Supplemental Materials

1 More Discussions on the Empirical Results

In this section, we present more figures of the thresholds learned by `NeuralFDR` and IHW.

Fig. 5 (a-c) shows the alternative proportion, NeuralFDR's learned threshold, and IHW's learned threshold for the 2D GM simulated data in Sec. 4. We can see the alternative proportion is well recovered by NeuralFDR. To some extent, IHW also recovers the structure but not with high resolution because its threshold is limited to have a constant threshold for each group. This causes a loss in resolution in informative directions.

Fig. 5 (e,d) shows the learned threshold for the GTEx-2D experiment, where we recall for this experiment, the features distance (GTEx-dist) and expression level (GTEx-exp) are used. We can see that `NeuralFDR` captures the structure that the alternative proportion is large when the distance is small and when the expression level is small. This matches the biological explanation as illustrated in Sec 4. However, IHW does not capture such structure very well.

Fig. 5 (f) shows the learned threshold for the GTEx-PhastCons experiment. The threshold is higher for more conserved regions but the difference is not very significant, showing that this covariate contains less information than distance (GTEx-dist) and expression (GTEx-exp). This is consistent with the observation that both IHW and `NeuralFDR` make fewer discoveries with PhastCons score than with distance or expression.

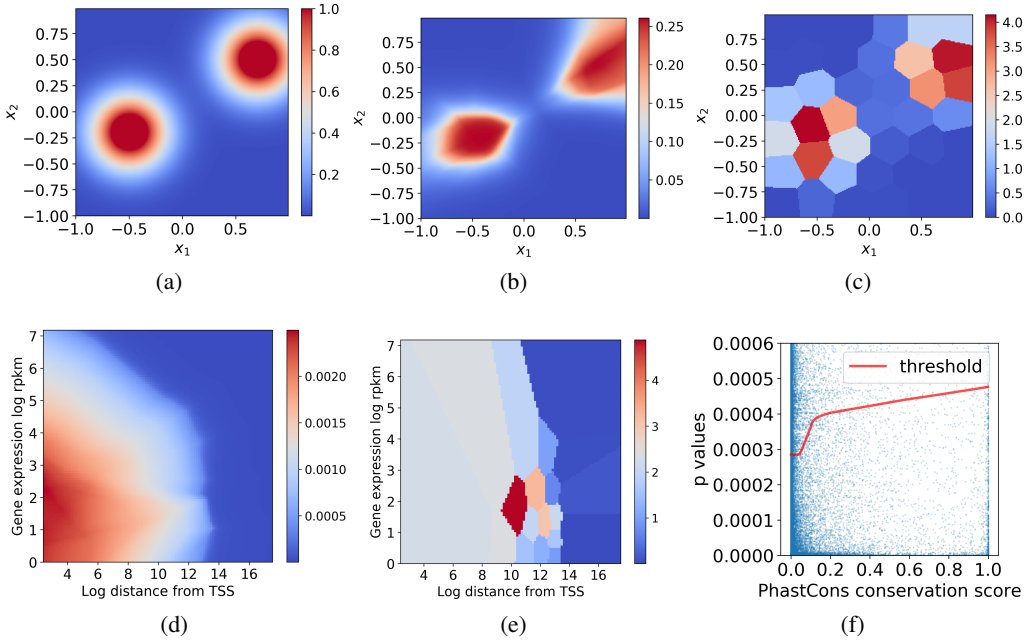

Figure 5: **(a-c)** Results for 2DGM: (a) the alternative proportion for 2Dslope; (b) NeuralFDR's learned threshold; (c) IHW's learned weights. **(d-e)** Results for GTEx-2D: (d) NeuralFDR's learned threshold;(e) IHW's learned weights. (f) NeuralFDR's learned threshold for GTEx-PhastCons.

As `NeuralFDR` uses neural network to do functional estimation, it has some randomness across mutiple runs. For example, the network could converge to bad local minimal. However, we show that `NeuralFDR` is stable across multiple runs. Fig. 1 shows the number of discoveries in Airway dataset in 10 parallel runs for each nominal FDR. The errorbar denotes standard deviation, i.e. 68.3% confidence interval. The coefficient of variation (CV) for each nominal FDR is smaller than 1% across experiments.

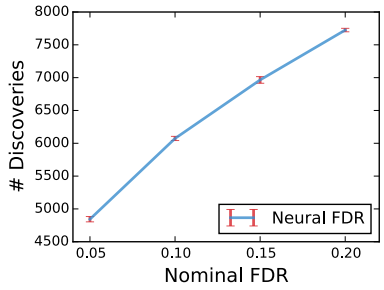

Figure 6: Resutls of parallel runs for airway dataset, and it demonstrates the variation aross runs is small

2 Implementation and Training Details

Objective function. We solve the constrained optimization problem 3 by the penalty method. We solve this optimization problem:

$$\text{maximize}_{\boldsymbol{\theta}} \sum_i \tilde{D}(t(\boldsymbol{\theta})) - \lambda_1 \max\left\{\widetilde{FD}(\gamma_i t_i^*(\boldsymbol{\theta})) - \alpha\tilde{D}(t(\boldsymbol{\theta})), 0\right\}. \tag{8}$$

To avoid using step function, we used sigmoid to approximate the counting. Denote the sigmoid function as σ. We define $\tilde{D}$ and $\widetilde{FD}$ to be the following.

$$\tilde{D}(t((\theta))) = \sum_j \sigma(\lambda_2(t((\theta); x_j) - p_j)) \tag{9}$$

$$\widetilde{FD}(t((\theta))) = \sum_j \sigma(\lambda_2(p_j - (1 - t((\theta); x_j)))) \tag{10}$$

In cross validation process, we don't use approximated version of D and FD. We use the actual number of points below the threshold and above the mirrored threshold as D and FD.

Initialization. As the optimization problem is highly non-convex, a good initialization is crucial for training. We used a smoothed version of k-mean clustering for initialization. The data is clustered into k clusters using k-mean clustering based on the hypothesis features. An optimal threshold for each cluster $t_{opt,k}$ is calculated following Theorem 2. For each hypothesis, the initial value of the threshold is set to be

$$t_{init,i} = \sum_{j=1}^{k} \frac{\exp(-\lambda_3||x_i - c_j||^2)}{\sum_{r=1}^{n}\exp(-\lambda_3||x_i - c_r||^2)} t_{opt,j} \tag{11}$$

where c_j is the center for cluster j.

Network architecture. We used a 10-layer of MLP, each layer has 10 nodes. For activation function, we used LeakyReLU with a slope of 0.2. In the output layer, we use a scaled version of Sigmoid function to make sure the output is in $(0, 0.5)$.

Implementation and Training. The algorithm is implemented in Python and the MLP is implemented using PyTorch. The optimization is solved using adaptive stochastic gradient method Adagrad [7].

For all the experiments, we split the data equally into $M = 3$ folds for cross validation. The learning rate is set to be 0.01. Because the optimization is driven by density, we use a large batch size of 10000. Penalty parameter λ_1 is set to 20, λ_2 is adaptively set depending on the BH threshold for a certain dataset, λ_3 is set to be 1. All hyper-parameters are not heavily tuned and work across datasets. Training to fitting converges at around 6000 iterations and for optimizing the number of discoveries converges at around 12000 iterations. The training is done on Nvidia Tesla K80 GPUs.

Notes for GTEx dataset. For GTEx dataset, the whole dataset is very large, so we filtered the p-values to get only hypothesis with $p < 0.005$ or $p > 0.995$, where the second part is for mirroring estimation. We also scale the network output to operate only in $[0, 0.005]$.

3 Asymptotic FDR Control Under Weak Dependence

Besides the i.i.d. case, `NeuralFDR` also controls FDR asymptotically under weak dependence [13, 21]. Extending the weak dependence definition in [13] from discrete groups to continuous features $\mathbf{X}$, the data are defined to be under weak dependence if the CDF of (P_i, X_i) for the null and the alternative proportion converge almost surely to their true values respectively. The linkage disequilibrium (LD) in GWAS and the correlated genes in RNA-Seq can be addressed by such dependence structure.

Definition 3. *(Weak dependence) For the data $\{(P_i, \mathbf{X}_i, H_i)\}_{i=0}^n$ with the marginal distribution described by (4), let $F_0(p, \mathbf{x})$ and $F_1(p, \mathbf{x})$ be the cumulative density function of the distributions over $(P_i, \mathbf{X}_i)$ defined as $\mathbb{P}(P_i \le p, \mathbf{X}_i \le \mathbf{x}, H_i = 0)$, $\mathbb{P}(P_i \le p, \mathbf{X}_i \le \mathbf{x}, H_i = 1)$ respectively, where the inequality for vectors are element-wise. The data is under weak dependence if $\forall(p, \mathbf{x})$,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{P_i \le p, \mathbf{X}_i \le \mathbf{x}, H_i = 0\}} \overset{a.s.}{\to} F_0(p, \mathbf{x}), \quad \sum_{i=1}^{n} \mathbb{I}_{\{P_i \le p, \mathbf{X}_i \le \mathbf{x}, H_i = 1\}} \overset{a.s.}{\to} F_1(p, \mathbf{x}).$$

Theorem 3. *(FDP control under weak dependence) Under weak dependence, `NeuralFDR` with weight clipping controls FDR asymptotically. The weight clipping refers to clamping the weights to a bounded set after each gradient update when training the neural network [2].*

Proof. (Proof of theorem 3.) Partition the space of $(p, \mathbf{x})$ into k small boxes $B_1, \cdots, B_k$. Under the weak dependence assumption Def. 3, the proportion of elements in each box B_j, $\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{(P_i, \mathbf{X}_i) \in B_j\}}$, converges uniformly to its true value, both for the CV set and the testing set. As $k \to \infty$, the boxes become smaller. Then for the family of Lipschitz continuous thresholds, their corresponding mirror estimates can be uniformly approximated by the proportion of elements in the boxes above the mirrored threshold. Hence, as $k \to \infty$, the mirror estimates converge uniformly to their true values for the family of Lipschitz continuous thresholds. Since `NeuralFDR` with weight clipping produces Lipschitz continuous thresholds, regardless of the value of L, the mirror estimates on the CV set and on the testing set will converge to their true values. Hence the difference of the mirror estimates on the CV set and on the testing set will converge to zero, giving that `NeuralFDR` controls FDR asymptotically. □

Remark 5. *(Lipschitz continuity) In the i.i.d. case, we do not need Lipschitz continuity because for any L learned thresholds based on the training data, their concentration on the CV data can be characterized by the concentration inequality and the union bound due to the i.i.d. structure. Therefore a FDR control guarantee can be established. For the weakly dependent case, however, the convergence rate is hard to characterize. All we have is a point-wise almost surely convergence, with rate unknown. Hence, we first establish a uniform convergence for the k boxes, or in other words simple functions with a fixed resolution depending on k, and use them to approximate the learned threshold. In this case, since we have no idea what the convergence rate of the L learned thresholds will be like, we seek a uniform approximation of the family of learned thresholds. Then this family should have some nice properties regarding the continuity in order for the approximation to be true, and Lipschitz continuity is one of the options.*

4 Proofs of the Theoretical Results

4.1 Proof of Lemma 1

Proof. (Proof of Lemma 1) $\forall \mathbf{x}$, given $\mathbf{X}_i = \mathbf{x}$ and $H_i = 0$, $P_i \sim \text{Unif}(0, 1)$. Then the joint distribution $f_{P\mathbf{X}H}(p, \mathbf{x}, 0)$ is also uniform w.r.t. p. We have

$$\mathbb{P}((P_i, X_i) \in C(t), H_i = 0) = \mathbb{P}((P_i, X_i) \in C^M(t), H_i = 0).$$

Then

$$\begin{aligned}
\mathbb{E}[\widehat{FD}(t)] &= \sum_{i=1}^{n} \mathbb{P}((P_i, X_i) \in C^M(t)) \\
&= \sum_{i=1}^{n} \mathbb{P}((P_i, X_i) \in C^M(t), H_i = 0) + \sum_{i=1}^{n} \mathbb{P}((P_i, X_i) \in C^M(t), H_i = 1) \\
&= \sum_{i=1}^{n} \mathbb{P}((P_i, X_i) \in C(t), H_i = 0) + \sum_{i=1}^{n} \mathbb{P}((P_i, X_i) \in C^M(t), H_i = 1) \\
&= \mathbb{E}[FD(t)] + \sum_{i=1}^{n} \mathbb{P}((P_i, X_i) \in C^M(t), H_i = 1).
\end{aligned}$$

□

4.2 Proof of Theorem 1

Proof. (Proof of Theorem 1) Consider fold j. Any decision rule candidate t_{jl} may depend on the training set $\mathcal{D}_{tr}(j)$ but is independent of the cross validation set $\mathcal{D}_{cv}(j)$. Thus assuming the first point is not in the training set, we can define

$$\begin{aligned}
p_{jl} &= \frac{1}{m}\mathbb{E}[FD(t_{jl}, \mathcal{D}_{cv}(j))] = \mathbb{P}(P_1 \le t_{jl}(X_1), H_1 = 0) \\
\overline{p}_{jl} &= \frac{1}{m}\mathbb{E}[\widehat{FD}(t_{jl}, \mathcal{D}_{cv}(j))] = \mathbb{P}(P_1 \ge 1 - t_{jl}(X_1)) \\
q_{jl} &= \frac{1}{m}\mathbb{E}[D(t_{jl}, \mathcal{D}_{cv}(j))] = \mathbb{P}(P_1 \le t_{jl}(X_1)).
\end{aligned}$$

Notice that $p_{jl} \le \overline{p}_{jl}$.

Let the mirror estimate $\widehat{FD}(t_{jl}, \mathcal{D}_{cv}(j))$ and the number of discoveries $D(t_{jl}, \mathcal{D}_{cv}(j))$ be the quantities evaluated on $\mathcal{D}_{cv}(j)$. We know $\widehat{FD}(t_{jl}, \mathcal{D}_{cv}(j)) \sim \text{Bin}(m, \overline{p}_{jl})$ and $D(t_{ij}, \mathcal{D}_{cv}(j)) \sim \text{Bin}(m, q_{jl})$. By Lemma 2,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{FD}(t_{jl}, \mathcal{D}_{cv}(j)) \le (1-\delta_1)m\overline{p}_{jl}\right) < \exp -\frac{\delta_1^2 m\overline{p}_{jl}}{2}, && \forall\, 0 < \delta_1 < 1 \\
&\mathbb{P}\left(D(t_{jl}, \mathcal{D}_{cv}(j)) \ge (1+\delta_2)mq_{jl}\right) < \exp -\frac{\delta_2^2 mq_{jl}}{2+\delta_2}, && \forall\, \delta_2 > 0
\end{aligned}$$

As $\widehat{FDP}(t_{jl}, \mathcal{D}_{cv}(j)) = \frac{\widehat{FD}(t_{jl}, \mathcal{D}_{cv}(j))}{D(t_{jl}, \mathcal{D}_{cv}(j))}$, we have

$$\mathbb{P}\left(\widehat{FDP}(t_{jl}, \mathcal{D}_{cv}(j)) \le \frac{1-\delta_1}{1+\delta_2}\frac{\overline{p}_{jl}}{q_{jl}}\right) < \exp -\frac{\delta_1^2 m\overline{p}_{jl}}{2} + \exp -\frac{\delta_2^2 mq_{jl}}{2+\delta_2}.$$

Consider the "bad" event rule t_{jl} such that $\frac{\overline{p}_{jl}}{q_{jl}} \ge \frac{1+\delta_2}{1-\delta_1}\alpha$ or $q_{jl} \le \frac{1}{1+\delta_2}c_0$,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{FDP}(t_{jl}, \mathcal{D}_{cv}(j)) \le \alpha, \frac{D(t_{jl}, \mathcal{D}_{cv}(j))}{m} \ge c_0\right) \\
&\le \mathbb{P}\left(\widehat{FDP}(t_{jl}, \mathcal{D}_{cv}(j)) \le \alpha, \frac{\overline{p}_{jl}}{q_{jl}} \ge \frac{1+\delta_2}{1-\delta_1}\alpha\right) \bigvee \mathbb{P}\left(\frac{D(t_{jl}, \mathcal{D}_{cv}(j))}{m} \ge c_0, q_{jl} \le \frac{1}{1+\delta_2}c_0\right) \\
&< \left(\exp -\frac{\delta_1^2 m\alpha q_{jl}}{2} + \exp -\frac{\delta_2^2 mq_{jl}}{2+\delta_2}\right) \bigvee \left(\exp -\frac{\delta_2^2 mq_{jl}}{2+\delta_2}\right) = \exp -\frac{\delta_1^2 m\alpha q_{jl}}{2} + \exp -\frac{\delta_2^2 mq_{jl}}{2+\delta_2},
\end{aligned}$$

where for the second inequality we note that $\overline{p}_{jl} \ge \frac{1+\delta_2}{1-\delta_1}\alpha q_{jl} > \alpha q_{jl}$.

Let $\underline{q} = \frac{1}{1+\delta_2}c_0$ and let $\mathcal{S} = \{t_{jl} : \frac{\overline{p}_{jl}}{q_{jl}} \ge \frac{1+\delta_2}{1-\delta_1}\alpha \text{ or } q_{jl} \le \underline{q}\}$. We know that there are at most L elements in $\mathcal{S}$. Then by the union bound,

$$\mathbb{P}\left(\exists\, l \in \mathcal{S}, s.t.\ \widehat{FDP}(t_{jl}, \mathcal{D}_{cv}(j)) \le \alpha, \frac{D(t_{jl}, \mathcal{D}_{cv}(j))}{m} \ge c_0\right) < L\left(\exp -\frac{\delta_1^2 \alpha m\underline{q}}{2} + \exp -\frac{\delta_2^2 m\underline{q}}{2+\delta_2}\right).$$

Furthermore, let the l^*-th rule, t_{jl^*}, be the rule selected for testing. Note here l^* is a random variable, $\widehat{FDP}(t_{jl^*}, \mathcal{D}_{cv}(j)) \leq \alpha$, and $\frac{D(t_{jl^*}, \mathcal{D}_{cv}(j))}{m} \geq c_0$. Therefore

$$\mathbb{P}(l^* \in \mathcal{S}) = \mathbb{P}\left(\frac{\overline{p}_{jl^*}}{q_{jl*}} \geq \frac{1+\delta_2}{1-\delta_1}\alpha \ \ or \ \ q_{jl^*} \leq \underline{q}\right) < L\left(\exp -\frac{\delta_1^2 \alpha m \underline{q}}{2} + \exp -\frac{\delta_2^2 m \underline{q}}{2+\delta_2}\right).$$

As $p_{jl^*} < \overline{p}_{jl^*}$, we have

$$\mathbb{P}\left(\frac{p_{jl^*}}{q_{jl^*}} \geq \frac{1+\delta_2}{1-\delta_1}\alpha \ \ or \ \ q_{jl^*} \leq \underline{q}\right) < L\left(\exp -\frac{\delta_1^2 \alpha m \underline{q}}{2} + \exp -\frac{\delta_2^2 m \underline{q}}{2+\delta_2}\right). \tag{12}$$

Now we move to the test data $\mathcal{D}_{te}(j)$. Again by Lemma 2, given t_{jl^*},

$$\mathbb{P}\left(FD(t_{jl^*}, \mathcal{D}_{te}(j)) \geq (1+\delta_3) m p_{jl^*} | t_{jl^*}\right) < \exp -\frac{\delta_3^2 m p_{jl^*}}{2+\delta_3}, \qquad \forall\, \delta_3 > 0$$

$$\mathbb{P}\left(D(t_j^*, \mathcal{D}_{te}(j)) \leq (1-\delta_4) m q_{jl^*} | t_{jl^*}\right) < \exp -\frac{\delta_4^2 m q_{jl^*}}{2}, \qquad \forall\, 0 < \delta_4 < 1.$$

Then, given t_j^*,

$$\mathbb{P}\left(FDP(t_{jl^*}, \mathcal{D}_{te}(j)) \geq \frac{1+\delta_3}{1-\delta_4}\frac{p_{ij^*}}{q_{jl^*}} \middle| t_{jl^*}\right) < \exp -\frac{\delta_3^2 m p_{jl^*}}{2+\delta_3} + \exp -\frac{\delta_4^2 m q_{jl^*}}{2}.$$

The probability that FDP is large can be decomposed as follows:

$$\mathbb{P}\left(FDP(t_{jl^*}, \mathcal{D}_{te}(j)) \geq \frac{1+\delta_2}{1-\delta_1}\frac{1+\delta_3}{1-\delta_4}\alpha\right) \tag{13}$$

$$\leq \mathbb{P}\left(FDP(t_{jl^*}, \mathcal{D}_{te}(j)) \geq \frac{1+\delta_2}{1-\delta_1}\frac{1+\delta_3}{1-\delta_4}\alpha \middle| \frac{p_{jl^*}}{q_{jl^*}} < \frac{1+\delta_2}{1-\delta_1}\alpha, q_{jl^*} \geq \underline{q}\right) \tag{14}$$

$$+ \mathbb{P}\left(\frac{p_{jl^*}}{q_{jl^*}} \geq \frac{1+\delta_2}{1-\delta_1}\alpha \ \ or \ \ q_{jl^*} \leq \underline{q}\right). \tag{15}$$

For the conditional probability in the first term, we have

$$\mathbb{P}\left(FDP(t_{jl^*}, \mathcal{D}_{te}(j)) \geq \frac{1+\delta_2}{1-\delta_1}\frac{1+\delta_3}{1-\delta_4}\alpha \middle| \frac{p_{jl^*}}{q_{jl^*}} < \frac{1+\delta_2}{1-\delta_1}\alpha, q_{jl^*} \geq \underline{q}\right) \tag{16}$$

$$\leq \mathbb{P}\left(FDP(t_j^*, \mathcal{D}_{te}(j)) \geq \frac{1+\delta_2}{1-\delta_1}\frac{1+\delta_3}{1-\delta_4}\alpha \middle| \frac{p_{jl^*}}{q_{jl^*}} = \frac{1+\delta_2}{1-\delta_1}\alpha, q_{jl^*} \geq \underline{q}\right) \tag{17}$$

$$\leq \exp -\frac{\delta_3^2 \alpha m \underline{q}}{2+\delta_3} + \exp -\frac{\delta_4^2 m \underline{q}}{2}. \tag{18}$$

Combining (12), (16), (13) can be written as

$$\mathbb{P}\left(FDP(t_{jl^*}, \mathcal{D}_{te}(j)) \geq \frac{1+\delta_2}{1-\delta_1}\frac{1+\delta_3}{1-\delta_4}\alpha\right)$$
$$< L\left(\exp -\frac{\delta_1^2 \alpha m \underline{q}}{2} + \exp -\frac{\delta_2^2 m \underline{q}}{2+\delta_2}\right) + \left(\exp -\frac{\delta_3^2 \alpha m \underline{q}}{2+\delta_3} + \exp -\frac{\delta_4^2 m \underline{q}}{2}\right).$$

Finally, by union bound over all M folds,

$$\mathbb{P}\left(\exists\, j,\ FDP(t_j^*, \mathcal{D}_{te}(j)) \geq \frac{1+\delta_2}{1-\delta_1}\frac{1+\delta_3}{1-\delta_4}\alpha\right) \tag{19}$$

$$< LM\left(\exp -\frac{\delta_1^2 \alpha m \underline{q}}{2} + \exp -\frac{\delta_2^2 m \underline{q}}{2+\delta_2}\right) + M\left(\exp -\frac{\delta_3^2 \alpha m \underline{q}}{2+\delta_3} + \exp -\frac{\delta_4^2 m \underline{q}}{2}\right), \tag{20}$$

for some $\delta_1, \delta_4 \in (0,1)$, $\delta_2, \delta_3 > 0$. Note that (19) also indicates that the overall FDP is smaller than $\frac{1+\delta_2}{1-\delta_1}\frac{1+\delta_3}{1-\delta_4}\alpha$.

Now let us derive an asymptotic bound when $\delta_1, \delta_2, \delta_3, \delta_4$ are close to 0. In this case, we have $\delta_1, \delta_2, \delta_3, \delta_4 \in (0,1)$ and (19) can be reduced to

$$\mathbb{P}\left(\exists\, j,\ FDP(t_j^*, \mathcal{D}_{te}(j)) \geq \frac{1+\delta_2}{1-\delta_1}\frac{1+\delta_3}{1-\delta_4}\alpha\right) \tag{21}$$

$$< LM\left(\exp -\frac{\delta_1^2 \alpha m \underline{q}}{2} + \exp -\frac{\delta_2^2 m \underline{q}}{3}\right) + M\left(\exp -\frac{\delta_3^2 \alpha m \underline{q}}{3} + \exp -\frac{\delta_4^2 m \underline{q}}{2}\right). \tag{22}$$

Let $\Delta = \min_j \delta_j$. Then $\frac{1+\delta_2}{1-\delta_1}\frac{1+\delta_3}{1-\delta_4} - 1 = O(\Delta)$.

For some $\beta > 0$, let the four terms in (21) be equal to $\frac{\beta}{4}$ so that the overall probability is β. This gives

$$\delta_1 = \sqrt{\frac{2}{\alpha m \underline{q}} \log \frac{4ML}{\beta}}, \quad \delta_2 = \sqrt{\frac{3}{m\underline{q}} \log \frac{4ML}{\beta}}, \quad \delta_3 = \sqrt{\frac{3}{\alpha m \underline{q}} \log \frac{4M}{\beta}}, \quad \delta_4 = \sqrt{\frac{2}{m\underline{q}} \log \frac{4M}{\beta}}.$$

Thus $\Delta = \min_j \delta_j = O(\sqrt{\frac{M}{\alpha n} \log \frac{ML}{\beta}})$, where we note that the constant $\underline{q}$ is hidden inside the big O term and $m = \frac{n}{M}$. This completes the proof. □

4.3 Proof of Theorem 2

Proof. (Proof of Theorem 2) We first identify the worse case null proportion π_0^*. Consider any rule t. As $f_{P\mathbf{X}}$ is fixed, the probability of discovery $P_D(\gamma t, f_{P\mathbf{X}})$ is determined. For any two null proportions π_0 and π_0', if $\forall \mathbf{x}, \pi_0(\mathbf{x}) \geq \pi_0'(\mathbf{x})$, the probability of false discovery $P_{FD}(t, f_{P\mathbf{X}}, \pi_0) \geq P_{FD}(t, f_{P\mathbf{X}}, \pi_0')$, giving $FDP(t, f_{P\mathbf{X}}, \pi_0) \geq FDP(t, f_{P\mathbf{X}}, \pi_0')$. Hence FDP is maximized when $\pi_0(\mathbf{x})$ is maximized for each point of $\mathbf{x}$. As $\forall \mathbf{x}$, $\pi_0(\mathbf{x}) \leq f_{P|\mathbf{X}}(1|\mathbf{x})$ and $\pi_0^*(\mathbf{x}) = f_{P|\mathbf{X}}(1|\mathbf{x})$ is attainable, we know for any rule t, FDP is maximized with $\pi_0^*(\mathbf{x})$. Then, problem (6) can be rewritten as

$$\max_t P_D(t, f_{P\mathbf{X}}) \ \ s.t.\ \ FDP(t, f_{P\mathbf{X}}, \pi_0^*) \leq \alpha.$$

For condition (7.2), we prove by contradiction. Suppose t is the optimal rule and $FDP(t, f_{P\mathbf{X}}, \pi_0^*) < \alpha$. Then there exists $\gamma > 1$ such that $FDP(\gamma t, f_{P\mathbf{X}}, \pi_0^*) \leq \alpha$. As $P_D(\gamma t, f_{P\mathbf{X}}) > P_D(t, f_{P\mathbf{X}})$, t can not be the optimal rule, giving the contradiction.

For condition (7.1), we also prove by contradiction. Again suppose t is the optimal rule where (7.1) is not met, then there exists $\mathcal{X}_1, \mathcal{X}_2 \subset \mathcal{X}$ with positive measure such that

$$\frac{\int_{\mathcal{X}_1} f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x}}{\int_{\mathcal{X}_1} f_{P\mathbf{X}}(t(\mathbf{x}), \mathbf{x}) d\mathbf{x}} < \frac{\int_{\mathcal{X}_2} f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x}}{\int_{\mathcal{X}_2} f_{P\mathbf{X}}(t(\mathbf{x}), \mathbf{x}) d\mathbf{x}}.$$

Note that $f_{P\mathbf{X}}(p, \mathbf{x})$ is monotonically decreasing w.r.t. p. Then there exists $\epsilon > 0$ such that for any $\epsilon_1, \epsilon_2 \in (0, \epsilon)$,

$$\frac{\epsilon_1 \int_{\mathcal{X}_1} f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x}}{\int_{\mathcal{X}_1} \int_{t(\mathbf{x})}^{t(\mathbf{x})+\epsilon_1} f_{P\mathbf{X}}(t(\mathbf{x}) + \epsilon_1, \mathbf{x}) dp\, d\mathbf{x}} < \frac{\epsilon_2 \int_{\mathcal{X}_2} f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x}}{\int_{\mathcal{X}_2} \int_{t(\mathbf{x})-\epsilon_2}^{t(\mathbf{x})} f_{P\mathbf{X}}(t(\mathbf{x}) - \epsilon_2, \mathbf{x}) dp\, d\mathbf{x}}. \tag{23}$$

Then we can pick $\epsilon_1, \epsilon_2 < \epsilon$ such that

$$\int_{\mathcal{X}_1} \int_{t(\mathbf{x})}^{t(\mathbf{x})+\epsilon_1} f_{P\mathbf{X}}(p, \mathbf{x}) dp\, d\mathbf{x} = \int_{\mathcal{X}_2} \int_{t(\mathbf{x})-\epsilon_2}^{t(\mathbf{x})} f_{P\mathbf{X}}(p, \mathbf{x}) dp\, d\mathbf{x} > 0 \tag{24}$$

Defining a new rule $t'(\mathbf{x})$ as

$$t'(\mathbf{x}) = \begin{cases} t(\mathbf{x}) + \epsilon_1, & \mathbf{x} \in \mathcal{X}_1 \\ t(\mathbf{x}) - \epsilon_2, & \mathbf{x} \in \mathcal{X}_2 \\ t(\mathbf{x}), & otherwise \end{cases}.$$

Then for the probability of discovery,

$$P_D(t', f_{P\mathbf{X}}) = \int_{\mathcal{X}} \int_0^{t'(\mathbf{x})} f_{P\mathbf{X}}(p, \mathbf{x}) dp \, d\mathbf{x} = \int_{\mathcal{X}} \int_0^{t(\mathbf{x})} f_{P\mathbf{X}}(p, \mathbf{x}) dp \, d\mathbf{x}$$
$$+ \int_{\mathcal{X}_1} \int_{t(\mathbf{x})}^{t(\mathbf{x})+\epsilon_1} f_{P\mathbf{X}}(p, \mathbf{x}) dp \, d\mathbf{x} - \int_{\mathcal{X}_2} \int_{t(\mathbf{x})-\epsilon_2}^{t(\mathbf{x})} f_{P\mathbf{X}}(p, \mathbf{x}) dp \, d\mathbf{x} = P_D(t, f_{P\mathbf{X}}).$$

Moreover, from (23) and (24) we also know

$$\epsilon_1 \int_{\mathcal{X}_1} f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x} < \epsilon_2 \int_{\mathcal{X}_2} f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x}.$$

Then

$$P_{FD}(t', f_{P\mathbf{X}}, \pi_0^*) = \int_{\mathcal{X}} t'(\mathbf{x}) \pi_0^*(\mathbf{x}) \mu(\mathbf{x}) d\mathbf{x} = \int_{\mathcal{X}} t'(\mathbf{x}) f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x}$$
$$= \int_{\mathcal{X}} t(\mathbf{x}) f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x} + \epsilon_1 \int_{\mathcal{X}_1} f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x} - \epsilon_2 \int_{\mathcal{X}_2} f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x}$$
$$< \int_{\mathcal{X}} t(\mathbf{x}) f_{P\mathbf{X}}(1, \mathbf{x}) d\mathbf{x} = P_{FD}(t, f_{P\mathbf{X}}, \pi_0^*)$$

Then $FDR(t', f_{P\mathbf{X}}, \pi_0^*) < FDR(t, f_{P\mathbf{X}}, \pi_0^*) = \alpha$. According to condition (7.2), t' can not be the optimal rule. As t and t' are both feasible and have the same discovery probability, t can not be the optimal rule either, giving the contradiction. □

4.4 Ancillary lemmas

Lemma 2. *(Chernoff bound) For i.i.d. random variables $X_1, \cdots, X_n \in [0, 1]$, let $X = \sum_{i=1}^n X_i$ and let $\mu = \mathbb{E}[X]$. Then*

$$\mathbb{P}(X \geq (1+\delta)\mu) < \exp -\frac{\delta^2 \mu}{2+\delta}, \qquad \forall\, \delta > 0$$
$$\mathbb{P}(X \leq (1-\delta)\mu) < \exp -\frac{\delta^2 \mu}{2}, \qquad \forall\, 0 < \delta < 1.$$