Technical Appendix

Derivation of Bias Scan Score $F(S)$

To obtain the score function for a given subgroup S, Bias Scan computes the generalized log-likelihood ratio $F(S) = \frac{\max_{\widetilde{q}} P(D \mid H_1(S,\widetilde{q}))}{P(D \mid H_0)}$, assuming the following hypotheses:

$$H_0: \quad \text{odds}(y_i) = \frac{\widetilde{p}_i}{1-\widetilde{p}_i}, \ \forall s_i \in D.$$

$$H_1(S,\widetilde{q}): \quad \text{odds}(y_i) = \frac{\widetilde{q}\,\widetilde{p}_i}{1-\widetilde{p}_i}, \ \forall s_i \in D_S,$$

$$\text{odds}(y_i) = \frac{\widetilde{p}_i}{1-\widetilde{p}_i}, \ \forall s_i \in D \setminus D_S.$$

This implies that the probability $P(y_i) = \frac{\text{odds}(y_i)}{1+\text{odds}(y_i)} = \frac{\widetilde{q}\,\widetilde{p}_i}{1-\widetilde{p}_i+\widetilde{q}\,\widetilde{p}_i}$ for $s_i \in D_S$ under $H_1(S,\widetilde{q})$, and $P(y_i) = \widetilde{p}_i$ otherwise. We denote these probabilities by p_i^1 and p_i^0 respectively, and derive the score $F(S)$:

$$\begin{aligned}
F(S) &= \max_{\widetilde{q}} \log \frac{\prod_{s_i \in D_S} (p_i^1)^{y_i}(1-p_i^1)^{1-y_i}}{\prod_{s_i \in D_S} (p_i^0)^{y_i}(1-p_i^0)^{1-y_i}} \\
&= \max_{\widetilde{q}} \left(\sum_{s_i \in D_S} y_i \log \frac{p_i^1}{p_i^0} + \sum_{s_i \in D_S} (1-y_i) \log \frac{1-p_i^1}{1-p_i^0} \right) \\
&= \max_{\widetilde{q}} \left(\sum_{s_i \in D_S} y_i \log \frac{\widetilde{q}}{1-\widetilde{p}_i+\widetilde{q}\,\widetilde{p}_i} \right. \\
&\qquad \left. + \sum_{s_i \in D_S} (1-y_i) \log \frac{1}{1-\widetilde{p}_i+\widetilde{q}\,\widetilde{p}_i} \right) \\
&= \max_{\widetilde{q}} \left(\sum_{s_i \in D_S} y_i \log \widetilde{q} - \sum_{s_i \in D_S} \log(1-\widetilde{p}_i+\widetilde{q}\,\widetilde{p}_i) \right).
\end{aligned}$$

Here we focus on the case where the probabilities $\widetilde{p}_i$ are over-estimated, i.e., we identify biases where $\widetilde{p}_i > P(y_i)$ and thus $0 < \widetilde{q} < 1$. Thus we define $F(S)$ as above for $\widetilde{q}_{MLE} = \arg\max_{\widetilde{q}} \left(\sum_{s_i \in D_S} y_i \log \widetilde{q} - \sum_{s_i \in D_S} \log(1-\widetilde{p}_i+\widetilde{q}\,\widetilde{p}_i) \right) < 1$, and otherwise we have $\widetilde{q}_{MLE} = 1$ and thus $F(S) = 0$.

Bias Scan Algorithm

As described in Zhang and Neill (2016), Bias Scan detects intersectional subgroups for which the classifier's probabilistic predictions $\widetilde{p}_i$ are significantly biased as compared to the observed binary outcomes y_i, by searching for the rectangular subgroup $S \in rect(X)$ which maximizes the log-likelihood ratio score $F(S)$ derived above.

To optimize $F(S)$ over rectangular subgroups, Bias Scan performs a coordinate ascent procedure, optimizing $F(S)$ over subsets of values for one attribute at a time until convergence. This coordinate ascent procedure is repeated for multiple iterations, starting from a different, randomly initialized rectangular subgroup on each iteration. Bias Scan returns the maximum score $\hat{F}^* = \max_S F(S)$ and the corresponding subgroup $\hat{S}^* = \arg\max_S F(S)$ over all iterations.

The Bias Scan algorithm consists of the following steps:

1. Initialize $\hat{F}^* = 0$ and $\hat{S}^* = \emptyset$.
2. Choose an initial rectangular subgroup $S = S_1 \times \ldots \times S_Q$ by randomly selecting a subset of values $S_j \subseteq V_j$, $S_j \neq \emptyset$, for each attribute X_j. Mark all attributes as "unvisited".
3. Randomly select an unvisited attribute X_j and find $S_j' = \arg\max_{S_j \subseteq V_j, S_j \neq \emptyset} F(S_1 \times \ldots \times S_Q)$. Let $S' = S_1 \times \ldots \times S_{j-1} \times S_j' \times S_{j+1} \times \ldots \times S_Q$.
4. If $F(S') > F(S)$, then set $S = S'$ and mark all attributes as "unvisited".
5. If $F(S) > \hat{F}^*$, then set $\hat{F}^* = F(S)$ and $\hat{S}^* = S$.
6. Mark attribute X_j as "visited".
7. Repeat steps 3-6 until all attributes are marked as "visited".
8. Repeat steps 2-7 for a fixed, large number of iterations I.

The optimization over subsets of values $S_j \subseteq V_j$ in Step 3 can be performed efficiently, requiring a number of computations of the score function which is linear rather than exponential in the arity $|V_j|$ of that attribute, thanks to the Linear Time Subset Scanning (LTSS) property of the Bias Scan score function (Neill 2012). The statistical significance of detected subgroups can be obtained by randomization testing, performing the same search procedure on a large number of datasets randomly generated under the null hypothesis H_0, and then comparing the score $\hat{F}^*$ for the original data to the $(1-\alpha)$ quantile of the distribution of $\hat{F}^*$ for the null data. Further details are provided by Zhang and Neill (2016).

Proofs of Theorem 1 and Corollary 1

Theorem 1. *Assume that a classifier is trained on data $\widetilde{D}$ with differential sampling bias $\Delta > 1$ for subgroup S and makes predictions $\widetilde{p}_i$ for unbiased test data $D = \{(x_i, y_i)\}$. If Bias Scan is used to assess bias in $\widetilde{p}_i$ as compared to y_i, then under assumptions (A1)-(A3), as the number of training data records $|\widetilde{D}| \to \infty$, the Bias Scan score $F(S)$ of subgroup S converges to:*

$$F(S) \to F_{old}(S) - \sum_{s_i \in D_S} y_i \log \Delta + \sum_{s_i \in D_S} \log(\Delta p_i + 1 - p_i),$$

if $\Delta > \hat{q}_{MLE}$, and $F(S) \to 0$ otherwise, where $\hat{q}_{MLE}$ is the maximum likelihood estimate of $\widetilde{q}$ for Bias Scan assuming no differential sampling bias ($\Delta = 1$), satisfying

$$\sum_{s_i \in D_S} y_i = \sum_{s_i \in D_S} \frac{\hat{q}_{MLE}\, p_i}{\hat{q}_{MLE}\, p_i + 1 - p_i},$$

and

$$F_{old}(S) = \sum_{s_i \in D_S} y_i \log \hat{q}_{MLE} - \sum_{s_i \in D_S} \log(1 - p_i + \hat{q}_{MLE} p_i)$$

is the Bias Scan score of subgroup S assuming no differential sampling bias ($\Delta = 1$).

Proof. As $|\widetilde{D}| \to \infty$ without differential sampling bias, the number of training data records tends to infinity for each $x \in support(\widetilde{f}_X)$. The classification model is consistent by assumption (A1), and thus the estimated probability $\hat{p}_i$

converges to $\mathbb{P}(Y = 1|X = x) = p_i$ for all $x \in S$ for the training data. By assumption (A2), $\hat{p}_i \rightarrow \mathbb{P}(Y = 1|X = x) = p_i$ for all $x \in S$ for the test data, and the corresponding set of test data records D_S is non-empty.

Similarly, as $|\widetilde{D}| \rightarrow \infty$ with differential sampling bias Δ for subgroup S, the number of training data records tends to infinity for each $x \in support(\widetilde{f}_X)$. The classification model is consistent by assumption (A1), and thus the estimated probability $\widetilde{p}_i$ converges to $\widetilde{\mathbb{P}}(Y = 1|X = x) = \frac{\Delta\mathbb{P}(Y=1|X=x)}{\Delta\mathbb{P}(Y=1|X=x)+\mathbb{P}(Y=0|X=x)} = \frac{\Delta p_i}{\Delta p_i+1-p_i}$ for all $x \in S$ for the training data. By assumption (A2), $\widetilde{p}_i \rightarrow \widetilde{\mathbb{P}}(Y = 1|X = x) = \frac{\Delta p_i}{\Delta p_i+1-p_i}$ for all $x \in S$ for the test data, and D_S is non-empty.

Next, we derive the relationship between the maximum likelihood estimate of the $\widetilde{q}$ parameter for Bias Scan, with and without differential sampling bias. We define:

$$\widetilde{q}_{MLE} = \arg\max_{\widetilde{q}} \left(\sum_{s_i \in D_S} y_i \log \widetilde{q} - \sum_{s_i \in D_S} \log(1 - \widetilde{p}_i + \widetilde{q}\,\widetilde{p}_i) \right),$$

$$\hat{q}_{MLE} = \arg\max_{\hat{q}} \left(\sum_{s_i \in D_S} y_i \log \hat{q} - \sum_{s_i \in D_S} \log(1 - \hat{p}_i + \hat{q}\,\hat{p}_i) \right).$$

By setting $\frac{dF(S)}{d\widetilde{q}} = 0$ and $\frac{dF(S)}{d\hat{q}} = 0$ for the cases with and without differential sampling bias respectively, we obtain:

$$\sum_{s_i \in D_S} \frac{y_i}{\widetilde{q}_{MLE}} = \sum_{s_i \in D_S} \frac{\widetilde{p}_i}{1 - \widetilde{p}_i + \widetilde{q}_{MLE}\,\widetilde{p}_i},$$

$$\sum_{s_i \in D_S} \frac{y_i}{\hat{q}_{MLE}} = \sum_{s_i \in D_S} \frac{\hat{p}_i}{1 - \hat{p}_i + \hat{q}_{MLE}\,\hat{p}_i},$$

and thus,

$$\sum_{s_i \in D_S} \frac{\widetilde{q}_{MLE}\,\widetilde{p}_i}{1 - \widetilde{p}_i + \widetilde{q}_{MLE}\,\widetilde{p}_i} = \sum_{s_i \in D_S} \frac{\hat{q}_{MLE}\,\hat{p}_i}{1 - \hat{p}_i + \hat{q}_{MLE}\,\hat{p}_i}.$$

Plugging in the values of $\hat{p}_i = p_i$ and $\widetilde{p}_i = \frac{\Delta p_i}{\Delta p_i+1-p_i}$ from above, and simplifying, we obtain:

$$\sum_{s_i \in D_S} \frac{\widetilde{q}_{MLE}\,\Delta p_i}{1 - p_i + \widetilde{q}_{MLE}\,\Delta p_i} = \sum_{s_i \in D_S} \frac{\hat{q}_{MLE}\,p_i}{1 - p_i + \hat{q}_{MLE}\,p_i},$$

and thus,

$$\widetilde{q}_{MLE} = \frac{\hat{q}_{MLE}}{\Delta}.$$

We can now derive the Bias Scan score without differential sampling bias as:

$$\begin{aligned} F_{old}(S) &= \sum_{s_i \in D_S} y_i \log \hat{q}_{MLE} - \sum_{s_i \in D_S} \log(1 - \hat{p}_i + \hat{q}_{MLE}\,\hat{p}_i) \\ &= \sum_{s_i \in D_S} y_i \log \hat{q}_{MLE} - \sum_{s_i \in D_S} \log(1 - p_i + \hat{q}_{MLE}\,p_i). \end{aligned}$$

Note that F_{old} is defined without enforcing the constraint $\widetilde{q} < 1$. Finally, we derive the Bias Scan score with differential sampling bias (for detecting over-estimated probabilities) as:

$$\begin{aligned} F(S) &= \sum_{s_i \in D_S} y_i \log \widetilde{q}_{MLE} - \sum_{s_i \in D_S} \log(1 - \widetilde{p}_i + \widetilde{q}_{MLE}\,\widetilde{p}_i) \\ &= \sum_{s_i \in D_S} y_i \log \frac{\hat{q}_{MLE}}{\Delta} - \sum_{s_i \in D_S} \log\left(1 + \left(\frac{\hat{q}_{MLE}}{\Delta} - 1\right)\widetilde{p}_i\right) \\ &= \sum_{s_i \in D_S} y_i \log \frac{\hat{q}_{MLE}}{\Delta} - \sum_{s_i \in D_S} \log\left(1 + \frac{(\hat{q}_{MLE} - \Delta)p_i}{\Delta p_i + 1 - p_i}\right) \\ &= \sum_{s_i \in D_S} y_i \log \frac{\hat{q}_{MLE}}{\Delta} - \sum_{s_i \in D_S} \log\left(\frac{1 - p_i + \hat{q}_{MLE}p_i}{\Delta p_i + 1 - p_i}\right), \\ &= F_{old}(S) - \sum_{s_i \in D_S} y_i \log \Delta + \sum_{s_i \in D_S} \log(\Delta p_i + 1 - p_i) \end{aligned}$$

if $\Delta > \hat{q}_{MLE}$ (and thus $\widetilde{q}_{MLE} = \hat{q}_{MLE}/\Delta < 1$), and otherwise we have $\widetilde{q}_{MLE} = 1$ and thus $F(S) = 0$. □

Corollary 1. *Under the conditions of Theorem 1, as the number of test data records $|D| \rightarrow \infty$, the normalized Bias Scan score $F(S)/|D|$ of subgroup S converges to:*

$$\frac{F(S)}{|D|} \rightarrow \mathbb{P}(x \in S)\mathbb{E}_{s_i \in D_S}[\log(\Delta p_i + 1 - p_i) - p_i \log \Delta],$$

an increasing function of Δ.

Proof. From Theorem 1, we have

$$F(S) \rightarrow F_{old}(S) - \sum_{s_i \in D_S} y_i \log \Delta + \sum_{s_i \in D_S} \log(\Delta p_i + 1 - p_i),$$

if $\Delta > \hat{q}_{MLE}$. As $|D| \rightarrow \infty$, $\hat{q}_{MLE} \rightarrow 1$, and thus we have both w.h.p. $\Delta > \hat{q}_{MLE}$ and $F_{old}(S) \rightarrow 0$:

$$F(S) \rightarrow \sum_{s_i \in D_S} (\log(\Delta p_i + 1 - p_i) - y_i \log \Delta),$$

and

$$\frac{F(S)}{|D|} \rightarrow \frac{|D_S|}{|D|}\mathbb{E}_{s_i \in D_S}[\log(\Delta p_i + 1 - p_i) - y_i \log \Delta].$$

As $|D| \rightarrow \infty$, $|D_S|/|D| \rightarrow \mathbb{P}(x \in S)$, and $\mathbb{E}[y_i] = \mathbb{E}[\mathbb{E}[y_i \mid x_i]] = \mathbb{E}[p_i]$ for $s_i \in D_S$. Plugging in these values, we obtain the given expression. To see that the expression increases with Δ, assumption (A2) implies $\mathbb{P}(x \in S) > 0$, and the first derivative

$$\frac{d(\log(\Delta p_i + 1 - p_i) - p_i \log \Delta)}{d\Delta} = \frac{p_i}{\Delta p_i + 1 - p_i} - \frac{p_i}{\Delta}$$

is positive for $\Delta > 1$, given $0 < p_i < 1$ by assumption (A3). □

Proof of Theorem 2 (and associated Lemmas)

In this section, we derive a critical value $h(\alpha)$ of the Bias scan score F^*, for a given Type-I error rate α, when no differential sampling bias is present. Our approach is to upper bound $F^* = \max_{S \in rect(X)} F(S)$ by the Bias Scan score maximized over *all* subgroups, $F^*_u = F(S^*_u)$, where $S^*_u = \arg\max_{S \subseteq V} F(S)$. Additionally, as the number of training data records $|\widetilde{D}| \rightarrow \infty$, with no differential sampling bias, the classifier's predictions $\hat{p}_i$ converge to p_i for all test records s_i, under assumptions (A1) and (A2), as in Theorem 1. Thus we derive the distribution of F^*_u under the null hypothesis, $H_0 : \mathbb{P}(y_i = 1) = p_i$ for all $s_i \in D$.

For any covariate profile x, we define the associated set of test data records $D_x = \{(x_i, y_i)\} \subseteq D : x_i = x$, and the aggregate quantities $y(x) = \sum_{s_i \in D_x} y_i$ and $n(x) = \sum_{s_i \in D_x} 1$. We assume that the predicted probabilities p_i are identical for all $s_i \in D_x$, since these data records have the same values for all predictor variables, and denote this probability by $p(x)$. We can then write the Bias Scan score of a subgroup $F(S)$ as

$$F(S) = \max_{0<\widetilde{q}<1} \sum_{x \in S} \gamma_x(\widetilde{q}),$$

where

$$\gamma_x(\widetilde{q}) = y(x) \log(\widetilde{q}) - n(x) \log(\widetilde{q}p(x) + 1 - p(x))$$

is the total contribution of data records with covariate profile x to the score of subgroup S for a given value of $\widetilde{q}$. Then the maximum score over all subgroups can be written as

$$\begin{aligned} F_u^* &= \max_{S \subseteq V} \max_{0<\widetilde{q}<1} \sum_{x \in S} \gamma_x(\widetilde{q}) \\ &= \max_{0<\widetilde{q}<1} \max_{S \subseteq V} \sum_{x \in S} \gamma_x(\widetilde{q}) \\ &= \max_{0<\widetilde{q}<1} \sum_{x \in V} \gamma_x(\widetilde{q}) \mathbf{1}\{\gamma_x(\widetilde{q}) > 0\}, \end{aligned}$$

thus including all and only those covariate profiles which make a positive contribution to the score for the given value of $\widetilde{q}^*_{mle}$. Given these definitions, we now consider the probability that a given covariate profile x will have $\gamma_x(\widetilde{q}^*_{mle}) > 0$, and thus be included in S_u^*:

Lemma 1. *Under the null hypothesis H_0, as $n(x) \to \infty$, the probability that covariate profile x is included in the highest scoring subgroup S_u^* converges to $1 - \Phi\left(\frac{Z}{2}\sqrt{p(x)(1-p(x))}\right)$, where $\widetilde{q}^*_{mle} = 1 - \frac{Z}{\sqrt{n(x)}}$, and Φ is the Gaussian cdf.*

Proof. Covariate profile x is included in S_u^* if and only if $\gamma_x(\widetilde{q}^*_{mle}) = y(x) \log(\widetilde{q}^*_{mle}) - n(x) \log(\widetilde{q}^*_{mle}p(x) + 1 - p(x)) > 0$. Given $0 < \widetilde{q}^*_{mle} < 1$, we have:

$$\begin{aligned} &\mathbb{P}(\gamma_x(\widetilde{q}^*_{mle}) > 0) \\ &= \mathbb{P}\left(\frac{y(x)}{n(x)} < \frac{\log(\widetilde{q}^*_{mle}p(x) + 1 - p(x))}{\log \widetilde{q}^*_{mle}}\right) \\ &\to \mathbb{P}\left(\frac{y(x)}{n(x)} < p(x) - \frac{(1-\widetilde{q}^*_{mle})p(x)(1-p(x))}{2}\right) \\ &= \mathbb{P}\left(\psi(x) < -\frac{Z}{2}\sqrt{p(x)(1-p(x))}\right), \end{aligned}$$

where $\psi(x) = \sqrt{\frac{n(x)}{p(x)(1-p(x))}}\left(\frac{y(x)}{n(x)} - p(x)\right)$. Here we have used a second order Taylor expansion for $\widetilde{q}^*_{mle}$, since $\widetilde{q}^*_{mle}$ converges to 1 as $n(x) \to \infty$. Next, since $y(x) \sim \text{Binomial}(n(x), p(x))$ under H_0, $\psi(x) \to \text{Gaussian}(0,1)$ as $n(x) \to \infty$, and $\mathbb{P}(\gamma_x(\widetilde{q}^*_{mle}) > 0)$ converges to $1 - \Phi\left(\frac{Z}{2}\sqrt{p(x)(1-p(x))}\right)$, where Φ is the Gaussian cdf. □

Lemma 2. *Under the null hypothesis H_0, as $n(x) \to \infty$, the expectation and variance of $\gamma_x(\widetilde{q}^*_{mle})\mathbf{1}\{\gamma_x(\widetilde{q}^*_{mle}) > 0\}$ are upper bounded by constants $k_1 \approx 0.202$ and $k_2^2 \approx 0.274$ respectively.*

Proof. From Lemma 1, as $n(x) \to \infty$, $\psi(x) = \sqrt{\frac{n(x)}{p(x)(1-p(x))}}\left(\frac{y(x)}{n(x)} - p(x)\right) \to \text{Gaussian}(0,1)$. Moreover, conditional on $\gamma_x(\widetilde{q}^*_{mle}) > 0$, $\psi(x)$ has its right tail truncated at $\beta = -\frac{Z}{2}\sqrt{p(x)(1-p(x))}$, giving $\mathbb{E}[\psi(x) \mid \gamma_x(\widetilde{q}^*_{mle}) > 0] = -h(-\beta)$, and $\text{Var}[\psi(x) \mid \gamma_x(\widetilde{q}^*_{mle}) > 0] = 1 - \beta h(-\beta) - h(-\beta)^2$, where $h(x) = \frac{\phi(x)}{1-\Phi(x)}$ is the Gaussian hazard function. Since $y(x) = n(x)p(x) + \psi(x)\sqrt{n(x)p(x)(1-p(x))}$, this implies:

$$\begin{aligned} &\mathbb{E}\left[\frac{y(x) - n(x)p(x)}{\sqrt{n(x)}} \mid \gamma_x(\widetilde{q}^*_{mle}) > 0\right] \\ &= -h(-\beta)\sqrt{p(x)(1-p(x))}; \\ &\text{Var}\left[\frac{y(x) - n(x)p(x)}{\sqrt{n(x)}} \mid \gamma_x(\widetilde{q}^*_{mle}) > 0\right] \\ &= (1 - \beta h(-\beta) - h(-\beta)^2)p(x)(1-p(x)). \end{aligned}$$

Next, as in Lemma 1, we can use a second-order Taylor expansion to write:

$$\begin{aligned} &\gamma_x(\widetilde{q}^*_{mle}) \\ &= y(x) \log \widetilde{q}^*_{mle} - n(x) \log(\widetilde{q}^*_{mle}p(x) + 1 - p(x)) \\ &= y(x) \log \widetilde{q}^*_{mle} - n(x) \log \widetilde{q}^*_{mle} \frac{\log(\widetilde{q}^*_{mle}p(x) + 1 - p(x))}{\log \widetilde{q}^*_{mle}} \\ &\to (y(x) - n(x)p(x)) \log \widetilde{q}^*_{mle} \\ &\quad + \frac{(1-\widetilde{q}^*_{mle})(\log \widetilde{q}^*_{mle})n(x)p(x)(1-p(x))}{2} \\ &\to (y(x) - n(x)p(x))\left(-\frac{Z}{\sqrt{n(x)}}\right) \\ &\quad + \frac{1}{2}\left(\frac{Z}{\sqrt{n(x)}}\right)\left(-\frac{Z}{\sqrt{n(x)}}\right) n(x)p(x)(1-p(x)) \\ &= -Z\left(\frac{y(x) - n(x)p(x)}{\sqrt{n(x)}}\right) - \frac{Z^2}{2}p(x)(1-p(x)), \end{aligned}$$

where we have used $\log \widetilde{q}^*_{mle} \to -(1 - \widetilde{q}^*_{mle}) = -\frac{Z}{\sqrt{n(x)}}$ as $n(x) \to \infty$.

Then for the expectation we have:

$$\begin{aligned} &\mathbb{E}[\gamma_x(\widetilde{q}^*_{mle}) \mid \gamma_x(\widetilde{q}^*_{mle}) > 0] \\ &\to -Z\mathbb{E}\left[\frac{y(x) - n(x)p(x)}{\sqrt{n(x)}}\right] - \frac{Z^2}{2}p(x)(1-p(x)) \\ &= Zh(-\beta)\sqrt{p(x)(1-p(x))} - \frac{Z^2}{2}p(x)(1-p(x)) \\ &= -2\beta h(-\beta) - 2\beta^2. \end{aligned}$$

Then, since $\mathbb{P}(\gamma_x(\widetilde{q}^*_{mle}) > 0) = 1 - \Phi(-\beta)$ from Lemma 1, we know:

$$\begin{aligned}
&\mathbb{E}[\gamma_x(\widetilde{q}^*_{mle})\mathbf{1}\{\gamma_x(\widetilde{q}^*_{mle}) > 0\}] \\
&\to (1 - \Phi(-\beta))(-2\beta h(-\beta) - 2\beta^2) \\
&= -2\beta\phi(-\beta) - 2\beta^2(1 - \Phi(-\beta)) \\
&\leq k_1,
\end{aligned}$$

since this expression attains a maximum value of $k_1 \approx 0.202456$ at $\beta \approx -0.61$.
For the variance, we have:

$$\begin{aligned}
&\text{Var}[\gamma_x(\widetilde{q}^*_{mle}) \mid \gamma_x(\widetilde{q}^*_{mle}) > 0] \\
&\to Z^2\text{Var}\left[\frac{y(x) - n(x)p(x)}{\sqrt{n(x)}}\right] \\
&= Z^2(1 - \beta h(-\beta) - h(-\beta)^2)p(x)(1 - p(x)) \\
&= 4\beta^2(1 - \beta h(-\beta) - h(-\beta)^2).
\end{aligned}$$

Then, we know:

$$\begin{aligned}
&\text{Var}[\gamma_x(\widetilde{q}^*_{mle})\mathbf{1}\{\gamma_x(\widetilde{q}^*_{mle}) > 0\}] \\
&= \text{Var}[\gamma_x(\widetilde{q}^*_{mle}) \mid \gamma_x(\widetilde{q}^*_{mle}) > 0]\mathbb{P}(\gamma_x(\widetilde{q}^*_{mle}) > 0) + \\
&\quad \mathbb{E}[\gamma_x(\widetilde{q}^*_{mle}) \mid \gamma_x(\widetilde{q}^*_{mle}) > 0]^2\mathbb{P}(\gamma_x(\widetilde{q}^*_{mle}) > 0) \\
&\quad (1 - \mathbb{P}(\gamma_x(\widetilde{q}^*_{mle}) > 0)) \\
&\to 4\beta^2(1 - \beta h(-\beta) - h(-\beta)^2)(1 - \Phi(-\beta)) + \\
&\quad (-2\beta h(-\beta) - 2\beta^2)^2(1 - \Phi(-\beta))\Phi(-\beta) \\
&= 4\beta^2(1 - \Phi(-\beta))(1 - \beta h(-\beta) - h(-\beta)^2 + \\
&\quad (\beta + h(-\beta))^2\Phi(-\beta)) \\
&\leq k_2^2,
\end{aligned}$$

since this expression attains a maximum value of $k_2^2 \approx 0.273709$ at $\beta \approx -0.98$.

□

Theorem 2. *Assume that a classifier is trained on unbiased training data $\widetilde{D}$ and makes predictions $\hat{p}_i$ for unbiased test data $D = \{(x_i, y_i)\}$, and Bias Scan is used to assess bias in $\hat{p}_i$ as compared to y_i. Let $F^* = \max_{S \in rect(X)} F(S)$ be the Bias Scan score, maximized over all rectangular subgroups S. Then under assumptions (A1)-(A4), as the number of training data records $|\widetilde{D}| \to \infty$ and the number of test data records $|D| \to \infty$, for a given Type-I error rate $\alpha > 0$, there exists a critical value $h(\alpha)$ and constants $k_1 \approx 0.202$, $k_2 \approx 0.523$ such that*

$$\mathbb{P}(F^* > h(\alpha)) \leq \alpha,$$

where

$$h(\alpha) = k_1 M + k_2 \Phi^{-1}(1 - \alpha)\sqrt{M}, \tag{3}$$

and Φ is the Gaussian cdf.

Proof. As $|\widetilde{D}| \to \infty$ without differential sampling bias, the number of training data records tends to infinity for each $x \in support(\widetilde{f}_X)$. The classification model is consistent by assumption (A1), and thus the estimated probability $\hat{p}_i$ converges to $\mathbb{P}(Y = 1|X = x) = p_i$ for all $x \in S$ for the training data. By assumption (A2), $\hat{p}_i \to \mathbb{P}(Y = 1|X = x) = p_i$ for all $x \in S$ for the test data, and the corresponding set of test data records D_S is non-empty. As shown above, $F^* \leq F^*_u = \sum_{x \in V} \gamma_x(\widetilde{q}^*_{mle})\mathbf{1}\{\gamma_x(\widetilde{q}^*_{mle}) > 0\}$, where $\widetilde{q}^*_{mle} = \arg\max_{0<\widetilde{q}<1} \sum_{x \in V} \gamma_x(\widetilde{q})\mathbf{1}\{\gamma_x(\widetilde{q}) > 0\}$ and $\gamma_x(\widetilde{q}) = y(x)\log\widetilde{q} - n(x)\log(\widetilde{q}p(x) + 1 - p(x))$. From Lemma 2, for each of the M unique covariate profiles x in the test data, we know that $\gamma_x(\widetilde{q}^*_{mle})\mathbf{1}\{\gamma_x(\widetilde{q}^*_{mle}) > 0\}$ is drawn from a censored Gaussian distribution, with mean $\mu_x \leq k_1$ and variance $\sigma_x^2 \leq k_2^2$, where $k_1 \approx 0.202$ and $k_2 \approx \sqrt{0.274} \approx 0.523$. Moreover, since a censored Gaussian with bounded variance has bounded fourth moment, we know that the Lyapunov condition holds. Thus, from the Lyapunov CLT, we know that for large M, $\frac{F^*_u - \sum_{x \in V} \mu_x}{\sqrt{\sum_{x \in V} \sigma_x^2}} \to \text{Gaussian}(0, 1)$, and by assumption (4) we know that M is large enough for F^*_u to be approximately Gaussian. Then since $F^* \leq F^*_u$, $\mu_x \leq k_1\ \forall x$, and $\sigma_x^2 \leq k_2^2\ \forall x$, we have:

$$\begin{aligned}
&\mathbb{P}(F^* > h(\alpha)) \\
&= \mathbb{P}(F^* > k_1 M + k_2\Phi^{-1}(1 - \alpha)\sqrt{M}) \\
&\leq \mathbb{P}(F^*_u > k_1 M + k_2\Phi^{-1}(1 - \alpha)\sqrt{M}) \\
&= \mathbb{P}\left(F^*_u > \left(\sum_{x \in V} k_1\right) + \Phi^{-1}(1 - \alpha)\sqrt{\sum_{x \in V} k_2^2}\right) \\
&\leq \mathbb{P}\left(F^*_u > \left(\sum_{x \in V} \mu_x\right) + \Phi^{-1}(1 - \alpha)\sqrt{\sum_{x \in V} \sigma_x^2}\right) \\
&= 1 - \Phi(\Phi^{-1}(1 - \alpha)) \\
&= \alpha.
\end{aligned}$$

□

Proofs of Theorems 3 and 4

Theorem 3. *Assume that a classifier is trained on data $\widetilde{D}$ with differential sampling bias $\Delta > 1$ for rectangular subgroup S^T and makes predictions $\widetilde{p}_i$ for unbiased test data $D = \{(x_i, y_i)\}$, and Bias Scan is used to assess bias in $\widetilde{p}_i$ as compared to y_i. Let $F^* = \max_{S \in rect(X)} F(S)$ be the Bias Scan score, and let $h(\alpha)$ be the score threshold for detection at a fixed Type-I error rate of α, as given in Equation (3). Then for any $\alpha > 0$ and $\Delta > 1$, under assumptions (A1)-(A4), as the number of training data records $|\widetilde{D}| \to \infty$ and the number of test data records $|D| \to \infty$, $\mathbb{P}(F^* > h(\alpha)) \to 1$.*

Proof. By Corollary 1, as $|D| \to \infty$, $F(S^T)/|D|$ converges to $\mathbb{P}(x \in S^T)\mathbb{E}_{s_i \in D_{S^T}}[\log(\Delta p_i + 1 - p_i) - p_i \log\Delta]$, which is greater than zero because $\mathbb{P}(x \in S^T) > 0$ by assumption (A2), $0 < p_i < 1$ by assumption (A3), and $\log(\Delta p_i + 1 - p_i) - p_i \log\Delta > 0$ when $\Delta > 1$ and $0 < p_i < 1$. By Theorem 2, as $|D| \to \infty$, for any $\alpha > 0$, $h(\alpha)$ converges to a constant independent of $|D|$. Thus $h(\alpha)/|D| \to 0$ and $\mathbb{P}(F(S^T) > h(\alpha)) \to 1$. Finally, since subgroup S^T is rectangular, $F^* = \max_{S \in rect(X)} F(S) \geq F(S^T)$, and $\mathbb{P}(F^* > h(\alpha)) \to 1$. □

Theorem 4. *Assume that a classifier is trained on data $\widetilde{D}$ with differential sampling bias $\Delta > 1$ for rectangular subgroup S^T and makes predictions $\widetilde{p}_i$ for unbiased test data $D = \{(x_i, y_i)\}$, and Bias Scan is used to assess bias in $\widetilde{p}_i$ as compared to y_i. Let $F^* = \max_{S \in rect(X)} F(S)$ be the Bias Scan score, and let $h(\alpha)$ be the score threshold for detection at a fixed Type-I error rate of α, as given in Equation (3). Further, assume D_{S^T} is fixed, with finite size $|D_{S^T}|$ and $\left(\sum_{s_i \in D_{S^T}} y_i\right) < |D_{S^T}|$. Then for any $\alpha > 0$, under assumptions (A1)-(A4), as the number of training data records $|\widetilde{D}| \to \infty$, there exists $\Delta_{thresh} \geq 1$ such that, if $\Delta > \Delta_{thresh}$, then $\mathbb{P}(F^* > h(\alpha)) \to 1$, where*

$$\Delta_{thresh} = \max(1, Q^{-1}(h(\alpha) - F_{old}(S^T))),$$

$$Q(\Delta) = \sum_{s_i \in D_{S^T}} (\log(\Delta p_i + 1 - p_i) - y_i \log \Delta),$$

and $F_{old}(S^T)$ is the Bias Scan score of subgroup S^T assuming no differential sampling bias ($\Delta = 1$).

Proof. From Theorem 1, for finite $|D_{S^T}|$, we have $F(S^T) \to F_{old}(S^T) + Q(\Delta)$ for $\Delta > \hat{q}_{MLE}$ and $|\widetilde{D}| \to \infty$. We derive:

$$\begin{aligned}\frac{dQ}{d\Delta} &= \sum_{s_i \in D_{S^T}} \left(\frac{p_i}{\Delta p_i + 1 - p_i} - \frac{y_i}{\Delta} \right) \\ &= \frac{1}{\Delta} \sum_{s_i \in D_{S^T}} (\widetilde{p}_i - y_i).\end{aligned}$$

Since $0 < p_i < 1$ by assumption (A3), all $\widetilde{p}_i$ are increasing with Δ. Moreover, since $\sum_{s_i \in D_{S^T}} (\widetilde{p}_i - y_i) = 0$ for $\Delta = \hat{q}_{MLE}$, $\sum_{s_i \in D_{S^T}} (\widetilde{p}_i - y_i) > 0$ for all $\Delta > \hat{q}_{MLE}$. This implies that $Q(\Delta)$ is increasing, and therefore invertible, on the interval $\Delta \geq \hat{q}_{MLE}$.

Next we show $Q(\Delta) \to \infty$ as $\Delta \to \infty$. For some small positive $\epsilon \approx 0$, let Δ_ϵ denote the minimum value of $\Delta > \hat{q}_{MLE}$ such that $\sum_{s_i \in D_{S^T}} (\widetilde{p}_i - y_i) \geq \epsilon$. Then for any $\Delta' > \Delta_\epsilon$, we have:

$$\begin{aligned}Q(\Delta') &= Q(\Delta_\epsilon) + \int_{\Delta_\epsilon}^{\Delta'} \frac{dQ}{d\Delta}\, d\Delta \\ &\geq Q(\Delta_\epsilon) + \int_{\Delta_\epsilon}^{\Delta'} \frac{\epsilon}{\Delta}\, d\Delta \\ &= Q(\Delta_\epsilon) + \epsilon(\log \Delta' - \log \Delta_\epsilon) \\ &= C_1 \log \Delta' + C_0\end{aligned}$$

for constants C_1 and C_0, and thus $Q(\Delta)$ increases as $o(\log \Delta)$ for $\Delta \geq \hat{q}_{MLE}$.

Now, since $F_{old}(S^T)$ is independent of Δ, we know that $F_{old}(S^T) + Q(\Delta)$ is continuous and increasing for $\Delta \geq \hat{q}_{MLE}$, and $\lim_{\Delta \to \infty} F_{old}(S^T) + Q(\Delta) = \infty$. Since $F_{old}(S^T) + Q(\Delta) = 0$ at $\Delta = \hat{q}_{MLE}$, there must exist a single intermediate value of $\Delta > \hat{q}_{MLE}$ such that $F_{old}(S^T) + Q(\Delta) = h(\alpha)$, i.e., $\Delta = Q^{-1}(h(\alpha) - F_{old}(S^T))$. Then we set $\Delta_{thresh} = \max(1, Q^{-1}(h(\alpha) - F_{old}(S^T)))$. This implies that $F(S^T) \to F_{old}(S^T) + Q(\Delta) > h(\alpha)$, and $\mathbb{P}(F(S^T) > h(\alpha)) \to 1$, for $\Delta > \Delta_{thresh}$. Finally, assuming that subgroup S^T is rectangular, $F^* = \max_{S \in rect(X)} F(S) \geq F(S^T)$, and $\mathbb{P}(F^* > h(\alpha)) \to 1$ for $\Delta > \Delta_{thresh}$. □