \exp\left\{ \frac{1}{2}||x - y_i||_2^2 - \frac{1}{2}||x - \hat{y}_i||_2^2 \right\} - 1 \\ &\approx 1 + \frac{1}{2}||x - y_i||_2^2 - \frac{1}{2}||x - \hat{y}_i||_2^2 - 1 \\ &= \frac{1}{2}\langle \hat{y}_i - y_i, 2x - y_i - \hat{y}_i \rangle \end{aligned} \tag{8}$$

$$\begin{aligned} &\leq \frac{1}{2}||\hat{y}_i - y_i|| \cdot ||2x - y_i - \hat{y}_i|| \\ &\leq ||\hat{y}_i - y_i|| \left(||x - y_i|| + ||x - \hat{y}_i|| \right) \\ &= ||\hat{y}_i - y_i|| \left(||x - y_i|| + ||x - y_i + y_i - \hat{y}_i|| \right) \\ &\leq ||\hat{y}_i - y_i|| \left(2||x - y_i|| + ||y_i - \hat{y}_i|| \right) \\ &= ||\hat{y}_i - y_i|| \left(2\sqrt{L + E_i} + ||y_i - \hat{y}_i|| \right). \end{aligned} \tag{9}$$

Next, we must control $E_i - \hat{E}_i$ in order to obtain the bound for $\frac{\hat{a}_i}{\hat{b}}$.

$$\begin{aligned} E_i - \hat{E}_i &= (||x - y_i||^2 - L) - (||x - \hat{y}_i||^2 - \hat{L}) \\ &= ||x - y_i||^2 - ||x - \hat{y}_i||^2 + ||x - \hat{y}_z||^2 - ||x - y_z||^2 \\ &= \langle \hat{y}_i - y_i, 2x - y_i - \hat{y}_i \rangle + \langle y_z - \hat{y}_z, 2x - \hat{y}_z - y_z \rangle \\ &\leq ||\hat{y}_i - y_i|| \cdot ||2x - y_i - \hat{y}_i|| + ||y_z - \hat{y}_z|| \cdot ||2x - \hat{y}_z - y_z|| \\ &\leq ||y_i - \hat{y}_i|| \left(||x - \hat{y}_i|| + ||x - y_i|| \right) + ||\hat{y}_z - y_z|| \left(||x - y_z|| + ||x - \hat{y}_z|| \right) \\ &\leq c\sqrt{\frac{d \log K/\delta}{n}} \left(||x - \hat{y}_i|| + ||x - y_i|| + ||x - y_z|| + ||x - \hat{y}_z|| \right) \\ &= c\sqrt{\frac{d \log K/\delta}{n}} \left(\sqrt{\hat{L}} + \sqrt{\hat{L} + \hat{E}_i} + \sqrt{L} + \sqrt{L + E_i} \right) \\ \hat{E}_i &\geq \max\left\{ 0, E_i - cT\sqrt{\frac{d \log K/\delta}{n}} \right\}. \end{aligned}$$

Using (6), we obtain

$$\begin{aligned} \frac{\hat{a}_i}{\hat{b}} &\leq \frac{1}{\exp\{\mathbf{1}_{i \neq z} \frac{1}{2}\hat{E}_i\}} \\ &\leq \frac{1}{\exp\left\{ \mathbf{1}_{i \neq z} \max\left\{ 0, \frac{1}{2}E_i - \frac{cT}{2}\sqrt{\frac{d \log K/\delta}{n}} \right\} \right\}} \end{aligned} \tag{10}$$

Plugging (6), (7), (9), and (10) into (5), we obtain

$$\left| \sum_{i \in [K]} \frac{a_i}{b} \left(\frac{\hat{a}_i}{a_i} - 1 \right) c_i \right| + \left| \left(\sum_{i \in [K]} \frac{a_i}{b} \left(\frac{\hat{a}_i}{a_i} - 1 \right) \right) \left(\sum_{i \in [K]} \frac{\hat{a}_i}{\hat{b}} c_i \right) \right|$$

$$
\begin{aligned}
&\leq \sum_{i\in[K]} \frac{||\hat{y}_i - y_i|| \left(2\sqrt{L+E_i} + ||y_i - \hat{y}_i||\right)}{\exp\{\mathbf{1}_{i\neq z}\frac{1}{2}E_i\}} 2(L + E_i + ||x - y_*||_2^2) \\
&\quad + \left(\sum_{i\in[K]} \frac{||\hat{y}_i - y_i|| \left(2\sqrt{L+E_i} + ||y_i - \hat{y}_i||\right)}{\exp\{\mathbf{1}_{i\neq z}\frac{1}{2}E_i\}}\right) \\
&\quad \cdot \left(\sum_{i\in[K]} \frac{2(L + E_i + ||x - y_*||_2^2)}{\exp\left\{\mathbf{1}_{i\neq z}\max\left\{0, \frac{1}{2}E_i - \frac{cT}{2}\sqrt{\frac{d\log K/\delta}{n}}\right\}\right\}}\right) \\
&\leq O\left(\sqrt{\frac{d\log K/\delta}{n}} \sum_{i\in[K]} \frac{L + E_i + ||x - y_*||_2^2}{\exp\left\{\mathbf{1}_{i\neq z}\frac{1}{2}E_i\right\}}\right)
\end{aligned}
$$

Next, recalling the fact that our class means form an ε-net on the radius-R ball, we use Lemma 5.2 from [40] to bound L as $L \leq \frac{2R}{\sqrt{d}K-1}$.

$$
\leq O\left(\sqrt{\frac{d\log K/\delta}{n}} \sum_{i\in[K]} \frac{\frac{R}{\sqrt{d}K} + E_i + ||x - y_*||_2^2}{\exp\left\{\mathbf{1}_{i\neq z}\frac{1}{2}E_i\right\}}\right). \tag{11}
$$

Next, we will consider cases on E_i. We will consider the cases in which $E_i < \log R^2$ and $E_i \geq \log R^2$.

We will begin with the case in which $E_i < \log R^2$, then Equation 11 becomes:

$$
\leq O\left(\sqrt{\frac{d\log K/\delta}{n}} \left(\sum_{i\in[K]} \frac{R}{\sqrt{d}K} + ||x - y_*||_2^2\right)\right).
$$

Next, the case in which $E_i \geq \log R^2$, then Equation 11 becomes:

$$
\begin{aligned}
&\leq O\left(\sqrt{\frac{d\log K/\delta}{n}} \sum_{i\in[K]} \frac{\frac{R}{\sqrt{d}K} + 2\log R + ||x - y_*||_2^2}{R}\right) \\
&\leq O\left(\sqrt{\frac{d\log K/\delta}{n}} \left(\sum_{i\in[K]} \frac{1}{\sqrt{d}K} + ||x - y_*||_2^2\right)\right).
\end{aligned}
$$

Setting $M \leq K$ to be the number of terms for which $E_i < \log R^2$ holds and by combining the two bounds, we obtain the following:

$$
\leq O\left(\sqrt{\frac{d\log K/\delta}{n}} \left(\left(\frac{MR}{\sqrt{d}K} + \frac{K-M}{\sqrt{d}K}\right) + K||x - y_*||_2^2\right)\right).
$$

Ultimately, our bound is

$$
|\Psi_{\widehat{\mathbb{P}}}(y_*) - \Psi_{\mathbb{P}}(y_*)| \leq O\left(\sqrt{\frac{d\log K/\delta}{n}} \left(\left(\frac{MR}{\sqrt{d}K} + \frac{K-M}{\sqrt{d}K}\right) + K||x - y_*||_2^2\right)\right). \tag{12}
$$

Next, we will obtain a lower bound on $\Psi_{\widehat{\mathbb{P}}}(\hat{y}_*) - \Psi_{\mathbb{P}}(y_*)$.

$$
\begin{aligned}
|\Psi_{\widehat{\mathbb{P}}}(\hat{y}_*) - \Psi_{\mathbb{P}}(y_*)| &\geq |\Psi_{\mathbb{P}}(y_*) - \Psi_{\mathbb{P}}(\hat{y}_*)| - |\Psi_{\widehat{\mathbb{P}}}(\hat{y}_*) - \Psi_{\mathbb{P}}(\hat{y}_*)| && \text{triangle inequality} \\
&\geq \alpha d^2(y_*, \hat{y}_*) - |\Psi_{\widehat{\mathbb{P}}}(y_*) - \Psi_{\mathbb{P}}(y_*)| && \text{Assumption A.2.}
\end{aligned}
$$

Combining both of these bounds with Equation 2, we obtain

$$\begin{aligned}
\alpha d^2(y_*, \hat{y}_*) - |\Psi_{\widehat{\mathbb{P}}}(y_*) - \Psi_{\mathbb{P}}(y_*)| \leq \Psi_{\widehat{\mathbb{P}}}(\hat{y}_*) - \Psi_{\mathbb{P}}(y_*) &\leq \Psi_{\widehat{\mathbb{P}}}(y_*) - \Psi_{\mathbb{P}}(y_*) \\
&\leq |\Psi_{\widehat{\mathbb{P}}}(y_*) - \Psi_{\mathbb{P}}(y_*)|
\end{aligned}$$

$$\begin{aligned}
&\Rightarrow \alpha d^2(y_*, \hat{y}_*) \leq 2|\Psi_{\widehat{\mathbb{P}}}(y_*) - \Psi_{\mathbb{P}}(y_*)| \\
&\Rightarrow d^2(y_*, \hat{y}_*) \leq O\left(\frac{1}{\alpha}\sqrt{\frac{d\log K/\delta}{n}}\left(\frac{MR}{\sqrt{d}K} + \frac{K-M}{\sqrt{d}K} + K||x - y_*||_2^2\right)\right) \\
&\Rightarrow \mathbb{E}_{(x,y_*)\sim\mathcal{D}_{\text{test}}}[d^2(y_*, \hat{y}_*)] \leq O\left(\frac{1}{\alpha}\sqrt{\frac{d\log K/\delta}{n}}\left(\frac{MR}{\sqrt{d}K} + \frac{K-M}{\sqrt{d}K} + dK\right)\right),
\end{aligned}$$

Next, we obtain a bound on M. Since $M \leq K$ is the number of terms for which $E_i < \log R^2$, we have that $M = VK$, where $V \in [0, 1]$ is the ratio of the volumes of the ball for which $E_i < \log R^2$, and the ball containing the entire set of classes $\mathcal{Y}$. We have

$$\begin{aligned}
E_i = \|x - y_i\|_2^2 - \|x - y_z\|_2^2 &< \log R^2 \\
\|x - y_i\|_2^2 &< \log R^2 + L \\
&\leq \log R^2 + \frac{2R}{\sqrt{d}K - 1} \leq R^2 \\
\Rightarrow \sqrt{\log R^2 + \frac{2R}{\sqrt{d}K - 1}} &\leq R,
\end{aligned}$$

which is an upper bound of the radius of the ball for which $E_i < \log R^2$ is true. Now we construct V:

$$V = \left(\frac{\sqrt{\log R^2 + \frac{2R}{\sqrt{d}K-1}}}{R}\right)^d.$$

Combining this with our error bound, we have

$$\mathbb{E}_{(x,y_*)\sim\mathcal{D}_{\text{test}}}[d^2(y_*, \hat{y}_*)] \leq$$
$$O\left(\frac{1}{\alpha}\sqrt{\frac{d\log K/\delta}{n}}\left(\frac{\left(\frac{\sqrt{\log R^2 + \frac{2R}{\sqrt{d}K-1}}}{R}\right)^d R}{\sqrt{d}} + \frac{1 - \left(\frac{\sqrt{\log R^2 + \frac{2R}{\sqrt{d}K-1}}}{R}\right)^d}{\sqrt{d}} + dK\right)\right).$$

However, we can choose K to weaken our dependence on R.

$$\begin{aligned}
V = \left(\frac{\sqrt{\log R^2 + \frac{2R}{\sqrt{d}K-1}}}{R}\right)^d &\leq \frac{1}{R} \\
\Rightarrow K &\geq \frac{2}{\sqrt{d}\left(R^{1-\frac{2}{d}} - \frac{\log R}{R}\right)} + 1.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\mathbb{E}_{(x,y_*)\sim\mathcal{D}_{\text{test}}}[d^2(y_*, \hat{y}_*)] &\leq O\left(\frac{1}{\alpha}\sqrt{\frac{d\log K/\delta}{n}}\left(\frac{1}{\sqrt{d}} + \frac{1 - \frac{1}{R}}{\sqrt{d}} + \frac{\sqrt{d}}{\left(R^{1-\frac{2}{d}} - \frac{\log R}{R}\right)} + d\right)\right) \\
&\leq O\left(\frac{1}{\alpha}\sqrt{\frac{d\log K/\delta}{n}}\left(\frac{1}{\sqrt{d}} + \frac{\sqrt{d}}{\left(R^{1-\frac{2}{d}} - \frac{\log R}{R}\right)} + d\right)\right) \\
&\leq O\left(\frac{d}{\alpha}\sqrt{\frac{\log K/\delta}{n}}\left(\frac{1}{\left(R^{1-\frac{2}{d}} - \frac{\log R}{R}\right)} + \sqrt{d}\right)\right)
\end{aligned}$$

which completes the proof. □

Discussion of Theorem A.3. The above bound contains several standard quantities, including the dimensionality of the inputs d, the radius R of the ball from which labels are drawn, the number of classes and samples used to estimate the model K and n respectively, and δ, which is used to control the probability that the bound holds. Naturally, the bound scales up with higher d and K, and improves with an increased number of samples n. The dependence on R is also related to its dependence on d — so long as we have that $K \geq 2(\sqrt{d}(R^{1-\frac{2}{d}} - \log R/R))^{-1} + 1$ and for $d \geq 2$, R increases, which improves the bound. The bound includes a metric space dependent quantity α, which is related to the amount that the Fréchet variance can change subject to changes in its argument, i.e., which node is considered.

A.2 Proof of Theorem A.4 (minimum locus cover for trees)

Theorem A.4 (Minimum locus cover for trees)**.** *The minimum locus cover for a tree graph T is $\Lambda =$ Leaves(T).*

Proof. We would like to show to show two things:

1. any node can be resolved by an appropriate construction of $\boldsymbol{w}$ using only the leaves and
2. removing any leaf makes that leaf node unreachable, no matter what the other nodes are in $\boldsymbol{y}$–i.e., we must use at least all of the leaf nodes.

If both of these conditions hold (i.e., that the leaves form a locus cover and that they are the smallest such set), then the the leaves of any tree form a minimum locus cover. To set up the problem, we begin with some undirected tree, T, whose nodes are our set of labels: $T = (\mathcal{Y}, E)$ where $\Lambda = \text{Leaves}(T) \subseteq \mathcal{Y}$ is the set of leaf nodes. Moreover, let $\boldsymbol{\Lambda}$ be a tuple of the leaf nodes. We will start by proving that Λ is a locus cover. We will show this by cases on $v \in \mathcal{Y}$, the node that we would like to resolve:

1. If $v \in \Lambda$, i.e. v is a leaf node, then setting $\boldsymbol{w}_i = \mathbf{1}\{\boldsymbol{\Lambda}_i = v\}$ yields

$$\begin{aligned} m_{\boldsymbol{\Lambda}}(\boldsymbol{w}) &= \operatorname*{arg\,min}_{y \in \mathcal{Y}} \sum_{i=1}^{K} \mathbf{1}\{\boldsymbol{\Lambda}_i = v\} d^2(y, \boldsymbol{\Lambda}_i) \\ &= \operatorname*{arg\,min}_{y \in \mathcal{Y}} d^2(y, v) = \{v\}. \end{aligned}$$

2. If $v \notin \Lambda$, we have a bit more work to do. Since v is an internal node, consider any pair of leaves, $l_1 \neq l_2$ such that v is along the (unique) path between l_1 and l_2: $v \in \Gamma(l_1, l_2)$. Then set $\boldsymbol{w}_i = \begin{cases} \frac{d(v,l_2)}{d(l_1,l_2)} \text{ if } \boldsymbol{\Lambda_i} = l_1, \\ \frac{d(v,l_1)}{d(l_1,l_2)} \text{ if } \boldsymbol{\Lambda_i} = l_2, \\ 0 \text{ otherwise} \end{cases}$ yields

$$\begin{aligned} m_{\boldsymbol{\Lambda}}(\boldsymbol{w}) &= \operatorname*{arg\,min}_{y \in \mathcal{Y}} \frac{d(v, l_2) d^2(y, l_1) + d(v, l_1) d^2(y, l_2)}{d(l_1, l_2)} \\ &= \operatorname*{arg\,min}_{y \in \mathcal{Y}} d(v, l_2) d^2(y, l_1) + d(v, l_1) d^2(y, l_2) \\ &= \{v\}. \end{aligned}$$

Thus Λ is a locus cover. Next, we will show that it is the smallest such set. Let $l \in \Lambda$ be any leaf node, and define $\Lambda' = \Lambda \setminus \{l\}$. We must show that Λ' is not a locus cover. Assume for contradiction that Λ' is a locus cover. This means that given some tuple $\boldsymbol{y}'$ whose entries are the elements of Λ', we have $\Pi(\boldsymbol{\Lambda}') = \mathcal{Y}$. This implies that the path between any two nodes in Λ is also contained in $\Pi(\boldsymbol{\Lambda}')$. Since the leaves form a locus cover and any $y \in \mathcal{Y}$ can be constructed by the appropriate choice of $\boldsymbol{w}$ along with a path between two leaf nodes, l must either be one of the entries of $\boldsymbol{\Lambda}'$ (one of the endpoints of the path) or it must be an internal node. Both cases are contradictions–$\boldsymbol{\Lambda}'$ cannot include l by assumption, and l is assumed to be a leaf node. It follows that Λ is a minimum locus cover. □

A.3 Proof of Lemma A.5 (tree pairwise decomposability)

Lemma A.5 (Tree pairwise decomposability)**.** *Let $T = (\mathcal{Y}, \mathcal{E})$ be a tree and $\Lambda \subseteq \mathcal{Y}$. Then $\Pi(\Lambda)$ is pairwise decomposable.*

Proof. Assume for contradiction that $\exists y^* \in \mathcal{Y}$ where $y^* \notin \Pi(\{\lambda_i, \lambda_j\}), \forall \lambda_i, \lambda_j \in \Lambda$, but $y^* \in \Pi(\Lambda)$. Then, $y^* \in m_\Lambda(\mathbf{w})$ for some $\mathbf{w}$. Note that $y^* \notin \Pi(\{\lambda_i, \lambda_j\})$ implies $y^* \notin \Gamma(\lambda_i, \lambda_j), \forall \lambda_i, \lambda_j \in \Lambda$, because $\Pi(\{\lambda_i, \lambda_j\}) = \Gamma(\lambda_i, \lambda_j)$ due to the uniqueness of paths in trees. As such, $\cap_{\lambda \in \Lambda} \Gamma(y^*, \lambda)$ must contain an immediate relative y' of y^* where

$$\begin{aligned}\sum_{i=1}^{|\Lambda|} \mathbf{w}_i d^2(y^*, \lambda_i) &= \sum_{i=1}^{|\Lambda|} \mathbf{w}_i (d(y', \lambda_i) + 1)^2 \\ &> \sum_{i=1}^{|\Lambda|} \mathbf{w}_i d^2(y', \lambda_i).\end{aligned}$$

Therefore, $y^* \notin \Pi(\Lambda)$ because y^* is not a minimizer of $\sum_{i=1}^{|\Lambda|} \mathbf{w}_i d^2(y, \lambda_i)$. So, it must be that $y^* \in \Pi(\{\lambda_i, \lambda_j\})$ for some $\lambda_i, \lambda_j \in \Lambda$ when $y \in \Pi(\Lambda)$. □

A.4 Proof of Theorem A.6 (Algorithm B.1 correctness)

Theorem A.6 (Algorithm B.1 correctness)**.** *Algorithm B.1 returns a locus cover for phylogenetic trees.*

Proof. We first prove that Algorithm B.1 will halt, then according to the stopping criterion, it is guaranteed that Algorithm B.1 returns a locus cover. Suppose the algorithm keeps running until it reaches the case in which all of the leaf nodes are included in Λ (i.e. $\Lambda = \mathcal{Y}$), this results in the trivial locus cover of the phylogenetic tree and the algorithm halts. □

A.5 Proof of Lemma A.7 (phylogenetic tree pairwise decomposability)

Lemma A.7 (Phylogenetic tree pairwise decomposability)**.** *Let $T = (\mathcal{V}, \mathcal{E})$ be a tree with $\mathcal{Y} =$ Leaves(T) and $\Lambda \subseteq \mathcal{Y}$. Then $\Pi(\Lambda)$ is pairwise decomposable.*

Proof. Suppose to reach a contradiction that $\exists y^* \in \mathcal{Y}$ where $y^* \in \Pi(\Lambda)$, but $y^* \notin \Pi(\{\lambda_i, \lambda_j\})$ $\forall \lambda_i, \lambda_j \in \Lambda$. In other words, y^* is a Fréchet mean for some weighting on Λ, but y^* is not a Fréchet mean for any weighting for any two vertices within Λ.

For arbitrary $\lambda_i, \lambda_j \in \Lambda$, define the closest vertex to y^* on the path between λ_i and λ_j as

$$v'(\lambda_i, \lambda_j) = \operatorname*{arg\,min}_{v \in \Gamma(\lambda_i, \lambda_j)} d(y^*, v).$$

Also, let

$$\{\lambda_1, \lambda_2\} = \operatorname*{arg\,min}_{\lambda_i, \lambda_j \in \Lambda} d(y^*, v'(\lambda_i, \lambda_j)),$$

the pair of vertices in Λ whose path $\Gamma(\lambda_1, \lambda_2)$ passes *closest* to y^*. To avoid notational clutter, define $v' = v'(\lambda_1, \lambda_2)$. Thus, v' represents the closest vertex to y^* lying on the path between any two $\lambda_i, \lambda_j \in \Lambda$.

Because v' lies on the path $\Gamma(\lambda_1, \lambda_2)$, there exists some weighting $\mathbf{w}$ for which

$$v' = \operatorname*{arg\,min}_{v \in \Gamma(\lambda_1, \lambda_2)} (\mathbf{w}_1 d^2(v, \lambda_1) + \mathbf{w}_2 d^2(v, \lambda_2)).$$

For this $\mathbf{w}$, we have that $m_{\{\lambda_1, \lambda_2\}}(\mathbf{w}) = \arg\min_{y \in \mathcal{Y}} d(y, v')$, the set of leaf nodes of smallest distance from v'. By the initial assumption that $y^* \notin \Pi(\{\lambda_1, \lambda_2\})$, we have that $y^* \notin m_{\{\lambda_1, \lambda_2\}}(\mathbf{w})$. Let y' be a vertex of $m_{\{\lambda_i, \lambda_j\}}(\mathbf{w})$ that is closest to y^*. Clearly, $d(y', v')^2 < d(y^*, v')^2$.

For any $v \in \mathcal{V}$, respectively define the sets of vertices (1) shared in common on the two paths $\Gamma(v, y^*)$ and $\Gamma(v, y')$, (2) on the path $\Gamma(v, y^*)$ that are not on $\Gamma(v, y')$, and (3) on the path $\Gamma(v, y')$ that are not on $\Gamma(v, y^*)$ as

$$\begin{aligned} d_C(v) &= \Gamma(v, y^*) \cap \Gamma(v, y'), \\ d_{y^*}(v) &= \Gamma(v, y^*) \setminus \Gamma(v, y'), \text{ and} \\ d_{y'}(v) &= \Gamma(v, y') \setminus \Gamma(v, y^*). \end{aligned}$$

We have that $|d_{y^*}(v)| + |d_{y^*}(v)| + 1 = |\Gamma(y^*, y')|$ (the difference by 1 represents the final vertex shared in common by both paths) and also that $|d_{y'}(v')| < |d_{y^*}(v')|$ since y' is closer to v' than y^* is.

Suppose $\exists \lambda_3, \lambda_4 \in \Lambda$ yielding

$$u' = v'(\lambda_3, \lambda_4) \in \Gamma(y^*, y') \text{ such that } d^2(y^*, u') \leq d^2(y', u').$$

This is to say that u' is defined analogously to v', with the additional restrictions that u' lie on the path $\Gamma(y^*, y')$ and u' be at least as close to y^* as to y'. Then, $|d_{y^*}(u')| < |d_{y^*}(v')|$ implying $d^2(y^*, u') < d^2(y^*, v')$, which contradicts the definition of v'.

Suppose now that $\exists \lambda_3, \lambda_4 \in \Lambda$ yielding

$$u' = v'(\lambda_3, \lambda_4) \notin \Gamma(y^*, y') \text{ such that } d^2(y^*, u') \leq d^2(y', u').$$

This is to say that u' is again defined analogously to v', but u' is not on $\Gamma(y^*, y')$ and is at least as close to y^* as to y'. Then, either $\cap_{\lambda \in \Lambda} d_C(\lambda) \neq \emptyset$ (i.e. all lambdas lie in the same branch off of $\Gamma(y', y^*)$), in which case the path between any two lambdas will have the same closest node v' to y^*, or it is possible to choose a $\lambda_5 \in \Lambda$ such that $d_C(\lambda_3) \cap d_C(\lambda_5) = \emptyset$. In this situation, it must be that $\Gamma(\lambda_3, \lambda_5) \cap \Gamma(y', y^*) \neq \emptyset$, which implies there exists a vertex $t \in \Gamma(\lambda_3, \lambda_5)$ such that $d^2(y^*, t) < d^2(y^*, u')$ which violates the definition of u'.

Altogether, for all $v \in \Gamma(\lambda_i, \lambda_j)$ with arbitrary $\lambda_i, \lambda_j \in \Lambda$, we have that $d^2(y', v) < d^2(y^*, v)$. This implies that for all $\lambda \in \Lambda$, there exists a $y' \in \mathcal{Y}$ such that $d^2(y', \lambda) < d^2(y^*, \lambda)$ and therefore that $y^* \notin \Pi(\Lambda)$ which contradicts that $y^* \in \Pi(\Lambda)$. It must be that $\forall y^* \in \mathcal{Y}$ where $y^* \in \Pi(\Lambda)$, we have $y^* \in \Pi(\{\lambda_i, \lambda_j\}) \; \forall \lambda_i, \lambda_j \in \Lambda$. □

A.6 Proof of Theorem A.8 (minimum locus cover for grid graphs)

Theorem A.8 (Minimum locus cover for the grid graphs)**.** *The minimum locus cover for a grid graph is the pair of vertices in the furthest opposite corners.*

Proof. We would like to show two things:

1. any vertex can be resolved by construction of $\boldsymbol{w}$ using the furthest pair of two vertices (i.e. opposite corners) and
2. there does not exist a locus cover of size one for grid graphs with more than one vertex.

Let $G = (\mathcal{V}, \mathcal{E})$ be a grid graph and let $\Lambda = \{\lambda_1, \lambda_2\}$ be a set of two vertices who are on opposite corners of G (i.e., perhipheral vertices achieving the diameter of G). For notational convenience, set $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2)$ We can reach any interior vertex by an appropriate setting the weight vector $\boldsymbol{w} = [\boldsymbol{w}_1, \boldsymbol{w}_2]$. Then set $\boldsymbol{w} = \left(\frac{d(v,\lambda_2)}{d(\lambda_1,\lambda_2)}, \frac{d(v,\lambda_1)}{d(\lambda_1,\lambda_2)} \right)$. Finally, the Fréchet mean is given by

$$\begin{aligned} m_{\boldsymbol{\Lambda}}(\boldsymbol{w}) &= \arg\min_{y \in \mathcal{Y}} \frac{d(v, \lambda_2) d^2(y, \lambda_1) + d(v, \lambda_1) d^2(y, \lambda_2)}{d(\lambda_1, \lambda_2)} \\ &= \arg\min_{y \in \mathcal{Y}} d(v, \lambda_2) d^2(y, \lambda_1) + d(v, \lambda_1) d^2(y, \lambda_2) \\ &\ni v. \end{aligned}$$

Hence Λ is a locus cover. We can clearly see that Λ is a minimum locus cover—the only way to obtain a smaller set Λ' is for it to include only a single vertex. However, if $\Lambda' = \{v'\}$ contains only a single vertex, it cannot be a locus cover so long as G contains more than one vertex—v' is always guaranteed to be a unique minimizer of the Fréchet mean under Λ' which misses all other vertices in G. Thus Λ is a minimum locus cover. □

A.7 Proof of Lemma A.9 (loci of grid subspaces)

Lemma A.9 (Locus of grid subspaces)**.** *Given any pair of vertices in Λ, we can find a subset G' of the original grid graph $G = (\mathcal{Y}, \mathcal{E})$ which takes the given pair as two corner. $\Pi(\mathbf{\Lambda})$ equals to all the points inside G'.*

Proof. The result follows from application of Theorem A.8 to the choice of metric subspace. □

A.8 Proof of Lemma A.10 (grid pairwise decomposability)

Lemma A.10 (Grid pairwise decomposability)**.** *Let $G = (\mathcal{Y}, \mathcal{E})$ be a grid graph and $\Lambda \subseteq \mathcal{Y}$. Then $\Pi(\Lambda)$ is pairwise decomposable.*

Proof. Suppose we have a grid graph $G = (\mathcal{Y}, \mathcal{E})$ and a vertex $\hat{y} \in \Pi(\Lambda)$ with $\Lambda \subseteq \mathcal{Y}$. Due to the fact that $\hat{y} \in \Pi(\Lambda)$ and the fact that G is a grid graph, we have that $\hat{y} \in \Gamma(\lambda_\alpha, \lambda_\beta)$ for some $\lambda_\alpha, \lambda_\beta \in \Lambda$. Then by Lemma A.10,

$$\hat{y} \in \Pi(\{\lambda_\alpha, \lambda_\beta\}) \subseteq \cup_{\lambda_i, \lambda_j \in \Lambda} \Pi(\{\lambda_i, \lambda_j\}).$$

Therefore loci on grid graphs are pairwise decomposable. □

A.9 Proof of Theorem A.11 (no nontrivial locus covers for complete graphs)

Theorem A.11 (Trivial locus cover for the complete graph)**.** *There is no non-trivial locus cover for the complete graph.*

Proof. We show that there is no nontrivial locus cover for complete graphs by showing that removing any vertex from the trivial locus cover renders that vertex unreachable. We proceed by strong induction on the number of vertices, n.

Base case We first set $n = 3$. Let $\boldsymbol{K_3} = (\mathcal{V}, \mathcal{V}^2)$ be the complete graph with three vertices: $\mathcal{V} = \{v_1, v_2, v_3\}$, and without loss of generality, let $\Lambda = \{v_2, v_3\}$ be our set of observed classes. There are two cases on the weight vector $\boldsymbol{w} = [w_2, w_3]$:

Case 1: Suppose $\boldsymbol{w} \notin \text{int}\Delta^2$. This means that either $w_2 = 1$ or $w_3 = 1$—which leads to the Fréchet mean being either v_2 or v_3, respectively. Neither of these instances correspond to v_1 being a minimizer.

Case 2: Suppose $\boldsymbol{w} \in \text{int}\Delta^2$. Then the Fréchet mean is given by

$$m_{\boldsymbol{\lambda}}(\boldsymbol{w}) = \arg\min_{y \in \mathcal{Y}} w_2 d^2(y, v_2) + w_3 d^2(y, v_3)$$

Assume for contradiction that $v_1 \in \Pi(\mathbf{\Lambda})$:

$$\begin{aligned} w_2 d^2(v_1, v_2) + w_3 d^2(v_1, v_3) &= w_2 + w_3 \\ &> w_3 && \text{because } \boldsymbol{w} \in \text{int}\Delta^2 \\ &= w_2 d^2(v_2, v_2) + w_3 d^2(v_2, v_3). \end{aligned}$$

Therefore, $v_1 \notin \Pi(\mathbf{\Lambda})$. This is a contradiction. Thus there is no nontrivial locus cover for $\boldsymbol{K_3}$.

Inductive step: Let $\boldsymbol{K_{n-1}} = (\mathcal{V}, \mathcal{V}^2)$ be the complete graph with $n - 1$ vertices. Assume that there is no nontrivial locus cover for $\boldsymbol{K_{n-1}}$. We will show that there is no nontrivial locus cover for $\boldsymbol{K_n}$. Let the weight vector be $\boldsymbol{w} = [w_1, ...w_{n-1}]$ corresponding to vertices $v_1, ..., v_{n-1} \in \mathcal{V}$ with $\Lambda = \{v_1, ..., v_{n-1}\}$. We want to show that $v_n \notin \Pi(\mathbf{\Lambda})$. We proceed by cases on $\boldsymbol{w}$.

Case 1: Suppose $\boldsymbol{w} \notin \text{int}\Delta^{n-2}$ where m entries of $\boldsymbol{w}$ are zero, then by strong induction we know that $v_n \notin \Pi(\mathbf{\Lambda}) = \{v_i\}_{i=1}^{n-m}$, i.e., there is no nontrivial locus cover for $\boldsymbol{K_{n-m}}$.

Case 2: Suppose $\boldsymbol{w} \in \text{int}\Delta^{n-2}$.

Assume for contradiction that $v_n \in \Pi(\boldsymbol{\Lambda})$. Using this assumption, we obtain the following

$$\begin{aligned}
\sum_{i=1}^{n-1} w_i d^2(v_n, v_i) &= \sum_{i=1}^{n-1} w_i \\
&> \sum_{i=1}^{n-2} w_i && \text{because } \boldsymbol{w} \in \text{int}\Delta^{n-2} \\
&= \sum_{i=1}^{n-1} w_i d^2(v_{n-1}, v_i).
\end{aligned}$$

Therefore, v_n is not a minimizer, so $v_n \notin \Pi(\Lambda)$. This is a contradiction, hence $\boldsymbol{K_n}$ has no nontrivial locus cover. □

A.10 Proof of Theorem 4.8 (active next-class selection for trees)

Proof. We will prove the result by showing that the two optimization problems are equivalent. Let v be a solution to the following optimization problem:

$$\begin{array}{cl}
\underset{y \in \mathcal{Y} \setminus \Pi(\Lambda)}{\arg\max} & d(y, b) \\
\text{s.t.} & b \in \partial_{\text{in}} T', \\
 & \Gamma(b, y) \setminus \{b\} \subseteq \mathcal{Y} \setminus \Pi(\Lambda),
\end{array}$$

where $\partial_{\text{in}} T'$ is the inner boundary of T', the subgraph of T whose vertices are $\Pi(\Lambda)$. This optimization problem can be equivalently rewritten as

$$\begin{array}{cl}
\underset{y \in \mathcal{Y} \setminus \Pi(\Lambda)}{\arg\max} & |\Gamma(y, b)| \\
\text{s.t.} & b \in \partial_{\text{in}} T', \\
 & \Gamma(b, y) \setminus \{b\} \subseteq \mathcal{Y} \setminus \Pi(\Lambda),
\end{array}$$

and we can furthermore introduce additional terms that do not change the maximizer

$$\begin{array}{cl}
\underset{y \in \mathcal{Y} \setminus \Pi(\Lambda)}{\arg\max} & |\cup_{\lambda_i, \lambda_j \in \Lambda} \Gamma(\lambda_i, \lambda_j) \cup \Gamma(b, y)| \\
\text{s.t.} & b \in \partial_{\text{in}} T', \\
 & \Gamma(b, y) \setminus \{b\} \subseteq \mathcal{Y} \setminus \Pi(\Lambda).
\end{array}$$

Equivalently, we can also connect b to one of the elements of Λ

$$\begin{array}{cl}
\underset{y \in \mathcal{Y} \setminus \Pi(\Lambda)}{\arg\max} & |\cup_{\lambda_i, \lambda_j \in \Lambda} \Gamma(\lambda_i, \lambda_j) \cup \Gamma(\lambda_i, b) \cup \Gamma(b, y)| \\
\text{s.t.} & b \in \partial_{\text{in}} T', \\
 & \Gamma(b, y) \setminus \{b\} \subseteq \mathcal{Y} \setminus \Pi(\Lambda).
\end{array}$$

Due to the uniqueness of paths in trees, this optimization problem also has the following equivalent form without any dependence on b:

$$\begin{aligned}
v \in \underset{y \in \mathcal{Y} \setminus \Lambda}{\arg\max} |\cup_{\lambda_i, \lambda_j \in \Lambda} \Gamma(\lambda_i, \lambda_j) \cup \Gamma(\lambda_i, y)| &= \underset{y \in \mathcal{Y} \setminus \Lambda}{\arg\max} |\cup_{\lambda_i, \lambda_j \in \Lambda} \Pi(\{\lambda_i, \lambda_j\}) \cup \Pi(\{\lambda_i, y\})| \\
&= \underset{y \in \mathcal{Y} \setminus \Lambda}{\arg\max} |\cup_{\lambda_i, \lambda_j \in \Lambda \cup \{y\}} \Pi(\{\lambda_i, \lambda_j\})| \\
&= \underset{y \in \mathcal{Y} \setminus \Lambda}{\arg\max} |\Pi(\Lambda \cup \{y\})| \\
&\quad \text{using Lemma A.5.}
\end{aligned}$$

Therefore v is a maximizer of $|\Pi(\Lambda \cup \{v\})|$, as required. □

B Algorithms and Time Complexity Analyses

We provide time complexity analyses for Algorithms B.1, B.2, and B.3.

B.1 Analysis of Algorithm B.1 (locus cover for phylogenetic trees)

We provide Algorithm B.1 with comments corresponding to the time complexity of each step.

Algorithm 1 Locus cover for phylogenetic trees

```
Require: phylogenetic tree $T = (\mathcal{V}, \mathcal{E})$, $\mathcal{Y} = \text{Leaves}(T)$
  $N \leftarrow |\mathcal{Y}|$
  $P \leftarrow \text{sortbylength}([\Gamma(y_i, y_j)]_{i,j \in [N]})$        $\triangleright N|\mathcal{E}| + N^2 \log N$
  $P \leftarrow \text{reverse}(P)$        $\triangleright O(N^2)$
  $\Lambda \leftarrow \emptyset$
  for $\Gamma(y_i, y_j)$ in $P$ do        $\triangleright O(N^2)$
    if $\Pi(\Lambda) = \mathcal{Y}$ then        $\triangleright O(K^2 D \max\{N|\mathcal{E}|, N^2 \log N\})$
      return $\Lambda$
    else
      $\Lambda \leftarrow \Lambda \cup \{y_i, y_j\}$
    end if
  end for
```

Combining these, we obtain the following time complexity:

$$O(N|\mathcal{E}| + N^2 \log N + N^2 + N^2 K^2 D \max\{N|\mathcal{E}|, N^2 \log N\}) = O(N^2 K^2 D \max\{N|\mathcal{E}|, N^2 \log N\}).$$

B.2 Analysis of Algorithm B.2 (computing a pairwise decomposable locus)

We first provide Algorithm B.2 here, with comments corresponding to the time complexity of each step.

Algorithm 2 Computing a pairwise decomposable locus

```
Require: $\Lambda, \mathcal{Y}, G = (\mathcal{V}, \mathcal{E})$
  $\Pi \leftarrow \emptyset$
  $D \leftarrow \text{diam}(G)$        $\triangleright O(N|\mathcal{E}| + N^2 \log N)$
  for $\lambda_i, \lambda_j \in \Lambda$ do        $\triangleright O(K^2)$
    for $w_1$ in $\{\frac{0}{D}, \frac{1}{D}, ..., \frac{D}{D}\}$ do        $\triangleright O(D)$
      $\mathbf{w} \leftarrow [w_1, 1 - w_1]$
      $\Pi \leftarrow \Pi \cup m_{\{\lambda_i, \lambda_j\}}(w)$        $\triangleright O(N|\mathcal{E}| + N^2 \log N)$
    end for
  end for
  return $\Pi$
```

We first compute the diameter of the graph $G = (\mathcal{Y}, \mathcal{E})$ with $N = |\mathcal{Y}|$, which is done in $O(N|\mathcal{E}| + N^2 \log N)$ time using Dijkstra's algorithm to compute the shortest paths between all pairs of vertices. We then iterate over all pairs of elements in Λ with $K = |\Lambda|$, which amounts to $O(K^2)$ iterations. Within this, we perform $O(D)$ computations of the Fréchet mean, for which each iteration requires $O(N|\mathcal{E}| + N^2 \log N)$ arithmetic operations or comparisons. Combining these, the final time complexity is

$$O(N|\mathcal{E}| + N^2 \log N + K^2 D(N|\mathcal{E}| + N^2 \log N)) = O(K^2 D \max\{N|\mathcal{E}|, N^2 \log N\}).$$

B.3 Analysis of Algorithm B.3 (computing a generic locus)

We provide Algorithm B.3 with comments corresponding to the time complexity of each step.

Following a similar argument from the analysis of Algorithm B.2, the time complexity is

$$O(N|\mathcal{E}| + N^2 \log N + D^K(N|\mathcal{E}| + N^2 \log N)) = O(D^K),$$

Algorithm 3 Computing a generic locus

```
Require: $\Lambda, \mathcal{Y}, G = (\mathcal{V}, \mathcal{E})$
  $\Pi \leftarrow \emptyset$
  $D \leftarrow \text{diam}(G)$                                            $\triangleright O(N|\mathcal{E}| + N^2 \log N)$
  for w in $\left\{\frac{0}{D}, \frac{1}{D}, ..., \frac{D}{D}\right\}^{|\Lambda|}$ do                     $\triangleright O(D^K)$
      $\Pi \leftarrow \Pi \cup m_{\mathbf{\Lambda}}(w)$                                  $\triangleright O(N|\mathcal{E}| + N^2 \log N)$
  end for
  return $\Pi$
```

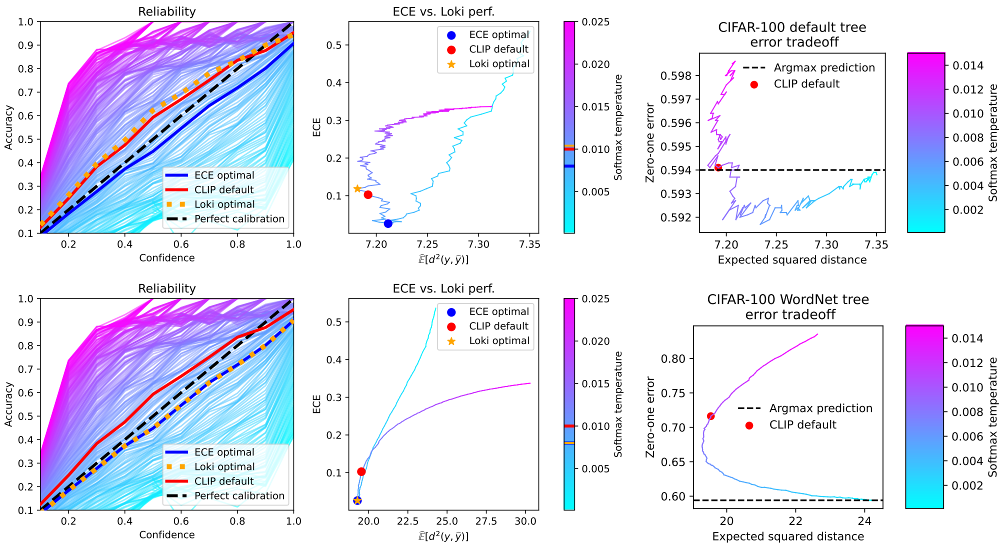

Figure 4: **CLIP on CIFAR-100 with the WordNet hierarchy.** (Left) Reliability diagrams across a range of Softmax temperatures, highlighting the CLIP default temperature, the optimal temperature for LOKI, and the minimizer of the Expected Calibration Error (ECE). All three are well-calibrated. (Center) Tradeoff between optimizing for ECE and the expected squared distance. As with the reliability diagrams, the CLIP default temperature, the LOKI-optimal temperature, and the ECE-optimal temperature are similar. (Right) Tradeoff between optimizing for zero-one error and the expected squared distance. Depending on the metric space, temperature scaling can improve *accuracy*.

where K is the number of observed classes.

C Additional Experiments

C.1 Additional calibration experiments

In the main text, we evaluated the effect of calibrating the Softmax outputs on the performance of LOKI using a metric space based on the internal features of the CLIP model. Here, we perform the same analysis using two external metric spaces.

Setup We perform our calibration analysis using the default and WordNet tree.

Results Our results are shown in Figure 4. Like our experiment using an internally-derived metric, we find that the optimal Softmax temperature is close to the CLIP default and the optimal temperature for calibration. For the default tree, we again found that temperature scaling can be used to improve accuracy using LOKI.

Setup We calibrate via Softmax temperature scaling [13] using CLIP on CIFAR-100. We do not use an external metric space, and instead use Euclidean distance applied to the CLIP text encoder.

Results The reliability diagram in Figure 3 shows that the optimal Softmax temperature for LOKI is both close to the default temperature used by CLIP and to the optimally-calibrated temperature. In Figure 3 (right), we find that appropriate tuning of the temperature parameter *can lead to improved accuracy with CLIP*, even when no external metric space is available.

C.2 Ablation of LOKI formulation

LOKI is based on the Fréchet mean, which is defined as $\arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{K} \mathbf{P}_{\lambda_i|x} d^2(y, \lambda_i)$. However, this is not the only approach that can be considered. For example, the Fréchet *median*, often used in robust statistics, is defined as $\arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{K} \mathbf{P}_{\lambda_i|x} d(y, \lambda_i)$, without squaring the distances. More generally, we define $\arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{K} \mathbf{P}_{\lambda_i|x} d^{\beta}(y, \lambda_i)$ and evaluate different choices of β.

Setup We conduct this experiment on ImageNet using SimCLR as our pretrained classifier with $K = 250$, 500, and 750 randomly selected classes.

Results From this analysis, we conclude that using the Fréchet mean is the optimal formulation for Loki, as it achieved the lowest mean squared distance for all settings of K.

Table 4: Expected squared distances of SimCLR+Loki alternatives on ImageNet. We ablate over the choice of distance exponent in LOKI (where $\beta = 2$, corresponding to the Fréchet mean), including the Fréchet median ($\beta = 1$). That is, we tune $\beta : \hat{y} \in \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{K} \mathbf{P}_{\lambda_i|x} d^{\beta}(y, \lambda_i)$ and find that the optimal setting is $\beta = 2$, corresponding to LOKI.

$\beta =$	$K = 250$	$K = 500$	$K = 750$
0.5	61.28	46.57	36.31
1	52.98	41.06	33.14
2 (Loki)	**46.78**	**37.76**	**29.88**
4	47.99	39.58	34.65
8	65.29	58.42	54.90

C.3 Comparison across metric spaces

Expected squared distances cannot be directly compared across metric spaces, as they may be on different scales. Our solution is to use a form of normalization: we divide the expected squared distance by the square of the graph diameter. This brings all of the values to the 0-1 range, and since $\mathbb{E}[d^2(y, \hat{y})]/\text{diam}(G)^2$ indeed also generalizes the 0-1 error, this enables comparison between 0-1 errors and those from different metric spaces. We provide these results in Table 1, again for our CLIP experiments on CIFAR-100. This evaluation metric enables us to determine which metric spaces have geometry best 'fit' to our pretrained models.

Setup Using $\mathbb{E}[d^2(y, \hat{y})]/\text{diam}(G)^2$, we re-evaluate our CLIP results on CIFAR-100 shown in Table 1 for the ResNet-50 architecture.

Results For CIFAR-100, we observe that the WordNet metric space resulted in the lowest error and therefore has the best geometry.

Table 5: Comparison across metric spaces for CLIP on CIFAR-100 by normalizing by the squared diameter of the metric space: $\mathbb{E}[d^2(y, \hat{y})]/\text{diam}(G)^2$.

Metric space G	$\text{diam}(G)^2$	$\mathbb{E}[d^2(y, \hat{y})]/\text{diam}(G)^2$
Complete graph	1	0.5941 (0-1 error)
Default tree	16	0.4493
WordNet tree	169	**0.1157**
CLIP features	0.9726	0.2686

D Experimental Details

In this section, we provide additional details about our experimental setup.[3]

D.1 CLIP experiments

All experiments are carried out using CLIP frozen weights. There are no training and hyperparameters involved in experiments involving CLIP, except for the Softmax temperature in the calibration analysis. We evaluted using the default CIFAR-100 test set. The label prompt we use is "a photo of a [class_name]."

D.2 ImageNet experiments

To construct our datasets, we randomly sample 50 images for each class from ImageNet as our training dataset then use the validation dataset in ImageNet to evaluate LOKI's performance. We extract the image embeddings by using SimCLRv1[4] and train a baseline one-vs-rest classifier. We use WordNet phylogenetic tree as the metric space. The structure of phylogenetic tree can be found at here[5]. We test different numbers of observed classes, K, from 1000 classes. Observed classes are sampled in two ways, uniformly and Gibbs distribution with the concentration parameter 0.5.

D.3 LSHTC experiments

We generate a summary graph of the LSHTC class graph (resulting in supernodes representing many classes) by iteratively:

1. randomly selecting a node or supernode
2. merging its neighbors into the node to create a supernode

until the graph contains at most 10,000 supernodes. In the LSHTC dataset, each datapoint is assigned to multiple classes. We push each class to its supernode, then apply majority vote to determine the final class. We test different numbers of observed classes, K, from 10,000 classes. We collect datapoints which are in the observed classes. Then split half of dataset as training dataset and make the rest as testing dataset, including those datapoints which are not in the observed classes. We train a baseline classifier using a 5-NN model and compare its performance with LOKI.

E Broader Impacts and Limitations

As a simple adaptor of existing classifiers, it is possible for our method to inherit biases and failure modes that might be present in existing pretrained models. On the other hand, if the possibility of harmful mis-classifications are known a priori, information to mitigate these harms can be baked into the metric space used by LOKI. Finally, while we have found that LOKI often works well in practice, it is possible for the per-class probabilities that are output by a pretrained model to be sufficiently mis-specified relative to the metric over the label space, or for the metric itself to be mis-specified. Nonetheless, we have found that off-the-shelf or self-derived metric spaces to work well in practice.

F Random Locus Visualizations

In Figures 1, 5, 6, 7, 8, and 9, we provide visualizations of classification regions of the probability simplex when using LOKI with only three observed classes out of 100 total classes, and different types of random metric spaces. The first example in Figure 1 shows the prediction regions on the probability simplex when using standard $\arg\max$ prediction—the three regions correspond to predicting one of the three classes (0, 39, and 99) and no regions corresponding to any of the other classes $\{1, ..., 38, 40, ..., 98\}$. We compute these regions using a version of Algorithm B.3, and while

[3] Code implementing all of our experiments is available here: `https://github.com/SprocketLab/loki`.

[4] Embeddings are extracted from the checkpoints stored at: `https://github.com/tonylins/simclr-converter`

[5] `https://github.com/cvjena/semantic-embeddings/tree/master/ILSVRC`

it does have exponential time complexity, the exponent is only three in this case since we consider only three observed classes.

On the other hand, the other three examples in Figure 1 and Figures 5, 6, 7, 8, and 9 all show the prediction regions on the probability simplex when using LOKI. Figure 5 shows this for random tree graphs. Here, the prediction regions are striped or contain a single triangle-shaped region in the center—these correspond, respectively, to intermediate classes along branches of the tree leading up from the observed class and the prediction region formed by the common-most parent node. Figure 6 shows similar regions, although these are more complex and are thus more difficult to interpret, as phylogenetic trees are metric subspaces equipped with the induced metric from trees. Furthermore, in order to generate phylogenetic trees with 100 leaves, we needed to create much larger trees than the ones used for Figure 5, which led to narrower prediction regions due to the higher graph diameter. Finally, Figures 7, 8, and 9 each show the prediction regions using LOKI when the metric space is a random graph produced using Watts-Strogatz, Erdős-Rényi, and Barabási–Albert, models respectively. These prediction regions are more complex and represent complex relationships between the classes.

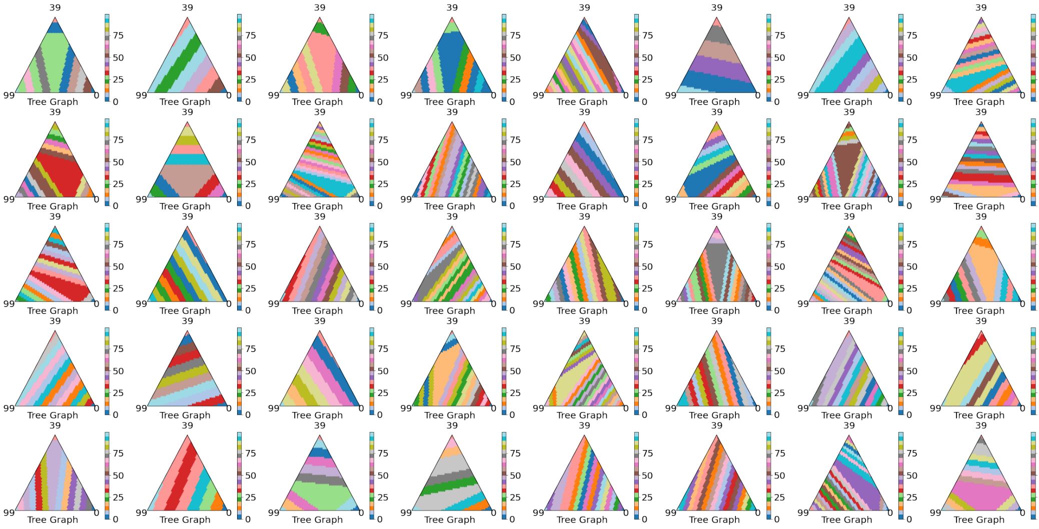

Figure 5: Classification regions in the probability simplex of 3-class classifiers faced with a 100-class problem where the classes are related by random trees graphs.

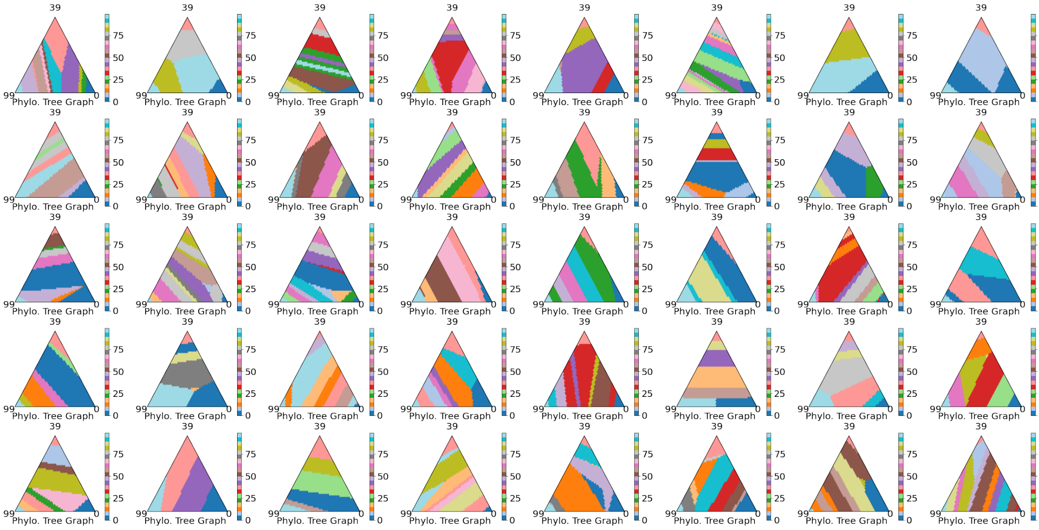

Figure 6: Classification regions in the probability simplex of 3-class classifiers faced with a 100-class problem where the classes are related by random phylogenetic trees graphs.

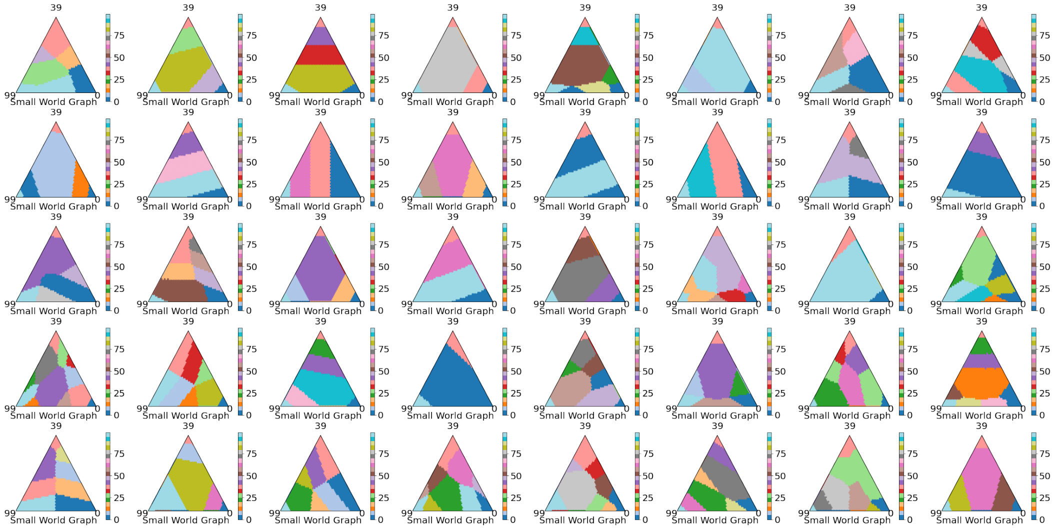

Figure 7: Classification regions in the probability simplex of 3-class classifiers faced with a 100-class problem where the classes are related by random small-world graphs generated by a Watts–Strogatz model.

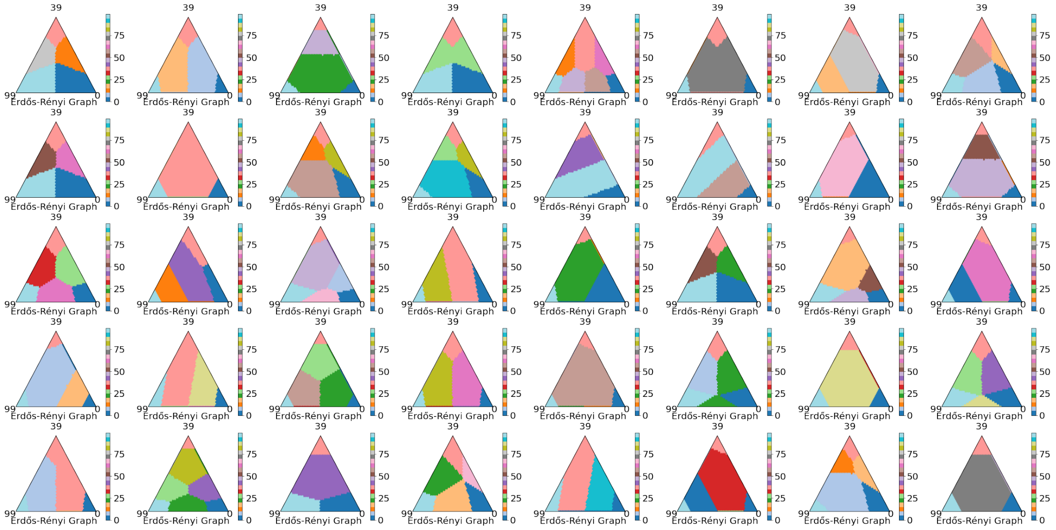

Figure 8: Classification regions in the probability simplex of 3-class classifiers faced with a 100-class problem where the classes are related by random Erdős-Rényi graphs.