WHEN MACHINE LEARNING GETS PERSONAL: EVALUATING PREDICTION AND EXPLANATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In high-stakes domains like healthcare, users often expect that sharing personal information with machine learning systems will yield tangible benefits, such as more accurate diagnoses and clearer explanations of contributing factors. However, the validity of this assumption remains largely unexplored. We propose a unified framework to quantify how *personalizing a model* influences both prediction and explanation. We show that its impacts on prediction and explanation can diverge: a model may become more or less explainable even when prediction is unchanged. For practical settings, we study a standard hypothesis test for detecting personalization effects on demographic groups. We derive a finite-sample lower bound on its probability of error as a function of group sizes, number of personal attributes, and desired benefit from personalization. This provides actionable insights, such as which dataset characteristics are necessary to test an effect, or the maximum effect that can be tested given a dataset. We apply our framework to real-world datasets, uncovering scenarios where effects are fundamentally untestable due to the dataset statistics. Our results highlight the need for joint evaluation of prediction and explanation in personalized models and the importance of designing models and datasets with sufficient information for such evaluation.

1 INTRODUCTION

In critical domains like healthcare and education, machine learning models are increasingly personalized by incorporating input attributes that encode personal characteristics. These attributes can be sensitive and linked to historical bias, such as sex or race, or costly, for example requiring expert-administered medical assessments.When users provide personal attributes to a model, they implicitly expect improved predictions, but does personalization consistently meet that expectation?

Personalization can indeed enhance predictive accuracy. For instance, cardiovascular risk prediction models often perform better when including sex (Paulus et al., 2016; Huang et al., 2024; Mosca et al., 2011) and race (Paulus et al., 2018). This is because men, women, and different racial groups exhibit different heart disease patterns. For example, hypertension is more common in African American populations (Flack et al., 2003). Hence, personalization enhances clinical predictions by capturing meaningful biological and sociocultural variation.

However, personalization can also pose risks. Including sensitive attributes such as race, gender, or age can amplify biases in machine learning and perpetuate damaging inequality. For example, Obermeyer et al. (2019) showed that a health algorithm relying on health care costs, an attribute shaped by racial inequities, systematically underestimated illness in Black patients compared to equally sick white patients. This reduced their access to extra care by over half.

Generally, personalization may benefit overall accuracy while harming specific groups, making such risks harder to detect. In sleep apnea classification, adding age and sex improved overall performance but increased errors for older women and younger men (Suriyakumar et al., 2023). Similar group disparities have been observed in explainable machine learning, where some users receive less faithful or reliable explanations than others (Balagopalan et al., 2022; Dai et al., 2022). However, these studies did not examine whether personalization itself contributes to explanation disparities, making it critical to assess whether model personalization may reduce explanation quality for some users. Hence, before personalizing a model, practitioners must consider if it delivers consistent gains across demographic groups in both prediction and explanation—see Fig. 1.

This showcases the need for a quantitative framework to rigorously assess the benefits and risks of personalization. We focus on two key goals of machine learning models in high-stakes settings like healthcare: *(i)* making accurate predictions and *(ii)* providing explanations for them. Our central question is: *how reliably can we evaluate whether personalization improves prediction accuracy and explanation quality, both overall and across groups?*

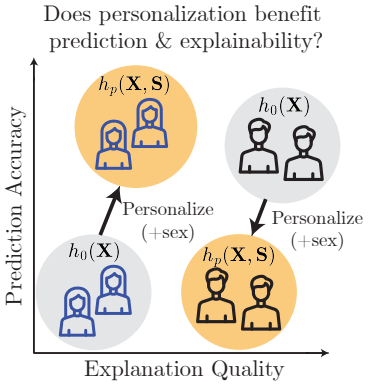

Figure 1: Impact of personalization on prediction and explanation: some groups benefit, others are harmed. h_0 is a generic model, h_p is a personalized model that takes an additional group attribute, $\mathbf{S}$.

Contributions. We propose a comprehensive study of the impact of personalization for prediction accuracy and explanation quality in machine learning models. Specifically:

1. We show that even when personalization does not improve prediction, it can enhance or degrade explainability, highlighting the need evaluate both independently in settings where accuracy and interpretability are critical (Section 4).
2. We derive distribution-aware limits on when personalization cannot be reliably tested, showing how many attributes or samples are needed in finite datasets. Our theory extends prior work beyond binary classification to general supervised learning, revealing key differences between evaluating prediction and explanation in classification versus regression (Section 5).
3. We apply our proposed framework on classification and regression tasks, illustrating how group-level gains from personalization are fundamentally untestable, thereby precluding statistical justification across different real-world scenarios (Section 6).

Overall, we offer a cautionary perspective on the promise of personalized medicine and the personalization of machine learning in other critical domains. Even when personalizing a machine learning model could be beneficial, it might be impossible to reliably prove it—thus limiting its practical use.

2 Related Works

Studies that investigate how personalizing machine learning models influences group outcomes (Suriyakumar et al., 2023) are limited to a narrow subset of performance measures and do not address explanation quality as described next. Extended related works are in Appendix A.

Theory. Few works theoretically characterize the impact of personalization. Monteiro Paes et al. (2022) define the Benefit of Personalization (BoP) as the minimum performance gain any group can expect. While the *definition* applies to any supervised learning task and "performance" measure, the theory supporting its use is confined to binary performance measures, such as accuracy in binary classification (0/1 loss) or false negative and positive rates (Bernoulli variables). Hence, it does not extend to continuous metrics like regression accuracy or explanation quality for regression and fails to provide a complete framework. Moreover, the theorems make unrealistic assumptions about dataset statistics (e.g., demographic groups of equal size) that further restrict their applicability in real-world settings. The general impact of personalization therefore remains theoretically uncharacterized.

Empirical Evidence. While the impact of personalization on explanation quality has never been measured, a few empirical studies have evaluated the fairness of explanations. Specifically, Balagopalan et al. (2022) train a human-interpretable model to imitate the behavior of a blackbox model, and characterize *fidelity* as how well it matches the blackbox model predictions. They found that the quality and reliability of explanations vary across different groups, but their experiments are restricted to binary classifiers, and to fidelity as the only explanation method. By contrast, Dai et al. (2022) evaluate various post hoc explanation methods across different evaluation metrics. They show that explanations can vary in quality across demographic groups, leading to fairness concerns, though their experiments are also restricted to binary classifiers. Neither work considers regression tasks or examines how personalization would affect differences in explanation quality across groups. These constraints limit the practical relevance of existing empirical results, as real-world scenarios do not always align with such settings.

Table 1: **Costs of model h for group s used to evaluate the impact of personalization** on data $(\tilde{\mathbf{X}}, \mathbf{Y})$ where $\tilde{\mathbf{X}} = \mathbf{X}$ for a generic model h_0, $\tilde{\mathbf{X}} = (\mathbf{X}, \mathbf{S})$ for a personalized model h_p, while $\mathbf{X}_{\setminus J}$ denotes the input when removing the most important features and $\mathbf{X}_J$ is its complement (see Section 4). Personalization benefits group $s \in \mathcal{S}$ if $C(h_0, s) - C(h_p, s) > 0$ and harms if $C(h_0, s) - C(h_p, s) < 0$. Incomprehensiveness is abbreviated as Incomp.

	$C(h, s)$	**Classification**	**Regression**
Predict	**Loss**	$\Pr(h(\tilde{\mathbf{X}}) \neq \mathbf{Y} \mid \mathbf{S} = s)$	$\mathbb{E}\left[\|h(\tilde{\mathbf{X}}) - \mathbf{Y}\|^2 \mid \mathbf{S} = s\right]$
	Evaluation metric	$-\text{AUC}(h, \mathbf{X}, \mathbf{Y} \mid \mathbf{S} = s)$	$-R^2(h, \mathbf{X}, \mathbf{Y} \mid \mathbf{S} = s)$
Explain	**Sufficiency**	$\Pr(h(\tilde{\mathbf{X}}) \neq h(\tilde{\mathbf{X}}_J) \mid \mathbf{S} = s)$	$\mathbb{E}\left[\|h(\tilde{\mathbf{X}}) - h(\tilde{\mathbf{X}}_J)\|^2 \mid \mathbf{S} = s\right]$
	Incomp.	$-\Pr\left(h(\tilde{\mathbf{X}}) \neq h(\tilde{\mathbf{X}}_{\setminus J}) \mid \mathbf{S} = s\right)$	$-\mathbb{E}\left[\|h(\tilde{\mathbf{X}}) - h(\tilde{\mathbf{X}}_{\setminus J})\|^2 \mid \mathbf{S} = s\right]$

Link to Fairness. Fairness in machine learning aims to mitigate biased outcomes affecting individuals or groups (Mehrabi et al., 2022). Past works have defined individual fairness, which seeks similar performance for similar individuals (Dwork et al., 2011), or group fairness (Dwork & Ilvento, 2019; Hardt et al., 2016), which seeks similar performance across different groups. Within this literature, most methods, metrics, and analyses are intended for classification tasks (Pessach & Shmueli, 2022). As for the fair regression literature, authors focus on designing fair learning methods (Hebert-Johnson et al., 2018; Berk et al., 2017; Fukuchi et al., 2013; Pérez-Suay et al., 2017; Calders et al., 2013), such as multicalbration, or defining fairness criteria for regression tasks (Gursoy & Kakadiaris, 2022; Agarwal et al., 2019). By contrast, our approach does not require equal performance across individuals or groups. Instead, we study a relaxed fairness notion: ensuring that no group is systematically harmed by personalization. We propose a framework to evaluate whether this weaker fairness criterion is satisfied, both theoretically and empirically, rather than proposing corrective algorithms.

3 Background: Benefit of Personalization Framework

Let $\mathcal{X}, \mathcal{S}, \mathcal{Y}$ denote, respectively, the input feature, group attribute, and outcome spaces. A *personalized model* $h_p : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ aims to predict an outcome variable $y \in \mathcal{Y}$ using both an input feature vector $x \in \mathcal{X}$ and a vector of group attributes $s \in \mathcal{S}$. In contrast, a *generic model* $h_0 : \mathcal{X} \to \mathcal{Y}$ does not use group attributes. We consider that a fixed data distribution $P = P_{\mathbf{X},\mathbf{S},\mathbf{Y}}$ is given, and that h_0 and h_p are trained to minimize a loss over a training dataset $\mathcal{D}_{train}$.

Cost. We first evaluate how a model h (generic or personalized) performs for a given group.

Definition 3.1 (Expected Group Cost)**.** The expected cost of model h for the group $s \in \mathcal{S}$ as measured by the cost function cost is defined as: $C(h, s) \triangleq \mathbb{E}_P[\text{cost}(h, \tilde{\mathbf{X}}, \mathbf{Y}) \mid \mathbf{S} = s]$, where $\tilde{\mathbf{X}} = \mathbf{X}$ for a generic model h_0, and $\tilde{\mathbf{X}} = (\mathbf{X}, \mathbf{S})$ for a personalized model h_p.

In what follows, we use cost and expected cost interchangeably, with the convention that lower cost means better performance. In practice, the cost is evaluated over a set, $\mathcal{D}$, that is independent from the train set. Costs of interest are shown in Table 1: top rows focus on prediction accuracy (loss and evaluation metrics), while bottom ones address explanation quality (sufficiency and incomprehensiveness). As explanation metrics are less common than accuracy metrics, we review them next.

Cost for Explanability. We assume access to an *auxiliary explanation method* that assigns importance scores to input features—e.g., based on the magnitude of input gradients. Then, the *explanation quality metric* measures whether the features with the highest importance scores are actually meaningful (see Nauta et al. (2023) for a review). We use *sufficiency* and *incomprehensiveness* as explanation quality metrics to illustrate our framework. They quantify the change in prediction when the most important features are removed or retained. For a comprehensive discussion of our rationale in selecting these metrics, see Appendix A. We emphasize that importance is defined relative to the explanation method, not to any ground truth. This is by design: the goal is not to assume a known set of truly important features, but to assess how well a given explanation method identifies features that meaningfully affect the model's prediction.

Benefit of Personalization. We can quantify the impact of a personalized model in terms of the benefit of personalization, defined next:

Definition 3.2 (Group Benefit of Personalization (G-BoP) (Monteiro Paes et al., 2022))**.** The gain from personalizing a model can be measured by $\text{G-BoP}(h_0, h_p, s) \triangleq C(h_0, s) - C(h_p, s)$, comparing the costs of the generic h_0 and personalized models h_p for group $s \in \mathcal{S}$. By convention, G-BoP > 0 if the personalized model performs better than the generic one.

We use G-BoP_P and G-BoP_X to refer to G-BoP for prediction and explanation respectively – see Appendix B for concrete examples. To evaluate whether all groups benefit from personalization, or if any are harmed, we use the following definition as our final assessment metric:

Definition 3.3 (Benefit of Personalization (BoP) (Monteiro Paes et al., 2022))**.** The BoP is defined as: $\gamma(h_0, h_p) \triangleq \min_{s \in \mathcal{S}}(\text{G-BoP}(h_0, h_p, s))$, i.e., the minimum group BoP value across groups $s \in S$ to capture the worst group improvement, or degradation, resulting from personalization.

A positive γ indicates that all groups receive better performance with respect to the cost function. Contrary to this, a negative γ reflects that at least one group is disadvantaged by personalization. When γ is small or negative, the practitioner might want to reconsider the use of personalized attributes in terms of fairness with respect to all groups. When γ is used to evaluate improvement in prediction and explanation, it is referred to as γ_P and γ_X, respectively.

Remark. The definitions of G-BoP and γ were originally introduced in Monteiro Paes et al. (2022). While formally applicable to any cost function, these definitions have only been studied and used with binary costs—such as 0-1 classification loss or false positive/negative rates—due to a theoretical gap that prevents their use with continuous costs, including an analysis of prediction and explanation for regression tasks. Since a holistic analysis of prediction and explanation across machine learning tasks is our primary focus, addressing this gap is central to our contribution in Section 5.

4 Impact of Personalization on Prediction And Explainability

This section provides the first formal analysis showing that personalization's effect on prediction does not determine its effect on explainability, highlighting the need to evaluate both. A common intuition is that if personalization improves prediction, it must also improve explanations (Del Giudice, 2024). This assumption underlies many applications, particularly in high-stakes domains where explanations are used to extract insights from high-performing models (Elmarakeby et al., 2021; Chereda et al., 2021). Yet this link has never been formally analyzed.

Theorems 4.1 and 4.2 prove that prediction gains and explanation gains can diverge, demonstrating that gains in prediction performance (measured by BoP_P) and gains in explanation quality (measured by BoP_X) need not align. Theorem 4.3 provides a partial converse, identifying an additive setting where the two align. Though idealized, this boundary case clarifies when practitioners can trust prediction and explanation to align. Proofs are in Appendix C.

No Prediction Benefit Does not Imply No Explainability Benefit. The following theorem shows that a personalized model may match a generic model in accuracy, yet offer better explanation. Thus, focusing only on prediction can overlook significant interpretability gains.

Theorem 4.1. *There exists a data distribution $P_{\mathbf{X},\mathbf{S},\mathbf{Y}}$ such that the Bayes optimal classifiers h_0 and h_p satisfy $\gamma_P(h_0, h_p) = 0$ (with γ_P measured by 0-1 loss) and $\gamma_X(h_0, h_p) > 0$ (with γ_X measured by sufficiency and incomprehensiveness).*

We illustrate Theorem 4.1 with a real-world example. Consider a model with many input features that are partially redundant, for instance, a loan approval model that uses credit score, income, and debt-to-income ratio. Adding a personal feature that is highly correlated with existing features may not change the predictions. However, it can alter the explanation if that feature is the most direct or informative input. For example, adding a binary feature like "pre-approved by another bank", which is strongly correlated with existing features, may leave predictions unchanged, but an explainer might now assign most importance to this new feature because it provides a clearer justification. Figure 5 illustrates the construction behind the proof for sufficiency, where both generic h_0 and personalized h_p models predict perfectly (left side), yet only keeping the most important feature for each (right side) shows that the personalized model is more explainable. For this distribution, $\text{G-BoP}_P(h_0, h_p, s) = 0$ and $\text{G-BoP}_X(h_0, h_p, s) > 0$ for each group s, so all groups are impacted similarly by personalization. Figure 6 illustrates the proof for incomprehensiveness.

No Prediction Harm Does Not Imply No Explainability Harm. A personalized model may match a generic model in accuracy yet offer worse explanations. Thus, focusing only on predictive performance can obscure significant harms to explainability.

Theorem 4.2. *There exists a data distribution $P_{\mathbf{X},\mathbf{S},\mathbf{Y}}$ such that the Bayes optimal classifiers h_0 and h_p satisfy $\gamma_P = 0$ (with γ_P measured by 0-1 loss) and $\gamma_X < 0$ (with γ_X measured by incomprehensiveness).*

To illustrate Theorem 4.2 consider a pneumonia detection model using chest X-ray findings that perfectly predict outcomes. Adding white blood cell count leaves accuracy unchanged, but the personalized model now splits importance between X-ray findings and white blood cell count. The explanation is worse because it's now split across two features, making it less clear which feature drives the decision, even though the X-ray alone was already perfectly predictive. Additionally, Theorem C.1 proves this phenomena for both sufficiency and incomprehensiveness by showing how personalization can affect explainability differently for different groups.

Together, Theorems 4.1, 4.2 and C.1 show that knowing $\gamma_P = 0$ provides no information about γ_X. This motivates the need to evaluate both prediction and explainability, as we offer to do in Section 5.

Absence of explainability benefit can imply absence of prediction benefit. We now ask the converse: can a lack of explainability benefit imply no predictive benefit? We show that this can be true, for a simple additive model, as long as two notions of explanability measures –sufficiency and incomprehensiveness– do not see any benefit.

Theorem 4.3. *Assume that h_0 and h_p are Bayes optimal regressors and $P_{\mathbf{X},\mathbf{S},\mathbf{Y}}$ follows an additive model, i.e., $\mathbf{Y} = \alpha_1\mathbf{X_1} + \cdots + \alpha_t\mathbf{X_t} + \alpha_{t+1}\mathbf{S_1} + \cdots + \alpha_{t+k}\mathbf{S_k} + \epsilon$, where $\mathbf{X_1}, \cdots, \mathbf{X_t}$ and $\mathbf{S_1}, \cdots, \mathbf{S_k}$ are independent, and ϵ is independent random noise. Then, if for $s \in \mathcal{S}$ we have $\text{G-BoP}_{\textit{suff}}(h_0, h_p, s) = \text{G-BoP}_{\textit{incomp}}(h_0, h_p, s) = 0$, then $\text{G-BoP}_{\text{P}}(h_0, h_p, s) = 0$. Consequently, if, for all groups s, $\text{G-BoP}_{\textit{suff}}(h_0, h_p, s) = \text{G-BoP}_{\textit{incomp}}(h_0, h_p, s) = 0$, then $\gamma_P = 0$.*

This theorem demonstrates that under an additive model, if there is no benefit in explanation quality, then there is also no benefit in prediction accuracy. Additionally, we get the following corollary:

Corollary 4.4. *Under the assumptions of Theorem 4.3, if for $s \in \mathcal{S}$, we have $\text{G-BoP}_P(h_0, h_p, s) \neq 0$, then it also holds that $\text{G-BoP}_{\textit{suff}}(h_0, h_p, s) \neq 0$ or $\text{G-BoP}_{\textit{incomp}}(h_0, h_p, s) \neq 0$. Consequently, if $\gamma_P \neq 0$, then there exists a group $s \in \mathcal{S}$ such that $\text{G-BoP}_{\textit{suff}}(h_0, h_p, s) \neq 0$ or $\text{G-BoP}_{\textit{incomp}}(h_0, h_p, s) \neq 0$.*

This theorem means that an effect of personalization on prediction necessarily means an effect on explanation for at least one explainability measure and for at least one demographic group. This result establishes a rare direct link between explanation and prediction, in a simplified linear setting. Proving this for general models remains an open question.

5 Testing Personalization's Impact on Prediction and Explanation

Having emphasized the importance of evaluating both prediction and explainability, we now introduce a methodology to assess them in practice. The true BoP γ, defined over the whole data distribution, is inaccessible and needs to be estimated from finite samples. Then, if its estimate $\hat{\gamma}$ is positive, one must consider whether the true γ is also likely to be positive. In scenarios where personalization incurs a price—such as requesting sensitive user information—one should determine how large $\hat{\gamma}$ must be to ensure that the true benefit exceeds a desired threshold $\gamma \geq \epsilon$. This section analyzes the validity of BoP hypothesis testing and provides guidelines for its application. Proofs for this section are in Appendices D.1, D.3, D.5, D.8.

5.1 Validity of Hypothesis Tests

Hypothesis Tests. Given an audit dataset $\mathcal{D}$ with k binary group attributes, we want to know whether personalization improves each group by at least $\epsilon > 0$. We formalize the null and the alternative hypotheses using a standard framework for the BoP (Monteiro Paes et al., 2022):

$$H_0 : \gamma(h_0, h_p; \mathcal{D}) \leq 0 \Leftrightarrow \text{Personalized } h_p \text{ does not bring any gain for at least one group,}$$
$$H_1 : \gamma(h_0, h_p; \mathcal{D}) \geq \epsilon \Leftrightarrow \text{Personalized } h_p \text{ yields at least } \epsilon \text{ improvement for all groups.}$$

Importantly, H_0 and H_1 are not complementary to each other, because we want to reject the null if the impact is both positive *and* practically meaningful, i.e., $\geq \epsilon$. With these hypotheses, we ask: can we rule out that there is no harm *and* assert a meaningful benefit of at least ϵ?

The improvement ϵ is in cost function units, and represents the improvement for the group that benefits the least from the personalized model. The value ϵ is domain-specific and should be chosen by the practitioner. For example, in healthcare, if personalization requires time-intensive and sensitive inputs—like mental health assessments—it may only be justified if it improves diagnostic accuracy by at least a few points, making ϵ a clinically and ethically meaningful threshold. In such cases, ϵ becomes a threshold for balancing speed and clinical value.

Once ϵ is chosen, the practitioner may run the hypothesis test by computing the estimate $\hat{\gamma}$ on $\mathcal{D}$ and follow the rule: $\hat{\gamma} \geq \epsilon \Rightarrow$ *Reject H_0: Conclude that personalization yields at least ϵ improvement for all groups*. We note that different testing strategies could also be used. To capture this generality, we define a decision function $\Psi : (h_0, h_p, \mathcal{D}, \epsilon) \to \{0, 1\}$, where $\Psi = 1$ indicates rejection of H_0. In our case, $\Psi(h_0, h_p, \mathcal{D}, \epsilon) = (\hat{\gamma} \geq \epsilon)$. Regardless of its specific form, our goal is to assess the validity of *any* test aiming to evaluate the impact of personalization γ.

Invalidity of the Tests: Probability of Error. We quantify the (in)validity of a test in terms of its probability of error: $P_e = \Pr(\text{Rejecting } H_0 | H_0 \text{ is true}) + \Pr(\text{Failing to reject } H_0 | H_1 \text{ is true})$.

We propose to derive a minimax lower bound on the error probability P_e. This involves considering the worst-case data distributions that maximizes P_e and the best possible decision function Ψ that minimizes it. Notably, a high lower bound guarantees a high error probability for *any* test with H_0 and H_1 on the BoP, flagging settings where testing the impact of personalization is unreliable.

Theorem 5.1. *Consider k binary group attributes, $\mathcal{S} \triangleq \{0, 1\}^k$, that specify $d \triangleq |\mathcal{S}| = 2^k$ groups, each containing m_j individuals, $j = 1, .., d$. Let H_0 (resp. H_1) denotes the data distributions under which the generic model h_0 (resp. the personalized model h_p) performs better, i.e., $\gamma \leq 0$ (resp. $\gamma \geq \epsilon$). Then, there exists $P_0 \in H_0$ (resp. $P_1 \in H_1$), for which the individual benefit of personalization $\mathbf{B} = cost(h_0, \tilde{\mathbf{X}}, \mathbf{Y}) - cost(h_p, \tilde{\mathbf{X}}, \mathbf{Y})$, follows a probability density p (resp. p_ϵ for one group), where $\mathbb{E}_p[\mathbf{B}] = 0$, and $\mathbb{E}_{p^\epsilon}[\mathbf{B}] = \epsilon$, such that:*

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\frac{p^\epsilon(\mathbf{B})}{p(\mathbf{B})} \right]^{m_j} - 1 \right]^{\frac{1}{2}}. \tag{1}$$

Crucially, this lower bound can be tailored to the practitioner's specific use case, i.e., to the distribution of the individual benefit $\mathbf{B}$ under H_0 and H_1. For example, if $\mathbf{B}$ is known or observed to follow a Laplace distribution with scale b, the practitioner should choose $p = \text{Laplace}(0, b)$ and $p^\epsilon = \text{Laplace}(\epsilon, b)$. Figure 3 shows the expression of the lower bound for the Laplace distribution. The next corollary expresses it for distributions in the exponential family.

Corollary 5.2. *The lower bound in Th. 5.1 for distributions p, p^ϵ in the exponential family (parameter θ, moment generating function M) is: $1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \left(\frac{M_p(2\Delta\theta)}{M_p(\Delta\theta)^2} \right)^{m_j} - 1 \right]^{\frac{1}{2}}$ with $\Delta\theta = \theta^\epsilon - \theta$.*

These results generalize and tighten an existing bound for categorical distribution only (Monteiro Paes et al., 2022) and provide the first general framework to evaluate the (in)validity of hypothesis tests on personalization for prediction and explanation, and across supervised machine learning tasks.

Experimental Design: Group Attributes, Sample Size, and Detectable Gain. We investigate how probability of error depends on the dataset, and how it determines their ability to test the impact of personalization. For example, with a fixed number of individuals N, a larger number of personal attributes k increases the number of groups $d = 2^k$, reducing the number of samples per group, which increases the risk of error. Accordingly, if the practitioner commits to a fixed k to test a desired gain ϵ (resp. fixed k and N), they need a minimum group size m, as shown next.

Corollary 5.3. *To ensure $\min \max P_e \leq v$ for a chosen threshold v, equal group sizes must satisfy $m \geq m_{\min}$, where: $m_{min} = \frac{\log\left(4 \cdot 2^k (1-v)^2 + 1\right)}{\log(1+4\epsilon^2)}$ for a categorical BoP, $m_{min} = \frac{\sigma^2}{\epsilon^2} \log\left(2^{2+k}\left(1 + 2^{-2-k} - 2v + v^2\right)\right)$ for a Gaussian BoP of variance σ^2, and $m_{\min} = \frac{b}{\epsilon} \log\left(2^{2+k}\left(1 + 2^{-2-k} - 2v + v^2\right)\right)$ for a Laplace BoP of scale b.*

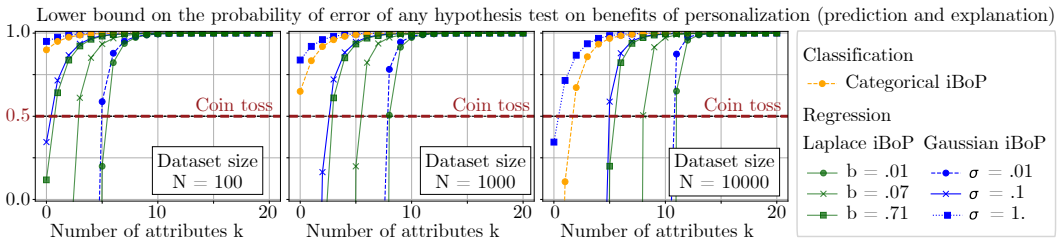

Figure 2: **Testing personalization for prediction and explanation depends on learning task.** Lower bound on the probability of error P_e with respect to number of personal attributes k, for dataset sizes $N = 10^2$, 10^3, and 10^4 with $\epsilon = 0.01$. In classification (orange), the bound is fixed by the categorical nature of the individual BoP (iBoP) and is identical for prediction and explanation. In regression (green and blue), P_e depends on the spread of individual BoPs—parameterized by variance σ^2 (Gaussian) or scale b (Laplace). Smaller variance or scale allows more attributes before testing becomes unreliable ($P_e \geq 0.5$). Computed for $m = \lfloor N/d \rfloor$ samples per group with $d = 2^k$ groups.

Appendix E provide practitioners with dataset-specific feasibility checks: Corollary E.1 bounds the maximum number of attributes that can be included before the lower bound error exceeds 50%, while Corollary E.2 specifies the minimum group size needed to keep the error bound below a desired level.

5.2 Practical Considerations when Testing Prediction and Explanation

We examine how the lower bound in Theorem 5.1 depends on the distribution of individual BoPs $\mathbf{B}$, and how this determines the practitioner's ability to test for prediction or explanation gains.

Testing Prediction and Explanation in Classification Tasks. When the task is classification with 0-1 loss, the individual BoPs follow categorical distributions with values in $\{-1, 0, 1\}$:

$$\mathbf{B}_P = (h_0(\mathbf{X}) \neq \mathbf{Y}) - (h_p(\mathbf{X}, \mathbf{S}) \neq \mathbf{Y}), \quad \mathbf{B}_X = (h_0(\mathbf{X}) \neq h_0(\mathbf{X}_J)) - (h_p(\mathbf{X}, \mathbf{S}) \neq h_0(\mathbf{X}_J, \mathbf{S}_J))$$

for prediction and explanation (e.g., sufficiency), respectively –see costs in Table 1. In this setting, the lower bound in Theorem 5.1 is identical for prediction and explanation (see Figure 3, bottom): either both are testable, or neither is.

Figure 2 shows the lower bound on the probability of error P_e as a function of k, for typical dataset sizes in medical settings $N \in \{10^2, 10^3, 10^4\}$. In classification (orange curves), even a small number of personal attributes k leads to high error lower bounds. For instance, at $N = 100$ and $k = 1$, the bound already exceeds 85%, making reliable testing impossible for both prediction and explanation.

Testing Prediction and Explanation in Regression Tasks. In regression, the situation is more nuanced. For instance, with MSE loss, we have continuously valued individual BoP random variables:

$$\mathbf{B}_P = |h_0(\mathbf{X}) - \mathbf{Y}|^2 - |h_p(\mathbf{X}, \mathbf{S}) - \mathbf{Y}|^2, \quad \mathbf{B}_X = |h_0(\mathbf{X}) - h_0(\mathbf{X}_J)|^2 - |h_p(\mathbf{X}, \mathbf{S}) - h_0(\mathbf{X}_J, \mathbf{S}_J)|^2,$$

for prediction and explanation, respectively. Suppose these follow Laplace distributions with scales b_P and b_X. Then, the lower bounds will differ for prediction and explanation (Figure 3, bottom): one could be testable while the other is not, highlighting an asymmetry absent in the classification case.

As illustrated in Figure 2, smaller scale values (b) allow for a larger number of personal attributes $k_{\max}$ to be tested without theoretical barriers. Unlike classification, there is no proof that regression tasks cannot support reliable testing of personalization for dataset sizes encountered in medical settings $N \in \{10^2, 10^3, 10^4\}$, even with many personal attributes k.

6 Case Studies: Evaluating Personalization on Real Datasets

We illustrate how to use our results to investigate the impact of personalization on prediction and explanation, to reveal the many cases where reliable testing is in fact impossible. This section focuses on one real-world healthcare scenario, while other scenarios are provided in Appendix G.
Remark. Across these hypothesis tests we always evaluate if there is a benefit of personalization, i.e.

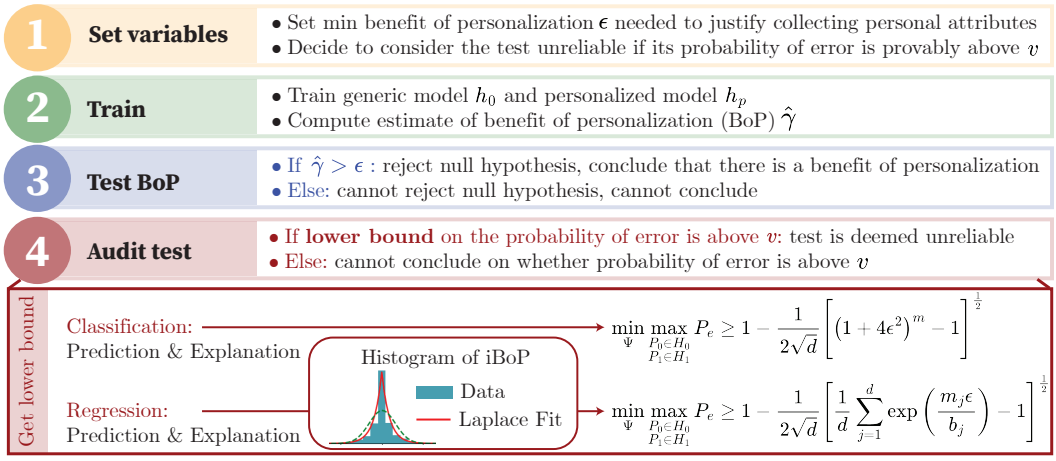

Figure 3: **Summary of the steps to test BoP for prediction and explanation**.

$\gamma > \epsilon > 0$, but interested practitioners may want to evaluate whether an existing machine learning model could harm one group. In that case the hypothesis test should be flipped, i.e. $\gamma < \epsilon < 0$.

Healthcare Scenario. Consider MIMIC-III (Medical Information Mart for Intensive Care) (Johnson et al., 2016), a dataset of patients admitted to critical care units at a large tertiary hospital –containing vital signs, medications, lab results, diagnoses, imaging reports, and outcomes such as length of stay. Suppose that a practitioner has developed a deep learning model to predict a patient's length of stay (regression) or whether the length of stay exceeds 3 days (classification) – see details in Appendix F.1. They are wondering whether their model should be personalized by including (or not) two personal attributes: Age $\times$ Race $\in \{18-45, 45+\} \times \{\text{White(W)}, \text{NonWhite(NW)}\}$. However, they are concerned this could disadvantage some groups, not only by reducing prediction accuracy but also by limiting the ability to uncover factors that explain critical care duration. We provide a step-by-step procedure to use our framework to evaluate the benefit of personalization (summarized in Figure 3).

① **Select ϵ and v, report empirical benefits of personalization.** The practitioner first chooses the minimum improvement they expect from personalization—ϵ_P for prediction and ϵ_X for explanation (e.g., $\epsilon_P = \epsilon_X = 0.002$). They then set a tolerance threshold v for the probability of error beyond which they will not trust the hypothesis test (e.g., $v = 50\%$).

② **Report empirical benefits of personalization** The practioner trains h_0 and h_p (with additional attributes age and race) and reports empirical personalization benefits in Table 2 (0–1 loss for classification, MSE for regression). In both tasks, some groups show benefits for prediction but harm for explanation, and vice versa. This should not be surprising given the results of Section 4, which show that prediction and explanation gains can diverge.

③ **Perform hypothesis test.** The practioner assesses whether $\hat{\gamma}$ exceeds ϵ_P or ϵ_X. It does for all metrics with a positive $\hat{\gamma}$, hence they can reject the null hypothesis for these cases.

④ **Assess reliability of the results.** Next, the practitioner assesses whether the empirical results are statistically meaningful using the framework from Section 5. *For the classification model*, the lower bound on the probability of error exceeds 80% (Figure 4, $\epsilon = 0.002$), indicating that it is not even possible to test whether personalization helps or harms performance. As a result, the practitioner would likely retain the generic classifier. *For the regression model*, they examine the distributions of individual BoPs, $\mathbf{B}_P$ and $\mathbf{B}_X$ (Figure 3, bottom, and Appendix F.1). Sufficiency is best fit by Gaussians with varying variances; prediction and incomprehensiveness align with Laplace distributions of different scales. The corresponding lower bounds on error exceed 80% for sufficiency—making it untestable—but fall below 10% for prediction and incomprehensiveness (Figure 4, $\epsilon = 0.002$). Now, we provide insights that were gained from applying our framework to this scenario, and others in Appendix G.

Insight: A high empirical benefit of personalization $\hat{\gamma}$ can be misleading. In the regression experiment, sufficiency showed the largest benefit ($\hat{\gamma} = 0.1914$), yet the data did not permit a valid test, making the result inconclusive. Prediction showed a much smaller benefit ($\hat{\gamma} = 0.0021$), but

Table 2: Benefits of personalization ($\hat{C}(h_0) - \hat{C}(h_p)$) on the MIMIC-III test set for predicting length of stay (LOS): regression or classification (LOS > 3 days). Incomprehensiveness is abbreviated as incomp. and population as pop. Values that are worsened by h_p are colored red.

Group	Classification				Regression			
	n	Prediction	Incomp.	Sufficiency	n	Prediction	Incomp.	Sufficiency
White, 45+	8443	0.0063	-0.0226	0.0053	8379	0.0021	-0.0906	0.1914
White, 18–45	1146	0.0044	0.0489	0.0244	1197	0.0023	0.1219	0.2223
NonWhite, 45+	3052	-0.0026	-0.0023	0.0029	3044	0.0108	-0.0501	0.3494
NonWhite, 18–45	696	-0.0216	0.0560	0.0072	717	0.0212	0.0441	0.3293
All Pop.	13337	0.0026	-0.0077	0.0065	13337	0.0051	-0.0550	0.2376
Minimal BoP	13337	-0.0216	-0.0226	0.0029	13337	0.0021	-0.0906	0.1914

our analysis found no barriers to testing, and the null was rejected. This shows that large $\hat{\gamma}$ does not guarantee a valid conclusion; empirical values must be paired with our framework to assess validity.

Insight: The choice of improvement threshold ϵ is key. Increasing ϵ reduces the lower bound on the probability of error P_e, making hypothesis testing potentially less unreliable (Figure 4), but also raises the bar for rejecting the null, requiring a larger $\hat{\gamma}$. Thus, ϵ trades off test validity against ability to detect effects.

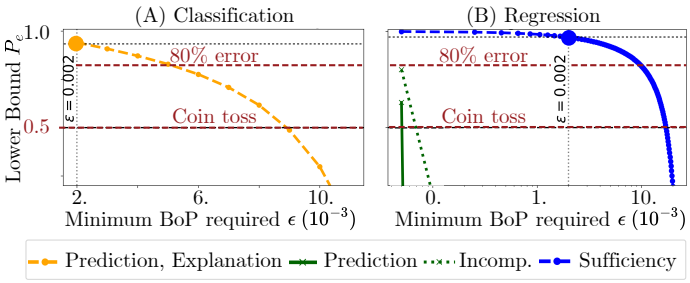

Figure 4: **Lower bound of P_e vs. ϵ on MIMIC-III**: classification (A) and regression with Laplace (green) and Gaussian (blue) models for the individual BoPs (B). At the minimum BoP set in this case study ($\epsilon = 0.002$), testing personalization for prediction and explanation is impossible for classification (same for sufficiency for regression) as $P_e \geq 80\%$ regardless of the hypothesis test.

Insight: Results do not depend on the explanation method. Table 2 reports results with Integrated Gradients (Sundararajan et al., 2017). Since our framework applies to any explanation method, we test whether this choice affects the evaluation of the impact of personalization. Appendix G analyzes Shapley Value Sampling Štrumbelj & Kononenko (2010) and DeepLIFT Shrikumar et al. (2017), finding substantial agreement across the methods—though effect sizes differ.

Insight: Personalization is hard to evaluate across medical datasets. To show the practicality of the framework, we also include experiments on the UCI Heart Dataset (Janosi et al., 1989) and the MIMIC-III Kidney injury cohort Suriyakumar et al. (2023), again utilizing a range of explanation methods (see Appendix G). Using the same ϵ as above, no test is valid for the S.V.S explainer on the UCI Heart dataset, showing the difficulty of reliably evaluating personalization. More generally, this analysis points to a limitation of personalized medicine and healthcare: while personalization may yield improvements, demonstrating them reliably can be infeasible—restricting applicability.

Concluding Remarks

We present a unified framework for evaluating the benefits of personalization with respect to both prediction accuracy and explanation quality, facilitating nuanced decisions regarding the use of personal attributes. Our analysis shows that in many practical settings, particularly classification tasks, the statistical conditions required to validate personalization are often unmet. As a result, even when personalization shows empirical gains, meaningful validation may not be feasible.

Limitations & Future Work. While we relax several assumptions relative to prior work, our theoretical results still rely on assumptions not always met in practice; further reducing them remains an important direction. Additionally, while we focused on explanation quality due to its importance in clinical adoption, our results in Section 5 extend to other goals. Future work can build on this framework to evaluate additional desiderata such as fairness, robustness, and uncertainty calibration.

REFERENCES

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms, 2019. URL https://arxiv.org/abs/1905.12843.

Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, June 2022. doi: 10.1145/3531146.3533179. URL http://dx.doi.org/10.1145/3531146.3533179.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression, 2017. URL https://arxiv.org/abs/1706.02409.

Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. *2013 IEEE 13th International Conference on Data Mining*, pp. 71–80, 2013. URL https://api.semanticscholar.org/CorpusID:16541789.

Hryhorii Chereda, Annalen Bleckmann, Katharina Menck, et al. Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Medicine*, 13(1):42, 2021. doi: 10.1186/s13073-021-00845-7. URL https://doi.org/10.1186/s13073-021-00845-7.

Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H. Bach, and Himabindu Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 203–214. ACM, July 2022. doi: 10.1145/3514094.3534159. URL http://dx.doi.org/10.1145/3514094.3534159.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations, 2022. URL https://arxiv.org/abs/2202.00734.

Marco Del Giudice. The prediction-explanation fallacy: A pervasive problem in scientific applications of machine learning. *Methodology*, 20(1):22–46, Mar. 2024. doi: 10.5964/meth.11235. URL https://meth.psychopen.eu/index.php/meth/article/view/11235.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

Cynthia Dwork and Christina Ilvento. Fairness under composition. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi: 10.4230/LIPICS.ITCS.2019.33. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2019.33.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011. URL https://arxiv.org/abs/1104.3913.

Haitham A. Elmarakeby, Jaeil Hwang, Rami Arafeh, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021. doi: 10.1038/s41586-021-03922-4. URL https://doi.org/10.1038/s41586-021-03922-4.

John M Flack, Keith C Ferdinand, and Samar A Nasser. Epidemiology of hypertension and cardiovascular disease in african americans. *The Journal of Clinical Hypertension*, 5(1):5–11, 2003.

Kazuto Fukuchi, Toshihiro Kamishima, and Jun Sakuma. Prediction with model-based neutrality. In *ECML/PKDD*, 2013. URL https://api.semanticscholar.org/CorpusID:6964544.

Furkan Gursoy and Ioannis A. Kakadiaris. Error parity fairness: Testing for group fairness in regression tasks, 2022. URL https://arxiv.org/abs/2208.08279.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016. URL https://arxiv.org/abs/1610.02413.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Bi Huang, Mayank Dalakoti, and Gregory Y H Lip. How far are we from accurate sex-specific risk prediction of cardiovascular disease? One size may not fit all. *Cardiovascular Research*, 120(11):1237–1238, 06 2024. ISSN 0008-6363. doi: 10.1093/cvr/cvae135. URL https://doi.org/10.1093/cvr/cvae135.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C52P4X.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):160035–160035, 2016. ISSN 2052-4463.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey, 2024. URL https://arxiv.org/abs/2209.11326.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022. URL https://arxiv.org/abs/1908.09635.

Lucas Monteiro Paes, Carol Long, Berk Ustun, and Flavio Calmon. On the epistemic limits of personalized prediction. *Advances in Neural Information Processing Systems*, 35:1979–1991, 2022.

Lori Mosca, Elizabeth Barrett-Connor, and Nanette Wenger. Sex/gender differences in cardiovascular disease prevention what a difference a decade makes. *Circulation*, 124:2145–54, 11 2011. doi: 10.1161/CIRCULATIONAHA.110.968792.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, July 2023. ISSN 1557-7341. doi: 10.1145/3583558. URL http://dx.doi.org/10.1145/3583558.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. ISSN 1095-9203. doi: 10.1126/science.aax2342. URL http://dx.doi.org/10.1126/science.aax2342.

Jessica Paulus, Benjamin Wessler, Christine Lundquist, Lana Yh, Gowri Raman, Jennifer Lutz, and David Kent. Field synopsis of sex in clinical prediction models for cardiovascular disease. *Circulation: Cardiovascular Quality and Outcomes*, 9:S8–S15, 02 2016. doi: 10.1161/CIRCOUTCOMES.115.002473.

Jessica Paulus, Benjamin Wessler, Christine Lundquist, and David Kent. Effects of race are rarely included in clinical prediction models for cardiovascular disease. *Journal of General Internal Medicine*, 33, 05 2018. doi: 10.1007/s11606-018-4475-x.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55 (3), feb 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL https://doi.org/10.1145/3494672.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning, 2017. URL https://arxiv.org/abs/1710.05578.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *CoRR*, abs/1704.02685, 2017. URL http://arxiv.org/abs/1704.02685.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. URL https://arxiv.org/abs/1706.03825.

Erik Štrumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(1):1–18, 2010. URL http://jmlr.org/papers/v11/strumbelj10a.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

Vinith M. Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms: Reconsidering the use of group attributes in prediction, 2023. URL https://arxiv.org/abs/2206.02058.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. On the sensitivity and stability of model interpretations in nlp, 2022. URL https://arxiv.org/abs/2104.08782.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey, 2022. URL https://arxiv.org/abs/2012.15445.

A Extended Related Works: Explainability

We provide additional extended works about explainability methods below.

Explainability Typical approaches to model explanation involve measuring how much each input feature contributes to the model's output, highlighting important inputs to promote user trust. This process often involves using gradients or hidden feature maps to estimate the importance of inputs (Simonyan et al., 2014; Smilkov et al., 2017; Sundararajan et al., 2017; Yuan et al., 2022). For instance, gradient-based methods use backpropagation to compute the gradient of the output with respect to inputs, with higher gradients indicating greater importance (Sundararajan et al., 2017; Yuan et al., 2022). We focus on feature-attribution explanations as they remain the most widely used form of post hoc interpretability in practice (Nauta et al., 2023). To reflect a range of underlying assumptions, we employ three distinct and widely adopted explainers: Integrated Gradients (gradient-based), DeepLIFT (backpropagation-based), and Shapley value sampling (perturbation-based).

The quality of these explanations is often evaluated using the principle of *faithfulness* (Lyu et al., 2024; Dasgupta et al., 2022; Jacovi & Goldberg, 2020), which measures how accurately an explanation represents the reasoning of the underlying model. Two key aspects of faithfulness are *sufficiency* and *comprehensiveness* (DeYoung et al., 2020; Yin et al., 2022); the former assesses whether the inputs deemed important are adequate for the model's prediction, and the latter examines if these features capture the essence of the model's decision-making process. We selected these metrics as they are widely-adopted, model-agnostic measures that directly assess explanation faithfulness through standard perturbation-based evaluation (Serrano & Smith, 2019), aligning with established principles of correctness and completeness in the explainability literature (Nauta et al., 2023).

B BoP

In the following table, we show how these abstract definitions can be used to measure BoP for both predictions and explanations, each across both classification and regression tasks. The empirical population and group BoP are defined as: $\hat{\text{BoP}}(h_0, h_p) = \hat{C}(h_0) - \hat{C}(h_p)$ and $\hat{\text{BoP}}(h_0, h_p, s) = \hat{C}(h_0, s) - \hat{C}(h_p, s)$, respectively.

Table 3: Formal definitions of the benefit of personalization for prediction and explanation metrics, evaluated for subgroup s.

Evaluation Type	**Benefit of personalization for group s**
Predict (Classification, 0-1 loss)	$\Pr(h_0(\mathbf{X}) \neq \mathbf{Y} \mid \mathbf{S} = s) - \Pr(h_p(\mathbf{X}, s) \neq \mathbf{Y} \mid \mathbf{S} = s)$
Predict (Regression, MSE)	$\mathbb{E}\left[\|h_0(\mathbf{X}) - \mathbf{Y}\|^2 \mid \mathbf{S} = s\right] - \mathbb{E}\left[\|h_p(\mathbf{X}, s) - \mathbf{Y}\|^2 \mid \mathbf{S} = s\right]$
Explain (Sufficiency, classification, 0-1 loss)	$\Pr(h_0(\mathbf{X}) \neq h_0(\mathbf{X}_J) \mid \mathbf{S} = s) - \Pr(h_p(\mathbf{X}, s) \neq h_p(\mathbf{X}_J, s_J) \mid \mathbf{S} = s)$
Explain (Sufficiency, regression, MSE)	$\mathbb{E}\left[\|h_0(\mathbf{X}) - h_0(\mathbf{X}_J)\|^2 \mid \mathbf{S} = s\right] - \mathbb{E}\left[\|h_p(\mathbf{X}, s) - h_p(\mathbf{X}_J, s_J)\|^2 \mid \mathbf{S} = s\right]$
Explain (Incomprehensiveness, classification, 0-1 loss)	$\Pr\left(h_p(\mathbf{X}, s) \neq h_p(\mathbf{X}_{\setminus J}, s_{\setminus J}) \mid \mathbf{S} = s\right) - \Pr\left(h_0(\mathbf{X}) \neq h_0(\mathbf{X}_{\setminus J}) \mid \mathbf{S} = s\right)$
Explain (Incomprehensiveness, regression, MSE)	$\mathbb{E}\left[\|h_p(\mathbf{X}, s) - h_p(\mathbf{X}_{\setminus J}, s_{\setminus J})\|^2 \mid \mathbf{S} = s\right] - \mathbb{E}\left[\|h_0(\mathbf{X}) - h_0(\mathbf{X}_{\setminus J})\|^2 \mid \mathbf{S} = s\right]$

C Comparison BoP for Prediction and BoP for Explainability Proofs

In this section, we present the full proofs comparing the impact of personalization on prediction accuracy versus explanation quality, highlighting situations under which their effects diverge or align.

C.1 Proof for Theorem 4.1

We provide the proof for theorem 4.1 for two metrics of explanation quality: sufficiency and incomprehensiveness, from Table 1. The proof for sufficiency is illustrated in Figure 5. The proof for incomprehensivess is illustrated in Figure 6

Proof. Let $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$ where $\mathbf{X_1}$ and $\mathbf{X_2}$ are independent and each follows Unif$(-\frac{1}{2}, \frac{1}{2})$. Let us define one binary personal attribute $s \in \{0, 1\}$ as $\mathbf{S} = \mathbb{1}(\mathbf{X_1} + \mathbf{X_2} > 0)$ and assume that we seek to predict $\mathbf{Y} = \mathbf{S}$. Then, $h_0(x) = \mathbb{1}(\mathbf{X_1} + \mathbf{X_2} > 0)$ and $h_p(x) = \mathbb{1}(\mathbf{S} > 0)$ are the generic and personalized classifiers of interest.

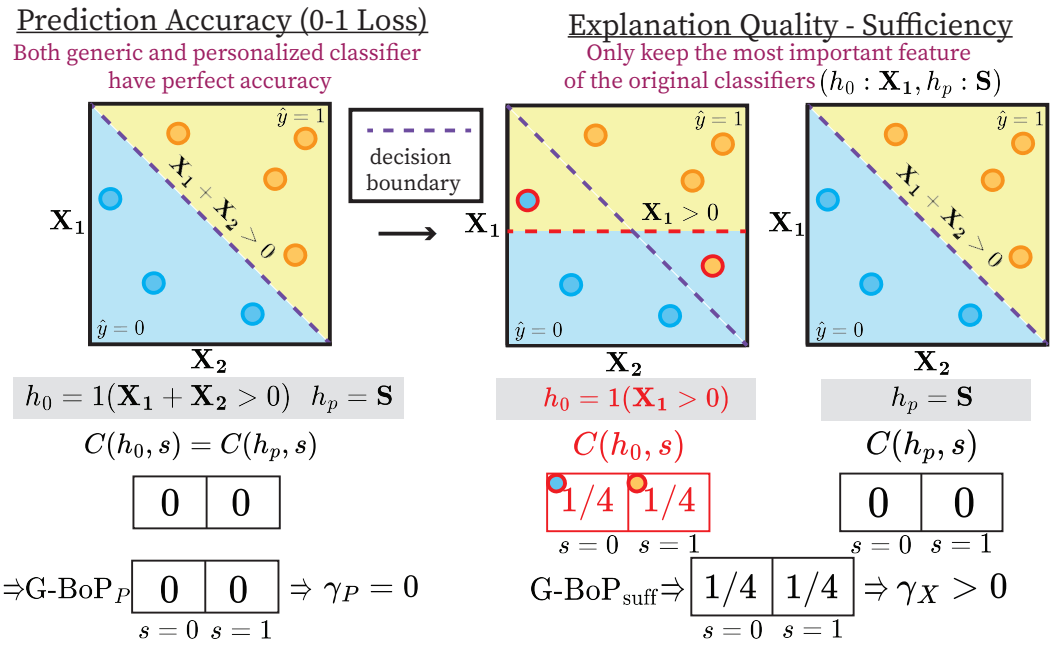

Figure 5: Comparing a generic model (h_0) and a personalized model (h_p) on prediction and explanation (sufficiency). Top-left: The generic model h_0 uses both $\mathbf{X_1}$ and $\mathbf{X_2}$ for predictions, with its decision boundary defined by $\mathbf{X_1} + \mathbf{X_2} > 0$. The personalized model, h_p, has access to the group attribute $\mathbf{S}$ (defined as $\mathbf{S} = \mathbb{1}(\mathbf{X_1} + \mathbf{X_2} > 0)$), and its prediction rule is to output $\mathbf{S}$. Bottom-left: Since both classifiers achieve perfect accuracy (on both groups $s = 0$ and $s = 1$), the Group Benefit of Personalization ($G - BoP_P$) is 0 on both groups, and thus: $\gamma_P = 0$. Top-right: In the sufficiency evaluation, where only the most important feature is kept, h_p achieves perfect prediction since it relies solely on $\mathbf{S}$, reaching a sufficiency cost of 0 for each group. In contrast, h_0, using only $\mathbf{X_1}$, now makes prediction errors and has a worst sufficiency cost of $\frac{1}{4}$ for each group. Bottom-right: Since the personalized model has better sufficiency than the generic model, the G-BoP is positive and equal to $\frac{1}{4}$ for both groups, and hence $\gamma_x = \frac{1}{4} > 0$. Hence, personalization can enhance explainability even though prediction accuracy remains the same.

Prediction. Both classifiers achieve perfect accuracy. Therefore, $\mathrm{BoP}_P(h_0, h_p) = 0$.

In particular, they also achieve perfect accuracy when we restrict the input $\mathbf{X}$ to any subgroup, subgroup $s = 0$ or subgroup $s = 1$, such that:

$$\begin{aligned}\text{G-BoP}_P(h_0, h_p, s = 0) = \text{G-BoP}_P(h_0, h_p, s = 1) = \mathrm{BoP}_P(h_0, h_p) = 0,\\ \Rightarrow \gamma_P(h_0, h_p) = \min_{s\in\{0,1\}} \text{G-BoP}_P(h_0, h_p, s) = 0.\end{aligned}$$

Explanation (sufficiency). We now test sufficiency by evaluating the accuracy of classifiers using only the important feature.

For model h_0, its important feature set J_0 is either $\{\mathbf{X_1}\}$ or $\{\mathbf{X_2}\}$. Without loss of generality, let $J_0 = \{\mathbf{X_1}\}$. For the personalized model, $J_p = \{\mathbf{S}\}$.

For sufficiency, we compute:

$$\begin{aligned}\Pr(h_0(\mathbf{X}) \neq h_0(\mathbf{X_{J_0}})) &= \Pr(\mathbf{X_1} + \mathbf{X_2} \leq 0 | \mathbf{X_1} > 0)\Pr(\mathbf{X_1} > 0)\\ &\quad + \Pr(\mathbf{X_1} + \mathbf{X_2} > 0 | \mathbf{X_1} \leq 0)\Pr(\mathbf{X_1} \leq 0)\\ &= \frac{1}{4},\end{aligned} \tag{2}$$

where the computation per group also gives:

$$\Pr(h_0(\mathbf{X}) \neq h_0(\mathbf{X_{J_0}}) | s = 0) = \Pr(h_0(\mathbf{X}) \neq h_0(\mathbf{X_{J_0}}) | s = 1) = \frac{1}{4}.$$

On the other hand, the sufficiency for h_p is

$$\Pr(h_p(\mathbf{X}, \mathbf{S}) \neq h_p(\mathbf{X_{J_p}}, \mathbf{S_{J_p}})) = 0,$$

as $J_p = \{\mathbf{S}\}$ is sufficient to make a prediction for h_p. The computation per group also gives 0, since the model makes perfect predictions independently of the value taken by $\mathbf{S}$.

Thus, BoP_X in terms of sufficiency is also $\frac{1}{4}$. Computing this quantity per group gives:

$$\begin{aligned}\text{G-BoP}_X(h_0, h_p, s = 0) = \text{G-BoP}_X(h_0, h_p, s = 1) = \frac{1}{4},\\ \Rightarrow \gamma_{\text{suff}}(h_0, h_p) = \min_{s\in\{0,1\}} \text{G-BoP}_X(h_0, h_p, s) = \frac{1}{4}.\end{aligned} \tag{3}$$

Explanation (incomprehensiveness) Incomprehensiveness is the opposite of comprehensiveness. For clarity, we provide the computations for comprehensiveness first.

Comprehensiveness of h_0 is

$$\begin{aligned}\Pr(h_0(\mathbf{X}) \neq h_0(\mathbf{X_{\setminus J_0}})) &= \Pr(\mathbf{X_1} + \mathbf{X_2} \leq 0 | \mathbf{X_2} > 0)\Pr(\mathbf{X_2} > 0)\\ &+ \ \Pr(\mathbf{X_1} + \mathbf{X_2} > 0 | \mathbf{X_2} \leq 0)\Pr(\mathbf{X_2} \leq 0) \qquad (4)\\ &= \Pr(\mathbf{X_1} + \mathbf{X_2} \leq 0 | \mathbf{X_2} > 0) \cdot \frac{1}{2} + \Pr(\mathbf{X_1} + \mathbf{X_2} > 0 | \mathbf{X_2} \leq 0) \cdot \frac{1}{2}\\ &= \Pr(\mathbf{X_1} + \mathbf{X_2} \leq 0 | \mathbf{X_2} > 0) \quad \text{(due to symmetry of the distribution)}\\ &= \int_{x_2>0, x_1+x_2\leq 0} \Pr(x_1, x_2) dx_1 dx_2 / \Pr(\mathbf{X_2} > 0)\\ &= 2 \cdot \int_{x_2=0}^{\frac{1}{2}} \Pr(x_2) \int_{x_1\leq -x_2} \Pr(x_1) dx_1 dx_2\\ &= 2 \cdot \int_{x_2=0}^{\frac{1}{2}} \Pr(x_2)(-x_2 + \frac{1}{2}) dx_2\\ &= 2 \cdot \left[-\frac{1}{2}x_2^2 + \frac{1}{2}x_2\right]_0^{\frac{1}{2}}\\ &= \frac{1}{4}.\end{aligned}$$

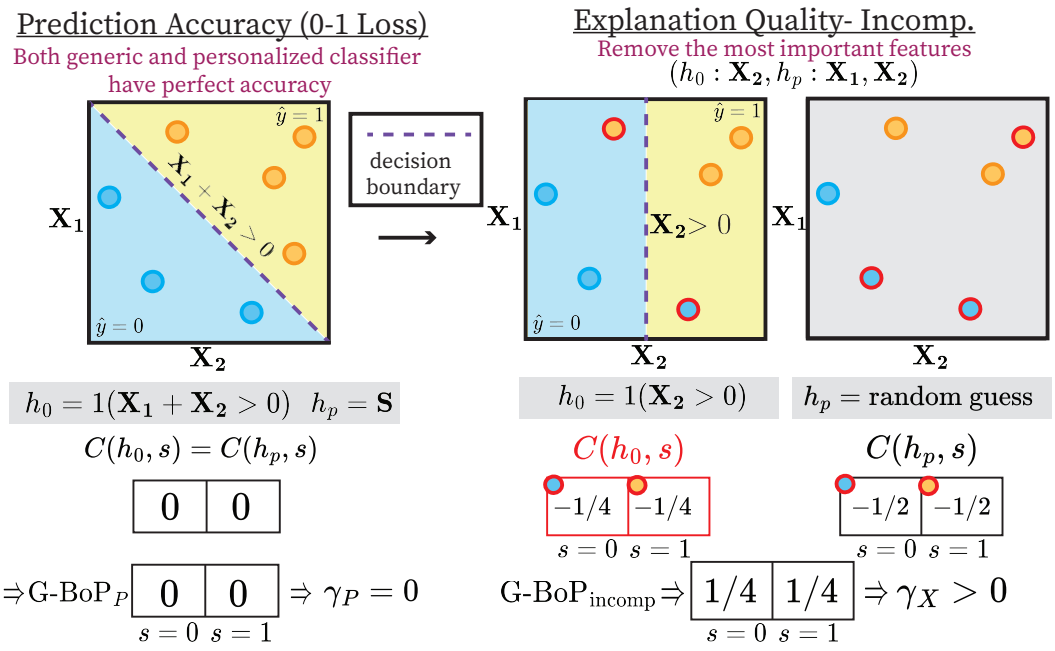

Figure 6: Comparing a generic model (h_0) and a personalized model (h_p) on prediction and explanation (incomprehensiveness). Both achieve perfect accuracy, but h_p relies solely on $\mathbf{S} = 1(\mathbf{X_1} + \mathbf{X_2} > 0)$, yielding higher incomprehensiveness. Hence, personalization can improve explainability even when accuracy is unchanged: here, $\gamma_P = 0$ and $\gamma_X > 0$.

Hence, incomprehensiveness of h_0 is $-\frac{1}{4}$.

Computing this quantity per group gives, by symmetry of the problem:

$$\begin{aligned}\Pr(h_0(\mathbf{X}) \neq h_0(\mathbf{X}_{\setminus \mathbf{J_0}}) \mid s = 0) &= \Pr(h_0(\mathbf{X}) \neq h_0(\mathbf{X}_{\setminus \mathbf{J_0}}) \mid s = 1) \\ &= \frac{1}{2}\Pr(h_0(\mathbf{X}) \neq h_0(\mathbf{X}_{\setminus \mathbf{J_0}})) \\ &= \frac{1}{4}.\end{aligned} \tag{5}$$

Hence, incomprehensiveness per group is also $-\frac{1}{4}$.

For h_p, comprehensiveness is:

$$\Pr(h_p(\mathbf{X}, \mathbf{S}) \neq h_p(\mathbf{X}_{\setminus \mathbf{J_p}}, \mathbf{S}_{\setminus \mathbf{J_p}})) = \frac{1}{2},$$

as without $\mathbf{S}$, h_p can only make a random guess. Hence, incomprehensiveness for each group is $-\frac{1}{2}$.

Computing this quantity per group also gives $\frac{1}{2}$ since h_p makes a random guess independently of the subgroup considered:

$$\begin{aligned}\Pr(h_p(\mathbf{X}) \neq h_p(\mathbf{X}_{\setminus \mathbf{J_p}}) \mid s = 0) &= \Pr(h_p(\mathbf{X}) \neq h_p(\mathbf{X}_{\setminus \mathbf{J_p}}) \mid s = 1) \\ &= \Pr(h_p(\mathbf{X}) \neq h_p(\mathbf{X}_{\setminus \mathbf{J_p}})) \\ &= \frac{1}{2}.\end{aligned} \tag{6}$$

while the incomprehensiveness per group is therefore $-\frac{1}{2}$.

Hence, BoP$_X$ in terms of incomprehensiveness is $\frac{1}{4}$.

Computing this quantity per group gives:

$$\text{G-BoP}_X(h_0, h_p, s=0) = \text{G-BoP}_X(h_0, h_p, s=1) = \frac{1}{4},$$
$$\Rightarrow \gamma_{\text{incomp}}(h_0, h_p) = \min_{s\in\{0,1\}} \text{G-BoP}_X(h_0, h_p, s) = \frac{1}{4}. \tag{7}$$

□

C.2 Proof for Theorem 4.2:

We provide the proof for Theorem 4.2, for explainability incomprehensiveness.

Proof. Let $\mathbf{X} = (\mathbf{X})$ where $\mathbf{X}$ follows Unif$(-\frac{1}{2}, \frac{1}{2})$. Define one binary personal attribute $s \in \{0, 1\}$ as $\mathbf{S} = \mathbf{X}$ and assume that the true label that we seek to predict is $\mathbf{Y} = \mathbf{X} > 0$. We define the classifiers of interest as:

$$h_0(\mathbf{X}) = \mathbb{1}(\mathbf{X} > 0), h_p(\mathbf{X}, \mathbf{S}) = \frac{1}{2}(\mathbf{X} + \mathbf{S}).$$

Prediction. Both h_0 and h_p are perfectly aligned with the ground truth and yield $\hat{y} = \mathbf{Y}$. Therefore, they achieve perfect accuracy. In particular, they also achieve perfect accuracy when we restrict the input $\mathbf{X}$ to any subgroup, subgroup $s = 0$ or subgroup $s = 1$, such that:

$$\text{G-BoP}_P(h_0, h_p, s=0) = \text{G-BoP}_P(h_0, h_p, s=1) = \text{BoP}_P(h_0, h_p) = 0,$$
$$\Rightarrow \gamma_P(h_0, h_p) = \min_{s\in\{0,1\}} \text{G-BoP}_P(h_0, h_p, s) = 0.$$

Therefore, $\text{BoP}_P(h_0, h_p) = 0$.

Explanation (sufficiency). For h_0, the most important feature is $\mathbf{X}$, while for h_p, the most important feature is $\mathbf{S}$.

We now test sufficiency by evaluating the accuracy of classifiers using only the important feature.

- For h_0, keeping $\mathbf{X}$ results in the original predictor. Therefore, prediction does not change at all and the feature is maximally sufficient for both groups ($\text{G-BoP}_{\text{suff}} = 0$ for $s = 0$ and $s = 1$, hence $\gamma_X = 0$.
- For h_p, keeping $\mathbf{S}$ does not change the prediction output because $\frac{1}{2}\mathbf{X} > 0 = \mathbf{X} > 0$. Therefore, prediction does not change at all and the feature is maximally sufficient for both groups ($\text{G-BoP}_{\text{suff}} = 0$ for $s = 0$ and $s = 1$, hence $\gamma_X = 0$

Therefore, $\text{BoP}_\text{X} = 0$ for sufficiency.

Explanation (incomprehensiveness) In this setting, we evaluate incomprehensiveness by measuring the degradation in model predictions when the most important feature is removed.

- **Removing X from h_0:** For h_0, incomprehensiveness is:

  $$\Pr(h_0(\mathbf{X}) \neq h_p())) = \frac{1}{2},$$

  as without $\mathbf{X}$, h_0 can only make a random guess. Hence, incomprehensiveness for each group is $\frac{1}{2}$ and $\gamma_X = \frac{1}{2}$.

- **Removing S from h_p:** For h_p, we compute:

  $$\begin{aligned}\Pr\big(h_p(\mathbf{X}, \mathbf{S}) \neq h_p(\mathbf{X})\big) &= \Pr(\mathbf{X} + \mathbf{S} \leq 0 \mid \mathbf{X} > 0) \Pr(\mathbf{X} > 0) \\ &\quad + \Pr(\mathbf{X} + \mathbf{S} > 0 \mid \mathbf{X} \leq 0) \Pr(\mathbf{X} \leq 0) \\ &= \tfrac{1}{4}. \end{aligned} \tag{8}$$

where the computation per group also gives:

$$\Pr(h_p(\mathbf{X},\mathbf{S}) \neq h_p(\mathbf{X})|s=0) = \Pr(h_p(\mathbf{X},\mathbf{S}) \neq h_p(\mathbf{X})|s=1) = \frac{1}{4}.$$

Hence, $\gamma_X = \frac{1}{4}$.

Therefore, BoP-X $= -\frac{1}{4}$.

□

C.3 Proof for Theorem C.1:

Personalization might not alter predictive accuracy across groups, but it might affect explainability differently for different groups, as emphasized in the next theorem.

Theorem C.1. *There exists a data distribution $P_{\mathbf{X},\mathbf{S},\mathbf{Y}}$ such that the Bayes optimal classifiers h_0 and h_p satisfy* G-BoP$_P(h_0, h_p, s) = 0$ *(measured by 0-1 loss) for all groups s, but some groups have* G-BoP$_P(h_0, h_p, s) > 0$ *while others have* G-BoP$_P(h_0, h_p, s) < 0$ *(measured by sufficiency and incomprehensiveness).*

We provide the proof for Theorem C.1, for two measures of explanability evaluation: sufficiency and incomprehensiveness, as illustrated in Figure 7 and Figure 8. Figure 7 illustrates the proof for sufficiency, where both generic h_0 and personalized h_p models predict perfectly (left), yet only keeping the most important feature for each (right) shows that the personalized model is more explainable for the group $(s'=1, s=0)$, and less explainable for group $(s'=0, s=1)$. Figure 8 illustrates the proof for incomprehensiveness.

Proof. Let $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$ where $\mathbf{X_1}$ and $\mathbf{X_2}$ are independent and follow Unif$(-1, 1)$. Define two binary personal attributes $s \in \{0, 1\}$ and $s' \in \{0, 1\}$ such that the true label that we seek to predict is $\mathbf{Y} = \mathbf{S} \cdot \mathbf{S}'$. We define the classifiers of interest as:

$$h_0(\mathbf{X}) = \mathbb{1}(\mathbf{X_1} + \mathbf{X_2} > 0) \cdot \mathbb{1}(\mathbf{X_2} < 0), \quad h_p(\mathbf{X}, \mathbf{S}) = \mathbf{S} \cdot \mathbf{S}'.$$

Prediction. Both h_0 and h_p are perfectly aligned with the ground truth and yield $\hat{y} = \mathbf{Y}$. Therefore, they achieve perfect accuracy. In particular, this holds for both values of $\mathbf{S}$ and $\mathbf{S}'$:

G-BoP$_P$

$s'\backslash s$	$s=0$	$s=1$
$s'=0$	0	0
$s'=1$	0	0

Such that we get:

$$\gamma_P(h_0, h_p) = \min_{s,s' \in \{0,1\}} \text{G-BoP}_P(h_0, h_p, s) = 0.$$

Explanation (sufficiency). For h_0, the most important feature is $\mathbf{X_1}$, while for h_p, the most important feature is $\mathbf{S}$.

We now test sufficiency by evaluating the accuracy of classifiers using only the important feature.

- For h_0, keeping only $\mathbf{X_1}$ results in a constant predictor $h_0(\mathbf{X_1}) = 0$. This fails to recover $\hat{y}$ when $s = 1$ and $s' = 1$ (red orange dot), leading to an error for the subgroup $(s = 1, s' = 1)$, while the three other subgroups still enjoy perfect prediction.
- For h_p, keeping only $\mathbf{S}$ yields $h_p(\mathbf{S}) = \mathbf{S}$, which fails to recover $\hat{y}$ when $s = 1$ and $s' = 0$ (red blue circles) but still correctly predicts for the other three subgroups.

Combining per-group values gives: such that we get:

$$\gamma_X(h_0, h_p) = \min_{s \in \{0,1\}} \text{G-BoP}_{\text{suff}}(h_0, h_p, s) = -1.$$

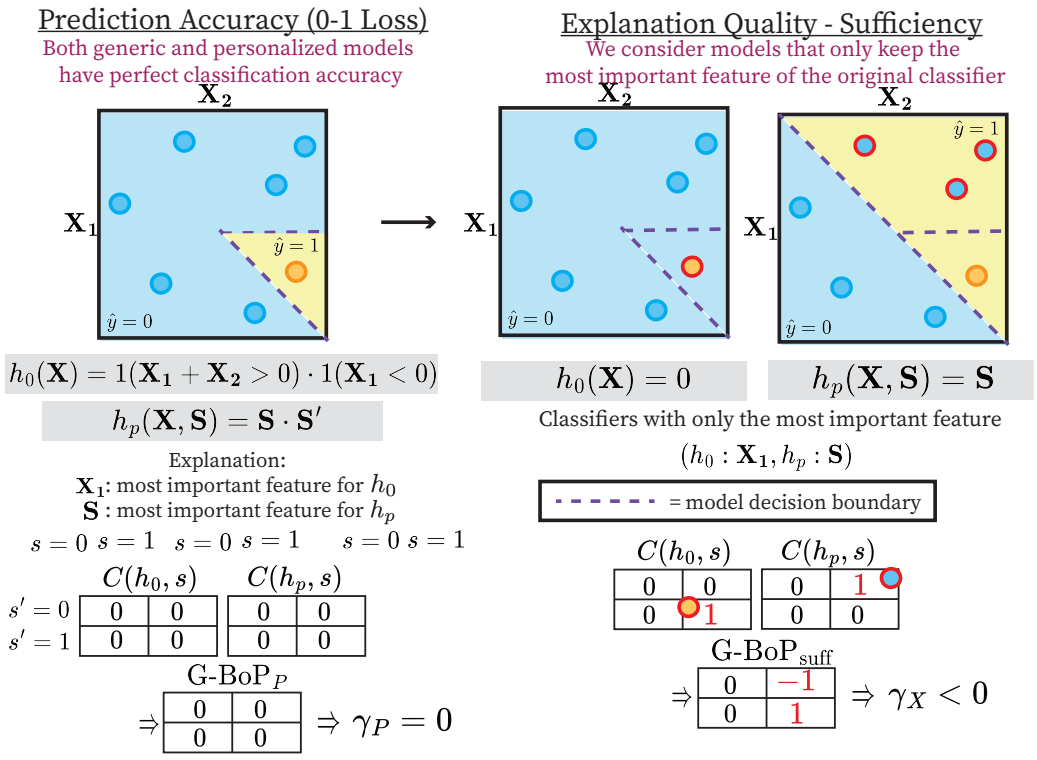

Figure 7: Comparing a generic model (h_0) and a personalized model (h_p) on prediction and explanation (sufficiency). Top-left: The generic model h_0 uses both $\mathbf{X_1}$ and $\mathbf{X_2}$ for predictions with its decision boundary defined by $1(\mathbf{X_1} + \mathbf{X_2} > 0) \cdot 1(\mathbf{X_1} < 0)$. The personalized model, h_p instead predicts using the binary group attributes $s \in 0, 1$ and $s' \in 0, 1$ via the rule $s \cdot s'$. Bottom-left: Both classifiers achieve perfect accuracy across all four groups, hence $\gamma_P = 0$. Top-right: Sufficiency evaluation reveals a difference in explanation quality. For h_0, keeping only the top feature $\mathbf{X_1}$ results in a constant prediction $h_0(\mathbf{X_1}) = 0$, causing an error for the group $s = s' = 1$ (orange circle). For h_p, keeping only $\mathbf{S}$ yields $h_p(\mathbf{S}) = \mathbf{S}$, which fails to recover the true $\mathbf{Y}$ for the group $(s = 1, s' = 0)$ (blue circles). Bottom-right: Thus, the G-BoP is positive for $s = s' = 1$ but negative for $s = 1, s' = 0$, yielding $\gamma_X < 0$. This shows that even with identical predictive performance, the models rely on different features, and personalization can reduce sufficiency-based explainability for some groups.

$\text{G-BoP}_{\text{suff}}$

$s'\backslash s$	$s=0$	$s=1$
$s'=0$	0	-1
$s'=1$	0	1

Explanation (incomprehensiveness) In this setting, we evaluate incomprehensiveness by measuring the degradation in model predictions when the most important feature is removed.

The generic classifier is $h_0(\mathbf{X}) = \mathbb{1}(\mathbf{X_1} + \mathbf{X_2} > 0) \cdot \mathbb{1}(X_1 < 0)$ and the personalized classifier is $h_p(\mathbf{X}, \mathbf{S}) = \mathbf{S} \cdot \mathbf{S}'$. The most important feature for h_0 is $\mathbf{X_1}$ and for h_p is $\mathbf{S}$.

- **Removing $\mathbf{X_1}$ from h_0:** Without $\mathbf{X_1}$, the classifier reduces to the constant function $h_0(\mathbf{X}_{\backslash \mathbf{X_1}}) = 0$. This leads to an incorrect prediction when $s = 1$ and $s' = 1$.
- **Removing S from h_p:** The personalized model becomes $h_p(\mathbf{X}, \mathbf{S}_{\backslash \mathbf{S}}) = \mathbf{S}'$, which ignores $\mathbf{S}$. This leads to an incorrect prediction when $s = 0$ and $s' = 1$, since the true label is $y = 0$ but $h_p = 1$.

All other combinations yield correct predictions even when the important feature is removed.

$\text{G-BoP}_{\text{incomp}}$

$s'\backslash s$	$s=0$	$s=1$
$s'=0$	0	0
$s'=1$	1	-1

This yields the minimum group benefit of personalization is:

$$\gamma_X^{\text{incomp}}(h_0, h_p) = \min_{s,s' \in \{0,1\}} \text{G-BoP}_{\text{incomp}}(h_0, h_p, s, s') = -1.$$

□

C.4 Proof for Theorem 4.3:

See Figure 9 for a visualization of Theorem 4.3 for a linear model with h_0 and h_p Bayes optimal regressors.

Proof. A Bayes optimal regressor using a subset of variables from indices in $J \subseteq [1, \ldots, t+k]$ would be given as:

$$\hat{y} = h_J^*(\mathbf{X}_J, \mathbf{S}_J) = \sum_{\substack{j \in J, \\ j \leq t}} \alpha_j \mathbf{X}_j + \sum_{\substack{j \in J, \\ j \geq t+1}} \alpha_j \mathbf{S}_{j-t}, \tag{9}$$

where h_J^* represents a Bayes optimal regressor for the given subset J, and $\mathbf{X}_J$ and $\mathbf{S}_J$ are sub-vectors of $\mathbf{X}$ and $\mathbf{S}$, using the indices in J.

In what follows, we denote $\backslash J$ as a shorthand notation for $[1, \ldots t+k] \setminus J$.

From equation 9 and the definition of the true response $\mathbf{Y} = \sum_{j \leq t} \alpha_j \mathbf{X}_j + \sum_{j \geq t+1} \alpha_j \mathbf{S}_{j-t}, +\epsilon$ we obtain:

$$\text{MSE}(h_0) = \sum_{j=t+1}^{t+k} \alpha_j^2 \text{Var}(\mathbf{S}_{t+j}) + \text{Var}(\epsilon), \tag{10}$$

$$\text{MSE}(h_p) = \text{Var}(\epsilon). \tag{11}$$

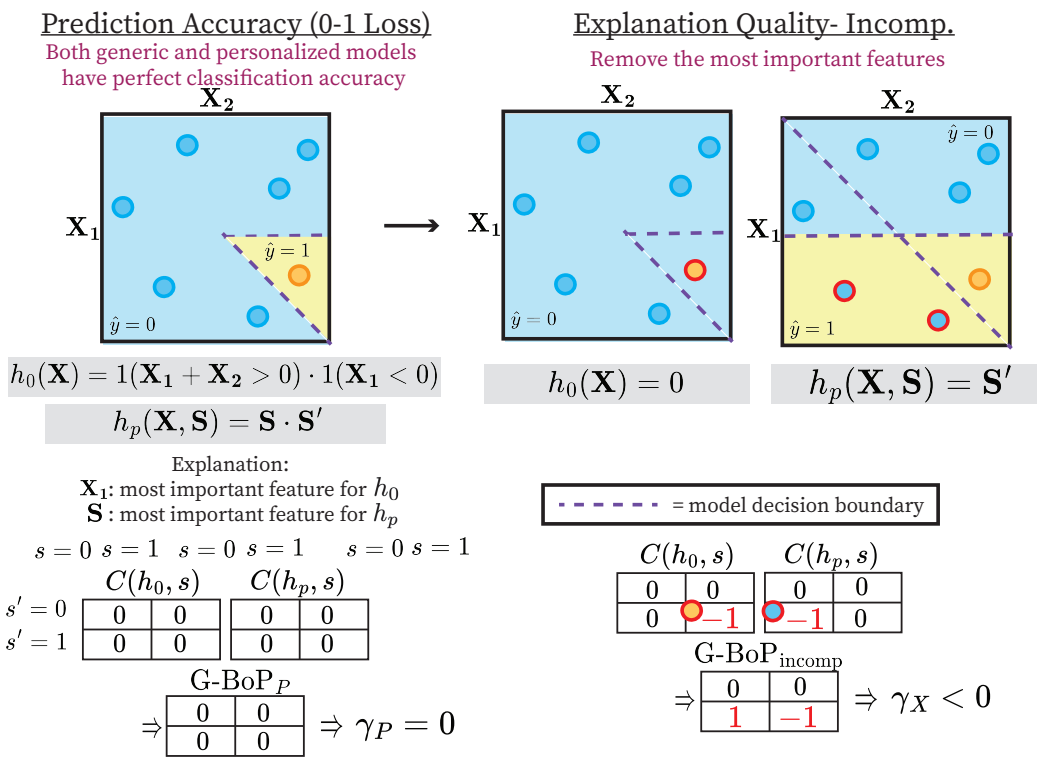

Figure 8: Comparing a generic model (h_0) and a personalized model (h_p) on prediction and explanation (incomprehensiveness). Both achieve perfect accuracy, but removing each most important features yields different prediction performances. We find that $\gamma_P = 0$ while $\gamma_X < 0$.

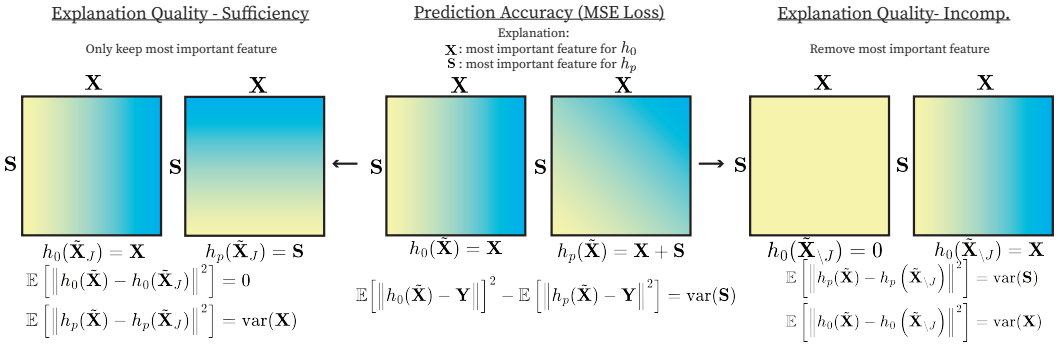

Figure 9: For a linear model, absence of benefit in explanation quality means that there is also an absence of benefit in prediction accuracy, as illustrated here (see Theorem 4.3). We consider a linear model $\mathbf{Y} = \mathbf{X} + \mathbf{S} + \epsilon$, with h_0 and h_p Bayes optimal regressors. In this example, absence of benefit of personalization for the explanation quality, BoP-X$^{\text{suff}} = 0$ evaluated in terms of sufficiency (left column) means: $\mathbb{E}[\|h_0(\tilde{\mathbf{X}}) - h_0(\tilde{\mathbf{X}}_\mathbf{J})\|^2] = \mathbb{E}[\|h_p(\tilde{\mathbf{X}}) - h_p(\tilde{\mathbf{X}}_\mathbf{J})\|^2] \Rightarrow \text{var}(\mathbf{X}) = 0$. Then, absence of benefit of personalization for the explanation quality, BoP-X$^{\text{comp}} = 0$ evaluated in terms of comprehensiveness (right column) means: $\mathbb{E}[\|h_0(\tilde{\mathbf{X}}) - h_0(\tilde{\mathbf{X}}_{\setminus\mathbf{J}})\|^2] = \mathbb{E}[\|h_0(\tilde{\mathbf{X}}) - h_0(\tilde{\mathbf{X}}_{\setminus\mathbf{J}})\|^2] \Rightarrow \text{var}(\mathbf{S}) = \text{var}(\mathbf{X}) \Rightarrow \text{var}(\mathbf{S}) = 0$. This allows us to conclude that, in terms of prediction accuracy (middle column): MSE$_0$ = MSE$_p$ and hence there is also no benefit of personalization in prediction :BoP-P $= 0$.

We define J_0 and J_p as a set of important features for h_0 and h_p. Note that J_0 and J_p are the same across all samples for the additive model. Then, the sufficiency of the explanation for h_0 and h_p is written as:

$$\mathbb{E}[\|h_0(\tilde{\mathbf{X}}) - h_0(\tilde{\mathbf{X}}_{\mathbf{J_0}})\|^2] = \sum_{\substack{j \in \setminus J_0, \\ j \leq t}} \alpha_j^2 \text{Var}(\mathbf{X_t}) \tag{12}$$

$$\mathbb{E}[\|h_p(\tilde{\mathbf{X}}) - h_p(\tilde{\mathbf{X}}_{\mathbf{J_p}})\|^2] = \sum_{\substack{j \in \setminus J_p, \\ j \leq t}} \alpha_j^2 \text{Var}(\mathbf{X_t}) + \sum_{\substack{j \in \setminus J_p, \\ j \geq t+1}} \alpha_j^2 \text{Var}(\mathbf{S_{j-t}}). \tag{13}$$

Similarly, the comprehensiveness of the explanation for h_0 and h_p is written as:

$$\mathbb{E}[\|h_0(\tilde{\mathbf{X}}) - h_0(\tilde{\mathbf{X}}_{\setminus J_0})\|^2] = \sum_{\substack{j \in J_0, \\ j \leq t}} \alpha_j^2 \text{Var}(\mathbf{X}_t) \tag{14}$$

$$\mathbb{E}[\|h_p(\tilde{\mathbf{X}}) - h_p(\tilde{\mathbf{X}}_{\setminus J_p})\|^2] = \sum_{\substack{j \in J_p, \\ j \leq t}} \alpha_j^2 \text{Var}(\mathbf{X}_t) + \sum_{\substack{j \in J_p, \\ j \geq t+1}} \alpha_j^2 \text{Var}(\mathbf{S}_{j-t}). \tag{15}$$

Then, our assumption of BoP-X $= 0$ for sufficiency becomes:

$$\mathbb{E}[\|h_0(\tilde{\mathbf{X}}) - h_0(\tilde{\mathbf{X}}_{J_0})\|^2] = \mathbb{E}[\|h_p(\tilde{\mathbf{X}}) - h_p(\tilde{\mathbf{X}}_{J_p})\|^2] \tag{16}$$

$$\Rightarrow \sum_{\substack{j \in \setminus J_0, \\ j \leq t}} \alpha_j^2 \text{Var}(\mathbf{X}_t) = \sum_{\substack{j \in \setminus J_p, \\ j \leq t}} \alpha_j^2 \text{Var}(\mathbf{X}_t) + \sum_{\substack{j \in \setminus J_p, \\ j \geq t+1}} \alpha_j^2 \text{Var}(\mathbf{S}_{j-t}) \tag{17}$$

Similarly, our assumption of BoP-X $= 0$ for comprehensiveness becomes:

$$\mathbb{E}[\|h_0(\tilde{\mathbf{X}}) - h_0(\tilde{\mathbf{X}}_{\setminus J_0})\|^2] = \mathbb{E}[\|h_p(\tilde{\mathbf{X}}) - h_p(\tilde{\mathbf{X}}_{\setminus J_p})\|^2] \tag{18}$$

$$\Rightarrow \sum_{\substack{j \in J_0, \\ j \leq t}} \alpha_j^2 \text{Var}(\mathbf{X}_t) = \sum_{\substack{j \in J_p, \\ j \leq t}} \alpha_j^2 \text{Var}(\mathbf{X}_t) + \sum_{\substack{j \in J_p, \\ j \geq t+1}} \alpha_j^2 \text{Var}(\mathbf{S}_{j-t}). \tag{19}$$

Summing both equations:

$$\begin{aligned} \sum_{\substack{j \in \setminus J_0 \\ j \leq t}} \alpha_j^2 \ \text{Var}(\mathbf{X}_t) + \sum_{\substack{j \in J_0 \\ j \leq t}} \alpha_j^2 \ \text{Var}(\mathbf{X}_t) &= \sum_{\substack{j \in \setminus J_p \\ j \leq t}} \alpha_j^2 \ \text{Var}(\mathbf{X}_t) + \sum_{\substack{j \in \setminus J_p \\ j \geq t+1}} \alpha_j^2 \ \text{Var}(\mathbf{S}_{j-t}) \\ &\quad + \sum_{\substack{j \in J_p \\ j \leq t}} \alpha_j^2 \ \text{Var}(\mathbf{X}_t) + \sum_{\substack{j \in J_p \\ j \geq t+1}} \alpha_j^2 \ \text{Var}(\mathbf{S}_{j-t}) \\ \Rightarrow \text{Var}(\mathbf{X}) &= \text{Var}(\mathbf{X}) + \text{Var}(\mathbf{S}) \\ \Rightarrow \text{Var}(\mathbf{S}) &= 0. \end{aligned} \tag{20}$$

Since $\text{Var}(\mathbf{S}) = 0$, we have that $\text{MSE}(h_0) = \text{MSE}(h_p)$ and thus: BoP-P $= 0$ which concludes the proof.

We can make the same claim with similar logic for a classifier where $\mathbf{Y}$ is given as:

$$\mathbf{Y} = \mathbb{1}(\alpha_1 \mathbf{X}_1 + \cdots \alpha_t \mathbf{X}_t + \alpha_{t+1} \mathbf{S}_1 + \cdots + \alpha_{t+k} \mathbf{S}_k + \epsilon > 0). \tag{21}$$

The derivations above are made at the population level, i.e., without distinguishing subgroups in the data. However, the reasoning also applies for subgroups, where we define subgroups to be defined by $\mathbb{1}(\mathbf{S} \geq 0)$ taking values in $\{0, 1\}$. In other words, if G-BoP$_{\text{suff}}(h_0, h_p, s) = 0$ and G-BoP$_{\text{incomp}}(h_0, h_p, s) = 0$ then G-BoP$_P(h_0, h_p, s) = 0$ for any $s \in \{0, 1\}$. However, we note that we can only make a statement on $\gamma(h_0, h_p)$ (prediction accuracy) for the case where $\gamma_{\text{sufficiency}}(h_0, h_p) = 0$ and $\gamma_{\text{incomprehensiveness}}(h_0, h_p) = 0$ if the following is true: the group realizing the minima in the three γ's is the same group. □

D Proof of Theorems on Lower Bounds for the Probability of Error

As in (Monteiro Paes et al., 2022), we will prove every theorem for the flipped hypothesis test defined as:

$$H_0: \quad \gamma(h_0, h_p; \mathcal{D}) \leq \epsilon \quad \Leftrightarrow \quad \text{Personalized } h_p \text{ performs worst: yields } \epsilon < 0 \text{ disadvantage}$$
$$H_1: \quad \gamma(h_0, h_p; \mathcal{D}) \geq 0 \quad \Leftrightarrow \quad \text{Personalized } h_p \text{ performs at least as good as generic } h_0.$$

where we emphasize that $\epsilon < 0$.

As shown in (Monteiro Paes et al., 2022), proving the bound for the original hypothesis test is equivalent to proving the bound for the flipped hypothesis test, since estimating γ is as hard as estimating $-\gamma$. In every section that follows, H_0, H_1 refer to the flipped hypothesis test.

Here, we first prove a proposition that is valid for all of the cases that we consider in the next sections.

Proposition D.1. *Consider $P_{\mathbf{X},\mathbf{S},y}$ is a distribution of data, for which the generic model h_0 performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) < 0$, and $Q_{\mathbf{X},\mathbf{S},y}$ is a distribution of data points for which the personalized model performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) > 0$. Consider a decision rule Ψ that represents any hypothesis test. We have the following bound on the probability of error P_e:*

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - TV(P \parallel Q),$$

for any well-chosen $P \in H_0$ and any well-chosen $Q \in H_1$. Here TV refers to the total variation between probability distributions P and Q.

Proof. Consider h_0 and h_p fixed. Take one decision rule Ψ that represents any hypothesis test. Consider a dataset such that H_0 is true, i.e., $\mathcal{D} \sim P_0$ and a dataset such that H_1 is true, i.e., $\mathcal{D} \sim P_1$.

It might seem weird to use two datasets to compute the same quantity P_e, i.e., one dataset to compute the first term in P_e, and one dataset to compute the second term in P_e. However, this is just a reflection of the fact that the two terms in P_e come from two different settings: H_0 true or H_0 false, which are disjoint events: in the same way that H_0 cannot be simultaneously true and false, yet each term in P_e consider one or the other case; then we use one or the other dataset.

We have:

$$\begin{aligned}
P_e &= \text{Pr}(\text{Rejecting } H_0 | H_0 \text{ true}) + \text{Pr}(\text{Failing to reject } H_0 | H_1 \text{ true}) \\
&= \text{Pr}(\Psi(h_0, h_p, \mathcal{D}, \epsilon) = 1 | \mathcal{D} \sim P_0) + \text{Pr}(\Psi(h_0, h_p, \mathcal{D}, \epsilon) = 0 | \mathcal{D} \sim P_1) \\
&= \text{Pr}(\Psi(\mathcal{D}) = 1 | \mathcal{D} \sim P_0) + \text{Pr}(\Psi(\mathcal{D}) = 0 | \mathcal{D} \sim P_1) \text{ simplifying notations} \\
&= 1 - \text{Pr}(\Psi(\mathcal{D}) = 0 | \mathcal{D} \sim P_0) + \text{Pr}(\Psi(\mathcal{D}) = 0 | \mathcal{D} \sim P_1) \text{ complementary event} \\
&= 1 - P_0(E_\Psi) + P_1(E_\Psi) \text{ writing } E_\Psi \text{ the event } \Psi(\mathcal{D}) = 0 \\
&= 1 - (P_0(E_\Psi) - P_1(E_\Psi))
\end{aligned}$$

Now, we will bound this quantity:

$$\begin{aligned}
\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e &= \min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} 1 - (P_0(E_\Psi) - P_1(E_\Psi)) \\
&\geq \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} \min_{\Psi} [1 - (P_0(E_\Psi) - P_1(E_\Psi))] \text{ using minmax inequality} \\
&= \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} \left[1 - \max_{\Psi} (P_0(E_\Psi) - P_1(E_\Psi))\right] \text{ to minimize over } \Psi, \text{ we maximize } (P_0(E_\Psi) - P_1(E_\Psi)) \\
&\geq \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} \left[1 - \max_{\text{events } A} (P_0(A) - P_1(A))\right] \text{ because the max is now over all possible events } A
\end{aligned}$$

The maximization is broadened to consider all possible events A. This increases the set over which the maximum is taken. Because Ψ is only a subset of all possible events, maximizing over all events

A (which includes Ψ) will result in a value that is at least as large as the maximum over Ψ. In other words, extending the set of possible events can only make the maximum greater or the same.

$$
\begin{aligned}
&= \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} [1 - TV(P_0 \parallel P_1)] \text{ by definition of the total variation (TV)} \\
&= 1 - \min_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} TV(P_0 \parallel P_1) \\
&\geq 1 - TV(P \parallel Q) \text{ for any } P \in H_0 \text{ and } Q \in H_1.
\end{aligned}
$$

This is true because the total variation distance $TV(P \parallel Q)$ for any particular pair P and Q cannot be smaller than the minimum total variation distance across all pairs. We recall that, by definition, the total variation of two probability distributions P, Q is the largest possible difference between the probabilities that the two probability distributions can assign to the same event A. □

Next, we prove a lemma that will be useful for the follow-up proofs.

Lemma D.2. *Consider a random variable a such that $\mathbb{E}[a] = 1$. Then:*

$$\mathbb{E}[(a-1)^2] = \mathbb{E}[a^2] - 1 \tag{22}$$

Proof. We have that:

$$
\begin{aligned}
\mathbb{E}[(a-1)^2] &= \mathbb{E}[a^2 - 2a + 1] \\
&= \mathbb{E}[a^2] - 2\mathbb{E}[a] + 1 \text{ (linearity of the expectation)} \\
&= \mathbb{E}[a^2] - 2 + 1(\mathbb{E}[a] = 1 \text{ by assumption}) \\
&= \mathbb{E}[a^2] - 1.
\end{aligned}
$$

□

D.1 Proof for any probability distribution and any number of samples in each group

Below, we find the lower bound for the probability of error for any probability distribution of the BoP, and any number of samples per group.

Theorem D.3 (Lower bound for any probability distribution BoP.)**.** *The lower bound writes:*

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\frac{p^\epsilon(B)}{p(B)} \right]^{m_j} - 1 \right]^{\frac{1}{2}} \tag{23}$$

where P_0 is a distribution of data, for which the generic model h_0 performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) < 0$, and P_1 is a distribution of data points for which the personalized model performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) \geq \epsilon$. Dataset $\mathcal{D}$ is drawn from an unknown distribution and has d groups where $d = 2^k$, with each group having m_j samples.

Proof. By Proposition D.1, we have that:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - TV(P \parallel Q)$$

for any well-chosen $P \in H_0$ and any well-chosen $Q \in H_1$. We will design two probability distributions P, Q defined on the N data points $(\mathbf{X_1}, \mathbf{S_1}, \mathbf{Y_1}), ..., (\mathbf{X_N}, \mathbf{S_N}, \mathbf{Y_N})$ of the dataset $\mathcal{D}$ to compute an interesting right hand side term. An "interesting" right hand side term is a term that makes the lower bound as tight as possible, i.e., it relies on distributions P, Q for which $TV(P \parallel Q)$ is small, i.e., probability distributions that are similar. To achieve this, we will first design the distribution $Q \in H_1$, and then propose P as a very small modification of Q, just enough to allows it to verify $P \in H_0$.

Mathematically, P, Q are distributions on the dataset $\mathcal{D}$, i.e., on N i.i.d. realizations of the random variables $\mathbf{X}, \mathbf{S}, \mathbf{Y}$. Thus, we wish to design probability distributions on $(\mathbf{X_1}, \mathbf{S_1}, \mathbf{Y_1}), ..., (\mathbf{X_N}, \mathbf{S_N}, \mathbf{Y_N})$.

However, we note that the dataset distribution is only meaningful in terms of how each triplet $(\mathbf{X_i}, \mathbf{S_i}, \mathbf{Y_i})$ impacts the value of the individual BOP $\mathbf{B_i}$. Indeed, since $\mathbf{B_i}$ is a function of the data point $\mathbf{Z_i} = (\mathbf{X_i}, \mathbf{S_i}, \mathbf{Y_i})$, that we denote f such that $\mathbf{B_i} = f(\mathbf{Z_i})$, any probability distribution on $\mathbf{Z_i}$ will yield a probability distribution on $\mathbf{B_i}$ and any distribution on the dataset $\mathbf{Z_1}, ..., \mathbf{Z_N}$ will yield a distribution on $\mathbf{B_1}, ..., \mathbf{B_N}$.

Conversely, let be given $\tilde{P}(b_1, ..., b_N) = \Pi_{i=1}^N \tilde{P}_i(b_i)$ a distribution on $\mathbf{B_1}, ..., \mathbf{B_N}$ defined by N independent distributions $\tilde{P}_i$ for $i = 1, .., N$, such that the support of each $\tilde{P}_i$ is restricted to the image of f. We propose to build a probability distribution $P(z_1, ..., z_N) = \Pi_{i=1}^N P_i(z_i)$ on $\mathbf{Z_1}, ..., \mathbf{Z_N}$ that will ensure that $f(\mathbf{Z_1}), ..., f(\mathbf{Z_N})$ is distributed as $\tilde{P}$.

First, for each P_i we restrict P_i so that, for every value b_i that $\mathbf{B_i}$ can take according to $\tilde{P}_i$, there exists a unique z_i with positive density, concentrated as a Dirac at z_i, and such that we have $f(z_i) = b_i$. Existence is guaranteed since $\mathbf{B_i}$ takes values in the image of f. Uniqueness is guaranteed because we can assign 0 mass to the potential non-unique values. Equivalently, f is a bijection from $\text{supp}(P_i)$ to the set of values taken by $\mathbf{B}_i$ for each i.

Next, for all $z_i \in \text{supp}(P_i)$, we explicitly construct $P_i(z_i)$ as follows:

$$P_i(z_i) = \tilde{P}_i(f_i(z_i)) \cdot \left| \left(\frac{df_i^{-1}(b_i)}{db_i} \right) \right|^{-1},$$

where f_i now denotes the restriction of f to $\text{supp}(P_i)$. We construct Q_i analogously for any $i = 1, ..., N$.

Now moving back to the full dataset of N samples, we relate the TV between P and Q over the full dataset $\mathbf{Z} = \mathbf{Z_1}, \cdots, \mathbf{Z_N}$ to the TV between $\tilde{P}$ and $\tilde{Q}$ over $\mathbf{B} = \mathbf{B_1}, \cdots, \mathbf{B_N}$ by a change of variables:

$$\begin{aligned}
TV(P \parallel Q) &= \tfrac{1}{2} \int |P(z_1, \ldots, z_N) - Q(z_1, \ldots, z_N)| \, dz_1 \cdots dz_N \\
&= \tfrac{1}{2} \int \left| \prod_{i=1}^N P_i(z_i) - \prod_{i=1}^N Q_i(z_i) \right| dz_1 \cdots dz_N \\
&\quad \text{where } (z_1, \ldots, z_N) = F(b_1, \ldots, b_N), \quad \text{and } F(b_1, \ldots, b_N) = \big(f_1^{-1}(b_1), \ldots, f_N^{-1}(b_N)\big). \\
&= \tfrac{1}{2} \int_{b_1 \cdots b_N} \left| \prod_{i=1}^N P_i\big(f_i^{-1}(b_i)\big) - \prod_{i=1}^N Q_i\big(f_i^{-1}(b_i)\big) \right| \cdot |\det\left(\mathbf{J_F}(b_1, \ldots, b_N)\right)| \, db_1 \ldots db_N \\
&= \tfrac{1}{2} \int_{b_1 \cdots b_N} \left| \prod_{i=1}^N P_i\big(f_i^{-1}(b_i)\big) - \prod_{i=1}^N Q_i\big(f_i^{-1}(b_i)\big) \right| \cdot \prod_{i=1}^N \frac{\partial z_i(b_i)}{\partial b_i} \, db_1 \cdots db_N \\
&= \tfrac{1}{2} \int_{b_1 \cdots b_N} \left| \prod_{i=1}^N P_i\big(f_i^{-1}(b_i)\big) - \prod_{i=1}^N Q_i\big(f_i^{-1}(b_i)\big) \right| \prod_{i=1}^N \left(\frac{df_i^{-1}(b_i)}{db_i} \right) db_1 \cdots db_N \\
&= \tfrac{1}{2} \int_{b_1 \cdots b_N} \left| \left[\prod_{i=1}^N \frac{df_i^{-1}}{d\mathrm{b}}(b_i) \right] \left[\prod_{i=1}^N P_i\big(f_i^{-1}(b_i)\big) \right] - \left[\prod_{i=1}^N \frac{df_i^{-1}}{db_i}(b_i) \right] \left[\prod_{i=1}^N Q_i\big(f_i^{-1}(b_i)\big) \right] \right| db_1 \cdots db_N \\
&= \tfrac{1}{2} \int_{b_1, \ldots, b_N} \left| \prod_{i=1}^N P_i\big(f_i^{-1}(b_i)\big) \cdot \left| \frac{df_i^{-1}(b_i)}{db_i} \right| - \prod_{i=1}^N Q_i\big(f_i^{-1}(b_i)\big) \cdot \left| \frac{df_i^{-1}(b_i)}{db_i} \right| \right| db_1 \cdots db_N \\
&= \tfrac{1}{2} \int_{b_1, \ldots, b_N} \left| \prod_{i=1}^N \widetilde{P}_i(b_i) - \prod_{i=1}^N \widetilde{Q}_i(b_i) \right| db_1 \cdots db_N \quad \text{by def. of } P_i \text{ and } \widetilde{P}_i \text{ for all } i.
\end{aligned}$$

Thus, we design probability distributions P, Q on n i.i.d. realizations of an auxiliary random variable $\mathbf{B}$, with values in $\mathbb{R}$, defined as:

$$\mathbf{B} = \ell(h_0(\mathbf{X}), \mathbf{Y}) - \ell(h_p(\mathbf{X}, \mathbf{S}), \mathbf{Y}). \tag{24}$$

Intuitively, $\mathbf{B}_i$ represents how much the triplet $(\mathbf{X_i}, \mathbf{S_i}, \mathbf{Y_i})$ contributes to the value of the BOP. $b_i > 0$ means that the personalized model provided a better prediction than the generic model on the triplet (x_i, s_i, y_i) corresponding to the data point i.

Consider the event $b = (b_1, ..., b_N) \in \mathbb{R}^N$ of N realizations of $\mathbf{B}$. For simplicity in our computations, we divide this event into the d groups, i.e., we write instead: $b_j = (b_j^{(1)}, ..., b_j^{(m)})$, since each group j has m_j samples. Thus, we have: $b = \{b_j^{(k)}\}_{j=1...d, k=1...m}$ indexed by j, k where $j = 1...d$ is the group in which this element is, and $k = 1...m_j$ is the index of the element in that group.

Design Q. Next, we continue designing a distribution Q (since we have justified that we can define them on $\mathbf{B}$) on this set of events that will (barely) verify H_1, i.e., such that the expectation of B according to Q will give $\gamma = 0$. We recall that $\gamma = 0$ means that the minimum benefit across groups is 0, implying that there might be some groups that have a > 0 benefit.

Given p as a distribution with mean $\mu = 0$, we propose the following distribution for Q

$$Q_j(b_j) = \prod_{k=1}^{m} p(b_j^{(k)}), \text{ for every group } j = 1....d$$

$$Q(b) = \prod_{j=1}^{d} Q_j(b_j).$$

We verify that we have designed Q correctly, i.e., we verify that $Q \in H_1$. When the dataset is distributed according to Q, we have:

$$\begin{aligned}
\gamma &= \min_{s \in S} C_s(h_0, s) - C_s(h_p, s) \\
&= \min_{s \in S} \mathbb{E}_Q[\ell(h_0(\mathbf{X}), \mathbf{Y}) \mid \mathbf{S} = s] - \mathbb{E}_Q[\ell(h_p(\mathbf{X}), \mathbf{Y}) \mid \mathbf{S} = s] \text{ (by definition of group cost)} \\
&= \min_{s \in S} \mathbb{E}_Q[\ell(h_0(\mathbf{X}), \mathbf{Y}) - \ell(h_p(\mathbf{X}), \mathbf{Y}) \mid \mathbf{S} = s] \text{ (by linearity of expectation)} \\
&= \min_{s \in S} \mathbb{E}_Q[B \mid \mathbf{S} = s] \text{ (by definition of random variable } \mathbf{B}) \\
&= \min_{s \in S} 0 \text{ (by definition of the probability distribution on } \mathbf{B}) \\
&= 0.
\end{aligned}$$

Thus, we find that $\gamma = 0$ which means that $\gamma \geq 0$, i.e., $Q \in H_1$.

Design P. Next, we design P as a small modification of the distribution Q, that will just be enough to get $P \in H_0$. We recall that $P \in H_0$ means that $\gamma \leq \epsilon$ where $\epsilon < 0$ in the flipped hypothesis test. This means that, under H_0, there is one group that suffers a decrease of performance of $|\epsilon|$ because of the personalized model.

Given p as a distribution with $\mu = 0$, and p^ϵ a distribution with mean $\mu = \epsilon < 0$, we have:

$$P_j(b_j) = \prod_{k=1}^{m_j} p(b_j^{(k)}), \text{ for every group } j = 1....d,$$

$$P_j^\epsilon(b_j) = \prod_{k=1}^{m_j} p^\epsilon(b_j^{(k)}), \text{ for every group } j = 1....d,$$

$$P(b) = \frac{1}{d} \sum_{j=1}^{d} P_j^\epsilon(b_j) \prod_{j' \neq j} P_{j'}(b_{j'}).$$

Intuitively, this distribution represents the fact that there is one group for which the personalized model worsen performances by $|\epsilon|$.We assume that this group can be either group 1, or group 2,

etc, or group d, and consider these to be disjoint events: i.e., exactly only one group suffers the $|\epsilon|$ performance decrease. We take the union of these disjoint events and sum of probabilities using the Partition Theorem (Law of Total Probability) in the definition of P above.

We verify that we have designed P correctly, i.e., we verify that $P \in H_0$. When the dataset is distributed according to P, we have:

$$\begin{aligned}\gamma &= \min_{s \in S} C_s(h_0, s) - C_s(h_p, s) \\ &= \min_{s \in S} \mathbb{E}_P[\mathbf{B} \mid \mathbf{S} = s] \text{ (same computations as for } Q \in H_1) \\ &= \min(\epsilon, 0, ..., 0) \text{ (since exactly one group has mean } \epsilon) \\ &= \epsilon \text{ (since } \epsilon < 0).\end{aligned}$$

Thus, we find that $\gamma = \epsilon$ which means that $\gamma \leq 0$, i.e., $P \in H_0$.

Compute total variation $TV(P \parallel Q)$. We have verified that $Q \in H_1$ and that $P \in H_0$. We use these probability distributions to compute the lower bound to P_e. First, we compute their total variation:

$$\begin{aligned}TV(P \parallel Q) &= \frac{1}{2} \int_{b_1,...,b_j} |P(b_1, ..., b_j) - Q(b_1, ..., b_j)| \, db_1 ... db_j \text{ (TV for probability density functions)} \\ &= \frac{1}{2} \int_{b_1,...,b_j} \left| \frac{1}{d} \sum_{j=1}^{d} P_j^{\epsilon}(b_j) \prod_{j' \neq j} P_{j'}(b_{j'}) - \prod_{j=1}^{d} Q_j(b_j) \right| db_1 ... db_j \text{ (definition of } P, Q) \\ &= \frac{1}{2} \int_{b_1,...,b_j} \left| \frac{1}{d} \sum_{j=1}^{d} \frac{P_j^{\epsilon}(b_j)}{P_j(b_j)} \prod_{j'=1}^{d} P_{j'}(b_{j'}) - \prod_{j=1}^{d} Q_j(b_j) \right| db_1 ... db_j \text{ (adding missing } j' = j) \\ &= \frac{1}{2} \int_{b_1,...,b_j} \left| \frac{1}{d} \sum_{j=1}^{d} \frac{P_j^{\epsilon}(b_j)}{P_j(b_j)} \prod_{j'=1}^{d} Q_{j'}(b_{j'}) - \prod_{j=1}^{d} Q_j(b_j) \right| db_1 ... db_j \text{ } (P_j = Q_j \text{ by construction}) \\ &= \frac{1}{2} \int_{b_1,...,b_j} \prod_{j=1}^{d} Q_j(b_j) \left| \frac{1}{d} \sum_{j=1}^{d} \frac{P_j^{\epsilon}(b_j)}{P_j(b_j)} - 1 \right| db_1 ... db_j \text{ (extracting the product)} \\ &= \frac{1}{2} \mathbb{E}_Q \left[\left| \frac{1}{d} \sum_{j=1}^{d} \frac{P_j^{\epsilon}(b_j)}{P_j(b_j)} - 1 \right| \right] \text{ (recognizing an expectation with respect to } Q) \\ &= \frac{1}{2} \mathbb{E}_Q \left[\left| \frac{1}{d} \sum_{j=1}^{d} \frac{\prod_{k=1}^{m_j} p^{\epsilon}(b_j^{(k)})}{\prod_{k=1}^{m_j} p(b_j^{(k)})} - 1 \right| \right] \text{ (definition of } P_j \text{ and } P_j^{(\epsilon)}) \\ &\leq \frac{1}{2} \mathbb{E}_Q \left[\left| \frac{1}{d} \sum_{j=1}^{d} \frac{\prod_{k=1}^{m_j} p^{\epsilon}(b_j^{(k)})}{\prod_{k=1}^{m_j} p(b_j^{(k)})} - 1 \right|^2 \right]^{1/2} \text{ (Cauchy-Schwartz)}\end{aligned}$$

Auxiliary computation to apply Lemma D.2 Next, we will apply Lemma D.2. For this, we need to prove that the expectation of the first term is 1. We have:

$$\begin{aligned}&\mathbb{E}_Q \left[\frac{1}{d} \sum_{j=1}^{d} \frac{\prod_{k=1}^{m_j} p^{\epsilon}(b_j^{(k)})}{\prod_{k=1}^{m_j} p(b_j^{(k)})} \right] \\ &= \frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_Q \left[\frac{\prod_{k=1}^{m_j} p^{\epsilon}(b_j^{(k)})}{\prod_{k=1}^{m_j} p(b_j^{(k)})} \right] \text{ (linearity of expectation)}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{d}\sum_{j=1}^{d}\mathbb{E}_Q\left[\prod_{k=1}^{m_j}\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right] \text{ (rearranging the product)}\\
&= \frac{1}{d}\sum_{j=1}^{d}\prod_{k=1}^{m_j}\mathbb{E}_Q\left[\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right] \text{ (product of independent variables)}\\
&= \frac{1}{d}\sum_{j=1}^{d}\prod_{k=1}^{m_j}\mathbb{E}_p\left[\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right] \text{ (definition of } Q)\\
&= \frac{1}{d}\sum_{j=1}^{d}\prod_{k=1}^{m_j}\int_{-\infty}^{+\infty}\frac{p^\epsilon(b)}{p(b)}p(b)db \text{ (definition of expectation in } p)\\
&= \frac{1}{d}\sum_{j=1}^{d}\prod_{k=1}^{m_j}\int_{-\infty}^{+\infty}p^\epsilon(b)db \text{ (simplify)}\\
&= \frac{1}{d}\sum_{j=1}^{d}\prod_{k=1}^{m_j}1 \text{ (probability density function integrates to 1)}\\
&= \frac{1}{d}\sum_{j=1}^{d}1 \text{ (term independent of } k)\\
&= \frac{1}{d}d \text{ (term independent of } j)\\
&= 1.
\end{aligned}
$$

Continue by applying Lemma D.2. This auxiliary computation shows that we meet the assumption of Lemma D.2. Thus, we continue the computation of the lower bound of the TV by applying Lemma D.2.

$$
\begin{aligned}
&TV(P \parallel Q)\\
&\leq \frac{1}{2}\mathbb{E}_Q\left[\left(\frac{1}{d}\sum_{j=1}^{d}\frac{\prod_{k=1}^{m_j}p^\epsilon(b_j^{(k)})}{\prod_{k=1}^{m_j}p(b_j^{(k)})}\right)^2-1\right]^{\frac{1}{2}} \text{ Lemma D.2}\\
&= \frac{1}{2}\mathbb{E}_Q\left[\left(\frac{1}{d}\sum_{j=1}^{d}z_j\right)^2-1\right]^{\frac{1}{2}} \text{ defining } z_j=\frac{\prod_{k=1}^{m_j}p^\epsilon(b_j^{(k)})}{\prod_{k=1}^{m_j}p(b_j^{(k)})}=\prod_{k=1}^{m_j}\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\\
&= \frac{1}{2}\mathbb{E}_Q\left[\frac{1}{d^2}\sum_{j,j'=1}^{d}z_jz_{j'}-1\right]^{\frac{1}{2}} \text{ expanding the square of the sum}\\
&= \frac{1}{2}\mathbb{E}_Q\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}z_j^2+\sum_{j,j'=1,j\neq j'}^{d}z_j.z_{j'}\right)-1\right]^{\frac{1}{2}},
\end{aligned}
$$

where we split the double sum to get independent variables in the second term.

We get by linearity of the expectation, $\mathbb{E}[aX+bY]=a\mathbb{E}[X]+b\mathbb{E}[Y]$:

$$
\begin{aligned}
&TV(P \parallel Q)\\
&\leq \frac{1}{2}\mathbb{E}_Q\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}z_j^2+\sum_{j,j'=1,j\neq j'}^{d}z_j.z_{j'}\right)-1\right]^{\frac{1}{2}}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}\mathbb{E}_Q[z_j^2] + \sum_{j,j'=1,j\neq j'}^{d}\mathbb{E}_Q[z_j.z_{j'}]\right) - 1\right]^{\frac{1}{2}} \\
&= \frac{1}{2}\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}\mathbb{E}_Q\left[\left(\prod_{k=1}^{m_j}\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right)^2\right] + \sum_{j,j'=1,j\neq j'}^{d}\mathbb{E}_Q\left[\left(\prod_{k=1}^{m_j}\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right).\left(\prod_{k=1}^{m_j}\frac{p^\epsilon(b_{j'}^{(k)})}{p(b_{j'}^{(k)})}\right)\right]\right) - 1\right]^{\frac{1}{2}} \\
&= \frac{1}{2}\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}\mathbb{E}_Q\left[\left(\prod_{k=1}^{m_j}\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right)^2\right]\right. \\
&\quad \left. + \sum_{\substack{j,j'=1\\ j\neq j'}}^{d}\mathbb{E}_Q\left[\prod_{k=1}^{m_j}\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right]\mathbb{E}_Q\left[\prod_{k=1}^{m_{j'}}\frac{p^\epsilon(b_{j'}^{(k)})}{p(b_{j'}^{(k)})}\right]\right) - 1\right]^{\frac{1}{2}} \quad \text{(product of independent variables)} \\
&= \frac{1}{2}\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}\prod_{k=1}^{m_j}\mathbb{E}_p\left[\left(\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right)^2\right]\right. \\
&\quad \left. + \sum_{\substack{j,j'=1\\ j\neq j'}}^{d}\prod_{k=1}^{m_j}\mathbb{E}_p\left[\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right]\prod_{k=1}^{m_{j'}}\mathbb{E}_p\left[\frac{p^\epsilon(b_{j'}^{(k)})}{p(b_{j'}^{(k)})}\right]\right) - 1\right]^{\frac{1}{2}} \quad \text{(product of independent variables and def. of } Q\text{)} \\
&= \frac{1}{2}\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}\prod_{k=1}^{m_j}\mathbb{E}_p\left[\left(\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right)^2\right] + \sum_{j,j'=1,j\neq j'}^{d}\prod_{k=1}^{m_j}1\prod_{k=1}^{m_j}1\right) - 1\right]^{\frac{1}{2}} \quad \text{(auxiliary computation below)} \\
&= \frac{1}{2}\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}\prod_{k=1}^{m_j}\mathbb{E}_p\left[\left(\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right)^2\right] + \sum_{j,j'=1,j\neq j'}^{d}1\right) - 1\right]^{\frac{1}{2}} \quad \text{(term independent of } k\text{)} \\
&= \frac{1}{2}\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}\prod_{k=1}^{m_j}\mathbb{E}_p\left[\left(\frac{p^\epsilon(b_j^{(k)})}{p(b_j^{(k)})}\right)^2\right] + (d^2-d)\right) - 1\right]^{\frac{1}{2}} \quad \text{(term independent of } j\text{)} \\
&= \frac{1}{2}\left[\frac{1}{d^2}\left(\sum_{j=1}^{d}\mathbb{E}_p\left[\left(\frac{p^\epsilon(B)}{p(B)}\right)^2\right]^{m_j} + (d^2-d)\right) - 1\right]^{\frac{1}{2}} \quad \text{(term independent of } k\text{)} \\
&= \frac{1}{2}\left[\frac{1}{d^2}\sum_{j=1}^{d}\mathbb{E}_p\left[\left(\frac{p^\epsilon(B)}{p(B)}\right)^2\right]^{m_j} + 1 - \frac{1}{d} - 1\right]^{\frac{1}{2}} \quad \text{(distribute } 1/d^2\text{)} \\
&= \frac{1}{2}\left[\frac{1}{d^2}\sum_{j=1}^{d}\mathbb{E}_p\left[\left(\frac{p^\epsilon(B)}{p(B)}\right)^2\right]^{m_j} - \frac{1}{d}\right]^{\frac{1}{2}} \quad \text{(simplify)} \\
&= \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\mathbb{E}_p\left[\left(\frac{p^\epsilon(B)}{p(B)}\right)^2\right]^{m_j} - 1\right]^{\frac{1}{2}} \quad \text{(extract } 1/\sqrt{d}\text{)} \\
&= \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(\int_{-\infty}^{+\infty}\left(\frac{p^\epsilon(b)}{p(b)}\right)^2 p(b)db\right)^{m_j} - 1\right]^{\frac{1}{2}} \quad \text{(definition of expectation)}
\end{aligned}
$$

$$= \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \left(\int_{-\infty}^{+\infty} \frac{p^\epsilon(b)^2}{p(b)} db \right)^{m_j} - 1 \right]^{\frac{1}{2}} \text{ (simplify } p(b)\text{)}$$

$$= \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\frac{p^\epsilon(B)}{p(B)} \right]^{m_j} - 1 \right]^{\frac{1}{2}} \text{ (def of expectation)}$$

Auxiliary computation in 1 We show that:

$$\begin{aligned} &\mathbb{E}_p \left[\frac{p^\epsilon(b_{j'}^{(k)})}{p(b_{j'}^{(k)})} \right] \\ &= \int_{-\infty}^{+\infty} \frac{p^\epsilon(b)}{p(b)} p(b) db \\ &= \int_{-\infty}^{+\infty} p^\epsilon(b) db \text{ simplify } p(b) \\ &= 1 \text{ probability density function } p^\epsilon \text{ integrates to 1.} \end{aligned}$$

Final result: This gives the final result:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - TV(P \parallel Q)$$

$$\Rightarrow \min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\frac{p^\epsilon(B)}{p(B)} \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$

□

D.2 Proof for distribution in an exponential family

We consider a fixed exponential family in it natural parameterization, i.e., probability distributions of the form:

$$f_X(x \mid \boldsymbol{\theta}) = h(x) \exp(\theta \cdot \mathbf{T}(x) - A(\boldsymbol{\theta})), \tag{25}$$

where θ is the only parameter varying between two distributions from that family, i.e., the functions η, T and A are fixed. We recall a few properties of any exponential family (EF) that will be useful in our computations.

First, the moment generating function (MGF) for the natural sufficient statistic $T(x)$ is equal to:

$$M^T(t) = \exp\left(A(\theta + t) - A(\theta)\right).$$

Then, the moments for $T(x)$, when θ is a scalar parameter, are given by:

$$\begin{aligned} \mathrm{E}[T] &= A'(\theta) \\ \mathrm{V}[T] &= A''(\theta). \end{aligned}$$

Since the variance is non-negative $\mathrm{V}[T] \geq 0$, this means that we have $A''(\theta) > 0$ and thus A' is monotonic and bijective. We will use that fact in the later computations.

In the following, we recall that the categorical distribution and the Gaussian distribution with fixed variance σ^2 are members of the exponential family.

Example: Categorical distributions as a EF The categorical variable has probability density function:

$$p(x \mid \pi) = \exp\left(\sum_{k=1}^{K} x_k \log \pi_k \right)$$

$$= \exp\left(\sum_{k=1}^{K-1} x_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} x_k\right) \log\left(1 - \sum_{k=1}^{K-1} \pi_k\right)\right)$$
$$= \exp\left(\sum_{k=1}^{K-1} \log\left(\frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k}\right) x_k + \log\left(1 - \sum_{k=1}^{K-1} \pi_k\right)\right)$$

where we have used the fact that$\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$.

We note that we need to use the PDF of the categorical that uses a minimal (i.e., $K-1$) set of parameters. We define $h(x), T(x), \theta \in \mathbb{R}^{K-1}$ and $A(\theta)$ as:

$$h(x) = 1$$
$$T(x) = x,$$
$$\theta_k = \log\left(\frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k}\right) = \log\left(\frac{\pi_k}{\pi_K}\right), \text{ for } k = 1, ..., K-1$$
$$A(\theta) = -\log\left(1 - \sum_{k=1}^{K-1} \pi_k\right) = \log\left(\frac{1}{1 - \sum_{k=1}^{K-1} \pi_k}\right) = \log\left(\frac{\sum_{k=1}^{K} \pi_k}{1 - \sum_{k=1}^{K-1} \pi_k}\right) = \log\left(\sum_{k=1}^{K} e^{\theta_k}\right),$$

which shows that the categorical distribution is within the EF. For convenience we have defined θ_K setting it to 0 as per the Equation above.

Now, we adapt these expressions for the case of a Categorical variable with only $K = 3$ values $x_1 = -1, x_2 = 1$ and $x_3 = 0$ such that $\pi_3 = 0$, i.e., there is no mass on the $x_3 = 0$, and we denote $\pi_1 = p_1$ and $\pi_2 = p_2$ and $\pi_3 = 1 - p_1 - p_2 = 0$. We get:

$$h(x) = 1$$
$$T(x) = x,$$
$$\theta_1 = \log\left(\frac{p_1}{p_2}\right), \text{ and } \theta_2 = 0 \text{ by convention, as above, } \theta_3 = \log\left(\frac{\pi_3}{p_2}\right) = -\infty$$
$$A(\theta_1) = \log\left(e^{\theta_1} + e^{\theta_2} + e^{\theta_3}\right) = \log\left(e^{\theta_1} + 1 + 0\right) = \log\left(e^{\log\left(\frac{p_1}{p_2}\right)} + 1\right) = \log\left(\frac{p_1}{p_2} + 1\right),$$

where, in the proofs, we will have $p_1 = \frac{1}{2} + \epsilon$ and $p_3 = \frac{1}{2} - \epsilon$ such that the expectation is $-1.(\frac{1}{2} + \epsilon) + 1.(\frac{1}{2} - \epsilon) = -2\epsilon$.

Example: Gaussian distribution with fixed variance as a EF The Gaussian distribution with fixed variance has probability density function:

$$p(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2}\right)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp\left(\frac{2x\mu - \mu^2}{2\sigma^2}\right)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp\left(\frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right).$$

We define $h(x), T(x), \theta \in \mathbb{R}$ and $A(\theta)$ as:

$$h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$
$$T(x) = x,$$
$$\theta = \frac{\mu}{\sigma^2}$$

$$A(\theta) = \frac{\mu^2}{2\sigma^2} = \frac{\sigma^2\theta^2}{2}.$$

which shows that the Gaussian distribution with fixed variance σ^2 is within the EF.

Proposition D.4. *The lower bound for the exponential family with any number of samples in each group writes:*

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \left(\frac{M_p(2\Delta\theta)}{M_p(\Delta\theta)^2} \right)^{m_j} - 1 \right]^{\frac{1}{2}}$$

Proof. By Theorem D.3, we have:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\frac{p^\epsilon(B)}{p(B)} \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$

Plug in the exponential family Under the assumption of an exponential family distribution for the random variable B, we have:

$$\begin{aligned}
&\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \\
&\geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\frac{h(B)\exp(\theta^\epsilon.T(B) - A(\theta^\epsilon))}{h(B)\exp(\theta^0.T(B) - A(\theta^0))} \right]^{m_j} - 1 \right]^{\frac{1}{2}} \\
&= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\frac{\exp(\theta^\epsilon.T(B) - A(\theta^\epsilon))}{\exp(\theta^0.T(B) - A(\theta^0))} \right]^{m_j} - 1 \right]^{\frac{1}{2}} \quad \text{simplifying } h \\
&= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\exp(\theta^\epsilon.T(B) - A(\theta^\epsilon)) \exp(-\theta^0.T(B) + A(\theta^0)) \right]^{m_j} - 1 \right]^{\frac{1}{2}} \quad \text{properties of } \exp \\
&= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\exp(A(\theta^0) - A(\theta^\epsilon)) \right. \right. \\
&\quad \left. \left. \cdot \exp\left((\theta^\epsilon - \theta^0) \cdot T(B)\right) \right]^{m_j} - 1 \right]^{\frac{1}{2}} \quad \text{(properties of } \exp \text{ and rearranging terms)} \\
&= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \exp(A(\theta^0) - A(\theta^\epsilon))^{m_j} \mathbb{E}_{p^\epsilon} \left[\exp((\theta^\epsilon - \theta^0)T(B)) \right]^{m_j} - 1 \right]^{\frac{1}{2}} \\
&= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \exp(A(\theta^0) - A(\theta^\epsilon))^{m_j} M_{p^\epsilon}(\Delta\theta)^{m_j} - 1 \right]^{\frac{1}{2}} \\
&\quad \text{(def. of MGF of } T(B)\text{: } M_{p^\epsilon}(t) = \mathbb{E}_{p^\epsilon}[\exp(t \cdot T(B))] \text{ with } \Delta\theta = \theta^\epsilon - \theta^0)
\end{aligned}$$

We define $\Delta\theta = \theta_\epsilon - \theta_0$. Here, we will apply the properties of EF regarding moment generating functions, i.e., for the p^ϵ with natural parameter θ_ϵ:

$$\begin{aligned}
M_{p^\epsilon}(t) = \exp\left(A(\theta_\epsilon + t) - A(\theta_\epsilon)\right) &\Rightarrow M_{p^\epsilon}(-\Delta\theta) = \exp\left(A(\theta_0) - A(\theta_\epsilon)\right), \\
&\Rightarrow M_{p^\epsilon}(\Delta\theta) = \exp\left(A(2\theta_\epsilon - \theta_0) - A(\theta_\epsilon)\right),
\end{aligned}$$

And, for p associated with natural parameter θ_0:

$$M_p(t) = \exp\left(A(\theta_0 + t) - A(\theta_0)\right) \Rightarrow M_p(-\Delta\theta) = \exp\left(A(2\theta_0 - \theta_\epsilon) - A(\theta_0)\right),$$

$$\Rightarrow M_p(\Delta\theta) = \exp\left(A(\theta_\epsilon) - A(\theta_0)\right),$$
$$\Rightarrow M_p(\Delta\theta)^2 = \exp\left(2A(\theta_\epsilon) - 2A(\theta_0)\right)$$
$$\Rightarrow M_p(2\Delta\theta) = \exp\left(A(2\theta_\epsilon - \theta_0) - A(\theta_0)\right)$$

So, that we have on the one hand:

$$M_{p^\epsilon}(-\Delta\theta)M_{p^\epsilon}(\Delta\theta) = \exp\left(A(\theta_0) - A(\theta_\epsilon)\right).\exp\left(A(2\theta_\epsilon - \theta_0) - A(\theta_\epsilon)\right)$$

and on the other hand:

$$\begin{aligned}
\frac{M_p(2\Delta\theta)}{M_p(\Delta\theta)^2} &= \frac{\exp\left(A(2\theta_\epsilon - \theta_0) - A(\theta_0)\right)}{\exp\left(2A(\theta_\epsilon) - 2A(\theta_0)\right)} \\
&= \frac{\exp\left(A(2\theta_\epsilon - \theta_0)\right)}{\exp\left(2A(\theta_\epsilon) - 2A(\theta_0)\right)} \cdot \frac{1}{\exp\left(A(\theta_0)\right)} \\
&= \frac{\exp\left(A(2\theta_\epsilon - \theta_0)\right)}{\exp\left(2A(\theta_\epsilon) - A(\theta_0)\right)} \\
&= \exp\left(A(2\theta_\epsilon - \theta_0) + A(\theta_0) - A(\theta_\epsilon) - A(\theta_\epsilon)\right) \\
&= \exp\left(A(\theta_0) - A(\theta_\epsilon) + A(2\theta_\epsilon - \theta_0) - A(\theta_\epsilon)\right) \\
&= \exp\left(A(\theta_0) - A(\theta_\epsilon)\right).\exp\left(A(2\theta_\epsilon - \theta_0) - A(\theta_\epsilon)\right)
\end{aligned}$$

Consequently, we get two equivalent expressions for our final result:

$$= 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\exp(A(\theta^0) - A(\theta^\epsilon))^{m_j}\exp\left(A(2\theta_\epsilon - \theta_0) - A(\theta_\epsilon)\right)^{m_j} - 1\right]^{\frac{1}{2}}$$

$$= 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(M_{p^\epsilon}(-\Delta\theta)M_{p^\epsilon}(\Delta\theta)\right)^{m_j} - 1\right]^{\frac{1}{2}} \quad \text{(first expression)}$$

$$= 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(\frac{M_p(2\Delta\theta)}{M_p(\Delta\theta)^2}\right)^{m_j} - 1\right]^{\frac{1}{2}} \quad \text{(second expression)}$$

We will use the second expression.

□

D.3 Proof for categorical BoP

Here, we apply the exponential family result found in D.2 to find the lower bound for a categorical distribution.

Corollary D.5. *[Lower bound for categorical individual BoP for any number of samples in each group (Monteiro Paes et al., 2022)] The lower bound writes:*

$$\min_{\Psi}\max_{\substack{P_0\in H_0\\P_1\in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(1+4\epsilon^2\right)^{m_j} - 1\right]^{\frac{1}{2}}$$

where $P_{\mathbf{X},\mathbf{S},Y}$ is a distribution of data, for which the generic model h_0 performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) < 0$, and $Q_{\mathbf{X},\mathbf{S},Y}$ is a distribution of data points for which the personalized model performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) \geq \epsilon$.

Proof. By Proposition D.4, we have:

$$\min_{\Psi}\max_{\substack{P_0\in H_0\\P_1\in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(\frac{M_p(2\Delta\theta)}{M_p(\Delta\theta)^2}\right)^{m_j} - 1\right]^{\frac{1}{2}}$$

Plug in Categorical assumption We find the bound for the categorical case. For the categorical, we have $\theta = \theta_1$ and:

$$\begin{aligned}
\theta_0 &= \log\left(\frac{p_1}{p_2}\right) = \log\frac{1/2}{1/2} = 0 \\
\theta_\epsilon &= \log\left(\frac{p_1^\epsilon}{p_2^\epsilon}\right) = \log\left(\frac{1/2+\epsilon}{1/2-\epsilon}\right) = \log\left(\frac{1+2\epsilon}{1-2\epsilon}\right) \\
A(\theta_0) &= \log\left(e^{\theta_0}+1\right) = \log(2) \\
A(\theta_\epsilon) &= \log\left(e^{\theta_\epsilon}+1\right) = \log\left(\frac{1+2\epsilon}{1-2\epsilon}+1\right) = \log\left(\frac{1+2\epsilon+1-2\epsilon}{1-2\epsilon}\right) = \log\left(\frac{2}{1-2\epsilon}\right) \\
A(2\theta_\epsilon) &= \log\left(e^{2\theta_\epsilon}+1\right) \\
&= \log\left((e^{\theta_\epsilon})^2+1\right) \\
&= \log\left(\left(\frac{1+2\epsilon}{1-2\epsilon}\right)^2+1\right) \\
&= \log\left(\frac{1+4\epsilon+4\epsilon^2}{1-4\epsilon+4\epsilon^2}+1\right) \\
&= \log\left(\frac{1+4\epsilon+4\epsilon^2+1-4\epsilon+4\epsilon^2}{1-4\epsilon+4\epsilon^2}\right) \\
&= \log\left(\frac{2+8\epsilon^2}{1-4\epsilon+4\epsilon^2}\right)
\end{aligned}$$

We also have: $\Delta\theta = \theta_\epsilon$.

Accordingly, we have:

$$\begin{aligned}
M_p(\Delta\theta) &= \exp\left(A(\theta_0+\Delta\theta) - A(\theta_0)\right) \\
&= \exp\left(A(\theta_\epsilon) - A(\theta_0)\right) \\
&= \exp\left(\log\left(\frac{2+\epsilon}{1-2\epsilon}\right) - \log(2)\right) \\
&= \exp\log\left(\frac{1}{2}\left(\frac{2}{1-2\epsilon}\right)\right) \\
&= \frac{1}{1-2\epsilon} \\
M_p(2\Delta\theta) &= \exp\left(A(\theta_0+2\Delta\theta) - A(\theta_0)\right) \\
&= \exp\left(A(2\theta_\epsilon\theta) - A(\theta_0)\right) \\
&= \exp\left(\log\left(\frac{2+8\epsilon^2}{1-4\epsilon+4\epsilon^2}\right) - \log(2)\right) \\
&= \exp\log\left(\frac{1}{2}\frac{2+8\epsilon^2}{1-4\epsilon+4\epsilon^2}\right) \\
&= \frac{1+4\epsilon^2}{1-4\epsilon+4\epsilon^2}
\end{aligned}$$

And the lower bound becomes:

$$\min_{\Psi}\max_{\substack{P_0\in H_0\\P_1\in H_1}} P_e \geq 1 - TV(P\parallel Q)$$

$$\Rightarrow \min_{\Psi}\max_{\substack{P_0\in H_0\\P_1\in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(\frac{M_p(2\Delta\theta)}{M_p(\Delta\theta)^2}\right)^{m_j} - 1\right]^{\frac{1}{2}}$$

$$= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \left(\frac{\frac{1+4\epsilon^2}{1-4\epsilon+4\epsilon^2}}{\left(\frac{1}{1-2\epsilon}\right)^2} \right)^{m_j} - 1 \right]^{\frac{1}{2}}$$

$$= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \left(\frac{\frac{1+4\epsilon^2}{1-4\epsilon+4\epsilon^2}}{\frac{1}{1-4\epsilon+4\epsilon^2}} \right)^{m_j} - 1 \right]^{\frac{1}{2}}$$

$$= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \left(1 + 4\epsilon^2\right)^{m_j} - 1 \right]^{\frac{1}{2}}$$

□

D.4 Maximum attributes (categorical BoP) for all people

In the case where dataset $\mathcal{D}$ is drawn from an unknown distribution and has d groups where $d = 2^k$, with each group having $m = \lfloor N/d \rfloor$ samples, Corollary D.5 becomes:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\left(1 + 4\epsilon^2\right)^m - 1 \right]^{\frac{1}{2}}$$

Corollary D.6 (Maximum attributes (categorical) for all people)**.** *Consider auditing a personalized classifier h_p to verify if it provides a gain of $\epsilon = 0.01$ to each group on an auditing dataset D. Consider an auditing dataset with $N = 8 \times 10^9$ samples, or one sample for each person on earth. If h_p uses more than $k \geq 18$ binary group attributes, then for any hypothesis test there will exist a pair of probability distributions $P_{X,G,Y} \in H_0, \quad Q_{X,G,Y} \in H_1$ for which the test results in a probability of error that exceeds 50%.*

$$k \geq 18 \implies \min_{\Psi} \max_{\substack{P_{X,G,Y} \in H_0 \\ Q_{X,G,Y} \in H_1}} P_e \geq \frac{1}{2}. \tag{26}$$

D.5 Proof for Gaussian BoP

Here, we do the proof assuming that the BoP is a normal variable with a second moment bounded by σ^2.

Corollary D.7. *[Lower bound for Gaussian BoP for any number of samples in each group] The lower bound writes:*

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \exp\left(\frac{m_j \epsilon^2}{\sigma^2} \right) - 1 \right]^{\frac{1}{2}}$$

where $P_{\mathbf{X},\mathbf{S},Y}$ is a distribution of data, for which the generic model h_0 performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) < 0$, and $Q_{\mathbf{X},\mathbf{S},Y}$ is a distribution of data points for which the personalized model performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) > 0$.

Proof. By Proposition D.4, we have:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \left(\frac{M_p(2\Delta\theta)}{M_p(\Delta\theta)^2} \right)^{m_j} - 1 \right]^{\frac{1}{2}}$$

Plug in Gaussian assumption We find the bound for the Gaussian case. For the Gaussian, we have:

$$\theta_0 = \frac{\mu_0}{\sigma^2} = 0$$
$$\theta_\epsilon = \frac{\mu_\epsilon}{\sigma^2} = \frac{\epsilon}{\sigma^2}$$
$$A(\theta_0) = \frac{\sigma^2\theta_0^2}{2} = 0$$
$$A(\theta_\epsilon) = \frac{\sigma^2\theta_\epsilon^2}{2} = \frac{\epsilon^2}{2\sigma^2}$$
$$A(2\theta_\epsilon) = \frac{\sigma^2 4\theta_\epsilon^2}{2} = \frac{2\epsilon^2}{\sigma^2}$$

because $\mu_0 = 0$ and $\mu_\epsilon = \epsilon$ by construction. Thus, we also have: $\Delta\theta = \theta_\epsilon$.

Accordingly, we have:

$$M_p(\Delta\theta) = \exp\left(A(\theta_0 + \Delta\theta) - A(\theta_0)\right) = \exp\left(A(\theta_\epsilon) - A(\theta_0)\right) = \exp\left(\frac{\epsilon^2}{2\sigma^2}\right)$$
$$M_p(2\Delta\theta) = \exp\left(A(\theta_0 + 2\Delta\theta) - A(\theta_0)\right) = \exp\left(A(2\theta_\epsilon - \theta_0)\right) = \exp\left(A(2\theta_\epsilon)\right) = \exp\left(\frac{2\epsilon^2}{\sigma^2}\right)$$

And the lower bound becomes:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - TV(P \parallel Q)$$

$$\Rightarrow \min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(\frac{M_p(2\Delta\theta)}{M_p(\Delta\theta)^2}\right)^{m_j} - 1\right]^{\frac{1}{2}}$$
$$= 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(\frac{\exp\left(\frac{2\epsilon^2}{\sigma^2}\right)}{\exp\left(\frac{\epsilon^2}{2\sigma^2}\right)^2}\right)^{m_j} - 1\right]^{\frac{1}{2}}$$
$$= 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(\frac{\exp\left(\frac{2\epsilon^2}{\sigma^2}\right)}{\exp\left(\frac{2\epsilon^2}{2\sigma^2}\right)}\right)^{m_j} - 1\right]^{\frac{1}{2}}$$
$$= 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\left(\frac{\exp\left(\frac{2\epsilon^2}{\sigma^2}\right)}{\exp\left(\frac{\epsilon^2}{\sigma^2}\right)}\right)^{m_j} - 1\right]^{\frac{1}{2}}$$
$$= 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\exp\left(\frac{\epsilon^2}{\sigma^2}\right)^{m_j} - 1\right]^{\frac{1}{2}}$$
$$= 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\exp\left(\frac{m_j\epsilon^2}{\sigma^2}\right) - 1\right]^{\frac{1}{2}}$$

In the case where each group has a different standard deviation of their BoP distribution, this becomes:

$$= 1 - \frac{1}{2\sqrt{d}}\left[\frac{1}{d}\sum_{j=1}^{d}\exp\left(\frac{m_j\epsilon^2}{\sigma_j^2}\right) - 1\right]^{\frac{1}{2}}$$

□

D.6 Maximum attributes (Gaussian BoP) for all people

In the case where dataset $\mathcal{D}$ is drawn from an unknown distribution and has d groups where $d = 2^k$, with each group having $m = \lfloor N/d \rfloor$ samples, Corollary D.7 becomes:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\exp\left(\frac{m\epsilon^2}{\sigma^2} \right) - 1 \right]^{\frac{1}{2}}$$

Corollary D.8 (Maximum attributes (Gaussian BoP) for all people)**.** *Consider auditing a personalized classifier h_p to verify if it provides a gain of $\epsilon = 0.01$ to each group on an auditing dataset D. Consider an auditing dataset with $\sigma = 0.1$ and $N = 8 \times 10^9$ samples, or one sample for each person on earth. If h_p uses more than $k \geq 22$ binary group attributes, then for any hypothesis test there will exist a pair of probability distributions $P_{X,G,Y} \in H_0, \quad Q_{X,G,Y} \in H_1$ for which the test results in a probability of error that exceeds 50%.*

$$k \geq 22 \implies \min_{\Psi} \max_{\substack{P_{X,G,Y} \in H_0 \\ Q_{X,G,Y} \in H_1}} P_e \geq \frac{1}{2}. \tag{27}$$

D.7 Proof for the symmetric generalized normal distribution

We solve for the the bound assuming the BoP is a symmetric generalized Gaussian distribution.

Symmetric Generalized Gaussian The symmetric generalized Gaussian distribution, also known as the exponential power distribution, is a generalization of the Gaussian distributions that include the Laplace distribution. A probability distribution in this family has probability density function:

$$p(x|\mu, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left(- \left(\frac{|x - \mu|}{\alpha} \right)^{\beta} \right), \tag{28}$$

with mean and variance:

$$\mathbb{E}[X] = \mu, \quad V[X] = \frac{\alpha^2 \Gamma(3/\beta)}{\Gamma(1/\beta)}. \tag{29}$$

We can write the standard deviation $\sigma = \alpha\sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}} = \alpha\gamma(\beta)$ where we introduce the notation $\gamma(\beta) = \sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}}$. This notation will become convenient in our computations.

Example: Laplace The Laplace probability density function is given by:

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b} \right) \tag{30}$$

which is in the family for $\alpha = b$ and $\beta = 1$, since the gamma function verifies $\Gamma(1) = (1-1)! = 0! = 1$.

Proposition D.9. *[Lower bound for symmetric generalized Gaussian BoP for any number of samples in each group] The lower bound writes:*

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^{\epsilon}} \left[\exp\left(-\frac{|B - \epsilon|^{\beta} - |B|^{\beta}}{\alpha^{\beta}} \right) \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$

where $P_{\mathbf{X},\mathbf{S},Y}$ is a distribution of data, for which the generic model h_0 performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) < 0$, and $Q_{\mathbf{X},\mathbf{S},Y}$ is a distribution of data points for which the personalized model performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) > 0$.

Proof. By Theorem D.3, we have:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^{\epsilon}} \left[\frac{p^{\epsilon}(B)}{p(B)} \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$

Plug in the symmetric generalized Gaussian distribution Under the assumption that the random variable B follows an exponential power distribution, we continue the computations as:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\frac{\exp\left(-\left(\frac{|B-\epsilon|}{\alpha} \right)^\beta \right)}{\exp\left(-\left(\frac{|B|}{\alpha} \right)^\beta \right)} \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$

$$= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\exp\left(-\left(\frac{|B-\epsilon|}{\alpha} \right)^\beta \right) \cdot \exp\left(\left(\frac{|B|}{\alpha} \right)^\beta \right) \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$

(property of exp)

$$= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\exp\left(-\left(\frac{|B-\epsilon|}{\alpha} \right)^\beta + \left(\frac{|B|}{\alpha} \right)^\beta \right) \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$ (property of exp)

$$= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\exp\left(-\frac{|B-\epsilon|^\beta - |B|^\beta}{\alpha^\beta} \right) \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$ (property of exp)

□

D.8 Proof for Laplace BoP

Here, we do the proof assuming that the BoP is a Laplace distribution (for more peaked than the normal variable).

Corollary D.10. *[Lower bound for a Laplace BoP for any number of samples in each group] The lower bound writes:*

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \exp\left(\frac{m_j \epsilon}{b} \right) - 1 \right]^{\frac{1}{2}}$$

where $P_{\mathbf{X},\mathbf{S},Y}$ is a distribution of data, for which the generic model h_0 performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) < 0$, and $Q_{\mathbf{X},\mathbf{S},Y}$ is a distribution of data points for which the personalized model performs better, i.e., the true γ is such that $\gamma(h_0, h_p, \mathcal{D}) > 0$.

Proof. By Proposition D.9, we have:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\exp\left(-\frac{|B-\epsilon|^\beta - |B|^\beta}{\alpha^\beta} \right) \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$

Plugging in our values of α and β shown to satisfy the Laplace probability density function we get:

$$= 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \mathbb{E}_{p^\epsilon} \left[\exp\left(-\frac{|B-\epsilon| - |B|}{b} \right) \right]^{m_j} - 1 \right]^{\frac{1}{2}}$$

Using bounds Since we are finding the worst case lower bound, we will find functions that upper and lower bound $|B-\epsilon| - |B|$. This function is lower bounded by ϵ and upper bounded by $-\epsilon$ since $\epsilon < 0$. Indeed, since $\epsilon < 0$, there are three cases:

- $0 < B < B - \epsilon$: this gives $|B-\epsilon| - |B| = B - \epsilon - B = -\epsilon$

- $B < 0 < B - \epsilon$: this gives $|B - \epsilon| - |B| = B - \epsilon + B = 2B - \epsilon > 2\epsilon - \epsilon = \epsilon$ since $0 < B - \epsilon$.
- $B < B - \epsilon < 0$: this gives $|B - \epsilon| - |B| = -B + \epsilon + B = \epsilon$.

Thus, we have: $\epsilon \leq |B - \epsilon| - |B| \leq -\epsilon$ and:

$$\begin{aligned}
&\epsilon \leq |B - \epsilon| - B| \leq -\epsilon \\
&\Rightarrow \frac{\epsilon}{b} \leq \frac{|B - \epsilon| - |B|}{b} \leq -\frac{\epsilon}{b} \\
&\Rightarrow -\frac{\epsilon}{b} \geq -\frac{|B - \epsilon| - |B|}{b} \geq \frac{\epsilon}{b} \\
&\Rightarrow \exp\left(-\frac{\epsilon}{b}\right) \geq \exp\left(-\frac{|B - \epsilon| - |B|}{b}\right) \geq \exp\left(\frac{\epsilon}{b}\right)
\end{aligned}$$

Thus, applying the expectation gives:

$$\begin{aligned}
&\mathbb{E}_{p^\epsilon}\left[\exp\left(-\frac{\epsilon}{b}\right)\right] \geq \mathbb{E}_{p^\epsilon}\left[\exp\left(-\frac{|B - \epsilon| - |B|}{b}\right)\right] \geq \mathbb{E}_{p^\epsilon}\left[\exp\left(\frac{\epsilon}{b}\right)\right] \\
&\Rightarrow \exp\left(-\frac{\epsilon}{b}\right) \geq \mathbb{E}_{p^\epsilon}\left[\exp\left(-\frac{|B - \epsilon| - |B|}{b}\right)\right] \geq \exp\left(\frac{\epsilon}{b}\right)
\end{aligned}$$

because the lower and upper bounds do not depend on B.

All the terms in these inequalities are positive, and the power function is increasing on positive numbers. Thus, we get:

$$\begin{aligned}
&\exp\left(-\frac{\epsilon}{b}\right)^{m_j} \geq \mathbb{E}_{p^\epsilon}\left[\exp\left(-\frac{|B - \epsilon| - |B|}{b}\right)\right]^{m_j} \geq \exp\left(\frac{\epsilon}{b}\right)^{m_j} \\
&\Rightarrow \frac{1}{d}\sum_{j=1}^{d}\exp\left(-\frac{\epsilon}{b}\right)^{m_j} \geq \frac{1}{d}\sum_{j=1}^{d}\mathbb{E}_{p^\epsilon}\left[\exp\left(-\frac{|B - \epsilon| - |B|}{b}\right)\right]^{m_j} \geq \frac{1}{d}\sum_{j=1}^{d}\exp\left(\frac{\epsilon}{b}\right)^{m_j} \\
&\Rightarrow \frac{1}{d}\sum_{j=1}^{d}\exp\left(-\frac{m_j\epsilon}{b}\right) \geq \frac{1}{d}\sum_{j=1}^{d}\mathbb{E}_{p^\epsilon}\left[\exp\left(-\frac{|B - \epsilon| - |B|}{b}\right)\right]^{m_j} \geq \frac{1}{d}\sum_{j=1}^{d}\exp\left(\frac{m_j\epsilon}{b}\right) \\
&\Rightarrow \frac{1}{d}\sum_{j=1}^{d}\exp\left(-\frac{m_j\epsilon}{b}\right) - 1 \geq \frac{1}{d}\sum_{j=1}^{d}\mathbb{E}_{p^\epsilon}\left[\exp\left(-\frac{|B - \epsilon| - |B|}{b}\right)\right]^{m_j} - 1 \geq \frac{1}{d}\sum_{j=1}^{d}\exp\left(\frac{m_j\epsilon}{b}\right) - 1 \\
&\Rightarrow \left(\frac{1}{d}\sum_{j=1}^{d}\exp\left(-\frac{m_j\epsilon}{b}\right) - 1\right)^{\frac{1}{2}} \\
&\geq \left(\frac{1}{d}\sum_{j=1}^{d}\mathbb{E}_{p^\epsilon}\left[\exp\left(-\frac{|B - \epsilon| - |B|}{b}\right)\right]^{m_j} - 1\right)^{\frac{1}{2}} \\
&\geq \left(\frac{1}{d}\sum_{j=1}^{d}\exp\left(\frac{m_j\epsilon}{b}\right) - 1\right)^{\frac{1}{2}} \\
&\Rightarrow -\frac{1}{2\sqrt{d}}\left(\frac{1}{d}\sum_{j=1}^{d}\exp\left(-\frac{m_j\epsilon}{b}\right) - 1\right)^{\frac{1}{2}} \\
&\leq -\frac{1}{2\sqrt{d}}\left(\frac{1}{d}\sum_{j=1}^{d}\mathbb{E}_{p^\epsilon}\left[\exp\left(-\frac{|B - \epsilon| - |B|}{b}\right)\right]^{m_j} - 1\right)^{\frac{1}{2}}
\end{aligned}$$

$$\leq -\frac{1}{2\sqrt{d}} \left(\frac{1}{d} \sum_{j=1}^{d} \exp\left(\frac{m_j \epsilon}{b}\right) - 1 \right)^{\frac{1}{2}}$$

Back to Probability of error To maximize P_e, we take the function that gives us the lower bound. Plugging this upper bound back into our equation for P_e:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \exp\left(-\frac{m_j \epsilon}{b}\right) - 1 \right]^{\frac{1}{2}}$$

In the case where each group has a different scale parameter of their BoP distribution, this becomes:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \exp\left(-\frac{m_j \epsilon}{b_j}\right) - 1 \right]^{\frac{1}{2}}$$

Such that for the unflipped hypothesis testing with $\epsilon > 0$ we get:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\frac{1}{d} \sum_{j=1}^{d} \exp\left(\frac{m_j \epsilon}{b_j}\right) - 1 \right]^{\frac{1}{2}}$$

□

D.9 Maximum attributes (Laplace BoP) for all people

In the case where dataset $\mathcal{D}$ is drawn from an unknown distribution and has d groups where $d = 2^k$, with each group having $m = \lfloor N/d \rfloor$ samples, Corollary D.10 becomes:

$$\min_{\Psi} \max_{\substack{P_0 \in H_0 \\ P_1 \in H_1}} P_e \geq 1 - \frac{1}{2\sqrt{d}} \left[\exp\left(\frac{m\epsilon}{b}\right) - 1 \right]^{\frac{1}{2}}$$

Corollary D.11 (Maximum attributes (Laplace) for all people)**.** *Consider auditing a personalized classifier h_p to verify if it provides a gain of $\epsilon = 0.01$ to each group on an auditing dataset D. Consider an auditing dataset with $\sigma = 0.1$ and $N = 8 \times 10^9$ samples, or one sample for each person on earth. If h_p uses more than $k \geq 26$ binary group attributes, then for any hypothesis test there will exist a pair of probability distributions $P_{X,G,Y} \in H_0, \quad Q_{X,G,Y} \in H_1$ for which the test results in a probability of error that exceeds 50%.*

$$k \geq 26 \implies \min_{\Psi} \max_{\substack{P_{X,G,Y} \in H_0 \\ Q_{X,G,Y} \in H_1}} P_e \geq \frac{1}{2}. \tag{31}$$

E Limits on Attributes and Sample Size

This section derives theoretical limits on the number of personal attributes and the sample size required per group to ensure that the probability of error remains below a practitioner-specified threshold.

Corollary E.1. *Let N be the number of participants, and assume that each group $j = 1, \ldots, d$ has $m_j = m = \left\lfloor \frac{N}{d} \right\rfloor$ samples. To ensure that the probability of error verifies $\min \max P_e \leq 1/2$, the number of binary attributes k must be chosen such that $k \leq k_{\max}$, where:*

$$k_{max} = \begin{cases} 1.4427W\left(N \log(4\epsilon^2 + 1)\right) \text{ (Categorical BoP)} \\ 1.4427W\left(\frac{\epsilon^2 N}{\sigma^2}\right) \text{ (Gaussian BoP, variance } \sigma^2\text{)} \\ 1.4427W\left(\frac{\epsilon N}{b}\right) \text{ (Laplace BoP, scale } b\text{)}, \end{cases}$$

where W is the Lambert W function.

Corollary E.2. *Let N be the number of participants, and assume that each group $j = 1, \ldots, d$ has $m_j = m = \left\lfloor \frac{N}{d} \right\rfloor$ samples. To ensure that the probability of error verifies $\min \max P_e \leq v$ where v is chosen by the practitioner, the size of the groups m must be $m \geq m_{\min}$, where:*

$$m_{min} = \begin{cases} \frac{\log\left(4 \cdot 2^k (1-v)^2 + 1\right)}{\log(1+4\epsilon^2)} \text{ (Categorical BoP)} \\ \frac{\sigma^2}{\epsilon^2} \log\left(2^{2+k}\left(1 + 2^{-2-k} - 2v + v^2\right)\right) \text{ (Gaussian BoP, variance } \sigma^2\text{)} \\ \frac{b}{\epsilon} \log\left(2^{2+k}\left(1 + 2^{-2-k} - 2v + v^2\right)\right) \text{ (Laplace BoP, scale } b\text{)}, \end{cases}$$

F MIMIC-III Experiment Results

Below is all supplementary material for the MIMIC-III experiment. This includes G-BoP distribution plots and plots showing how incomprehensiveness and sufficiency change over the number of features removed.

F.1 Experiment Plots

Experiment Setup. We assume that the practitioner uses a 70/30 train-test split for both tasks and compare two neural network models: a personalized model with one-hot encoded group attributes (h_p) and a generic model without them (h_0). Regression outputs are normalized to zero mean and unit variance.

Explanation Method and Explanation Evaluation metric. We assume that the practitioner generates the most important features of our models using Integrated Gradients from Captum as our explanation method (Sundararajan et al., 2017). We assume that they use sufficiency and incomprehensiveness as our explanation evaluation metrics, where 50% of features are either kept or removed.

Integrated Gradients extracts the most important features of each model by computing input-feature attributions by integrating gradients along a path from a baseline to the input. To evaluate BoP_X using sufficiency and incomprehensiveness, we set r such that 50% of features are kept or removed. Plots below depict how sufficiency and incomprehensiveness change for different values of r, as well as show the individual BoP distributions. We use Integrated Gradients for its efficiency, interpretability, and broad adoption, though our framework supports any attribution method.

In the following section, we show supplementary plots for the regression task on the auditing dataset. We show the distribution of the BoP across participants for all three metrics we evaluate. We overlay Laplace and Gaussian distributions to see which fit the individual BoP distribution best, illustrating that prediction and incomprehensiveness are best fit by Laplace distributions and sufficiency by a Gaussian distribution. Additionally, we show how incomprehensiveness and sufficiency change for the number of important attributes r that are kept are removed.

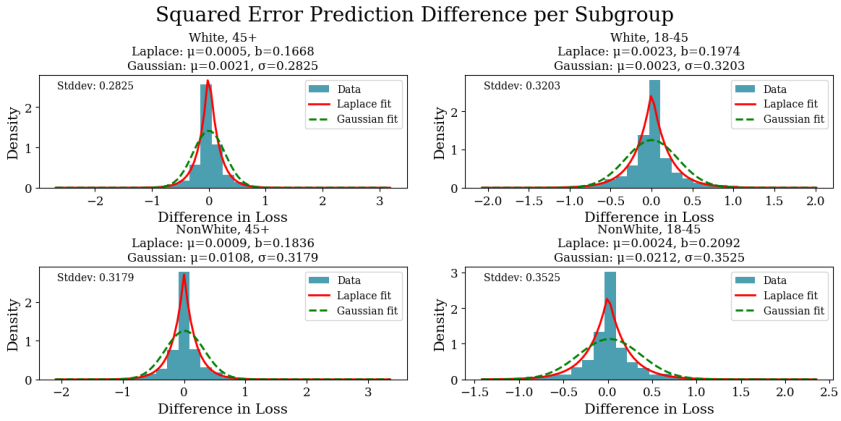

Figure 10: Individual prediction cost for all groups using the square error loss function.

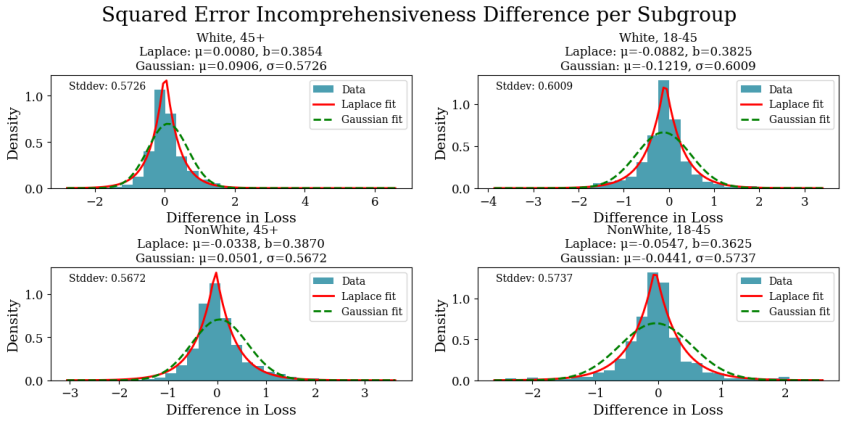

Figure 11: Individual incomprehensiveness cost for all groups using the square error loss function.

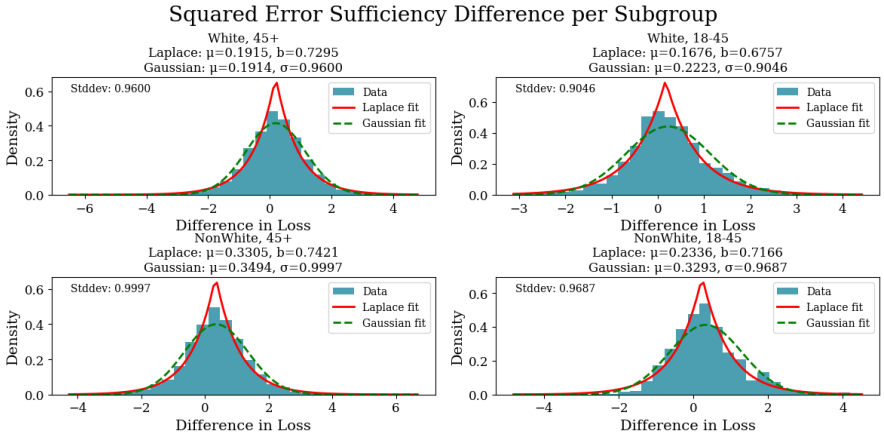

Figure 12: Individual sufficiency cost for all groups using the square error loss function.

G Additional Dataset Results

The following is the experiment results G-BoP$_P$ and G-BoP$_X$ on the UCI Heart (Janosi et al., 1989) and MIMIC-III Kidney injury dataset (Johnson et al., 2016) utilizing three explainer methods through Captum: Integrated Gradients Sundararajan et al. (2017), Shapley Value Sampling (Štrumbelj & Kononenko, 2010), and Deeplift (Shrikumar et al., 2017). Interestingly, we see a large amount of agreement across these explainer methods: in nearly all cases, groups that benefited or were harmed remain consistent across methods, although the amount by which this occurs varies. We compute ϵ_{lim}, the value of ϵ for which the lower bound of P_e surpasses 50% for the Shapley Value Sampling Method on the UCI Heart dataset to illustrate the full pipeline.

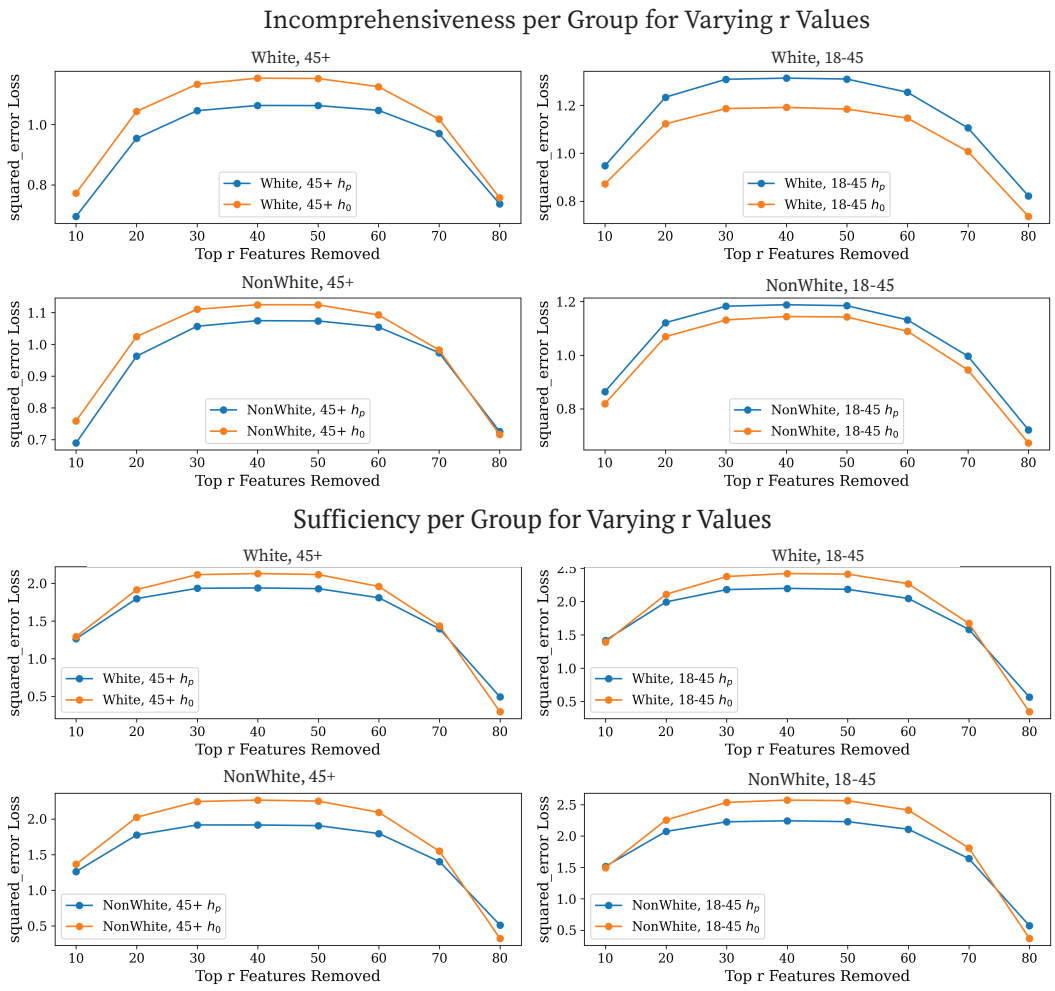

Figure 13: Values of Sufficiency and Incomprehensiveness across varying r top features selected using the square error loss function. Values are found for h_0 and h_p.

Table 4: Experimental results on the UCI Heart test set, with columns for DeepLift (D.L.), Integrated Gradients (I.G.), and Shapley Value Sampling (S.V.S.). The classification task is predicting heart disease presence and the regression task is predicting ST depression induced by exercise. All available features are used, and negative entries appear in red. Using our framework, we computed ϵ_{lim} (for the S.V.S. explainer method) where the lower bound on P_e surpasses 50%. In classification, $\epsilon_{\text{lim}} = 0.1156$ for all metrics; in regression, $\epsilon_{\text{lim}} = 0.0163$ for prediction (Laplace), 0.02 for incomprehensiveness (Laplace), and 0.153 for sufficiency (Gaussian).

Classification Results

Group	Prediction	Incomp. D.L	Suff. D.L	Incomp. I.G.	Suff. I.G.	Incomp. S.V.S.	Suff. S.V.S.
Female, 45+	0.0000	0.0000	-0.0435	0.0000	-0.0435	0.0000	-0.0870
Female, 18–45	0.0000	-0.1429	0.0000	-0.1429	0.0000	0.0000	0.0000
Male, 45+	0.0588	-0.0588	-0.0784	-0.0588	-0.1373	-0.0588	-0.1176
Male, 18–45	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000
All Pop.	0.0440	-0.0330	-0.0440	-0.0330	-0.0750	-0.0220	-0.0769
Minimal BoP	0.0000	-0.1429	-0.0784	-0.1429	-0.1373	-0.0588	-0.1176

Regression Results

Group	Prediction	Incomp. D.L	Suff. D.L	Incomp. I.G	Suff. I.G	Incomp. S.V.S.	Suff. S.V.S.
Female, 45+	-0.3077	0.3528	0.1385	0.0980	0.2040	0.1747	0.3332
Female, 18–45	0.0521	-0.0004	0.1067	-0.0438	0.1774	-0.0207	0.0222
Male, 45+	0.0914	0.0286	0.0531	0.0173	0.1381	0.0315	0.1617
Male, 18–45	-0.1410	0.1239	0.4293	0.1384	0.4365	0.1360	0.3592
All Pop.	-0.0363	0.0791	0.1833	0.0523	0.2035	0.0779	0.2258
Minimal BoP	-0.3077	-0.0004	0.0531	-0.0438	0.1381	-0.0207	0.0222

Table 5: Experimental results on the MIMIC-III Kidney test set, with columns for DeepLift (D.L.), Integrated Gradients (I.G.), and Shapley Value Sampling (S.V.S.); negative values appear in red. The regression task predicts hours to the next continuous renal replacement therapy (CRRT). For classification, the target is patient mortality during the same hospital admission. Features include recent lab measurements (e.g., sodium, potassium, creatinine) prior to CRRT, along with patient age, hours in the ICU at CRRT administration, and the Sequential Organ Failure Assessment (SOFA) score at admission.

Classification Results

Group	Prediction	Incomp. D.L	Suff. D.L	Incomp. I.G	Suff. I.G	Incomp. S.V.S.	Suff. S.V.S.
Female, 45+	0.0392	0.0392	-0.0784	0.0392	-0.0784	0.0392	-0.0196
Female, 18–45	0.0000	0.0000	0.3636	0.0000	0.3636	0.0000	0.3636
Male, 45+	0.0164	-0.0164	0.0820	-0.0164	0.0984	-0.0164	0.0000
Male, 18–45	0.0000	0.0000	-0.0833	0.0000	-0.0833	0.0000	0.1667
All Pop.	0.0224	0.0074	0.0296	0.0074	0.0370	0.0074	0.0370
Minimal BoP	0.0000	-0.0164	-0.0833	-0.0164	-0.0833	-0.0164	-0.0196

Regression Results

Group	Prediction	Incomp. D.L	Suff. D.L	Incomp. I.G	Suff. I.G	Incomp. S.V.S.	Suff. S.V.S.
Female, 45+	0.7582	0.1440	-0.5722	0.1322	-0.6185	0.1380	-0.5414
Female, 18–45	0.5639	0.0177	-0.3325	0.0404	-0.2543	0.0649	-0.3107
Male, 45+	0.3449	0.0258	-0.1180	0.0299	-0.1368	0.0310	-0.1518
Male, 18–45	0.4869	-0.1016	-0.1639	-0.0997	-0.1571	-0.0892	-0.2124
All Pop.	-0.0093	0.0595	-0.3097	0.0584	-0.3311	0.0635	-0.3167
Minimal BoP	-0.0093	-0.1016	-0.5722	-0.0997	-0.6185	-0.0892	-0.5414