6. Appendix

6.1. Notation

$w_0$ is the weights after the seed round.

$A_{-1}$ is the matrix without the first row and column. $A_{1,-1}$ is the vector from the first row and all columns except the first column.

Generally, the $O(f(n))$ notation hides constants that only depend on the dataset, such as $\|w^*\|$, $s$, $B$, etc.

For the order of things going to zero, we first choose $\alpha$ to be small, then $r$ to be small, then $n$ to be large.

$w_0$ is weight vector after seed round

$$\epsilon_{\text{active}}(n) = \mathbb{E}_{f \sim \text{active, npoints}}[\text{Err}(f)]$$

$$\epsilon_{\text{passive}}(n) = \mathbb{E}_{f \sim \text{passive, npoints}}[\text{Err}(f)]$$

$$DE(\epsilon) = \max\{n : \epsilon_{\text{passive}}(n) \geq \epsilon\} / \max\{n : \epsilon_{\text{active}}(n) \geq \epsilon\}$$

Without loss of generality, assume $w^* = \|w^*\| e_1$, $w^*_0 = 0$, and $\mathbb{E}[x_2] = 0$.

With an abuse of notation, let $\sigma = \sigma(w^* \cdot x) = \sigma(\|w^*\| x_1)$.

6.2. Losses

Define $\sigma(x) = \frac{1}{1 + \exp(x)}$.

The loss (negative log-likelihood) for a single data point under logistic regression is

$$l_w(x, y) = \log(1 + \exp(-w \cdot yx))$$

and so the gradient is

$$\nabla l_w(x, y) = -\frac{yx \exp(-w \cdot yx)}{1 + \exp(-w \cdot yx)} = -yx \sigma(-w \cdot yx)$$

and the Hessian is

$$\nabla^2 l_w(x, y) = \frac{(yx)(yx)^T \exp(w \cdot yx)}{(1 + \exp(w \cdot yx))^2}$$

$$= \frac{xx^T}{(1 + \exp(w \cdot yx))(1 + \exp(-w \cdot yx))}$$

$$= \sigma(w \cdot yx) \sigma(-w \cdot yx) xx^T$$

Note that $\sigma(-x) = 1 - \sigma(x)$.

6.3. Decision Boundary

Lemma 6.1. For sufficiently small $r$, if $\|w' - w^*\|_2 \leq 2r$, then

$$|\int_{w' \cdot x = 0} p(x) - \int_{w^* \cdot x = 0} p(x)| = O(r)$$
Proof. Without loss of generality (rotation and translation), let \( w_0^* = 0, w^* = \|w^*\|c_1 \) and let \( w' = w'_1c_1 + w'_2c_2 \).

We sample from places where \( w_0' + w'_1x_1 + w'_2x_2 = 0 \) which occurs when \( x_1 = -w'_12x_2 + \frac{w'_2}{w'_1}x_2 = ax_2 + b \). From the theorem assumption, we know that \( |w_0'|, |w'_2| \leq r \) and \( |w'_1| \geq \|w^*\| - r \geq \frac{1}{2}\|w^*\| \) (for sufficiently small \( r \)) so we know that \( |a|, |b| \leq O(r) \)

Note that

\[
\left| \int_{w':x=0} p(x) - \int_{w^*:x=0} p(x) \right| = \left| \int_x p(x_1 = ax + b, x_2 = x) - p(x_1 = 0) \right|
\]

(Note that the Jacobian of the change of variables has the following matrix which has determinant 1)

\[
\begin{bmatrix}
1 & 0 \\
-a & 1
\end{bmatrix}
\]

\[
\left| \int_{w':x=0} p(x) - \int_{w^*:x=0} p(x) \right| \leq \int_x |p(x_1 = ax + b|x_2 = x)p(x_2 = x) - p(x_1 = 0|x_2 = x)p(x_2 = x)|
\]

With the assumption that the conditional probabilities are Lipschitz,

\[
\leq \int_x |L(ax + b)p(x_2 = x) |
\]

\[
\leq aLB + bL
\]

\[
= O(r)
\]

Lemma 6.2. For sufficiently small \( r \), if \( \|w_0 - w^*\| \leq r \), then with probability going to 1 exponentially fast, all points from two-stage uncertainty sampling are from some hyperplane \( w' \) such that \( \|w' - w^*\| \leq 2r \).

Proof. For small enough \( r \), then \( \int_{w':x=0} p(x) > p_0/2 \) from the above lemma if \( \|w_0 - w^*\| \leq 2r \). Thus, the probability of an unlabeled point within the parallel plane with bias less than \( r \) different from \( w_0 \) such that \( \|w' - w_0\| \leq r \) is at least \( 2\frac{r}{\|w_0\|}p_0/2 \geq rP_0 \Theta(r) \) (for sufficiently small \( r \)).

Recall that \( n_{pool} = \omega(n) \) and \( n_{seed} = o(n) \).

For sufficiently large \( n \), the probability of at least \( n \) points from the \( n_{pool} - n_{seed} \) unlabeled points falling in this range is

\[
\Pr[Binomial(n_{pool} - n_{seed}, \text{probability of falling}) \geq n] \geq \Pr[Binomial(n_{pool}/2, C_1r) \geq n]
\]

for some constant \( C_1 \).

We can use a Chernoff bound (standard with \( \delta = 1/2 \)) since \( n_{pool} = \omega(n) \) to bound by \( \exp(-\omega(n)) \). Thus the probability that the planes we choose from are farther than \( r \) away from \( w_0 \) goes to 0 with rate faster than \( \exp(-n) \).

6.4. Convergence

Lemma 4.2. Both two-stage uncertainty sampling and random sampling converge to \( w^* \).

Proof. For passive learning, the Hessian of the population loss is positive definite because the data covariance is non-singular (Assumption 8). Thus, the population loss has a unique optimum. By the definition of \( w^* \), \( w^* \) is the minimizer. Since the sample loss converges to the population loss, the result of passive learning converges to \( w^* \).
By a similar argument, the weight vector $w_0$ after the seed round converges to $w^*$ since $n_{\text{seed}}$ is super-constant (Assumption 2). Thus, for any $r > 0$, with probability converging to 1 as $n \to \infty$, $\|w_0 - w^*\| \leq r \leq \lambda/2$. By Lemma 6.2, with probability going to 1, all points selected are from hyperplanes $w$ where $\|w - w^*\| \leq 2r \leq \lambda$. Thus, by Assumption 5, $E_{w,x \sim 0}(\nabla l_{w^*}(x,y)) = 0$. In the second stage, because of the $\alpha$ proportion of randomly selected points, the loss from the new uncertainty sampling population has a unique optimum. And because the expectation of the gradient of the loss is 0 for the points near the decision boundary (with probability going to 1), the result of two-stage uncertainty sampling converges in probability to $w^*$.

\[ \square \]

6.5. Rates

**Lemma.** If $\Sigma$ exists, and for any $\epsilon > 0$, $n \Pr[\|A_n - A\| \geq \epsilon] \to 0$ and $n \Pr[\|w_n - w^*\| \geq \epsilon] \to 0$, then there exist vectors $c_k \neq 0$ that depend only on the data distribution such that,

\[ n(\epsilon(n) - \text{Err}) \to \sum_k c_k^T \Sigma_{-1} c_k \]

**Proof.** The zero-one error is

\[ Z(w_n) = \Pr[y x \cdot w_n < 0] \]

Since $Z$ is twice differentiable at $w^*$, by Taylor’s theorem,

\[ Z(w_n) = Z(w^*) + (\nabla Z(w^*))^T (w_n - w^*) + (w_n - w^*)^T (\frac{1}{2} \nabla^2 Z(w^*)) (w_n - w^*) + (w_n - w^*)^T R(w_n - w^*) (w_n - w^*)^T \]

where $R(w) \to 0$ as $w \to 0$.

Since $Z$ has a local optimum at $w^*$, $\nabla Z(w^*) = 0$. Also $Z(w^*) = \text{Err}$. Additionally, denote $H = \frac{1}{2} \nabla^2 Z(w^*)$,

\[ Z(w_n) = \text{Err} + (w_n - w^*)^T (H + R(w_n - w^*)) (w_n - w^*) \]

Choose any $\epsilon > 0$. Since $R(w) \to 0$ as $w \to 0$, there is $\delta_\epsilon$ such that $\|w\| \leq \delta_\epsilon$ implies $\|R(w)\| \leq \epsilon$. Define $\text{nearn}(n)$ to be the event that $\|A_n - A\| \geq \epsilon \& \|w_n - w^*\| \geq \delta_\epsilon$. Note that from the theorem assumption, $n \Pr[\neg \text{nearn}(n)] \to 0$.

\[ \epsilon(n) = \mathbb{E}[Z(w_n)] = \Pr[\neg \text{nearn}(n)] \mathbb{E}[Z(w_n) | \neg \text{nearn}(n)] + \Pr[\text{nearn}(n)] \mathbb{E}[Z(w_n) | \text{nearn}(n)] \]

\[ |n\epsilon(n) - n\mathbb{E}[Z(w_n) | \text{nearn}(n)]| \leq n \Pr[\neg \text{nearn}(n)] \mathbb{E}[Z(w_n) | \neg \text{nearn}(n)] - \mathbb{E}[Z(w_n) | \text{nearn}(n)] \]

\[ \leq n \Pr[\neg \text{nearn}(n)] \to 0 \]

Thus,

\[ n(\epsilon(n) - \text{Err}) \to n(\mathbb{E}[Z(w_n) | \text{nearn}(n)] - \text{Err}) \]

So we need to just worry about the convergence of the right side,

\[ \mathbb{E}[Z(w_n) | \text{nearn}(n)] = \text{Err} + \frac{1}{n} \mathbb{E}[(A_n^{-1} b_n)^T (H + R(w_n - w^*)) (A_n^{-1} b_n) | \text{nearn}(n)] \]

\[ n(\mathbb{E}[Z(w_n) | \text{nearn}(n)] - \text{Err}) = \mathbb{E}[b_n^T A_n^{-1} (H + R(w_n - w^*)) A_n^{-1} b_n | \text{nearn}(n)] \]
Because we conditioned on $\text{near}(n)$, $\|A_n - A\| \leq \epsilon$ and $\|w_n - w^*\| \leq \delta_n$ and therefore $\|R(w_n - w^*)\| \leq \epsilon$. So $\|A_n^{-1}(H + R(w_n - w^*))A_n^{-1} - A^{-1}HA^{-1}\| = O(\epsilon)$. Using this, we get,

$$
\|n(E[Z(w_n)\mid \text{near}(n)] - \text{Err}) - E[b_n^T A^{-1} H A^{-1} b_n \mid \text{near}(n)]\| \leq O(\epsilon)\|E[\|b_n\|^2 \mid \text{near}(n)]\|
$$

$$
\leq O(\epsilon)\|E[|b_n b_n^T| \mid \text{near}(n)]\|
$$

Note that,

$$
E[b_n b_n^T] = E[b_n b_n^T \mid \text{near}(n)] \Pr[\text{near}(n)] + E[b_n b_n^T \mid \text{not near}(n)] \Pr[\text{not near}(n)]
$$

and the later two expectations exist since the left exists and the matrices are positive semidefinite. Passing through the limit, we see that $E[b_n b_n^T \mid \text{near}(n)] \to B$.

Thus, noting that we can drive $\epsilon \to 0$,

$$
n(E[Z(w_n)\mid \text{near}(n)] - \text{Err}) \to E[b_n^T A^{-1} H A^{-1} b_n \mid \text{near}(n)]
$$

$$
\to \sum_{i,j} [A^{-1}HA^{-1}]_{i,j} E[b_n b_n^T \mid \text{near}(n)]_{i,j}
$$

$$
\to \sum_{i,j} [A^{-1}HA^{-1}]_{i,j} B_{i,j}
$$

Thus, putting this together, we see that

$$
n(\epsilon(n) - \text{Err}) \to \sum_{i,j} [A^{-1}HA^{-1}]_{i,j} B_{i,j}
$$

Doing manipulations on the indices, we find,

$$
\sum_{i,j} [A^{-1}HA^{-1}]_{i,j} B_{i,j} = \sum_{i,j} H_{i,j}(A^{-1}BA^{-1})_{i,j}
$$

$$
= \sum_{i,j} H_{i,j} \Sigma_{i,j}
$$

Therefore,

$$
n(\epsilon(n) - \text{Err}) \to \sum_{i,j} H_{i,j} \Sigma_{i,j}
$$

and we are most of the way there, just need to use some properties to show the final form.

Since $w^*$ is a local optimum, $H \succeq 0$ (and symmetric) and since the Hessian is not identically zero at $w^*$, $H \neq 0$.

Without loss of generality, let $w^* = \|w^*\|e_1$ and $w_0^* = 0$ as assumed before. Note that $Z(w^* + \alpha e_1) = Z(w^*)$ for $\alpha \in (-\|w^*\|/2, \infty)$. Since it is constant along this line, $(\nabla^2 Z(w^*))_{1,1} = 0$, and so $H_{1,1} = 0$.

So $H \succeq 0$, $H$ is symmetric, $H \neq 0$, and $H_{1,1} = 0$. Since $H \succeq 0$ and $H_{1,1} = 0$, $H_{1,i} = 0$ for all $i$.

Since $H \succeq 0$ and $H \neq 0$,

$$
H = \sum_k c_k c_k^T
$$

for some vectors $c_k$ (where there is at least one). And further, $(c_k)_1 = 0$. 

On the Relationship between Data Efficiency and Error for Uncertainty Sampling
On the Relationship between Data Efficiency and Error for Uncertainty Sampling

\[
\sum_{i,j} H_{i,j} \Sigma_{i,j} = \sum_{i,j} \left( \sum_k c_k e_k^T \right)_{i,j} \Sigma_{i,j} = \sum_k e_k^T \Sigma e_k
\]

We can remove the first elements of \( c_k \) and the first row and column of \( \Sigma \) without changing anything, so

\[
\sum_{i,j} H_{i,j} \Sigma_{i,j} = \sum_k e_k^T \Sigma_{i,j} - 1 c_k
\]

And thus the theorem is proved.

**Lemma.** If we have two algorithms \( a \) and \( b \) that satisfy the conditions of Lemma 2, and

\[ \Sigma_{a,-1} \succ c \Sigma_{b,-1} \]

then there exists \( \epsilon_0 \) such that for \( \text{Err} < \epsilon < \epsilon_0 \),

\[ n_a(\epsilon) \geq c n_b(\epsilon) \]

**Proof.**

\[ \Sigma_{a,-1} \succ \alpha \Sigma_{b,-1} \]

\[ \sum_k e_k^T \Sigma_{a,-1} c_k > \alpha \sum_k e_k^T \Sigma_{b,-1} c_k \]

so, for \( n > n_0, n' > n_0 \),

\[ n(\epsilon_a(n) - \text{Err}) > \alpha n'(\epsilon_b(n') - \text{Err}) \]

setting \( n' = n/\alpha \) and for \( n > \max(n_0, n_0/\alpha) \),

\[ n(\epsilon_a(n) - \text{Err}) > n(\epsilon_b(n/\alpha) - \text{Err}) \]

So for sufficiently large \( n \),

\[ \epsilon_a(n) > \epsilon_b(n/\alpha) \]

For any \( \epsilon > \text{Err} \) such that \( n_a(\epsilon) \) is sufficiently large, (we know this exists since \( n_a(\epsilon) = \Theta(\frac{1}{\epsilon - \text{Err}}) \))

\[ \epsilon_a(n) \leq \epsilon \text{ for } n \geq n_a(\epsilon) \]
\[ \epsilon_b(n/\alpha) \leq \epsilon \text{ for } n \geq n_a(\epsilon) \]
\[ \epsilon_b(n') \leq \epsilon \text{ for } n' \geq \frac{1}{\alpha} n_a(\epsilon) \]
\[ n_b(\epsilon) \leq \frac{1}{\alpha} n_a(\epsilon) \]
\[ n_a(\epsilon) \geq \alpha n_b(\epsilon) \]
Lemma 4.1. If we have two algorithms with $\Sigma_a$ and $\Sigma_b$, and for any $\epsilon > 0$ and both estimators, $n \Pr[\|A_n - A\| \geq \epsilon] \rightarrow 0$ and $n \Pr[\|w_n - w^*\| \geq \epsilon] \rightarrow 0$, then

$$\Sigma_{a,-1} > c\Sigma_{b,-1}$$

implies that for some $\epsilon_0$ and any $\text{Err} < \epsilon < \epsilon_0$,

$$n_a(\epsilon) \geq c n_b(\epsilon)$$

Proof. This is a straightforward application of the above lemmas, Lemma 2 and Lemma 3.

6.6. Conditions satisfied

Lemma 4.3. For our active and passive learning algorithms, for any $\epsilon > 0$, $n \Pr[\|A_n - A\| \geq \epsilon] \rightarrow 0$ and $n \Pr[\|w_n - w^*\| \geq \epsilon] \rightarrow 0$

Proof. Recall that

$$A_n = \frac{1}{n} \sum_i \nabla^2 l_w(x_i, y_i)$$

$$b_n = \frac{1}{\sqrt{n}} \sum_i \nabla l_{w^*}(x_i, y_i)$$

where $\|w' - w^*\| \leq \|w_n - w^*\|$.

For passive learning, by CLT, for any $\epsilon$, $\Pr[\|w_n - w^*\| > \epsilon] = O\left(\frac{\epsilon}{\sqrt{n}}\right)$. Thus, we find that $n \Pr[\|w_n - w^*\| \geq \epsilon] \rightarrow 0$. We also need this fact to bound $w'$. Then, with a Hoeffding bound on the sum of $A_n$, we can get that $\Pr[\|A_n - A\| \geq \epsilon] = O\left(\frac{\epsilon}{\sqrt{n}}\right)$ and thus $n \Pr[\|A_n - A\| \geq \epsilon] \rightarrow 0$.

For active learning, we need to be careful because if $\|w_0 - w^*\| > \lambda/2$, we are not even guaranteed that the final result converges (see Lemma 6.2). However, by the CLT, we find that $\Pr[\|w_0 - w^*\| > \lambda/2] = O\left(\frac{\epsilon}{\sqrt{n_{\text{seed}}}}\right)$. Because $n_{\text{seed}} = \Omega(n^\alpha)$ (see Assumption 2), this converges exponentially fast and $n \Pr[\|w_0 - w^*\| > \lambda/2] \rightarrow 0$.

Because of the $\sigma$ random sampling, and conditioned on the probability that $\|w_0 - w^*\| < \lambda/2$, we can get the same results for active learning as for passive learning. Note that from Lemma 6.2, there is exponentially small probability of not sampling all points from $w'$ where $\|w' - w^*\| < \lambda$.

6.7. COV calculation for passive

Lemma 6.3. For passive learning, $\mathbb{E}[\nabla l_{w^*}(x, y)(\nabla l_{w^*}(x, y))^T] = \mathbb{E}[\sigma(1 - \sigma)x x^T]$.

Proof. Since the mean of the derivative of the loss is 0 at $w^*$,

$$\mathbb{E}[\nabla l_{w^*}(x, y)(\nabla l_{w^*}(x, y))^T]_{i,j} = \mathbb{E}[x_i x_j \sigma(-\|w^*\|y x_1)^2]$$

$$= \mathbb{E}[x_i \mathbb{E}[x_i x_j | x_1] \mathbb{E}[\sigma(\|w^*\|y x_1)^2 | x_1]]$$

$$= \mathbb{E}[x_i | \mathbb{E}[x_i x_j | x_1] \mathbb{E}[P(y = 1 | x_1) \sigma(-\|w^*\|x_1)^2 + P(y = 1 | x_1) \sigma(\|w^*\|x_1)^2]]$$

from the calibrated assumption,
\[
E[x_1]\mathbb{E}[x_i x_j | x_1]|\sigma(\|w^*\| | x_1)\sigma(-\|w^*\| | x_1)^2 + \sigma(\|w^*\| | x_1)^2] \\
= E[x_1]\mathbb{E}[x_i x_j | x_1]|\sigma(\|w^*\| | x_1)\sigma(-\|w^*\| | x_1)\sigma(\|w^*\| | x_1) + \sigma(\|w^*\| | x_1)] \\
= E[x_1]\mathbb{E}[x_i x_j | x_1]|\sigma(\|w^*\| | x_1)(\|w^*\| | x_1)] \\
= E[x_i x_j \sigma(\|w^*\| | x_1)\sigma(-\|w^*\| | x_1)] \\
= E[\sigma(1 - \sigma)xx^T]_{i,j}
\]

Lemma 4.4.
\[\Sigma_{\text{passive}} = [E[\sigma(1 - \sigma)xx^T]]^{-1}\]

Proof. For passive learning, by the convergence of \(w^n \to w^*\) and by the law of large numbers,

\[A_n \to A = E[\sigma(1 - \sigma)xx^T]\]

Further, by independence of draws,

\[E[b_n b_n^T] = E[\nabla l_w^*(x,y)(\nabla l_w^*(x,y))^T]\]

so by Lemma 6.3,

\[E[b_n b_n^T] = E[\sigma(1 - \sigma)xx^T]
B = E[\sigma(1 - \sigma)xx^T]
B = A\]

Thus,

\[\Sigma_{\text{passive}} = A^{-1}BA^{-1}\]

\[= A^{-1}\]

\[= [E[\sigma(1 - \sigma)xx^T]]^{-1}\]

\[\square\]

6.8. COV calculation for active

Lemma 6.4. For sufficiently small \(r\) (small with respect to dataset-only dependent constants), if \(\|w' - w^*\|_2 \leq 2r\), then

\[\|E_{w' \cdot x=0}[\sigma(1 - \sigma)xx^T] - E_{w \cdot x=0}[\sigma(1 - \sigma)xx^T]\| = O(r)\]

and

\[\|E_{w' \cdot x=0}[\sigma(-y x_1\|w^*\|)^2xx^T] - E_{w \cdot x=0}[\sigma(-y x_1\|w^*\|)^2xx^T]\| = O(r)\]

Proof. Without loss of generality (rotation and translation), let \(w_0' = 0, w^* = \|w^*\|e_1\) and let \(\hat{w} = c_1 e_1 + c_2 e_2\).

We sample from places where \(w_0' + w_1' x_1 + w_2' x_2 = 0\) which occurs when \(x_1 = \frac{w_0'}{w_3} x_2 + \frac{w_0'}{w_1} = ax_2 + b\). From the theorem assumption, we know that \(|w'_0|, |w'_1| \leq r\) and \(|w'_2| \geq \|w^*\| - r \geq \frac{1}{2}\|w^*\|\) (for sufficiently small \(r\)) so we know that \(|a|, |b| \leq O(r)\)
Define \( Q(x_1) = \sigma(||w^*||x_1)\sigma(-||w^*||x_1) \) or \( Q(x_1) = \sigma(-yx_1||w^*||)^2 \) (abuse of notation). Both these functions are Lipschitz around \( x_1 = 0 \), and bounded (since support bounded by \( B \)).

First, we compute the joint (not the conditionals) and then we can divide by the marginals from the previous lemma, Let \( i_1, i_2, \ldots, i_d \) be indicators for the indices \( i, j \) that are non-zero. Thus, \( i_1 + i_2 + \ldots + i_d \leq 2 \),

\[
\mathbb{E}_{w',x=0}[\sigma(1-\sigma)x x^T]_{i,j} = \\
= \mathbb{E}_{w',x=0}[Q(x_1)(x_1)^{i_1}(x_2)^{i_2}(x_3)^{i_3} \ldots] =
\]

(As before, the Jacobian of the change of variables has determinant 1)

\[
\int_x p(x_1 = ax + b, x_2 = x)Q(ax + b)(ax + b)^{i_1}(x)^{i_2}\mathbb{E}[x^{i_3} \ldots | x_1 = ax + b, x_2 = x] = \\
= \int_x p(x_2 = x)(x)^{i_2}F(ax + b, x)
\]

where \( F(x_1, x_2) = p(x_1 | x_2)Q(x_1)x^{i_1} \) \( O \left( \mathbb{E}[x^{i_3} \ldots | x_1, x_2] \right) \)

All three components of \( F \) are bounded, since support bounded, Assumption 3. Further, all three components are Lipschitz, because of Assumption 4 and bounded support as well. Therefore, \( F \) is Lipschitz.

\[
| \int_x p(x_2 = x)(x)^{i_2}F(ax + b, x) - \int_x p(x_2 = x)(x)^{i_2}F(0, x) | \\
\leq \int_x p(x_2 = x)|x|^{i_2}L|ax + b| \\
\leq aLB^{i_2+1} + bLB^2 \\
= O(r)
\]

Thus, for any \( i, j \),

\[
\| \mathbb{E}_{w',x=0}[Qxx^T]_{i,j} - \mathbb{E}_{w',x=0}[Qxx^T]_{i,j} \| = O(r)
\]

We can use this to bound the matrix norm,

\[
\| \mathbb{E}_{w',x=0}[Qxx^T] - \mathbb{E}_{w',x=0}[Qxx^T] \| = O(r)
\]

Since the probabilities (see Lemma 6.1) and conditionals are both off by only \( O(r) \) (from above) and since the probabilities are bounded away from 0 (see Lemma 6.1 and Assumption 8), the conditional distribution is off by \( O(r) \). We can plug in both functions of \( Q \) to get the statement of the theorem.

\[ \square \]

**Lemma 4.5.**

\[
\Sigma_{active} = [(1-\alpha)\mathbb{E}_{x_1=0}[\sigma(1-\sigma)x x^T] + \alpha\mathbb{E}[\sigma(1-\sigma)x x^T]]^{-1}
\]

**Proof.** Because \( w_n \to w^* \), and by the law of large numbers,

\[
A_n \to (1-\alpha)\mathbb{E}_{w'}[\mathbb{E}_{w',x=0}[\sigma(-yx_1||w^*||)^2xx^T]] + \alpha\mathbb{E}[\sigma(-yx_1||w^*||)^2xx^T]
\]

From Lemma 6.4,
On the Relationship between Data Efficiency and Error for Uncertainty Sampling

\[ \|E_{w',x=0}\sigma(1-\sigma)xx^T - E_{w^*,x=0}\sigma(1-\sigma)xx^T\| = O(r) \]

and \( \|w' - w^*\| < 2r \) with probability going to 1.

\[ A_n \to \frac{n - n_{\text{seed}}}{n}[(1 - \alpha)E_{w^*,x=0}\sigma(1-\sigma)xx^T] + O(r) + \alpha E(\sigma(1-\sigma)xx^T) \]

Since \( w_n \to w^* \), \( r \to 0 \), and since \( n_{\text{seed}} = o(n) \) (see Assumption 2) so

\[ A_n \to A = (1 - \alpha)E_{x,=0}\sigma(1-\sigma)xx^T + \alpha E(\sigma(1-\sigma)xx^T) \]

The same line of argument with using Lemma 6.4 and Lemma 6.3 yields

\[ B = A \]

So

\[ \Sigma_{\text{active}} = A^{-1}BA^{-1} = A^{-1} \]

\[ = [(1 - \alpha)E_{x,=0}\sigma(1-\sigma)xx^T + \alpha E(\sigma(1-\sigma)xx^T)]^{-1} \]

\[ \square \]

6.9. Inverses Without First Coordinate

Lemma 6.5.

\[ \begin{bmatrix} a & \bar{a}^T \\ \bar{a} & A \end{bmatrix}^{-1} = \begin{bmatrix} b & \bar{b}^T \\ \bar{b} & B \end{bmatrix} \]

Where

\[ b = \frac{1}{a - \bar{a}^T A^{-1} \bar{a}} \]

\[ \bar{b} = -bA^{-1} \bar{a} \]

\[ B = A^{-1} + b(A^{-1} \bar{a})(A^{-1} \bar{a})^T \]

Proof. Matrix algebra. \[ \square \]

Lemma 6.6.

\[ (A^{-1})^{-1} = (A^{-1})^{-1} + \frac{(A^{-1})^{-1}A_{-1,1})(A^{-1})^{-1}A_{-1,1})^T}{A_{1,1} - A_{1,1}^T A_{-1,1}(A^{-1})^{-1}A_{-1,1}} \]

Proof. Use the above theorem and note that \( b > 0 \) so

\[ b(A^{-1} \bar{a})(A^{-1} \bar{a})^T \geq 0 \]

\[ \square \]
6.10. Relating $\text{Err}$ to expectation of sigmoid

**Lemma 6.7.**

\[ \frac{\text{Err}}{2} < \mathbb{E}[\sigma(1 - \sigma)] < \text{Err} \]

**Proof.**

\[ \text{Err} = P(y|x_1\|w^\star\| < 0) \]
\[ = P(x_1 < 0 \land y = 1) + P(x_1 > 0 \land y = -1) \]

From Assumption 7,

\[ = \int_{-\infty}^{0} p_{x_1}(x_1)\sigma(-w_1^\star x_1) + \int_{0}^{\infty} p_{x_1}(x_1)\sigma(w_1^\star x_1) \]
\[ = \int_{0}^{\infty} [p_{x_1}(-x_1) + p_{x_1}(x_1)]\sigma(w_1^\star x_1) \]

Additionally,

\[ \mathbb{E}[\sigma(1 - \sigma)] = \mathbb{E}[\sigma(y|x_1\|w^\star\|)\sigma(-y|x_1\|w^\star\|)] \]
\[ = \mathbb{E}[\sigma(\|w^\star\| x_1)\sigma(-\|w^\star\| x_1)] \]
\[ = \int_{-\infty}^{0} p_{x_1}(x_1)\sigma(\|w^\star\| x_1)\sigma(-\|w^\star\| x_1) + \int_{0}^{\infty} p_{x_1}(x_1)\sigma(\|w^\star\| x_1)\sigma(-\|w^\star\| x_1) \]
\[ = \int_{0}^{\infty} [p_{x_1}(-x_1) + p_{x_1}(x_1)]\sigma(\|w^\star\| x_1)\sigma(-\|w^\star\| x_1) \]

Note that for $x_1 > 0$, $\frac{1}{2} < \sigma(-\|w^\star\| x_1) < 1$. Comparing equations, we get,

\[ \frac{\text{Err}}{2} < \mathbb{E}[\sigma(1 - \sigma)] < \text{Err} \]

6.11. Main DE bound

**Theorem 4.1.** For sufficiently small constant $\alpha$ (that depends on the dataset) and for $\text{Err} < \epsilon < \epsilon_0$,

\[ \text{DE}(\epsilon) > \frac{s}{4\epsilon \text{Err}} \]

**Proof.** For convenience, define

\[ Q = \mathbb{E}_{x_1 = 0}[\sigma(1 - \sigma)x x^T] \]
\[ R = \mathbb{E}[\sigma(1 - \sigma)x x^T] = COV_{\text{passive}} \]
\[ S = \alpha R + (1 - \alpha)Q = COV_{\text{active}} \]

By the definition of $s$,

\[ \mathbb{E}_{x_1 = 0}[x_{-1}x_{-1}^T] \geq \frac{s \mathbb{E}[\sigma(1 - \sigma)x_{-1}x_{-1}^T]}{\mathbb{E}[\sigma(1 - \sigma)]]} \]
By Lemma 6.7,

\[ 4Q_1 \succ \frac{s}{Err} R_1 \]

For small enough \( \alpha \),

\[ Q_1 \succ \frac{s / (4Err) - \alpha}{1 - \alpha} R_1 \]

\[ \alpha R_1 + (1 - \alpha)Q_1 \succ \frac{s}{4Err} R_1 \]

\[ S_1 \succ \frac{s}{4Err} R_1 \]

\[ \frac{s}{4Err} (S_1)^{-1} \times (R_1)^{-1} \preceq (R^{-1})_1 \]

The last step comes from noting that the right hand side of Lemma 6.6 positive semidefinite for \( A \) positive semidefinite.

Additionally, note that the first row and column of \( Q \) is 0, so \( S_{-1,1} = \alpha R_{-1,1} \) and \( S_{1,1} = \alpha R_{1,1} \).

An examination yields,

\[ \frac{(S_1)^{-1}S_{-1,1}(S_1)^{-1}S_{-1,1}^T}{S_{1,1} - S_{-1,1}^T(S_1)^{-1}S_{-1,1}} = O(\alpha) \]

Using Lemma 6.6, we find that we can make \( \alpha \) small enough so that

\[ \frac{s}{4Err} (S^{-1})_1 \prec (R^{-1})_1 \]

\[ \frac{s}{4Err} COV_{active, -1} \prec COV_{passive, -1} \]

so by Lemma 4.1, for \( Err < \epsilon \),

\[ DE(\epsilon) > \frac{s}{4Err} \]

\[ \square \]

6.12. DE Bound Given Decomposition

We actually get a slightly more general result from the following lemma.

**Lemma 6.8.** If \( p(x) = p(x_1)p(x_{-1}) \), then for sufficiently small constant \( \alpha \) (that depends on the dataset), and for \( Err < \epsilon < \epsilon_0 \),

\[ \frac{1}{4Err} < DE(\epsilon) < \frac{1}{2Err}(1 + \frac{E[\tilde{X}]}{Var(X)}) \]

where

\[ p(\tilde{X} = x) \propto \sigma(\|w^*\|_x)(1 - \sigma(\|w^*\|_x))p(x_1 = x) \]
Proof. With the decomposition, in the Theorem 4.1, \( s = 1 \). So we get for free that for \( Err \prec \epsilon \prec \epsilon_0 \),

\[
DE(\epsilon) > \frac{1}{4\epsilon r}
\]

As before, for convenience, define

\[
Q = E_{x_1=0}[\sigma(1-\sigma)x x^T]
\]

\[
R = E[\sigma(1-\sigma)x x^T] = COV_{\text{passive}}
\]

\[
S = \alpha R + (1-\alpha)Q = COV_{\text{active}}
\]

Because of the decomposition,

\[
R_{2,2} = E[\sigma(1-\sigma)]E[x_2 x_2^T] \succ \frac{Err}{2} E[x_2 x_2^T]
\]

\[
Q_{2,2} = \frac{1}{4} E[x_2 x_2^T]
\]

\[
Q_{2,2} \prec \frac{1}{2E\epsilon} R_{2,2}
\]

For sufficiently small \( \alpha \),

\[
Q_{2,2} \prec \frac{1/(2Err) - \alpha}{1-\alpha} R_{2,2}
\]

\[
\alpha R_{2,2} + (1-\alpha)Q_{2,2} \prec \frac{1}{2E\epsilon} R_{2,2}
\]

\[
S_{2,2} \prec \frac{1}{2E\epsilon} R_{2,2}
\]

Because of the decomposition, and because \( E[x_2] = 0 \) (without loss of generality by translation),

\[
R_{0:1,2} = 0
\]

\[
Q_{0:1,2} = 0
\]

\[
\frac{1}{2E\epsilon} (A^{-1})_{2,2} \succ (R^{-1})_{2,2}
\]

Now, let us examine the upper left corners,

\[
R_{0:1,0:1} = \begin{bmatrix}
E[\sigma(1-\sigma)] & E[\sigma(1-\sigma)x_1] \\
E[\sigma(1-\sigma)x_1] & E[\sigma(1-\sigma)x_1^2]
\end{bmatrix}
\]

\[
S_{0:1,0:1} = \begin{bmatrix}
(1-\alpha)/4 + \alpha E[\sigma(1-\sigma)] & \alpha E[\sigma(1-\sigma)x_1] \\
\alpha E[\sigma(1-\sigma)x_1] & \alpha E[\sigma(1-\sigma)x_1^2]
\end{bmatrix}
\]

Denote

\[
D = E[\sigma(1-\sigma)]E[\sigma(1-\sigma)x_1^2] - E[\sigma(1-\sigma)x_1]^2
\]

Then,
\[
\begin{align*}
(R^{-1})_{0,0} &= \frac{\mathbb{E}[\sigma(1-\sigma)x_1^2]}{D} \\
(S^{-1})_{0,0} &= \frac{\alpha\mathbb{E}[\sigma(1-\sigma)x_1^2]}{\alpha(1-\alpha)(1/4)\mathbb{E}[\sigma(1-\sigma)x_1^2] + \alpha^2D} \\
(R^{-1})_{0,0}/(S^{-1})_{0,0} &= \frac{1-\alpha}{4\mathbb{E}[\sigma(1-\sigma)]}(1 + \frac{\mathbb{E}[\sigma(1-\sigma)x_1^2]}{D}) + \alpha
\end{align*}
\]

For small enough \(\alpha\),

\[
(R^{-1})_{0,0}/(S^{-1})_{0,0} < \frac{1}{2\text{Err}}(1 + \frac{\mathbb{E}[\sigma(1-\sigma)x_1^2]}{D})
\]

Combining the bounds on the two blocks of the matrices, we get that

\[
\frac{1}{2\text{Err}}(1 + \frac{\mathbb{E}[\sigma(1-\sigma)x_1^2]}{D})(S^{-1})_{-1} \succ (R^{-1})_{-1}
\]

\[
\frac{1}{2\text{Err}}(1 + \frac{\mathbb{E}[\sigma(1-\sigma)x_1^2]}{D})\text{COV}_{active,-1} \succ \text{COV}_{passive,-1}
\]

So for \(\epsilon < \epsilon_0\),

\[
\text{DE}(\epsilon) < \frac{1}{2\text{Err}}(1 + \frac{\mathbb{E}[\sigma(1-\sigma)x_1^2]}{D})
\]

if we define \(\tilde{X}\) such that \(p_{\tilde{X}}(x) \propto \sigma(1-\sigma)p_{x_1}(x)\),

\[
\text{DE}(\epsilon) < \frac{1}{2\text{Err}}(1 + \frac{\mathbb{E}[\tilde{X}]^2}{\text{Var}(\tilde{X})})
\]

**Theorem 4.2.** If \(p(x) = p(x_1)p(x_{-1})\) and \(p(x_1) = p(-x_1)\), then for sufficiently small constant \(\alpha\) (that depends on the dataset), and for \(\text{Err} < \epsilon < \epsilon_0\),

\[
\frac{1}{4\text{Err}} < \text{DE}(\epsilon) < \frac{1}{2\text{Err}}
\]

**Proof.** If \(p(x_1) = p(-x_1)\), then \(p(\tilde{X}) = p(-\tilde{X})\) and so \(\mathbb{E}[\tilde{X}] = 0\).

Using Lemma 6.8, we arrive at the conclusion. \(\square\)