Robust Bayesian Recourse

Tuan-Duy H. Nguyen[1] **Ngoc Bui**[1] **Duy Nguyen**[1] **Man-Chung Yue**[2] **Viet Anh Nguyen**[1]

[1]VinAI Research, Vietnam
[2]The University of Hong Kong

Abstract

Algorithmic recourse aims to recommend an informative feedback to overturn an unfavorable machine learning decision. We introduce in this paper the Bayesian recourse, a model-agnostic recourse that minimizes the posterior probability odds ratio. Further, we present its min-max robust counterpart with the goal of hedging against future changes in the machine learning model parameters. The robust counterpart explicitly takes into account possible perturbations of the data in a Gaussian mixture ambiguity set prescribed using the optimal transport (Wasserstein) distance. We show that the resulting worst-case objective function can be decomposed into solving a series of two-dimensional optimization subproblems, and the min-max recourse finding problem is thus amenable to a gradient descent algorithm. Contrary to existing methods for generating robust recourses, the robust Bayesian recourse does not require a linear approximation step. The numerical experiment demonstrates the effectiveness of our proposed robust Bayesian recourse facing model shifts. Our code is available at `https://github.com/VinAIResearch/robust-bayesian-recourse`.

1 INTRODUCTION

Human constantly embark on multiple temporally-extended planning problems throughout the course of their lifespan, and we have several layers of means-end planning in order to achieve the desired goals. For example, to have a successful career as a machine learning researcher, an individual needs to put in persistent effort from their early education to their post-graduate studies, which may span over the course of over twenty years with numerous significant milestones to achieve. Two of these important milestones are the PhD admission and the job application, and arguably, a favorable outcome at these two milestones may propel an individual's career on a more auspicious trajectory than a negative outcome. To aid the committee to make better decisions, machine learning models are increasingly used in both university admission [Waters and Miikkulainen, 2014] and job hiring [Sajjadiani et al., 2019]. A similar trend takes place in credit loan applications [Siddiqi, 2012], healthcare [Mertes et al., 2021] and many others.

The increasing reliance on and the long impact of algorithmic decisions raise significant requirements on the trustworthiness and explainability of the machine learning models. These requirements become more urgent as black-box, complex models are also gaining spotlight attraction due to their superior performance [Garisch and Merchant, 2019]. Post-hoc explanations, which extracts human-understandable explanations, may benefit individuals to understand machine-produced decisions [Kenny et al., 2021]. A post-hoc method must demonstrate why unfavorable predictions are made, and possibly how an input would have been to obtain a favorable predicted outcome. If the inputs encode the characteristics of human individuals, then a possible post-hoc explanation may come in the form of a recourse. A recourse recommends the actions that an individual should take in order to receive an alternate algorithmic outcome [Ustun et al., 2019]. Consider an applicant who is rejected for a particular job, a recourse may come in the form of personalized recommendations such as "complete a 6-month full-stack engineer internship" or "score 20 more points in the ability test", along with the promise that if the applicant successfully implement the necessary action then the algorithm will return a favorable outcome.

Several approaches has been proposed to provide recourses for machine learning models [Karimi et al., 2021, Stepin et al., 2021, Mishra et al., 2021, Artelt and Hammer, 2019, Pawelczyk et al., 2021]. Wachter et al. [2018] used a gradient-based approach to find nearest counterfactual to the original instance. Ustun et al. [2019] proposed an integer

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

programming approach to generate actionable recourses for linear classifiers. Karimi et al. [2020] proposed a model-agnostic approach to generate nearest counterfactual explanations while Poyiadzi et al. [2020] generates counterfactuals that are actionable and supported by the "feasible paths" of actions. Pawelczyk et al. [2020] find a counterfactual explanation with an upper bound for the costs of counterfactual explanations under predictive multiplicity. Mothilal et al. [2020] proposed a framework for generating and evaluating a diverse set of counterfactual explanations based on determinantal point processes. Bui et al. [2022] proposed an uncertainty quantification tool to compute the bounds of the probability of validity of a set of counterfactual explanations and enhanced the validity of this set via a correction tool.

These aforementioned approaches all assume that the underlying machine learning models do not change over time. In practice, this assumption is easily violated as experts update the machine learning system frequently due to data distribution shifts [Quionero-Candela et al., 2009, Y et al., 2019]. As such, an individual may have accomplished all the recommended actions but the next time they apply for the job, the parameters of the model may already change and the updated model may still recommend a negative outcome. In that case, the recourse becomes useless: it is ineffective in overturning a negative prediction, it incurs cost to the applicant, and at the same time it raises substantial doubts about the recourse [Rawal et al., 2020]. Following this line, a recourse is considered to be robust if it is effective at reversing the algorithmic outcome even under model shifts.

To construct a robust recourse, Upadhyay et al. [2021] proposed ROAR, a framework that leverages adversarial training to hedge against the perturbation of the model parameter. ROAR considers only linear classifiers; for *non*linear classifiers, ROAR first generates a locally linear approximation of the underlying model (e.g., by using LIME [Ribeiro et al., 2016]), then applies the adversarial training procedure with respect to this locally linear surrogate. However, there are multiple downsides when a locally linear model is used to approximate the nonlinear classifier. Recent works have shown that the locally linear model of LIME has some limitations with both its fidelity and robustness. LIME may not be faithful to the underlying model since it might be influenced by input features at a global scale rather than a local scale [White and Garcez, 2019, Laugel et al., 2018]. At the same time, several works [Alvarez-Melis and Jaakkola, 2018, Slack et al., 2020, Agarwal et al., 2021] point out that the explanations generated by LIME and other explanation methods may change significantly for nearby original inputs. Moreover, these explanations are even sensitive to the sampling distribution, and the can deliver different explanations of the same input in different simulation runs. Finally, a recourse which is robust for the linear approximation model may not necessarily be robust respective to the original nonlinear model.

Contributions. The goal of this paper is to formulate a model-agnostic recourse, which is also valid subject to potential future shifts of the machine learning models. Compared to existing methods such as ROAR [Upadhyay et al., 2021], our method does not depend on the linear surrogate of the *non*linear predictive model. Instead, our method looks directly into the sampled data points, and employs a Bayesian approach to generate recourses. Potential shifts of the predictive models are engendered by "perturbing" these data samples in an adversarial manner. We contribute concretely the followings.

- In Section 2, we propose the notion of a Bayesian recourse, which minimizes the odds ratio between the posterior probability of negative and positive predicted outcomes. In a non-parametric setting, the likelihood can be approximated using a kernel density estimator built around the data sampled in the neighborhood of the boundary point. This results in the KDE-Bayesian recourse, which can be found by (projected) gradient descent.
- In Section 3, we propose the robust counterpart of the Bayesian recourse problem. This robustification involves smoothing the samples by an isotropic Gaussian convolution, then solving a min-max optimization problem over a Wasserstein-Gaussian mixture conditional ambiguity set. Section 4 details our method of using the optimal transport to form the ambiguity sets on the space of Gaussian mixtures.
- In Section 5, we show that the robust Bayesian recourse problem is amenable to separability and dimensionality reduction, thus the recourse can be constructed efficiently even in high dimensions. Section 6 demonstrates that our recourse also performs competitively on both synthetic and real datasets.

Notations. We use δ_s to denote a Dirac measure supported on point s. The space of p-by-p symmetric, positive semidefinite matrix is denoted by $\mathbb{S}_+^p$.

2 BAYESIAN RECOURSE

We consider a generic covariate $X \in \mathcal{X} = \mathbb{R}^p$ and a binary predicted label $\hat{Y} \in \mathcal{Y} = \{0, 1\}$, where class 0 denotes an *un*favorable outcome while class 1 denotes a favorable one. Given a pre-specified black-box classifier $\mathcal{C}$ and an input x_0 with unfavorable prediction, i.e., $\mathcal{C}(x_0) = 0$, the goal of algorithmic recourse is to devise an alternative x' in the vicinity of x_0 that satisfies $\mathcal{C}(x') = 1$. The Bayesian recourse imposes a probabilistic viewpoint into this problem: the goal of Bayesian recourse is to devise an alternative in the vicinity of x_0 that has high *favorable posterior probability*. In technical terms, consider the joint random vector of covariate-label $(X, \hat{Y}) \in \mathcal{X} \times \mathcal{Y}$, then the class posterior probability of any input x can be represented by the

conditional random variable $\hat{Y}|X = x$.[1]

Definition 2.1 (Bayesian recourse)**.** *Given an input x_0, let $\mathbb{X}$ be a neighborhood around x_0. A Bayesian recourse $x_{\text{Bayes}} \in \mathbb{X}$ is an alternative that minimizes the Bayesian posterior odds ratio, i.e.,*

$$x_{\text{Bayes}} \triangleq \arg\min_{x \in \mathbb{X}} \frac{\mathbb{P}(\hat{Y} = 0|X = x)}{\mathbb{P}(\hat{Y} = 1|X = x)},$$

for some joint distribution $\mathbb{P}$ of $(X, \hat{Y})$ induced by the sampling of the synthetic covariate X and the synthetic predicted label $\hat{Y} = \mathcal{C}(X)$.

The ratio $\mathbb{P}(\hat{Y} = 0|X = x)/\mathbb{P}(\hat{Y} = 1|X = x)$ is a well-known quantity in Bayesian classification. The posterior probability odds is also a popular ratio in Bayesian statistics, and it has been applied for comparing regression hypotheses Zellner [1981], econometric models Geweke [1994], asset pricing theories McCulloch and Rossi [1991] and collaborative evaluations Hicks et al. [2018].

As x_{Bayes} minimizes the Bayesian posterior odds ratio, we can argue that $\mathbb{P}(\hat{Y} = 0|X = x_{\text{Bayes}})$ tends to be low, while $\mathbb{P}(\hat{Y} = 1|X = x_{\text{Bayes}})$ tends to be high. We next describe how we can solve the optimization problem to get x_{Bayes}. Note that the posterior probability can be calculated using the Bayes' theorem [Schervish, 1995, Theorem 1.31], and we can instead solve the fractional optimization problem

$$\min_{x \in \mathbb{X}} \frac{\mathbb{P}(\hat{Y} = 0)\mathbb{P}(X = x|\hat{Y} = 0)}{\mathbb{P}(\hat{Y} = 1)\mathbb{P}(X = x|\hat{Y} = 1)}.$$

It is now clear that to find x_{Bayes}, we need the marginal probability of $\hat{Y}$ and the likelihood of $X|\hat{Y}$. Suppose that we can use a sampling mechanism to sample n covariates $\widehat{x}_i$, then query the given classifier to obtain the predicted labels $\widehat{y}_i = \mathcal{C}(\widehat{x}_i)$ to form n pairs $(\widehat{x}_i, \widehat{y}_i)$, $i = 1, \ldots, n$. Using these synthetic, labelled samples, we now can formulate the empirical version of Bayesian recourse problem. Let $\mathcal{I}_y = \{i \in [n] : \widehat{y}_i = y\}$ be the indices of samples in class $y \in \mathcal{Y}$. Let $N_y = |\mathcal{I}_y|$ be the number of training samples with class y, then we can use $\gamma_y = N_y/n$, the empirical proportion of data for class y, as an estimate of $\mathbb{P}(\hat{Y} = y)$.

Next, we take the nonparametric approach to estimate the likelihood $\mathbb{P}(X = x|\hat{Y} = y)$ using a kernel density estimator [Tsybakov, 2008, Section 1]. As a concrete example, we choose the Gaussian kernel with bandwidth $h > 0$, thus the kernel density estimate of the quantity $\mathbb{P}(X = x|\hat{Y} = y)$ is

$$L_{\text{KDE}}(x|\hat{Y} = y) = \frac{1}{N_y} \sum_{i \in \mathcal{I}_y} \exp\left(-\frac{1}{2h^2}\|x - \widehat{x}_i\|_2^2\right).$$

[1] In algorithmic recourse, the random variable of interest is the predicted label $\hat{Y}$ induced by the classifier $\mathcal{C}$, not the true label Y of the data-generating process. It is important to keep in mind that the (robust) Bayesian recourse is formulated with respect to the predicted label $\hat{Y}$.

Thus, the empirical version of the Bayesian recourse, termed the KDE-Bayesian recourse, can be found by solving

$$\min_{x \in \mathbb{X}} \frac{\gamma_0 \times L_{\text{KDE}}(x|\hat{Y} = 0)}{\gamma_1 \times L_{\text{KDE}}(x|\hat{Y} = 1)}. \tag{1}$$

This problem further simplifies to

$$\min_{x \in \mathbb{X}} \frac{\sum_{i \in \mathcal{I}_0} \exp\left(-\frac{1}{2h^2}\|x - \widehat{x}_i\|_2^2\right)}{\sum_{i \in \mathcal{I}_1} \exp\left(-\frac{1}{2h^2}\|x - \widehat{x}_i\|_2^2\right)}$$

by exploiting the definition of L_{KDE} and γ_y. In this form, a (projected) gradient descent algorithm can be employed to find the KDE-Bayesian recourse.

There remain two elements to be specified about the formulation of the Bayesian recourse: the sampling scheme to generate covariates $\widehat{x}_i$ and the feasible set $\mathbb{X}$. We discuss these components in the remainder of this section.

Sampling scheme. The goal of the sampling scheme is to synthesize covariate data $\widehat{x}_i$ around the boundary to obtain *local* information from the black-box classifier. Toward this goal, we use a local sampling method, similar to Vlassopoulos et al. [2020] and Laugel et al. [2018] as follows.

- Given an instance x_0, we choose K nearest counterfactuals $x_1, \ldots, x_K$ from the training data that have favorable predicted outcome, that is, $\mathcal{C}(x_k) = 1$ for $k = 1, \ldots, K$.
- For each counterfactual x_k, we perform a line search to find a point x_k^b that is on the decision boundary and on the line segment joining x_0 and x_k.
- Among these points x_k^b, we choose the nearest point to x_0 by setting $x^b \triangleq \arg\min_{x_i^b}\{c(x_i^b, x_0)\}$, where $c(\cdot)$ is the cost function. We then sample $\widehat{x}_i$ uniformly in a neighborhood determined by an ℓ_2-ball with radius r_p centered on x^b.

Feasible set $\mathbb{X}$. It is desirable to constrain the recourse in a *strict* neighborhood of distance δ from the input Venkatasubramanian and Alfano [2020]. Thus, we can impose a feasible set of the form

$$\mathbb{X} = \{x \in \mathcal{X} \ : \ \varphi(x, x_0) \leq \delta\},$$

where φ is a measure of dissimilarity on the covariate space $\mathcal{X}$. Alternatively, if we use a boundary sampler as previously discussed, we may also opt for the constraint $\varphi(x, x^b) \leq \delta'$ around the boundary point x^b. A good choice of φ is the ℓ_1 distance, which promotes sparse modifications to the input.

In order to construct plausible and meaningful recourses, we could additionally consider the actionability constraints that forbid unrealistic recourses. For example, the gender or race of a person should be considered immutable. Likewise, recourse should not suggest an individual reduce their age to achieve a favorable outcome. These constraints could be easily injected into the definition of the feasible set $\mathbb{X}$, similar

to Upadhyay et al. [2021]. Finding the optimal actionable recourse restricted to this feasible set could be addressed effectively by a projected gradient descent algorithm [Mothilal et al., 2020, Upadhyay et al., 2021].

3 ROBUST BAYESIAN RECOURSE

The Bayesian recourse in Definition 2.1 depends on the classifier $\mathcal{C}$ as we query $\mathcal{C}$ to label the samples $\widehat{x}_i$ via $\widehat{y}_i = \mathcal{C}(\widehat{x}_i)$. Thus, inherently, the recourse would possess high posterior probability of favorable outcome with respect to the *present* classifier $\mathcal{C}$. Because the parameters defining $\mathcal{C}$ may be updated, the Bayesian recourse does not guarantee a high probability of favorable outcome with respect to the *future* classifier $\tilde{\mathcal{C}}$. Devising a recourse that has a high probability of future favorable outcome encounters two critical difficulties: first, the classifiers $\mathcal{C}$ and $\tilde{\mathcal{C}}$ are possibly nonlinear, and second, it is nontrivial to predict the shifts in the parameters of $\tilde{\mathcal{C}}$ from the present model $\mathcal{C}$. Existing robust recourse methods such as ROAR [Upadhyay et al., 2021] need to approximate a nonlinear model by a linear model using LIME [Ribeiro et al., 2016], then robustness is represented by perturbations of the parameters of the linear surrogate.

The robust Bayesian recourse takes a completely different path to ensure robustness by removing the need for an intermediate linear surrogate model. The robust Bayesian recourse aims to perturb directly the empirical conditional distributions of $X|\hat{Y} = y$, which then reshapes the decision boundary in the covariate space in an adversarial manner. Holistically, our approach can be decomposed into the following steps:

1. Forming the empirical conditional distributions of $X|\hat{Y} = y$, then smoothen them by convoluting an isotropic Gaussian noise to each data point.
2. Formulating the ambiguity set for each conditional distributions of $X|\hat{Y} = y$.
3. Solving a min-max problem to find the recourse that minimizes the worst-case Bayesian posterior odds ratio.

We now dive into the technical specifications of the robust Bayesian recourse. Remind that the sampling procedure equips us with the samples $(\widehat{x}_i, \widehat{y}_i)_{i=1,\ldots,n}$, and $\mathcal{I}_y$ are indices of samples with predicted label y. Let $\widehat{\mathbb{P}}_y^\sigma = N_y^{-1} \sum_{i\in\mathcal{I}_y} \delta_{\widehat{x}_i} * \mathcal{N}(0, \sigma^2 I)$ be the *smoothed* empirical conditional distribution of $X|Y = y$, in which $*$ denotes the convolution. Notice that $\widehat{\mathbb{P}}_y^\sigma$ is a mixture of Gaussian with N_y components located at the covariate $\widehat{x}_i$ with isotropic variance $\sigma^2 I$. Smoothing the empirical distribution by convoluting a noise to each sample is also attracting attention recently thanks to its possibility to quantify and enhance the robustness of machine learning models Cohen et al. [2019].

We assume now that the conditional distribution can be perturbed in an ambiguity set $\mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}_y^\sigma)$. This set $\mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}_y^\sigma)$ is defined as a neighborhood of radius $\varepsilon_y \geq 0$ centered at the nominal distribution $\widehat{\mathbb{P}}_y^\sigma$. The robust Bayesian recourse is defined as the optimal solution of the following problem

$$\min_{x\in\mathbb{X}} \max_{\mathbb{Q}_0\in\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma), \mathbb{Q}_1\in\mathbb{B}_{\varepsilon_1}(\widehat{\mathbb{P}}_1^\sigma)} \frac{\gamma_0 \mathbb{Q}_0(X = x)}{\gamma_1 \mathbb{Q}_1(X = x)}. \tag{2}$$

Notice that $\mathbb{Q}_y$ is a *conditional* probability measure of X given $\hat{Y} = y$, and thus it is a measure supported on $\mathbb{R}^p$. The value $\mathbb{Q}_y(X = x)$ is also the likelihood of x under the conditional measure $\mathbb{Q}_y$, thus problem (2) can be view as a robust likelihood ratio minimization problem. Here, robustness is defined with respect to the conditional sets $\mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}_y^\sigma)$ in the specific sense: the optimal value of problem (2) constitutes a uniform upper bound of the likelihood ratio over all possible choices of conditional distributions in the sets $\mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}_y^\sigma)$. Further, we have explicitly used γ_y as an estimator of the marginal distribution of $\hat{Y}$ in problem (2).

There is an intimate relationship between the KDE-Bayesian recourse problem (1) and the robust Bayesian recourse problem (2). This relationship is established thanks to the smoothing of the empirical conditional distributions, and is highlighted in the following remark.

Remark 3.1 (Recovery of the KDE-Bayesian recourse)**.** *The smoothed conditional distribution $\widehat{\mathbb{P}}_y^\sigma$ is a mixture of Gaussians, and the likelihood of x under $\widehat{\mathbb{P}}_y^\sigma$ is*

$$\frac{1}{N_y (2\pi)^{\frac{p}{2}} \sigma^p} \sum_{i\in\mathcal{I}_y} \exp\left(-\frac{1}{2\sigma^2}\|x - \widehat{x}_i\|_2^2\right).$$

As a consequence, if the ambiguity sets $\mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}_y^\sigma)$ collapse into singletons, that is, $\mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}_y^\sigma) = \{\widehat{\mathbb{P}}_y^\sigma\}$, then problem (2) *coincides with the KDE-Bayesian recourse problem* (1)*. Thus, problem* (2) *can be considered as a robustification of the KDE-Bayesian recourse formulation.*

4 WASSERSTEIN-GAUSSIAN MIXTURE AMBIGUITY SETS

The central notion underlying the robust Bayesian recourse problem (2) is the set of probability measures for the covariate X conditional that $Y = y$. A suitable design of the ambiguity set $\mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}_y^\sigma)$ is critical to enable an efficient resolution of problem (2). We here propose a novel design of the ambiguity set by merging ideas from the theory of optimal transport and Gaussian mixtures.

Note that any Gaussian distribution is fully characterized by its mean vector and its covariance matrix. As the smoothed measure $\widehat{\mathbb{P}}_y^\sigma$ is a Gaussian mixture, it is associated with the discrete distribution $\widehat{\nu}_y = N_y^{-1} \sum_{i\in\mathcal{I}_y} \delta_{(\widehat{x}_i, \sigma^2 I)}$ on the space of mean vector and covariance matrix $\mathbb{R}^p \times \mathbb{S}_+^p$.[2]

[2]Associated with any mixture of Gaussians $\mathbb{Q}_y$ on $\mathbb{R}^p$ is a

Moreover, define the set

$$\mathbb{S}^p_{\geq\sigma} \triangleq \{\Sigma \in \mathbb{S}^p_+ : \Sigma \succeq \sigma^2 I\} \subset \mathbb{S}^p_+$$

of covariance matrices whose eigenvalues are lower bounded by $\sigma^2 > 0$, where σ^2 is the isotropic variance of the smoothing convolution. Notice that we explicitly constrain the covariance matrices to be invertible so that the likelihood function of each Gaussian component is well-defined. For any $y \in \{0, 1\}$, we formally define the ambiguity set as

$$\mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}^\sigma_y) \triangleq \left\{ \mathbb{Q}_y : \begin{array}{l} \nu_y \in \mathcal{P}(\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}),\ \mathbb{W}_c(\nu_y, \widehat{\nu}_y) \leq \varepsilon_y \\ \mathbb{Q}_y \text{ is a Gaussian mixture associated with } \nu_y \end{array} \right\}.$$

Here, $\mathcal{P}(\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma})$ denotes the set of all possible distributions supported on $\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}$. Intuitively, $\mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}^\sigma_y)$ contains all Guassian mixtures $\mathbb{Q}_y$ associated with some ν_y having a distance less than or equal to ε_y from the nominal measure $\widehat{\nu}_y$. Thus each measure $\mathbb{Q}_y$ of the random vector $X|Y = y$ is a Gaussian mixture. Each distribution ν_y is a measure on the space of mean vector-covariance matrix $\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}$, and the distance between ν_y and $\widehat{\nu}_y$ is measured by an optimal transport distance $\mathbb{W}_c$. We will use in this paper the type-∞ Wasserstein distance, which is defined as follows.

Definition 4.1 (Type-∞ Wasserstein distance)**.** *Let c be a nonnegative, symmetric and continuous ground transport cost on $\Xi \triangleq \mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}$. The type-$\infty$ Wasserstein distance between two distributions $\nu_1,\ \nu_2 \in \mathcal{P}(\Xi)$ amounts to*

$$\mathbb{W}_c(\nu_1, \nu_2) \triangleq \inf_{\lambda \in \Lambda(\nu_1, \nu_2)} \left\{ \operatorname*{ess\,sup}_\lambda \left\{ c(\xi_1, \xi_2) : (\xi_1, \xi_2) \in \Xi \times \Xi \right\} \right\},$$

where $\Lambda(\nu_1, \nu_2)$ is the set of all couplings of ν_1 and ν_2.

It remains to specify the ground metric c on the space $\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}$. Because the space $\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}$ aims to model the mean vectors and the covariance matrices of Gaussian distributions, it is also natural to use a ground metric c that is inspired by the Wasserstein distance between Gaussian distributions. Fortunately, the Wasserstein type-2 distance between Gaussian measures is known in closed form [Olkin and Pukelsheim, 1982, Givens and Shortt, 1984].

Proposition 4.2 (Wasserstein type-2 distance between Gaussian distributions)**.** *The Wasserstein type-2 distance between two p-dimensional Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$ under the Euclidean ground metric amounts to $\mathbb{G}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})) = \sqrt{\|\mu - \widehat{\mu}\|_2^2 + \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]}$.*

probability measure ν_y on the mean-covariance space of $\mathbb{R}^p \times \mathbb{S}^p_+$ such that for any measurable set $\mathcal{S} \subseteq \mathbb{R}^p$

$$\mathbb{Q}_y(X \in \mathcal{S}) = \int_{\mathbb{R}^p \times \mathbb{S}^p_+} \int_{\mathcal{S}} f(\tilde{x}|\mu, \Sigma)\, \mathrm{d}\tilde{x}\, \nu_y(\mathrm{d}\mu, \mathrm{d}\Sigma),$$

where $f(\cdot\,|\mu, \Sigma)$ is the density function of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$.

Motivated by the above result, we endow the space $\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}$ with the cost function c defined as

$$c((\mu, \Sigma), (\widehat{\mu}, \widehat{\Sigma})) \triangleq \sqrt{\|\mu - \widehat{\mu}\|_2^2 + \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]}.$$

It is easy to see that c is non-negative, symmetric and continuous on $\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}$ and thus c is a valid ground cost for the Wasserstein distance $\mathbb{W}_c$ on $\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}$. We should point out that the Wasserstein distance has also been heavily used to construct ambiguity sets in the context of distributionally robust machine learning Nguyen et al. [2019b], Taskesen et al. [2021], Vu et al. [2022]. Our formulation of $\mathbb{W}_c$ is related with the family of optimal transport for Gaussian mixtures, which we discuss in the next remark.

Remark 4.3 (OT between Gaussian mixtures)**.** *Our construction relies on representing a Gaussian mixture distribution as a discrete distribution on the mean vector and covariance matrix space. This construction is motivated by recent work on optimal transport between Gaussian mixtures in Chen et al. [2019] and Delon and Desolneux [2020]. A clear distinction is that we use $\mathbb{W}_c$ as the type-∞ distance in Definition 4.1, while the existing literature focuses on type-1 and type-2 distance. As we later demonstrate in Lemma 5.3, the type-∞ construction is critical for the separability of the resulting problem.*

5 COMPUTATION

In this section, we delineate the solution procedure to find a robust Bayesian recourse with the Wasserstein-Gaussian mixture ambiguity sets formalized in Section 4. Fix any measure $\mathbb{Q}_y \in \mathbb{B}_{\varepsilon_y}(\widehat{\mathbb{P}}^\sigma_y)$, then $X|Y = y$ follows a mixture of Gaussian under $\mathbb{Q}_y$, and we let $L(x, \mathbb{Q}_y)$ be the Gaussian mixture likelihood of a point x under $\mathbb{Q}_y$. By internalizing the maximization term inside the fraction and replacing $\mathbb{Q}_y(X = x)$ by the likelihood $L(x, \mathbb{Q}_y)$, problem (2) is equivalent to

$$\min_{x \in \mathbb{X}} F(x), \quad F(x) \triangleq \frac{\gamma_0 \times \max\limits_{\mathbb{Q}_0 \in \mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}^\sigma_0)} L(x, \mathbb{Q}_0)}{\gamma_1 \times \min\limits_{\mathbb{Q}_1 \in \mathbb{B}_{\varepsilon_1}(\widehat{\mathbb{P}}^\sigma_1)} L(x, \mathbb{Q}_1)}.$$

In the sequence, we discuss how to evaluate the objective value $F(x)$, sketch the necessary proof and provide further insights to the likelihood evaluation problems.

5.1 REFORMULATIONS OF THE LIKELIHOOD EVALUATION PROBLEMS AND ROUTINES

For any $x \in \mathbb{X}$, evaluating its objective value $F(x)$ requires solving the maximization of the likelihood in the numerator

$$\max\{L(x,\mathbb{Q}_0):\mathbb{Q}_0\in\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma)\} \tag{3}$$

and the minimization of the likelihood in the denominator

$$\min\{L(x,\mathbb{Q}_1):\mathbb{Q}_1\in\mathbb{B}_{\varepsilon_1}(\widehat{\mathbb{P}}_1^\sigma)\}. \tag{4}$$

At this stage, it is important to relate problems (3) and (4) to the existing literature on (Bayesian) likelihood estimation/approximation. Problem (3) searches for a distribution that *maximizes* the likelihood of x over the set $\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma)$, and it is also known in the machine learning literature as an *optimistic* likelihood Nguyen et al. [2019a, 2020]. There is, however, a clear distinction between the existing results and the results of this paper: Nguyen et al. [2019a] use a Gaussian feasible set prescribed using the Fisher-Rao distance and Nguyen et al. [2020] use a moment-based feasible set using the Kullback-Leibler type divergence; in contrast, our set $\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma)$ is a mixture of Gaussian feasible set prescribed using a hierarchical Wasserstein distance. The attractiveness of the existing optimistic likelihood methods lies in their computational tractability. Next, we show that our optimistic likelihood under the Wasserstein-Gaussian mixture ambiguity set also possesses this tractability.

Theorem 5.1 (Optimistic likelihood)**.** *For each $i\in\mathcal{I}_0$, let α_i be the optimal value of the following two-dimensional optimization problem*

$$\min_{\substack{a\in\mathbb{R}_+,\ d_p\in[\sigma,+\infty)\\ a^2+(d_p-\sigma)^2\le\varepsilon_0^2}} \log d_p+\frac{(\|x-\widehat{x}_i\|_2-a)^2}{2d_p^2}+(p-1)\log\sigma.$$

Then, we have

$$\max\{L(x,\mathbb{Q}_0):\mathbb{Q}_0\in\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma)\}=\frac{\sum_{i\in\mathcal{I}_0}\exp(-\alpha_i)}{N_0(2\pi)^{p/2}}.$$

Theorem 5.1 asserts that we can solve problem (3) by solving N_0 individual subproblems, each subproblem is a two-dimensional minimization problem. Notice that the feasible set of each subproblem is relatively simple: it contains an ellipsoidal constraint and lower bounds on the variables. Hence, it is easy to devise a projection operator for this feasible set. Note that the objective function of the subproblem is non-convex.

Let us now focus our attention on problem (4): it searches for a distribution that *minimizes* the likelihood of x over all candidate distributions in $\mathbb{B}_{\varepsilon_1}(\widehat{\mathbb{P}}_1^\sigma)$, and it is termed the *pessimistic* likelihood. It has been previously noticed that the pessimistic likelihood is not easy to solve due to non-convexity [Nguyen et al., 2020, Appendix A]. Surprisingly, for our Wasserstein-Gaussian mixture set, we still can obtain the reformulation below.

Theorem 5.2 (Pessimistic likelihood)**.** *For each $i\in\mathcal{I}_1$, let α_i be the optimal value of the following two-dimensional optimization problem*

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in[\sigma,+\infty)\\ a^2+p(d_1-\sigma)^2\le\varepsilon_1^2}} \left\{-\log d_1-\frac{(\|x-\widehat{x}_i\|_2+a)^2}{2d_1^2} -(p-1)\log\left(\sigma+\sqrt{\frac{\varepsilon_1^2-a^2-(d_1-\sigma)^2}{p-1}}\right)\right\}.$$

Then, we have

$$\min\{L(x,\mathbb{Q}_1):\mathbb{Q}_1\in\mathbb{B}_{\varepsilon_1}(\widehat{\mathbb{P}}_1^\sigma)\}=\frac{\sum_{i\in\mathcal{I}_1}\exp(\alpha_i)}{N_1(2\pi)^{p/2}}.$$

Theorem 5.2 asserts that the pessimistic likelihood problem (4) admits a similar decomposable structure: solving (4) is equivalent to solving N_1 individual subproblems, each subproblem is a two-dimensional minimization problem with a non-convex objective function. Further, the feasible set of the subproblem is also of tractable form for projection.

Numerical routines. Equipped with Theorems 5.1 and 5.2, we can design an iterative scheme to solve the robust Bayesian recourse problem. For any value $x\in\mathbb{X}$, we can use a projected gradient descent to solve a series of two-dimensional subproblems to evaluate the objective value $F(x)$. In Appendix C, we elaborate on the construction of the projection operator as well as the algorithm to evaluate $F(x)$. To optimize $F(x)$ to find the robust recourse, we can also apply a similar algorithm, provided that the projection onto the feasible region $\mathbb{X}$ is easy to solve.

5.2 SKETCH OF PROOFS

We sketch here the main steps leading to the results in Section 5.1. Because $\widehat{\mathbb{P}}_y^\sigma$ is a Gaussian mixture and we are using a type-∞ Wasserstein distance to prescribe the neighborhood around the representable distribution, the likelihood evaluation problems admit a decomposable structure. This decomposability has also been exploited previously in the literature of operations management [Bertsimas et al., 2021], chance constrained programming [Xie, 2020] and fair classification [Wang et al., 2021]. In the sequel, we denote $f(x|\mu_i,\Sigma_i)$ the likelihood of x under the p-dimensional Gaussian distribution with a mean vector μ_i and a covariance matrix Σ_i:

$$f(x|\mu_i,\Sigma_i)=\frac{\exp\big(-\frac{1}{2}(x-\mu_i)^\top\Sigma_i^{-1}(x-\mu_i)\big)}{(2\pi)^{\frac{p}{2}}\det(\Sigma_i)}.$$

The next lemma asserts that the likelihood evaluation problem can be decomposed into solving smaller subproblems, each subproblem is an optimization problem over the mean vector - covariance matrix space $\mathbb{R}^p\times\mathbb{S}_{\succeq\sigma}^p$.

Lemma 5.3 (Separability)**.** *There exists a distribution $\mathbb{Q}_0^\star$ that solves* (3) *and is a mixture of at most N_0 Gaussian components. Moreover, problem* (3) *is equivalent to a separable*

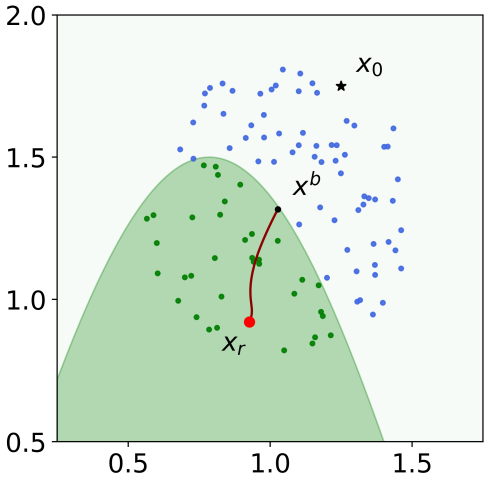

Figure 1: An example of the robust Bayesian recourse on a toy 2-dimensional instance. The star denotes the input x_0, and the black circle denotes the boundary point x^b. Green and blue circles are locally sampled data with favorable and unfavorable predicted values, respectively. The red circle denotes the robust Bayesian recourse, and the curved line denotes the continuum of intermediate solutions of the gradient descent algorithm. The robust Bayesian recourse moves to the interior of the favorable region (green), and thus is more likely to be valid subject to model shifts.

problem of the form

$$\begin{aligned}&\max\left\{L(x,\mathbb{Q}_0):\mathbb{Q}_0\in\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma)\right\}\\ &=\left\{\begin{array}{ll}\max & \frac{1}{N_0}\sum_{i\in\mathcal{I}_0} f(x|\mu_i,\Sigma_i)\\ \text{s.t.} & (\mu_i,\Sigma_i)\in\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}\\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2\bar{I}))\leq\varepsilon_0\quad\forall i\in\mathcal{I}_0.\end{array}\right.\end{aligned}$$

An analogous result holds for problem (4) *with the corresponding subscript* $y=1$.

Lemma 5.3 leverages the essential supremum in the definition of the type-∞ Wasserstein distance in Definition 4.1 to separate the problem into subproblem for each component. This separability is *not* obtainable under other types of the Wasserstein distance. It is important to bear in mind that each subproblem is still not easy: the objective function is neither convex nor concave in Σ_i. Further, we also need to evaluate both the maximization and the minimization counterparts, and tractability is difficult to be established simultaneously in both directions. Despite these difficulties, we can show that each subproblem, which is originally on the $\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}$ space, can be reduced to a 2-dimensional subproblem. This is in fact a significant reduction of dimensionality, and this reduction does not depend on the dimension p. First, we provide the reformulation for the maximization counterpart.

Proposition 5.4 (Maximization subproblem)**.** *Fix any index* $i\in\mathcal{I}_0$. *For any* $\widehat{x}_i\in\mathbb{R}^p$, $x\in\mathbb{R}^p$ *and* $\varepsilon_0\in\mathbb{R}_+$, *we have*

$$\frac{\exp(-\alpha_i)}{(2\pi)^{p/2}}=\left\{\begin{array}{ll}\max & f(x|\mu_i,\Sigma_i)\\ \text{s.t.} & (\mu_i,\Sigma_i)\in\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}\\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2\bar{I}))\leq\varepsilon_0,\end{array}\right.$$

where α_i *is the optimal value of the two-dimensional optimization problem*

$$\min_{\substack{a\in\mathbb{R}_+,\ d_p\in[\sigma,+\infty)\\ a^2+(d_p-\sigma)^2\leq\varepsilon_0^2}}\log d_p+\frac{(\|x-\widehat{x}_i\|_2-a)^2}{2d_p^2}+(p-1)\log\sigma.$$

The two auxiliary variables a and d_p have a specific meaning which can be explained as follows. Let $(\mu_i^\star,\Sigma_i^\star)$ be the optimal solution of the original maximization problem over (μ_i,Σ_i), and let $(a^\star,d_p^\star)$ be the optimal solution of the reduced problem over (a,d_p). We then have $\|\mu_i^\star-\widehat{x}_i\|_2=a^\star$ and $d_p^\star$ coincides with the *largest* eigenvalues of $\Sigma_i^\star$. Next, we expose the reformulation for the minimization problem.

Proposition 5.5 (Minimization subproblem)**.** *Fix any index* $i\in\mathcal{I}_1$. *For any* $\widehat{x}_i\in\mathbb{R}^p$, $x\in\mathbb{R}^p$ *and* $\varepsilon_1\in\mathbb{R}_+$, *we have*

$$\frac{\exp(\alpha_i)}{(2\pi)^{p/2}}=\left\{\begin{array}{ll}\min & f(x|\mu_i,\Sigma_i)\\ \text{s.t.} & (\mu_i,\Sigma_i)\in\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}\\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2\bar{I}))\leq\varepsilon_1,\end{array}\right.$$

where α_i *is the optimal value of the two-dimensional optimization problem*

$$\begin{aligned}\min_{\substack{a\in\mathbb{R}_+,\ d_1\in[\sigma,+\infty)\\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}}\Bigg\{&-\log d_1-\frac{(\|x-\widehat{x}_i\|_2+a)^2}{2d_1^2}\\ &-(p-1)\log\left(\sigma+\sqrt{\frac{\varepsilon^2-a^2-(d_1-\sigma)^2}{p-1}}\right)\Bigg\}.\end{aligned}$$

There is a similar relationship between $(\mu_i^\star,\Sigma_i^\star)$ which solves the original minimization problem and $(a^\star,d_1^\star)$ which solves the reduced problem: we have $\|\mu_i^\star-\widehat{x}_i\|_2=a^\star$ and $d_1^\star$ coincides with the *smallest* eigenvalues of $\Sigma_i^\star$.

The above discussion reveals that we can fully reconstruct the distribution $\mathbb{Q}_0^\star$ that solves (3) and $\mathbb{Q}_1^\star$ that solves (4) from the solutions of the reduced subproblems, we provide this reconstruction in Appendix D.

6 NUMERICAL EXPERIMENT

We evaluate in this section the robustness to model shifts of different recourses, together with the trade-off against the cost of adopting the recourse's recommendation. We compare our proposed robust Bayesian recourse method, namely RBR, against the counterfactual explanation of Wachter [Wachter et al., 2017] and against the robust recourse generated by ROAR [Upadhyay et al., 2021] using

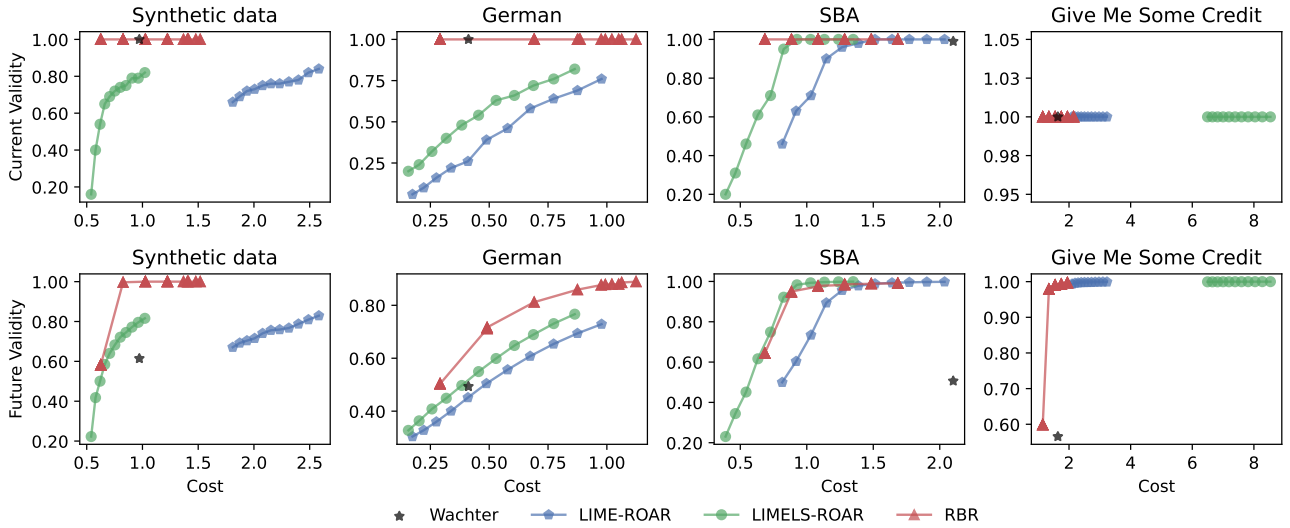

Figure 2: Pareto frontiers of the cost-validity trade-off with the MLP classifier, on synthetic, German Credit, Small Business Administration, and Give Me Some Credit datasets.

either LIME [Ribeiro et al., 2016] and LIMELS [Laugel et al., 2018] as a surrogate model[3].

6.1 EXPERIMENTAL SETUP

Datasets. We examine the recourse generators on both a synthetic dataset and the real-world datasets: *German Credit* [Dua and Graff, 2017, Groemping, 2019], *Small Bussiness Administration (SBA)* [Li et al., 2018], and *Give Me Some Credit (GMC)*. Each dataset contains two sets of data: D_1 and D_2. The former is the current data which is used to train current classifier to generate recourses. The latter represents the possible data arriving in the future.

For each dataset, we use 80% of the instances in the current data D_1 to train the underlying predictive model and fix this classifier to construct recourses for the remaining 20% of the instances. The future data D_2 will be used to train future classifiers, which are for evaluation only.

Classifier. We use a three-layer MLP with 20, 50 and 20 nodes, respectively with a ReLU activation in each consecutive layer. The sigmoid function is used in the last layer to produce predictive probabilities. The performance of the MLP classifier is reported in Table 1.

Sampling procedure. We employ the sampling scheme described in Section 2. We choose the number of counterfactuals $K = 1000$ and sample 200 synthetic samples uniformly with a sampling radius $r_p = 0.2$.

[3]While LIME samples synthetic data *globally* and train a weighted ridge regression, LIMELS generates the local surrogate model by training a (unweighted) ridge regression on the data sampled *locally* near by the closest counterfactual of the input instance (similar to the sampling procedure described in Section 2).

Metrics. To measure the ease of adopting a recourse, we use the ℓ_1-distance as the cost function φ on the covariate space $\mathcal{X}$, this choice is similar to Ustun et al. [2019] and Upadhyay et al. [2021]. We define the *current validity* as the validity of the recourses with respect to the current classifier $\mathcal{C}$. To evaluate the robustness of recourses to the changes in model's parameters, we sample 20% of the instances in the data set D_2 as the arrival data. We then re-train the classifier with the old data (80% of D_1) coupled with this arrival data to simulate the future classifiers $\tilde{\mathcal{C}}$. We repeat this procedure 100 times to obtain 100 future classifiers and report the *future validity* of a recourse as the fraction of the future classifiers with respect to which the recourse is valid.

6.2 EXPERIMENTAL DETAILS

We use both synthetic and real-world datasets.

Synthetic dataset. We synthesize the 2-dimensional data by sampling 1000 instances uniformly in a rectangle $[-2, 4] \times [-2, 7]$. For each sample, we label using the function $f(x) = 1$ if $x_2 \geq 1 + x_1 + 2x_1^2 + x_1^3 - x_1^4 + \varepsilon$, and $f(x) =$ otherwise, where ε is a random noise. We set $\varepsilon = 0$ when generating the present set D_1 and $\varepsilon \sim \mathcal{N}(0, 1)$ for the future set D_2.

Real-world datasets. Three real-world datasets are used.

- *German Credit* [Dua and Graff, 2017]. The dataset contains the information (e.g. age, gender, financial status,...) of 1000 customers who took a loan from a bank. The classification task is to determine the risk (good or bad) of an individual. There is another version of this dataset regarding to corrections of coding error [Groemping, 2019]. We use the corrected version of

Dataset	Present data D_1		Shift data D_2	
	Accuracy	*AUC*	*Accuracy*	*AUC*
Synthetic data	0.99 ± 0.00	1.00 ± 0.00	0.94 ± 0.01	0.99 ± 0.01
German Credit	0.67 ± 0.02	0.60 ± 0.03	0.66 ± 0.23	0.60 ± 0.04
SBA	0.96 ± 0.00	0.99 ± 0.00	0.98 ± 0.01	0.96 ± 0.01
GMC	0.94 ± 0.00	0.84 ± 0.00	0.94 ± 0.00	0.84 ± 0.00

Table 1: Accuracy and AUC results of the MLP classifier on the synthetic and real-world datasets.

this dataset as a shifted data to capture correction shift. The features we used in this dataset include 'duration', 'amount', 'personal_status_sex', and 'age'.

- *Small Bussiness Administration (SBA)* [Li et al., 2018]. This dataset includes 2102 observations of small business loan approvals from 1987 to 2014. We divide it into two datasets (one is instances from 1989 - 2006 and one is instances from 2006 - 2014) to capture temporal shift. We use the following features: 'Term', 'NoEmp', 'CreateJob', 'RetainedJob', 'UrbanRural', 'ChgOffPrinGr', 'GrAppv', 'SBA_Appv', 'New', 'RealEstate', 'Portion', 'Recession'.
- *Give Me Some Credit (GMC)*[4]. This dataset is used to predict if a person would experience financial distress in the next two years. Given 150000 entries from the available dataset, we randomly shuffle and partition the data equally into the current set D_1 and the shifted set D_2. Each entry contains 10 features: 'RevolvingUtilizationOfUnsecuredLines', 'age', 'NumberOfTime30-59DaysPastDueNotWorse', 'DebtRatio', 'MonthlyIncome', 'NumberOfOpenCreditLinesAndLoans', 'NumberOfTimes90DaysLate', 'NumberRealEstateLoansOrLines', 'NumberOfTime60-89DaysPastDueNotWorse', 'NumberOfDependents'.

6.3 COST-VALIDITY TRADE-OFF

We obtain the Pareto front for the trade-off between the cost of adopting recourses produced by RBR and their validity by varying the ambiguity sizes ε_1 and ε_0, along the maximum recourse cost δ, with $\delta = \|x_0 - x_b\|_1 + \delta_+$. Particularly, we consider $\sigma = 1.0$, $\varepsilon_0, \varepsilon_1 \in \{0.5k \mid k = 0, \ldots, 2\}$, and $\delta_+ \in \{0.2l \mid l = 0, \ldots, 5\}$. The frontiers for ROAR-based methods are obtained by varying $\delta_{\max} \in \{0.02m \mid m = 0, \ldots, 10\}$, where $\delta_{\max}$ is the tuning parameter of ROAR. As shown in Figure 2, increasing ε_1 and δ_+ generally increase the future validity of recourses yielded by RBR at the sacrifice of the cost, while sustaining the current validity. Yet, the frontiers obtained by RBR either dominate or comparable to other frontiers of Wachter, LIME-ROAR, and LIMELS-ROAR.

Conclusions. In this work, we proposed the robust Bayesian recourse which aims to be effective at reversing algorithmic outcome under potential model shifts. It is a model-agnostic approach that does not require approximating the nonlinear classifier by a linear surrogate. Instead, the robust Bayesian recourse minimizes directly the worst-case posterior probability odds ratio subject to the cost constraint bound. The robustness is designed with respect to the Wasserstein-Gaussian mixture ambiguity sets of the conditional distributions, in which the neighborhood is prescribed using an optimal transport (type-∞ Wasserstein) distance. We showed that the min-max recourse problem can be optimized using a gradient descent algorithm, which exploits separability and dimensionality reduction when evaluating the objective value. Our experiments on synthetic and real-world datasets demonstrate that the robust Bayesian recourse is more robust at a lower cost than other baselines.

While this paper focus on algorithmic transparency, we note that transparency may lead to the tension between transparency and gaming-the-system behaviors: the greater transparent be the decision process the more opportunity for exploitative manipulations [Yan and Zhang, 2022]. We envision that robustness techniques may alleviate these gaming behaviors and may lead to more trustworthy guarantees of (machine learning) algorithms.

Acknowledgments. Man-Chung Yue is supported by the HKRGC under the General Research Fund project 15305321.

[4]https://www.kaggle.com/competitions/GiveMeSomeCredit/data

References

Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Steven Wu, and Himabindu Lakkaraju. Towards the unification and robustness of perturbation and gradient based explanations. In *Proceedings of the 38th International Conference on Machine Learning*, pages 110–119, 2021.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

André Artelt and Barbara Hammer. On the computation of counterfactual explanations - A survey. *arXiv preprint arXiv:1911.07749*, 2019.

Amir Beck. *First-order Methods in Optimization*. SIAM, 2017.

Dimitris Bertsimas, Shimrit Shtern, and Bradley Sturt. Technical note–Two-stage sample robust optimization. *Operations Research*, 2021.

J Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Science & Business Media, 2013.

Ngoc Bui, Duy Nguyen, and Viet Anh Nguyen. Counterfactual plans under distributional ambiguity. In *International Conference on Learning Representations*, 2022.

Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2019.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1310–1320, 2019.

Julie Delon and Agnès Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/`.

Gordon Garisch and Akber Merchant. Model lifecycle transformation: How banks are unlocking efficiencies, April 2019. URL `https://financialservicesblog.accenture.com/model-lifecycle-transformation-how-banks-are-unlocking-efficiencies`.

John Geweke. Bayesian comparison of econometric models. Technical report, 1994.

C.R. Givens and R.M. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

U Groemping. South German credit data: Correcting a widely used data set. *Reports in Mathematics, Physics and Chemistry, Department II, Beuth University of Applied Sciences Berlin*, 2019.

Tyler Hicks, Liliana Rodríguez-Campos, and Jeong Hoon Choi. Bayesian posterior odds ratios: Statistical tools for collaborative evaluations. *American Journal of Evaluation*, 39(2):278–289, 2018.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. *arXiv preprint arXiv:1905.11190*, 2020.

Amirhossein Karimi, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *arXiv preprint arXiv:2010.04050*, 2021.

Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, 294:103459, 2021. ISSN 0004-3702.

Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity. *arXiv preprint arXiv:1806.07498*, 2018.

Min Li, Amy Mickel, and Stanley Taylor. "Should this loan be approved or denied?": A large dataset with class assignment guidelines. *Journal of Statistics Education*, 26(1):55–66, 2018.

Robert McCulloch and Peter E. Rossi. A Bayesian approach to testing the arbitrage pricing theory. *Journal of Econometrics*, 49(1):141–168, 1991.

Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. GANterfactual - counterfactual explanations for medical non-experts using generative adversarial learning. 2021.

Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *ArXiv*, abs/2111.00358, 2021.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. Calculating optimistic likelihoods using (geodesically) convex optimization. In *Advances in Neural Information Processing Systems 32*, pages 13942–13953, 2019a.

Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. Optimistic distributionally robust optimization for nonparametric likelihood approximation. In *Advances in Neural Information Processing Systems 32*, pages 15872–15882, 2019b.

Viet Anh Nguyen, Nian Si, and Jose Blanchet. Robust Bayesian classification using an optimistic score ratio. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *UAI*, 2020.

Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. CARLA: A Python library to benchmark algorithmic recourse and counterfactual explanation algorithms. *arXiv preprint arXiv:2108.00783*, 2021.

Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodríguez, Tijl De Bie, and Peter A. Flach. FACE: Feasible and actionable counterfactual explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv preprint arXiv:2012.11788*, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

Sima Sajjadiani, Aaron Sojourner, John Kammeyer-Mueller, and Elton Mykerezi. Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, 104, 03 2019.

M. J. Schervish. *Theory of Statistics*. Springer, 1995.

Naeem Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, 2012.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post-hoc explanation methods. *arXiv preprint arXiv:1911.02508*, 2020.

Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

Bahar Taskesen, Man-Chung Yue, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. Sequential domain adaptation by synthesizing distributionally robust experts. In *International Conference on Machine Learning*, pages 10162–10172. PMLR, 2021.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. In *Advances in Neural Information Processing Systems*, 2021.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 284–293, 2020.

Georgios Vlassopoulos, Tim van Erven, Henry Brighton, and Vlado Menkovski. Explaining predictions by approximating the local decision boundary. *arXiv preprint arXiv:2006.07985*, 2020.

Hieu Vu, Toan Tran, Man-Chung Yue, and Viet Anh Nguyen. Distributionally robust fair principal components via geodesic descents. *arXiv preprint arXiv:2202.03071*, 2022.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841, 2017.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 2018.

Yijie Wang, Viet Anh Nguyen, and Grani Hanasusanto. Wasserstein robust support vector machines with fairness constraints. *arXiv preprint arXiv:2103.06828*, 2021.

Austin Waters and Risto Miikkulainen. Grade: Machine learning support for graduate admissions. *AI Magazine*, 35(1):64–64, 2014.

Adam White and Artur d'Avila Garcez. Measurable counterfactual local explanations for any classifier. *arXiv preprint arXiv:1908.03020*, 2019.

Weijun Xie. Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ Wasserstein ball. *Operations Research Letters*, 48(4):513–523, 2020.

Geeta Dharani. Y, Nimisha G Nair, Pallavi Satpathy, and Jabez Christopher. Covariate shift: A review and analysis on classifiers. In *2019 Global Conference for Advancement in Technology (GCAT)*, pages 1–6, 2019.

Tom Yan and Chicheng Zhang. Margin-distancing for safe model explanation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 5104–5134, 2022.

Arnold Zellner. Posterior odds ratios for regression hypotheses: General considerations and some specific results. *Journal of Econometrics*, 16(1):151–152, 1981.

A PROOFS OF SECTION 5

Lemma 5.3 (re-stated). *There exists a distribution $\mathbb{Q}_0^\star$ that solves* (3) *and is a mixture of at most N_0 Gaussian components. Moreover, problem* (3) *is equivalent to a separable problem of the form*

$$\max\left\{L(x,\mathbb{Q}_0):\mathbb{Q}_0\in\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma)\right\}=\left\{\begin{array}{ll}\max & \frac{1}{N_0}\sum_{i\in\mathcal{I}_0}f(x|\mu_i,\Sigma_i)\\ \text{s.t.} & (\mu_i,\Sigma_i)\in\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}\\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2\bar{I}))\leq\varepsilon_0\quad\forall i\in\mathcal{I}_0.\end{array}\right.$$

An analogous result holds for problem (4) *with the corresponding subscript $y=1$.*

Proof of Lemma 5.3. There exists a distribution $\mathbb{Q}_0^\star$ that solves (3) and is a mixture of at most N_0 Gaussian components. Moreover, problem (3) is equivalent to a separable problem of the form

$$\begin{array}{l}\max\left\{L(x,\mathbb{Q}_0):\mathbb{Q}_0\in\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma)\right\}\\ =\left\{\begin{array}{ll}\max & \frac{1}{N_0}\sum_{i\in\mathcal{I}_0}f(x|\mu_i,\Sigma_i)\\ \text{s.t.} & (\mu_i,\Sigma_i)\in\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}\\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2\bar{I}))\leq\varepsilon_0\quad\forall i\in\mathcal{I}_0.\end{array}\right.\end{array}$$

We use $\forall i$ implies $\forall i\in\mathcal{I}_0$, and $\sum_i$ is also taken over the same set. Given any x, the likelihood of x under any Gaussian mixture $\mathbb{Q}_0$ can be written using the corresponding measure ν_0 as

$$L(x,\mathbb{Q}_0)=\int_{\mathbb{R}^p\times\mathbb{S}^p_+}f(x|\mu,\Sigma)\nu_0(\mathrm{d}\mu,\mathrm{d}\Sigma).$$

Recall that $\Xi=\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}$. Using the definition of the type-∞ Wasserstein, we find

$$\begin{array}{ll}&\mathbb{W}_c(\nu_0,\widehat{\nu}_0)\leq\varepsilon_0\\ \Leftrightarrow&\exists\lambda\in\Lambda(\nu_0,\widehat{\nu}_0)\text{ such that}\\ &\operatorname{ess\,sup}_\lambda\left\{c((\mu,\Sigma),(\mu',\Sigma')):(\mu,\Sigma,\mu',\Sigma')\in\Xi\times\Xi\right\}\leq\varepsilon_0\\ \Leftrightarrow&\forall i\ \exists\lambda_i\in\mathcal{P}(\Xi)\text{ such that}\\ &\operatorname{ess\,sup}_{\lambda_i}\left\{c((\mu,\Sigma),(\widehat{x}_i,\sigma I)):(\mu,\Sigma)\in\Xi\right\}\leq\varepsilon_0\\ \Leftrightarrow&\forall i\ \exists\lambda_i\in\mathcal{P}(\Xi)\text{ such that}\\ &c((\mu,\Sigma),(\widehat{x}_i,\sigma I))\leq\varepsilon_0\quad(\mu,\Sigma)\in\operatorname{supp}(\lambda_i),\end{array}$$

where the second equivalence follows from that $\widehat{\nu}_0=\frac{1}{N_0}\sum_i\delta_{(\widehat{x}_i,\sigma^2 I)}$ and hence any $\lambda\in\Lambda(\nu_0,\widehat{\nu}_0)$ takes the form $\frac{1}{N_0}\sum_i\lambda_i\otimes\delta_{(\widehat{x}_i,\sigma^2 I)}$ for some probability measures $\lambda_i\in\mathcal{P}(\Xi)$, and the third equivalence follows from Lemma A.1. Hence, problem (3) is equivalent to

$$\begin{array}{l}\left\{\begin{array}{ll}\max & \int_{\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}}f(x|\mu,\Sigma)\nu_0(\mathrm{d}\mu,\mathrm{d}\Sigma)\\ \text{s.t.} & \nu_0\in\mathcal{P}(\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma})\\ & \mathbb{W}_c(\nu_0,\widehat{\nu}_0)\leq\varepsilon_0\end{array}\right.\\ =\left\{\begin{array}{ll}\max & \frac{1}{N_0}\sum_i\int_{\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}}f(x|\mu_i,\Sigma_i)\lambda_i(\mathrm{d}\mu_i,\mathrm{d}\Sigma_i)\\ \text{s.t.} & \lambda_i\in\mathcal{P}(\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma})\qquad\forall i\\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2 I))\leq\varepsilon_0\quad\forall(\mu_i,\Sigma_i)\in\operatorname{supp}(\lambda_i)\qquad\forall i.\end{array}\right.\end{array}$$

It is easy now to employ a greedy argument to show that the optimal solution for λ_i should be a Dirac delta distribution supported on one point in the space of $\mathbb{R}^p\times\mathbb{S}^p_{\geq\sigma}$. This leads to the conclusion regarding the maximization problem (3).

An similar argument can be applied for the minimization problem (4), the detailed proof is omitted. □

Lemma A.1. *For any $\lambda\in\mathcal{P}(\Xi)$, $\widehat{x}\in\mathbb{R}^p$, $\sigma,\varepsilon>0$ and any function $c:\Xi\times\Xi\to\mathbb{R}$ such that the map $(\mu,\Sigma)\mapsto c((\mu,\Sigma),(\widehat{x},\sigma^2 I))$ is continuous, we have* $\operatorname{ess\,sup}_\lambda c((\mu,\Sigma),(\widehat{x},\sigma^2 I))\leq\varepsilon$ *if and only if $c((\mu,\Sigma),(\widehat{x},\sigma^2 I))\leq\varepsilon$ for any $(\mu,\Sigma)\in\operatorname{supp}(\lambda)$.*

Proof of Lemma A.1. We first prove the "only if" direction. Suppose that there exists $(\mu', \Sigma') \in \operatorname{supp}(\lambda)$ such that

$$c((\mu', \Sigma'), (\widehat{x}, \sigma^2 I)) > \varepsilon.$$

By continuity of the map $(\mu, \Sigma) \mapsto c((\mu, \Sigma), (\widehat{x}, \sigma^2 I))$, there exists an open neighbourhood $U \subseteq \Xi$ containing (μ', Σ') such that

$$c((\mu, \Sigma), (\widehat{x}, \sigma^2 I)) > \varepsilon \quad \forall (\mu, \Sigma) \in U.$$

By the definition of of support, $\lambda(U) > 0$. Therefore,

$$\Pr_{\lambda}(c((\mu, \Sigma), (\widehat{x}, \sigma^2 I)) \leq \varepsilon) = 1 - \Pr_{\lambda}(c((\mu, \Sigma), (\widehat{x}, \sigma^2 I)) > \varepsilon) \leq 1 - \lambda(U) < 1,$$

which contradicts to that $\operatorname{ess\,sup}_{\lambda} c((\mu, \Sigma), (\widehat{x}, \sigma^2 I)) \leq \varepsilon$.

We next prove the "if" direction. By the law of total probability and the fact that $c((\mu, \Sigma), (\widehat{x}, \sigma^2 I)) \leq \varepsilon$ for any $(\mu, \Sigma) \in \operatorname{supp}(\lambda)$,

$$\begin{aligned}
&\Pr_{\lambda}\left(c((\mu, \Sigma), (\widehat{x}, \sigma^2 I)) \leq \varepsilon\right) \\
=&\Pr_{\lambda}\left(c((\mu, \Sigma), (\widehat{x}, \sigma^2 I)) \leq \varepsilon | (\mu, \Sigma) \in \operatorname{supp}(\lambda)\right) \lambda(\operatorname{supp}(\lambda)) \\
&+ \Pr_{\lambda}\left(c((\mu, \Sigma), (\widehat{x}, \sigma^2 I)) \leq \varepsilon | (\mu, \Sigma) \notin \operatorname{supp}(\lambda)\right) (1 - \lambda(\operatorname{supp}(\lambda))) \\
=&1 \cdot 1 + \Pr_{\lambda}\left(c((\mu, \Sigma), (\widehat{x}, \sigma^2 I)) \leq \varepsilon | (\mu, \Sigma) \notin \operatorname{supp}(\lambda)\right) \cdot 0 = 1,
\end{aligned}$$

which completes the proof. □

Proposition 5.4 (re-stated). *Fix any index $i \in \mathcal{I}_0$. For any $\widehat{x}_i \in \mathbb{R}^p$, $x \in \mathbb{R}^p$ and $\varepsilon_0 \in \mathbb{R}_+$, we have*

$$\frac{\exp(-\alpha_i)}{(2\pi)^{p/2}} = \left\{ \begin{array}{ll} \max & f(x|\mu_i, \Sigma_i) \\ \text{s.t.} & (\mu_i, \Sigma_i) \in \mathbb{R}^p \times \mathbb{S}^p_{\geq \sigma} \\ & c((\mu_i, \Sigma_i), (\widehat{x}_i, \sigma^2 I)) \leq \varepsilon_0, \end{array} \right.$$

where α_i is the optimal value of the two-dimensional optimization problem

$$\min_{\substack{a \in \mathbb{R}_+,\ d_p \in [\sigma, +\infty) \\ a^2 + (d_p - \sigma)^2 \leq \varepsilon_0^2}} \log d_p + \frac{(\|x - \widehat{x}_i\|_2 - a)^2}{2 d_p^2} + (p-1) \log \sigma.$$

Proof of Proposition 5.4. Let α_i be the optimal value of the negative log-likelihood minimization problem

$$\alpha_i = \left\{ \begin{array}{ll} \min & \frac{1}{2} \log\det \Sigma_i + \frac{1}{2} (x - \mu_i)^\top \Sigma_i^{-1} (x - \mu_i) \\ \text{s.t.} & \mu_i \in \mathbb{R}^p,\ \Sigma_i \in \mathbb{S}^p_+ \\ & \|\mu_i - \widehat{x}_i\|_2^2 + \operatorname{Tr}\left[\Sigma_i + \sigma^2 I - 2\big((\sigma^2 I)^{\frac{1}{2}} \Sigma_i (\sigma^2 I)^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \leq \varepsilon_0^2 \\ & \Sigma_i \succeq \sigma^2 I. \end{array} \right.$$

It is easy to see that

$$\max\{f(x|\mu_i, \Sigma_i) : (\mu_i, \Sigma_i) \in \mathbb{R}^p \times \mathbb{S}^p_{\geq \sigma},\ c((\mu_i, \Sigma_i), (\widehat{x}_i, \sigma^2 I)) \leq \varepsilon_0\} = \frac{1}{\sqrt{(2\pi)^p}} \exp(-\alpha_i).$$

It remains to provide a simpler formulation to determine α_i. To simplify the notation, we omit the index i on all variables and parameters. We reparameterize $\Sigma = V \operatorname{diag}(d^2) V^\top$ for a vector $d \in \mathbb{R}^p_+$, where $\operatorname{diag}(d^2)$ denotes a $\mathbb{R}^{p \times p}$ diagonal matrix with its j-th diagonal entries equals to d_j^2, and $\mathrm{O}(p)$ is the set of p-dimensional orthogonal matrices

$$\mathrm{O}(p) = \{V \in \mathbb{R}^{p \times p} : V^\top V = I_p\}.$$

The negative log-likelihood minimization problem is further equivalent to

$$\begin{array}{ll} \min & \sum_{j=1}^p \log d_j + \frac{1}{2} (V^\top (x - \mu))^\top \operatorname{diag}(d^{-2}) (V^\top (x - \mu)) \\ \text{s.t.} & d \in \mathbb{R}^p_+,\ V \in \mathrm{O}(p),\ \mu \in \mathbb{R}^p \\ & \|\mu - \widehat{x}\|_2^2 + \sum_{j=1}^p (d_j - \sigma)^2 \leq \varepsilon_0^2 \\ & d \geq \sigma, \end{array}$$

where $d \geq \sigma$ implies the element-wise constraints $d_j \geq \sigma$ for any $j = 1, \ldots, p$. We introduce an auxiliary variable $a \in \mathbb{R}_+$ and rewrite the optimization problem in an equivalent way as

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}^p_+,\ d\geq\sigma \\ a^2+\sum_j (d_j-\sigma)^2\leq\varepsilon_0^2}} \ \min_{\substack{\mu\in\mathbb{R}^p \\ \|\mu-\widehat{x}\|_2^2=a^2}} \ \min_{V\in \mathrm{O}(p)} \ \sum_{j=1}^p \log d_j + \frac{1}{2}(V^\top(x-\mu))^\top \operatorname{diag}(d^{-2})(V^\top(x-\mu)).$$

Notice that the above optimization problem is invariant to the ordering of the entries of d. As a consequence, without any loss of generality, we can assume that d_p is the maximum value across all d_j. By Lemma B.1, the above optimization problem becomes

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}^p_+,\ d\geq\sigma \\ a^2+\sum_j (d_j-\sigma)^2\leq\varepsilon_0^2 \\ d_p=\max\{d\}}} \ \min_{\substack{\mu\in\mathbb{R}^p \\ \|\mu-\widehat{x}\|_2^2=a^2}} \ \sum_{j=1}^p \log d_j + \frac{1}{2d_p^2}\|x-\mu\|_2^2.$$

Using Lemma B.2, we obtain the equivalent optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}^p_+,\ d\geq\sigma \\ a^2+\sum_j (d_j-\sigma)^2\leq\varepsilon_0^2 \\ d_p=\max\{d\}}} \ \sum_{j=1}^p \log d_j + \frac{1}{2d_p^2}(\|x-\widehat{x}\|_2 - a)^2.$$

Rewriting the above problem into a two-layer optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d_p\in\mathbb{R}_+,\ d_p\geq\sigma \\ a^2+(d_p-\sigma)^2\leq\varepsilon_0^2}} \left\{ \log d_p + \frac{1}{2d_p^2}(\|x-\widehat{x}\|_2 - a)^2 + \min_{\substack{d_j\in\mathbb{R}_+,\ d_j\geq\sigma\ \forall j=1,\ldots,p-1 \\ \sum_{j=1}^{p-1}(d_j-\sigma)^2\leq\varepsilon_0^2-a^2-(d_p-\sigma)^2 \\ d_j\leq d_p\ \forall j=1,\ldots,p-1}} \sum_{j=1}^{p-1} \log d_j \right\}. \tag{5}$$

Notice that for any d_p that is feasible for the outer minimization problem, the inner minimization problem over d_j, $\forall j = 1, \ldots, p-1$ admits a non-empty feasible set. Indeed, because $d_p \geq \sigma$, the value $d_j = \sigma$, $j = 1, \ldots, p-1$ is a feasible solution for the inner problem. We now focus on solving the inner minimization problem. As $\log(\,\cdot\,)$ is an increasing function, for any $s \geq 0$, we find

$$\min_{\substack{d_p\geq d_j\geq\sigma\ \forall j=1,\ldots,p-1 \\ \sum_{j=1}^{p-1}(d_j-\sigma)^2\leq s}} \ \sum_{j=1}^{p-1} \log d_j = (p-1)\log\sigma,$$

which holds because the optimization problem on the left hand side admits the optimal solution $d_j^\star = \sigma$ for all $j = 1, \ldots, p-1$. This completes the proof. □

Proposition 5.5 (re-stated). *Fix any index $i \in \mathcal{I}_1$. For any $\widehat{x}_i \in \mathbb{R}^p$, $x \in \mathbb{R}^p$ and $\varepsilon_1 \in \mathbb{R}_+$, we have*

$$\frac{\exp(\alpha_i)}{(2\pi)^{p/2}} = \left\{ \begin{array}{ll} \min & f(x|\mu_i, \Sigma_i) \\ \text{s.t.} & (\mu_i, \Sigma_i) \in \mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma} \\ & c((\mu_i, \Sigma_i), (\widehat{x}_i, \sigma^2 I)) \leq \varepsilon_1, \end{array} \right.$$

where α_i is the optimal value of the two-dimensional optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in[\sigma,+\infty) \\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}} \left\{ -\log d_1 - \frac{(\|x-\widehat{x}_i\|_2 + a)^2}{2d_1^2} - (p-1)\log\left(\sigma + \sqrt{\frac{\varepsilon^2 - a^2 - (d_1-\sigma)^2}{p-1}}\right) \right\}.$$

Proof of Proposition 5.5. Let α_i be the optimal value of the log-likelihood *minimization* problem

$$\alpha_i = \left\{ \begin{array}{ll} \min & -\frac{1}{2}\log\det\Sigma_i - \frac{1}{2}(x-\mu_i)^\top\Sigma^{-1}(x-\mu_i) \\ \text{s.t.} & \mu_i \in \mathbb{R}^p,\ \Sigma_i \in \mathbb{S}^p_+ \\ & \|\mu_i - \widehat{x}_i\|_2^2 + \operatorname{Tr}\left[\Sigma_i + \sigma^2 I - 2\big((\sigma^2 I)^{\frac{1}{2}}\Sigma_i(\sigma^2 I)^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \leq \varepsilon_1^2 \\ & \Sigma_i \succeq \sigma^2 I. \end{array} \right.$$

It is easy to see that

$$\min\{f(x|\mu_i,\Sigma_i) : (\mu_i,\Sigma_i) \in \mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma},\ c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2 I)) \leq \varepsilon_1\} = \frac{1}{(2\pi)^{p/2}}\exp(\alpha_i).$$

It remains to provide the computational routine to determine α_i. To simplify the notation, we omit the index i on all variables and parameters. We reparameterize $\Sigma = V\,\mathrm{diag}(d^2)V^\top$ for a vector $d \in \mathbb{R}^p_+$, where $\mathrm{diag}(d^2)$ denotes a $\mathbb{R}^{p\times p}$ diagonal matrix with its j-th diagonal entries equals to d_j^2, and $\mathrm{O}(p)$ is the set of p-dimensional orthogonal matrices

$$\mathrm{O}(p) = \{V \in \mathbb{R}^{p\times p} : V^\top V = I_p\}.$$

The log-likelihood minimization problem is further equivalent to

$$\begin{array}{ll}\min & -\sum_{j=1}^p \log d_j - \frac{1}{2}(V^\top(x-\mu))^\top \mathrm{diag}(d^{-2})(V^\top(x-\mu)) \\ \text{s.t.} & d \in \mathbb{R}^p_+,\ V \in \mathrm{O}(p),\ \mu \in \mathbb{R}^p \\ & \|\mu - \widehat{x}\|_2^2 + \sum_{j=1}^p (d_j-\sigma)^2 \leq \varepsilon_1^2 \\ & d \geq \sigma,\end{array}$$

where $d \geq \sigma$ implies the element-wise constraints $d_j \geq \sigma$ for any $j = 1,\ldots,p$. We introduce an auxiliary variable $a \in \mathbb{R}_+$ and rewrite the optimization problem in an equivalent way as

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}^p_+,\ d\geq\sigma \\ a^2+\sum_j(d_j-\sigma)^2\leq\varepsilon_1^2}} \ \min_{\substack{\mu\in\mathbb{R}^p \\ \|\mu-\widehat{x}\|_2^2=a^2}} \ \min_{V\in\mathrm{O}(p)} -\sum_{j=1}^p \log d_j - \frac{1}{2}(V^\top(x-\mu))^\top \mathrm{diag}(d^{-2})(V^\top(x-\mu)).$$

Notice that the above optimization problem is invariant to the ordering of the entries of d. As a consequence, without any loss of generality, we can assume that d_1 is the minimum value across all d_j. By Lemma B.1, the above optimization problem becomes

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}^p_+,\ d\geq\sigma \\ a^2+\sum_j(d_j-\sigma)^2\leq\varepsilon_1^2 \\ d_1=\min\{d\}}} \ \min_{\substack{\mu\in\mathbb{R}^p \\ \|\mu-\widehat{x}\|_2^2=a^2}} -\sum_{j=1}^p \log d_j - \frac{1}{2d_1^2}\|x-\mu\|_2^2.$$

Using Lemma B.2, we obtain the equivalent optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}^p_+,\ d\geq\sigma \\ a^2+\sum_j(d_j-\sigma)^2\leq\varepsilon_1^2 \\ d_1=\min\{d\}}} -\sum_{j=1}^p \log d_j - \frac{1}{2d_1^2}(\|x-\widehat{x}\|_2 + a)^2.$$

Notice that the constraint $\sigma \leq d_1 = \min\{d\}$ implies that $p(d_1-\sigma)^2 \leq \sum_j (d_j-\sigma)^2$. As a consequence, any feasible value for d_1 should satisfy $a^2 + p(d_1-\sigma)^2 \leq \varepsilon_1^2$. Separating the variable d into two groups d_1 and $d_2,\ldots,d_p$ leads to a two-layer optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in\mathbb{R}_+,\ d_1\geq\sigma \\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}} \left\{ -\log d_1 - \frac{1}{2d_1^2}(\|x-\widehat{x}\|_2 + a)^2 + \min_{\substack{d_j\in\mathbb{R}_+,\ d_j\geq d_1\ \forall j=2,\ldots,p \\ \sum_{j=2}^p (d_j-\sigma)^2 \leq \varepsilon_1^2 - a^2 - (d_1-\sigma)^2}} -\sum_{j=2}^p \log d_j \right\}. \tag{6}$$

Consider momentarily the minimization problem

$$\min_{\substack{d_j\in\mathbb{R}_+ \\ \forall j=2,\ldots,p \\ \sum_{j=2}^p (d_j-\sigma)^2 \leq \varepsilon_1^2 - a^2 - (d_1-\sigma)^2}} -\sum_{j=2}^p \log d_j,$$

where the constraints $d_j \geq d_1$ have been intentionally omitted. Proposition B.3 asserts that this optimization problem has the optimal value

$$-(p-1)\log\left(\sigma + \sqrt{\frac{\varepsilon_1^2 - a^2 - (d_1-\sigma)^2}{p-1}}\right)$$

at the optimal solution $d_j^\star = \sigma + \sqrt{\frac{\varepsilon_1^2 - a^2 - (d_1-\sigma)^2}{p-1}}$, which also by the outer constraint $a^2 + p(d_1-\sigma)^2 \le \varepsilon_1^2$ satisfies $d_j \ge d_1 \; \forall j = 2, \ldots, p$. Thus it is indeed the optimal solution to the inner minimization problem in (6). As a consequence, problem (6) is equivalent to

$$\min_{\substack{a\in\mathbb{R}_+,\, d_1\in\mathbb{R}_+,\, d_1\ge\sigma \\ a^2+p(d_1-\sigma)^2\le\varepsilon_1^2}} \left\{ -\log d_1 - \frac{1}{2d_1^2}(\|x-\widehat{x}\|_2 + a)^2 - (p-1)\log\left(\sigma + \sqrt{\frac{\varepsilon_1^2 - a^2 - (d_1-\sigma)^2}{p-1}}\right)\right\}.$$

This completes the proof. □

B AUXILIARY RESULTS

The following preparatory results are necessary to prove Propositions 5.4 and 5.5.

Lemma B.1 (Eigenbasis solution)**.** *Let $E \in \mathbb{R}^{p\times p}$ be a diagonal matrix satisfying $E_{11} \le \cdots \le E_{pp}$. Then, for any $w \in \mathbb{R}^p$, we have*

$$\max_{V\in \mathrm{O}(p)} \; w^\top V E V^\top w = E_{pp}\|w\|_2^2.$$

Proof of Lemma B.1. The claim holds trivially when $w = 0$. Consider now any $w \in \mathbb{R}^p\backslash\{0\}$. Since $VEV^\top \preceq E_{pp}\cdot I_p$, we find

$$\max_{V\in \mathrm{O}(p)} \; w^\top V E V^\top w \le \max_{V\in \mathrm{O}(p)} \; w^\top V (E_{pp}\cdot I_p) V^\top w = E_{pp}\|w\|_2^2.$$

On the other hand, taking $V^\star = [v_1^\star, \ldots, v_p^\star] \in \mathrm{O}(p)$ with $v_p^\star = \frac{w}{\|w\|_2}$, and using the orthogonality of the columns of $V^\star$, we have

$$w^\top V^\star E {V^\star}^\top w = E_{pp}\|w\|_2^2.$$

This shows that $V^\star$ is an optimal solution and completes the proof. □

Lemma B.2 (Quadratic optimization)**.** *For any $x \in \mathbb{R}^p$, $\widehat{x} \in \mathbb{R}^p$ and $a \in \mathbb{R}_+$, the following assertions hold.*

- *Convex quadratic minimization:*

$$\min_{\mu\in\mathbb{R}^p:\|\mu-\widehat{x}\|_2^2=a^2} \; \|x-\mu\|_2^2 = (\|x-\widehat{x}\|_2 - a)^2,$$

  where the minimum is attained at $\mu^\star = \frac{a}{\|x-\widehat{x}\|_2}x + (1 - \frac{a}{\|x-\widehat{x}\|_2})\widehat{x}$.
- *Convex quadratic maximization:*

$$\max_{\mu\in\mathbb{R}^p:\|\mu-\widehat{x}\|_2^2=a^2} \; \|x-\mu\|_2^2 = (\|x-\widehat{x}\|_2 + a)^2,$$

  where the maximum is attained at $\mu^\star = -\frac{a}{\|x-\widehat{x}\|_2}x + (1 + \frac{a}{\|x-\widehat{x}\|_2})\widehat{x}$.

The results in Lemma B.2 are dispersed in the literature. An elementary proof is provided here for completeness.

Proof of Lemma B.2. By the triangle inequality, for any μ such that $\|\mu - \widehat{x}\|_2 = a$, we have

$$\|x-\mu\|_2 \ge |\|x-\widehat{x}\| - \|\mu-\widehat{x}\|| = |\|x-\widehat{x}\| - a|\,,$$

where the lower bound can be attained by taking $\mu = \frac{a}{\|x-\widehat{x}\|_2}x + (1 - \frac{a}{\|x-\widehat{x}\|_2})\widehat{x}$. Therefore,

$$\min_{\mu\in\mathbb{R}^p:\|\mu-\widehat{x}\|_2^2=a^2} \; \|x-\mu\|_2^2 = (\|x-\widehat{x}\|_2 - a)^2$$

Similarly, by the triangle inequality we have

$$\|x-\mu\|_2 \le \|x-\widehat{x}\| + \|\widehat{x}-\mu\| = \|x-\widehat{x}\| + a,$$

and the upper bound can be attained by $\mu = -\frac{a}{\|x-\widehat{x}\|_2}x + (1 + \frac{a}{\|x-\widehat{x}\|_2})\widehat{x}$. This completes the proof. □

Proposition B.3 (Logarithm maximization)**.** *For any $s, \sigma \geq 0$ and positive integer k, we have*

$$k \log\left(\sqrt{\frac{s}{k}} + \sigma\right) = \left\{ \begin{array}{ll} \max\limits_{e \in \mathbb{R}^k_+} & \sum\limits_{j=1}^{k} \log e_j \\ \text{s.t.} & \sum\limits_{j=1}^{k} (\sigma - e_j)^2 \leq s. \end{array} \right. \tag{7}$$

Moreover, the optimal solution $e^\star$ satisfies $e^\star_j = \sqrt{\frac{s}{k}} + \sigma$ for any $j = 1, \ldots, k$.

Proof of Proposition B.3. Let $e^\star \in \mathbb{R}^k_+$ be an optimal solution to the maximization problem (7). Suppose there exist two indices m and n such that $e^\star_m \neq e^\star_n$. Consider e' defined by

$$e'_j = \begin{cases} \frac{1}{2}(e^\star_m + e^\star_n), & \text{if } j \in \{m, n\}, \\ e^\star_j, & \text{otherwise.} \end{cases}$$

By the convexity of the function $x \mapsto (x - \sigma)^2$,

$$(e'_m - \sigma)^2 + (e'_n - \sigma)^2 = 2\left(\frac{e^\star_m + e^\star_n}{2} - \sigma\right)^2 \leq (e^\star_m - \sigma)^2 + (e^\star_n - \sigma)^2,$$

which implies that e' is a feasible solution to problem (7). Furthermore, since $e^\star_m \neq e^\star_n$, by the concavity of the function $x \mapsto \log x$, we have that

$$\log e^\star_m + \log e^\star_n < 2 \log\left(\frac{e^\star_m + e^\star_n}{2}\right) = \log e'_m + \log e'_n,$$

which violates the optimality of $e^\star$. Therefore, any optimal solution $e^\star$ must have all entries identical. Using this, we get from the constraint that

$$|e^\star_j - \sigma| \leq \sqrt{\frac{s}{k}} \quad \forall j = 1, \ldots, k.$$

By continuity of the objective and constraint functions, we must have

$$|e^\star_j - \sigma| = \sqrt{\frac{s}{k}} \quad \forall j = 1, \ldots, k.$$

Since the objective function is increasing in $e^\star_j$, the optimal solution is given by

$$e^\star_j = \sigma + \sqrt{\frac{s}{k}} \quad \forall j = 1, \ldots, k.$$

The optimal value can then be obtained by direct computation. This completes the proof. □

C FIRST-ORDER ALGORITHMS

C.1 OPTIMISTIC LIKELIHOOD PROBLEM

For the optimistic likelihood problem, Theorem 5.1 reduces the task to solving the 2-dimensional problem

$$\min_{\substack{a \in \mathbb{R}_+,\ d_p \in [\sigma, +\infty) \\ a^2 + (d_p - \sigma)^2 \leq \varepsilon_0^2}} \log d_p + \frac{(\|x - \widehat{x}_i\|_2 - a)^2}{2d_p^2} + (p-1)\log \sigma.$$

By letting

$$d_p = v_2 + \sigma, \quad \text{and} \quad a = v_1,$$

we can obtain the equivalent form

$$\min_{\substack{v_1, v_2 \geq 0 \\ v_1^2 + v_2^2 \leq \varepsilon_0^2}} F(v), \tag{8}$$

where the objective function is given by

$$F(v) = \log(v_2 + \sigma) + \frac{(\|x - \widehat{x}_i\|_2 - v_1)^2}{2(v_2 + \sigma)^2} + (p-1)\log\sigma.$$

If we denote by $\mathcal{V} = \{v \in \mathbb{R}^2 : v_1, v_2 \geq 0, v_1^2 + v_2^2 \leq \varepsilon_0^2\}$ the feasible region of the above minimization problem, then the projection $\mathrm{Proj}_{\mathcal{V}}(v)$ can be computed in closed-form via

$$\mathrm{Proj}_{\mathcal{V}}(v) = \begin{cases} v, & \text{if } v_1, v_2 \geq 0, v_1^2 + v_2^2 \leq \varepsilon_0^2, \\ \frac{\varepsilon_0}{\|v\|_2} v, & \text{if } v_1, v_2 \geq 0, v_1^2 + v_2^2 > \varepsilon_0^2, \\ (0, \varepsilon_0)^\top, & \text{if } v_1 < 0, v_2 > \varepsilon_0, \\ (0, v_2)^\top, & \text{if } v_1 < 0, 0 \leq v_2 \leq \varepsilon_0, \\ (\varepsilon_0, 0)^\top, & \text{if } v_1 > \varepsilon_0, v_2 < 0, \\ (v_1, 0)^\top, & \text{if } 0 \leq v_1 \leq \varepsilon_0, v_2 < 0, \\ (0, 0)^\top, & \text{if } v_1, v_2 < 0. \end{cases}$$

Algorithm 1 is a projected gradient descent routine to solve problem (8). The convergence guarantee for Algorithm 1 follows from Beck [2017, Theorem 10.15].

Algorithm 1 Projected Gradient Descent Algorithm with Backtracking Line-Search

Algorithm parameters: Line search parameters $\theta \in (0,1)$, $\beta > 0$
Initialization: Set $v^0 \leftarrow 0$
for $t = 0, 1, \ldots$ **do**
Find the smallest integer $k \geq 0$ such that

$$F\left(\mathrm{Proj}_{\mathcal{V}}(v^t - \theta^k \beta \nabla F(v^t))\right) \leq F(v^t) - \frac{1}{2\theta^k \beta}\|v^t - \mathrm{Proj}_{\mathcal{V}}(v^t - \theta^k \beta \nabla F(v^t))\|_2^2$$

Set $s^t = \theta^k \beta$ and set $v^{t+1} = \mathrm{Proj}_{\mathcal{V}}(u^t - s^t \nabla F(v^t))$.
end for

C.2 PESSIMISTIC LIKELIHOOD PROBLEM

For the pessimistic likelihood problem, Theorem 5.2 reduces the task to solving the 2-dimensional problem

$$\min_{\substack{a \in \mathbb{R}_+,\ d_1 \in [\sigma, +\infty) \\ a^2 + p(d_1 - \sigma)^2 \leq \varepsilon_1^2}} \left\{ -\log d_1 - \frac{1}{2d_1^2}(\|x - \widehat{x}_i\|_2 + a)^2 - (p-1)\log\left(\sigma + \sqrt{\frac{\varepsilon_1^2 - a^2 - (d_1 - \sigma)^2}{p-1}}\right)\right\}.$$

Note that the gradient of the objective function is a non-Lipschitz function. Worse still, the gradient is even undefined on at the feasible point $(d_1, a) = (\sigma, \varepsilon_1)$. These properties induce numerical issues for the optimization algorithm. Therefore, we solve the following perturbed problem instead:

$$\min_{\substack{a \in \mathbb{R}_+,\ d_1 \in [\sigma, +\infty) \\ a^2 + p(d_1 - \sigma)^2 \leq \varepsilon_1^2}} \left\{ -\log d_1 - \frac{1}{2d_1^2}(\|x - \widehat{\mu}\|_2 + a)^2 - (p-1)\log\left(\sigma + \sqrt{\frac{\zeta + \varepsilon_1^2 - a^2 - (d_1 - \sigma)^2}{p-1}}\right)\right\}, \tag{9}$$

for some small $\zeta > 0$. By Bonnans and Shapiro [2013, Proposition 4.4], the optimal value of problem (9) is continuous in ζ and the optimal solution set is upper semi-continuous in ζ as a set-valued mapping, see Bonnans and Shapiro [2013, Section 4.1].

We now derive a projected gradient descent algorithm with backtracking line search for solving problem (9). First, by letting

$$d_1 = u_2 + \sigma, \quad \text{and} \quad a = \sqrt{p} u_1,$$

we can equivalently transform problem (9) to the following one:

$$\min_{\substack{u_1,u_2\geq 0\\ u_1^2+u_2^2\leq(\varepsilon_1/\sqrt{p})^2}} F(u), \tag{10}$$

where the objective function is given by

$$F(u) = -\log(u_2+\sigma) - \frac{1}{2(u_2+\sigma)^2}(\|x-\widehat{x}_i\|_2+\sqrt{p}u_1)^2 - (p-1)\log\left(\sigma+\sqrt{\frac{\zeta+\varepsilon_1^2-pu_1^2-u_2^2}{p-1}}\right).$$

The upshot of problem (10) is that the feasible region is the intersection of the non-negative orthant with a circular disk of radius $\varepsilon_1/\sqrt{p}$ centered at the origin. As we will see below, this enables easy computation of the projection and linear optimization oracle. Indeed, denoting by $\mathcal{U} = \{u \in \mathbb{R}^2 : u_1, u_2 \geq 0, u_1^2+u_2^2 \leq (\varepsilon_1/\sqrt{p})^2\}$ the feasible region of problem (10), the projection $\mathrm{Proj}_{\mathcal{U}}(u)$ can be computed in closed-form via

$$\mathrm{Proj}_{\mathcal{U}}(u) = \begin{cases} u, & \text{if } u_1,u_2\geq 0, u_1^2+u_2^2\leq(\varepsilon_1/\sqrt{p})^2,\\ \frac{(\varepsilon_1/\sqrt{p})}{\|u\|_2}u, & \text{if } u_1,u_2\geq 0, u_1^2+u_2^2>(\varepsilon_1/\sqrt{p})^2,\\ (0,\frac{\varepsilon_1}{\sqrt{p}})^\top, & \text{if } u_1<0, u_2>\frac{\varepsilon_1}{\sqrt{p}},\\ (0,u_2)^\top, & \text{if } u_1<0, 0\leq u_2\leq\frac{\varepsilon_1}{\sqrt{p}},\\ (\frac{\varepsilon_1}{\sqrt{p}},0)^\top, & \text{if } u_1>\frac{\varepsilon_1}{\sqrt{p}}, u_2<0,\\ (u_1,0)^\top, & \text{if } 0\leq u_1\leq\frac{\varepsilon_1}{\sqrt{p}}, u_2<0,\\ (0,0)^\top, & \text{if } u_1,u_2<0. \end{cases}$$

A projected gradient descent algorithm can now be employed to solve problem (10).

D RECOVERY OF THE ADVERSARIAL DISTRIBUTION

It is often instructive to recover and analyze the optimal distribution that maximizes the posterior probability odds ratio, or more directly, the likelihood ratio in (2). Equivalent, it suffices to characterize the distribution $\mathbb{Q}_0^\star$ that maximizes (3), and the distribution $\mathbb{Q}_1^\star$ that minimizes (4).

Lemma D.1 (Likelihood maximizer)**.** *For each $i \in \mathcal{I}_0$, let $(a_i^\star, d_{pi}^\star)$ be the optimal solution of the following two-dimensional optimization problem*

$$\min_{\substack{a\in\mathbb{R}_+,\ d_p\in[\sigma,+\infty)\\ a^2+(d_p-\sigma)^2\leq\varepsilon_0^2}} \log d_p + \frac{(\|x-\widehat{x}_i\|_2-a)^2}{2d_p^2} + (p-1)\log\sigma.$$

Then, the maximizero $\mathbb{Q}_0^\star$ of problem (3) *is a Gaussian mixture with N_0 components, and for $i \in \mathcal{I}_0$, the i-th components has mean*

$$\mu_i^\star = \frac{a_i^\star}{\|x-\widehat{x}_i\|_2}x + \left(1-\frac{a_i^\star}{\|x-\widehat{x}_i\|_2}\right)\widehat{x}_i,$$

and covariance matrix

$$\Sigma_i^\star = V_i^\star\,\mathrm{diag}(\sigma,\ldots,\sigma,d_{pi}^\star)^2(V_i^\star)^\top,$$

where $V_i^\star$ is any orthogonal matrix with the p-th column given by $\frac{x-\mu_i^\star}{\|x-\mu_i^\star\|_2}$.

Proof of Lemma D.1. The result follows directly by inspecting the proofs of Proposition 5.4, Lemma B.1 and Lemma B.2. □

Lemma D.2 (Likelihood minimizer)**.** *For each $i \in \mathcal{I}_1$, let $(a_i^\star, d_{1i}^\star)$ be the optimal solution of the following two-dimensional optimization problem*

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in[\sigma,+\infty)\\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}} \left\{-\log d_1 - \frac{(\|x-\widehat{x}_i\|_2+a)^2}{2d_1^2} - (p-1)\log\left(\sigma+\sqrt{\frac{\varepsilon_1^2-a^2-(d_1-\sigma)^2}{p-1}}\right)\right\}.$$

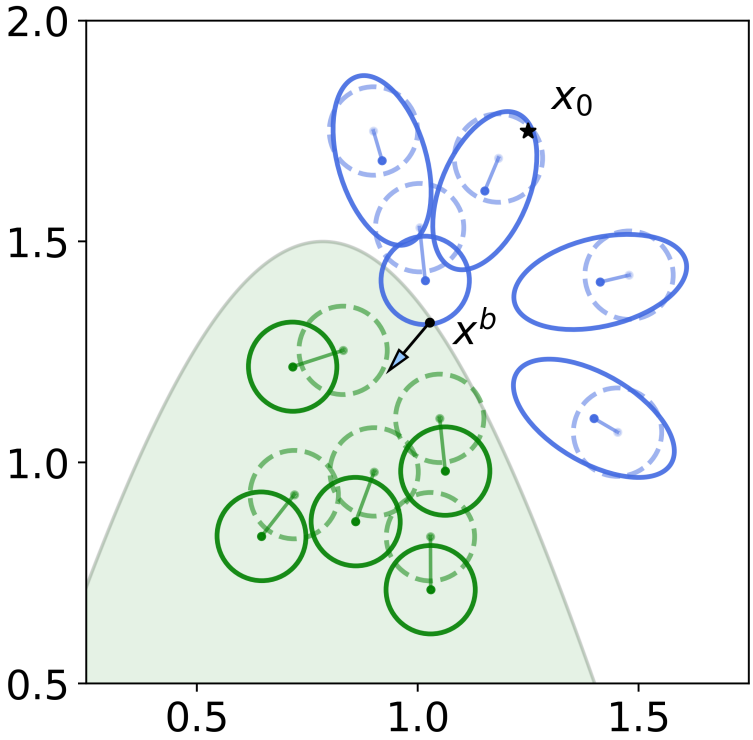

Figure 3: Visualization of the worst-case distributions on a toy dataset, color codes are similar to Figure 1. The dashed, opaque dots and circles represent the isotropic Gaussian around each data sample. The solid dots and circles represent the worst-case distributions corresponding to the boundary point x^b. For blue (unfavorably predicted) samples, the worst-case distribution is formed by perturbing the distribution towards x^b – which leads to maximizing the posterior probability of unfavorable prediction. For green (favorably predicted) samples, the worst-case distribution is formed by perturbing the distribution away from x^b – which leads to minimizing the posterior probability of favorable prediction. These worst-case distributions will maximize the posterior probability odds ratio.

Then, the minimizer $\mathbb{Q}_1^\star$ of problem (4) *is a Gaussian mixture with N_1 components, and for $i \in \mathcal{I}_1$, the i-th components has mean*

$$\mu_i^\star = -\frac{a_i^\star}{\|x - \widehat{x}_i\|_2} x + \left(1 + \frac{a_i^\star}{\|x - \widehat{x}_i\|_2}\right) \widehat{x}_i,$$

and covariance matrix

$$\Sigma_i^\star = V_i^\star \operatorname{diag}\left(d_{1i}^\star, \sigma + \sqrt{\frac{\varepsilon_1^2 - a_i^{\star 2} - (d_{1i}^\star - \sigma)^2}{p-1}}, \ldots, \sigma + \sqrt{\frac{\varepsilon_1^2 - a_i^{\star 2} - (d_{1i}^\star - \sigma)^2}{p-1}}\right)^2 (V_i^\star)^\top,$$

where $V_i^\star$ is any orthogonal matrix with the 1*st column given by* $\frac{x-\mu_i^\star}{\|x-\mu_i^\star\|_2}$.

Proof of Lemma D.2. The result follows directly by inspecting the proofs of Proposition 5.5, Lemma B.1 and Lemma B.2. □