Consider the dataset $\{X_i, Y_i\}_{i=1}^n \sim^{iid} F_{\mu_0}$ where the distributions F_{μ_0} is specified by $Y \sim Bernoulli(1/2)$ and $X|Y \sim \mathcal{N}(\mu_0 + (-1)^y, 1)$. The maximum likelihood estimate for μ_0 is given by,

$$\hat{\mu_0} = \frac{1}{n}\sum_{i=1}^{n}(X_i - (-1)^{Y_i}) \tag{1}$$

Hence, the estimated decision rule would be $\hat{g}(X; \{X_i, Y_i\}_{i=1}^n) = \mathbb{I}\{X < \hat{\mu_0}\}$.

Now we have an additional m training examples $\{X_i, Y_i\}_{i=n+1}^{n+m} \sim^{iid} F_{\mu_{ood}}$ and we are estimating μ_0 using the combined dataset $\{X_i, Y_i\}_{i=1}^{n+m}$ using (1). This estimate is given by,

$$\hat{\mu_0}' = \frac{1}{n+m}\sum_{i=1}^{n+m}(X_i - (-1)^{Y_i}) \tag{2}$$

This yield the decision rule $\hat{g}(X; \{X_i, Y_i\}_{i=1}^{n+m}) = \mathbb{I}\{X < \hat{\mu_0}'\}$ We are interested in finding a $c(n)$ such that when $|\mu_0 - \mu_{ood}| < c(n)$ the performance of $\hat{g}(X; \{X_i, Y_i\}_{i=1}^{n+m})$ improves (does not degrade) and when $|\mu_0 - \mu_{ood}| > c(n)$ the performance of $\hat{g}(X; \{X_i, Y_i\}_{i=1}^{n+m})$ degrades.

Let's begin with (2),

$$\begin{aligned}
\hat{\mu_0}' &= \frac{1}{n+m}\sum_{i=1}^{n+m}(X_i - (-1)^{Y_i}) \\
\hat{\mu_0}' &= \frac{1}{n+m}\left[\sum_{i=1}^{n}(X_i - (-1)^{Y_i}) + \sum_{i=n+1}^{n+m}(X_i - (-1)^{Y_i})\right] \\
\hat{\mu_0}' &= \frac{1}{n+m}\left[n\hat{\mu_0} + m\hat{\mu}_{ood}\right] \quad \left(\because \hat{\mu}_{ood} = \frac{1}{m}\sum_{j=1}^{m}(X_j - (-1)^{Y_j}\right) \\
\hat{\mu_0}' &= \frac{n\hat{\mu_0} + m\hat{\mu}_{ood}}{n+m}
\end{aligned}$$

Since $\hat{\mu}_0 \sim \mathcal{N}(\mu_0, 1/n)$ and $\hat{\mu}_{ood} \sim \mathcal{N}(\mu_{ood}, 1/m)$, we notice

$$\mathbb{E}[\hat{\mu_0}'] = \frac{n\mu_0 + m\mu_{ood}}{n+m}$$

$$\text{var}[\hat{\mu_0}'] = \frac{1}{n+m}$$

Therefore,

$$\hat{\mu_0}' \sim \mathcal{N}\left(\frac{n\mu_0 + m\mu_{ood}}{n+m}, \frac{1}{n+m}\right)$$

In order for $\hat{g}(X; \{X_i, Y_i\}_{i=1}^{n+m})$ improve (or maintain) performance within a selected precision ϵ and tolerance δ,

$$\begin{aligned}
P(|\hat{\mu_0}' - \mu_0| < \epsilon) &> 1 - \delta \\
P(\mu_0 - \epsilon < \hat{\mu_0}' < \mu_0 + \epsilon) &> 1 - \delta \\
P\left(\frac{\mu_0 - \epsilon - \frac{n\mu_0 + m\mu_{ood}}{n+m}}{\frac{1}{\sqrt{n+m}}} < Z < \frac{\mu_0 + \epsilon - \frac{n\mu_0 + m\mu_{ood}}{n+m}}{\frac{1}{\sqrt{n+m}}}\right) &> 1 - \delta
\end{aligned}$$

Let $M > 0$ such that $P(-M < Z < M) = 1 - \delta$. Then, we have

$$\frac{\mu_0 - \epsilon - \frac{n\mu_0 + m\mu_{ood}}{n+m}}{\frac{1}{\sqrt{n+m}}} < -M$$

and

$$\frac{\mu_0 + \epsilon - \frac{n\mu_0 + m\mu_{ood}}{n+m}}{\frac{1}{\sqrt{n+m}}} > M$$

Simplifying the above, we arrive at

$$\mu_0 - \mu_{ood} < \frac{n+m}{m}\epsilon - \frac{\sqrt{n+m}}{m}M < \frac{n+m}{m}\epsilon$$

and

$$\mu_0 - \mu_{ood} > -\frac{n+m}{m}\epsilon + \frac{\sqrt{n+m}}{m}M < -\frac{n+m}{m}\epsilon$$

Therefore,

$$-\frac{n+m}{m}\epsilon < \mu_0 - \mu_{ood} < \frac{n+m}{m}\epsilon$$

$$|\mu_0 - \mu_{ood}| < \frac{n+m}{m}\epsilon = c(n)$$

$$\therefore c(n) = \frac{n+m}{m}\epsilon$$

Figure 1: These plots demonstrate the effectiveness of the bound $c(n)$. The red line and black line correspond to the Bayes optimal error and $c(n)$ value. The parameters used in the experiment are as follows: $\mu_0 = 0, m = 100, \epsilon = 0.3125$