A Model-free Closeness-of-influence Test for Features in Supervised Learning

Mohammad Mehrabi [1] **Ryan A. Rossi** [2]

Abstract

Understanding the effect of a feature vector $x \in \mathbb{R}^d$ on the response value (label) $y \in \mathbb{R}$ is the cornerstone of many statistical learning problems. Ideally, it is desired to understand how a set of collected features combine together and influence the response value, but this problem is notoriously difficult, due to the high-dimensionality of data and limited number of labeled data points, among many others. In this work, we take a new perspective on this problem, and we study the question of assessing the difference of influence that the two given features have on the response value. We first propose a notion of closeness for the influence of features, and show that our definition recovers the familiar notion of the magnitude of coefficients in the parametric model. We then propose a novel method to test for the closeness of influence in general model-free supervised learning problems. Our proposed test can be used with finite number of samples with control on type I error rate, no matter the ground truth conditional law $\mathcal{L}(Y|X)$. We analyze the power of our test for two general learning problems i) linear regression, and ii) binary classification under mixture of Gaussian models, and show that under the proper choice of score function, an internal component of our test, with sufficient number of samples will achieve full statistical power. We evaluate our findings through extensive numerical simulations, specifically we adopt the datamodel framework (Ilyas, et al., 2022) for CIFAR-10 dataset to identify pairs of training samples with different influence on the trained model via optional black box training mechanisms.

[1]Department of Data Sciences and Operations, University of Southern California, Los Angeles, USA [2]Adobe, San Jose, USA. Correspondence to: Mohammad Mehrabi <mehrabim@marshall.usc.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

In a classic supervised learning problem, we are given a dataset of n iid data points $\{(x_i, y_i)\}_{i=1:n}$ with feature vectors $x \in \mathbb{R}^d$ and response value (label) $y \in \mathbb{R}$. From the inferential point of view, understanding the influence of each individual feature $i \in \{1, \ldots, d\}$ on y is of paramount importance. Considering a parametric family of distributions for $\mathcal{L}(Y|X)$ is among the most studied techniques for this problem. In this setting, the influence of each feature can be seen by their corresponding coefficient value in the parametric model. Essentially such methods can result in spurious statistical findings, mainly due to model misspecification, where in the first place the ground-truth data generating law $\mathcal{L}(Y|X)$ does not belong to the considered parametric family. A natural remedy for this problem is to relax the parametric family assumption, removing concerns about model misspecification. Besides the difficulties with the new model-free structure of the problem, we need a new notion to capture the influence of features, as there is no longer a coefficient vector as per class of parametric models.

In this paper, we follow the model-free structure, but take a new perspective on the generic problem of investigating the influence of features on the response value. In particular, as a first step towards this notoriously hard question under no class of parametric distribution assumption or whatsoever, we are specifically interested in assessing the closeness of influence of features. For this end, we posit the following fundamental question:

(*) In a general model-free supervised learning problem, for two given features, is it possible to assess the closeness of their influence on the response value (label) in a statistically sound way?

In this paper, we answer question (*) affirmatively. We characterize a notion of closeness for the influence of features on y under the general model-free framework. We show that this notion aligns perfectly well with former expectations in parametric models, where small difference in the coefficient values imply close influence on the response value. We then cast the closeness of influence question as a hypothesis testing problem, and show that we can control associated type I error rate with finite number of samples.

1.1. Motivation Behind Question (*)

Beyond the inferential nature of Question (*) that helps to better understand the data-generating process of on-hand data, being able to answer this question has a myriad of applications for other classic machine learning tasks. In fact, inspired by the recent advancements in interpretable machine learning systems, it is desired to strike a balance between model flexibility in capturing the ground-truth law $\mathcal{L}(Y|X)$ and using few number of explanatory variables. For this goal, feature aggregation has been used to distill a large amount of feature information into a smaller number of features. In several parametric settings, features with equal coefficients are naturally grouped together, e.g, in linear regression new feature $x_1 + x_2$ is considered rather than (x_1, x_2), in case that x_1, x_2 have equal corresponding regression coefficients (Yan & Bien, 2021). In addition, identifying features with near influence on the response value can be used for tree-based aggregation schemes (Shao et al., 2021; Bien et al., 2021; Wilms & Bien, 2022). This is of paramount importance in learning problems involving rare features, such as the count of microbial species (Bien et al., 2021). In addition, in many learning problems, an honest comprehensive assessment for characterizing the behavior of Y with respect to a certain attribute A is desired. This can be used to assess the performance of model with respect to a sensitive attribute (fair machine learning), or to check if two different treatments (different values of A) have close influence on potential outcomes.

1.2. Related Work

In machine learning, the problem of identifying a group of features that have the largest influence on the response value is often formulated as variable selection. With a strong parametric assumption, the conditional law $\mathcal{L}(Y|X)$ is considered to belong to a known class of parametric models, such as linear regression. For variable selection in the linear regression setting, the LASSO (Tibshirani, 1996) and Dantzig selector (Candes & Tao, 2007) are the most widely used. In fact, there are several other works for variable selection in the linear regression setting with output solutions satisfying certain structures, such as (Bogdan et al., 2015; Tibshirani et al., 2005). There has been another complimentary line in the past years from model-X perspective (Candes et al., 2018). In this setting, despite the classical setup, in which a strong parametric assumption is considered on the conditional law, it shifts the focus to the feature distribution X and assumes an extensive knowledge on the distribution of the features. This setting arises naturally in many learning problems. For example, we can get access to distributional information on features in learning scenarios where the sampling mechanism can be controlled, e.g,. in datamodel framework (Ilyas et al., 2022), and gene knockout experiments (Peters et al., 2016; Cong et al., 2013). Other settings include problems where an abundant number of unlabeled data points (unsupervised learning) are available.

The other related line of work is to estimate and perform statistical inference on certain statistical model parameters. Specifically, during the past few years, there have been several works (Javanmard & Montanari, 2014; Van de Geer et al., 2014; Deshpande et al., 2019; Fei & Li, 2021) for inferential tasks on low-dimensional components of model parameters in high-dimensional $(d > n)$ settings of linear and generalized linear models. Another complementary line of work, is the conditional independence testing problem $X_j \perp\!\!\!\perp Y|X_{-j}$ to test if a certain feature X_j is independent of the response value Y, while controlling for the effect of the other features. This problem has been studied in several recent works for both parametric (Crawford et al., 2018; Belloni et al., 2014), and model-X frameworks (Candes et al., 2018; Javanmard & Mehrabi, 2021; Liu et al., 2022; Shaer & Romano, 2022; Berrett et al., 2020).

Here are couple of points worth mentioning regarding the scope of our paper.

1. *(Feature selection methods)* However Question (*) has a complete different nature from well-studied variable selection techniques– with the goal of removing redundant features, an assessment tool provided for (*) can be beneficial for post-processing of feature selection methods as well. Specifically, we expect that two redundant features have close (zero) influence on the response value, therefore our closeness-of-influence test can be used to sift through the set of redundant features and potentially improve the statistical power of the baseline feature selection methods.

2. *(Regression models)* We would like to emphasize that however fitting any class of regression models would yield an estimate coefficient vector, but comparing the magnitude of coefficient values for answering Question (*) is not statistically accurate and would result in invalid findings, mainly due to model misspecification. Despite such inaccuracies of fitted regression models, our proposed closeness-of-influence test works under no parametric assumption on the conditional law.

3. *(Hardness of non-parametric settings)* The finite-sample guarantee on type-I error rate for our test does not come free. Specifically, this guarantee holds when certain partial knowledge on the feature distributions $\mathcal{L}(X)$ is known. This setup is often referred as model-X framework (Candes et al., 2018), where on contrary to the classic statistic setups, the conditional law $\mathcal{L}(Y|X)$ is optional, and adequate amount of information on features distribution $\mathcal{L}(X)$ is known. Such requirements for features distribution makes the scope

of our work distant from completely non-parametric problems.

1.3. Summary of contributions and organization

In this work, we propose a novel method to test the closeness of influence of a given pair of features on the response value. Here is the organization of the three major parts of the paper:

- In Section 2, we propose the notion of symmetric influence and formulate the question (*) as a tolerance hypothesis testing problem. We then introduce the main algorithm to construct the test statistic, and the decision rule. We later show that the type-I error is controlled for finite number of data points.
- In Section 3, for two specific learning problems: 1) linear regression setup, and 2) binary classification under a mixture of Gaussians, we analyze the statistical power of our proposed method. Our analysis reveals guidelines on the choice of the score function, that is needed for our procedure.
- In Section 5, we combine our closeness-of-influence test with datamodels (Ilyas et al., 2022) to study the influence of training samples on the trained black box model. We consider CIFAR-10 dataset and identify several pairs of training samples with different influence on the output models.

Finally, we empirically evaluate the performance of our method in several numerical experiments, we show that our method always controls type-I error with finite number of data points, while it can achieve high statistical power. We end the paper by providing concluding remarks and interesting venues for further research.

1.4. Notation

For a random variable X, we let $\mathcal{L}(X)$ denote the probability density function of X. For two density functions p, q let $d_{\mathsf{TV}}(p, q)$ denote the total variation distance. We use $\Phi(t)$ and $\varphi(t)$ respectively for cdf and pdf of standard normal distribution. For and integer n let $[n] = \{1, \dots, n\}$ and for a vector $x \in \mathbb{R}^d$ and integers $i, j \in [d]$ let $x_{\mathsf{swap}(i,j)}$ be a vector obtained by swapping the coordinates i and j of x. We let $\mathsf{N}(\mu, \Sigma)$ denote the probability density function of a multivariate normal distribution with mean μ and covariance matrix Σ.

2. Problem Formulation

We are interested in investigating that if two given features i, j have close influence on the response value y. Specifically, in the case of the linear regression setting $\mathcal{L}(Y|X) = \mathsf{N}(X^\mathsf{T}\theta, \sigma^2)$, two features i and j have an equal effect on the response variable y, if the model parameter θ has equal coordinates in i and j. In this parametric problem, the close influence analysis can be formulated as the following hypothesis testing problem

$$H_0 : |\theta_i - \theta_j| \le \tau \,, \quad H_A : |\theta_i - \theta_j| > \tau \,.$$

In practice, the considered parametric model may not hold, and due to model misspecification, the reported results are not statistically sound and accurate. Our primary focus is to extend the definition of close influence of features on the response value to a broader class of supervised learning problems, ideally with no parametric assumption on $\mathcal{L}(Y|X)$ (model-free). For this end, we first propose the notion of *symmetric influence*.

Definition 2.1 (Symmetric influence)**.** *We say that two features $i, j \in [d]$ have a symmetric influence on the response value y if the conditional law $p_{Y|X}$ does not change once features i and j are swapped in x. More precisely, if $\mathcal{L}(Y|X) = \mathcal{L}(Y|X_{\mathsf{swap}(i,j)})$, where $X_{\mathsf{swap}(i,j)}$ is obtained from swapping coordinates i and j in X.*

While the perfect alignment between density function $p_{Y|X}$ and $p_{Y|X_{\mathsf{swap}(i,j)}}$ is considered as equal influence, it is natural to consider small (but nonzero) average distance of these two density functions as having close influence of features i, j on the response value. Inspired by this observation, we cast the problem of closeness-of-influence testing as a tolerance hypothesis testing problem 1. Before further analyzing this extended definition, for two simple examples we show that the symmetric influence definition recovers the familiar equal effect notion in parametric problems. It is worth noting that this result can be generalized to a broader class of parametric models.

Proposition 2.2. *Consider the logistic model $\mathbb{P}(Y = 1|X = x) = \frac{1}{1+\exp(-x^\mathsf{T}\theta)}$. In this model, features i and j have symmetric influence on y if and only if $\theta_i = \theta_j$. In addition, for the linear regression setting $y = x^\mathsf{T}\theta + \varepsilon$ with $\varepsilon \sim \mathsf{N}(0, \sigma^2)$, features i and j have symmetric influence on y if and only if $\theta_i = \theta_j$.*

We refer to Appendix A for proofs of all propositions and theorems.

2.1. Closeness-of-influence testing

Inspired by the definition of symmetric influence given in Definition 2.1, we formulate the problem of testing the closeness of the influence of two features i, j on y as the following:

$$\begin{aligned} \mathcal{H}_0 :\ & \mathbb{E}\left[d_{\mathsf{TV}}(p_{Y|X}, p_{Y|X_{\mathsf{swap}(i,j)}})\right] \le \tau \,, \\ \mathcal{H}_A :\ & \mathbb{E}\left[d_{\mathsf{TV}}(p_{Y|X}, p_{Y|X_{\mathsf{swap}(i,j)}})\right] > \tau \,. \end{aligned} \tag{1}$$

Specifically, this hypothesis testing problem allows for general non-negative τ values. We can test for symmetric influence by simply selecting $\tau = 0$. In this case, we must have $p_{Y|X} = p_{Y|X_{\mathsf{swap}(i,j)}}$ almost surely (with respect to some measure on $\mathcal{X}$). For better understanding of the main quantities in the left-hand-side of 1, it is worth to note that $p_{Y|X_{\mathsf{swap}(i,j)}}(y|x) = p_{Y|X}(y|x_{\mathsf{swap}}(i,j))$ and the quantity of interest can be written as

$$\mathbb{E}\left[d_{\mathsf{TV}}(p_{Y|X}, p_{Y|X_{\mathsf{swap}(i,j)}})\right] = \frac{1}{2}\int \left|p_{Y|X}(y|x) - p_{Y|X}(y|x_{\mathsf{swap}(i,j)})\right| p_X(x)\mathrm{d}y\mathrm{d}x\,.$$

We next move to the formal process to construct the test statistics of this hypothesis testing problem.

Test statistics. We first provide high-level intuition behind the test statistics used for testing 1. In a nutshell, for two i.i.d. data points $(x^{(1)}, y^{(1)})$ and $(x^{(2)}, y^{(2)})$, if the density functions $p_{Y|X}$ is close to $p_{Y|X_{\mathsf{swap}(i,j)}}$, then for an optional score functions applied on $(x^{(1)}, y^{(1)})$ and $(x^{(2)}_{\mathsf{swap}(i,j)}, y^{(2)})$, with equal chance (50%) one should be larger than the other one. This observation is subtle though. Since we intervene in the features of the second data point (by swapping its coordinates), this shifts the features distribution, thereby the joint distribution of $(x^{(1)}, y^{(1)})$ and $(x^{(2)}_{\mathsf{swap}(i,j)}, y^{(2)})$ are not equal. This implies that we must control for such distributional shifts on features as well. The formal process for constructing the test statistics U_n is given in Algorithm 1. We next present the decision rule for hypothesis problem 1.

Algorithm 1 Test statistic for hypothesis testing 1

Input: n data points $\{(x^{(m)}, y^{(m)})\}_{m=1:n}$ with $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ (for n being even–if not, remove one sample), two features $i, j \in \{1, 2, ..., d\}$, and a score function $T : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Output: A test statistic U_n .

For $1 \le m \le \frac{n}{2}$ define

$$\tilde{x}^{(m)} = x^{(m+\frac{n}{2})}_{\mathsf{swap}(i,j)}\,, \quad \tilde{y}^{(m)} = y^{(m+\frac{n}{2})}\,.$$

Define tests statistic U_n

$$U_n = \frac{2}{n}\sum_{m=1:\frac{n}{2}} \mathbb{I}\left(T\big(x^{(m)}, y^{(m)}\big) \ge T\big(\tilde{x}^{(m)}, \tilde{y}^{(m)}\big)\right)\,.$$

Decision rule. For the data set $(\mathbf{X}, \mathbf{Y})$ of size n and test statistic U_n as per Algorithm 1 at significance level α consider the following decision rule

$$\psi_n(\mathbf{X}, \mathbf{Y}) = \mathbb{I}\left(\left|U_n - \frac{1}{2}\right| \ge \tau + \tau_X + \sqrt{\frac{\log(2/\alpha)}{n}}\right)\,, \tag{2}$$

with τ_X being an upper bound on the total variation distance between the original feature distribution, and the obtained distribution by swapping coordinates i, j. More precisely, for two independent features vectors $X^{(1)}, X^{(2)}$ let τ_X be such that $\tau_X \ge d_{\mathsf{TV}}\left(\mathcal{L}(X^{(1)}), \mathcal{L}(X^{(2)}_{\mathsf{swap}(i,j)})\right)$. In fact, in several learning problems when features have a certain symmetric structure, the quantity τ_X is zero. For instance, when features are multivariate Gaussian with isotropic covariance matrix. More on this can be seen in Section 2.2.

Size of the test. In this section, we show that the obtained decision rule 2 has control on type I error with finite number of samples. More precisely, we show that the probability of falsely rejecting the null hypothesis 1 can always be controlled such that it does not exceed a predetermined significance level α.

Theorem 2.3. *Under the null hypothesis 1, decision rule 2 has type-I error smaller than α. More precisely*

$$\mathbb{P}_{\mathcal{H}_0}(\psi(\mathbf{X}, \mathbf{Y}) = 1) \le \alpha\,.$$

Based on decision rule 1, we can construct p-values for the hypothesis testing problem 1. The next proposition gives such formulation.

Proposition 2.4. *Consider*

$$p = \begin{cases} 1\,, & |U_n - 1/2| \le \tau + \tau_X\,, \\ 1 \wedge \eta_n(U_n, \tau, \tau_X)\,, & \textit{otherwise}\,, \end{cases} \tag{3}$$

with function $\eta_n(u, \tau_1, \tau_2)$ being defined as

$$\eta_n(u, \tau_1, \tau_2) = 2\exp\left(-n\left(\left|u - \frac{1}{2}\right| - \tau_1 - \tau_2\right)^2\right)\,.$$

In this case, the p-value p is super-uniform. More precisely, under the null hypothesis 1 for every $\alpha \in [0, 1]$ we have

$$\mathbb{P}(p \le \alpha) \le \alpha\,.$$

2.2. Effect of feature swap on features distribution

From the formulation of the decision rule given in 2, it can be seen that an upper bound on total variation distance between density functions of $X^{(1)}$ and $X^{(2)}_{\mathsf{swap}(i,j)}$ is required. This quantity shows up as τ_X in 2. Regarding this change on X distribution, two points are worth mentioning. First, in several classes of learning problems the feature vectors follow a symmetric structure which renders the quantity

τ_X to zero. For instance, when features have an isotropic Gaussian distribution (Proposition 2.5), or in the datamodel sampling scheme (Ilyas et al., 2022), the formal statement is given in Proposition 2.6. Secondly, the value of τ_X can be computed when adequate amount of information is available on distribution of X, the so-called model-X framework (Candes et al., 2018). We would also like to emphasize that indeed we do not need the direct access to entire density function p_X information, and an upper bound on the quantity $d_{\mathsf{TV}}(\mathcal{L}(X^{(1)}), \mathcal{L}(X^{(2)}_{\mathsf{swap}(i,j)}))$ is sufficient. In the next proposition, for the case that features follow a general multivariate Gaussian distribution $\mathsf{N}(\mu, \Sigma)$ we provide a valid closed-form value for τ_X.

Proposition 2.5. *Consider a multivariate Gaussian distribution with the mean vector $\mu \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, for two features i and j the following holds:*

$$d_{\mathsf{TV}}\left(\mathcal{L}(X^{(1)}), \mathcal{L}(X^{(2)}_{\mathsf{swap}(i,j)})\right) \leq \frac{1}{2}\Big[\mathsf{tr}\big(-I_d + P_{ij}\Sigma^{-1}P_{ij}\Sigma\big) + (\mu - P_{ij}\mu)^{\mathsf{T}}\Sigma^{-1}(\mu - P_{ij}\mu)\Big]^{1/2}, \quad (4)$$

where P_{ij} is the permutation matrix that swaps the coordinates i and j. More precisely, for every $x \in \mathbb{R}^d$ we have $P_{ij}x = x_{\mathsf{swap}(i,j)}$.

It is easy to observe that in the case of isotropic Gaussian distribution with zero mean, we can choose $\tau_X = 0$. More concretely, when $\mu = 0$, and $\Sigma = \sigma^2 I$, then Proposition 2.5 reads $\tau_X = 0$. We next consider a setting with binary feature vectors that arise naturally in datamodels (Ilyas et al., 2022), and will be used later in experiments of Section 5.

Proposition 2.6. *Consider a learning problem with binary features vector $x \in \{0,1\}^d$. For a positive integer m, we suppose that x is sampled uniformly at random from the space $S_m = \{x \in \{0,1\}^d : \sum x_i = m\}$. This means that the output sample has binary entries with exactly m nonzero coordinates. Then, in this setting for two independent features vectors $x^{(1)}, x^{(2)}$, the following holds*

$$d_{\mathsf{TV}}\left(\mathcal{L}\big(X^{(1)}\big), \mathcal{L}\big(X^{(2)}_{\mathsf{swap}(i,j)}\big)\right) = 0\,.$$

3. Power Analysis

In this section, we provide a power analysis for our method. For a fixed score function $T : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ and two i.i.d. data points $(x^{(1)}, y^{(2)})$ and $(x^{(2)}, y^{(2)})$ consider the following cumulative distribution functions:

$$F_T(t) = \mathbb{P}\left(T(X^{(1)}, Y^{(1)}) \leq t\right)\,,$$
$$G_T(t) = \mathbb{P}\left(T(X^{(2)}_{\mathsf{swap}(i,j)}, Y^{(2)}) \leq t\right)\,.$$

In the next theorem, we show that the power of our test depends on the average deviation of the function $F_T \circ G_T^{-1}$ from the identity mappinp on the interval $[0,1]$.

Theorem 3.1. *Consider the hypothesis testing problem 1 at significance level α with n data points $(\mathbf{X}, \mathbf{Y})$. In addition, suppose that score function $T : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ satisfies the following condition for some $\beta \in (0,1)$:*

$$\left|\int_0^1 (F_T(G_T^{-1}(u)) - u)\mathrm{d}u\right| \geq \rho_n(\alpha,\beta,\tau) + \tau_X\,,$$

with $\rho_n(\alpha,\beta,\tau) = 2\exp(-n\beta^2) + \sqrt{\frac{\log(2/\alpha)}{n}} + \tau$. In this case, the decision rule 2 used with the score function T has type II error not exceeding β. More precisely $\mathbb{P}\left(\Psi_n(\mathbf{X},\mathbf{Y}) = 1\right) \geq 1-\beta$.

The function $F_T \circ G_T^{-1}$ is called *ordinal dominance curve* (ODC) (Hsieh & Turnbull, 1996; Bamber, 1975). It can be seen that the ODC is the population counterpart of the PP plot. A direct consequence of the above theorem is that if the ODC has a larger distance from the identity map $i(u) = u$, then it would be easier for our test to flag smaller gaps between the influence of features. We next focus on two learning problems: 1) linear regression setting, and 2) binary classification under Gaussian mixture models. For each problem, we use Theorem 3.1 and provide lower bounds on the statistical power of our closeness-of-influence test.

Linear regression setup. In this setting, we suppose that $y = x^{\mathsf{T}}\theta^* + \varepsilon$ for $\varepsilon \sim \mathsf{N}(0,\sigma^2)$ and feature vectors drawn iid from a multivariate normal distribution $\mathsf{N}(0, I_d)$. Since features are isotropic Gaussian with zero mean, by an application of Theorem 2.5 we know that τ_X is zero. In the next theorem, we provide an upper bound for hypothesis testing problem 1 with n data points and the score function $T(x,y) = |y - x^{\mathsf{T}}\widehat{\theta}|$ for some model estimate $\widehat{\theta}$. We show that in this example, the power of the test highly depends on the value $|\theta_i^* - \theta_j^*|$ and the quality of the model estimate $\widehat{\theta}$. Indeed, the higher the contrast between the coefficient values θ_i^* and θ_j^*, the easier it is for our test to reject the null hypothesis.

Theorem 3.2. *Under the linear regression setting $y = x^{\mathsf{T}}\theta^* + \varepsilon$ with $\varepsilon \sim \mathsf{N}(0,\sigma^2)$ with feature vectors coming from a normal population $x \sim \mathsf{N}(0, I_d)$, consider the hypothesis testing problem 1 for features i and j with $\tau \in (0,1)$. We run Algorithm 1 at the significance level α with the score function $T(x,y) = |y - x^{\mathsf{T}}\widehat{\theta}|$ for a model estimate $\widehat{\theta} \in \mathbb{R}^d$. For $\beta \in (0,1)$ such that $\tan(\frac{\pi}{2}\rho_n(\alpha,\beta,\tau)) \leq \frac{1}{2}$, suppose that the following condition holds*

$$|\theta_i^* - \theta_j^*| \geq \frac{2\tan(\frac{\pi}{2}(\rho_n(\alpha,\beta,\tau)))}{1 - 2\tan(\frac{\pi}{2}(\rho_n(\alpha,\beta,\tau)))}\frac{\left(\sigma^2 + \|\widehat{\theta} - \theta^*\|_2^2\right)}{|\widehat{\theta}_i - \widehat{\theta}_j|}\,,$$

for $\rho_n(\alpha,\beta,\tau)$ as per Theorem 3.1. Then, the type II error is bounded by β. More precisely, we have $\mathbb{P}(\Psi_n(\mathbf{X},\mathbf{Y}) = 1) \geq 1-\beta$.

We refer to Appendix for the proof of Theorem 3.2. It can be seen that the right-hand-side of the above expression can be decomposed into two major parts. The first part involves the problem parameters, such as the number of samples n, and error tolerance values α and β. This quantity for a moderately large number of samples n, and small tolerance value τ can get sufficiently small. On the other hand, the magnitude of the second part depends highly on the quality of the model estimate $\widehat{\theta}$ and the inherent noise value of the problem σ^2 which basically indicates how structured is the learning problem. Another interesting observation is regarding the $|\widehat{\theta}_i - \widehat{\theta}_j|$. Indeed, it can be inferred that small values of this quantity renders the problem of discovering deviation from the symmetric influence harder. This conforms to our expectation, given that in the extreme scenario that $\widehat{\theta}_i = \widehat{\theta}_j$ it is impossible for the score function to discern θ^*_i and θ^*_j, because of the additive nature of the considered score function.

Binary classificaiton. In this section, we provide power analysis of our method for a binary classification setting. Specifically, we consider the binary classification under a mixture of Gaussian model. More precisely, in this case the data generating process is given by

$$y = \begin{cases} +1\,, & \text{w.p } q\,, \\ -1\,, & \text{w.p } 1-q\,. \end{cases}, \quad x \sim \mathsf{N}(y\mu, I_d)\,. \tag{5}$$

We consider the influence testing problem 1 with $\tau = 0$. In the next theorem, we provide a lower bound on the statistical power of our method used under this learning setup.

Theorem 3.3. *Under the binary classification setup 5, consider the hypothesis testing problem 1 for $\tau = 0$. We run Algorithm 1 with the score function $T(x, y) = yx^\mathsf{T}\theta$ at the significance level α, and suppose that for some nonnegative value β the following holds*

$$|\mu_i - \mu_j| \geq \Phi^{-1}\left(\frac{1}{2} + \rho_n(\alpha, \beta, 0)\right) \frac{\sqrt{2}\|\widehat{\theta}\|_2}{|\widehat{\theta}_i - \widehat{\theta}_j|}\,,$$

where $\rho_n(\alpha, \beta, \tau)$ is given as per Theorem 3.1. Then the type-II error in this case is bounded by β. More concretely, we have $\mathbb{P}(\Psi_n(\mathbf{X}, \mathbf{Y}) = 1) \geq 1 - \beta$.

It is important to note that in this particular setting, the features do not follow a Gaussian distribution with a zero mean. Instead, they are sampled from a mixture of Gaussian distributions with means μ and $-\mu$. The reason why $\tau_X = 0$ can be utilized is not immediately obvious. However, we demonstrate that when testing for $\tau = 0$ under the null hypothesis, it is necessary for μ_i to be equal to μ_j, and the distribution of features remains unchanged when the coordinates i and j are swapped. As a result, we can employ $\tau_X = 0$ in this scenario. This argument is further elaborated upon in the proof of Theorem 3.3.

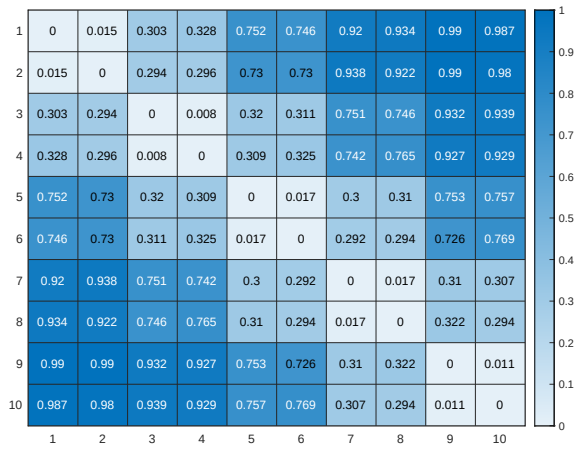
From the above expression it can be observed that for sufficiently large number of data points n and a small value τ, the value $\Phi^{-1}(1/2 + \rho_n)$ will get smaller and converge to zero. In addition, it can be inferred that an ideal model estimate $\widehat{\theta}$ must have small norm and high contrast between $\widehat{\theta}_i$ and $\widehat{\theta}_j$ values. An interesting observation can be seen on the role of other coordinate values in $\widehat{\theta}$. In fact, it can be realized that for the choice of the score function $T(x, y) = yx^\mathsf{T}\widehat{\theta}$, the support of the model estimate $\widehat{\theta}$ must be a subset of two features i and j, since this would decrease $|\widehat{\theta}|$ and increases the value of $|\widehat{\theta}_i - \widehat{\theta}_j|$.

4. Experiments

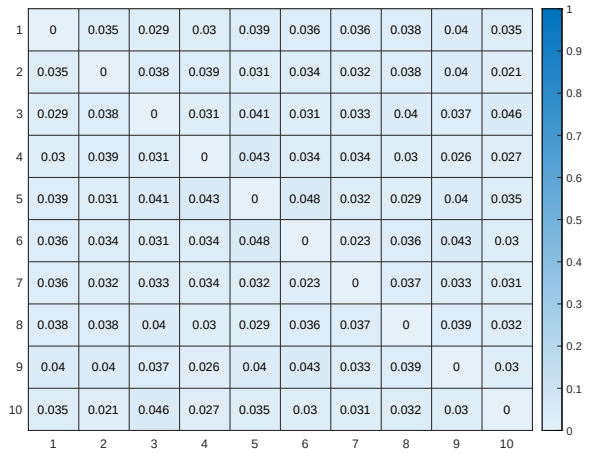
In this section, we evaluate the performance of our proposed method for identifying the symmetric influence across features. We start by the Isotropic Gaussian model for feature vectors. More precisely, we consider $x \sim \mathsf{N}(0, I_d)$ with $d = 10$. In this case, we have $\tau_X = 0$ and we consider the hypothesis testing problem 1 for $\tau = 0$ (symmetric influence).

Size of the test. We first start by examining the size of our proposed method. For this end, we consider the conditional law $y|x \sim \mathsf{N}(x^\mathsf{T}Sx, 1)$, for a semi-positive definite matrix S with coordinate (i, j) being $S_{i,j} = 1 + \mathbb{I}(i = j)$. The conditional mean of $y|x$ is a quadratic form and it is easy to observe that in this case for every two features $i, j \in \{1, \ldots, 10\}$ we have $x^\mathsf{T}Sx = x^\mathsf{T}_{\mathsf{swap}(i,j)}Sx_{\mathsf{swap}(i,j)}$, and therefore the null hypothesis holds. We test for the symmetric influence of each pair of features ($\binom{10}{2}$ number of tests). We run our method with the score function $T(x, y) = |y - \widehat{\theta}^\mathsf{T}x|$ with $\widehat{\theta} \sim \mathsf{N}(0, I_d)$. The estimate $\widehat{\theta}$ is fixed across all 45 tests. We suppose that we have access to 1000 data points, and we consider three different significance levels $\alpha = 0.1, 0.15$, and 0.2. The results of this experiment can be seen in Figure 1(b) where the reported numbers (rejection rates) are averaged over 1000 independent experiments. It can be observed that, in this case for all three significance levels, the rejection rates are smaller than α, and therefore the size of the test is controlled.

Power analysis. The linear regression setting is considered, in which $y|x \sim \mathsf{N}(x^\mathsf{T}\theta^*, 1)$, for $\theta^* \in \mathbb{R}^d$ with $d = 10$. We consider the following pattern for signal strength $\theta^*_1 = \theta^*_2 = 1$, $\theta^*_3 = \theta^*_4 = 2$, $\theta^*_5 = \theta^*_6 = 3$, $\theta^*_7 = \theta^*_8 = 4$, $\theta^*_9 = \theta^*_{10} = 5$. In this example, it can be observed that the following pairs of features $\mathcal{I} = \{(1, 2), (3, 4), (5, 6), (7, 8), (9, 10)\}$ have symmetric influence, and for any other pair the null hypothesis 1 must be rejected. We use the score function $T(x, y) = |y - x^\mathsf{T}\widehat{\theta}|$ at significance level $\alpha = 0.1$ for three different choices of $\widehat{\theta}$. We follow this probability distribution $\widehat{\theta} \sim \mathsf{N}(\theta_0, \sigma^2 I_d)$ for three different σ values $\sigma = 1, 2$, and 3. A smaller value of σ implies a better estimation of θ_0. The average rejection rates are depicted in Figure 1(a),



(a) Average rejection rate of the null hypothesis of 1 for $\tau = 0$ and features with isotropic Gaussian distribution $x \sim \mathsf{N}(0, I_{10})$. In this experiment, we consider $y|x \sim \mathsf{N}(x^\mathsf{T}\theta^*, 1)$ for $\theta^* = (1, 1, 2, 2, 3, 3, 4, 4, 5, 5)$.



(b) Average rejection rate of the null hypothesis 1 for $\tau = 0$ and features coming from an isotropic Gaussian distribution $x \sim \mathsf{N}(0, I_{10})$. In this experiment, we consider $y|x \sim \mathsf{N}(x^\mathsf{T}Sx, 1)$ for a positive definite matrix $S_{i,j} = 1 + \mathbb{I}(i = j)$ (2 on diagonal and 1 on off-diagonal entries).

Figure 1: Average Rejection Rates for Different Settings

where each 10×10 square corresponds to a different σ value (three plots in total). Specifically, (i, j)-th cell in each plot denotes the average rejection rate of the symmetric influence hypothesis for features i and j. The rejection rates are obtained by averaging over 1000 independent experiments. First, it can be inferred that for pairs belonging to the set $\mathcal{I}$ the rejection rate is always smaller than the significance level $\alpha = 0.1$, thereby the size of the test is controlled. In addition, by decreasing the σ value (moving from right to left), it can be inferred that the test achieves higher power (more dark blue regions). It is consistent with our prior expectation that the statistical power of our method depends on the quality of the score function T and model estimate $\widehat{\theta}$; see Theorem 3.2. More on the statistical power of our method, it can be observed that within each plot, pairs that have higher contrast in the difference of coefficient magnitudes have higher statistical power. For instance, this pair of features $(1, 10)$ with coefficient values $\theta_1^* = 1, \theta_{10}^* = 5$ has rejection rates of $0.987, 0.768, 0.543$ (for $\sigma = 1, 2, 3$, respectively) while the other pair of features $(6, 8)$ with coefficient values $\theta_6^* = 3, \theta_8^* = 4$ has rejection rate of $0.294, 0.097, 0.055$ (for $\sigma = 1, 2, 3$, respectively).

5. Influence of Training Data on Output Model

In this section, we combine our closeness-of-influence test with datamodel framework (Ilyas et al., 2022) to analyze the influence of training samples on the evaluations of the trained model on certain target examples. We first provide a brief overview on datamodels and later describe the experiments setup.

5.1. Datamodels

For training samples $\mathcal{D}^{\mathsf{train}} = \{(x_i, y_i)\}_{i=1:N}$ consider a class of learning algorithm $\mathcal{A}$, where by class we mean a training mechanism (potentially randomized), such as training a fixed geometry of deep neural networks via gradient descent and a fixed random initialization scheme. In datamodels (Ilyas et al., 2022), a new learning problem is considered, where feature vectors S are binary 0-1 vectors with size N with $\gamma \in (0, 1)$ portion one entries, selected uniformly at random. Here S is an indicator vector for participation of N data points $\mathcal{D}^{\mathsf{train}}$ in the training mechanism, i.e,. $S_i = 1$ if and only if the $i-$th sample of $\mathcal{D}^{\mathsf{train}}$ is considered for the training purpose via $\mathcal{A}$. For a fixed target example x, the response value is the evaluation (will be described later) of the output model (trained with samples indicated in S) on x, denoted by $f_{\mathcal{A}}(x; S)$. This random sampling of data points from $\mathcal{D}^{\mathsf{train}}$ is repeated m times, therefore data for the new learning problem is $\{(S_i, f_{\mathcal{A}}(x, S_i))\}_{i=1:m}$. The ultimate goal of datamodels is to learn the mapping $S \to f_{\mathcal{A}}(x, S)$ via surrogate modeling and a class of much less complex models. In the seminal work of (Ilyas et al., 2022), they show that using linear regression with ℓ_1 penalty (LASSO (Tibshirani, 1996)) performs surprisingly well in learning the highly complex mapping of $S \to f_{\mathcal{A}}(x, S)$.

5.2. Motivation

We are specifically interested in analyzing the influence of different pairs of training samples on a variety of test targets, and discover pairs of training samples that with high cer-

Figure 2: Summary of discoveries on CIFAR-10 dataset via datamodels used with our closeness-of-influence test. For each pair of 10 classes (can be similar), we choose random samples from the training data along with a random target image from dog pictures in the test data, and we repeat this process 20 times. After running the Benjamini–Yekutieli procedure on output p-values (2000 in total) at $\alpha = 0.2$ three significant results are reported. The images of these findings are plotted above, with their associated p-values. This implies that with high certainty images in each pair influence the target example differently.

tainty influence the test target differently. We use the score function $(f_{\mathcal{A}}(x, S) - x^{\mathsf{T}}\widehat{\theta})^2$ for our closeness-of-influence test, where $\widehat{\theta}$ is the learned datamodel. We adopt this score function, mainly due to the promising performance of linear surrogate models in (Ilyas et al., 2022) for capturing the dependency rule between S and $f_{\mathcal{A}}(x; S)$. In addition, the described sampling scheme in datamodels satisfies the symmetric structure as per Proposition 2.6 (so $\tau_X = 0$). We would like to emphasize that despite the empirical success of datamodels, the interpretation of training samples with different coefficient magnitude in the obtained linear datamodel $\widehat{\theta}$ is *not* statistically accurate. Here we approach this problem through the lens of hypothesis testing and output p-values, to project the level of confidence in our findings.

5.3. Experimental Setups and Results

We consider the CIFAR-10 dataset (Krizhevsky et al., 2009), which has $N = 50000$ training samples along with 10000 test datapoints and 10 classes [1]. We consider $\gamma = 0.5$ (portion of ones in S_i samples), and follow the same heuristics provided for $f_{\mathcal{A}}(x; S)$ in (Ilyas et al., 2022), which is the correct-class margin, defined as the logit value of the true class minus the highest logit value among incorrect classes. We use the datamodel data given in `https://github.com/MadryLab/datamodels-data`. The provided data has $310k$ samplings, where for each target example x (in the test data) the datamodel parameter $\widehat{\theta} \in \mathbb{R}^N$ is estimated via the first $300k$ samples (10000 total number of datamodels $\widehat{\theta}$ for each test data). We use the additional $10k$ samples to run our closeness-of-fit test with the linear score function $(f_{\mathcal{A}}(x; S) - x^{\mathsf{T}}\widehat{\theta})^2$. Now, for each pair of training samples and a specific target test example, we can test for their closeness of influence. In the first experiment, for each two classes (can be the same) we choose two pictures as the training pair (randomly from the two classes), and for the target sample, we select randomly from the class of dog pictures. For each two classes, we repeat this process 20 times, and run our test 1 with $\tau = 0$, and report all p-values (2000 in total). After running the Benjamini–Yekutieli procedure (Benjamini & Yekutieli, 2001) (with log factor correction to control for dependency among p-values), we find three statistically significant results at $\alpha = 0.2$ with p-value=5×10^{-5} (for all three discoveries). Surprisingly, all three findings correspond to a similar test image, the pictures of training pairs and the one test image can be seen in Figure 2. It can be observed that in all findings one of the reported images is visually closer to the target image. This conforms well to obtained results that the null hypothesis 1 which states that the two training images have equal influence on the target sample is rejected. We refer to Appendix B for the rest of experiments.

6. Concluding Remarks

In this paper, we proposed a novel method to test the closeness of influence of a given pair of features on the response value. This procedure makes no assumption on the conditional law between the response value and features ($\mathcal{L}(Y|X)$). We first proposed a notion called "symmetric influence" that generalized the familiar concept of equal coefficient in parametric models. This notion is motivated to characterize the sensitivity of the conditional law with respect to swapping the features. We then formulated the closeness-of-influence testing problem as a tolerance hypothesis testing. We provide theoretical guarantees on type-I error rate. We then analyzed statistical power of our method for a general score function T, and show that for two specific learning problems i) linear regression settings, and 2) binary classification under a mixture of Gaussian models with a certain choice of score functions we can achieve full statistical power. Finally, we adopt the datamodel framework and use our closeness-of-influence test to find training samples that have different influence on the trained model.

Several interesting venues for future research are in order. In particular, extending this framework for multiple testing (testing for multiple number of pairs) and still achieving valid statistical results. This can be done with generic mul-

[1] airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck

tiple testing frameworks (similar to Benjamini–Yekutieli procedure used in Section 5) on the obtained p-values, but a method that is crafted for this setting can be more powerful. In addition, extending this framework for studying influence of a group of features (more that two) can be of great interest.

References

Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415, 1975.

Belloni, A., Chernozhukov, V., and Hansen, C. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pp. 1165–1188, 2001.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 175–197, 2020.

Bien, J., Yan, X., Simpson, L., and Müller, C. L. Tree-aggregated predictive modeling of microbiome data. *Scientific Reports*, 11(1):1–13, 2021.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9 (3):1103, 2015.

Candes, E. and Tao, T. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

Candes, E., Fan, Y., Janson, L., and Lv, J. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., et al. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–823, 2013.

Crawford, L., Wood, K. C., Zhou, X., and Mukherjee, S. Bayesian approximate kernel regression with variable selection. *Journal of the American Statistical Association*, 113(524):1710–1721, 2018.

Deshpande, Y., Javanmard, A., and Mehrabi, M. Online debiasing for adaptively collected high-dimensional data. *arXiv preprint arXiv:1911.01040*, 2019.

Duchi, J. Derivations for linear algebra and optimization. *Berkeley, California*, 3(1):2325–5870, 2007.

Fei, Z. and Li, Y. Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach. *J. Mach. Learn. Res.*, 22:58–1, 2021.

Hsieh, F. and Turnbull, B. W. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The annals of statistics*, 24(1):25–40, 1996.

Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., and Madry, A. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

Javanmard, A. and Mehrabi, M. Pearson chi-squared conditional randomization test. *arXiv preprint arXiv:2111.00027*, 2021.

Javanmard, A. and Montanari, A. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(2):277–293, 2022.

Peters, J. M., Colavin, A., Shi, H., Czarny, T. L., Larson, M. H., Wong, S., Hawkins, J. S., Lu, C. H., Koo, B.-M., Marta, E., et al. A comprehensive, crispr-based functional analysis of essential genes in bacteria. *Cell*, 165(6):1493–1506, 2016.

Shaer, S. and Romano, Y. Learning to increase the power of conditional randomization tests. *arXiv preprint arXiv:2207.01022*, 2022.

Shao, S., Bien, J., and Javanmard, A. Controlling the false split rate in tree-based aggregation. *arXiv preprint arXiv:2108.05350*, 2021.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Wilms, I. and Bien, J. Tree-based node aggregation in sparse graphical models. *Journal of Machine Learning Research*, 23(243):1–36, 2022.

Yan, X. and Bien, J. Rare feature selection in high dimensions. *Journal of the American Statistical Association*, 116(534):887–900, 2021.

A. Proof of Theorems and Technical Lemmas

A.1. Proof of Theorem 2.3

Consider two data points $z^{(1)} = (x^{(1)}, y^{(1)})$, $z^{(2)} = (x^{(2)}, y^{(2)})$ drawn i.i.d. from the density function $p_{X,Y}$. For two features i, j, define

$$\pi = \mathbb{P}\left(T\big(X^{(1)}, Y^{(1)}\big) \geq T\big(X^{(2)}_{\mathsf{swap}(i,j)}, Y^{(2)}\big)\right) .$$

We want to show that under the null hypothesis, the value π is concentrated around $1/2$ with maximum distance of τ_X. First, from the symmetry between two i.i.d. data points we have

$$\mathbb{P}\left(T\big(X^{(1)}, Y^{(1)}\big) \geq T\big(X^{(2)}, Y^{(2)}\big)\right) = 1/2 .$$

The underlying assumption is that in the case of equal values the tie is broken randomly. We introduce $\widetilde{z}^{(2)} = (x^{(2)}_{\mathsf{swap}(i,j)}, y^{(2)})$. This brings us

$$\begin{aligned}
\pi - \frac{1}{2} &= \mathbb{P}\left(T\big(X^{(2)}_{\mathsf{swap}(i,j)}, Y^{(2)}\big) \leq T(Z_1)\right) \\
&\quad - \mathbb{P}\left(T\big(X^{(2)}, Y^{(2)}\big) \leq T(Z_1)\right) \\
&= \mathbb{E}\left[\mathbb{P}(T(\widetilde{Z}^{(2)}) \leq T(Z^{(1)}) | Z^{(1)}, Y^{(2)})\right] \\
&\quad - \mathbb{E}\left[\mathbb{P}(T(Z^{(2)}) \leq T(Z^{(1)}) | Z^{(1)}, Y^{(2)})\right] .
\end{aligned}$$

In the next step, we let $T^{(1)} = T(Z^{(1)})$, $T^{(2)} = T(Z^{(2)})$, and $\widetilde{T}^{(2)} = T(\widetilde{Z}^{(2)})$. Then, by an application of Jenson's inequality we get

$$\left|\pi - \frac{1}{2}\right| \leq \mathbb{E}\Big[\big|\mathbb{P}(\widetilde{T}^{(2)} \leq T^{(1)} | Z^{(1)}, Y^{(2)}) - \mathbb{P}(T^{(2)} \leq T^{(1)} | Z^{(1)}, Y^{(2)})\big|\Big] \tag{6}$$

On the other hand, for some values $z \in \mathbb{R}^{d+1}, y \in \mathbb{R}$ consider the following measurable set:

$$A_{z,y} = \{x \in \mathbb{R}^d : T(x, y) \leq T(z)\} .$$

By using this definition of set $A_{z,y}$ in 6 and shorthands $W = (Z^{(1)}, Y^{(2)})$ we arrive at

$$\begin{aligned}
\left|\pi - \frac{1}{2}\right| &\leq \mathbb{E}\Big[\big|\mathbb{P}(X^2_{\mathsf{swap}(i,j)} \in A_W | W) - \mathbb{P}(X^{(2)} \in A_W | W)\big|\Big] \\
&\leq \mathbb{E}\Big[d_{\mathsf{TV}}\big(p_{X^{(2)}_{\mathsf{swap}(i,j)}|W}, p_{X^{(2)}|W}\big)\Big] ,
\end{aligned} \tag{7}$$

where the last inequality follows the definition of the total variation distance. Since $Z^{(1)}$ and $Z^{(2)}$ are independent random variables, we get that

$$\begin{aligned}
d_{\mathsf{TV}}\Big(p_{X^{(2)}_{\mathsf{swap}(i,j)}|W}, p_{X^{(2)}|W}\Big) &= d_{\mathsf{TV}}\Big(p_{X^{(2)}_{\mathsf{swap}(i,j)}|Y^{(2)}}, p_{X^{(2)}|Y^{(2)}}\Big) \\
&= d_{\mathsf{TV}}\big(p_{X_{\mathsf{swap}(i,j)}|Y}, p_{X|Y}\big) ,
\end{aligned}$$

where the last relation comes from the fact that random variable $(x, y) \sim p_{x,y}$ and $(x^{(2)}, y^{(2)})$ has a similar density function. Using the above relation in 7 yields

$$\begin{aligned}
\left|\pi - \frac{1}{2}\right| &\leq \mathbb{E}\left[d_{\mathsf{TV}}\Big(p_{X_{\mathsf{swap}(i,j)}|Y}, p_{X|Y}\Big)\right] \\
&= d_{\mathsf{TV}}\left(p_{X_{\mathsf{swap}(i,j)},Y}, p_{X,Y}\right) .
\end{aligned}$$

In the next step, for $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ let $p(x, y)$ and $q(x, y)$ respectively denote the density functions of $(X_{\mathsf{swap}(i,j)}, Y)$ and (X, Y). From the above relation we get

$$\left|\pi - \frac{1}{2}\right| \leq \frac{1}{2}\int |p(x, y) - q(x, y)| \, \mathrm{d}x\mathrm{d}y .$$

On the other hand, by rewriting the total variation distance of the joint random variables get

$$\begin{aligned}|p(x,y)-q(x,y)| &= |p(x)p(y|x)-q(x)q(y|x)| \\ &= |p(x)p(y|x)-p(x)q(y|x) \\ &\quad + p(x)q(y|x)-q(x)q(y|x)| \\ &\leq p(x)|p(y|x)-q(y|x)| \\ &\quad + |p(x)-q(x)|q(y|x)\,.\end{aligned}$$

Plugging this into the above relation yields

$$\begin{aligned}\left|\pi-\frac{1}{2}\right| &\leq \frac{1}{2}\int p(x)|p(y|x)-q(y|x)|\mathrm{d}x\mathrm{d}y \\ &\quad + \frac{1}{2}\int |p(x)-q(x)|q(y|x)\mathrm{d}x\mathrm{d}y\,.\end{aligned}$$

In the next step, by integration with respect to y we get

$$\begin{aligned}\left|\pi-\frac{1}{2}\right| &\leq \frac{1}{2}\int p(x)|p(y|x)-q(y|x)|\mathrm{d}x\mathrm{d}y \\ &\quad + \frac{1}{2}\int |p(x)-q(x)|\mathrm{d}x\,.\end{aligned}$$

This implies that

$$\left|\pi-\frac{1}{2}\right| \leq \mathbb{E}_X\left[d_{\mathsf{TV}}(p_{Y|X_{\mathsf{swap}(i,j)}}, p_{Y|X})\right] + d_{\mathsf{TV}}(p_X, p_{X_{\mathsf{swap}(i,j)}})\,.$$

Finally, under the null hypothesis 1 and the fact that $\tau_X \geq d_{\mathsf{TV}}(p_X, p_{X_{\mathsf{swap}(i,j)}})$ we get

$$\left|\pi-\frac{1}{2}\right| \leq \tau_X + \tau\,. \tag{8}$$

Any deviation from this range is accounted as evidence against the null hypothesis 1. In Algorithm 1, for each $1 \leq m \leq n/2$, it is easy to observe that each random variable $\mathbb{I}\left(T(X^{(m)}, Y^{(m)}) \geq T(\widetilde{X}^{(m)}, \widetilde{Y}^{(m)})\right)$ is a Bernoulli with success probability π. In the next step, by an application of Hoeffding's inequality for every $t \geq 0$ and sum of $n/2$ independent Bernoulli random variables we get

$$\mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^{n/2}\mathbb{I}\big(T(x_i,y_i) \leq T(\tilde{x}_i,\tilde{y}_i)\big) - \pi\right| \geq t\right) \leq 2\exp(-nt^2)\,.$$

Therefore, for statistics U_n as per Algorithm 1 we get

$$\mathbb{P}(|U_n-\pi| \geq t) \leq 2\exp(-nt^2)\,, \quad \forall t \geq 0 \tag{9}$$

We next consider $\delta \geq \tau + \tau_X$ and use triangle inequality to obtain

$$\begin{aligned}\mathbb{P}\left(\left|U_n-\frac{1}{2}\right| \geq \delta\right) &\leq \mathbb{P}\left(\left|U_n-\pi\right| + \left|\pi-\frac{1}{2}\right| \geq \delta\right) \\ &\leq \mathbb{P}(|U_n-\pi| \geq \delta-\tau-\tau_X) \\ &\leq 2\exp(-n(\delta-\tau-\tau_X)^2)\,.\end{aligned}$$

Where in the penultimate relation we used 8, and the last relation follows 9. By letting $\alpha = \delta - \tau - \tau_X$, we get

$$\mathbb{P}\left(\left|U_n-\frac{1}{2}\right| \geq \tau + \tau_X + \sqrt{\frac{\log\frac{2}{\alpha}}{n}}\right) \leq \alpha\,.$$

This completes the proof.

A.2. Proof of Proposition 2.2

We start with $\theta_i = \theta_j$, and we want to show that the symmetric influence property holds. We have

$$\begin{aligned} p_{Y|X_{\mathsf{swap}(i,j)}}(y|x) &= p_{Y|X}(y|x_{\mathsf{swap}(i,j)}) \\ &= \mathbb{P}(Y=1|X=x_{\mathsf{swap}(i,j)}) \\ &= \Big(1+\exp(-x_{\mathsf{swap}(i,j)}^{\mathsf{T}}\beta)\Big)^{-1} \\ &= \Big(1+\exp\big(-\beta_i x_j - \beta_j x_i - \sum_{\ell\neq i,j} x_\ell\beta_\ell\big)\Big)^{-1}. \end{aligned}$$

Using $\beta_i = \beta_j$ yields

$$\begin{aligned} p_{Y|X_{\mathsf{swap}(i,j}}(y|x) &= \Big(1+\exp\big(-\beta_j x_j - \beta_i x_i - \sum_{\ell\neq i,j} x_\ell\beta_\ell\big)\Big)^{-1} \\ &= \Big(1+\exp\big(-\sum_{\ell} x_\ell\beta_\ell\big)\Big)^{-1} \\ &= p_{Y|X}(y|x)\,. \end{aligned}$$

This completes the proof for the first part. For the other direction, suppose that the symmetric influence for i, j holds, thereby for every $x \in \mathbb{R}^d$ we have

$$\mathbb{P}(Y=+1|X_{\mathsf{swap}(i,j)}=x) = \mathbb{P}(Y=+1|X=x)\,.$$

By using $p_{Y|X_{\mathsf{swap}(i,j)}}(y|x) = p_{Y|X}(y|x_{\mathsf{swap}(i,j)})$ along with the logistic regression relation, we get

$$\begin{aligned} &\Big(1+\exp\big(-\beta_i x_j - \beta_j x_i - \sum_{\ell\neq i,j} x_\ell\beta_\ell\big)\Big)^{-1} \\ &= \Big(1+\exp\big(-\sum_{\ell} x_\ell\beta_\ell\big)\Big)^{-1}. \end{aligned}$$

In the next step, using the function $\log(\frac{u}{1-u})$ on the both sides, we get

$$\beta_i x_i + \beta_j x_j = \beta_i x_j + \beta_j x_i\,.$$

Since this must hold for all x_i, x_j values, we must have $\beta_i = \beta_j$. The proof for the linear regression setting follows the exact similar argument.

A.3. Proof of Proposition 2.5

Since x is a multivariate Gaussian, it means that its coordinates are jointly Gaussian random variables, therefore swapping the location of two coordinates i and j does not change the joint Gaussian property. On the other hand, from the linear transform $x_{\mathsf{swap}(i,j)} = P_{ij}x$ it is easy to arrive at $x_{\mathsf{swap}(i,j)} \sim \mathsf{N}(P_{ij}\mu, P_{ij}\Sigma P_{ij})$. We are only left with upper bounding the KL divergence of density functions $\mathsf{N}(\mu,\Sigma)$ and $\mathsf{N}(P_{ij}\mu, P_{ij}\Sigma P_{ij})$. For this end, we borrow a result from (Duchi, 2007) for kl-divergence of multivariate Gaussian distributions. Formally we have,

$$\begin{aligned} &d_{\mathsf{kl}}\left(\mathcal{N}(\mu_1,\Sigma_1)\|\mathcal{N}(\mu_2,\Sigma_2)\right) \\ &= \frac{1}{2}\left(\log\frac{\det\Sigma_2}{\det\Sigma_1} - d + \mathsf{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^{\mathsf{T}}\Sigma_2^{-1}(\mu_2-\mu_1)\right)\,. \end{aligned}$$

By replacing $\Sigma_2 = \Sigma$ and $\Sigma_1 = P_{ij}\Sigma P_{ij}$ along with the fact that $\det P_{ij} = -1$, we arrive at

$$d_{\mathsf{KL}}\left(\mathcal{L}(X_{\mathsf{swap}(i,j)})\|\mathcal{L}(X)\right) = \frac{1}{2}\left(-d + \mathsf{tr}(\Sigma^{-1}P_{ij}\Sigma P_{ij}) + (\mu - P_{ij}\mu)^{\mathsf{T}}\Sigma^{-1}(\mu - P_{ij}\mu)\right)\,.$$

Finally using Pinsker's inequality[2] completes the proof.

[2]For two denisty functions p, q this holds $d_{\mathsf{TV}}(p,q) \leq \sqrt{\frac{d_{\mathsf{kl}}(p\|q)}{2}}$

A.4. Proof of Proposition 2.6

In this setup, form the construction of the feature vector $x \in \{0,1\}^d$ it is easy to get that for every $\alpha \in \{0,1\}^d$ we have

$$\mathbb{P}(x = \alpha) = \frac{\mathbb{I}(|\alpha| = m)}{\binom{d}{m}} .$$

From this structure, since swapping the coordinates does not change the number of non-zero entries of the binary feature vector, we get $|\alpha| = |\alpha_{\mathsf{swap}(i,j)}|$. Thereby, we get

$$\mathbb{P}(x = \alpha) = \mathbb{P}(x_{\mathsf{swap}(i,j)} = \alpha) , \quad \forall \alpha \in \{0,1\}^d.$$

Therefore $d_{\mathsf{TV}}\left(\mathcal{L}(X), \mathcal{L}(X_{\mathsf{swap}(i,j)})\right) = 0$.

A.5. Proof of Theorem 3.1

Let

$$\pi = \mathbb{P}\left(T(X^{(1)}, Y^{(1)}) \geq T(X^{(2)}_{\mathsf{swap}(i,j)}, Y^{(2)})\right) .$$

For the sake of simplicity, we adopt the following shorthands: $T_1 = T(X^{(1)}, Y^{(1)})$ and $T_2 = T(X^{(2)}_{\mathsf{swap}(i,j)}, Y^{(2)})$. This gives us

$$\begin{aligned}
\pi &= \mathbb{P}(T_1 \geq T_2) \\
&= \mathbb{E}_{T_1}\left[\mathbb{P}(T_1 \geq T_2 | T_1)\right] \\
&= \mathbb{E}_{T_1}\left[G_T(T_1)\right] \\
&= \int G_T(t) \mathrm{d}F_T(t) \\
&= \int_0^1 G_T(F_T^{-1}(u)) \mathrm{d}u .
\end{aligned}$$

In the next step, we let $\delta = 2\exp(-n\beta^2)$, then by plugging this relation in the given condition in Theorem 3.1 we arrive at

$$|\pi - 1/2| \geq \delta + \tau + \tau_X + \sqrt{\frac{\log(2/\alpha)}{n}} . \tag{10}$$

We now focus on the decision rule 2. Let $\tau' = \tau + \tau_X$, then we get

$$\mathbb{P}(\Psi(\mathbf{X}, \mathbf{Y}) = 1) = \mathbb{P}\left(|U_n - 1/2| \geq \tau' + \sqrt{\frac{\log(2/\alpha)}{n}}\right) . \tag{11}$$

On the other hand, from triangle inequality we have $|U_n - 1/2| \geq |\pi - 1/2| - |U_n - \pi|$. Plugging this into 10 yields

$$|U_n - 1/2| \geq \delta + \tau' + \sqrt{\frac{\log(2/\alpha)}{n}} - |U_n - \pi|$$

Combining this with 11 gives us

$$\begin{aligned}
\mathbb{P}(\Psi(\mathbf{X}, \mathbf{Y}) = 1) &\geq \mathbb{P}\left(\delta \geq |U_n - \pi|\right) \\
&= 1 - \mathbb{P}\left(\delta \leq |U_n - \pi|\right) .
\end{aligned} \tag{12}$$

In the next step, we return to the given relation for U_n in Algorithm 1. From the definition of π, for each m we have

$$\mathbb{P}\left(T\big(X^{(m)}, Y^{(m)}\big) \leq T\big(\widetilde{X}^{(m)}, \widetilde{Y}^{(m)}\big)\right) = \pi .$$

Therefore by an application of the Hoeffding's inequality we get

$$\mathbb{P}\left(|U_n - \pi| \geq \delta\right) \leq \sqrt{\frac{\log(2/\delta)}{n}} .$$

Finally, recalling $\delta = 2\exp(-n\beta^2)$ yields

$$\mathbb{P}\left(|U_n - \pi| \geq \delta\right) \leq \beta .$$

Using this in 12 completes the proof. In this case, statistical power not smaller than $1 - \beta$ can be achieved.

A.6. Proof of Theorem 3.2

From the isotropic Gaussian distribution, we have $\tau_X = 0$. We next start by the ODC function $G_T \circ F_T^{-1}$. For this end, we start by the definition of F_T where for some non-negative t we have:

$$\begin{aligned} F_T(t) &= \mathbb{P}(|Y^{(1)} - \widehat{\theta}^\mathsf{T} X^{(1)}| \leq t) \\ &= \mathbb{P}(|(\theta^* - \widehat{\theta})^\mathsf{T} X^{(1)} + \varepsilon_1| \leq t) \\ &= \mathbb{P}(-t \leq (\theta^* - \widehat{\theta})^\mathsf{T} X^{(1)} + \varepsilon_1 \leq t)\,. \end{aligned}$$

On the other hand, we know that $x^\mathsf{T}(\widehat{\theta} - \theta^*) + \varepsilon$ has a Gaussian distribution $\mathsf{N}(0, \|\theta^* - \widehat{\theta}\|_2^2 + \sigma^2)$. This brings us

$$\begin{aligned} F_T(t) &= \mathbb{P}(-t \leq (\theta^* - \widehat{\theta})^\mathsf{T} X^{(1)} + \varepsilon_1 \leq t) \\ &= \mathbb{P}\left(\frac{-t}{\sqrt{\|\theta^* - \widehat{\theta}\|_2^2 + \sigma^2}} \leq \frac{(\theta^* - \widehat{\theta})^\mathsf{T} X^{(1)} + \varepsilon_1}{\sqrt{\|\widehat{\theta} - \theta^*\|_2^2 + \sigma^2}} \leq \frac{t}{\sqrt{\|\widehat{\theta} - \theta^*\|_2^2 + \sigma^2}}\right) \\ &= \Phi\left(\frac{t}{\sqrt{\sigma^2 + \|\theta^* - \widehat{\theta}\|_2^2}}\right) - \Phi\left(-\frac{t}{\sqrt{\sigma^2 + \|\theta^* - \widehat{\theta}\|_2^2}}\right) \\ &= 2\Phi\left(\frac{t}{\sqrt{\sigma^2 + \|\theta^* - \widehat{\theta}\|_2^2}}\right) - 1\,, \end{aligned} \tag{13}$$

where the last line comes from the fact that $\Phi(t) + \Phi(-t) = 1$ for every real value t. We introduce the shorthand $\widehat{\theta}_{\mathsf{swap}} = \widehat{\theta}_{\mathsf{swap}(i,j)}$, then by a similar argument we get

$$\begin{aligned} G_T(t) &= \mathbb{P}(|Y^{(2)} - \widehat{\theta}^\mathsf{T} X^{(2)}_{\mathsf{swap}(i,j)}| \leq t) \\ &= \mathbb{P}(|(\theta^* - \widehat{\theta}_{\mathsf{swap}})^\mathsf{T} X^{(2)} + \varepsilon_2| \leq t) \\ &= 2\Phi\left(\frac{t}{\sqrt{\sigma^2 + \|\theta^* - \widehat{\theta}_{\mathsf{swap}}\|_2^2}}\right) - 1\,, \end{aligned} \tag{14}$$

By combining 13 and 14 we get

$$F_T \circ G_T^{-1}(u) = 2\Phi\left(\frac{\sigma_2}{\sigma_1}\Phi^{-1}\Big(\frac{u+1}{2}\Big)\right) - 1\,,$$

for σ_2 and σ_1 given by

$$\sigma_1^2 = \sigma^2 + \|\theta^* - \widehat{\theta}\|_2^2\,, \quad \sigma_2^2 = \sigma^2 + \|\theta^* - \widehat{\theta}_{\mathsf{swap}}\|_2^2\,.$$

We consider $\gamma = \frac{\sigma_2}{\sigma_1}$. Plugging this into the power expression in Theorem 3.1 we arrive at

$$F_T(G_T^{-1}(u)) - u = 2\left[\Phi\Big(\gamma\Phi^{-1}\Big(\frac{u+1}{2}\Big)\Big) - \frac{u+1}{2}\right]\,.$$

In the next step, by using the change of variable $v = \frac{u+1}{2}$ we get

$$\int_0^1 [F_T(G_T^{-1}(u)) - u]\mathrm{d}u = 4\int_{\frac{1}{2}}^1 \left[\Phi\big(\gamma\Phi^{-1}(v)\big) - v\right]\mathrm{d}v\,.$$

We then introduce function $\psi : [0, +\infty] \to \mathbb{R}$ as following

$$\psi(\gamma) = 4\int_{\frac{1}{2}}^1 \Phi\Big(\gamma\Phi^{-1}(v)\Big)\mathrm{d}v\,.$$

This implies that

$$\psi(\gamma) - \psi(1) = \int_0^1 [F_T(G_T^{-1}(u)) - u]\mathrm{d}u\,. \tag{15}$$

By differentiating $\psi(.)$ with respect to γ in its original definition we obtain

$$\begin{aligned}\frac{\mathrm{d}\psi}{\mathrm{d}\gamma} &= 4\frac{\partial}{\partial\gamma}\int_{\frac{1}{2}}^{1}\Phi\Big(\gamma\Phi^{-1}(v)\Big)\mathrm{d}v \\ &= 4\int_{\frac{1}{2}}^{1}\Phi^{-1}(v)\varphi\Big(\gamma\Phi^{-1}(v)\Big)\mathrm{d}v\,.\end{aligned}$$

We next use $s=\Phi^{-1}(v)$ to arrive at the following

$$\begin{aligned}\frac{\mathrm{d}\psi}{\mathrm{d}\gamma} &= 4\int_{0}^{+\infty}s\varphi\Big(\gamma s\Big)\varphi(s)\mathrm{d}s \\ &= \frac{4}{2\pi}\int_{0}^{+\infty}s\exp\Big(-\frac{s^2}{2}(1+\gamma^2)\Big)\mathrm{d}s \\ &= \frac{2}{\pi(\gamma^2+1)}\int_{0}^{+\infty}s\exp(-s^2/2)\mathrm{d}s \\ &= \frac{2}{\pi(\gamma^2+1)}\,.\end{aligned}$$

Since the differentiation of ψ with respect to γ is provided above, we then can use this and obtain the closed form equation for $\psi(u)$. This indeed is given by

$$\psi(\gamma) = C + \frac{2}{\pi}\arctan(\gamma)\,,$$

For some constant value C. In order to find C, note that $\psi(1) = 4\int_{\frac{1}{2}}^{1} v\mathrm{d}v = \frac{3}{2}$. This brings us $\psi(\gamma) = 1+\frac{2}{\pi}\arctan(\gamma)$. Using this in 15 yields

$$\begin{aligned}\left|\int_0^1[F_T(G_T^{-1}(u))-u]\mathrm{d}u\right| &= |\psi(\gamma)-\psi(1)| \\ &= \frac{2}{\pi}\left|\arctan(\gamma)-\arctan(1)\right|\,.\end{aligned}$$

On the other hand, from the identity $\arctan(x)-\arctan(y)=\arctan\frac{x-y}{1+xy}$ we arrive at:

$$\begin{aligned}\left|\int_0^1[F_T(G_T^{-1}(u))-u]\mathrm{d}u\right| &= \frac{2}{\pi}\left|\arctan\left(\frac{\gamma-1}{1+\gamma}\right)\right| \\ &= \frac{2}{\pi}\arctan\left(\frac{|\gamma-1|}{1+\gamma}\right),\end{aligned}$$

where in the last relation we used $\arctan(|.|) = |\arctan(.)|$ (note that $\gamma\geq 0$). We next use $\gamma=\sigma_2/\sigma_1$ to get

$$\left|\int_0^1[F_T(G_T^{-1}(u))-u]\mathrm{d}u\right| = \frac{2}{\pi}\arctan\left(\frac{|\sigma_1-\sigma_2|}{\sigma_1+\sigma_2}\right)\,. \tag{16}$$

On the other hand, from $\sigma_1^2+\sigma_2^2\geq 2\sigma_1\sigma_2$ we get

$$\Delta_T = \frac{|\sigma_1-\sigma_2|}{|\sigma_1+\sigma_2|}\geq\frac{|\sigma_1^2-\sigma_2^2|}{2(\sigma_1^2+\sigma_2^2)}$$

We then use this with the definition of σ_1,σ_2 to get

$$\begin{aligned}\Delta_T &\geq \frac{1}{2}\frac{\left|\|\theta^*-\widehat{\theta}\|_2^2-\|\theta^*-\widehat{\theta}_{\mathsf{swap}}\|_2^2\right|}{2\sigma^2+\|\theta^*-\widehat{\theta}\|_2^2+\|\theta^*-\widehat{\theta}_{\mathsf{swap}}\|_2^2} \\ &= \frac{1}{2}\frac{\left|-2\widehat{\theta}^{\mathsf{T}}\theta^*+2\widehat{\theta}_{\mathsf{swap}}^{\mathsf{T}}\theta^*\right|}{2\sigma^2+2\|\theta^*\|_2^2+2\|\widehat{\theta}\|_2^2-2\widehat{\theta}^{\mathsf{T}}\theta^*-2\widehat{\theta}_{\mathsf{swap}}^{\mathsf{T}}\theta^*}\,,\end{aligned}$$

where we used $\|\theta\|_2 = \|\theta_{\mathsf{swap}}\|_2$. In the next step, since $\widehat{\theta}_{\mathsf{swap},\ell} = \widehat{\theta}_\ell$ for all $\ell \neq i, j$ we get

$$\Delta_T \geq \frac{1}{2} \frac{\left|-\widehat{\theta}_i\theta_i^* - \widehat{\theta}_j\theta_j^* + \widehat{\theta}_i\theta_j^* + \widehat{\theta}_j\theta_i^*\right|}{\sigma^2 + \|\theta^*\|_2^2 + \|\widehat{\theta}\|_2^2 - \widehat{\theta}^\mathsf{T}\theta^* - \widehat{\theta}_{\mathsf{swap}}^\mathsf{T}\theta^*}$$
$$= \frac{1}{2} \frac{\left|-\widehat{\theta}_i\theta_i^* - \widehat{\theta}_j\theta_j^* + \widehat{\theta}_i\theta_j^* + \widehat{\theta}_j\theta_i^*\right|}{\sigma^2 + \|\theta^* - \widehat{\theta}\|_2^2 + \widehat{\theta}^\mathsf{T}\theta^* - \widehat{\theta}_{\mathsf{swap}}^\mathsf{T}\theta^*}$$

In the next step, by using the observation that $\widehat{\theta}_{\mathsf{swap},\ell} = \widehat{\theta}_\ell$ for all $\ell \neq i, j$ another time we get

$$\Delta_T \geq \frac{1}{2} \frac{|\widehat{\theta}_i - \widehat{\theta}_i||\theta_i^* - \theta_j^*|}{\sigma^2 + \|\theta^* - \widehat{\theta}\|^2 + (\theta_i^* - \theta_j^*)(\widehat{\theta}_i - \widehat{\theta}_j)}$$

Thereby we get

$$\Delta_T \geq \frac{1}{2} \frac{|\widehat{\theta}_i - \widehat{\theta}_i||\theta_i^* - \theta_j^*|}{\sigma^2 + \|\theta^* - \widehat{\theta}\|^2 + |\theta_i^* - \theta_j^*||\widehat{\theta}_i - \widehat{\theta}_j|}$$

Using the above relation in 16 we get

$$\left|\int_0^1 [F_T(G_T^{-1}(u)) - u]\mathrm{d}u\right| \geq \frac{2}{\pi} \arctan\left(\frac{1}{2} \frac{|\widehat{\theta}_i - \widehat{\theta}_i||\theta_i^* - \theta_j^*|}{\sigma^2 + \|\theta^* - \widehat{\theta}\|^2 + |\theta_i^* - \theta_j^*||\widehat{\theta}_i - \widehat{\theta}_j|}\right) \tag{17}$$

By recalling the given condition in Theorem 3.2 we have

$$|\theta_i^* - \theta_j^*| \geq \frac{2\tan(\frac{\pi}{2}(\rho_n(\alpha,\beta,\tau)))}{1 - 2\tan(\frac{\pi}{2}(\rho_n(\alpha,\beta,\tau)))} \frac{\left(\sigma^2 + \|\widehat{\theta} - \theta^*\|_2^2\right)}{|\widehat{\theta}_i - \widehat{\theta}_j|},$$

By using $\tan(\frac{\pi}{2}(\rho_n(\alpha,\beta,\tau))) \leq \frac{1}{2}$ in the above relation we get

$$\frac{2}{\pi} \arctan\left(\frac{1}{2} \frac{|\widehat{\theta}_i - \widehat{\theta}_i||\theta_i^* - \theta_j^*|}{\sigma^2 + \|\theta^* - \widehat{\theta}\|^2 + |\theta_i^* - \theta_j^*||\widehat{\theta}_i - \widehat{\theta}_j|}\right) \geq \rho_n(\alpha,\beta,\tau). \tag{18}$$

By combining 17 and 18 we get

$$\left|\int_0^1 [F_T(G_T^{-1}(u)) - u]\mathrm{d}u\right| \geq \rho_n(\alpha,\beta,\tau).$$

Finally using Theorem 3.2 completes the proof,

A.7. Proof of Theorem 3.3

We first show that in this case, ($\tau = 0$) for mixture of Gaussians, under the null hypothesis, we have $\tau_X = 0$. For this end, from the Bayes' formula it is easy to get $\mathcal{L}(Y|X) = \mathsf{Bern}(g(x,\mu))$ with

$$g(x,\mu) = \frac{1}{1 + \frac{1-q}{q} e^{-x^\mathsf{T}\mu}}.$$

With a similar argument, it can be observed that

$$\mathcal{L}(Y|X) = \mathsf{Bern}(g(x, \mu_{\mathsf{swap}(i,j)})).$$

Given that $d_{\mathsf{TV}}(\mathsf{Bern}(a), \mathsf{Bern}(b)) = |a - b|$, under the null hypothesis (with $\tau = 0$) we must have $g(x,\mu) = g(x, \mu_{\mathsf{swap}(i,j)})$ almost surely for all x values. This implies that $x^\mathsf{T}\mu = x^\mathsf{T}\mu_{\mathsf{swap}(i,j)}$ almost surely, thereby we have $\mu_i = \mu_j$. In the next step, we show that if $\mu_i = \mu_j$ then $\tau_X = 0$. We then note that

$$\mathcal{L}(X) = q\mathsf{N}(+\mu, I_d) + (1-q)\mathsf{N}(-\mu, I_d),$$
$$\mathcal{L}(X_{\mathsf{swap}(i,j)}) = q\mathsf{N}(+\mu_{\mathsf{swap}(i,j)}, I_d) + (1-q)\mathsf{N}(-\mu_{\mathsf{swap}(i,j)}, I_d).$$

In the next step, using $\mu_i = \mu_j$ we realize that $\mu_{\mathsf{swap}(i,j)} = \mu$, therefore $\mathcal{L}(X) = \mathcal{L}(X_{\mathsf{swap}(i,j)})$. This implies that $\tau_X = 0$.

For the rest of the proof, we follow a similar argument as per proof of Theorem 3.2 and we first characterize cdf functions F_T and G_T. In this case we have

$$\begin{aligned} F_T(t) &= \mathbb{P}(Y^{(1)}\widehat{\theta}^\mathsf{T} X^{(1)} \leq t) \\ &= q\mathbb{P}(\widehat{\theta}^\mathsf{T} X^{(1)} \leq t | Y^{(1)} = +1) + \\ &+ (1-q)\mathbb{P}(-\widehat{\theta}^\mathsf{T} X^{(1)} \leq t | Y^{(1)} = -1) \\ &= q\mathbb{P}(Z^+ \leq t) + (1-q)\mathbb{P}(Z^- \leq t)\,, \end{aligned}$$

where $Z_+ \sim \mathsf{N}(\mu^\mathsf{T}\widehat{\theta}, \|\widehat{\theta}\|_2^2)$ and $Z_- \sim \mathsf{N}(-\mu^\mathsf{T}\widehat{\theta}, \|\widehat{\theta}\|_2^2)$. This yields

$$\begin{aligned} F_T(t) &= q\Phi\Big(\frac{t - \widehat{\theta}^\mathsf{T}\mu}{\|\widehat{\theta}\|_2}\Big) + (1-q)\left(1 - \Phi\Big(\frac{-t + \widehat{\theta}^\mathsf{T}\mu}{\|\widehat{\theta}\|_2}\Big)\right) \\ &= \Phi\left(\frac{t - \widehat{\theta}^\mathsf{T}\mu}{\|\widehat{\theta}\|_2}\right)\,, \end{aligned}$$

where in the last line we used $\Phi(t) + \Phi(-t) = 1$. We next introduce the shorthands $\widehat{\theta}_{\mathsf{swap}} = \widehat{\theta}_{\mathsf{swap}(i,j)}$ and $\mu_{\mathsf{swap}} = \mu_{\mathsf{swap}(i,j)}$, then by a similar argument we arrive at

$$G_T(t) = \Phi\left(\frac{t - \widehat{\theta}_{\mathsf{swap}}^\mathsf{T}\mu}{\|\widehat{\theta}_{\mathsf{swap}}\|_2}\right)$$

Since $\widehat{\theta}_{\mathsf{swap}}^\mathsf{T}\mu = \mu_{\mathsf{swap}}^\mathsf{T}\widehat{\theta}$ and $\|\widehat{\theta}_{\mathsf{swap}}\| = \|\widehat{\theta}\|$ the expression for $G_T(t)$ can be written as the following:

$$G_T(t) = \Phi\left(\frac{t - \widehat{\theta}^\mathsf{T}\mu_{\mathsf{swap}}}{\|\widehat{\theta}\|_2}\right)$$

In the next step, it is easy to compute the quantile function $G_T^{-1}(u) = \|\widehat{\theta}\|_2\Phi^{-1}(u) + \widehat{\theta}^\mathsf{T}\mu_{\mathsf{swap}}$. This brings us

$$F_T(G_T^{-1}(u)) = \Phi\left(\Phi^{-1}(u) + \frac{\widehat{\theta}^\mathsf{T}(\mu_{\mathsf{swap}} - \mu)}{\|\widehat{\theta}\|_2}\right)\,.$$

By introducing $\lambda = \frac{\widehat{\theta}^\mathsf{T}(\mu_{\mathsf{swap}} - \mu)}{\|\widehat{\theta}\|_2}$ and the function $\rho(\lambda) = \int_0^1 \Phi(\Phi^{-1}(u) + \lambda)\mathrm{d}u$ we obtain

$$\int_0^1 F_T(G_T^{-1}(u))\mathrm{d}u = \rho(\lambda)\,.$$

On the other hand, by differentiating $\rho(\lambda)$ with respect to λ we get

$$\begin{aligned} \frac{\partial\rho}{\partial\lambda} &= \frac{\partial}{\partial\lambda}\int_0^1 \Phi(\Phi^{-1}(u) + \lambda)\mathrm{d}u \\ &= \int_0^1 \varphi(\Phi^{-1}(u) + \lambda)\mathrm{d}u\,. \end{aligned}$$

In the next step, by using the change of variable $s = \Phi^{-1}(u)$ we get that

$$\begin{aligned} \frac{\partial\rho}{\partial\lambda} &= \int_{-\infty}^{\infty} \varphi(s+\lambda)\varphi(s)\mathrm{d}s \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-\frac{(s+\lambda)^2}{2} - \frac{s^2}{2}\right)\mathrm{d}s \\ &= \frac{\exp(-\lambda^2/4)}{2\pi}\int_{-\infty}^{+\infty} \exp\left(-\frac{(\sqrt{2}s + \lambda/\sqrt{2})^2}{2}\right)\mathrm{d}s \\ &= \frac{\exp(-\lambda^2/4)}{2\sqrt{2}\pi}\int_{-\infty}^{+\infty} \exp\left(-\frac{(t + \lambda/\sqrt{2})^2}{2}\right)\mathrm{d}t = \frac{\exp(-\lambda^2/4)}{2\sqrt{\pi}}\,. \end{aligned}$$

Therefore we get $\rho(\lambda) = \rho(0) + \int_0^\lambda \frac{\exp(-s^2/4)}{2\sqrt{\pi}} = \rho(0) + \Phi(\frac{\lambda}{\sqrt{2}}) - \frac{1}{2}$, Since $\rho(0) = 1/2$, we arrive at $\rho(\lambda) = \Phi\left(\frac{\lambda}{\sqrt{2}}\right)$. Next from the definition of $\rho(\lambda)$ we have

$$\int_0^1 \Big[F_T(G_T^{-1}(u)) - u\Big]\mathrm{d}u = \rho(\lambda) - \rho(0)\,.$$

In the next step, we use the equivalent value of λ in the function $\rho(\lambda)$ to get

$$\int_0^1 \Big[F_T(G_T^{-1}(u)) - u\Big]\mathrm{d}u = \Phi\left(\frac{\widehat{\theta}^\mathsf{T}(\mu_{\mathsf{swap}} - \mu)}{\sqrt{2}\|\widehat{\theta}\|_2}\right) - \Phi(0)\,.$$

Therefore we get

$$\left|\int_0^1 \Big[F_T(G_T^{-1}(u)) - u\Big]\mathrm{d}u\right| = \left|\Phi\left(\frac{\widehat{\theta}^\mathsf{T}(\mu_{\mathsf{swap}} - \mu)}{\sqrt{2}\|\widehat{\theta}\|_2}\right) - \Phi(0)\right|\,.$$

On the other hand, the normal cdf satisfies the following property

$$\left|\Phi(t) - \frac{1}{2}\right| = \Phi(|t|) - \frac{1}{2}\,, \forall t \in \mathbb{R}$$

By using this we get

$$\left|\int_0^1 \Big[F_T(G_T^{-1}(u)) - u\Big]\mathrm{d}u\right| = \Phi\left(\left|\frac{\widehat{\theta}^\mathsf{T}(\mu_{\mathsf{swap}} - \mu)}{\sqrt{2}\|\widehat{\theta}\|_2}\right|\right) - \frac{1}{2}\,. \tag{19}$$

In the next step, by using the fact that $\mu_{\mathsf{swap},\ell} = \mu_\ell$ for $\ell \neq i, j$ we get that

$$\begin{aligned}\widehat{\theta}^\mathsf{T}(\mu_{\mathsf{swap}} - \mu) &= \widehat{\theta}_i(\mu_{\mathsf{swap},i} - \mu_i) + \widehat{\theta}_j(\mu_{\mathsf{swap},j} - \mu_j)\\ &= \widehat{\theta}_i(\mu_j - \mu_i) + \widehat{\theta}_j(\mu_i - \mu_j)\\ &= -(\widehat{\theta}_i - \widehat{\theta}_j)(\mu_i - \mu_j)\,.\end{aligned}$$

Using this in 19 yields

$$\left|\int_0^1 \Big[F_T(G_T^{-1}(u)) - u\Big]\mathrm{d}u\right| = \Phi\left(\frac{|(\widehat{\theta}_i - \widehat{\theta}_j)(\mu_i - \mu_j)|}{\sqrt{2}\|\widehat{\theta}\|_2}\right) - \frac{1}{2} \tag{20}$$

On the other hand, by recalling the condition on $|\mu_i - \mu_j|$ from Theorem 3.3 we have

$$|\mu_i - \mu_j| \geq \Phi^{-1}\left(\rho_n(\alpha, \beta, 0) + 2\Phi\left(\frac{|\mu_i - \mu_j|}{\sqrt{2}}\right) - \frac{1}{2}\right)\frac{\sqrt{2}\|\widehat{\theta}\|_2}{|\widehat{\theta}_i - \widehat{\theta}_j|} \tag{21}$$

Combining 20 and 21 yields

$$\left|\int_0^1 \Big[F_T(G_T^{-1}(u)) - u\Big]\mathrm{d}u\right| \geq \rho_n(\alpha, \beta, 0)\,.$$

Finally, using Theorem 3.1 completes the proof.

B. Additional Numerical Experiments

B.1. Size of the test (full experiments)

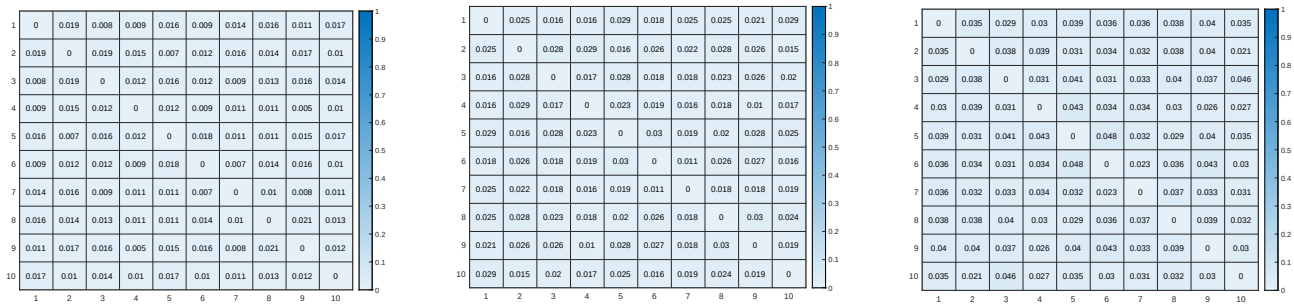
We refer to Figure 3 for experiment on the size of the test.

B.2. Power of the test (full experiments)

We refer to Figure 4 for experiment on power of the test.

B.3. binary classification under mixture of Gaussians

In this section, we consider the problem of testing for symmetric influence for binary classification under a mixture of Gaussian model. We consider the data generative law 5 with $q = 1/2$ and feature dimension $d = 10$. We consider $\widetilde{\mu} = [1, 2, 3, \ldots, 10]$ and let $\mu = \frac{\widetilde{\mu}}{\|\widetilde{\mu}\|_2}$. We follow the score function given in Theorem 3.3 and consider $T(x, y) = y\widehat{\theta}^\mathsf{T}x$ for some $\widehat{\theta} \sim \mathsf{N}(0, I_d)$. We consider three different number of samples $n = 5000, 20000, 50000$ for this experiment. Figure 5 denote the results. Each number is averaged over 1000 independent experiments. It can be observed that pairs with higher contrast between their μ values are rejected more often.



(a) $\alpha = 0.1$ (b) $\alpha = 0.15$ (c) $\alpha = 0.2$

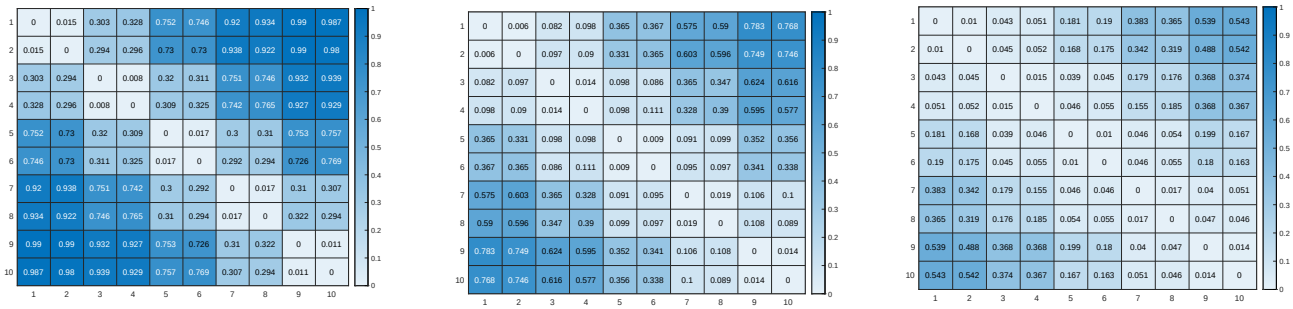
Figure 3: Average rejection rate of the null hypothesis 1 for $\tau = 0$ and features coming from an isotropic Gaussian distribution $x \sim \mathsf{N}(0, I_{10})$. In this experiment, we consider $y|x \sim \mathsf{N}(x^\mathsf{T} S x, 1)$ for a positive definite matrix $S_{i,j} = 1 + \mathbb{I}(i = j)$ (2 on diagonal and 1 on off-diagonal entries). The structure of S implies that the symmetric influence holds for every pair of features. We consider three significance levels $\alpha = 0.1, 0.15, 0.2$ (from left to right). The small cell (i, j) in each plot, represents rejection rates for testing symmetric influence for features i and j. In this experiment, the number of data points is 1000 and the method is run with the score function $T(x, y) = |y - x^\mathsf{T}\widehat{\theta}|$ for $\widehat{\theta} \sim \mathsf{N}(0, I_{10})$. The reported numbers are averaged over 1000 experiments. It can be seen that the size of the test is controlled at the pre-determined significance levels.

Train pair	Target 1	Target 2	Target 3	Target 4	Target 5
	pval=0.0023	pval=0.0026	pval=0.0092	pval=0.0173	pval=0.2108
	pval=0.0023	pval=0.0026	pval=0.0084	pval=0.0188	pval=0.2108

Table 2. Verifying the robustness of our findings for two pairs of training samples, that are highly close to each other.

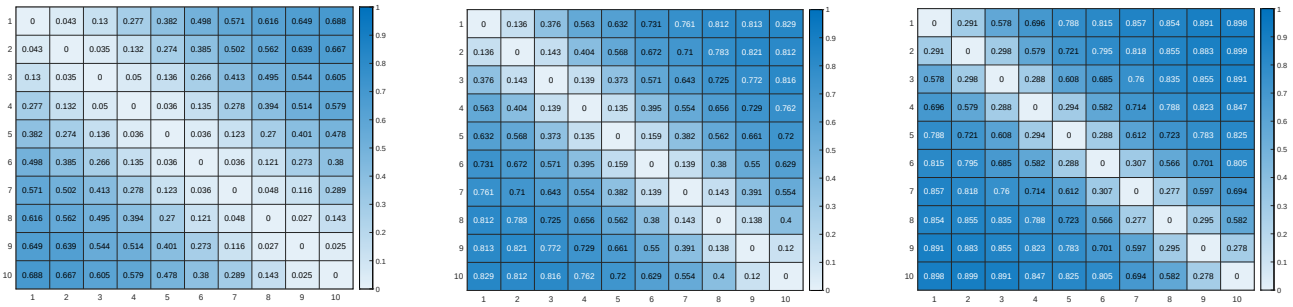
B.4. robustness of data models experiment

In the second experiment, we consider a pair of training samples with 5 target examples. The first four targets are statistically significant (at level $\alpha = 0.05$), while the target 5 gives $pval = 0.21$. We then replace the two training samples with some of their close other pictures, and compute the p-values for the new pair of images. We can see that the obtained p-values are somewhat close to the previous examples, which indicates the robustness of output results. The images along with p-values can be seen in Table 2.



(a) $\sigma = 1$ (b) $\sigma = 2$ (c) $\sigma = 3$

Figure 4: Average rejection rate of the null hypothesis of 1 for $\tau = 0$ and features with isotropic Gaussian distribution $x \sim \mathsf{N}(0, I_{10})$. In this experiment, we consider $y|x \sim \mathsf{N}(x^\mathsf{T}\theta^*, 1)$ for $\theta^* = [1, 1, 2, 2, 3, 3, 4, 4, 5, 5]$. In this experiment the symmetric influence holds for pairs of features $(1, 2), (3, 4), (5, 6), (7, 8)$, and $(9, 10)$. The small cell (i, j) in each plot, represents rejection rates for testing symmetric influence for features i and j at significance level $\alpha = 0.1$. In this experiment, the number of data points is 1000 and the method is run with the score function $T(x, y) = |y - x^\mathsf{T}\widehat{\theta}|$ for $\widehat{\theta} \sim \mathsf{N}(\theta^*, \sigma^2 I_{10})$ for three different σ values $\sigma = 1, 2, 3$ (from left to right). The reported numbers are averaged over 1000 experiments.



(a) $n = 5000$ (b) $n = 20000$ (c) $n = 50000$

Figure 5: Average rejection rate of the null hypothesis of 1 for $\tau = 0$ and features with isotropic Gaussian distribution $x \sim \mathsf{N}(0, I_{10})$ In this experiment, we consider binary classification under the mixture of Gaussian model 5 for $q = 1/2$ and $\mu = \frac{\mu}{\|\widetilde{\mu}\|_2}$ for $\widetilde{\mu} = [1, 2, \ldots, 10]$. The small cell (i, j) in each plot, represents rejection rates for testing symmetric influence for features i and j at significance level $\alpha = 0.1$. In this experiment, three different values for number of data points is considered $n = 5000, 20000, 50000$ (from left to right). We run Algorithm 1 with the score function $T(x, y) = yx^\mathsf{T}\widehat{\theta}$ for $\widehat{\theta} \sim \mathsf{N}(0, I_{10})$. The reported numbers are averaged over 1000 experiments.