Theorem 10. Suppose that we have a finite set of hypotheses $\mathcal{H}$ (i.e., $|\mathcal{H}| < \infty$) and $\hat{h}_n = \arg\min_{h \in \mathcal{H}} R_n(h)$. Also, assume that the data is separable (i.e., the perfect classifier $h^*$ with no error exists). For any $\epsilon, \delta \in (0, 1)$, we have $\Pr\left(R(\hat{h}_n) > \epsilon\right) \leq \delta$ whenever $n \geq \epsilon^{-1} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$. In other words, for any $\epsilon, \delta \in (0, 1)$, with probability $1 - \delta$, we have $R(\hat{h}_n) \leq \frac{\log(\frac{2|\mathcal{H}|}{\delta})}{n}$.

$$(x_i, y_i) \in \mathcal{X} \times \{0, 1\}$$

Assume $h^* = \arg\min_h R(h)$ and $R(h^*) = 0$.

$$R(h) = \mathbb{E}_{(x,y) \in \mathcal{D}} \left[ I\{h(x) \neq y\}\right]$$

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} I\{h(x_i) \neq y_i\}$$

Active Learner for streaming/sampling-oracle setting.

**Algorithm 1 CAL**

1. Initialize: $Z_0 = \emptyset$, $V_0 = \mathcal{H}$
2. for $t = 1, 2, \ldots, n$ do
3. Nature reveals unlabeled data point $x_t$
4. if $x_t \in DIS(V_{t-1})$ then
5. Query $y_t$, and set $Z_t = Z_{t-1} \cup (x_t, y_t)$
6. else
7. $Z_t = Z_{t-1}$
8. end if
9. $V_t = \{h \in \mathcal{H} : h(x_i) = y_i \ \forall (x_i, y_i) \in Z_t\}$
10. end for
11. return any $h \in V_n$

**Definition 6.** For some hypothesis class $\mathcal{H}$ and subset $V \subset \mathcal{H}$ where for each $h \in \mathcal{H}, h : \mathcal{X} \to \{0, 1\}$, the region of disagreement is defined as

$$DIS(V) = \{x \in \mathcal{X} : \exists h, h' \in V \text{ s.t. } h(x) \neq h'(x)\}$$

which is the set of unlabeled examples $x$ for which there are hypotheses in $V$ that disagree on how to label $x$. 
Algorithm 1 CAL

1: Initialize: $Z_0 = \emptyset$, $V_0 = \mathcal{H}$
2: for $t = 1, 2, \ldots, n$ do
3: Nature reveals unlabeled data point $x_t$
4: if $x_t \in DIS(V_{t-1})$ then
5: Query $y_t$, and set $Z_t = Z_{t-1} \cup \{x_t, y_t\}$
6: else
7: $Z_t = Z_{t-1}$ // $y_t = h^*(x_t)$ for all $h \in V_t$
8: end if
9: $V_t = \{h \in \mathcal{H} : h(x_t) = y_t \land (x_t, y_t) \in Z_t\}$
10: end for
11: return any $h \in V_n$

Algorithm 2 Efficient CAL

1: Initialize: $Z_0 = \emptyset$
2: for $t = 1, 2, \ldots, n$ do
3: Nature reveals unlabeled data point $x_t$
4: if for $\tilde{y} \in \{0, 1\}$ $\exists \tilde{h}_{\tilde{y}} \in \mathcal{H}$: $h_{\tilde{y}}(x_t) = y_t$, $\forall (x_s, y_s) \in Z_{t-1} \cup \{x_t, \tilde{y}\}$ then
5: Query $y_t$, and set $Z_t = Z_{t-1} \cup \{x_t, y_t\}$
6: else
7: $Z_t = Z_{t-1}$
8: end if
9: end for
10: return $\arg\min_{h \in \mathcal{H}} \sum_{(x, y) \in Z_t} 1\{h(x) \neq y\}$

\[ B(h, r) = \sup h' \in \mathcal{H} : \mathbb{E}_X \left[ \mathbb{I} h(x) \neq h'(x) \right] \leq r \]

**Definition 7.** The disagreement coefficient of $h \in \mathcal{H}$ with respect to a hypothesis class $\mathcal{H}$ and distribution $\mathcal{D}_X$ is defined as

\[ \theta^*_h = \sup_r \mathbb{P}_{X \sim \mathcal{D}_X} \left( X \in DIS(B(h, r)) \right) \]

**Theorem 11.** Let $h^* = \arg\min_{h \in \mathcal{H}} R(h)$ and assume $R(h^*) = 0$. Suppose $n$ iid labeled examples \{(x_i, y_i)\}_{i=1}^n are drawn from $\mathcal{D}$ and $V_n = \{h \in \mathcal{H} : h(x_i) = y_i \land y_i \in [n]\}$. If we request $\lambda$ additional labels only when the samples lie in the disagreement region $DIS(V_n)$, where $\lambda = 20h^* \log(|\mathcal{H}|/\delta)$, then, with probability greater than $1 - \delta$ we have $\sup_{h \in \mathcal{V}_{n+\lambda}} R(h) \leq \sup_{h \in V_n} \frac{1}{2} R(h)$.

\[
\mathbb{P}_X \left( X \in DIS(V_n) \right) \leq \mathbb{P}_X \left( X \in DIS \left( B(h^*, \frac{r}{\theta^*_h \circ R(h)}) \right) \right) \leq \Theta^*_h \Rightarrow \mathbb{P}_X \left( X \in DIS(V_n) \right) \leq \Theta^*_h \circ R(h)
\]

\[ \mathbb{E}_X \left[ \mathbb{I} h(x) \neq h^*(x) \right] = \mathbb{E}_X \left[ \mathbb{I} h(x) \neq Y \right] = R(h) \]
\[
\sup_{h \in V_n} R(h) = \sup_{h \in V_n} P(Y \neq h(x))
\]

\[
= \sup_{h \in V_n} P(Y \neq h(x) | x \in \text{DIS}(U_n)) P(x \in \text{DIS}(U_n))
\]

\[
\leq \log \left( \frac{12141}{d} \right) \leq \Theta^*_h \cdot \sup_{h \in V_n} R(h)
\]

\[
\leq \frac{1}{2} \sup_{h \in V_n} R(h).
\]

Conclude that every \( \lambda \) labels we halve the risk. If we take

\( n = k \lambda \) labels total then

w.p. \( 1 - k \delta \), \( R(\hat{h}) \leq 2^{-k} = 2^{-n/\lambda}
\]

\[
\leq \exp \left( -\frac{n}{2\Theta^*_h \log(12141/d)} \right)
\]

Equivalently, \( R(h) \leq \epsilon \) after \( \Theta^*_h \log(12141/d) \log(1/\epsilon) \) labels.
3 cases where $\Theta^* = \Theta(1)$.

Ex. $H$ are thresholds on unit interval ($D_x$ is uniform)

$$\mathbb{E}(h^*, r)$$

$$\text{Pr}(X \in \mathcal{B}(h^*, r)) = 2r$$

$$\Rightarrow \Theta^*(\varepsilon) \leq \frac{2r}{\varepsilon} = 2.$$

If $H$ halfspaces going through origin

and $D_x$ is uniform on sphere

then $\Theta^*(\varepsilon) \leq \sqrt{d}$
\[ \mathcal{H} \text{ are halfspaces that go through origin that are} \]
\[ \text{tilted to chop off a tiny bit of } \Gamma. \]

Assume \( h^* \in \mathcal{H} \) w/ \( \gamma = h^*(x) \)

\[ \Theta(x) = (1 - \delta) \Gamma + \delta \Gamma' \text{ for } \delta \ll 1 \]

If \( \delta = 1 \) CAL achieves \( \log(1/\varepsilon) \) labels.
If \( \delta = 0 \) any algorithm requires \( 1/\varepsilon \) labels.

If you run CAL and wait for samples only from \( \Gamma' \), requires \( \log(1/\varepsilon) \) labels and \( \frac{1}{\delta} \log(1/\varepsilon) \) unlabelled data.
Consider a finite hypothesis space \( \mathcal{H} \) and consider any \( Q \subset \binom{\mathcal{H}}{2} \) where \((h, h') \in Q\) can be considered an edge connecting any two hypotheses. For any \( \hat{y} \in \{0, 1\} \) define \( \mathcal{H}_{(x, \hat{y})} = \{ h \in \mathcal{H} : h(x) = \hat{y}\} \). We say an example \( x \) \( \rho \)-splits \( Q \) if requesting its label reduces the number of edges by at least a fraction \( \rho \in (0, 1) \):

\[
\max\{|Q \cap \mathcal{H}_{(x,0)}|, |Q \cap \mathcal{H}_{(x,1)}|\} \leq (1 - \rho)|Q|.
\]

We are now ready to introduce the splitting index.

**Definition 8.** Fix any subset \( S \subset \mathcal{H} \) and \( Q \subset \binom{\mathcal{S}}{2} \) such that \( \mathbb{P}(h(X) \neq h'(X)) \geq \epsilon, \forall (h, h') \in Q \). Then we say \( S \) is \((\rho, \epsilon, \tau)\)-splittable if \( \mathbb{P}(X \text{ splits } Q) \geq \tau \).

Basically, the definition is saying that to reduce the number of pairs of hypotheses that differ by at least \( \epsilon \) by a fraction at least \( \rho \), requires \( 1/\tau \) unlabeled data. If \( \mathcal{H} \) is finite, and \( \mathcal{H} \) is \((\rho, \epsilon, \tau)\)-splittable, then it is almost immediate that there exists an algorithm that requires \( 1/(\tau \rho) \) unlabeled data and \( 1/\rho \) labels to identify an \( \epsilon \)-good classifier ([Dasgupta, 2005b] suggests one, though it is computationally intractable). What is more important is the reproduced lower bound:

**Theorem 12 ([Dasgupta, 2005b]).** Fix any hypothesis space \( \mathcal{H} \) and distribution \( \mathcal{D} \) over \( \mathcal{X} \times \{0, 1\} \). Suppose that for some \( \rho \in (0, 1) \), \( \epsilon \in (0, 1) \) and some \( \tau \in (0, 1/2) \), the set \( S \subset \mathcal{H} \) is not \((\rho, \epsilon, \tau)\)-splittable. Then any active learning strategy that achieves an accuracy of \( \epsilon/2 \) on all target hypotheses in \( S \) must, with probability at least \( 3/4 \) (taken over the random sampling of data), either draw \( \geq 1/\tau \) unlabeled samples, or must request \( \geq 1/\rho \) labels.